






Universitat Autònoma de Barcelona

ADVERTIMENT. L'accés als continguts d'aquesta tesi queda condicionat a l'acceptació de les condicions d'ús establertes per la següent llicència Creative Commons:  http://cat.creativecommons.org/?page_id=184

ADVERTENCIA. El acceso a los contenidos de esta tesis queda condicionado a la aceptación de las condiciones de uso establecidas por la siguiente licencia Creative Commons:  <http://es.creativecommons.org/blog/licencias/>

WARNING. The access to the contents of this doctoral thesis it is limited to the acceptance of the use conditions set by the following Creative Commons license:  <https://creativecommons.org/licenses/?lang=en>

Universitat Autònoma de Barcelona

Facultat de Biociències

Departament de Biologia Animal, Biologia Vegetal i Ecologia

Programa de Doctorat en Biologia i Biotecnologia Vegetal

TESIS DOCTORAL

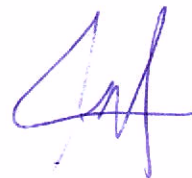
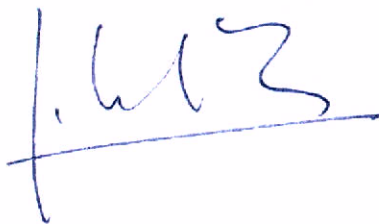
**DYNAMICS AND IMPACT OF MOBILE GENETIC ELEMENTS IN
THE MOSS *PHYSCOMITRIUM PATENS***

Memòria presentada per Pol Vendrell Mir per optar al títol de doctor per la Universitat Autònoma de Barcelona

Treball realitzat en el Programa de Genòmica de Plantes i Animals del Centre de Recerca en Agrigenòmica (CRAG), Campus UAB Bellaterra, sota la direcció del Doctor Josep M^a Casacuberta Suñer

El director i tutor de la tesis:

El candidat a doctor:



Dr. Josep M^a Casacuberta Suñer

Pol Vendrell Mir

Cerdanyola del Vallès, 2022

Als Pares, a l'Arnau i a l'Eva

INDEX

RESUM	I
RESUMEN	II
SUMMARY	III
LIST OF FIGURES	IV
LIST OF TABLES	VII
PREFACE.....	1
GENERAL INTRODUCTION	5
CHAPTER 1: DEVELOPMENT OF BIOINFORMATIC TOOLS TO IDENTIFY TRANSPOSABLE ELEMENTS MOBILIZATION AND TRANSCRIPTION	25
Chapter 1.1: Introduction.....	25
Chapter 1.2. Objectives.....	28
Chapter 1.3: Comparison of different available tools to detect Transposon Insertion polymorphisms using short-read data	29
Chapter 1.4: Detection of Transposable Element transcription from short-read data	54
CHAPTER 2: TEs DYNAMICS IN <i>PHYSCOMITRIUM PATENS</i>	77
Chapter 2.1: Introduction.....	77
Chapter 2.2: Objectives.....	80
Chapter 2.3: Expression and mobilization of TEs in <i>Physcomitrium patens</i>	81
Chapter 2.4: Complementary results and discussion	95
CHAPTER 3: IMPACT OF TEs IN <i>PHYSCOMITRIUM PATENS</i> GENOME.....	99
Chapter 3.1: Introduction.....	99
Chapter 3.2: Objectives	100
Chapter 3.3: Impact of Transposable Elements in the structure of <i>Physcomitrium patens genome</i>	101
Chapter 3.4: Impact of Transposable Elements in genic regions.....	125
CHAPTER 4: AN ENDOGENOUS VIRUS IN THE MOSS <i>PHYSCOMITRIUM PATENS</i>	157
Chapter 4.1: Introduction.....	157
Chapter 4.2: Objectives.....	160
Chapter 4.3: A vertically transmitted amalgavirus is present in certain accessions of the bryophyte <i>Physcomitrium patens</i>	161
Chapter 4.4: Complementary results and discussion	174
GENERAL DISCUSSION AND FUTURE PRESPECTIVES	179
General discussion	179

Future perspectives	182
CONCLUSIONS	191
MATERIALS AND METHODS	195
Chapter 1: Development of bioinformatic tools to identify transposable elements mobilization and transcription	195
Chapter 2: TEs dynamics in <i>Physcomitrium patens</i>	198
Chapter 3: Impact of TEs in <i>Physcomitrium patens</i> genome	199
Chapter 4: An endogenous virus in the moss <i>Physcomitrium patens</i>	215
BIBLIOGRAPHY	219
ACKNOWLEDGMENTS	237
ANNEXES	241

RESUM

Els elements genètics mòbils son material genètic amb la capacitat de moure's en el genoma o, en alguns casos, entre diferents organismes o cèl·lules. Podem distingir dues classes d'elements genètics mòbils, els virus, que tenen la capacitat de transferir el seu material gènic entre organismes, i els transposons, que es mouen i es repliquen dins el genoma de l'hoste. Els transposons ocupen una fracció important dels genomes eucariotes i mitjançant el seu moviment poden alterar-ne la seva estructura, tenint un paper clau en l'evolució dels genomes. En aquesta tesi s'ha estudiat la dinàmica i l'impacte en *Physcomitrium patens* de diferents elements mòbils.

El primer capítol es centra en l'anàlisi i la comparació de diferents metodologies per tal de detectar la transcripció i mobilització dels transposons utilitzant dades de seqüenciació basades en tecnologies de *short-reads*. Aquesta anàlisi ens va permetre definir les metodologies que millor s'adapten als objectius d'aquesta tesi. En el segon capítol es van emprar les metodologies seleccionades amb la finalitat d'estudiar la dinàmica dels transposons en el genoma de *Physcomitrium patens*, detectant varies famílies de retrotransposons i transposons d'ADN que son transcripcionalment actives i que en alguns casos son polimòrfiques a la població. En el tercer capítol es descriu l'anàlisi de l'impacte dels transposons, i en concret d'un retrotransposó amb LTRs anomenat RLG1, tant en l'estructura del genoma com en els gens de *Physcomitrium patens*, eliminant o, en algun cas, introduint de nou elements RLG1 en llocs concrets del genoma. Finalment, el darrer capítol es centra en la detecció i dinàmica del primer virus descrit a *P. patens*. Aquest virus, al que hem anomenat *Physcomitrium patens Amalgavirus 1* (PPAV1) és un virus endogen present només en algunes accessions de *P. patens* i que es transmet verticalment, tant per la línia paterna com materna.

RESUMEN

Los elementos genéticos móviles son material genético con la capacidad de moverse en el genoma o, en algunos casos, entre diferentes organismos o células. Pueden distinguirse dos clases de elementos genéticos móviles, los virus, que tienen la capacidad de transferir su material génico entre organismos, y los transposones, que se mueven y replican dentro del genoma del organismo huésped. Los transposones ocupan una fracción importante de los genomas eucariotas y mediante su movimiento pueden alterar su estructura, teniendo un papel clave en la evolución de los genomas. En esta tesis se ha estudiado la dinámica y el impacto de *Physcomitrium patens* de diferentes elementos móviles.

El primer capítulo se centra en el análisis y la comparación de diferentes metodologías para detectar la transcripción y movilización de los transposones utilizando datos de secuenciación basados en tecnologías de *short-reads*. Este análisis nos permitió definir las metodologías que mejor se adaptan a los objetivos de esta tesis. En el segundo capítulo se utilizaron estas metodologías seleccionadas con la finalidad de estudiar la dinámica de los transposones en el genoma de *Physcomitrium patens*, detectando varias familias de retrotransposones y transposones de ADN que son transcripcionalmente activas y que en algunos casos son polimórficas en la población. En el tercer capítulo se describe el análisis del impacto de los transposones, y en concreto de un retrotransposón con LTRs nombrado RLG1, tanto en la estructura del genoma como en los genes de *Physcomitrium patens*, eliminando o en algunos casos, introduciendo de nuevo elementos RLG1 en lugares concretos del genoma. Finalmente, el último capítulo se centra en la detección y dinámica del primer virus descrito en *P. patens*. Este virus, que hemos nombrado *Physcomitrium patens Amalgavirus 1* (PPAV1) es un virus endógeno presente únicamente en algunas accesiones de *P. patens* y que se transmite verticalmente, tanto por la línea paterna como materna.

SUMMARY

Mobile genetic elements are genetic material with the ability to move within the genome or, in some cases, between different organisms or cells. Two classes of mobile genetic elements can be distinguished, viruses, which have the ability to transfer their genetic material between organisms, and transposons, which move and replicate within the genome of the host organism. Transposons occupy an important fraction of eukaryotic genomes and through their movement can alter their structure, playing a key role in the evolution of genomes. In this thesis we have studied the dynamics and impact of *Physcomitrium patens* of different mobile elements.

The first chapter focuses on the analysis and comparison of different methodologies to detect transcription and mobilization of transposons using sequencing data based on *short-reads* technologies. This analysis allowed us to define the methodologies that best fit the objectives of this thesis. In the second chapter, these selected methodologies were used to study transposon dynamics in the genome of *Physcomitrium patens*, detecting several families of retrotransposons and DNA transposons that are transcriptionally active and, in some cases, polymorphic in the population. The third chapter describes the analysis of the impact of transposons, and in particular of a retrotransposon with LTRs named RLG1, on both the genome structure and genes of *Physcomitrium patens*, eliminating or in some cases, reintroducing RLG1 elements at specific locations in the genome. Finally, the last chapter focuses on the detection and dynamics of the first virus described in *P. patens*. This virus, which we have named *Physcomitrium patens* Amalgavirus 1 (PPAV1), is an endogenous virus present only in some *P. patens* accessions and transmitted vertically, both from the paternal and maternal lines.

LIST OF FIGURES

Figure 1: Main mechanisms of mobilization of the two classes of Transposable Elements	7
Figure 2: Main classes of transposable elements and different orders of Transposable Elements	8
Figure 3: Genomic size variation across different species of all plant divisions and the percentage of TEs with respect to the total of genomic for each species.....	12
Figure 4: Different examples of impacts introduced by Transposable Elements on a genome..	13
Figure 5: Different examples of phenotypes in different crops caused by polymorphic TEs altering gene expression.....	15
Figure 6: Phylogenetic tree of Land plants with a time tree at a scale based on phylogenetic analysis and fossil data.....	17
Figure 7: Life cycle of <i>P. patens</i>	19
Figure 8: Scheme of the strategies used to detect TIPs from short-reads data	29
Figure 9: Example of different orders of TE transcription.....	55
Figure 10: Expression of the different LTR-RT families detected by TETRanscripts	59
Figure 11: Phylogenetic tree of the different LTR-RTs clusters of the Copia superfamily	62
Figure 12: Phylogenetic tree of the different LTR-RT clusters of the Gypsy superfamily	62
Figure 13: Expression of the different LTR-RT clusters detected by TETRanscripts.....	63
Figure 14: Summary of the approaches used to detect TE expression using TETRanscripts.....	65
Figure 15: Expression of the different LTR-RT clusters detected by TETools	66
Figure 16: Assembly of Tandem repetitions of LTR-RT using the de novo assembly approach.....	69
Figure 17: Expression detected for the different families using the assembly approach	70
Figure 18: Example of the distribution of the RLG1 islands in a chromosome, in this case in the chromosome 27 of <i>P. patens</i>	101
Figure 19: Strategy followed to Introduce of the cassette containing the CRISPR/Cas9 system targeting the RLG1 elements to the genome.	106
Figure 20: Agarose gel of the PCR products of 5 different RLG1 elements located at 5 different chromosomes.	108
Figure 21: RLG1 quantification compared to the single copy gene APT for Gransden Wt compared to two clones that have been transiently transformed with the CRISPR/Cas9 targeting the RLG1 elements.	109
Figure 22: Use of a DNA template to replace the RLG1 islands that have been cut by the CRISPR/Cas9 system.	110
Figure 23: Explanation of the APT system to estimate the efficiency of edition in <i>P. patens</i>	111
Figure 24: Phenotypic differences between the clone 6 transformed with the CRISPR/Cas9 system targeting the APT+ the RLG1 elements (right) compared to the clone G2.4 that has only been edited the APT gene.....	112

Figure 25: Relative quantification of the Number of RLG1 elements compared to the Pp3c10_20470V3.1 gene for the different samples that have been transformed with the CRISPR/Cas9 targeting the RLG1 and the APT gene compared to Gransden Wt and to G2.4 (APT KO)	113
Figure 26: A) Chromosome 27 with all the RLG1 islands, heterochromatic regions in blue. B) In blue 15 biggest RLG1 islands of chromosome 27 ranging in size from 40 kb to 160 kb.	115
Figure 27: Replacement of Island D and Island O for two recombination templates by homologous recombination	117
Figure 28: Selected clones of the Island D (D) and island O (O) and both replacements (D/O) selected for sanger sequencing.	119
Figure 29: In top a scheme the two main hypothesis of integration of the constructs into the locus D. At the bottom PCRs using internal primers to verify the presence of the constructs.	119
Figure 30: Relative expression of the genes Pp3c27_3930V3.1 and Pp3c27_3970V3.1 compared to the 60S ribosomal protein expression in protonemata of 7 days old.	121
Figure 31: Distribution of the different <i>P. patens</i> accessions that have been resequenced using short-reads.	126
Figure 32: Polymorphic TE insertion predicted in Reute inside the eighth exon of the gene Pp3c17_3870V3.1 compared to the gene structure in Gransden.	131
Figure 33: Expression of the gene Pp3c17_3870V3.1 during the development of <i>P. patens</i> image extracted from the transcriptome atlas (Ortiz-Ramírez et al., 2016) that can be accessed through the following webpage http://bar.utoronto.ca/efp_physcomitrella/cgi-bin/efpWeb.cgi	132
Figure 34: Alignment of RNA-seq short reads to the gene in Gransden (upper part) and Reute (bottom part). We observe the absence of reads at the end of the 3' CDS in reute while in Gransden we detect reads covering all the transcript.	133
Figure 35: Polymorphic TE insertion predicted in Villersexel at the end of the 3' UTR of the gene Pp3c4_24710V3.1 compared to the gene structure in Gransden.	133
Figure 36: Expression of the gene Pp3c4_24710V3.1 during the development of <i>P. patens</i>	134
Figure 37: Verification of the RLG1 insertion in the Pp3c7_24710V3.1 by PCR amplification.	135
Figure 38: Relative expression of the gene Pp3c4_24710V3.1 between Gransden and Villersexel compared to the housekeeping gene APT in protonemata and in gametophores.	136
Figure 39: Relative expression of the gene Pp3c4_24710V3.1 between Gransden and Villersexel compared to the housekeeping gene APT in protoplasts.	136
Figure 40: Independent colonies of Gransden, Pp3c4_24710V3.1 (Mutants lines 26, 29 and 30) and Villersexel after 30 days of growth in BCD.	137
Figure 41: Growth differences between Gransden Wt and Villersexel Wt in BCD medium under different concentrations of ABA (0 µg/L, 1 µg/L and 2 µg/L).	138
Figure 42: Polymorphic RLG1 insertion predicted in Kaskaskia 229 bp of the 5' UTR of the gene Pp3c14_9040V3.1 and at 400 bp of the CDS sequence, compared to the gene structure in Gransden	138
Figure 43: Phylogenetic tree of the representative sequences of representative sequences of all the plant kingdom.	139

Figure 44: Expression of the gene Pp3c14_9040V3.1 during the development of <i>P. patens</i> . Observing a high expression in the spores.....	140
Figure 45: Gene expression over the development of <i>Arabidopsis thaliana</i> (Klepikova et al., 2016) and <i>Medicago truncatula</i> (Benedito et al., 2008) homolog genes of Pp3c14_9040V3.1.....	142
Figure 46: PCR amplification of the locus corresponding to the polymorphism in Gransden and in Kaskaskia.....	143
Figure 47: Expression of the gene Pp3c14_9040V3.1 when compared to the housekeeping APT (Top) and the 60S ribosomal protein (Bottom).	144
Figure 48: Scheme of the replacement strategy used to swap the RLG1 TE between Gransden and Kaskaskia.....	146
Figure 49: Verification of the RLG1 replacement in the 7 obtained clones of Kaskaskia (top) and the insertion of the TE at the expected locus in transformed clones of Gransden (bottom).....	147
Figure 50: Verification by PCR of the presence of integration of multiple copies of the RLG1 insertion in Gransden.....	148
Figure 51: Relative expression of the gene Pp3c14_9040V3.1 compared to the housekeeping genes APT (top) and 60S Ribosomal Protein (bottom) of protonemata 7 days old.....	150
Figure 52: Relative expression of the Pp3c14_9040V3.1 gene when compared to the housekeeping gene 60S Ribosomal Protein, during 14 days of development of the protonemata,	151
Figure 53: Images of the different clones after 21 days of growth in BCDA medium.	151
Figure 54: At the top structure of the PPAV1 wt virus, where the two blue boxes represent the two Open Reading Frames (ORF) of the virus. After this, the two different constructs that were developed in the lab. In the mid of the figure, the construct that we modified that had slight changes in the nucleotide sequence located at the second ORF that does not lead to any aminoacid change At the bottom the second construct that we developed with a fusion of a GFP at the end of the second ORF in the same frame. ...	175
Figure 55: Phylogenetic tree of the different LTR-RT of the copia family of rice	186
Figure 56: Protein alignment of the Integrase and Reverse transcriptase of two Copia families of LTR-RTs (R1418 and P116) that are closely related.	186
Figure 57: Replacement of the gRNA sequence by using two mutagenic primers.....	204

LIST OF TABLES

Table 1: Differential expressed LTR-RT families detected between protonemata samples treated with heat shock and protonemata control samples expressed in Log ₂ Fold Change (Log ₂ FC) values detected by Tetranscripts.....	60
Table 2: Total number of TEs by each TE family compared to the number of LTR-RT that have a minimum length of 3 Kbp.	60
Table 3: Differential expressed LTR-RT clusters detected between protonemata samples treated with heat shock and protonemata control samples expressed in Log ₂ Fold Change (Log ₂ FC) values detected by Tetranscripts.....	63
Table 4: Differential expressed clusters detected between protonemata samples treated with heat shock and protonemata control samples expressed in Log ₂ Fold Change (Log ₂ FC) values detected by Tertools.	67
Table 5: Differential expressed clusters detected between protonemata samples treated with heat shock and protonemata control samples expressed in Log ₂ Fold Change (Log ₂ FC) values detected by the TE assembly approach.	70
Table 6: Sequences selected to produce elimination of RLG1 elements all over the genome filtered with a minimum of 2500 cuts over the genome.	104
Table 7: Prediction by CRISPOR of the efficiency of the designed guide sequence..	105
Table 8: Results of the transformations comparing the regeneration transforming with the CRISPR/Cas9 targeting the APT gene compared to transforming with the CRISPR/Cas9 targeting the APT and the RLG1 elements at the same time.....	111
Table 9: Selected RLG1 islands from chromosome 27 with the position, number of RLG1 elements, bp length and the name given to the island.	116
Table 10: Results of the transformations done to replace the Island D, island O or both islands at the same time.....	118
Table 11: Number of Genes with a TIPs in the three accessions Reute, Kaskaskia and Villersexel as compared to the Gransden Reference genome.	127
Table 12: Selected TE Polymorphisms insertions between the different accessions selected based on the previous criteria.	128
Table 13: Type of mutation for the 39 sequenced clones of the CRISPR/Cas9 transformation targeting the Pp3c14_9040V3.1 gene. In the third and sixth column, length of the nucleotide insertion or deletion and/or sequence that has been substituted.	145
Table 14: Read counts and Transcripts per Million (TPM) normalized counts of the different RNAseq libraries mapped to the Amalgavirus sequence of juvenile gametophores with libraries produced with a polyA purification and a total RNA purification (Total RNA).	176
Table 15: Number of clones obtained by transforming with the miniTnt1 system combined with the Wt proteins of Tnt1 or the proteins with the mutations over the catalytic domain of the integrase of Tnt1 in <i>A. thaliana</i> and on <i>P. patens</i>	184
Table 16: Number of clones obtained by transforming with the miniTnt1 system combined with the Wt proteins of Tnt1 or the proteins with the mutations over the catalytic domain of the integrase of Tnt1 in <i>Wt P. patens</i> and in Δ Rad51 <i>P. patens</i> ...	185
Table 17: Primers used to genotype different RLG1	202

Table 18: gRNAs used to eliminate RLG1 islands of chromosome 27	203
Table 19: Primers used to check the modifications over the selected RLG1 islands of chromosome 27	206
Table 20: Primers used to do the qRT PCRs over the genes Pp3c27_3930V3.1 and Pp3c27_3970V3.1	207
Table 21: Primers used to confirm the insertion over the different polymorphic TEs Pp3c14_9040V3.1 and Pp3c7_24710V3.1 and to genotype the KOs and the TE replacement in the different <i>P. patens</i> accessions.	211
Table 22: Primers used to perform the qRT-PCRs over the genes Pp3c14_9040V3.1 and Pp3c7_24710V3.1.	212
Table 23: gRNAs and cDNAs used to perform the KOs and the RLG1 replacement in the different accessions.....	213

PREFACE

PREFACE

A living being is usually defined as an organism with the capacity to reproduce, interact with the environment, grow, adapt, and maintain homeostasis. All living organisms have a genome with the genetic information needed for their development and function. But where are the limits of life?

In a gray area between what we classify as living or non-living beings we find viruses. A cellular infectious agent that cannot reproduce by itself but that reproduces when infects hosts, living cells of other organisms. Viruses do not have the ability to maintain homeostasis but require the organisms they infect to maintain it; they do not have their own metabolism but require the metabolism of the host organism in order to propagate. However, they can adapt to the environment and contain genetic material.

And it is in this gray zone that we also find Transposable Elements (TEs). TEs are sequences of DNA found in virtually every living organism's genome. In many organisms they represent an important fraction of the genomes not occupied by genes. For example, in the human genome they represent 47.6% of the genome while genes occupy only 2% of the genome (Hoyt et al., 2022). These sequences have the capacity to mobilize and propagate, behaving similarly to viruses. However, unlike viruses that are mostly transmitted horizontally between different living organisms, TEs restrict their movement to the genome of their host, being transmitted vertically from the parental lines to the progeny.

TEs and viruses are also called mobile genetic elements. The study of these elements allows us to study the evolutionary processes of living beings from a different perspective. These elements through their interaction with their host organism cause alterations, modifying their host and therefore represent an engine of evolution of all living beings as proposed by Barbara Mc.Clintock.

In this thesis dissertation I have focused on the study of transposons and viruses in the model plant *Physcomitrium patens* to better understand their evolution and the impact that they have on this organism.

GENERAL INTRODUCTION

GENERAL INTRODUCTION

An introduction to Transposable Elements

The genome is the set of all the genetic information of an organism that provides all the information required for the organism to function. The analysis of genomes is therefore a key area of research in biology. Although an important amount of genome data was already available before, the publication of the first drafts of the *Arabidopsis thaliana* genome in 2000 (Kaul et al., 2000) and the human genome in 2001 (Lander et al., 2001) represented a tremendous breakthrough in the area. These were non-continuous draft genomes containing most of the gene information but lacking an important fraction of repetitive sequences. The centromeres of *Arabidopsis thaliana* were not completely sequenced until very recently (Naish et al., 2021) and the first human genome almost completed from telomere to telomere was released this year (Nurk et al., 2022). Both cases illustrate the difficulty of characterizing the repeated fraction of the genome and suggests that there is still a lot to be learned about these regions in the future.

We should consider that in the case of the human genome only 2% of the genome corresponds to genes. The rest is formed by a mixture of sequences that do not encode for any gene product and mostly repetitive sequences, that due to their repetitive nature are difficult to properly assemble in a genome. In most eukaryotic genomes we see a pattern similar to that of the human genome, with only a small part of the genome being composed of genes and most of the genomes being formed by repetitive sequences (Aurélien et al., 2017; Civián et al., 2011).

The most abundant fraction in these repetitive sequences is comprised by TEs, that in the case of the human genome account for a 47.6% of all the genome (Hoyt et al., 2022).

TEs are mobile genetic elements that can change their location within the genome that, in some cases, generate new copies of themselves. This process known as transposition is an important source of genetic variability and can induce changes in the structure and size of the genomes. For this reason, TEs are considered key contributors of genome evolutionary processes (Kidwell & Lisch, 2000). TEs can be found in virtually all

organisms. Their size can vary from a few hundred nucleotides to more than 12000 base pairs (bp) (Wicker et al., 2009).

The discovery of Transposable elements

TEs were first described by Barbara McClintock during the 1940s and 1950s through the study of the variation of color pattern of *Zea mays* kernels. She discovered that the differences in the color of the kernel of Maize that she observed were caused by a chromosome breakage event in a particular locus of chromosome 9 that she named the dissociation (Ds) locus (Mcclintock, 1940). During her study she discovered that this locus could change its position within the chromosome and that for this to occur a second dominant locus present in a different site or chromosome, was needed. She named this second locus Activator (Ac) (Mcclintock, 1950). Through this work she described the first transposable elements to be discovered (Ac and Dc elements), that the movement of TEs could lead to phenotypic events and that the movement of these TEs could restore the function of the mutated genes.

It was not until the 1980s decade that after cloning the Ac and Ds TEs it was discovered that Ac is an autonomous element encoding for a transposase that could mobilize itself and the Ds elements, non-autonomous derivatives of the Ac elements. In 1983 the Norwegian Nobel Committee recognized her work and awarded her with the Nobel Prize in Physiology or Medicine for her discovery of mobile genetic elements.

Transposable Elements classification

TEs are structurally and functionally diverse. There have been different proposals to classify them according to their mechanism of mobilization and structure.

Through this work we will follow the classification first introduced by David Finnegan (Finnegan, 1989) and further refined by Wicker et al., (2007) that was published with the goal to standardize transposable element annotation.

TEs are divided in two major classes based on their transposition mechanism (Figure 1): Class I, or retrotransposons, are TEs that transpose through an RNA intermediate that is retrotranscribed and integrated into a new place of the genome, leading to the generation of a new copy. As the original copy remains at its location and a new copy is generated,

this mechanism is also referred to as *copy and paste*. Class II or DNA transposons are TEs that transpose without an RNA intermediate. Most of the DNA transposons are excised from their original place and integrated into a new place of the genome by a mechanism also referred to as *cut and paste* (Figure 1). Although the *cut and paste* mechanism is the most frequent in class II TEs, there are other mechanisms such as the rolling circle mechanism used by Helitrons to transpose (Kapitonov & Jurka, 2007). Class I transposons increase the number of copies on the genome through their replicative mechanism of movement while Class II transposons, as it is the same copy that is mobilized, the number of TE copies is usually maintained.

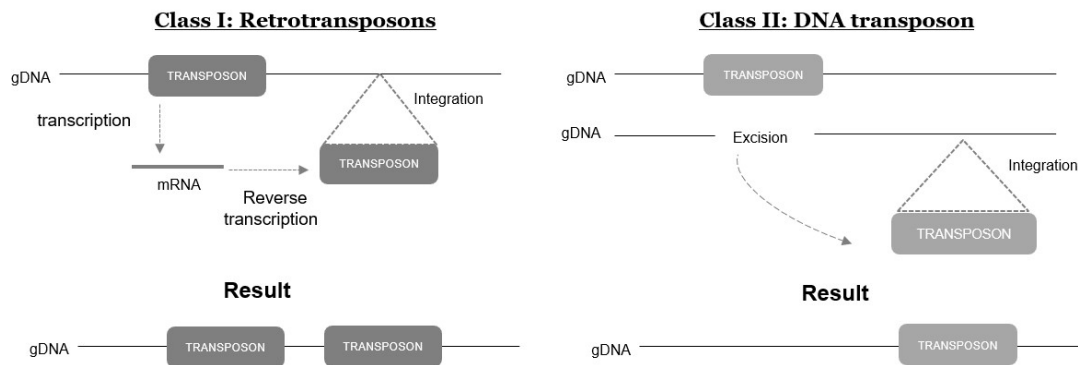
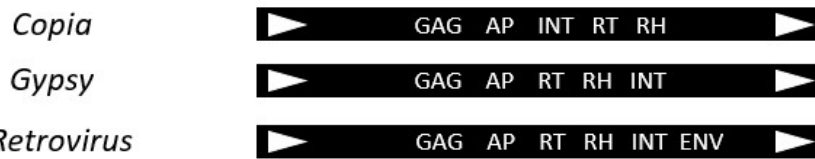


Figure 1: Main mechanisms of mobilization of the two classes of Transposable Elements. Class I or retrotransposons are transcribed, reverse transcribed and integrated into a new place of the genome originating a new TE copy. Class II transposons are excised by the encoded transposase and integrated into a new place of the genome.

Each TE class can be further categorized into superfamilies and families according to their encoded proteins and their non-coding regions (Figure 2). Moreover, as it was shown for the Ac and Ds TEs, TEs can be mobilized through the proteins encoded by their own coding sequences, and are then named as autonomous TEs, or by the proteins encoded by other related TEs, and are then named as non-autonomous TEs.

Class I: Retrotransposons

LTR-Retrotransposons:



non LTR-Retrotransposons :



Class II: DNA transposons:

Subclass I:



Subclass II



Figure 2: Main classes of transposable elements and different orders of Transposable Elements. Arrows indicate the long terminal repeats of LTR when are in the same direction. In opposite direction indicates the tandem inverted repeats of TIR transposons. Gag indicates the protein Gag forming the virus-like particle of the LTR-RTs. AP the aspartyl protease, INT the integrase, RT the reverse transcriptase, RH the RNaseH and ENV the Envelope protein. In LINEs EN indicates the endonuclease, RT the reverse transcriptase and in the case of SINEs the grey box indicates the tRNA union side. In the case of Helitrons RpA indicates the replication protein A and HEL the helicase. Based on (Casacuberta & Santiago, 2003; Wicker et al., 2009).

Class I: Retrotransposons

Class I retrotransposons can be divided in two major orders (Wicker et al., 2009): the LTR-retrotransposons (LTR-RT) and all the others, named as non-LTR retrotransposons.

LTR-RT are Class I TEs whose coding sequence is flanked by long terminal repeats (LTRs). They are closely related to the retroviruses. They are mobilized through the reverse transcription of their mRNA and the integration of this cDNA to a new place in the genome. The promoter region of these TEs is located inside the LTRs. The internal

part encodes for the gag protein that will encapsulate the mRNA, and a Polyprotein (Pol) encoding a peptidase, an integrase (INT), a reverse transcriptase (RT) and a ribonuclease H (RNAseH). Most LTR-RT lack the envelope (Env) protein typical of retroviruses. The LTR-RT RNA is encapsulated in particles named Virus Like Particles (VLPs), as they resemble the virus particles of retroviruses. Within the VLP the encoded peptidase processes the polyprotein into the different mature proteins and the reverse transcription takes place. Upon import into the nucleus, the integrase will integrate the new LTR-RT copy to the genome.

During retrotransposition, the sequence of the U3 region of the LTR, which is present in the 3' end of the mRNA is used as a template to synthesize that of the 5' LTR, and the U5 region present at the 5' end of the RNA molecule is used to synthesize that of the 3' LTR. As a consequence, the two newly produced LTRs are identical (Lynch & Walsh, 2007). This is commonly used to deduce the age of an LTR-RT, as the number of differences between the two LTRs of a particular LTR-RT copy is assumed to be proportional to the time passed since its insertion. Together with MITEs, LTR-RT are the most abundant TEs in plant genomes (Casacuberta & Santiago, 2003).

We can group most of the LTR-RTs found in two superfamilies, Copia and Gypsy, that differ in the order of the encoded proteins (Figure 2). However, there are also some non-coding LTR-RTs (Havecker et al., 2004) such as LARDs (from Large Retrotransposon Deletion derivatives) and TRIMs (Terminal repeat Retrotransposons In Miniature) that are mobilized by autonomous LTR-RTs.

Non-LTR Retrotransposons constitute the second order of Retrotransposons. LINEs are the most common non-LTR RT. They encode for an endonuclease and a reverse transcriptase and are mobilized through a process known as target-primed reverse transcription (Finnegan, 1997). In this case the TE is transcribed and translated, is imported into the nucleus and the encoded TE nuclease produces a nick into the genomic DNA, and the TE mRNA is reverse transcribed at the site where the nick is produced, resulting in a new integration of the TE into the nuclear genome. The non-autonomous counterparts of LINEs are named SINEs (from Short Interspersed Nuclear Elements) that are mobilized through the machinery encoded from the LINEs.

LINEs and SINEs have proliferated extensively in some mammalian genomes, such as in humans where the L1 LINE and the Alu SINE are highly prevalent, but they are usually

much less frequent than LTR-RTs in plant genomes, although it is possible to find active plant LINEs, such as the Karma LINE in rice (Komatsu et al., 2003).

Class II: DNA transposons

Class II or DNA transposons are divided in two subclasses that differ on the number of DNA strands that are cut during the transposition event.

Subclass 1 contains TEs that transpose through a *cut-and-paste* mechanism. The most abundant order is the TIR TEs. TIR TEs are characterized for the presence of terminal inverted repeats (TIRs) sequences flanking the internal coding sequence. TIR TEs encode a transposase that will recognize the TIR and cut the two strands at the end of both TIR sequences. After that, in a new position of the genome they will produce a cut and integrate the TE and generating a Target Side Duplication (TSD) at their integration point. This mechanism is conservative as the number of copies of the TE does not usually increase upon transposition. However, the number of copies can increase if the TE, for example, moves during DNA replication, mobilizing an already replicated copy to a region still not replicated by the DNA replication machinery (Fricker & Peters, 2014). TIR TEs can be further classified into superfamilies according to the transposase motif, the TIR sequences or the length or sequence of the TSD. Some of these superfamilies are the Tc1/Mariner TEs, hAT TEs, CACTA, MULE or PIF/Harbinger (Wicker et al., 2009). In general, TIR TEs can exist as autonomous elements or as non-autonomous elements, that are deletional or mutational derivatives of the former that can be mobilized in trans. The first and best known example of an autonomous/non-autonomous pair is the Ac/Ds system first described by McClintock (McClintock, 1950).

An interesting type of non-autonomous TIR TEs are MITEs (from Miniature Inverted-repeat Transposable Elements), especially in plants where they have been quite successful colonizing their genomes (Santiago et al., 2002, Feschotte et al., 2003). These elements are short (typically from 100 bp to 700 bp) and can contain TIRs that can be associated to the TIRs of autonomous DNA elements, being probably mobilized through their encoded transposase. MITEs can be amplified reaching high copy numbers, by a still to be described amplification mechanism, although some replicative mechanisms have been proposed (Izsvák et al., 1999).

A second order of this subclass contains the less known Crypton TEs. Originally found in fungi, they have further been identified in animals and oomycetes, but they have not been identified in plants. They encode for a tyrosine recombinase that will mobilize the TE through the genome.

Subclass 2 elements contains DNA TEs that replicate without the need of a double-stranded cleavage of the DNA. The helitron order belongs to this subclass. They replicate through a rolling-circle mechanism, producing a single strand cut during the process. Helitrons have been characterized mainly in plants, but they have been also found in animals and fungi (Kapitonov & Jurka, 2007). A second order of this subclass are the Maverick TEs, large TEs (between 10 to 20 kb) that encode a DNA polymerase and an integrase. They have been found in diverse eukaryotes but not in plants.

Impact of transposable elements on the host genome

As was observed by Barbara McClintock, TEs are an important source of mutations and can change the structure of the genome. These mutations can be created by their insertion but also by recombination between TE copies.

A direct impact of the mobilization of transposable elements, especially of Class I TEs, are changes in genome size (Kidwell, 2002; Vicient & Casacuberta, 2017). As Class I TEs generate new copies of themselves in new places of the genome, they increase their copy number and the size of the genome. These variations in genomic size caused by TE amplification have been observed in numerous eukaryote species. A paradigmatic example is the variability observed in the plant kingdom across species, with a clear correlation between the number of TEs and the genome size (Figure 3).

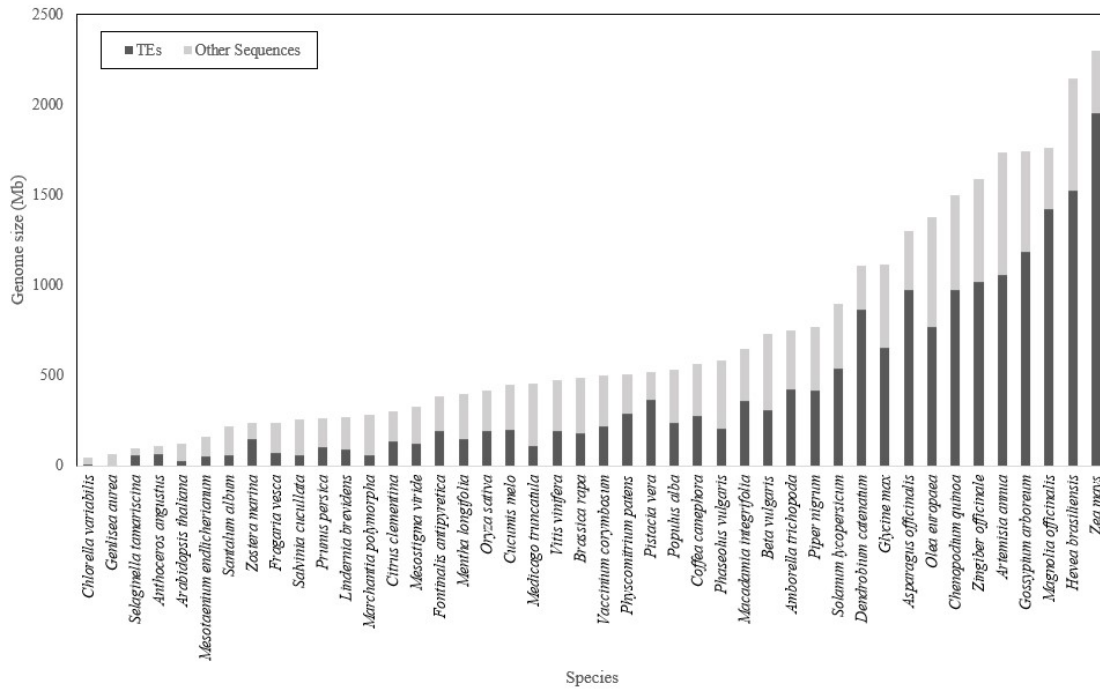


Figure 3: Genomic size variation across different species of all plant divisions and the percentage of TEs with respect to the total of genomic for each species. Data extracted from the plaBiPD webpage (https://www.plabipd.de/plant_genomes_pa.ep) and the manuscripts of each genome publication.

This correlation between genomic size and TE content has also been observed between closely related species, such as between *Oryza sativa* and *Oryza australiensis*. *Oryza australiensis* has a genome of double size as compared to *Oryza sativa*. This difference has been associated to a different activity of Retrotransposons between the two genomes (Piegu et al., 2006).

This increase of genomic size can be counteracted by the DNA recombination mechanism that tend to remove part of these repetitive regions. This process of TE mobilization and removal of TEs can also lead to genome rearrangements (Hedges & Deininger, 2007). These genome rearrangements can be a byproduct of TE mobilization itself or can be produced by the recombination between highly homologous DNA TE sequences, that are not necessarily active, located at distant places of the genome. They can range from chromosome breakages to duplications, inversions or deletions than can have both an effect on the global structure of the genome and to the genic regions.

Apart from this global changes of genome size and structure changes of the genome, TEs can induce more local mutations affecting gene coding capacity or gene expression (Figure 4) that can produce phenotypic effects on the host organism(Lisch, 2013).

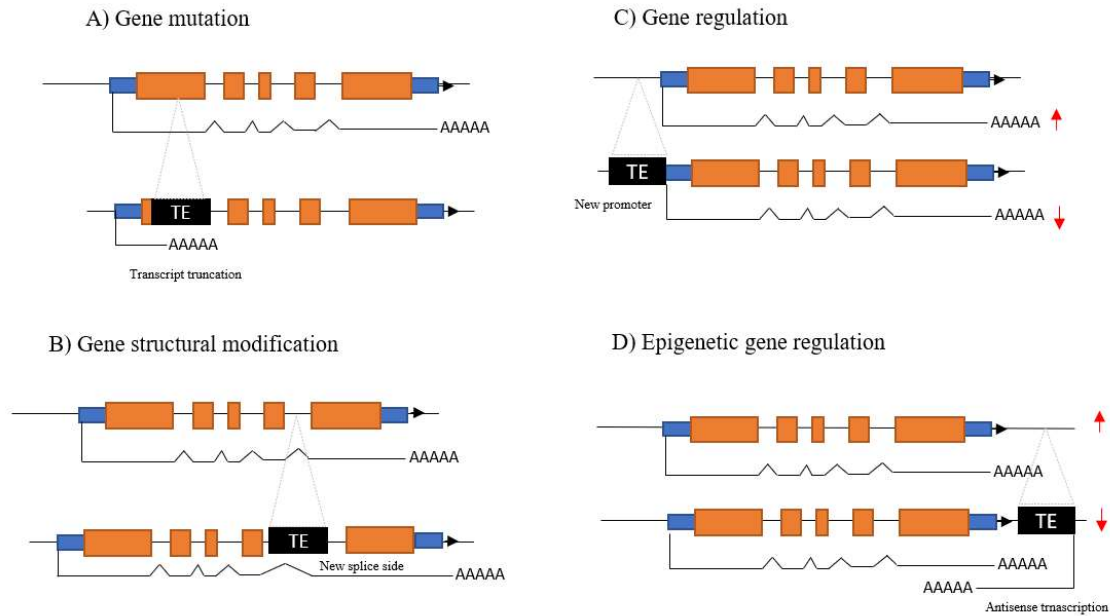


Figure 4: Different examples of impacts introduced by Transposable Elements on a genome. In A) we observe the truncation of a transcript induced by the insertion of a TE in an exon. In B) we observe the generation of a new transcript produced by the integration of a TE in an intron. In C) an example of a reduction of the transcription induced by the insertion of a TE in the promoter region of a gene. In D) an example of the production of an antisense transcript produced by a TE insertion at the 3' of the gene that leads to a reduction of the gene transcription.

An obvious impact that may be caused by TE movement is gene truncation, caused by the integration of a TE in the gene coding sequence leading to a premature stop codon. However, the insertion of TEs within genes can also induce other effects, such as the introduction of new splice sites altering the structure of the genes, or the modification of the 5' or 3' UTRs altering the transcription or the stability of the mRNA. TEs can also directly impact the expression of a gene by integrating into the promoter of genes mutating enhancers or repressors and even introducing new transcription factors elements carried by the TEs (Hénaff et al., 2014). In addition, TEs can also introduce epigenetic changes that may alter gene function.

TEs are usually targeted by the silencing machinery of the host organisms to prevent their mobilization, being methylated and accumulating histone modifications to keep the TEs

inactive (Lerat et al., 2019). These epigenetic modifications can alter the regions surrounding the TEs and modify the transcription of the surrounding genes. In addition, TEs can integrate in opposite direction to a given gene and produce antisense transcription of the gene that can result in gene silencing (Saze & Kakutani, 2007).

TEs can also capture coding sequences and mobilize them to a new genome region, which could lead to the production of new gene isoforms. Moreover, during evolution some TEs have been domesticated, acquiring new gene functions (Jangam et al., 2017). Two examples of that are, the gene encoding for the DAYSLEEPER transcription factor from *Arabidopsis thaliana* (*A. thaliana*) that was originated from a hAT transposase (Bundock & Hooykaas, 2005) or the domestication of a DNA TE to form the RAG1 and RAG2 recombinase genes in mammals, essential for the generation of mature lymphocytes B and T in jawed vertebrates (Y. Zhang et al., 2019).

For all the above reasons TEs are considered as major drivers of genomes evolution (Lisch, 2013).

Impact of transposable elements in crop genomes

TEs have been an important source of genetic variation for domestication and breeding of crops. In the last decades, many examples of TE-induced mutations selected during the process of domestication and breeding have been described (Wei & Cao, 2016). For example, the insertion of a DNA TE of the hAT superfamily in the promoter region of the DcMBY7 genes alters the production of anthocyanin resulting in different carrot colors (Xu et al., 2019), or the insertion of a Gypsy LTR-RT in the promoter region of the apple MdMYB1 gene resulting in an overexpression of the genes that lead to the red color phenotype of the fruit (L. Zhang et al., 2019). Another interesting example is the retrotransposon-mediated gene duplication of the SUN gene that is one of the major genes controlling the elongation of the fruit, which results in elongated tomatoes (H. Xiao et al., 2008). There are also other examples of polymorphic TEs related to an increase resistance to biotic and abiotic stresses such as in rice were the integration of an LTR-RT of Copia type at the promoter region of the Pit gene results in an enhanced resistance of the plant to the fungal pathogen *Magnaporthe grisea* (Hayashi & Yoshida, 2009) (Figure 5).

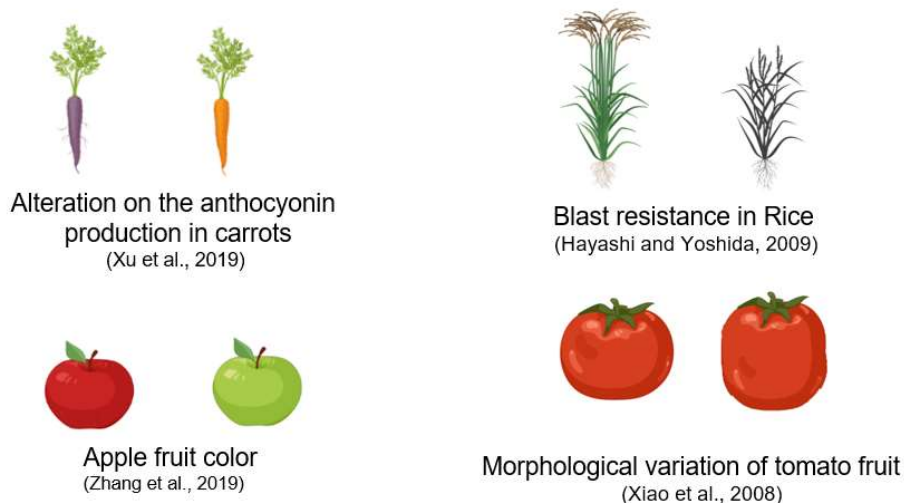


Figure 5: Different examples of phenotypes in different crops caused by polymorphic TEs altering gene expression.

Obviously, the impact of TEs into the generation of variability in populations is not restricted to crop plants. There are other examples of the impact of TEs into natural populations of plants. One example of that is a polymorphic TEs found at the vicinity of the Flowering Locus *C* gene in *Arabidopsis thaliana*. This TE polymorphism has been found associated with the attenuation of the expression of this gene and this TE may have been selected in the nature to regulate the flowering time between different populations (Strange et al., 2011). However, the impact of TEs in wild populations has remained less studied compared to cultivated plants. Nowadays, with the decrease of sequencing cost and the improvement of sequencing technologies, there has been a dramatic increase in the number of sequenced plant genomes. These new genomic resources are allowing to draw a bigger picture of the impact of TEs in plant evolution and on the generation of variability in these organisms (Coletta et al., 2021; Danilevicz et al., 2020).

Transposable Elements distribution in plant genomes

The distribution of TEs in a genome is the result of the balance of two opposing forces; the integration of the TE in the genome, that may be or may not be specifically targeted to particular places of the genome, and the posterior process of selection that will eliminate the insertions that have a negative effect on the fitness of the organism and will maintain insertions that are neutral or beneficial for the host (Sultana et al., 2017). In general, in angiosperms genomes, TEs tend to accumulate in the pericentromeric regions, especially LTR-RTs. Despite that, some TE families are found enriched close to genes.

For example, it has long been known that Copia LTR-RTs and MITEs are usually found close to genic regions (Casacuberta & Santiago, 2003), whereas Gypsy LTR-RTs are frequently found in the pericentromeric regions (Alioto et al., 2020; Naish et al., 2021). However, during the last 5 years with the release of chromosome-assembly genomes of non-angiosperms genomes it has been observed that this pattern is not conserved in other non-seed plants such as bryophytes. The sequenced genomes of different bryophytes suggest that the pericentromeric regions are not enriched for TEs as it is the case for angiosperms (Diop et al., 2020; Lang et al., 2018). Moreover, in this bryophytes genomes, TEs and genes are relatively evenly distributed along the chromosomes (Szövényi et al., 2021). A similar pattern, has been observed in other non-seed plants such as in the lycophyte *Selaginella* (VanBuren et al., 2018). This could be a major difference in terms of genome structure between non-seed plants and angiosperms, but more chromosome assemblies are required to be able to conclude (Szövényi et al., 2021).

As we have seen, most of the knowledge on the dynamics and impact of TEs, is focused on seed plants, especially in angiosperms, while there is much less knowledge about TE dynamics and the impact of TEs into the genome of non-vascular plants, such as bryophytes. One of the goals of this dissertation is to study the dynamics and impact of TEs in the bryophyte *Physcomitrium patens* (*P. patens*) a moss widely used as a model organism and it was the only bryophyte with a genome assembled at a chromosome scale at the time that this thesis was started.

Physcomitrium patens as a model organism

Physcomitrium patens (previously known as *Physcomitrella patens*) is a model organism widely used in plant development and evolution studies (for an extensive review read Rensing et al., 2020). It belongs to the division of the non-vascular land plants (Bryophyta or bryophytes). Bryophytes is divided in three major classes, liverworts, hornworts and mosses, where *P. patens* belongs (Figure 6). Specifically, *P. patens* belongs to the subclass Funariade and to the genus *Physcomitrium* that contain more than 80 species (Medina et al., 2019).

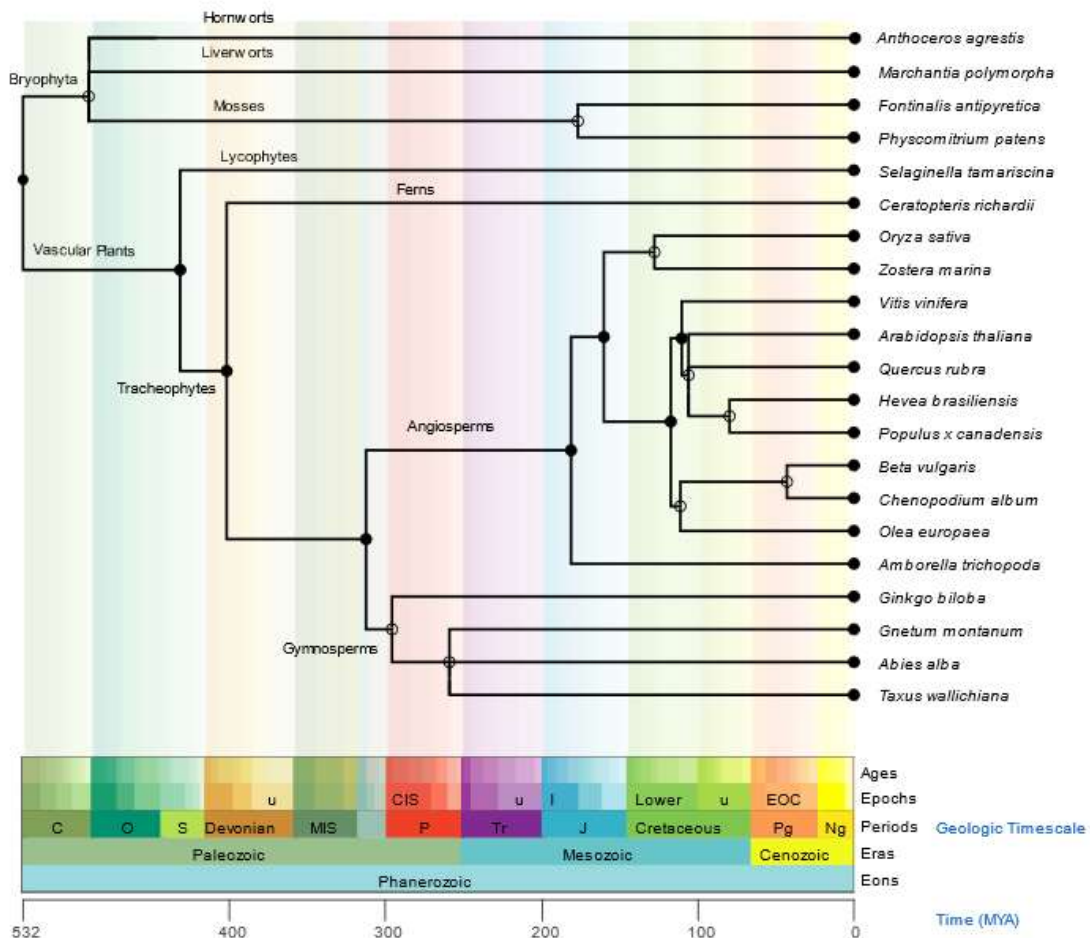


Figure 6: Phylogenetic tree of Land plants with a time tree at a scale based on phylogenetic analysis and fossil data. Tree developed using the timetree database (Kumar et al., 2017) .

P. patens is widely distributed in the northern hemisphere. It has been identified in America, Europe and East Asia. In the wild it is found seasonally in humid places, in low to middle altitudes.

P. patens was isolated from a forest close to the city of Gransden during the 1960s, from which the reference accession takes the name. Since then, it has been used to study moss morphology, gravitropism, response to hormones, among other studies. During the last two decades *P. patens* has been established as a widely used model plant. In part because its high homologous recombination rate allows for gene targeting approaches and that is haploid during most of the life cycle. These, combined with the possibility to induce its sexual reproduction in the laboratory (Hohe et al., 2002), the initial studies of transcriptomics (Rensing et al., 2002) and the publication of a reference genome (Rensing et al., 2008; Lang et al., 2018) together with the phylogenic position of the moss, belonging to the non-vascular plants lineage, and that can be easily propagated and manipulated in the laboratory, positioned *P. patens* as an interesting organism to study the development evolution of plants and to develop functional genomics studies.

P. patens, as most plants, has a life cycle divided in two phases, a gametophytic phase (haploid) and a sporophytic phase (diploid). However, compared to seed plants where the dominant phase during the life cycle is the diploid sporophytic phase, in the case of *P. patens* during most of the life cycle the dominant phase is the haploid gametophytic phase. The sporophytic phase only occurs during a short time period during the development of the progeny.

The life cycle of *P. patens* (Figure 7) starts from a single spore that germinates generating a primary tissue called protonemata. Protonemata are filaments that spread by branching growing in a bidimensional plane. Protonemata is composed of two major cell types: chloronema and caulonema. Chloronema is the initial tissue that emerges from the spores, rich in chloroplasts. This tissue as it grows transitions to caulonema, which has less chloroplasts, is thinner and grows much faster compared to chloronema (Rensing et al., 2020). From there it will develop side branches that will generate buds that are the initial stage of the three-dimensional phase, called gametophores.

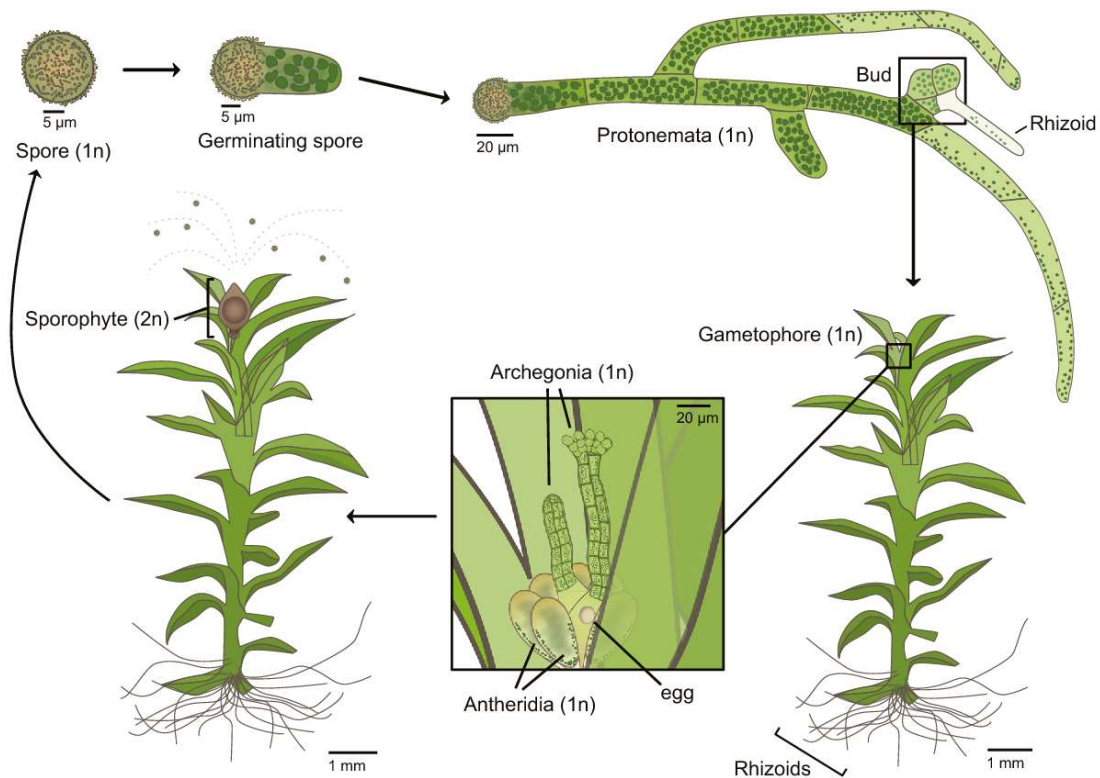


Figure 7: Life cycle of *P. patens*: spores (1n: haploid) would germinate to form protonemata divided in two main phases chloronema (dark green) and caulonema (light green). From there it would emerge a bud that would produce gametophores formed by leaflets (aerial part of the gametophores, similar to the leaves of vascular plants) and rhizoids (bottom of the gametophores, similar to the roots of vascular plants). The adult gametophores produce the male and female gametangia that after fecundation will generate the sporophyte (2n:diploid) where meiosis will take place and it will produce the next generation of spores that will be released closing the cycle. Reprinted with permission from Springer Nature Biophysical Reviews by Wu, Shu-Zon, Moé Yamada, Darren R Mallett and Magdalena Bezanilla. “Cytoskeletal discoveries in the plant lineage using the moss *Physcomitrella patens*.” Biophysical Reviews 10 (2018): 1683-1693. Copyright © 2018, International Union for Pure and Applied Biophysics (IUPAB) and Springer-Verlag GmbH Germany, part of Springer Nature (2018).

Gametophores are similar in structure to the sporophytic phase of vascular plants. They are structured into a stem that generates leaflets (similar to the leaves of vascular plants). From the bottom of the gametophores emerges a tissue similar to the roots of seed plants called rhizoids. Under short-day conditions, low temperature, and high humidity they will start the production of the male and female gametangia, generating a reproductive gametophore. The male gametangia will generate flagellated gametes that will require water to swim to the female gametangia and fertilize them. In the case of *P. patens* it can reproduce by selfing. After this, the zygote will develop into an embryo and will generate the sporophyte in the form of a capsule. Meiosis will occur in the immature sporophytes

(that are green), generating haploid spores. The sporophytes will mature (turning to a brownish color) and open, releasing the spores that after germination will generate again protonemata tissue closing the life cycle.

The moss can also be propagated asexually by taking part of the tissue and letting it regenerate in a new medium. This is the main mechanism of how it has been propagated in the laboratory since the establishment of the moss as a model organism.

The genome of *P. patens*

The genome of *P. patens* has a total length of 518 Mbp divided in 27 chromosomes. 57% of all the genome is formed by repetitive regions. The biggest fraction of these repetitions is occupied by LTR-RT (51.5% of all the genome). Most of these LTR-RTs are of the Gypsy superfamily, occupying a 48% of all the genome while Copia elements only occupy a 3.5% of all the genome. The other TE families occupy less than a 0.2% of the genome. Remarkably, there is a single LTR- RT family of the Gypsy superfamily called RLG1 that occupies almost the 25% of the genome and represents a 47% of all the TEs of the genome. This family is mostly found accumulated in the heterochromatic regions of the genome, although there are some copies that are located at the vicinities of genes. These heterochromatic regions of the genome mainly occupied by RLG1 elements are named in this dissertation as RLG1 islands.

Compared to flowering plants, the genome of *P. patens* has some peculiarities. As explained previously, TEs and the heterochromatic regions of the genome are usually found accumulated in the pericentromeric regions in flowering plants while in *P. patens* the heterochromatin is dispersed all along the chromosomes with the TEs and genes being almost homogeneously distributed among the chromosomes (Lang et al., 2018). Another interesting characteristic of *P. patens* genome is that it has an even recombination rate along chromosomes, differing from what is observed in other plants, where recombination is reduced in heterochromatic regions such as in flowering plants or in other bryophytes such as *Marchantia polymorpha* (Diop et al., 2020).

In spite of having a disperse heterochromatin, *P. patens* has a unique centromere per chromosome (Rensing et al., 2020). It has been hypothesized that the centromere may coincide with a region where there is an accumulation of a Copia LTR-RT called RLC5

elements, as well as putative non-autonomous elements of the same TE family called tRLC5 elements (Lang et al., 2018). In *M. polymorpha* a similar pattern is observed with an RT of the LINE superfamily marking the putative centromere (Diop et al., 2020).

Although there were initial studies trying to determine which TEs are being actively transcribed in *P. patens* and could be polymorphic in the population (Lang et al., 2018), the availability of new transcriptomic data in several conditions (Perroud et al., 2018) and the availability of new resequencing data of different accessions pushed us to dig deeper in the dynamics and impact of mobile genetic elements in this plant.

With this goal we first developed and studied which are the best approaches to detect TE insertion polymorphisms and the transcription of TEs from short-read sequencing data, that are further explained in Chapter 1.

In Chapter 2 we used the approaches developed in Chapter 1 to first study the dynamics of TEs in *P. patens*.

In Chapter 3, we studied the impact of TEs both in genome structure and in gene expression using CRISPR/Cas9 approaches to eliminate TEs of the genome and study their potential impact in *P. patens*.

Finally, in Chapter 4 we further investigated the presence of other mobile genetic elements, in this case viruses, in *P. patens* from transcriptomic datasets. Describing the first virus known to infect *P. patens* in the wild.

CHAPTER 1: DEVELOPMENT OF
BIOINFORMATIC TOOLS TO
IDENTIFY TRANSPOSABLE
ELEMENTS MOBILIZATION AND
TRANSCRIPTION

CHAPTER 1: DEVELOPMENT OF BIOINFORMATIC TOOLS TO IDENTIFY TRANSPOSABLE ELEMENTS MOBILIZATION AND TRANSCRIPTION

Chapter 1.1: Introduction

The publication of the human genome at the beginning of this century and the posterior drop of price to perform next generation sequencing have completely revolutionized the genomics field. In the case of plant genomes, the first land plant to be sequenced was the model plant *Arabidopsis thaliana* in 2000 (Kaul et al., 2000), two years before the human genome was published. Since then, more than 1000 plant genomes have been sequenced, corresponding to more than 800 different plant species (Kress et al., 2022). The number of plant species sequenced every year, and the quality of these genome sequences have grown exponentially in the last decade. This has been possible due to the decrease of costs required to sequence an organism and the improvement in the sequencing technologies, specially of long read sequencing technologies, that had improved their precision and length enabling to obtain high quality chromosome assemblies of several plant genomes (Michael & Vanburen, 2020). This has enabled a huge increase in the number of plant genomes sequenced as most of the plant genomes (74% of all the plant genomes published) have been sequenced in the last 4 years (Kress et al., 2022). Despite that, most plant genomes published during this period correspond to plants that have an agricultural interest, which represent a small portion of the vast diversity of organisms in the plant kingdom. There is still an underrepresentation of some land plant clades such as bryophytes, pteridophytes or gymnosperms. There are efforts to solve this unbalance, such as the Earth Biogenome Project, that has the goal to sequence all the eukaryotic organisms on earth (Lewin et al., 2018).

Plant genomes can be particularly challenging to assemble due to the ploidy level, the big genome size, and the high repetitive content of the genome of some species. As already introduced, one of the main components of these highly repetitive and big genomes are TEs. Next Generation Sequencing has allowed new approaches to study TEs. The publication of thousands of new genomes and the increase of resequencing data availability for these genomes (such as DNaseq, Bisulphite Seq, ChipSeq or RNAseq data) has allowed the study of the dynamics of TEs at a population level and has increased

exponentially the knowledge on the impact that TEs have on their host genome and their regulation.

One of the main problems when analyzing the repetitive regions and TE fraction of a genome is that they are more challenging to study than genic regions. Unique sequences, including genes, are simpler to assemble as compared to TE rich regions. In the same way it is easier to identify the transcription, methylation state, chromatin state or polymorphisms (such as SNPs or small INDELs) in the genic regions than in the TE-rich regions. As explained in the general introduction, most of the TEs in eukaryotic genomes, including plant genomes, are mostly found in abundance in highly repetitive regions. They form complex structures of recent TE copies flanked by fragmented and degenerated copies of other TEs. Their repetitive nature hinders all the analysis processes of these regions, from the identification of complete TE copies to the identification of TE transcription or their transpositional activity. To overcome all these challenges during these last two decades different tools and approaches have been developed to study the different impacts and roles of TEs using next generation sequencing data (Elliott et al., 2021). It is expected that the recent advancements and price decreases of long read technologies will facilitate the analysis of the TE rich regions of the genomes (Shahid & Slotkin, 2020). Despite that, at the start of this work, most of the data available for organisms with an assembled genome were essentially based on Illumina short-read sequencing. For this reason, almost all the work presented in this manuscript is based on these technologies.

Several problems may arise when studying TEs *in silico* using short-reads, but in this chapter, we will only focus on the use of NGS short-reads to detect TE activity. To detect TE activity several methods have been developed in the past years based on the use of NGS short-reads. One of the most used methods is the detection of polymorphic insertions in a population using DNA resequencing data (Hénaff et al., 2015; Keane et al., 2013), which can indicate recent TE activity. Another approach to detect TE activity is through the detection of the transcription of the TEs (Jin et al., 2015; Lerat et al., 2017) as this is the first step required for most of the TEs to transpose, although TE transcription does not necessarily translate into a new transposition event. Other examples of approaches that have been developed in the recent years are the analysis of the mobilome through the detection of TE DNA circles (Lanciano et al., 2021) or the purification of Virus Like

Particles (Satheesh et al., 2021) and posterior sequencing. Both methods are limited to some TE families as not all the TEs form Virus Like Particles or DNA circles. Moreover, although these methods can detect transposition intermediates, they cannot be taken as a direct indication of transposition.

In the first part of this chapter, we will focus on the comparison of different tools to identify TE insertion Polymorphisms (from now on TIPs) from DNaseq short-read data. To do that, we benchmarked several publicly available tools by using a high confidence TIP dataset that we generated based on the comparison of two rice assemblies. In the second part of this chapter, we will focus on methods to identify TE transcription that can potentially produce a new transposition event in the genome. We used previously published methods to analyze TE transcription and a novel method developed in the lab based on the assembly of short reads to identify putatively active TE copies.

Chapter 1.2. Objectives:

The main objectives of this Chapter are:

- Compare different available tools to detect Transposon Insertion Polymorphisms and discuss their strength and weaknesses.
- Identify the best approaches to identify Transposable Element transcription that can potentially lead to a new transposition event.

Chapter 1.3: Comparison of different available tools to detect Transposon Insertion polymorphisms using short-read data

Introduction

Several tools have been developed to detect TE Insertion Polymorphisms (TIPs) using short-read resequencing data. More than 80 tools have been published to detect TE insertion polymorphisms (Elliott et al., 2021). These tools can be used to look for the presence of both reference TE insertions (present in an assembled genome used as a reference) and non-reference TE insertions (not present in the assembled genome used as reference) on resequenced genomes of different samples, individuals, populations or varieties of the same species.

Most of the developed tools are based on the detection of discordant reads (paired-end reads mapping to the reference genome at discordant positions) and/or the presence of clipped reads (reads that map partially to the reference genome). Explained briefly, in the case of discordant reads a TIP is called when one of the two reads of the pair maps to a non-TE unique sequence in the reference genome while the second pair maps to a sequence included in a TE library (Figure 8). In the case of clipped reads part of a read maps uniquely to a single position of the genome while the other part of the part maps to a sequence included in a TE library (Figure 8). Both approaches allow the identification of TIPs using resequencing data.

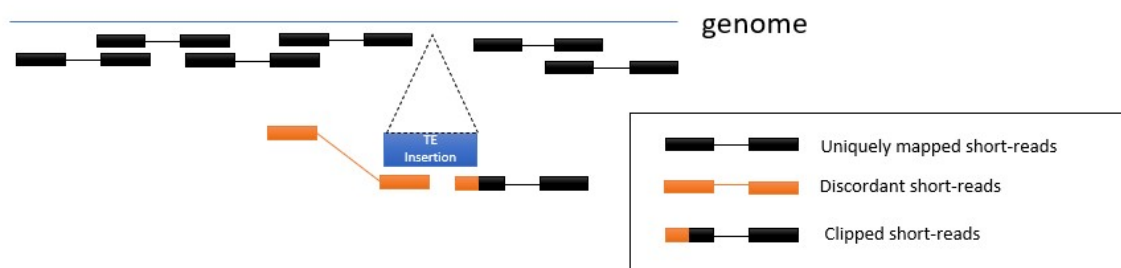


Figure 8: Scheme of the strategies used to detect TIPs from short-reads data. Both paired-end reads in black represents uniquely mapped short-reads, where both reads map uniquely to the expected place at the genome. Both paired-end reads in orange, represents discordant-short reads, where one read map uniquely to the genome and the second read map to a TE library, indicating the presence of a TE insertion in this region. Lastly, the read pair that is partially orange and black represents a clipped read where both reads map uniquely to the expected place of the genome, but part of one of the reads (in orange) maps to a TE library, indicating the presence of a polymorphic insertion at this place.

Although several tools have been developed, there is limited information on the performance of these tools when compared to other similar tools beyond the articles where the methods were published. In the case that there are some comparisons, these have been done using simulated datasets (Rishishwar et al., 2017). However, very little was known about the performance of these tools on real data, how do they compare between them under these conditions and which are their main limitations.

To address this, we performed a benchmark of different tools previously published taking the opportunity that two high quality *O. Sativa* genome assemblies were publicly available. To do that we annotated and located all the possible TIPs by comparing two genome assemblies of Rice (Nipponbare (Japonica) and MH63(Indica)) for two different TE families: LTR-RTs and MITEs. The main reasons to work with these two TEs are that both were known to be active in rice(Hirochika, 2001; Jiang et al., 2003), they are the most abundant TEs in plants and that they have different structure and distributions in the genome. While LTR-RTs are big (more than 3Kb in general) and are known to be mostly accumulated in the pericentromeric regions, MITEs are small sized TEs (~600bp) that are usually found close to genes.

Using these datasets, we run all the different tools to detect TIPs using four different resequencing coverages (5X,10X,20X and 40X). We then compared the results obtained using the different coverages, tools and TEs (LTR-RTs and MITEs). Finally, we complemented this study with previously published datasets of TIPs of *Homo sapiens* and *Drosophila melanogaster*, observing that the performance of the different tools was similar despite using different genomes.

This work was led by Josep M^a Casacuberta and Raúl Castanera and was collectively done by all the authors. Most of the experimental part has been done by Raúl Castanera, Fabio Barteri and myself. In my case I have run most of the tools on the different species (*Oryza sativa*, *Homo sapiens* and *Drosophila melanogaster*), manually curate the dataset with Raúl Castanera and analyzed the data together with Raúl Castanera and Fabio Barteri. Moreover, we collaborated to test the performance of the new version of T-lex3 that has been recently published using the dataset that we generated on rice (Bogaerts-Marquez et al., 2020)(see the annexes).

As first part of the chapter, a copy of the published article is included. The article was published at the end of the year 2019 at Mobile DNA (Vendrell-Mir et al., 2019), all the supplementary material cited in the article can be access through the following DOI:


<https://doi.org/10.1186/s13100-019-0197-9>

METHODOLOGY

Open Access



A benchmark of transposon insertion detection tools using real data

Pol Vendrell-Mir¹, Fabio Barteri¹, Miriam Merenciano², Josefa González², Josep M. Casacuberta^{1*} and Raúl Castanera^{1*} 

Abstract

Background: Transposable elements (TEs) are an important source of genomic variability in eukaryotic genomes. Their activity impacts genome architecture and gene expression and can lead to drastic phenotypic changes. Therefore, identifying TE polymorphisms is key to better understand the link between genotype and phenotype. However, most genotype-to-phenotype analyses have concentrated on single nucleotide polymorphisms as they are easier to reliably detect using short-read data. Many bioinformatic tools have been developed to identify transposon insertions from resequencing data using short reads. Nevertheless, the performance of most of these tools has been tested using simulated insertions, which do not accurately reproduce the complexity of natural insertions.

Results: We have overcome this limitation by building a dataset of insertions from the comparison of two high-quality rice genomes, followed by extensive manual curation. This dataset contains validated insertions of two very different types of TEs, LTR-retrotransposons and MITEs. Using this dataset, we have benchmarked the sensitivity and precision of 12 commonly used tools, and our results suggest that in general their sensitivity was previously overestimated when using simulated data. Our results also show that, increasing coverage leads to a better sensitivity but with a cost in precision. Moreover, we found important differences in tool performance, with some tools performing better on a specific type of TEs. We have also used two sets of experimentally validated insertions in *Drosophila* and humans and show that this trend is maintained in genomes of different size and complexity.

Conclusions: We discuss the possible choice of tools depending on the goals of the study and show that the appropriate combination of tools could be an option for most approaches, increasing the sensitivity while maintaining a good precision.

Keywords: Benchmark, Transposable elements, Polymorphism, Transposon insertion, Resequencing

Background

Transposable elements (TEs) constitute a very important fraction of eukaryotic genomes, and their ability to transpose, excise and produce complex genomic rearrangements make them a key source of genomic diversity. Previous work done over the last decades has uncovered their enormous potential as gene regulators, a role that TEs play through a variety of genetic and epigenetic mechanisms [12, 43]. Certain TEs, such as Long Terminal repeat (LTR)-retrotransposon carry their own promoters, and their

insertion close to genes can generate new gene expression patterns. In addition, TEs, and in particular LTR-retrotransposons and MITEs (Miniature Inverted Transposable Elements), have been shown to contain transcription factor binding sites, which can be mobilized by transposition rewiring new genes into pre-existing transcriptional networks [5, 12, 20]. As a consequence, TEs have the potentiality to generate important genomic and transcriptional variability, and the interest in these elements has drastically increased in the last years.

Due to their repetitive nature and their sequence diversity, the annotation of TEs is more complex than that of protein coding genes. Nevertheless, thanks to the development of tools such as Repeatmasker (<http://www.repeatmasker.org>) and sophisticated pipelines such

* Correspondence: josep.casacuberta@cragenomica.es; raul.castanera@cragenomica.es

¹Centre for Research in Agricultural Genomics CSIC-IRTA-UAB-UB, Campus UAB, Edifici CRAG, Bellaterra, 08193 Barcelona, Spain

Full list of author information is available at the end of the article



as REPET [16], methodologies of TE detection and annotation in assembled genomes are today robust. The availability of high-quality reference genomes coupled with the exponential increment of resequencing data has boosted our capacity to evaluate intraspecific variability. By obtaining accurate maps of genetic variation, characterizing the genetic basis of phenotypic variance is now possible at a genome-wide scale thanks to association studies (GWAS). Until now, most of the efforts have been focused on analyzing the variability at the nucleotide level (SNPs, single nucleotide polymorphisms), as there are robust algorithms to perform variant calling. However, TEs generate an important part of the genetic variability present in a particular species. Moreover, the timing of occurrence of TE and SNP mutations is different, as the former can amplify in bursts generating a great amount of diversity in short periods of time, whereas SNP mutation rates are more constant in time. Therefore, the identification of Transposon Insertion Polymorphisms (TIPs) is of high interest. Nevertheless, our capacity to accurately identify TIPs using re-sequencing data is hampered by the structural complexity of TEs.

In the last few years, many laboratories have developed bioinformatic tools to look for TIPs and have started to analyze their impact in intra-species variability, including crop plants [7, 10, 42]. There are two main approaches that can be used to detect TIPs in whole-genome sequence data: i) inference from discordant read-pair mappings, and ii) clustering of ‘split’ reads sharing common alignment junctions [2, 15]. Most of the recently developed tools incorporate both methodologies, and in some cases TIPs have been experimentally validated [27]. Moreover, in some cases the authors have evaluated their sensitivity and precision (also known as positive predictive value) [11, 24]. However, in most cases these evaluations were performed by generating simulated insertions that are randomly placed in the genome, and then used to compare with tool predictions. Simulated insertions are far from representing the complexity of “natural” TIPs, as many of their features are difficult or impossible to mimic accurately (i.e.: element degeneration, nested insertions, insertion preferences, etc.). As a consequence, the benchmarks done with simulated data tend to overestimate performance of the tools analyzed [21]. An example of such benchmarks is the one reported by the developers of McClintock, a pipeline that integrates six tools [36] (Table 1). In their study, the authors provided a detailed comparison of their component’s performance in sensitivity and positional accuracy based on simulated LTR-retrotransposon insertions, which also includes some real resequencing data, in the yeast *Saccharomyces cerevisiae*. In spite of the interest of such comparative analysis, the direct translation of these results to other eukaryotic models with bigger and more

repetitive genomes is uncertain. This is especially relevant as *S. cerevisiae* contains only 51 full LTR-retrotransposons in the whole genome [8], whereas in most plant and animal genomes the LTR-retrotransposon load is several orders of magnitude higher. Also, a recent study focused on simulated but also real human AluY, L1 and SVA families revealed huge differences in the ability of seven tools to detect TIPs [41]. In spite of the importance of these families for human research, they do not represent the diversity of the TE landscape of other animals and plants, which is far more complex, with many families from different orders being potentially active, and where the amount of truncated non-autonomous elements greatly outnumbers the active copies.

In plants, TEs are at the origin of important agronomic traits, such as apical dominance in maize [45], the skin and flesh colors in grape [28] and blood oranges [4]. Different efforts have been made recently to identify TIPs that could be responsible for important variability in plants. Carpentier *et. al* [7] screened the presence of 32 rice LTR-retrotransposon families in the 3000-rice genome dataset and uncovered more than 50,000 TIPs, most of them occurring at a very low frequency, which is indicative of recent activity. Besides LTR-retrotransposons, MITEs are probably the most prevalent group of transposons in plants, including rice, where they have experienced recent massive amplification bursts [10, 35]. MITEs are structurally very different from LTR-retrotransposons, as they are non-autonomous, usually non-coding, and relatively small. They are of particular interest because they tend to integrate close to genes and may carry regulatory domains [20], having the potential to create or rewire regulatory networks [12]. In the present study, we have taken advantage of the existence of several high-quality assembled genomes of different rice varieties to create a validated dataset of natural LTR-retrotransposon and MITE insertions obtained by direct comparison between the assembled genomes (Nipponbare and MH63), that we have used to benchmark the performance of 12 TIP calling tools. Moreover, we have also analyzed the sensitivity of the best performing tools to detect experimentally validated TIPs in *Drosophila* and humans. Our results evidence that tool performance is in general lower than estimated by previous simulations, and highly variable depending on sequencing coverage and TE type. Also, we show that an appropriate combination of tools can increase the sensitivity of predictions while maintaining high precision levels.

Results

Tools selected for benchmarking

We selected 12 of the most widely used tools for the detection of TIPs (Table 1). Among them, four were

Table 1 Tools selected for the benchmark of TE insertions

Tool	Target	Prediction	Input	Output format	Perceived difficulty		Manual
					Installation	Input preparation	
RelocaTE2	Non-reference insertions	All families	fastq	gff file	Easy	Easy	https://github.com/JinfengChen/RelocaTE2
Jitterbug	Non-reference insertions	All families	Bam	gff file	Medium	Medium	https://github.com/elzbth/jitterbug
Retroseq ^a	Non-reference insertions	All families	Bam	vcf file	Easy	Difficult	https://github.com/tk2/RetroSeq/wiki/RetroSeq-Tutorial
ITIS	Non-reference insertions	Single-family	fastq	Bed file	Easy	Medium	https://github.com/Chuan-Jiang/ITIS
MELT	Reference and non-reference insertion	Single-family	Bam	vcf file	Easy	Medium	http://melt.igs.umaryland.edu/manual.php
PopoolationTE2	Reference and non-reference insertions	All families	fastq	Tool-specific	Easy	Easy	https://sourceforge.net/projects/popoolation-te2/
Teflon	Reference and non-reference insertions	All families	fastq	Tool-specific	Medium	Medium	https://github.com/jradrion/TEFLoN
Trackposon	Reference and non-reference insertions	Single-family	fastq	Bed file	Easy	Easy	http://gamay.univ-perp.fr/~Panaudlab/TRACKPOSON.tar.gz
TEMP ^a	Reference and non-reference insertions	All families	Bam	Tool-specific	Easy	Difficult	https://github.com/JialiUMassWengLab/TEMP/blob/master/Manual
TE-locate ^a	Reference and non-reference insertions	All families	Sam	Tool-specific	Easy	Difficult	https://sourceforge.net/projects/te-locate/
PopoolationTE ^a	Reference and non-reference insertions	All families	fastq	Tool-specific	Easy	Difficult	https://sourceforge.net/p/popoolationte/wiki/Workflow/
ngs_te_mapper ^a	Reference and non-reference insertions	All families	fastq	Bed file	Easy	Difficult	https://github.com/bergmanlab/ngs_te_mapper
McClintock	Reference and non-reference insertion	All families	fastq	Bed file	Easy	Difficult	https://github.com/bergmanlab/mcclintock

^a These tools were run as part of the McClintock pipeline. Perceived difficulty refers to McClintock and not the original methods

Installation: Easy = available in Conda, or automatic / semi-automatic installation. Medium = Needs several dependencies or specific versions of packages that need manual installation. Input preparation: Easy = can be run using common formats (ie fasta, bed) without the need of specific formatting. Medium = Needs specific formatting. Difficult = Needs very specific formatting

specifically designed to detect non-reference insertions (not present in the reference genome) (RelocaTE2 [11], Jitterbug [21], Retroseq [27] and ITIS [24]), and eight were able to detect reference (present in the reference genome) and non-reference insertions (MELT [18], Popoolation TE2 [29], Teflon [1], Trackposon [7], TEMP [48], TE-locate [37], Popoolation TE [30], and ngs_te_mapper [32]). Tools specifically designed to detect presence/absence of reference TE insertions in re-sequenced genomes (i.e.: T-lex 3) [3] were not benchmarked here.

In addition to their different targets, some of the tools were family-specific (meaning that they run with one TE family at a time only), whereas most of them are able to detect insertions from all the families in the same run (broad-spectrum). Five out of the 12 tested tools were run as components of McClintock, a pipeline that combines the use of several TIP detection tools and standardizes their outputs into the commonly used BED format (Table 1).

The first difficulty that the user has to face is properly installing and making the tools run, often in a computer

cluster. This can be sometimes complex due to the number of different dependencies, and especially due to the specificity of the input file preparation. In this regard, we found that RelocaTE2, PopoolationTE2 and Trackposon were the less problematic tools (Table 1). One possibility that would make the installation of these tools much easier would be to have them integrated in an environment such as Conda. This is a possibility that future developers should take into account.

LTR-retrotransposon and MITE landscape in Nipponbare and MH63 genomes

In order to perform a benchmarking exercise that could be representative of as much as possible TIP detection in eukaryotes, we decided to use rice as a model as it has a genome of 430 Mb, which is relatively large and complex in terms of TE landscape, and that has already been considered as being as close as possible to a representative genome for angiosperms [7]. Moreover, there are several good-quality assemblies and short-read datasets of rice varieties available [23, 47]. In terms of the

TEs to be detected, we concentrated on LTR-retrotransposons and MITEs as, in addition to be the most prevalent TE types in plant genomes, they are functionally and structurally very different. Indeed, whereas LTR-retrotransposons are relatively long elements (typically several Kb-long) and contain many structural features relatively easy to detect (e.g.: long LTRs at their extremities, coding capacity for several well conserved enzymatic activities), MITEs are short (typically 100–800 nt), are non-coding and do not contain structural features (except for short inverted-repeats in most cases) allowing for structural detection.

We used a combination of structural and homology-based approaches to annotate a high-quality dataset of 3733 and 3787 full-length LTR-retrotransposons in Nipponbare and MH63 (Minghui 63) assemblies, respectively (Table 2). These elements contain intact Target Site Duplications (TSDs), Long Terminal Repeats as well as coding domains. All of them were clustered at 80% similarity over 80% length to obtain families and we derived a consensus for each family. RepeatMasker was then run with such consensuses to identify all the LTR-retrotransposon copies of the genome (including fragments and non-autonomous elements) related to the characterized families. A similar strategy was used to identify ~46,000 full-length MITEs, as well as ~200,000 partial MITE copies (see methods section). Whereas full-length LTR-retrotransposons represent a very small proportion of the total number of LTR-retrotransposon copies detected, (3%, Table 2), full-length MITEs represent an important fraction (23%). The distribution along the chromosomes of the two transposon groups is also different, with LTR-retrotransposons being more abundant in the centromeric and pericentromeric regions and MITEs populating evenly the rest of the chromosome (Fig. 1).

Table 2 Annotation of LTR-retrotransposons and MITEs in rice assemblies

TE Classification	Nipponbare	MH63
LTR-all ^a	131,905	117,362
LTR full-length ^b	3733	3787
LTR- Gypsy	1354	1303
LTR- Copia	944	759
LTR- Unclassified ^c	1435	1725
MITE-all ⁽¹⁾	211,732	191,113
MITE full-length ^d	45,963	46,725

^a Repeatmasker fragments. Includes both intact and truncated elements

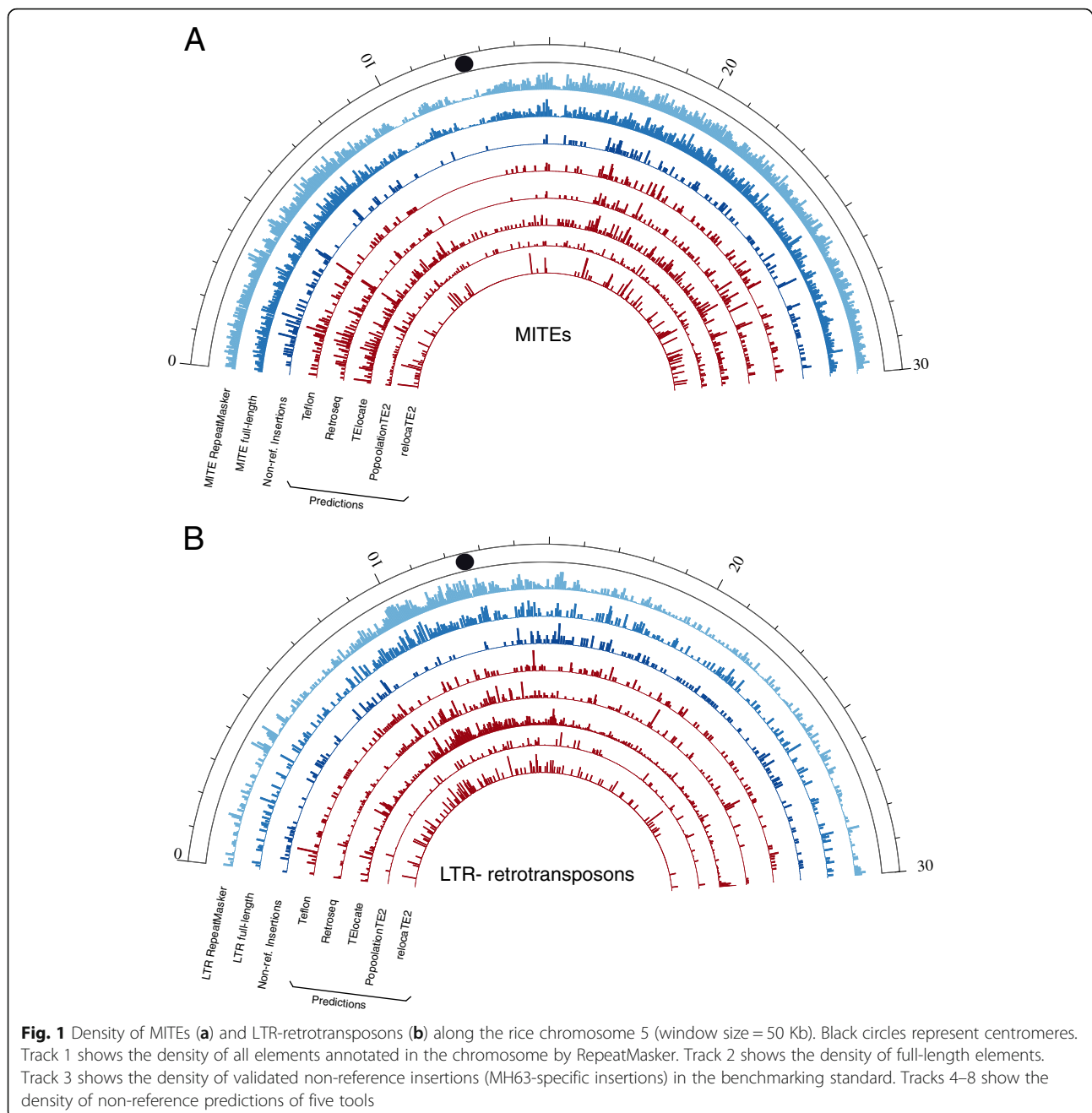
^b High confidence elements containing intact LTRs, TSDs and coding domains

^c Intact elements whose poor coding domain conservation doesn't allow proper classification

^d Elements spanning more than 80% of its family consensus length

Annotation of standard transposon insertion datasets for tool benchmarking

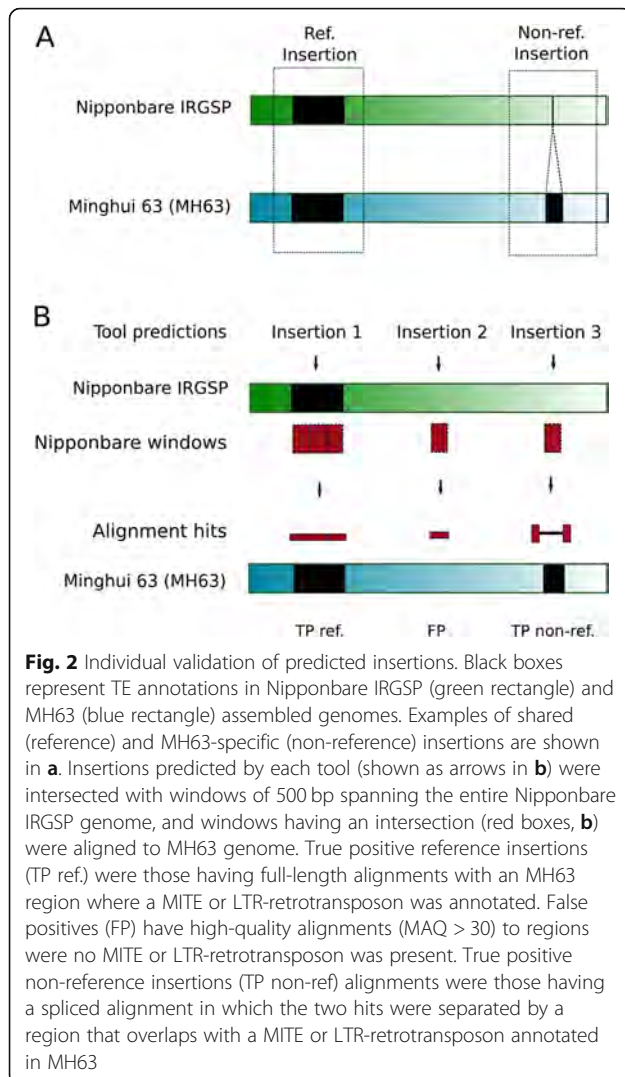
The most straightforward way of identifying an insertion polymorphism “in silico” when two high quality assembled genomes are available (as it is here the case), is by aligning orthologous loci. To identify the Nipponbare orthologous loci to those that in MH63 contain a TE insertion, we mapped the flanking regions of each MH63 full-length LTR-retrotransposon and MITE insertion against the Nipponbare genome. As sequence diversity and structural differences between the two genomes may complicate this analysis, we tested different flanking sequence lengths and found that 500 nt was the one that allow to identify more reference and non-reference insertions (Additional file 6: Figure S1). By inspecting the distance between the two mapped flanks, we could assign the orthology status to the locus (ie, empty site or full site). Using this approach, we were able to assign an orthology status to 86% of the MITE loci, but only to 41% of the LTR-retrotransposons loci. This was probably due to the difficulty to identify the orthologous loci of insertions siting in repetitive sequences, which is much more frequent for LTR-retrotransposons than for MITEs. Therefore, although this strategy seems the more straightforward, it has clear limitations. Moreover, as defining the precise TE-genome junctions for non-full length elements (ie, degenerated or partial elements, which are the vast majority of LTR-retrotransposons, Table 1) is challenging, we could not use this strategy to analyze the possible polymorphisms arising from non-full-length LTR-retrotransposons. To overcome those limitations and increase the dataset of curated insertions, we developed a strategy aimed at complementing the TIPs dataset with TIPs predicted with the 12 tools analyzed here (Table 2), which were individually validated. To this end we ran the different TIP-prediction tools using MH63 paired-end reads mapped to the Nipponbare reference genome. We divided the Nipponbare genome in 500 nt windows and mapped the windows containing predicted insertions (red boxes, Fig. 2) to the MH63 genome. An inspection of the aligned sections allowed determining whether the predicted insertion corresponded to a reference (shared) or non-reference (MH63 specific) insertion or if it should be considered a false positive (Fig. 2b). Indeed, in case of reference (shared) insertions, the Nipponbare and the corresponding MH63 sequences would perfectly align, showing that the sequence, which contains a TE insertion is conserved in both genomes (Fig. 2b, left); in case of a non-reference (MH63 specific) insertion, the alignment will be split by an insertion in the MH63 sequence corresponding to an annotated TE (Fig. 2b, right); and in case where the two sequences show a continuous alignment in the absence of an annotated TE insertion in



Nipponbare, this will indicate that the TE prediction is a false positive (Fig. 2b, middle). After running all tools, adjacent windows corresponding to TIP predictions of the same category were merged to produce a final dataset. LTR-retrotransposon insertions are frequently more complex than MITEs (i.e.: length, tendency to form nested insertions and extremely high amount of truncated and degenerated elements, Table 2). Because of this, it was difficult in many cases to automatically validate the insertions. Therefore, manual inspection of the alignments of LTR-retrotransposons TIPs was

performed, and we decided to restrict the dataset of LTR-retrotransposons to a single chromosome (chr5).

This strategy combined the power of detection of read-based methods (useful for uncovering polymorphisms derived from both full and degenerated elements), with the reliability of the validation based on alignments between high-quality assembled genomes. By using this combined approach, we increased the number of validated non-reference MITE insertions from 1898 to 3117 whereas for LTR-retrotransposons (chr5) the amount of non-reference insertions in our validated



dataset increased from 22 to 239 (Additional file 2: Table S1). The result was a high-quality dataset of True Positive (TP) and False Positive (FP) reference and non-reference insertions (Additional file 2: Table S1). In addition, there were predicted insertions that did not match neither with TP nor FP (i.e.: cases that did not fit in the scenarios described in Fig. 2b). We analyzed the specific cases of unclassified non-reference insertions and found that 86% of these LTR-retrotransposon predicted TIPs and 92% of such MITE TIPs overlapped with other transposons annotated in the reference. These cases were not used for downstream analyses, as most tools specifically indicate in their manuals that they cannot properly detect nested insertions. In order to evaluate the performance of each tool, we intersected the windows corresponding to the TE insertions predicted by the tool (both reference and non-reference TE insertions) with those of the curated dataset to identify TP and FP (Fig. 2b). Insertions present in the curated

dataset of TE insertions that were not detected by the evaluated tool were counted as False Negatives (FN).

Most of the tools analyzed here are able to detect insertions from all the families in the same run (broad-spectrum). Some of these tools are able to detect reference and non-reference insertions, whereas others only detect non-reference insertions. The programs use different strategies to identify these two types of insertions, and consequently we analyzed their performance separately.

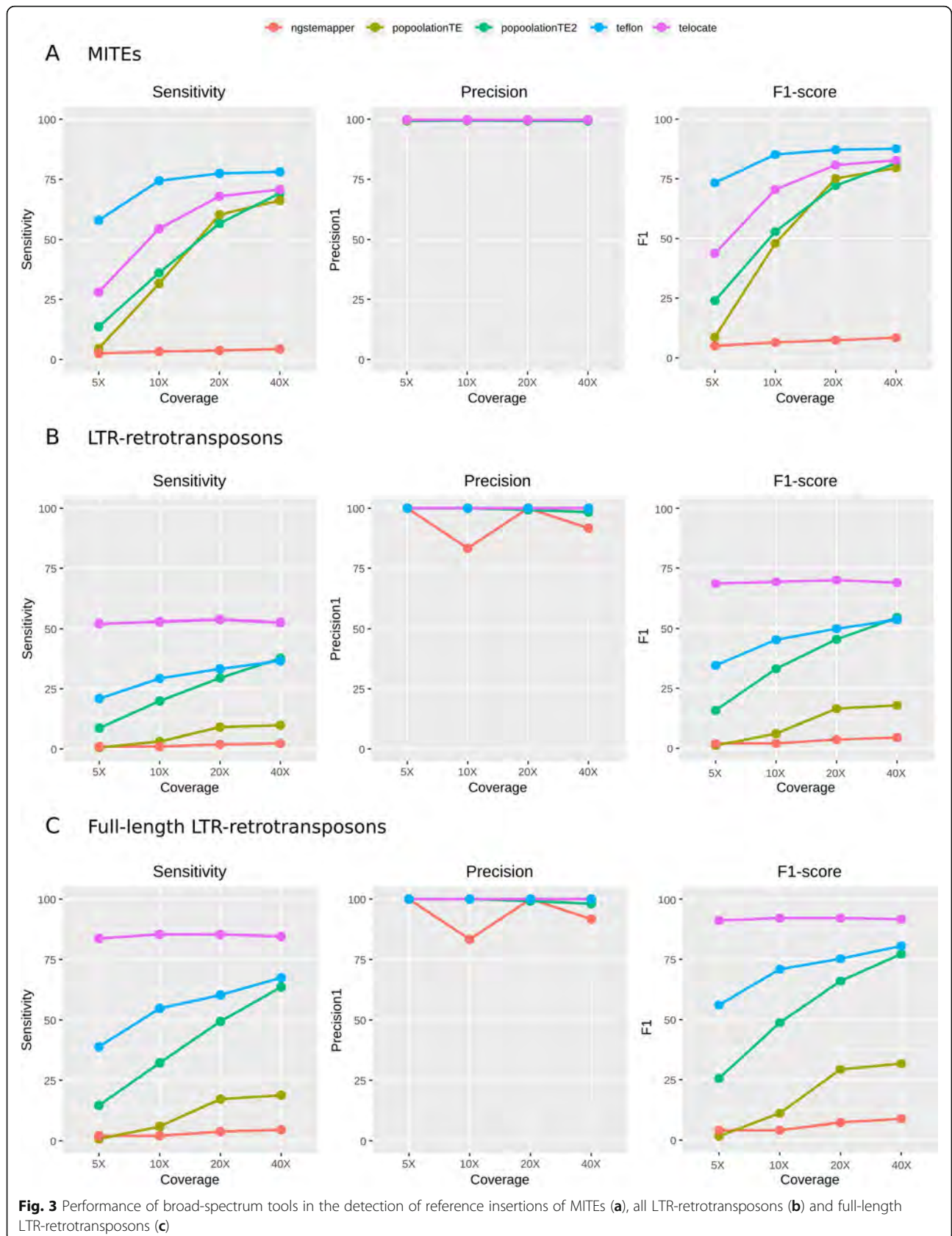
Detection of reference insertions by broad-spectrum tools

We observed that whereas the precision detecting MITE and LTR-retrotransposon reference insertions was very high for both types of elements, the sensitivity levels of most of the tools were much higher for MITEs (Fig. 3). For MITEs, the sensitivity of most tools increased with coverage and tended to stabilize at 20-40X coverage (Fig. 3a). Teflon had consistently the best sensitivity and overall performance (F1-score) in the detection of reference MITE insertions even at low coverage, reaching a sensitivity of 74% at 10X with an almost 100% precision (Fig. 3a). All tools showed precision levels higher than 99% at all coverages, and all tools except *ngs_te_mapper* yielded a sensitivity higher than 60% at 40X (Fig. 3a, Additional file 3: Table S2). By contrast, the sensitivity at 5X was low in general, with Teflon being the only tool reaching more than 50% (Fig. 3a).

Regarding the detection of reference LTR-retrotransposons, the general tool performance was much lower than for MITEs (Fig. 3b). In this case, TE-locate reached the maximum sensitivity followed by Teflon and was only slightly higher than 50% (Fig. 3b), and the other tools remained below 40% sensitivity. The sensitivity of TE-locate was higher than 50% in all the coverages, whereas in Teflon, PopoolationTE2 and PopoolationTE it increased with coverage (Fig. 3b). When we focused only on the detection of full-length LTR-retrotransposons, the performance of all tools increased considerably, reaching a maximum sensitivity of 85.4% (Fig. 3c). TE-locate was again the best performer showing a sensitivity over 80% for all the coverages. We excluded the predictions of TEMP for reference insertions, as this tool is based on the detection of absences assuming the presence as default, which leads to an overestimation of the number of insertions, especially at a very low coverage.

Detection of non-reference insertions by broad-spectrum tools

All the benchmarked tools are able to detect non-reference insertions, a task that is more challenging than detecting reference insertions, as the former are not present in the reference genome to which the reads are

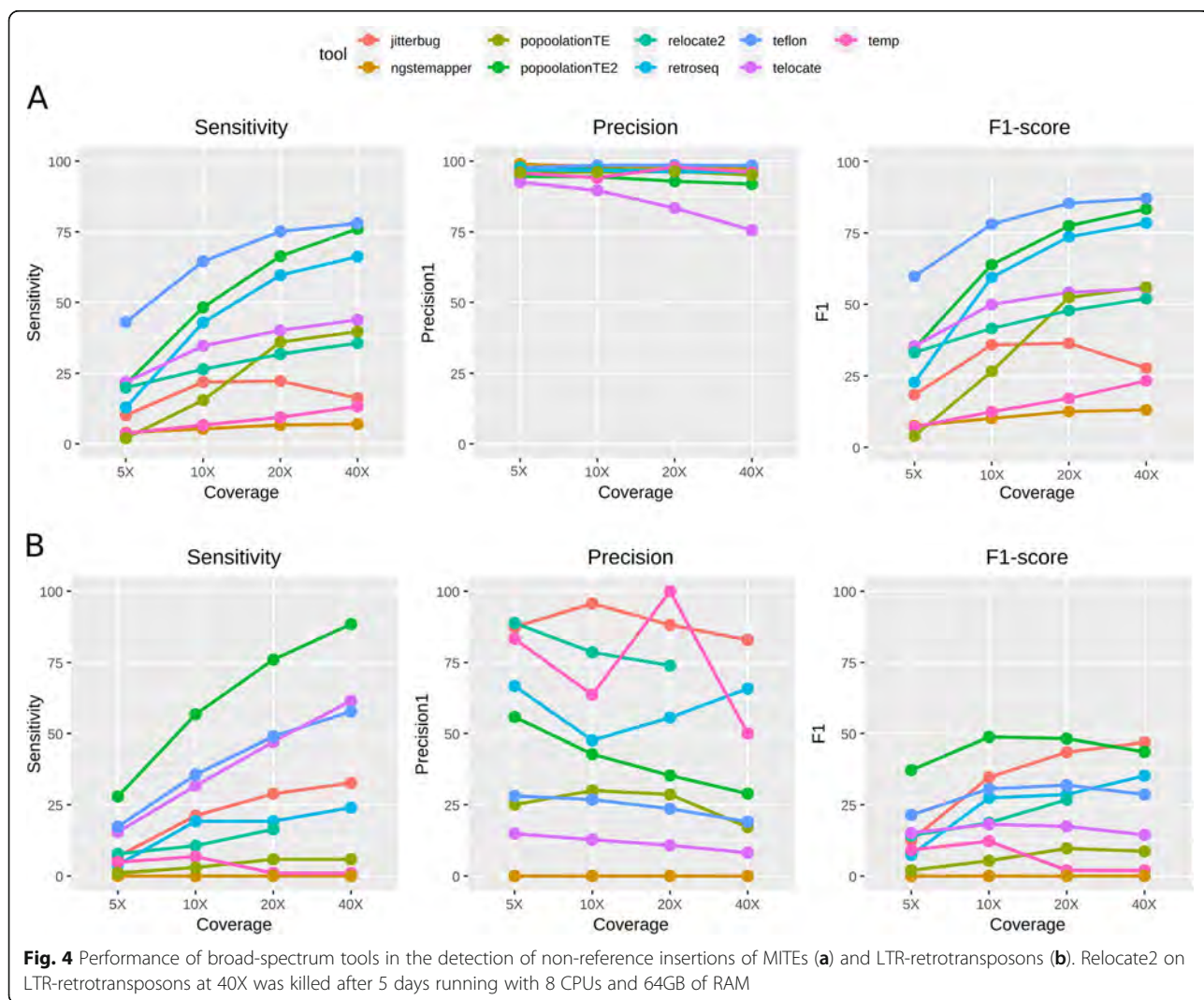


mapped. In this case sensitivity was strongly dependent on coverage (Fig. 4). Precision was very different for MITE and LTR-retrotransposon predictions, showing a tendency to decrease at high coverage (Fig. 4). Regarding MITEs, Teflon was the best performer followed by PopoolationTE2 and Retroseq (Fig. 4a). These tools reached a sensitivity close to 75% (up to 75.6% in 40X coverage for Teflon), whereas the rest of the tools had a much lower sensitivity (Fig. 4a). The precision was very high (>95%) for most tools with the exception of Teflon, which dropped from 92.5% in 5X to 75.6% in 40X. All the tools improved their performance when the coverage increased (except Jitterbug, which performed the best at 20X), with PopoolationTE2 and Retroseq showing the steepest increase, especially between 5X and 20X (Fig. 4a).

Regarding LTR-retrotransposons, PopoolationTE2 achieved the highest sensitivity, reaching a maximum of 88.5% at 40X (Fig. 4b). Nevertheless, these tools yielded

a high number of false positives, which translates into low precision levels (Fig. 4b). In general, the precision detecting LTR-retrotransposons with respect to MITEs was much lower for all tools. Jitterbug was the only program with a moderate precision (>75%) across all coverage levels, although its sensitivity was low (maximum of 32.7% at 40X) (Fig. 4b). According to the F1-score, PopoolationTE2 and Teflon were the best performers at low coverages (5X-10X), whereas at higher coverages PopoolationTE2 and Jitterbug showed the best balance between sensitivity and precision (Fig. 4b). Differently to what we previously did for reference insertions, we did not compute the performance of the tools using only full-length LTR-retrotransposons because they represent only a small fraction of the non-reference annotated insertions.

The output of most tools contains information that can be used for filtering the putative insertions to achieve more precise detection levels. We checked



different filters for each program looking for gains in precision with a low cost in sensitivity. In some cases, such as Jitterbug, the precision was already very high, and the filtering was not needed. In others, the cost in sensitivity was too high and the filtering was not considered useful. For the two best-performing tools, PopoolationTE2 and Teflon, filtering did result in significant gains in precision without an excessive cost in sensitivity. For PopoolationTE2 we applied a zygosity filter of 0.7 (based on the fraction of reads supporting the insertion) which led to a drop of sensitivity for both MITEs (from 76 to 63%) and LTR-retrotransposons detection (from 88 to 65%, Additional file 7: Figure S2), but with an increase of precision, which was particularly striking for LTR-retrotransposons (from 28.9 to 91.9% at 40X). For Teflon, a zygosity filter of 1 resulted in a drop of sensitivity for MITEs (from 78 to 61.5%) and LTR-retrotransposons (from 57.7 to 44.2%) but with important gain in precision for LTR-retrotransposons (from 15.2 to 70.8%), which was not significant for MITEs (98.4 to 98.5%) (not shown). In summary, based on the F1-score, filtering by zygosity greatly improved the overall performance of PopoolationTE2 and Teflon for LTR-retrotransposon detection, whereas the effect of this filter on MITEs detection was much less pronounced due to the already high precision of the unfiltered results.

Detection of non-reference insertions by family-specific tools

Some tools have been designed to look only for TIPs of a single TE family instead of all families at the same time (i.e., ITIS, MELT and Trackposon). In order to analyze the performance of such tools, we used the largest MITE and LTR-retrotransposon families, which contain 194 (whole genome) and 22 (chr5) MH63-specific insertions, respectively (Additional file 7: Table S1). The analysis of MITE TIPs showed that ITIS and MELT did not perform well and displayed low sensitivity and overall F1-score levels (Fig. 5a). By contrast, Trackposon performed well, displaying up to 72.8% sensitivity with 93.1 precision at 40X coverage. In line with the results found for broad-spectrum tools, sensitivity in the detection of LTR-retrotransposons was strongly dependent on the coverage. Trackposon and MELT showed moderate sensitivity levels at 40X (58.6 and 55.2%, respectively) whereas ITIS reached a maximum of sensitivity of 13.8%. Regarding precision, Trackposon was the best performer with values ranging between 76.9 and 100% (Fig. 5b).

Overlap between TIP prediction tools

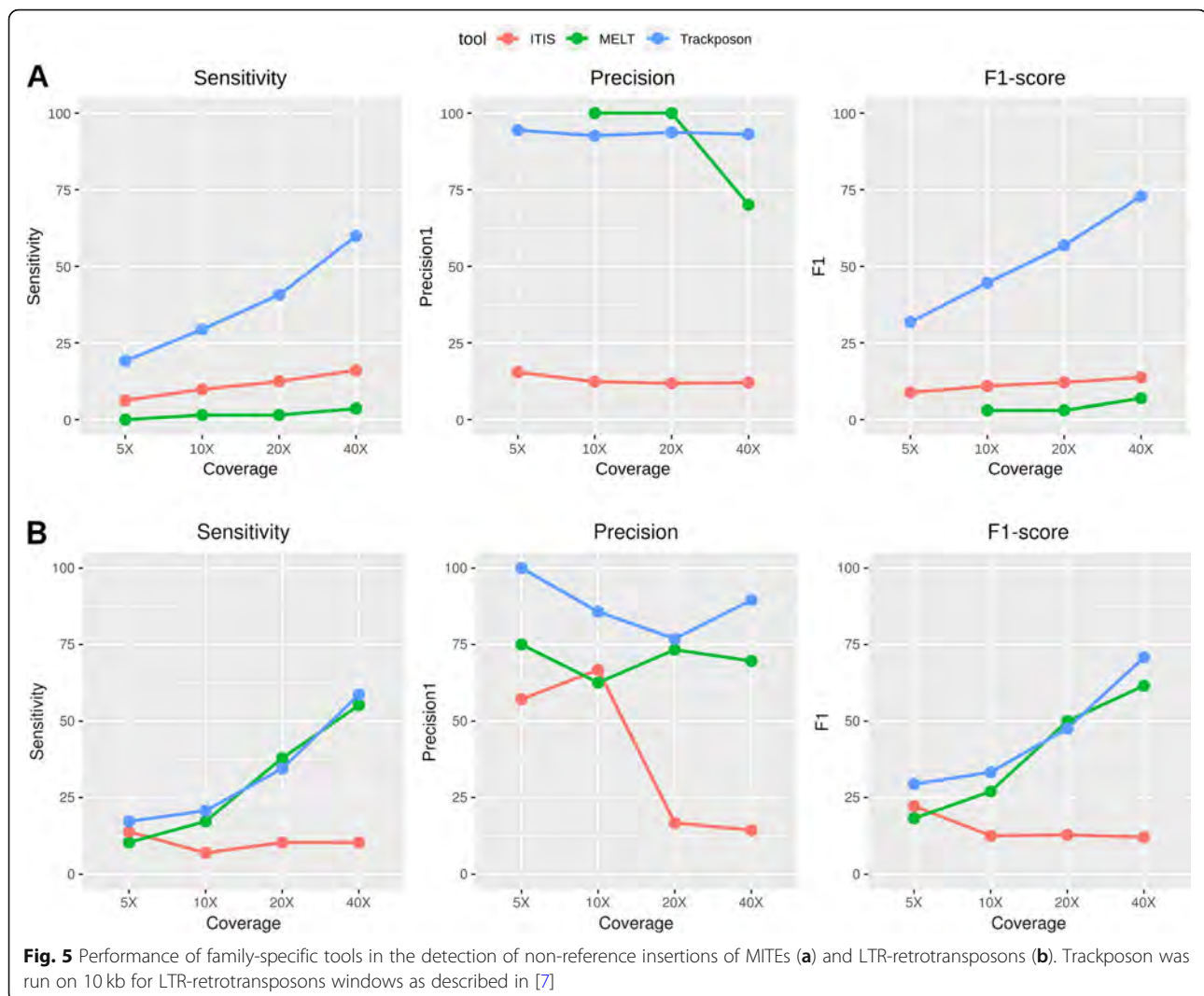
As there is no tool showing 100% sensitivity, we asked whether the predictions of the different tools were common or specific for each tool. We evaluated the overlap

of the detected non-reference true and false positives for the five better performing tools for MITE or LTR-retrotransposon TIP predictions (40X), taking into account their sensitivity and precision. In spite of the difference in the amount of predictions between MITEs and LTR-retrotransposons, the results showed very similar trends: 54% of TP were detected only by one tool for both MITE and LTR-retrotransposon insertions (Fig. 6). As expected, the FP detected were tool-specific in the vast majority of the cases (90.2% were detected by only one tool for MITEs and 98% for LTR-retrotransposons). The number of insertions detected by all tools was very low (1.3% of all TIPs detected for MITEs and 1.4% for LTR-retrotransposons). These results suggest that combining tools may increase the sensitivity of the TIP detection, although this may come with the cost of decreasing precision, as false positives are highly tool-specific.

Combining tools to improve sensitivity

Our previous results suggest that a combination of tools could be useful to increase the sensitivity in identifying non-reference transposon insertions. To this end, we combined the predictions of PopoolationTE2 (the overall best performer) sequentially with up to four tools selected based on their sensitivity and/or precision levels. As a general trend, the combination of tools led to higher sensitivity levels, reaching more than 90% for both MITEs and LTR-retrotransposons at 40X coverage when combining five different tools (Fig. 7). However, the increase in sensitivity comes with a decrease in precision, particularly clear for LTR-retrotransposons, that approaches 10% for 40X coverage when combining five different tools. The results presented suggest that the combination of two tools provided the best balance between sensitivity and precision. Specifically, the combination of zygosity-filtered PopoolationTE2 and Teflon for MITEs reached 82.1% sensitivity and 97.4% precision at 40X. Regarding LTR-retrotransposons, the combination of zygosity-filtered PopoolationTE2 and Jitterbug reached 75% sensitivity and 86.7% precision at 40X.

As already mentioned, McClintock is an available pipeline that combines several tools. Therefore, we compared the performance of the combination of tools here proposed with that of the McClintock pipeline, which combines the use of Retroseq, TEMP, TE-locate, PopoolationTE and ngs_te_mapper (we excluded RelocaTE from the pipeline due to excessive running time). The combination of tools here proposed (PopoolationTE2 and Jitterbug for LTR-retrotransposon insertions and PopoolationTE2 and Teflon for MITEs) yielded consistently a better sensitivity and much better precision and F1-scores than McClintock at all coverages (especially in the case of LTR-retrotransposons, Fig. 8). The most



important differences were found in precision at intermediate and high coverages. As an example, for MITEs at 40X PoPoolationTE2-Teflon had 97.4% precision whereas McClintock had 83.8% (Fig. 8a). Regarding LTR-retrotransposons at 40X, PoPoolationTE2-Jitterbug precision was 86.7%, whereas that of McClintock dropped to 9% (Fig. 8b).

Evaluation of best-performing tools using *Drosophila* and human datasets

In order to evaluate whether the benchmarking results using rice data could be extrapolated to data obtained from other species, we benchmarked the best performing tools (PoPoolationTE2, Teflon and Jitterbug) using PCR-validated TIPs from *Drosophila* and humans. The *Drosophila* dataset consisted of 81 TIPs from ten *Drosophila* lines sequenced at an average coverage of 42X [22]. This

dataset contained TIPs from 12 different transposon families, including retrotransposons (LTR and LINE) and cut-and-paste DNA transposons (TIR) experimentally validated by Lerat et al. [31] Merenciano et al. [33] and Ullastres et al. [46] (Additional file 4: Table S3). The human dataset consisted of 148 TIPs obtained from one human individual at a coverage of 20X [44]. This dataset consisted of TIPs related to ALU, SVA and LINE-1 retroelements. In the analysis of human insertions, we also included MELT, as it is the best-established tool for the detection of human TE polymorphisms. The detection levels of PoPoolationTE2 and Teflon in *Drosophila* were moderately high (69.1% of the insertions, Table 3 and Additional file 5: Table S4), and substantially higher than Jitterbug (44.4% of the insertions). Using the combination of the three tools, we were able to detect 81.5% of the insertions. These results are in high concordance with the sensitivity

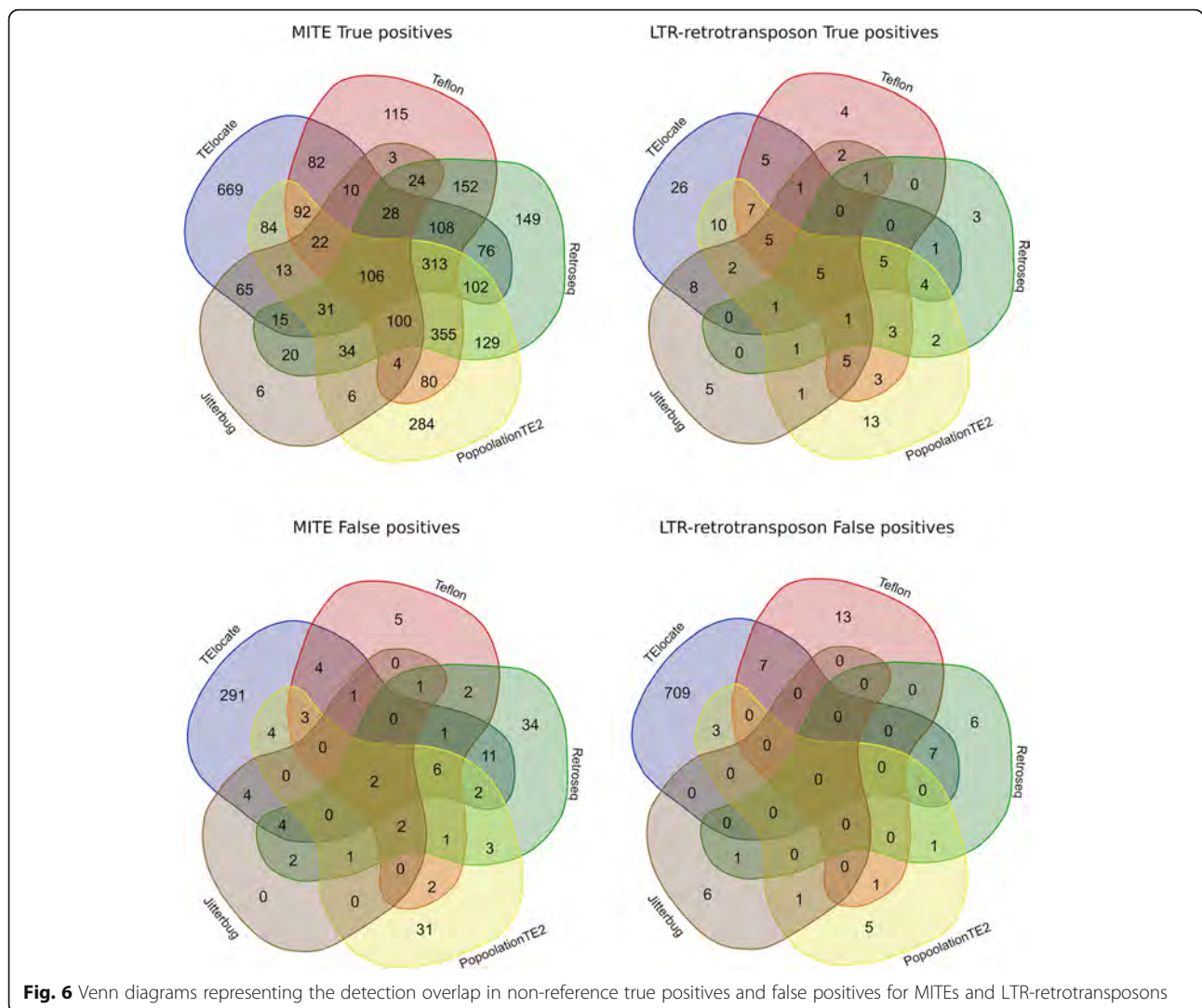


Fig. 6 Venn diagrams representing the detection overlap in non-reference true positives and false positives for MITEs and LTR-retrotransposons

levels found using rice data with LTR-retrotransposons and MITEs, where PoPoolationTE2 and Teflon showed superior detection levels to Jitterbug (Fig. 4). Regarding the human sample, MELT was the best tool identifying homozygous insertions (97.8%, Table 4), whereas PoPoolationTE2 was the best detecting heterozygous insertions (88.2%). Taking into account both kind of insertions, PoPoolationTE2 outperformed MELT, displaying an average detection level of 90.5%. The detection rate of these two programs was higher on human data than in *Drosophila* or rice, where sensitivity levels rarely exceeded 70% using 20X coverage (Fig. 4). The detection levels of Jitterbug were similar to those found using *Drosophila* and rice, ranging from 47.8 to 51%. Teflon was unable to complete the task and the process was killed after five running days. Using the combination of tools, the detection rate increased only 3.4% for the human dataset, reaching up to 93.9% (Table 4).

Running time

Computation time is a limiting factor when running TIP detection tools in large datasets. Therefore, it is an important criterion to take into consideration for selecting the most appropriate tool for a specific experiment. We tested the time needed by the tools to finish the prediction with a 10X dataset and 432 MITE families as input. It is important to mention that three tools (Trackposon, ITIS and MELT) work on a per-family basis. In these cases, the reported time was that needed to finish the prediction for a single family. By contrast, the remaining tools work with all the annotated TE families at the same time. According to our results, Trackposon was the fastest tool, with only 1.7 CPU hours needed to finish (Fig. 9). Among the general tools, ngs_te_mapper, TE-locate and PoPoolationTE2 were the fastest tools, with 8.6, 9.6 and 9.7 CPU hours needed to finish the prediction for the 432 families. RelocaTE2 took the

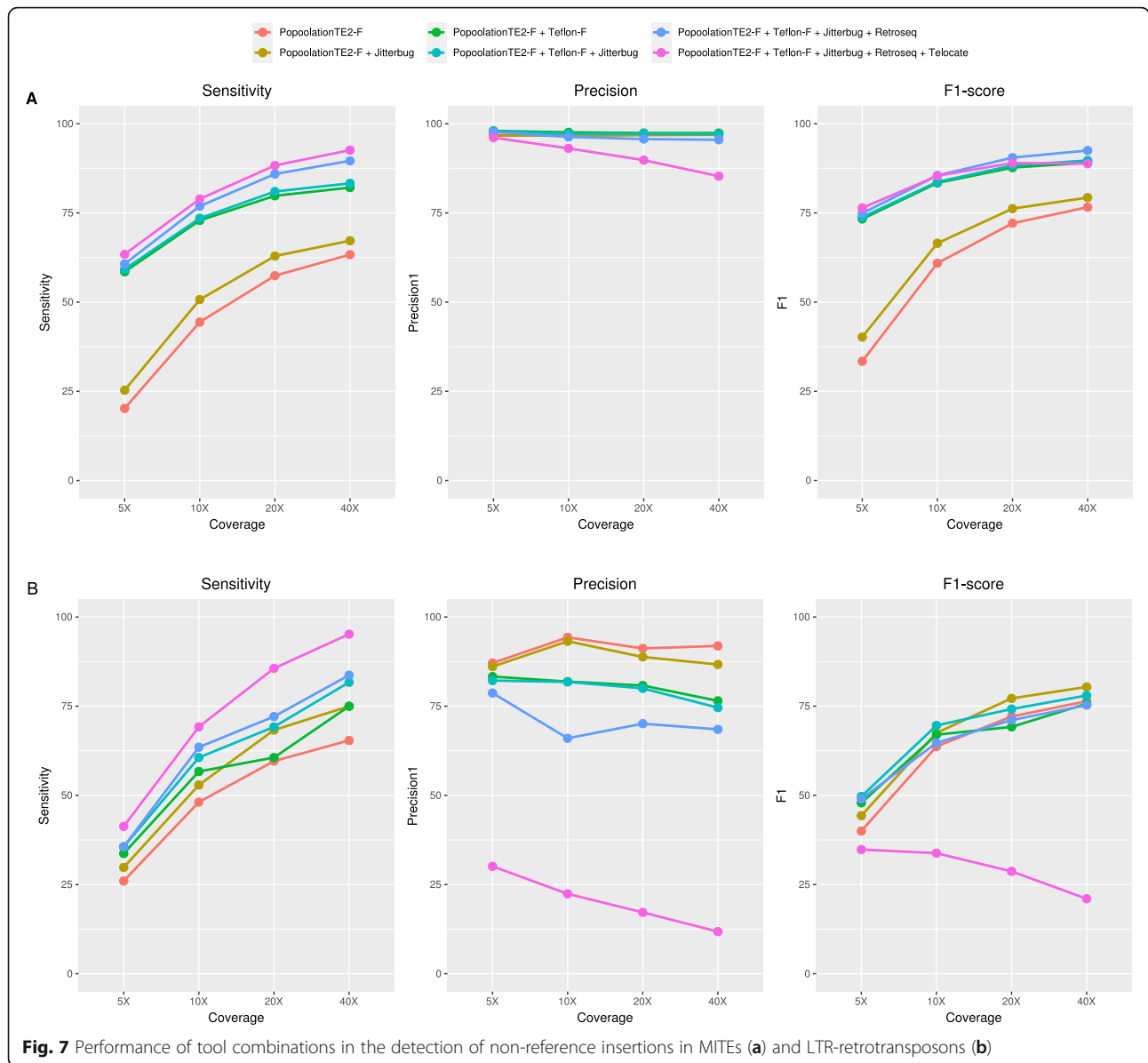


Fig. 7 Performance of tool combinations in the detection of non-reference insertions in MITEs (a) and LTR-retrotransposons (b)

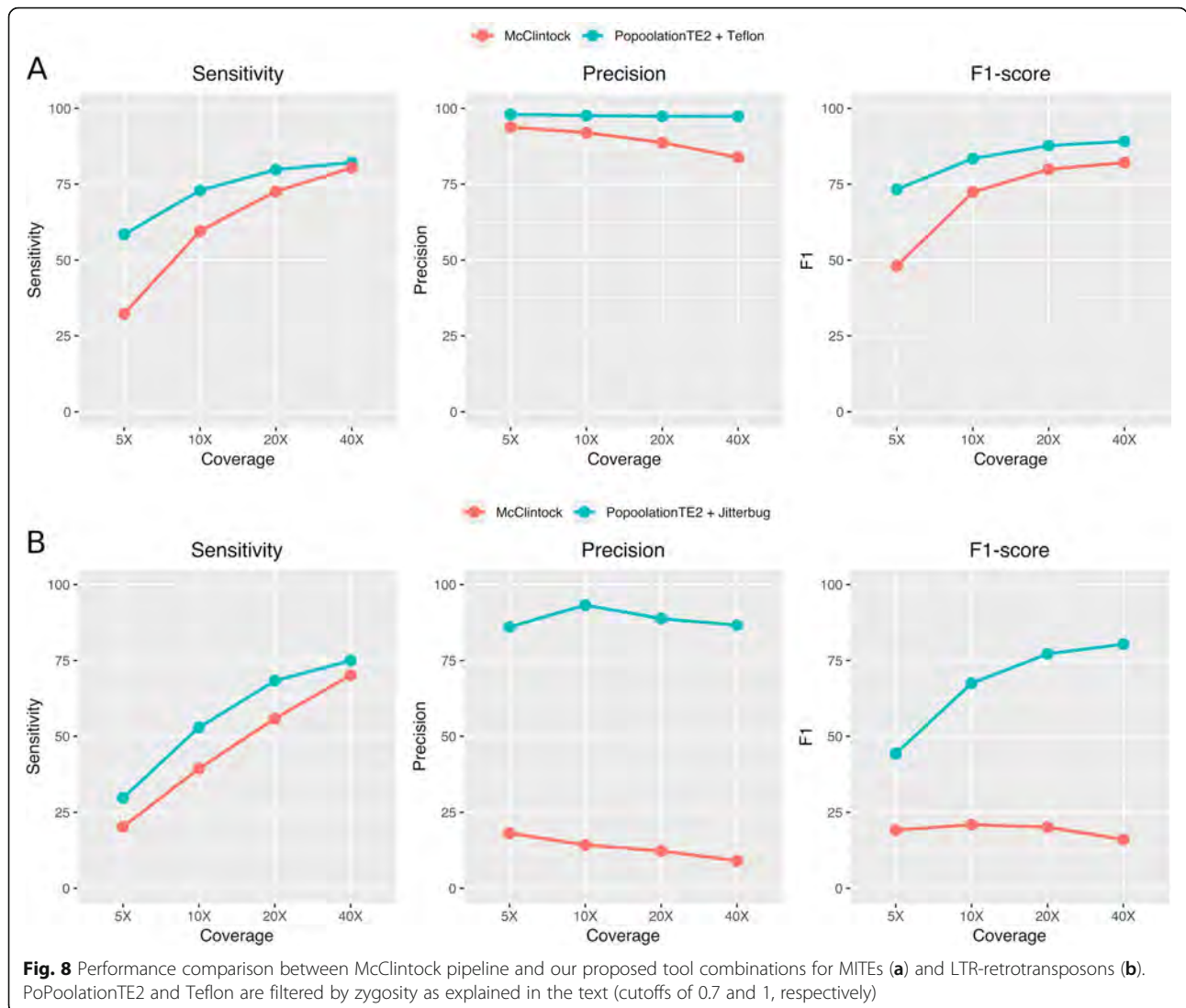
largest amount of time to finish the prediction (59.1 CPU hours) (Fig. 9).

Discussion

The use of real data is essential for an accurate benchmarking of TE insertion detection tools

There are several tools available to detect TIPs from short-read resequencing data, and some efforts have been made to validate the performance of such tools [36, 41]. However, their benchmarking has been essentially based on simulated TE insertions and simulated short reads. It is challenging to perfectly simulate sequencing errors, local coverage variations, biases due to GC content or other genome specific biases that real short-read datasets contain. Similarly, the heterogeneity of real

transposon insertions, with polymorphic truncated or degenerated elements and elements inserted in highly repetitive regions, among other confounding effects, are also difficult to simulate. As a consequence, the benchmarking using simulated data may be overestimating the performance of the TIP prediction tools. Indeed, our results show that, most of the tools here analyzed have a lower sensitivity than previously reported. For example, RelocateTE2 and TEMP were previously benchmarked on simulated rice data, and the sensitivity of both tools was estimated to be higher than 99% at 10X [11]. On the contrary, our results using a dataset of real insertions and real short-read data show that both programs perform very different, with TEMP having a maximum sensitivity of only 13.3% for MITE detection and



RelocateTE2 showing a 35.6% sensitivity. Similarly, we previously reported a sensitivity of close to 90% for Jitterbug, a program developed in our laboratory, using real short reads on simulated TE insertions [21]. Our results now show that for the dataset analyzed (real TIPs and real short reads) the maximal sensitivity is of 32.7% (Fig. 4, LTR-retrotransposons), although it does so with a relatively high precision. Therefore, our results suggest

that the sensitivity and precision previously reported for TIPs detection tools, determined using simulated data, are probably overestimated and that the real performance of these tools is probably lower. We think that the performance levels of the different tools presented here are a much better estimation of their detection ability on real datasets. It is important to note, however, that depending on the genome to be analyzed, parameters used

Table 3 Number of insertions detected by PoPoolationTE2, Jitterbug and Teflon using a validated *Drosophila melanogaster* dataset

	RAL-737	RAL-40	RAL-801	RAL-802	RAL-850	RAL-502	RAL-508	RAL-491	RAL-235	RAL-21	TOTAL	%
Validated insertions	17	16	9	7	4	5	7	5	4	7	81	
PoPoolationTE2	12	5	9	5	3	3	6	5	3	5	56	69,1
Jitterbug	11	2	3	5	2	2	4	2	3	2	36	44,4
Teflon	12	6	9	4	3	4	5	4	4	5	56	69,1
Combination	15	6	9	7	3	4	7	5	4	6	66	81,5

Total number of insertions detected by each tool on each line is provided in Additional file 5: Table S4

Table 4 Number of insertions detected by Jitterbug, MELT and PoPoolationTE2 using a validated human dataset

Tool	Homozygous	Heterozygous	Total
Validated insertions	46	102	148
PoPoolationTE2	44 (95,7%)	90 (88,2)	134 (90,5%)
Jitterbug	22 (47,8%)	52 (51,0%)	74 (50,0%)
Teflon ^a	–	–	–
MELT	45 (97,8%)	84 (82,4%)	129 (87,2%)
Combination	45 (97,8%)	94 (92,2%)	139 (93,9%)

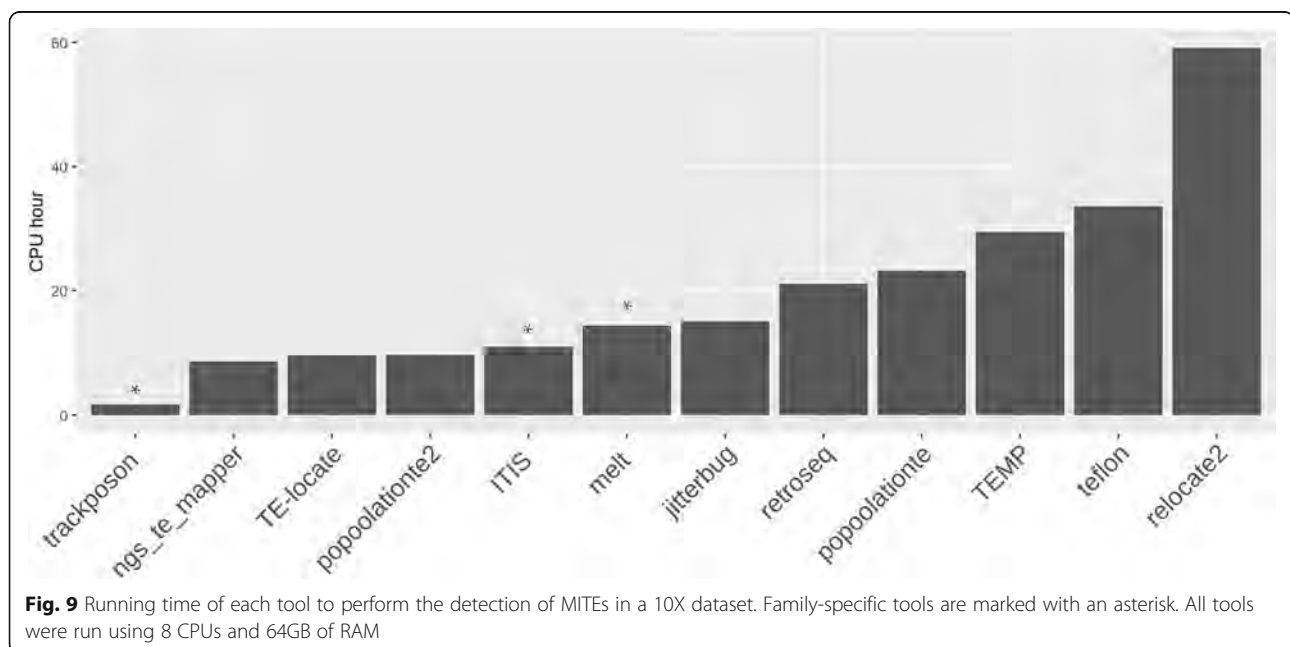
^a Teflon was killed after 5 days running with 12 CPU and 300GB of RAM
 Total number of insertions detected: PoPoolationTE2 (ref and non-ref) = 186,038; Jitterbug (non-ref) = 624; MELT (non-ref) = 1297

and especially on the quality of the annotation of the reference genome the performance of the programs may vary. All the programs benchmarked here are based on the detection of discordant paired-end reads and/or split-reads at the junction of TE insertions. Among the different confounding factors that can interfere with the detection process, the quality of the TE annotation of the reference genome and in particular of the proper definition of the TE-genome junctions, is an important one. Therefore, it is important to work on refining the annotation of the TEs (or at least the more interesting TE families for the purpose of the study) before searching for TIPs.

Tool performance varies depending on TE family

Eukaryote genomes contain a high diversity of TE elements with very different copy numbers and functional and structural characteristics, which may impact on the ability of TIP detecting programs to reliably identify

their insertions. Because of that, we decided to benchmark the different programs using two very different types of TEs that, in addition, are the most prevalent in plants: MITEs and LTR-retrotransposons. The results presented here show that, as expected, the analyzed tools do not detect different TE types with the same sensitivity and precision. MITEs and LTR-retrotransposons represent extreme examples based on their length and complexity, and the performance of the tools when used with other TEs will probably be in the range of this case-study. The analysis of the sensitivity of the best performing tools in detecting TIPs produced by different types of transposons (including LINEs, LTR-retrotransposons and cut-and paste TIR transposons) in *Drosophila* and humans suggests that this is indeed the case. Our results indicate that MITEs are detected with better sensitivity and precision than LTR-retrotransposons. The difference is especially relevant in the detection of non-reference insertions, where most tools show low precision levels for LTR-retrotransposons. In the present study, we ran all samples in default mode or using the parameters described by the authors in the corresponding manuscripts or manuals (Additional file 1). Nevertheless, we show that precision can be increased by applying specific filters to the results. For example, we show that, for some programs, LTR-retrotransposon detection can be drastically improved by applying a zygosity filtering. Applying such filtering may be a good strategy when not intending to study somatic insertions which should in most cases be heterozygous. The difficulties of detecting LTR-retrotransposons come from the complexity of the



elements and from the local regions where they insert. It is known that LTR-retrotransposons (especially those of the Gypsy superfamily) tend to integrate in heterochromatic regions enriched in other TEs. These repetitive regions are likely a source of false positives that affects all the programs tested. These repetitive regions are, in fact, difficult to annotate and polymorphisms within these regions may be challenging to detect even using long-read data or when aligning good-quality assemblies. By contrast, MITEs tend to integrate close to genes [25] and their flanking regions are more likely to be unique in the genome. The presence of non-repetitive TE flanks greatly simplifies the detection of TIPs, as the probability of finding multimapping reads in these regions is minimal.

Another important consideration related to the different TE families is the quality of the annotation. MITEs are easy to annotate and usually have well defined boundaries. By contrast, LTR-retrotransposons form nested insertions and are often degenerated. This makes very difficult to accurately define their boundaries, and as a consequence many chimeric elements are usually annotated. As already mentioned, an accurate TE annotation is essential to increase the capacity of the tools to identify TE insertions based on short-read data. In this context, it could be a good strategy to identify and remove chimeric transposons from the annotation prior to using any of these tools (ie, when working with consensus or with the actual annotation). A chimeric or nested transposon, for example an LTR-retrotransposon with a MITE inserted inside, will be targeted by reads arising from the two elements, and other MITE insertions of the same family present elsewhere in the genome could be wrongly identified as LTR-retrotransposons insertions by the TIP detection tools.

Influence of the type of genome on the performance of the tools

The ability of any of the tools to detect TIPs depend on the nature of the transposon insertion itself and its flanking genome sequence, and none of them can detect new transposon insertions in repetitive regions. Therefore, in addition to the type of transposon generating the TIP, as already discussed, the performance of the tools may depend on the genome under study. For this reason, we have analyzed the sensitivity of the tools that performed the best using rice data on *Drosophila* and human data and compared their performance on the different datasets. The sensitivity of the different programs analyzed in *Drosophila* was very similar to the one obtained in rice. As the genomes of rice and *Drosophila* are relatively different, the former being much bigger (430 Mb vs 175 Mb) and with a higher content of repetitive sequences (37% vs 20%), this suggests that the

performance of the tools is relatively independent of the genome used, and that the benchmarking here presented could be of use for TIP analysis in many different systems.

This analysis also showed that the tools that performed best on rice had an even better sensitivity on human data. The difference of sensitivity was particularly clear for PoPoolationTE2 and MELT. Although this could indicate a difference of the performance of these tools in the two genomes, it could also be due to the particular nature of the human dataset. Indeed, the dataset of validated TIPs in humans contains insertions from TE families (LINE-1, ALU, SVA) that were detected in the first place using only one method, based on split-read and read-pair information [44] and therefore the sensitivity of the programs on this dataset could be overestimated. It is worth mentioning that the PCR-validated *Drosophila* and human insertions have been predicted using a small number of tools in the original publications, and therefore it includes only a subset of all the insertions present in these genomes. Moreover, the human and *Drosophila* datasets were validated by PCR, which could have introduced a bias in the TEs that were included in these datasets. However, note that the number of families included in the human and *Drosophila* validation datasets are similar or bigger than the ones included in the rice dataset and contain both full-length and truncated TEs.

Sequencing coverage critically impacts TIP detection

Independently of the different performance found between TE families, we found that coverage has a major impact on tool performance for all the TE families tested. In general sensitivity increases with increasing coverage. Therefore, homogenization of sample coverage is essential when using TIPs prediction tools to quantitatively compare the transposition rates between organisms or populations. Some tools like PopoolationTE2 have internal steps to carry out this task. Nevertheless, for qualitative studies coverage homogenization is discouraged as down-sampling high-coverage datasets leads to a smaller number of detected insertions. It is important to note that the increase of sensitivity with increasing coverage comes, in most cases, with a decrease in precision. Therefore, depending on the goals of the study a different level of coverage may be suitable. From the data presented here it seems that a coverage below 20X is probably not suited for most analyses, as the probability of missing true insertions is very high.

Strategies to increase tool performance

The fact that an important fraction of the insertions detected by the different tools are not shared supports the fact that combining different tools may increase the

quality of the results [36]. However, simply increasing the number of tools does not necessarily increase the quality of predictions, due to the accumulation of tool-specific false positives (ie, the combination of five tools yielded 95% of sensitivity but only 11.8% precision in non-ref LTR-retrotransposon detection, Fig. 7). This is due to the fact that whereas many true insertions are detected by several tools, most false positives are tool-specific (Fig. 6). Combining a limited number of well-performing tools may be the best approach. Indeed, our results show that with the dataset used, the combination of PoPoolationTE2 and Jitterbug to detect LTR-retrotransposon insertions, or PoPoolationTE2 and Teflon to detect MITEs yielded superior TIP annotations (better F1-score) than the tools alone. Also, the performance of these tool combinations was better than that of the McClintock pipeline, especially regarding LTR-retrotransposons. In this sense, we recommend combining tools based on their high precision and not only on their high sensitivity (ie, PoPoolationTE2 and Jitterbug). Nevertheless, there can be situations in which sensitivity has a priority over precision (ie, re-sequencing of a single individual, or interest only on a few families). In such cases, running more tools can be an alternative and manual curation should be considered.

Selecting the appropriate tools for detecting TE insertions in resequencing data

Depending on the objective of the analysis, a family-specific tool could be more interesting than a broad-spectrum tool. For example, when tracking the effect of certain treatment in a concrete set of elements. Another important consideration is that the amount of storage needed is smaller in comparison to broad-spectrum tools, due to the smaller size of the alignment files. For such cases, a tool such as Trackposon could be a good option due to its fast speed, moderate sensitivity and high precision. Nevertheless, as a drawback, Trackposon does not report the exact insertion point and, which could be a limitation for some studies. In those cases, MELT can be an interesting alternative, although it requires adjusting family-specific parameters to produce high-quality results. This might be indeed the cause why MELT did not perform well on the detection of rice MITEs. In general, it is possible that the tools analyzed here, which were not specifically designed for MITEs and LTR-retrotransposons, may work better for other types of TEs or with modifications in the parameters used. Based on our results, if the objective of the study is to analyze insertions of more than one family, and the storage space is not a major limitation, using some of the top broad-spectrum tools such as PoPoolationTE2 is probably a better option as those programs can also be

relatively fast and show high sensitivity and precision independently of the species and TE type analyzed.

Conclusions

Besides the important efforts of tool developers, our results suggest that the identification of TIPs is still challenging. We propose here a number of approaches, such as combining tools, which can be followed depending on the purpose of the study and the TE families to be analyzed, that can provide good results. However, it is important to note that in the best scenario (combining optimal tools at best coverage, Fig. 7) and having a good TE annotation of the reference genome, the sensitivity could be around 70% with a precision of 80–90% for non-reference insertions. These numbers may be enough for most studies, but it is important to keep in mind that some insertions will be missed, especially when estimating insertion frequencies or when using TIPs for GWAS, for example. There are major limitations like the length of the reads that can be resolved with current technologies (ie long-read sequencing) and will certainly improve in the following years. But there is still the need to develop new algorithms specifically designed to identify TIPs from long reads, to generate highly curated TE annotations of reference genomes and also more independent benchmarks on real data to evaluate the performance of tools under different conditions.

Methods

Sequence data used

We used the available data for the japonica Nipponbare (GCA_000005425.2) and the indica MH63 (GCA_001623365.1) assemblies, and the short-read resequencing of MH63 (SRX1639978), which were used to generate the original assembly.

MITE annotation

MITE-hunter [19] was run on Nipponbare and MH63 assemblies to detect MITEs families, which were then combined with the high-quality predictions available in PMITE database [9] (only families carrying TSD). Clustering at 90% was performed to remove redundancy using cd-hit [17] and produce a final library. RepeatMasker (<http://www.repeatmasker.org/>) was run to annotate all regions having significant homology with any of the MITE families. The annotations were further screened to discriminate full-length elements (consensus length \pm 20%) from truncated hits.

LTR-retrotransposon annotation

LTR-retrotransposons were identified by running LTRharvest [14] on IRGSP and MH63 assemblies with default parameters. The internal conserved domains of these elements were obtained running hmmscan [13],

and only coding elements were retained for further analyses. The identified elements were clustered with Silix [34] according to the 80–80 rule. All the elements in each family were aligned with Mafft [26] and trimmed with Trimal [6]. Consensus sequences were built from the alignments using the EMBOSS package [40].

Determination of benchmarking standards

We took advantage of the availability of two high quality rice genome assemblies (IRGSP and MH63, the former used as reference) in order to obtain a curated dataset of real “reference” (orthologous) and “non-reference” (specific to MH63) insertions as explained in Fig. 2. Mapping of reference and non-reference windows to MH63 genome was performed using Bbmap (<https://sourceforge.net/projects/bbmap/>). Intersections between annotations were done with BEDtools [38].

Drosophila and human benchmarking datasets

The *Drosophila* dataset consisted of 81 TIPs from ten *Drosophila* lines sequenced at an average coverage of 42X [22], and validated by PCR by Lerat et al. [31], Merenciano et al. [33] and Ullastres et al. [46] (Additional file 4: Table S3). In Lerat et al. [31], TIPs were predicted using TIDAL [39] and PoPoolationTE2 [29] using 14 European *D. melanogaster* pooled populations (average coverage of 90X). Briefly, validated TIPs were present in the DGRP population and at least in one European population at >10% frequency, not present in the Y chromosome, and with a predicted length of <6 kb to avoid problems with PCR amplification. In Ullastres et al. [46], TIPs were predicted by TIDAL in the DGRP population [39]. Validated TIPs were inserted in regions with recombination rates >0, and present in at least 15 DGRP strains. Finally, in Merenciano et al. [33] TIPs were also predicted by TIDAL in the DGRP population [39] and all belonged to the *roo* family. Both full-length and truncated copies were validated, as no TE length filter was applied.

The human dataset consisted of 148 TIPs obtained from a human individual (NA12891, SRA accession SRX207113) [44]. Original sequencing coverage of the human genome was down sampled to 20X.

TIP prediction

Predictions of transposon insertions were done using the 12 tools shown in Table 2 using the default parameters and / or following the recommendations of the authors. The scripts used for running each of the tools are shown in Additional file 1.

Evaluation parameters

We used the following parameters for evaluating the ability of each tool to detect MITEs and LTR-

retrotransposons: True positives (TP): Insertions detected by any tool matching with our curated dataset of TPs. False positives (FP): Insertions detected by any tool matching with our curated dataset of FPs. False negatives (FN): Insertions present in our curated dataset of TPs, not detected by the evaluated tool. These primary parameters were used for calculating the final benchmarking ratios that have been previously used for assessing the performance of similar tools [41].

$$\text{Sensitivity} = \text{TP} / (\text{TP} + \text{FN}).$$

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

$$\text{F1-score} = 2 \times [(\text{Precision} \times \text{Sensitivity}) / (\text{Precision} + \text{Sensitivity})]$$

Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s13100-019-0197-9>.

Additional file 1. Scripts used to run all TIP detection tools. (.sh) (SH 13 kb)

Additional file 2 : Table S1. Insertion dataset used for benchmarking. Contains all the TP reference and non-reference windows, as well as all FP windows. (.xlsx)

Additional file 3 : Table S2. Numerical benchmark results. (.xlsx)

Additional file 4 : Table S3. *Drosophila melanogaster* TE insertions validated by PCR. TE names are provided in the table when authors gave names to the non-reference insertions. All insertions were validated based on PCR band sizes. Validated insertion sites are provided when authors sequenced PCR bands evidencing the presence of a particular insertion. ND: not determined. (.xlsx)

Additional file 5 : Table S4. Total number of insertions detected by PoPoolationTE2, Jitterbug and Teflon in ten *Drosophila* lines. (.xlsx)

Additional file 6 : Figure S1. Number of MH63 reference and non-reference insertions detected by direct comparison of 1000 LTR-retrotransposon flanking sites of different sizes from MH63 and Nipponbare genomes. (.pdf)

Additional file 7 : Figure S2. Application of zygosity filtering to PoPoolationTE2. PoPoolationTE2-F means that it was run and filtered at zygosity 0.7. PoPoolationTE2-R corresponds to the raw results. (.png)

Acknowledgements

Not applicable.

Authors' contributions

FB ran Jitterbug, MELT and produced the LTR-retrotransposon annotation. PV ran all TIP detection tools and analyzed the data. RC produced the MITE annotation and the final dataset of TE insertions and analyzed the data together with PV. MM and JG participated in the benchmarking using *Drosophila* and human datasets. JMC and RC conceived the study and wrote the manuscript. All the authors reviewed the final manuscript. All authors read and approved the final manuscript.

Funding

This work was supported in part by grants from the Ministerio de Economía y Competitividad (AGL2016–78992-R). Fabio Barteri and Pol Vendrell hold a FPI (Formación de Personal Investigador) fellowship from the Spanish Ministerio de Economía y Competitividad. Raúl Castanera holds a Juan de la Cierva Postdoctoral fellowship from the Spanish Ministerio de Economía y Competitividad. JG is funded by the European Commission (H2020-ERC-2014-CoG-647900) and the Spanish Ministerio de Ciencia, Innovación y Universidades/AEI/FEDER, EU (BFU2017–82937-P).

Availability of data and materials

The datasets analyzed during the current study are available in the NCBI repository:

- Nipponbare Assembly: GCA_000005425.2
- MH63 assembly: GCA_001623365.1
- Short-read resequencing data of MH63: SRX1639978
- Human resequencing reads: SRX207113
- *Drosophila* resequencing reads: PRJNA36679

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Centre for Research in Agricultural Genomics CSIC-IRTA-UAB-UB, Campus UAB, Edifici Crag, Bellaterra, 08193 Barcelona, Spain. ²Institute of Evolutionary Biology (CSIC-Universitat Pompeu Fabra), Passeig Maritim Barceloneta 37-49, 08003 Barcelona, Spain.

Received: 28 June 2019 Accepted: 17 December 2019

Published online: 30 December 2019

References

1. Adrion JR, Song MJ, Schrider DR, Hahn MW, Schaack S. Genome-wide estimates of transposable element insertion and deletion rates in *Drosophila melanogaster*. *Genome Biol Evol.* 2017;9(5):1329–40.
2. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. *Nat Rev Genet.* 2011;12(5):363–76.
3. Bogaerts-Márquez M, Barrón MG, Fiston-Lavier A-S, et al. T-lex3: an accurate tool to genotype and estimate population frequencies of transposable elements using the latest short-read whole genome sequencing data. *Bioinformatics.* 2019; btz727.
4. Butelli E, Licciardello C, Zhang Y, et al. Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. *Plant Cell.* 2012;24(3):1242–55.
5. Cao Y, Chen G, Wu G, et al. Widespread roles of enhancer-like transposable elements in cell identity and long-range genomic interactions. *Genome Res.* 2019;29(1):40–52.
6. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics.* 2009;25(15):1972–3.
7. Carpentier M-C, Manfroi E, Wei F-J, et al. Retrotranspositional landscape of Asian rice revealed by 3000 genomes4. *Nat Commun.* 2019;10(1):2.
8. Carr M, Bensasson D, Bergman CM. Evolutionary genomics of transposable elements in *Saccharomyces cerevisiae*. *Plos One.* 2012;7(11):e50978.
9. Chen J, Hu Q, Zhang Y, Lu C, Kuang H. P-MITE: a database for plant miniature inverted-repeat transposable elements. *Nucleic Acids Res.* 2014; 42(Database issue):D1176–81.
10. Chen J, Lu L, Benjamin J, et al. Tracking the origin of two genetic components associated with transposable element bursts in domesticated rice. *Nat Commun.* 2019;10(1):641.
11. Chen J, Wrightsman TR, Wessler SR, Stajich JE. RelocaTE2: a high resolution transposable element insertion site mapping tool for population resequencing. *PeerJ.* 2017;5:e2942.
12. Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: from conflicts to benefits. *Nature Reviews. Genetics.* 2017;18(2): 71–86.
13. Eddy SR. Accelerated profile HMM searches. *PLoS Comput Biol.* 2011;7(10): e1002195.
14. Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics.* 2008;9:18.
15. Ewing AD. Transposable element detection from whole genome sequence data. *Mob DNA.* 2015;6:24.
16. Flutre T, Duprat E, Feuillet C, Quesneville H. Considering transposable element diversification in de novo annotation approaches. *Plos One.* 2011; 6(1):e16526.
17. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics.* 2012;28(23):3150–2.
18. Gardner EJ, Lam VK, Harris DN, et al. The Mobile element locator tool (MELT): population-scale mobile element discovery and biology. *Genome Res.* 2017;27(11):1916–29.
19. Han Y, Wessler SR. MITE-hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res.* 2010;38(22):e199.
20. Hénaff E, Vives C, Desvoyes B, et al. Extensive amplification of the E2F transcription factor binding sites by transposons during evolution of Brassica species. *Plant J.* 2014;77(6):852–62.
21. Hénaff E, Zapata L, Casacuberta JM, Ossowski S. Jitterbug: somatic and germline transposon insertion detection at single-nucleotide resolution. *BMC Genomics.* 2015;16:768.
22. Huang W, Massouras A, Inoue Y, et al. Natural variation in genome architecture among 205 *Drosophila melanogaster* genetic reference panel lines. *Genome Res.* 2014;24(7):1193–208.
23. International Rice Genome Sequencing Project. The map-based sequence of the rice genome. *Nature.* 2005;436(7052):793–800.
24. Jiang C, Chen C, Huang Z, Liu R, Verdier J. ITIS, a bioinformatics tool for accurate identification of transposon insertion sites using next-generation sequencing data. *BMC Bioinformatics.* 2015;16(1):72.
25. Jiang N, Wessler SR. Insertion preference of maize and rice miniature inverted repeat transposable elements as revealed by the analysis of nested elements. *Plant Cell.* 2001;13(11):2553–64.
26. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 2013; 30(4):772–80.
27. Keane TM, Wong K, Adams DJ. RetroSeq: transposable element discovery from next-generation sequencing data. *Bioinformatics.* 2013;29(3):389–90.
28. Kobayashi S, Goto-Yamamoto N, Hirochika H. Retrotransposon-induced mutations in grape skin color. *Science.* 2004;304(5673):982.
29. Kofler R, Gómez-Sánchez D, Schlötterer C. PoPoolationTE2: comparative population genomics of transposable elements using Pool-Seq. *Mol Biol Evol.* 2016;33(10):2759–64.
30. Kofler R, Orozco-terWengel P, De Maio N, et al. PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *Plos One.* 2011;6(1):e15925.
31. Lerat E, Goubert C, Guirao-Rico S, et al. Population-specific dynamics and selection patterns of transposable element insertions in European natural populations. *Mol Ecol.* 2019;28(6):1506–22.
32. Linheiro RS, Bergman CM. Whole genome resequencing reveals natural target site preferences of transposable elements in *Drosophila melanogaster*. *Plos One.* 2012;7(2):e30008.
33. Merenciano M, Iacometti C, González J. A unique cluster of roo insertions in the promoter region of a stress response gene in *Drosophila melanogaster*. *Mob DNA.* 2019;10:10.
34. Miele V, Penel S, Duret L. Ultra-fast sequence clustering from similarity networks with SILiX. *BMC Bioinformatics.* 2011;12:116.
35. Naito K, Zhang F, Tsukiyama T, et al. Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. *Nature.* 2009;461(7267):1130–4.
36. Nelson MG, Linheiro RS, Bergman CM. McClintock: an integrated pipeline for detecting transposable element insertions in whole-genome shotgun sequencing data. *G3.* 2017;7(8):2763–78.
37. Platzer A, Nizhynska V, Long Q. TE-locate: a tool to locate and group transposable element occurrences using paired-end next-generation sequencing data. *Biology.* 2012;1(2):395–410.
38. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics.* 2010;26(6):841–2.
39. Rahman R, Chirn G, Kanodia A, et al. Unique transposon landscapes are pervasive across *Drosophila melanogaster* genomes. *Nucleic Acids Res.* 2015;43(22):10655–72.
40. Rice P, Longden I, Bleasby A. EMBOS: the european molecular biology open software suite. *Trends Genet.* 2000;16(6):276–7.
41. Rishishwar L, Mariño-Ramírez L, Jordan IK. Benchmarking computational tools for polymorphic transposable element detection. *Brief Bioinform.* 2017; 18(6):908–18.

42. Sanseverino W, Hénaff E, Vives C, et al. Transposon insertions, structural variations, and snps contribute to the evolution of the melon genome. *Mol Biol Evol.* 2015;32(10):2760–74.
43. Slotkin RK, Martienssen R. Transposable elements and the epigenetic regulation of the genome. *Nat Rev Genet.* 2007;8(4):272–85.
44. Stewart C, Kural D, Strömberg MP, et al. A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS Genet.* 2011;7(8): e1002236.
45. Studer A, Zhao Q, Ross-Ibarra J, Doebley J. Identification of a functional transposon insertion in the maize domestication gene *tb1*. *Nat Genet.* 2011; 43(11):1160–3.
46. Ullastres A, Merenciano M, González J. Natural transposable element insertions drive expression changes in genes underlying *Drosophila* immune response. *BioRxiv.* 2019, 655225.
47. Zhang J, Chen L-L, Sun S, et al. Building two indica rice reference genomes with PacBio long-read and Illumina paired-end sequencing data. *Sci Data.* 2016;3:160076.
48. Zhuang J, Wang J, Theurkauf W, Weng Z. TEMP: a computational method for analyzing transposable element polymorphism in populations. *Nucleic Acids Res.* 2014;42(11):6826–38.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions



Discussion and complementary results

The work presented in the previous article has been useful to the lab to properly detect TIPs and compare the TE landscape between different samples in several species. For example, we used the knowledge extracted from this study to compare different samples of *P. patens* as it can be seen in the chapters 2 and 3 of this dissertation. We also used it to perform a comparison of the transposition landscape in 1059 *Oryza sativa* varieties (Castanera et al., 2021) (See Annexes), work in which I also participated annotating part of the TIPs in the different varieties. This work has also been useful to other members of the lab to find the best approaches to compare the transposition events between different samples of peach and almond (Alioto et al., 2020).

This work indicated the need of a medium to high coverage (more than 10X) to properly identify the TIPs. Low coverages may be useful to identify SNPs but they make it difficult to properly identify TIPs as only a small fraction of them can be detected. We also identified the main limitations of these approaches, that explain that even at relative high coverages there is an important fraction of TIPs that cannot be identified, mostly those found in highly repetitive regions. Therefore, when using these approaches there is a bias towards non-repetitive regions of the genomes. This can be relevant to consider when performing some studies, such as the comparison of the preference of integration of different TEs.

Thanks to this study we also identified the need to normalize the short-reads coverage among samples to not bias the results to compare the number of polymorphisms between different samples.

Moreover, we also identified that other approaches can be used depending on the goal that one may have in his study. For example, when trying to detect the maximum number of TIPs in a population. In this case, it is possible to combine different tools and filter them using the criteria explained in the benchmark. We have followed this approach in chapter 3 to identify the maximum number of TIPs on *P. patens*.

As supplementary results, we also benchmarked the only available tool, at the moment that this work was done, that use Long Reads to detect TIPs; LORTE (Disdero & Filée, 2017). We took the opportunity that there were long-reads data available for the MH63 variety to test the performance of this tool. We ran LORTE using a coverage of 20X of

Pacbio long-reads (SRA: SRR5456657). LORTE only managed to detect a 22.2% of the manually curated TIPs of LTR-RTs between MH63 and the Nipponbare present in Chr05, while using short-reads we manage to resolve up to 41% of them. Moreover, 97.7% of the TIPs detected by LORTE were also detected by short reads. From this data we concluded that, for the moment the use of long-reads to detect TIPs can still be challenging. It is expected that in the future long-reads data will outperform short-read data to call TIPs, but at the moment there is still a lack of tools to call TIPs using this data and most of the publicly available data are based on short-read data for this reason it is still important to know which the best tools and approaches are to detect TIPs using short-reads resequencing data.

Another relevant information that we realized after performing the benchmark is that the preparation and the kind of short-read libraries used for sequencing may bias the results. For example, in the annotation of the 1059 varieties Raúl Castanera identified the presence of a bias between two libraries (CAAS and IRIS) that were prepared using the same methodologies but in different locations. Similarly, the use of paired-end libraries that have different paired-end fragment sizes can affect the obtained result. This is especially relevant when using Nextera Illumina short-reads libraries that do not have a separation between the two paired-end reads. We realized that in these libraries there was a decrease in the number of discordant reads that directly affected the performance of the tools to properly call TIPs on these samples.

Moreover, it is important to consider in which organism the study is going to be developed and if there are tools previously developed and optimized for this organism, to call the TIPs. In each organism it should be taken into account the resources necessary to call the TIPs depending on the TE content, the genome size and complexity. Despite that, in this work we proved that the best performing tools could be a good option when working with a non-model organism without prior knowledge of their TIPs as the best performing tools had a good performance in the different tested organisms (*Homo sapiens*, *Drosophila melanogaster* and *Oryza sativa*) that have wide differences in terms of TE content and TE distribution. It should also be considered that the benchmark was done comparing two different rice varieties where, apart from the TIPs, there will be other differences in the genome, such as SNPs, INDELS and other structural variants that can difficult the mapping of the reads. The methods benchmarked in this study could have a better performance, probably, when comparing the same sample sequenced different times on

different conditions (control vs mutant or different tumor samples, for example). As we would not expect that many differences on the genomes as in our case.

Overall, the results presented in this study are a good guide for users interested into the detection TEs insertion polymorphism using short-read data. Although we tried to incorporate to the benchmark the most relevant tools that existed at that time, we could not include all the tools that were available. Moreover, since then, new tools have been published. Despite that, the dataset published in this article can be useful to benchmark and compare these tools to the ones already benchmarked by us, indistinctly if they use short-reads or long-reads technologies as both libraries are available for this dataset.

We can conclude that to properly detecting TEs from sequencing data is a challenging process. It requires high resequencing coverage to detect an important fraction of the polymorphisms and despite that, it is still representing a fraction of the total of polymorphisms present in the population. That this is probably due to the results are biased towards the non-repetitive regions of the genome and this explains the difference on performance between different TE orders such as LTR-RT that are usually located far from genes mostly found in repetitive regions (in particular those of the Gypsy superfamily) from MITEs that are located at the gene vicinities in unique sequences.

Chapter 1.4: Detection of Transposable Element transcription from short-read data

Introduction

One of the first steps required for the mobilization of a TE is the expression of the molecular machinery necessary to produce a transposition event. We can use RNAseq approaches to identify TE transcription.

RNA seq data has been reliably used to analyze gene transcription. This is quite straightforward as an important fraction of the gene content in a genome is unique and there are well-established protocols to align the reads, estimate the abundance and compare the levels of expression between different conditions (Love et al., 2014; M. D. Robinson et al., 2009). However, it is not as easy to estimate the expression of TEs. Most TEs can be found in multiple places in the genome as highly similar copies, resulting in a difficult process to properly assign the transcription to these regions of the genome. Moreover, the biggest fraction of TEs in eukaryote genomes are truncated and degenerated copies (Ou et al., 2019). These copies, although they may not be able to transpose, may be included in transcripts, for example, as a readthrough from genes.

When looking for the transcription that can lead to a new transposition event, we should distinguish between the different classes of TEs:

In the case of Class I transposons, the expression of a transcript covering most of the complete sequence of the TE would be required. This transcript will be then retrotranscribed and integrated into a new place of the genome.

The main active class I TEs of plant genomes, LTR-RTs, initiate their transcription within the 5' LTR and terminate it within the 3'LTR, generating a transcript that does not cover the whole element, but contains all its sequence. The transcription of the TE starts at a region localized inside the 5' LTR known as region R followed by a region known as U5 and finish at the 3' LTR transcribing a region known as U3 and a second and identical sequence of the region R (Figure 9A). The complete sequence of the two LTRs is reconstituted during reverse transcription generated using the r-U5 and the r-U3 regions of the transcript as a template to synthesize two identical LTRs (Boeke et al., 1985). In the case of the non-LTR retrotransposons, specifically in LINES, to mobilize an autonomous copy they require the transcription of all the TE until the poly A tail (Figure

9B). Then, the encoded reverse transcriptase and nuclease will be associated to their encoded mRNA and reintroduced to the nucleus. The endonuclease will produce a single strand cut into the genomic DNA. At this point the associated reverse transcriptase will start retrotranscribing the TE from the 3' polyA tail of the TE transcript. Finally, the second strand of the DNA will be produced, generating a new TE copy in the genome.

DNA TEs do not transpose through an RNA intermediate, and therefore the whole TE sequence does not need to be transcribed (Figure 9C). However, to transpose they need to expression of a transposase or other TE-encoded proteins, such as an helicase in the case of Helitrons, to mobilize the TE from one position to another.

A) LTR-Retrotransposon transcription:



B) LINEs transcription:



C) DNA transposon transcription:



Figure 9: Example of different orders of TE transcription. On A) transcription of an LTR-RT that could potentially produce a new transposition event being transcribed from the R region of the 5' LTR to the R region of the 3' LTR. On B) the transcript that should be detected from a LINE Retrotransposon. On C) example of a transcription of a DNA TE, that will express the machinery necessary to mobilize the transposition, in this case a transposase.

Distinguishing between the truncated and degenerated copies expression and the transcription that could lead to a transposition event is not straightforward. Reads can map indistinctly to the truncated copies and to complete copies of a genome, especially when using short-reads data, hampering the analysis of the TE transcription that could lead to a new transposition event.

It should be noted that, even when detecting the transcription of all the required machinery to generate a transposition event we cannot directly deduce that a new transposition event will occur. This should be used just as an indication of under which conditions the TE could transpose. There are other mechanisms, such as transcriptional and posttranscriptional gene silencing, or other posttranscriptional control mechanisms that can interfere in the mobilization of the TE, regulating the transpositional process (Fultz et al., 2015). Moreover, in some cases there could be other mechanisms of replication of TEs that do not require the transcription of any protein derived from the machinery of any TE, such as the mechanisms that have been proposed for the amplification of some TEs like MITEs (Izsvák et al., 1999), not being possible to predict their mobilization based on the transcription of any TE derived sequence.

To try to overcome the different challenges related to TEs expression, multiples tools have been developed such as TETRanscripts (Jin et al., 2015), TETools (Lerat et al., 2017), TESalmon (Jeong et al., 2018) or LIONS (Babaian et al., 2019).

In this part of the chapter, we will focus on the identification of the best approaches to identify the TE expression that could potentially lead to a transposition event. We used three different methods to detect TE transcription in *P. patens*, focusing on the expression of LTR-RT, which represent the 51.4% of all the genome (Lang et al., 2018). Moreover, LTR-RT are better annotated than the other TE orders in this genome, as most of the other TEs from the other orders are old and degenerated TEs.

Among the different RNAseq data available for *P. patens*, we decided to use the heat shock RNAseq experiment published in Perroud et al., 2018 as a test case, as there seem to be a general association of stress and TE activation (Grandbastien, 1998), and in particular heat shock has been shown to induce the activation of some retrotransposons in other plant species, such as in *Arabidopsis thaliana* (Cavrak et al., 2014).

We used three different approaches to analyze the TE transcription. The first approach is based on the mapping of the reads mapped to the reference genome and counting then all these reads that are mapped to the TE fraction to estimate their expression. Although there are different methods that have been published based on this approach, such as TELESCOPE or TEsalmon (Schwarz et al., 2022), we decided to use Tetranscripts (Jin et al., 2015) as it is one of the most used tools to quantify TE expression and we have previous experience using the tool in the laboratory to quantify TE expression in *P. patens* (Lang et al., 2018).

Tetranscripts was published in 2015 to estimate simultaneously gene expression and TE expression from RNAseq data aligned to the reference genome. To do that, Tetranscripts estimates the number of reads mapped to the genes, considering only the reads that map in a single position in the genome (uniquely mapped reads), and at the same time estimates the number of reads mapped to TEs. To estimate the number of TEs reads, it counts the uniquely mapped reads and applies an expectation-maximization algorithm to estimate the contribution of each TE copy to the multimapping reads. The program assigns to each TE copy where the multimapping reads are mapped a value based on the effective length of the reads and the number of mapping places of the reads, and calculates the expression level of the given TEs clusters or families that the user has previously defined. The program will not provide information of the expression at the TE copy level. After that, with all the counts assigned to the TE families and the genes, a differential expression analysis can be done using the same methods used in a classical RNAseq experiment (such as DESeq2 or EdgeR) having at the end the differentially expressed genes and TE clusters as an output of the analysis.

The second selected approach, TETools (Lerat et al., 2017), does not require the alignment of the reads to the genome and can quantify the TE transcription by mapping to a selected TE library. This tool has also been used for non-model organisms and we thought that could potentially help us overcome the problem of the identification of the TE transcription on the highly repetitive regions of the genome.

TETools was designed to analyze only TE expression and not gene expression. This tool is divided in two modules: The first module known as TEcount maps the raw reads to the

sequence of the different TE copies and counts the total amount of reads mapped to each TE cluster, family or subfamily previously defined by the user. The second module, TEdiff, analyzes the TEs that are differentially expressed using DESeq2. This tool does not require the mapping of the reads to the reference genome. This may facilitate the identification and quantification of TE transcription in genomes that have complex and nested TE structures where is difficult to identify the TE copies, only mapping to a selection of the copies previously defined by the user.

Finally, we developed a third approach based on the de novo assembly of the reads mapping to the TE fraction of the genome. We used this approach to identify all the possible TE complete transcripts that can arise a new transposition event and overcome the main limitations that we identified when using the other approaches.

Results

Quantifying LTR-RTs transcription using Tetranscripts

Tetranscripts allows the study of gene expression and TE expression at the same time. As introduced previously, we used the libraries of *P. patens* RNA-seq of protonemata treated with heat shock and protonemata not treated, grown on the same medium (Perroud et al., 2018).

To estimate TE expression, Tetranscripts counts the fraction of reads mapped to the TE annotated sequences of the reference genome. In this case, we used the published LTR-RT annotation from the last version of the genome (Lang et al., 2018). This annotation classified the LTR-RT in 5 different families: 3 Gypsy LTR-RT families (RLG1, RLG2 and RLG3) and 2 Copia LTR-RT families (RLC4 and RLC5).

Using this method, we observed that the LTR-RT family RLG1 was detected as highly expressed in both conditions while the other TE families were detected as lowly expressed in both conditions (Figure 10):

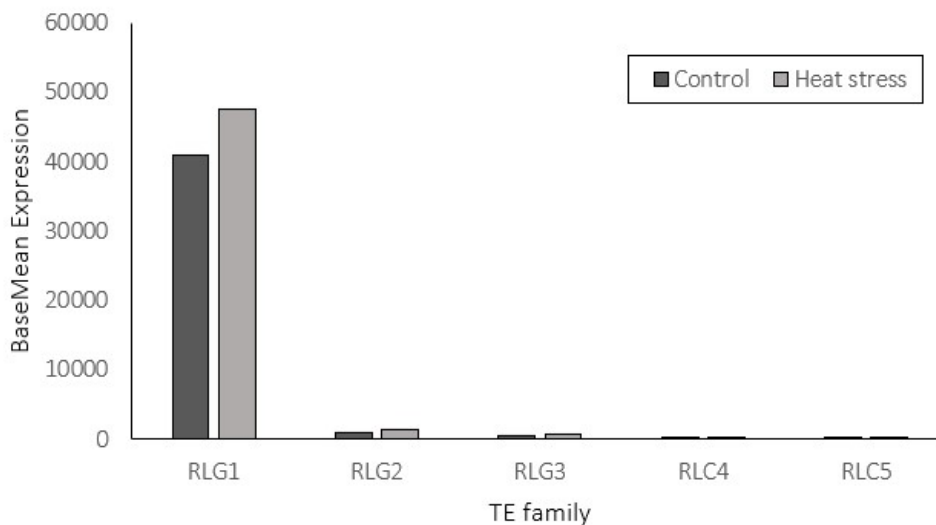


Figure 10: Expression of the different LTR-RT families detected by Tetranscripts expressed in normalized Deseq2 normalized values (BaseMean) between control condition (dark grey) compared to heat stress (light grey).

Moreover, we observed that 4 of the 5 families (RLG1, RLG2, RLG3 and RLC5) were detected as differentially expressed, detecting a small overexpression of the RLG1, RLG2, and RLG3 families under heat stress, and an induction of the RLC5 family in heat

stress with a Log2FC value of 1.46(Table 1). This is barely visible in Figure 10 due to their low expression.

Table 1: Differential expressed LTR-RT families detected between protonemata samples treated with heat shock and protonemata control samples expressed in Log2 Fold Change (Log2FC) values detected by Tetranscripts. Positive Log2FC correspond to samples that are overexpressed in heat stress, negative values samples that are repressed in the heat stress treatment. Differential expressed clusters were selected using a padjusted value of less than 0.05. padjusted value are shown in the right column for each cluster.

LTR-RT Family	log2FC	padj
RLG1	0,22	3,25E-02
RLG2	0,54	4,48E-07
RLG3	0,77	4,03E-12
RLC5	1,46	2,28E-05

The output of Tetranscripts does only give a normalized value of expression for all the genes and TE families and the differentially expressed genes and TEs. But the TE expression value is at the level of the given TE family annotation. In this case we used a LTR-RT annotation that does not separate between the truncated and degenerated fraction of the LTR-RT families that comprised more than half of the genome from the putative complete copies for each family, that are only a small fraction of the TEs in the genome (Table 2).

Table 2: Total number of TEs by each TE family compared to the number of LTR-RT that have a minimum length of 3 Kbp.

LTR-RT Family	Total TEs	LTR-RT >3kb
RLG1	77987	13489
RLG2	19108	2277
RLG3	22708	4252
RLC4	2569	413
RLC5	7786	975

Although we can visualize the alignments of the reads to the reference genome, it is not possible to discriminate between the expression of the truncated or degenerated copies from the expression of complete TEs that could lead to a new transposition event. For this reason, to try to accurately assign the transcription to a given number of possible complete elements, we defined clusters for the different LTR-RT families of the genome to try to define the possible complete copies and estimate the expression only over these elements.

Clustering of the identified LTR-RT families to improve the LTR-RT transcription detection

As explained previously, the published annotation does not distinguish in each family the putatively complete copies from the truncated or degenerated TE copies present in the genome. Moreover, in the annotation there are unclassified copies that could not be assigned to any of the 5 families in the reference TE annotation as they lack part of the domains that could allow a classification to a given family. It is possible that part of these unclassified elements are non-autonomous copies of the TE families. To try to distinguish which groups of copies are transcribed in each family, we have further divided all the different families into clusters using all the copies of the annotation. The clusters were defined based on a threshold of similarity of 80% of identity over 80% of the sequence, with a minimum of three TE copies per cluster. We performed the cluster using SILIX. In total we obtained 47 different clusters. To classify the obtained TE clusters into the previously published TE families, we aligned the consensus sequence obtained by each cluster and using the reverse transcriptase sequence we built a phylogenetic tree. We built one tree for the Copia TEs and one tree for the Gypsy TEs, using the Tos17 rice LTR-RT as an outgroup for the Copia LTR-RT tree and the CRR rice LTR-RT as an outgroup for the Gypsy LTR-RT tree. We observed that some clusters that contained previously unclassified elements with the published TE families (RLG1, RLG2, RLG3, RLC4 and RLC5) could now be associated to some of these TE families (Figure 11 and Figure 12).

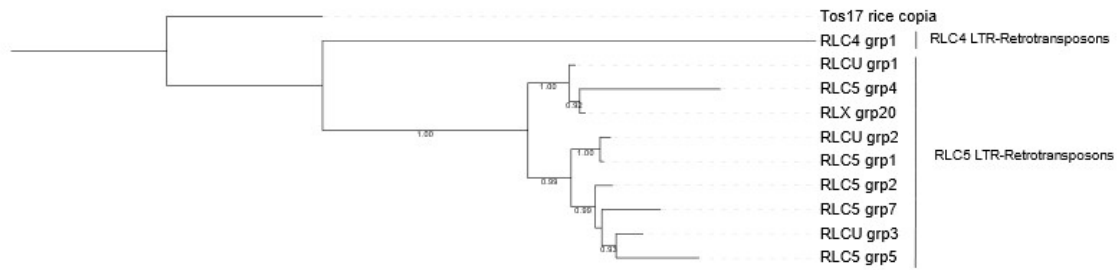


Figure 11: Phylogenetic tree of the different LTR-RTs clusters of the Copia superfamily.

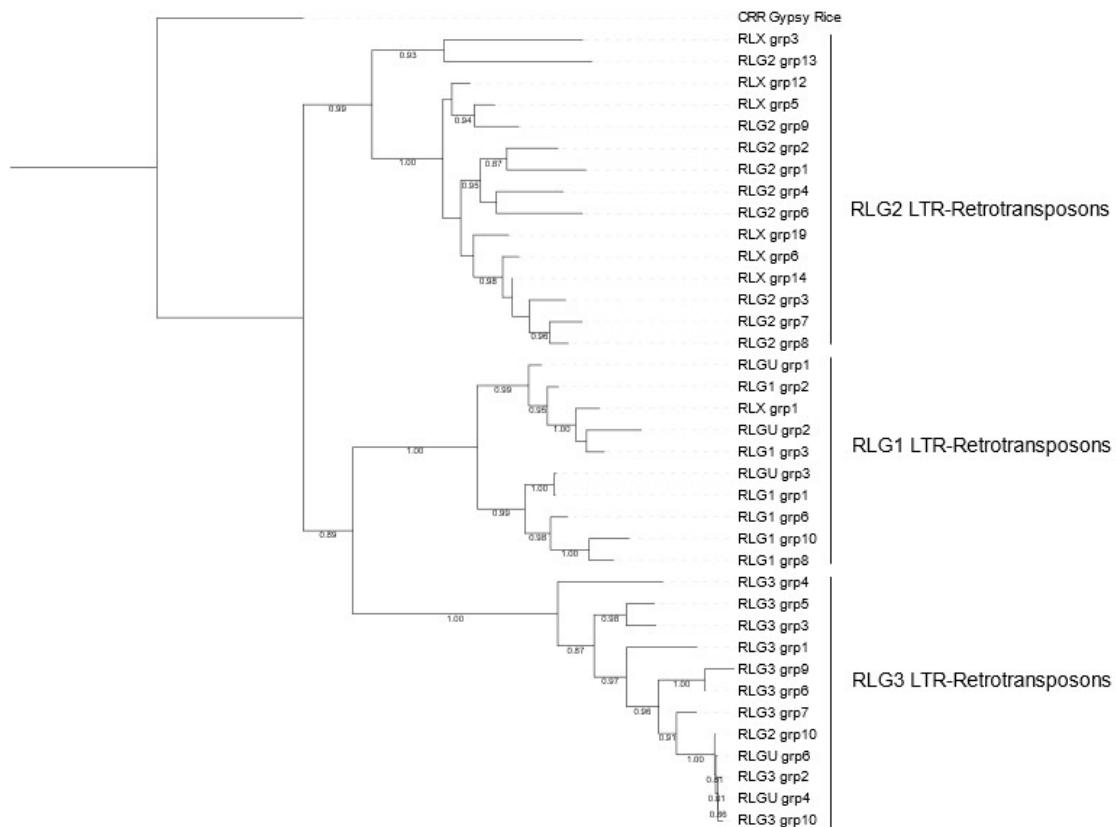


Figure 12: Phylogenetic tree of the different LTR-RT clusters of the Gypsy superfamily .

Quantifying the transcription using Tetranscripts with an improved LTR-RT classification

We quantified the transcription using the new annotation that only contains copies belonging to the 47 clusters of the 5 LTR-RT families. Using this approach, we detected 32 clusters as expressed of all the 5 LTR-RT families (Figure 13).

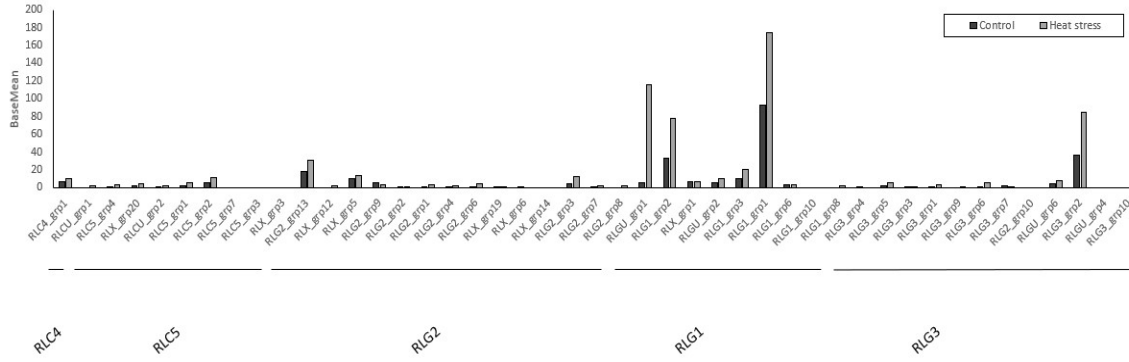


Figure 13: Expression of the different LTR-RT clusters detected by Tetrascripts expressed in normalized Deseq2 normalized values (BaseMean), comparing the expression in the control condition (dark grey) with the heat stress conditions (light grey).

Most of the 32 clusters that were detected as expressed had a low level of expression (Figure 13). The only clusters that were detected with a higher expression belong to the RLG1 family, as previously found, although the expression values are now much lower when compared to the quantification against the complete TE annotation.

We observed an increase of expression between the control condition and the heat stress condition for 10 clusters (Figure 13), but only 5 clusters were detected as differentially expressed between the two conditions from the analysis using DESeq2, belonging to the RLG2, RLG1 and RLG3 families (Table 3).

Table 3: Differential expressed LTR-RT clusters detected between protonemata samples treated with heat shock and protonemata control samples expressed in Log2 Fold Change (Log2FC) values detected by Tetrascripts. Positive Log2FC correspond to samples that are overexpressed in heat stress, negative values samples that are repressed in the heat stress treatment. Differential expressed clusters were selected using a padjusted value of less than 0.05. padjusted value are shown in the right column for each cluster.

TE Cluster	log2FC	Padj
RLG2_grp3	2,43	4,00E-02
RLGU_grp1	18,85	1,84E-11
RLG1_grp2	1,45	1,03E-03
RLG1_grp1	0,81	1,43E-05
RLG3_grp2	-1,24	7,71E-05

The highest induction was observed for the cluster RLGU_grp1 belonging to the RLG1 family. The only cluster that was detected as repressed was the cluster RLG3_grp2 from the RLG3 family that was detected as repressed under heat stress.

Using this approach, we could limit the quantification to a given number of elements from the LTR-RT annotation belonging to the most probable complete elements. Despite that, although it is possible to visualize the mapping of the reads to the regions of the reference genome that belong to the selected elements of each cluster and check if the reads cover most of the TE copies, is nearly impossible to do it for all the clusters due to the high number of elements that are represented in each cluster (for example the cluster RLG1_grp1 contains at least 3189 elements). When we manually inspected a group of copies, visualizing the mapping of the reads to individual copies of the clusters around the genome, we realized that most of the copies were not completely covered by reads. These copies were only partially covered by unique reads and multimapping reads mapping to multiple positions around the genome.

We also realized that following this approach, the multimapping reads that map to the complete copies of the clusters but at the same time map to truncated copies were not included now on the counting. This underestimates the quantification of the expression of these clusters as the value of expression was distributed between all the copies (truncated or not). An example is given in Figure 14:

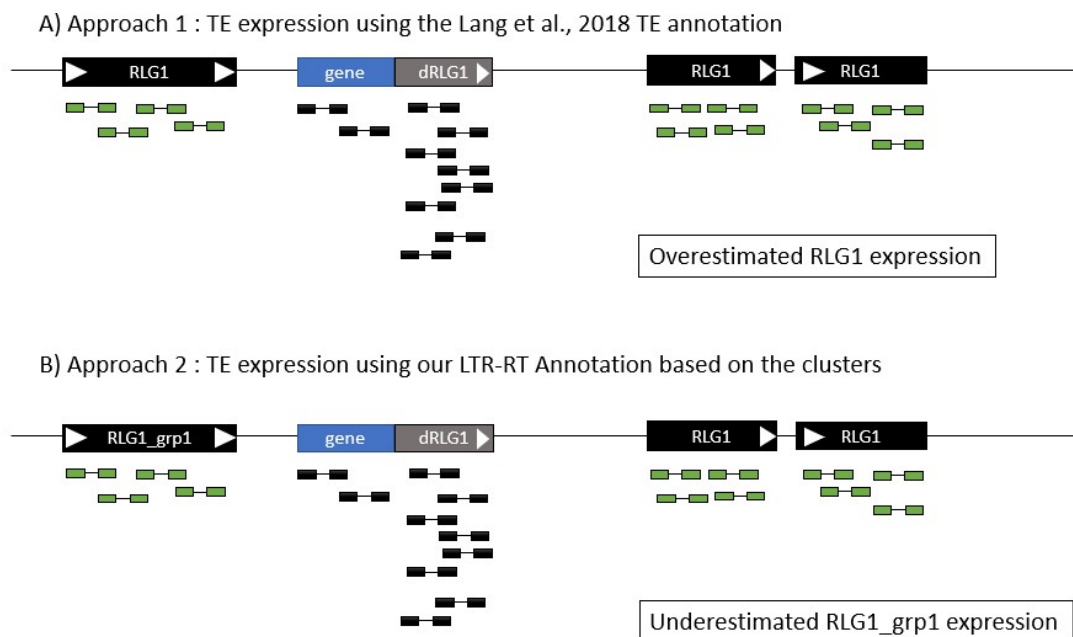


Figure 14: The panel A) represents the first approach used considering the use of the published TE annotation and classification. Represented in black the RLG1 elements, and in grey a degenerated RLG1 element. Under this scheme a representation of the alignment of the paired-end reads. Represented in black the uniquely mapped reads and in green the multimapping reads. When using the complete annotation and counting the expression of all the reads mapping to the RLG1 elements we include the reads mapping to the degenerated copy that is the byproduct of readthrough from a gene leading to an overestimation of the RLG1 expression. Panel B represents the second approach used, using only the annotation and classification of the different clusters. After counting the expression over the cluster RLG1_grp1 we do not count the reads mapping to the degenerated copy located at the end of the gene, as they uniquely map to a copy that is no longer considered part of the annotation, but the multimapping reads that map to the copies of the cluster they can also map to highly similar but truncated copies that are not considered as part of the annotation now, leading to an underestimation of the expression when counting the TE expression over this cluster.

For the above limitations, we decided to test TEtools (Lerat et al., 2017), which is a reference free approach to assess the TE expression. It counts only the reads that map to the selected copies of the 47 clusters and performs a differentially expression analysis only for the 47 different clusters between the two conditions.

Quantifying LTR-RT transcription using TEtools

To use TEtools to quantify the TE expression we first mapped the RNAseq reads to the individual copies belonging to the 47 LTR-RT clusters. TEcounts, the first module of TEtools, estimates the number of reads mapped to each TE cluster. Using the second module, TEDiff, we estimated the number of clusters that are differentially expressed between the two conditions (heat stress vs control).

As the normalization of the reads in this case only considers the number of reads mapped to the selected copies and not to the reference genome, the normalized values are not directly comparable to the ones obtained by TEtranscripts. However, as in this case the reads are only mapped to individual copies (all the copies of each cluster), it is possible to check the alignment of the reads to the individual copies and validate if the reads cover most of the LTR-RT sequences.

Using this method, we observed that 2 clusters, belonging to the families RLG1 (RLG1_grp1) and RLG2 (RLG2_grp13) seemed to be highly expressed compared to the other clusters in both conditions (control and heat stress) (Figure 15). Although we detected expression for 30 LTR-RT additional clusters, the level of expression for these clusters was very low in the two conditions (Figure 15):

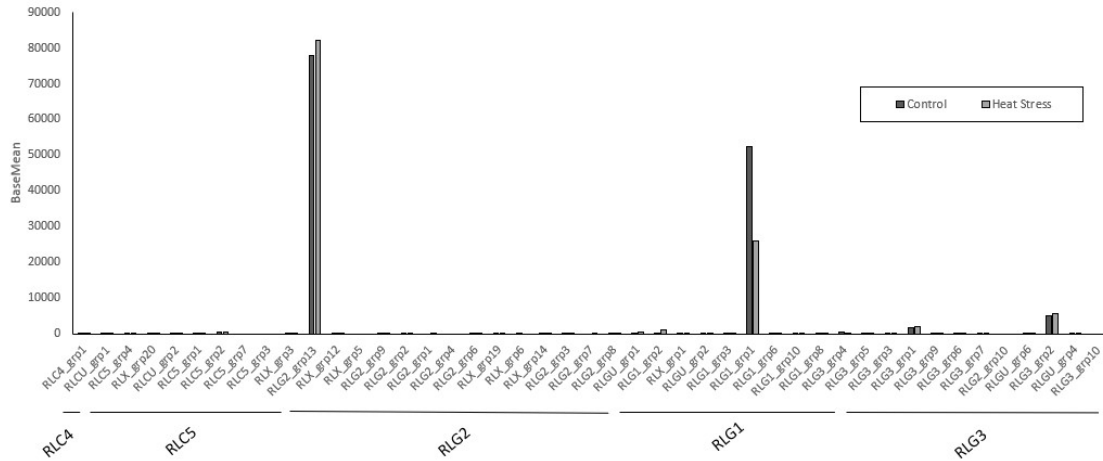


Figure 15: Expression of the different LTR-RT clusters detected by TETools expressed in normalized BaseMean values. In dark grey the expression in the control condition and in light grey the expression in heat stress.

Among the clusters that were detected as highly expressed, only RLG1_grp1 was detected as differentially expressed between the two conditions (Table 4) being repressed under heat stress. Regarding the 30 clusters expressed at low level, 13 were detected as differentially expressed between heat stress and the control condition. 7 clusters were repressed under heat stress while 6 of them were induced under heat stress conditions.

Table 4: Differential expressed clusters detected between protonemata samples treated with heat shock and protonemata control samples expressed in Log2 Fold Change (Log2FC) values detected by TTools. Positive Log2FC correspond to samples that are overexpressed in heat stress, negative values samples that are repressed in the heat stress treatment. Differential expressed clusters were selected using a padjusted value of less than 0.05.

TE cluster	log2FC	Padj
RLC4_grp1	-2,320	1,06E-08
RLC5_grp4	-1,85	8,32E-29
RLX_grp20	1,73	6,27E-03
RLG2_grp1	-4,81	2,86E-03
RLGU_grp2	-3,30	3,39E-06
RLG1_grp3	-4,92	5,80E-08
RLGU_grp1	-6,12	2,42E-06
RLG1_grp1	-1,02	3,99E-05
RLGU_grp3	0,81	1,94E-03
RLG1_grp2	-4,85	8,22E-19
RLG3_grp9	1,60	4,31E-07
RLG3_grp6	1,04	1,72E-04
RLG3_grp4	0,93	2,86E-03

We then further investigated if for the clusters that were detected as expressed the reads were covering most of the sequence of the individual copies in each cluster. We manually inspected the mapping of the reads using IGV to the different clusters observing that in some cases the reads were not covering the complete copies. For several clusters that were detected as expressed the reads were only covering part of the sequences, such as the LTRs. This manual curation is a time-consuming and nearly impossible to perform for those clusters that contain many elements that are highly similar, such as the cluster RLG_grp1 with 3189 elements.

The results obtained using Tetranscripts and TTools are essentially not concordant. For example, we detect expression in both cases for the RLG1_grp1 cluster but in Tetranscripts we detect the cluster as being overexpressed under heat stress and in TTools as being repressed under heat stress. We observed other clusters that were previously detected as expressed using TE transcripts that are now not detected as expressed using TTools. The results obtained between the two methods are not

comparable and with the observed limitations for both methods it is not possible to directly deduce which of the two methods is better to identify the transcription that could arise a new transposition event.

RNaseq de novo TE assembly to identify TE expression

As the main objective of the work presented here is to reliably quantify the TE expression that could be linked to transposition, the methods described above do not seem suitable. For this reason, we opted to develop an approach based on the assembly of short-reads to form contigs that could represent full-length transcripts of TEs, similar to what has previously been used for the analysis of TE expression in humans (Guffanti et al., 2018). These contigs were then compared to TE sequences potentially representing complete TEs in order to select transcripts corresponding to the entire TE sequence (in case of LTR-RTs) or the entire coding region (in case of DNA TEs).

Explained briefly, we first mapped all the RNaseq reads to the TE annotations in the genome. We then extracted the reads that mapped and performed a de novo assembly to form contigs. After that we selected contigs covering most of the length of the potentially full-length elements, in the case of retrotransposons, by mapping the contigs to the TE annotation or containing all the coding domains (such as the transposase sequences), in the case of DNA TEs.

In order to maximize the possibility of identifying the full-length TE transcripts, we pooled the reads from different RNaseq libraries from different developmental and stress conditions available from Perroud et al., 2018. The reads were de novo assembled to contigs using Trinity (Grabherr et al., 2011). We obtained a total of 696 assembled contigs. To identify to which TE clusters or TE families they may belong, all the contigs were aligned to the TE library using BLASTn. 94% of the contigs showed similarity to previously annotated LTR-RTs. Most of the contigs aligned to LTR-RTs (72%) had a length of less than 1000 nucleotides and corresponded to degenerated or fragmented TE copies such as solo LTRs. For the above reason we discarded all the assembled contigs corresponding to LTR-RT of less than 1 kbp, discarding 524 contigs. We kept all the contigs that mapped to other TE classes, such as DNA TEs and LINES, as they were only representing a 0.3% of the total of assembled contigs (20 contigs) and could be easily checked manually.

One of the problems that we observed using this approach is the formation of contigs that contained several repetitions of the same LTR-RT with the same sequence, which do not correspond to structures present in the reference genome. This was an artifact introduced due to the presence of reads covering both R regions of the LTR-RT. As the R regions are identical, when assembling the reads mapping to these regions it leads to the formation of repetitions of the same TE (Figure 16).

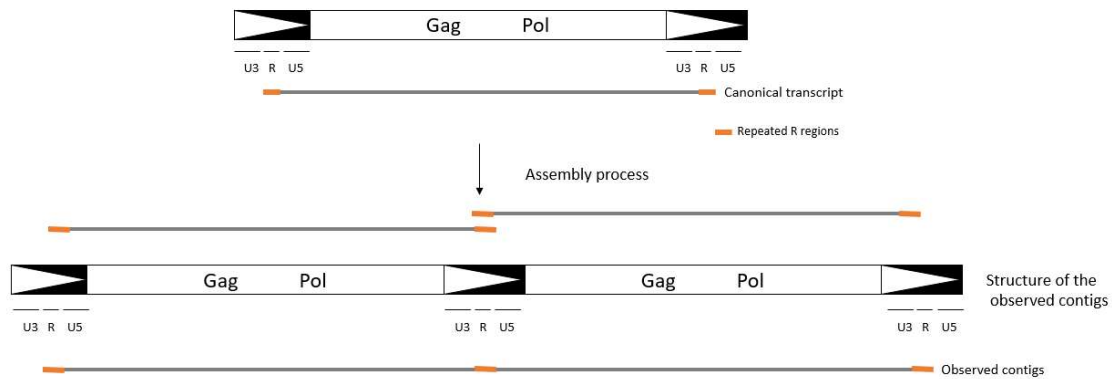


Figure 16: Assembly of Tandem repetitions of LTR-RT using the de novo assembly approach. Complete transcripts of the LTR-RT (from the R region of the first LTR to the R region of the second LTR) are formed, leading to the formation of tandem repetitions where both copies had the same nucleotide sequence.

To solve this problem, we manually trimmed these cases discarding one of the two tandem repetitions in our final selected contigs. The remaining 172 contigs were manually inspected discarding assemblies corresponding to poorly annotated TEs (i.e. repetitive genes like Leucine-Rich Repeat genes), solo LTR or chimeric TEs, ending up with 22 contigs.

As a last step, the RNA-seq short reads were mapped to the 22 selected contigs and we discarded those with only antisense mapped reads as they cannot correspond to TE transcription that can potentially lead to a transposition event. This reduced the number of contigs to 9, which corresponded to 9 different potentially active TEs. Five of them belonged to the RLG1, RLG2, RLC4 and RLC5 previously characterized LTR-RT families. Two LTR-RT contigs that were identified represented the complete RLC5 copies and the truncated RLC5 copies. These 5 clusters belonged to the previously defined clusters RLG1_grp1, RLG2_grp13, RLC4_grp1 and RLC5_grp2 (containing

both the truncated and the complete copies). The other 4 contigs corresponded to 2 LINES sequences and 2 DNA TEs (PpTc1 and PpTc2).

We aligned the RNAseq libraries from the protonemata treated with heat stress vs protonemata grown in control conditions to the 4 LTR-RT assembled contigs and estimated their expression (Figure 17 and Table 5).

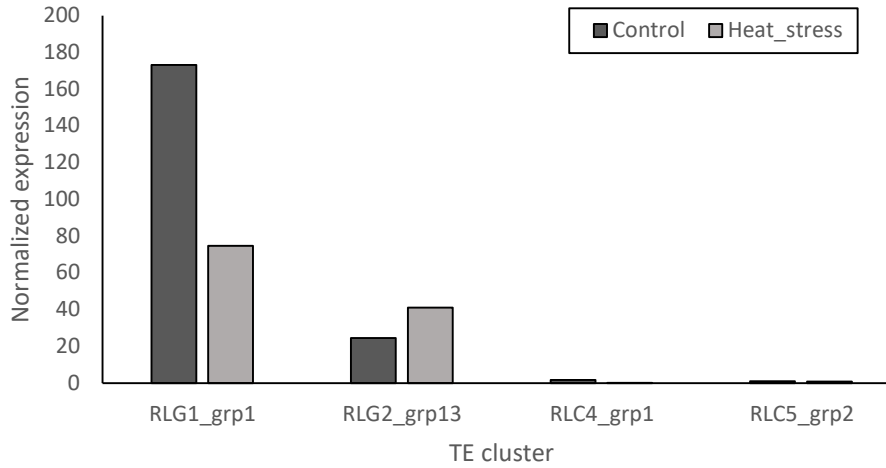


Figure 17: Expression detected for the different families using the assembly approach. Number of reads mapped to the different contigs divided by the total read number of the libraries (6 libraries). Observing a high expression for the RLG1_grp1 and RLG2_grp13 clusters while almost no expression for the RLC4_grp1 and RLC5_grp2 clusters.

Table 5: Differential expressed clusters detected between protonemata samples treated with heat shock and protonemata control samples expressed in Log₂ Fold Change (Log₂FC) values detected by the TE assembly approach. Positive Log₂FC correspond to samples that are overexpressed in heat stress, negative values samples that are repressed in the heat stress treatment. Differential expressed clusters were selected using a padjusted value of less than 0.05.

TE Cluster	log ₂ FC	padj
RLG1_grp1	-1,37	1,68E-02
RLG2_grp13	0,69	1,37E-01
RLC4_grp1	-2,31	1,25E-20
RLC5_grp2	-0,79	4,33E-02

Two clusters belonging to the RLG1 and RLG2 clusters were detected as expressed under both conditions with a repression of the expression for the RLG1_grp1 under heat stress

and an increase of the expression of the RLG2_grp13 cluster under heat stress. In both cases the mapped reads of the RNAseq were mapping all over the complete contigs.

The other two LTR-RT Copia clusters that were detected from the TE assembly approach had a low transcription level although in the case of RLC4_grp1, the reads were covering all the element. This element seems to be expressed under control condition, and we observed a decrease of expression under heat stress. In the case of the RLC5_grp2 cluster only a few reads were detected and did not cover the entire sequence.

Finally, we compared the results obtained using this approach with those of Tetranscripts and Tertools. The number of expressed clusters was reduced from 32 and 33 clusters (with Tetranscripts and Tertools respectively) to only 5 LTR-RT. Moreover, we could detect expression for other TE classes that were not detected by Tertools and Tetranscripts due to that they were not properly annotated in the reference TE annotation, such as the LINE TEs or the DNA TEs.

Discussion

The detection of the transcription that can lead to a transposition event is not straightforward. TEs are highly repetitive regions and the use of short-reads to detect the transcription over these regions can be particularly challenging as most of the reads will map to multiple positions in the genome. This multimapping reads hinders the capacity to distinguish between the TE transcription that can lead to a transposition event from the transcription of short TE fragments that are not related to the transposition. Therefore, it requires the design of strategies to be able to distinguish between these two events using short-reads data.

The detection of the TE transcription can also be highly influenced by the TE annotation of the genome as we observed in our analysis when comparing the use of the annotation of the published LTR-RT annotation (Lang et al., 2018) to our defined LTR-RT clusters.

As previously introduced, the TE fraction of the genome of *P. patens* has a high complexity. *P. patens* presents an heterochromatin distributed all along the chromosome, mainly composed of LTR-RTs of the Gypsy superfamily that are forming complex structures of nested TE insertions (Lang et al., 2018), probably a fraction of these regions are also misassembled. The presence of only a few LTR-RT families highly repetitive all around the genome hinders the approaches that requires the mapping of the reads to the genome. For this reason, we opted, to use TEtools, which does not require the mapping of the reads to the genome. TEtools allows the quantification of the TE transcription using previously defined TE clusters that may facilitate the quantification using only the detected complete elements of these families, allowing an easier visualization of the mapping of reads to these clusters.

Despite that, we observed that several clusters were detected as expressed, although they did not correspond to a complete TE transcription. As in our case we were interested mainly to find the transcription that could potentially lead to a new transposition event, this was not the best approach for our goal.

For this reason, we opted to develop a method that allows an easier identification of the transcripts that could potentially lead to a transposition event, based on the assembly of the short-reads that was done to detect TE expression in humans (Guffanti et al., 2018).

In this publication they describe a RNAseq *de novo* assembly approach using all the RNAseq data to identify TE copies that are active and chimeric transcripts between TEs and genes in human samples with the goal to identify the impact that the TEs can have on the gene regulatory networks. As we were not interested at that moment into the identification of TE-gene chimeric transcripts and as a *de novo* transcriptome assembly is a time consuming and needs high computational resources, we followed a different approach by only assembling the reads mapping to TEs. This allows a decrease of the CPU time as compared with the one required for a complete *de novo* assembly of a transcriptome.

This approach allowed us to decrease the number of identified clusters as expressed from the TETRANSCRIPTS and TETOOLS approach to only a few clusters. In the case of the LTR-RT to only 4 different clusters. These clusters belonged to all those elements that had all the TE coding sequence covered by reads and are probably more prone to be active copies. This method also allowed us an easier visualization of the mapping of the reads to these assembled transcripts to confirm their expression and under which conditions or stresses they are induced.

Despite that, the TE *de novo* transcriptome assembly approach has some limitations such as the formation of artificial LTR-RT repetitions of the same copy due to the presence of reads covering the R regions that leads to the formation of these contigs. There is also a need of a manual curation of the contigs to solve these repetitions that do not correspond to the real transcripts.

Since the development of the *de novo* TE transcriptome assembly approach to identify the expression of the TEs in *P. patens*, there has been further improvements by other members of the lab to analyze the expression of TEs in other species (Amelie Bardil and Carlos de Tomás in *Prunus persica* and *Prunus dulcis* and Raúl Castanera in *Arabidopsis thaliana* and *Oryza sativa*). One of the main limitations of the method was the need of a manual curation of the contigs formed by the assembly approach. To solve this problem, they aligned the assembled contigs to the annotation of TEs and selected for each contig the most similar complete TE copy to posteriorly map the reads to this selected TE copies. When this method was developed to normalize the expression, we were using a non-canonical normalization based on FPKM values. We were normalizing the expression by the total number of reads mapped to the assembly divided by the total length of the

assembled contig and the total raw reads of the library. This method has been replaced for the use of a standard method. They concatenated the selected expressed TE copies to the gene transcriptome. In this way it was possible to quantify and normalize the expression using FPKM or TMM values and analyze the gene expression and the TE expression in the same experiment, and still, being easy to visualize the alignments of the reads to the selected TE copies.

Overall, the work done in this chapter can be useful to identify the transcription of TEs that could putatively produce a new transposition event from short-read data. Probably these methods would be outperformed by cDNA or direct RNA long read sequencing that can produce complete transcripts in a single read. Most of the generated data up to the elaboration of this thesis is still based on short-reads and there are a few long-reads libraries publicly available, with none of them for *P. patens*, up to date. For this reason, this method can be useful to overcome the main limitations of the short-reads data to identify active TEs and allow the identification of TE transcription from publicly available short-reads dataset.

CHAPTER 2: TE_s DYNAMICS IN
PHYSCOMITRIUM PATENS

CHAPTER 2: TE_s DYNAMICS IN *PHYSCOMITRIUM PATENS*

Chapter 2.1: Introduction

As already mentioned, *P. patens* is a model organism widely used to answer several questions of life science and to understand the evolution of biological processes of land plants.

Interestingly, with the publication of the most recent version of the genome during the year 2018 (Lang et al., 2018) it was observed that the genome of *P. patens* has some particularities when compared with the available genomes of vascular plants. Although, the TE coverage is quite similar to other plant genomes of similar size, such as the one of *Cucumis melo* (Castanera et al., 2020) or of *Oryza sativa* (Kawahara et al., 2013), the distribution of TEs is different of what is typically observed in vascular plant genomes.

As introduced, the main component of the heterochromatic regions dispersed along the chromosomes of *P. patens* are LTR-RT Gypsy elements, which represent a 48% of the genome (Lang et al., 2018). The most prevalent Gypsy LTR-RT family is the RLG1 LTR-RT, which constitutes the 25% of the genome and the 51% of all the *P. patens* TEs. The other two LTR-RT Gypsy families described, RLG2 and RLG3, represent a 23% of all the genome, from which RLG2 occupies a 5.6% and RLG3 a 9.26% of the genome. There is a 7.2% of the genome that is occupied by unclassified Gypsy elements that are also located mostly in these heterochromatic regions. Copia LTR-RTs only account for a small fraction of the genome (3.5%). Interestingly, one of the two families of Copia elements, RLC5, is mostly found accumulated in a single position per chromosome that could coincide with the centromere and it has been hypothesized that these elements may have an essential structural role for the maintenance of the genome (Lang et al., 2018). RLC5 elements are present both as complete copies (with a complete ORF coding for all the proteins necessary for the mobilization of the TEs) and as putatively non-autonomous copies, called tRLC5 (from truncated RLC5), that have identical LTRs but lack part of the polyprotein sequence. These elements, share mostly the same structure and are potentially mobilized by the autonomous RLC5 elements (Lang et al., 2018).

With respect to other TE types, only one family of Helitron was identified in the first draft genome published in 2008 (Rensing et al., 2008). In the following years it was described 4 families of MITEs, comprising 3718 elements, and occupying less than a 0.12% of the genome (Chen et al., 2014). In addition, two families of DNA TEs were detected from RNAseq data and were named as PpTc1 and PpTc2 (Liu & Yang, 2014).

P. patens TEs accumulate histone marks related to gene silencing and typically marking heterochromatic regions, such as the histone mark H3K9me2 (Lang et al., 2018; Widiez et al., 2014). Moreover, as expected from what has been observed in other plant genomes, TEs are enriched in methylation in the three methylation contexts (CG, CHG and CHH) compared to genic regions (Lang et al., 2018). Moreover, recent results indicate that the absence of the symmetric CG context and in the asymmetric (CHG and CHH) contexts on the moss result in an increase of the expression of certain TE families (Domb et al., 2020). Small RNAs (sRNAs) targeting the RLG1 and RLG3 LTR-RTs have been detected (Coruh et al., 2015), these sRNAs could have a role in the silencing processes of these families. All this data suggests that the main mechanisms of TE silencing are conserved between *P. patens* and vascular plants.

Previous results obtained in the lab indicated that RLG1 elements are transcriptionally active in protonemata (Vives et al., 2016). Moreover, the work of Cristina Vives and Jordi Morata, previous members of the lab, also showed low expression of the RLC5 family in protonemata tissue (Lang et al., 2018). They also analyzed the presence of polymorphic TE insertions between Gransden, the accession from which the reference genome was assembled, and the Villersexel accession. Observing that most of the TIPs belonged to the RLG1 family, although some polymorphic RLG3 and RLC5 insertions were also detected.

The publication of a gene atlas with expression data from several development conditions and stresses (Perroud et al., 2018) and the availability of two new genome resequencing data corresponding to the *P. patens* accessions Reute and Kaskaskia opened the doors to study the expression of TEs in several development conditions and stresses and look for their potential mobilization during *P. patens* evolution. In this context, the development of the methods explained in the previous chapter to improve our capacity to detect TE expression and to study TIPs was essential. In this chapter we will focus on the use of

these approaches to study the dynamics of *P. patens* TEs using the gene atlas database and the different resequencing data to study the transcriptional and transpositional landscape of *P. patens*.

As a result, we detected different families of Retrotransposons and DNA TEs that are transcriptionally active in different development conditions and stresses and are polymorphic among different *P. patens* accessions. This work constitutes the first part of this chapter and was published in 2020 in *Frontiers in plant science* (Vendrell-Mir et al., 2020). It was a collaborative work between our group and that of Dr. Fabien Nogu e from IJPB-INRA Versailles with the help of Dr. Mauricio Lopez-Obando that provided us with sporophyte RNA samples. From this study, I performed all the experimental analysis except for the sporophyte RNA extractions and I also performed the bioinformatic analysis.

Hereunder, as part of the second chapter a copy of the published article is included. All the supplementary material cited in the article can be access through the following DOI: [10.3389/fpls.2020.01274](https://doi.org/10.3389/fpls.2020.01274)

Chapter 2.2: Objectives

- Study whether some TE families are transcriptionally active and are polymorphic in the population of *Physcomitrium patens*.

Chapter 2.3: Expression and mobilization of TEs in *Physcomitrium patens*



Different Families of Retrotransposons and DNA Transposons Are Actively Transcribed and May Have Transposed Recently in *Physcomitrium (Physcomitrella) patens*

OPEN ACCESS

Edited by:

Meenu Kapoor,
Guru Gobind Singh Indraprastha
University, India

Reviewed by:

Matej Lexa,
Masaryk University, Czechia
Rita Sharma,
Jawaharlal Nehru University, India

*Correspondence:

Fabien Nogué
fabien.nogue@inrae.fr
Josep M. Casacuberta
josep.casacuberta@cragenomica.es

Specialty section:

This article was submitted to
Plant Systematics and Evolution,
a section of the journal
Frontiers in Plant Science

Received: 21 February 2020

Accepted: 05 August 2020

Published: 19 August 2020

Citation:

Vendrell-Mir P, López-Obando M,
Nogué F and Casacuberta JM (2020)
Different Families of Retrotransposons
and DNA Transposons Are
Actively Transcribed and
May Have Transposed
Recently in *Physcomitrium*
(*Physcomitrella*) *patens*.
Front. Plant Sci. 11:1274.
doi: 10.3389/fpls.2020.01274

Pol Vendrell-Mir¹, Mauricio López-Obando², Fabien Nogué^{3*}
and Josep M. Casacuberta^{1*}

¹ Centre for Research in Agricultural Genomics CSIC-IRTA-UAB-UB, Campus UAB, Edifici CRAG, Barcelona, Spain,

² Department of Plant Biology, Swedish University of Agricultural Sciences, The Linnean Centre of Plant Biology in Uppsala, Uppsala, Sweden, ³ Institut Jean-Pierre Bourgin, INRAE, AgroParisTech, Université Paris-Saclay, Versailles, France

Similarly to other plant genomes of similar size, more than half of the genome of *P. patens* is covered by Transposable Elements (TEs). However, the composition and distribution of *P. patens* TEs is quite peculiar, with Long Terminal Repeat (LTR)-retrotransposons, which form patches of TE-rich regions interleaved with gene-rich regions, accounting for the vast majority of the TE space. We have already shown that RLG1, the most abundant TE in *P. patens*, is expressed in non-stressed protonema tissue. Here we present a non-targeted analysis of the TE expression based on RNA-Seq data and confirmed by qRT-PCR analyses that shows that, at least four LTR-RTs (RLG1, RLG2, RLC4 and tRLC5) and one DNA transposon (*PpTc2*) are expressed in *P. patens*. These TEs are expressed during development or under stresses that *P. patens* frequently faces, such as dehydration/rehydration stresses, suggesting that TEs have ample possibilities to transpose during *P. patens* life cycle. Indeed, an analysis of the TE polymorphisms among four different *P. patens* accessions shows that different TE families have recently transposed in this species and have generated genetic variability that may have phenotypic consequences, as a fraction of the TE polymorphisms are within or close to genes. Among the transcribed and mobile TEs, tRLC5 is particularly interesting as it concentrates in a single position per chromosome that could coincide with the centromere, and its expression is specifically induced in young sporophyte, where meiosis takes place.

Keywords: *Physcomitrium (Physcomitrella) patens*, transposable element, transcription, genetic variability, centromere

INTRODUCTION

Mosses are one of the oldest groups of land plants, forming a sister clade with vascular plants (Leebens-Mack et al., 2019). Since the demonstration, in 1997, that gene targeting *via* homologous recombination was possible in *Physcomitrium* (*Physcomitrella*) *patens* (Schaefer and Zrýd, 2001) this moss has become a leading plant model for answering essential questions in life sciences and in particular for understanding the evolution of biological processes of land plants. The draft of the *P. patens* genome was published in 2008 (Rensing et al., 2008), and a chromosome-scale assembly of the *P. patens* genome has been published (Lang et al., 2018), highlighting the similarities and differences with other plant genomes. Transposable Elements (TEs) account for the 57% of the 462.3 Mb of the assembled *P. patens* genome. This TE coverage is not very different from that of other plant genomes of similar size (Tenaillon et al., 2010). On the contrary, the distribution of TEs in *P. patens* is unusual as compared to other plants. TE-rich regions alternate with gene-rich regions all along the *P. patens* chromosomes (Lang et al., 2018) whereas in most plant genomes TEs accumulate in pericentromeric heterochromatic region on each chromosome. Interestingly, in spite of the general patchy TE distribution, a family of retrotransposons of the *copia* superfamily, RLC5 (comprised of full length, from now on RLC5, and truncated, tRLC5, elements), clusters at a single location in each chromosome that could correspond to the centromere (Lang et al., 2018). The TE-rich regions distributed all along the chromosomes are mainly composed of a single family of LTR-retrotransposons of the *gypsy* superfamily named RLG1 (Lang et al., 2018). RLG1 integrase contains a chromodomain, a type of protein domain that has been previously found to direct retrotransposon integration into heterochromatin (Gao et al., 2008), suggesting that RLG1 could target heterochromatic TE islands for integration. Although most TE copies are located in heterochromatic TE islands, gene-rich regions also contain some TE copies, with some of them that inserted recently and are polymorphic between the Gransden and Villersexel accessions (Lang et al., 2018). Moreover, the RLG1 retrotransposon is transcribed in *P. patens* protonema cells, suggesting that it can transpose during *P. patens* development (Vives et al., 2016; Lang et al., 2018). Although these data suggest that TE activity may have shaped the genome of *P. patens* and may continue to generate variability that potentially impact *P. patens* evolution, the global analysis of the capacity of *P. patens* TEs to be expressed and transpose is still lacking. Here we present an unbiased analysis of TE expression in *P. patens* based on RNA-Seq analyses and confirmed by qRT-PCR, that has allowed uncovering the developmentally or stress-related expression of different TE families, including class I (retrotransposons) and class II (DNA transposons) TEs. The data presented here reinforce the idea that TEs have shaped the genome of *P. patens* and show that they continue to drive its evolution.

MATERIALS AND METHODS

RNA-Seq Data Used

RNA-Seq data were obtained from the *P. patens* Gene Atlas library (Perroud et al., 2018). In particular, we used RNA-Seq data obtained from stress-treated tissues (protoplasts, ammonium treatment, de- and rehydration, heat stress, and UV-B), different developmental stages, including protonemata in BCD, BCDA or in Knopp medium, protonemata in liquid and solid medium, gametophores, leaflets, and sporophytes (green and brown stages) and some hormonal treatments (Auxin, ABA or the Jasmonic acid precursor OPDA). A complete list of the data set used can be found in **Supplementary Table 1**.

Transposable Element Transcriptome Assembly and Quantification

All selected reads were trimmed by quality using BBduk (<https://sourceforge.net/projects/bbmap/>). Reads mapping to the chloroplast, mitochondria or rRNA were discarded from the analysis. The remaining reads were mapped to the transposable element annotation (Hiss et al., 2017) using Bowtie2 (Langmead, 2013). All the reads that mapped were extracted using Samtools (Li et al., 2009). These reads were assembled to contigs using Trinity (Grabherr et al., 2011). In order to characterize and filter the assemblies, we aligned them to the TE library described in (Lang et al., 2018) using BLASTn (Altschul et al., 1990) with an e-value cutoff of 10^{-5} . For transcripts corresponding to class I TEs, we kept only those showing alignments longer than 1000 nt. Manual inspection allowed discarding assemblies corresponding to poorly annotated TEs (i.e. repetitive genes like Leucine-Rich Repeat genes), solo LTR or chimeric TEs. The potentially coding domains of the selected assemblies were identified by a CDD-search, which allowed defining the orientation of the potentially expressed TEs (Marchler-Bauer et al., 2015).

In order to estimate the levels of expression of the elements corresponding to the selected assemblies, RNA-Seq reads were mapped to the selected assemblies using bowtie2 and only the reads potentially corresponding to sense transcripts were kept. To quantify the expression the number of mapping reads was normalized by the length of the assembly (Kb) and the total amount of trimmed reads for each condition without aligning the reads to the genome. The normalized expression data of each transcript and the sequence of the selected transcripts can be found in **Supplementary Table 1**.

Plant Material

P. patens Gransden accession was used for all the samples used, with exception of the protonema vs sporophytes induction test where the *P. patens* Reute accession (Hiss et al., 2017) was used.

Protonemata were fragmented and plated on BCDAT medium overlaid with a cellophane disk in long-day conditions (16 h light 15 W m⁻² to 8 h darkness) at 24°C for 7 days. Samples were collected at day 7 after 4 h of light. All the samples were frozen in liquid nitrogen immediately after harvesting and were kept at -80°C.

Protoplasts were isolated from 6 days old protonemal cultures after 30 min incubation in 1% driselase (Sigma D8037), 0.48 M mannitol. The suspension was filtered through two superposed sieves of 80 and 40 μm . Protoplasts were sedimented by low-speed centrifugation (600g for 5 min) and washed in 0.48 M mannitol.

The ABA treatment was performed as previously described (Perroud et al., 2018). Briefly, protonemal cultures were grown for 6 days on a cellophane disk on BCD medium. At day 6, the cellophane disks containing the protonemata tissues were transferred to BCD medium as control or to BCD containing 50 μM abscisic acid (Sigma A1049) for 24 h before harvesting.

Sporophyte RNA was obtained from Reute *P. patens*. Seven days old regenerated tissue from two consecutive rounds of a week old grinded material grown on solid BCDAT medium covered with cellophane was used as starting material. Six similar size small dots of moss tissue were plated in a 25 mm height petri dish (WVR international) containing BCD solid medium. They were grown for 40 days at 30 $\mu\text{mol m}^{-2} \text{s}^{-1}$ constant white light regime and 25°C in a Sanyo MLR chamber. Then, plants were transferred to a Sanyo MLR chamber at an 8-h to 16-h light-dark cycle, 30 $\mu\text{mol m}^{-2} \text{s}^{-1}$ light intensity and 15°C for reproductive gametangia induction and sporophyte development. After 20 days of post-reproductive induction (dpri), plants were submerged overnight in water to increase fertilization. Sporophyte samples were collected at 45 dpri showing a green round shape developmental stage. Each sporophyte was dissected under a Leica MZ16 stereomicroscope. Gametophyte tissue was discarded as much as possible and sporophyte was quickly frozen in liquid nitrogen. 40 dissected sporophytes were collected and used for RNA extraction.

RNA Extraction and cDNA Production

Sporophyte RNA was obtained using the QIAGEN RNeasy mini kit following manufacturer's protocol. DNA was removed by treating the samples with Ambion™ DNaseI kit (AM2222) following the manufacturer's protocol. For all other tissues, RNA extraction and DNase treatment was done using the Maxwell® RSC Plant Kit (Promega). 500 ng of total RNA was used to synthesize the first-strand cDNA using the SuperScript™ III reverse transcriptase (ThermoFisher).

qRT-PCR

Quantitative real-time PCR were done in 96-well plates using the Roche LightCycler II instrument. SYBER Green I Master Mix (Roche Applied Science), primers at 1 μM and 1/20 dilution of the cDNA obtained from the reverse transcription were used for the qRT PCR. Each sample was run per triplicate with negative reverse transcriptase and non-template controls. The amplification conditions were: 95°C for 5 min, followed by 95°C for 10 s, 56°C for 10 s, and 72°C for 10 s, ending with the melting curve to check the specificity of the qRT-PCR. The housekeeping gene adenine phosphoribosyl transferase (APT) (Schaefer et al., 2010) was used to normalize the qRT-PCR results.

The primers used to check TE expression were designed using the Primer3plus software (Untergasser et al., 2012). The list of

the primers used in this study can be found in **Supplementary Table S2**.

Detection of Potentially Expressed TE Copies in the Genome and LTR-Retrotransposon Age Estimation

The TE copies most similar to the RNA assemblies, potentially representing the expressed elements, were identified by aligning the assemblies to the genome using Blastn with an e-value cutoff of 10E^{-90} . However, in many cases the RNA assembly is obtained from the assembly of reads potentially generated by the expression of similar but different copies, and therefore, this approach may not be suitable. In order to identify the subset of elements potentially expressed in those cases, we also searched for elements showing a similarity of 80% over 80% of the sequence of the assembly. In those cases, we estimated the age of the subset of elements most similar to the assembled transcript and compared it to the age of all the complete elements of the same family annotated in the genome. To do that, we estimated first the Kimura two-parameter distance (Kimura, 1980) between the two Long Terminal Repeats (LTRs) and estimated the age using the formula $T = K/2 \times r$, where T = time of divergence, K = divergence and r = substitution rate (Bowen and McDonald, 2001). Taking into account an estimated substitution rate of $9\text{E}-09$ (Rensing et al., 2007).

Transposable Element Polymorphisms Annotation

The publicly available DNA-seq resequencing data of three accessions of *P. patens* (Kaskaskia, SRX2234698; Reute, SRX1528135 and Villersexel, SRX030894) was used to look for TE polymorphisms with respect to the Gransden reference genome. Paired-end reads were mapped to the reference genome using BWA SW (Li and Durbin, 2009). TE insertions were detected using PoPoolationTE2 (Kofler et al., 2016) using the separate mode. To perform the analysis we kept only the non-reference insertions (insertions absent from the Gransden reference genome) predicted with a zygosity of at least 0.7. To establish the distance of these insertions to the closer genes the polymorphic TEs positions were intersected with that of the annotated genes using bedtools (Quinlan and Hall, 2010) using the function closestBed.

Phylogenetic Analyses

To look for sequences similar to *P. patens* TEs in other genomes we first performed a blastn search against the complete NCBI nucleotide database. As this only retrieve sequences with significant similarity to RLG1 element we complemented this search with a blastx search of the *P. patens* TEs first against the complete NCBI non-redundant protein sequence database excluding *P. patens* and subsequently, in order to increase the chance to detect plant sequences, to the NCBI green plant database (taxid:33090). We performed the tblastx with the default parameters with a maximum target sequence of 250. The most similar sequence for each species was chosen as

representative of the species. All the protein sequences were aligned using Mafft (Katoh and Standley, 2013) and trimmed using TrimAl (Capella-Gutiérrez et al., 2009). A phylogenetic tree was constructed using FastTree (Price et al., 2010) and visualized in iTOL (Letunic and Bork, 2019).

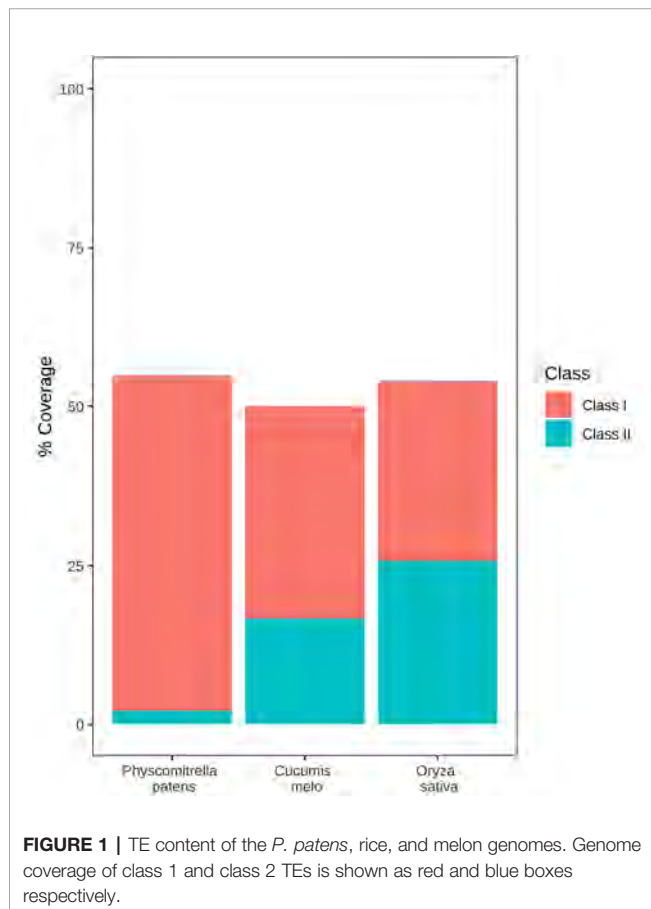
RESULTS

A New Approach to Measure the Expression of *P. patens* Transposable Elements

More than half of the *P. patens* genome (57%) is occupied by TEs, a figure that is similar to that of other genomes of similar size (Tenaillon et al., 2010). As an example, the *P. patens* TE content is similar to that of two other genomes of similar sizes and for which the TE content has been annotated using the same REPET package (Flutre et al., 2011), such as rice (46.6%) (Ou et al., 2019) and melon (45.2%) (Castanera et al., 2020). However, *P. patens* has a very different TE composition as compared with these two genomes. Indeed, class II TEs account for 21.06% of the rice genome and 15.42% of the melon genome, in *P. patens* they only represent 6% of the genome (**Figure 1**). More strikingly, a single retrotransposon family, RLG1 accounts for

almost half (47.44%) of the genome space occupied by class I elements (Lang et al., 2018). RLG1 is actively expressed in non-stressed protonema cells, and it may have transposed recently during *P. patens* evolution, as some of its copies are polymorphic between *P. patens* Gransden and Villersexel ecotypes (Vives et al., 2016; Lang et al., 2018).

RLG1 copies are concentrated in TE-rich heterochromatic islands and RLG1 transposition has therefore a limited capacity to induce gene variability. In order to explore the possibility that other TE families, apart from RLG1, could be expressed in particular developmental stages or stress situations, and could therefore generate new variability in gene regions, we took advantage of the large collection of *P. patens* RNA-Seq data available from the recently published *P. patens* gene atlas (Perroud et al., 2018), which includes data from different developmental stages and stress conditions. In addition of complete TEs, eukaryote genomes, and in particular those of plants, usually contain large amounts of defective and truncated elements that may be included in transcripts that are not the result of a genuine TE expression (Anderson et al., 2019). These transcripts can be sense or antisense with respect to the TE orientation and may in some cases participate in TE regulation, but cannot be considered as productive TE transcripts potentially involved in transposition. In *P. patens*, as it is common in eukaryote genomes and in particular in plants (Hoen et al., 2015; Bennetzen and Park, 2018), the fragmented and degenerated copies of TEs outnumber the complete and potentially functional copies. As a consequence, a quantification of the level of expression based on the number of RNA-Seq reads mapping to all TE-related sequences can lead to an overestimation of the expression of the different TE families. We have therefore decided to follow a strategy based on the detection of potentially complete transcripts obtained from an assembly of RNA-Seq reads, similar to what has previously been described for the analysis of the expression of human TEs (Guffanti et al., 2018). We used Trinity RNA-seq *de novo* assembly (Grabherr et al., 2011) to assemble reads showing similarity to annotated TEs (Lang et al., 2018). The 696 assemblies obtained were blasted back to the TE annotation to classify them. The vast majority (94%) of these 696 assemblies showed similarity to LTR-RT annotations, and an important fraction of them (72%) were short (less than 1000 nt) and corresponded to fragments of LTR-RTs, such as the LTRs. As an example, the assembly TRINITY_DN331_c0_g1 showed high sequence similarity to the LTR of RLC5 elements. A search for the genomic sequence most similar to that of the assembly identified a RLC5 solo-LTR located in the downstream proximal region of the Pp3c4_32070 gene annotation (**Supplementary Figure 1**). Interestingly, an analysis of the expression data available from the *P. patens* gene atlas (Perroud et al., 2018) showed that both the RLC5 solo-LTR and the Pp3c4_32070 annotated are specifically induced in gametophores treated with ABA, which strongly suggests that this solo-LTR is expressed as a consequence of read-through transcription from the gene promoter. In order to eliminate assemblies corresponding to the expression of fragments of LTR-RTs, and taking into account that



typical complete LTR-RTs are several kb long, we discarded all the LTR-RT assemblies shorter than 1,000 nt. The remaining 172 transcripts were analyzed for the potential presence of regions coding for the typical class I and class II TE protein domains and their alignments to annotated TE sequences were manually inspected to discard those showing similarities to poorly annotated transposable elements, and truncated or chimeric elements. As an example, **Supplementary Figure 2** shows the analysis of TRINITY_DN99_c0_g1_i5 that corresponds to a complex region containing different degenerated TE fragments that seem to be transcribed as a single transcription unit. Among the 22 assemblies retained, some corresponded to the antisense strand of annotated TEs. After manual inspection, some of these were shown to correspond to LINE elements (see **Supplementary Figure 3** for an example). These transcripts may participate in the control (e.g. silencing) of TE expression but cannot be considered as genuine TE transcription. The assemblies corresponding to potential antisense transcripts were discarded. An analysis of the remaining assemblies showed that they corresponded to 9 different potentially complete annotated TEs and were selected for further analysis.

Both Retrotransposons and DNA Transposons Are Expressed in *P. patens*

The analysis of the transcript assemblies showed that they correspond to 9 different *P. patens* TEs: 2 LTR retrotransposons of the *gypsy* superfamily (RLG1 and RLG2), two of the *copla* superfamily (RLC4 and RLC5), with one of them potentially corresponding to the two different forms of RLC5, the full-length and the truncated form (RLC5/tRLC5) and two different DNA TEs belonging to the Mariner superfamily, that were not properly annotated in the *P. patens* TE annotation (Lang et al., 2018), but had been previously identified as *PpTc1* and *PpTc2* (Liu and Yang, 2014). In addition, the manual inspection of the alignments of the transcript assemblies with the annotated TEs allowed refining the annotation of two elements annotated as unclassified non-LTR retrotransposon that we could identify as a potentially expressed complete LINEs (LINE-1 and LINE-2). The RNA-Seq reads obtained from the RNAs generated by the expression of a TE family show a certain degree of sequence variability, and therefore, they are not all of them identical to the assembly that represents the complete RNA of the family. On the other hand, this assembly is in most cases not identical to any of the of the TE copies of that particular TE family. This suggests that, for most TE families, different elements are concomitantly expressed and that the RNA assembly should be considered as a consensus of the expressed RNAs.

These results suggest that different families of both retrotransposons and DNA transposons are transcribed in *P. patens*.

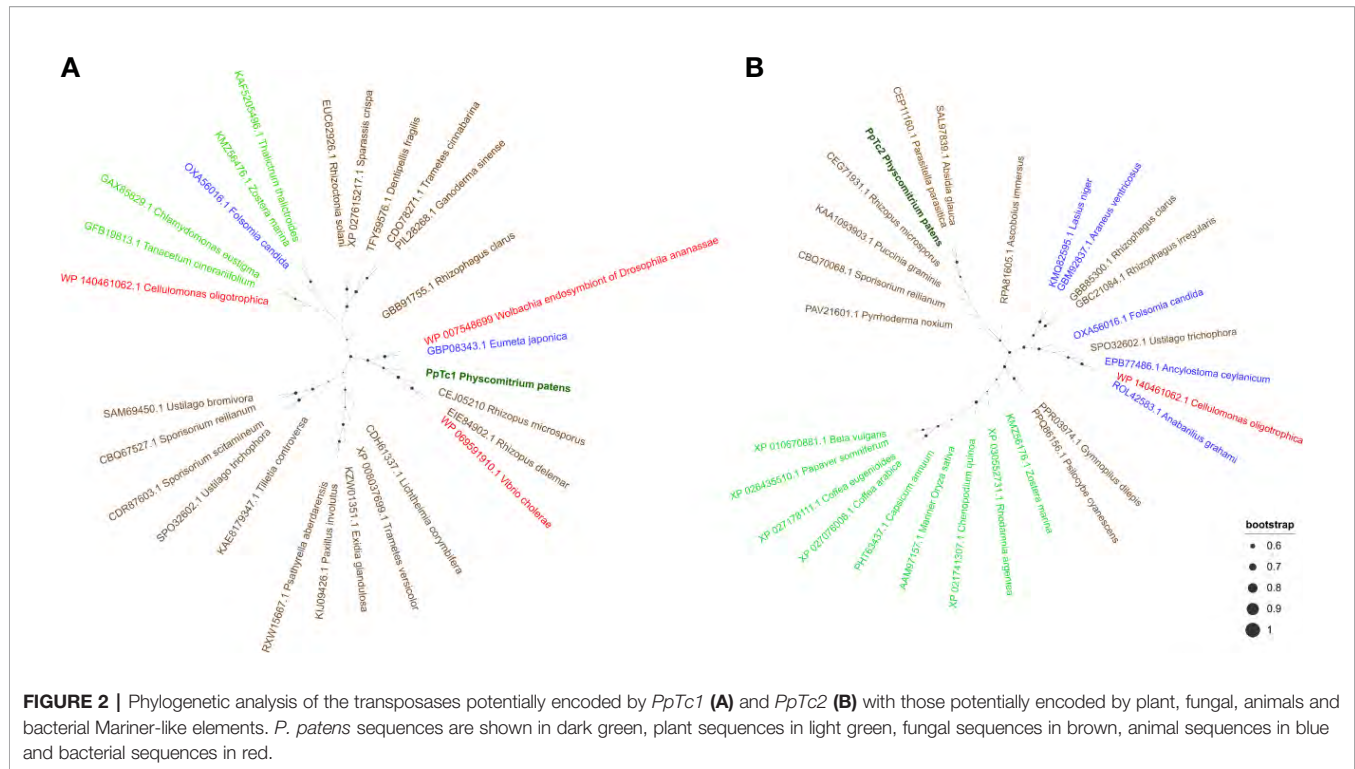
P. patens Contains TEs Closely Related to Fungal TEs

A preliminary characterization of the two Mariner-like elements found to be potentially expressed suggested that these elements were different from other plant Mariner-like elements, they being more closely related to fungal TEs of the

Mariner superfamily. As this result was somehow surprising, we searched for sequences potentially corresponding to transposases of similar elements in the phylogenetically related liverwort *Marchantia polymorpha* and in well-characterized dicotyledonous and monocotyledonous plants such as *Arabidopsis* and rice. These searches did not retrieve significant hits, suggesting that these genomes do not contain sequences related to Mariner-like elements similar to those found in *P. patens*. A phylogenetic analysis of the potential transposases of Mariner-like sequences present in public databases more similar to those of the two Mariner-like elements found in *P. patens*, and including other Mariner-like sequences from plants, shows that the *P. patens* elements are closely related to elements found in fungal genomes, and are not related to *Marchantia polymorpha* or other plant sequences (**Figure 2**). These results may indicate a horizontal transfer of these TEs from fungi. In order to explore whether other TEs may have also experienced a similar phenomenon, we extended the phylogenetic analysis performed for the two Mariner-like elements to the other *P. patens* TE families here described. These analyses showed that, in contrast to what happens for the two Mariner-like elements, databases contain plant sequences with significant similarity to the rest of TE families here described. However, the phylogenetic analyses performed show that whereas the trees obtained for *P. patens* RLG2, RLC4, LINE-1 and LINE-2 retrotransposons are congruent with the phylogenetic relationships of the species, this is less obvious for RLG1 and tRLC5 (**Supplementary Figures 4–8**). This may suggest that, in addition to the two Mariner-like elements, other *P. patens* TEs may have been transferred horizontally from fungal species.

Developmental and Stress-Related Expression of *P. patens* TEs

The availability of RNA-Seq data from different developmental stages and stress conditions (Perroud et al., 2018) allowed us to perform an unbiased analysis of the patterns of expression of the different transcribed *P. patens* TEs. We have previously shown that RLG1 is expressed in non-stressed protonema cells and its expression is reduced in protonema-derived protoplasts. RLG1 seems thus to be repressed by stress, in clear contrast with the stress-related expression of most TEs, as already discussed (Vives et al., 2016; Lang et al., 2018). Here we confirm that RLG1 is expressed in protonema, its expression increasing as the protonema develops and decreasing when gametophores develop, and is repressed in protoplasts (**Figures 3 and 4**). On the other hand, RLG1 does not seem to be expressed in other tissues and it is repressed by several of the stresses analyzed, in particular by heat shock and UV-B light (**Figures 3 and 4**). We confirmed the RLG1 expression in protonema cells and its repression in protonema-derived protoplasts by qRT-PCR (**Supplementary Figure 9**). A comparison of the RLG1 assembled RNA with all the RLG1 genomic copies suggests that only a subset of the RLG1 elements is expressed (**Table 1**). An analysis of the putative ages of these elements, by analyzing the sequence differences between the two LTRs of each element,



suggests that only the youngest RLG1 elements are transcribed (Figure 5A).

RLG1 is the TE expressed at the highest level in *P. patens* but, as already mentioned, we show here that other TEs are also expressed during *P. patens* development or under particular environmental conditions. The second Gypsy-like LTR-RT family found to be expressed, RLG2, is also expressed in protonema cells, and its expression increases in gametophores (Figure 3). On the other hand, the expression of RLG2 is strongly induced by ABA and heat stress in protonema, and repressed when gametophores are submitted to dehydration and rehydration (Figure 4). We confirmed the induction of RLG2 expression by ABA by qRT-PCR analyses (Supplementary Figure 10). Similarly, to RLG1, the comparison of the RLG2 assembled RNA with the RLG2 genomic copies shows that only the youngest RLG2 elements are transcribed in the conditions tested (Table 1 and Figure 5B).

The two *copia* retrotransposon families found here to be expressed, show low levels of expression during *P. patens* development. RLC4 seems to be particularly expressed in gametophores, whereas tRLC5 seems to be more expressed in sporophytes. RLC4 expression seems to be repressed in most stress conditions, although the levels of expression are very low in all cases.

tRLC5 is a particularly interesting family of TEs, as tRLC5 copies have been proposed to mark the centromere and participate in the centromeric function (Lang et al., 2018). The data presented suggest that tRLC5 may be particularly expressed in green sporophytes (Figure 3). In order to confirm this pattern of expression we performed qRT-PCR experiments. As the Gransden ecotype produces few sporophytes, which makes it

difficult to analyze sporophyte-specific expression, we used Reute tissues, as this ecotype produces many more sporophytes in laboratory conditions (Hiss et al., 2017). This analysis confirmed that tRLC5 expression is induced in young sporophytes (Supplementary Figure 11). A comparison of the tRLC5 assembled RNA with the tRLC5 genomic copies suggests that only the youngest tRLC5 elements are transcribed (Table 1 and Figure 5C).

LINE-1 seems to be expressed at a very low level in all conditions and we have not detected any relevant change in expression upon stress (not shown). On the other hand, LINE-2 is also expressed at a low level in most tissues but shows an increased expression in sporophytes and germinating spores (Figure 3). A comparison of the LINE-2 assembled RNA with the genomic copies suggests that the expressed LINE-2 is located in the close vicinity of an annotated gene (Pp3c16_3270) and the mapping of the RNA seq reads to this region suggests that LINE-2 could be expressed as the result of a readthrough transcription of this gene (Supplementary Figure 12). Indeed, although there are some minor differences, the patterns of expression of Pp3c16_3270 and LINE-2 during development or under particular stress conditions are mostly coincident (not shown).

Finally, of the two Mariner-like elements analyzed, only *PpTc1* is expressed in non-stressed tissues, with a particularly high expression in gametophores and leaflets (Figure 4), but both *PpTc1* and *PpTc2* are strongly induced by stress. *PpTc1* expression is particularly induced by heat stress, whereas *PpTc2* is only expressed after ABA induction or after dehydration or rehydration of gametophores (Figure 4). A comparison of the two Mariner-like assembled RNAs with their genomic copies identified the two elements potentially transcribed. Both

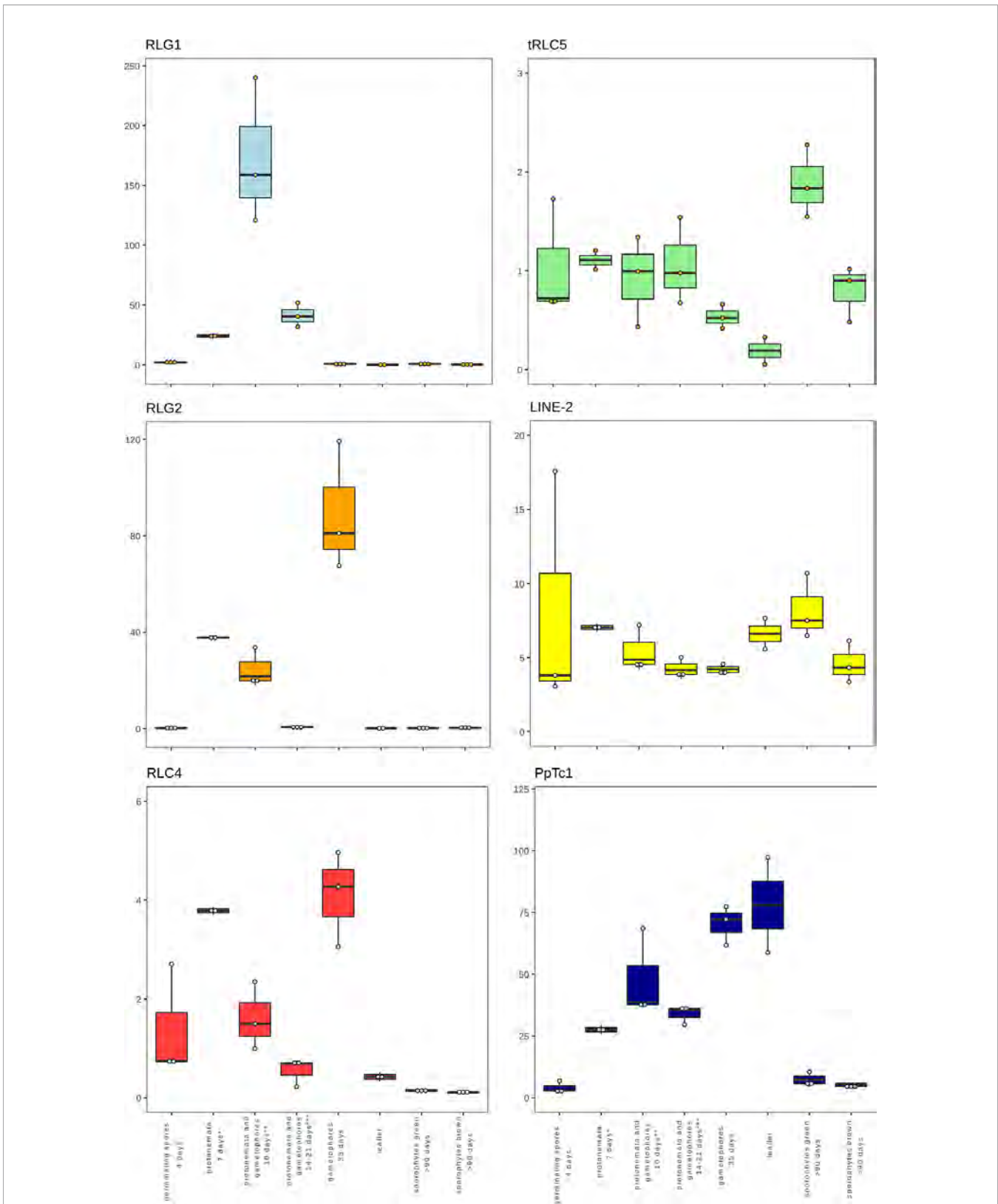


FIGURE 3 | Developmental expression of *P. patens* TEs. Normalized TE expression (see methods) in different developmental conditions selected from the *P. patens* Gene Atlas library (Perroud et al., 2018).

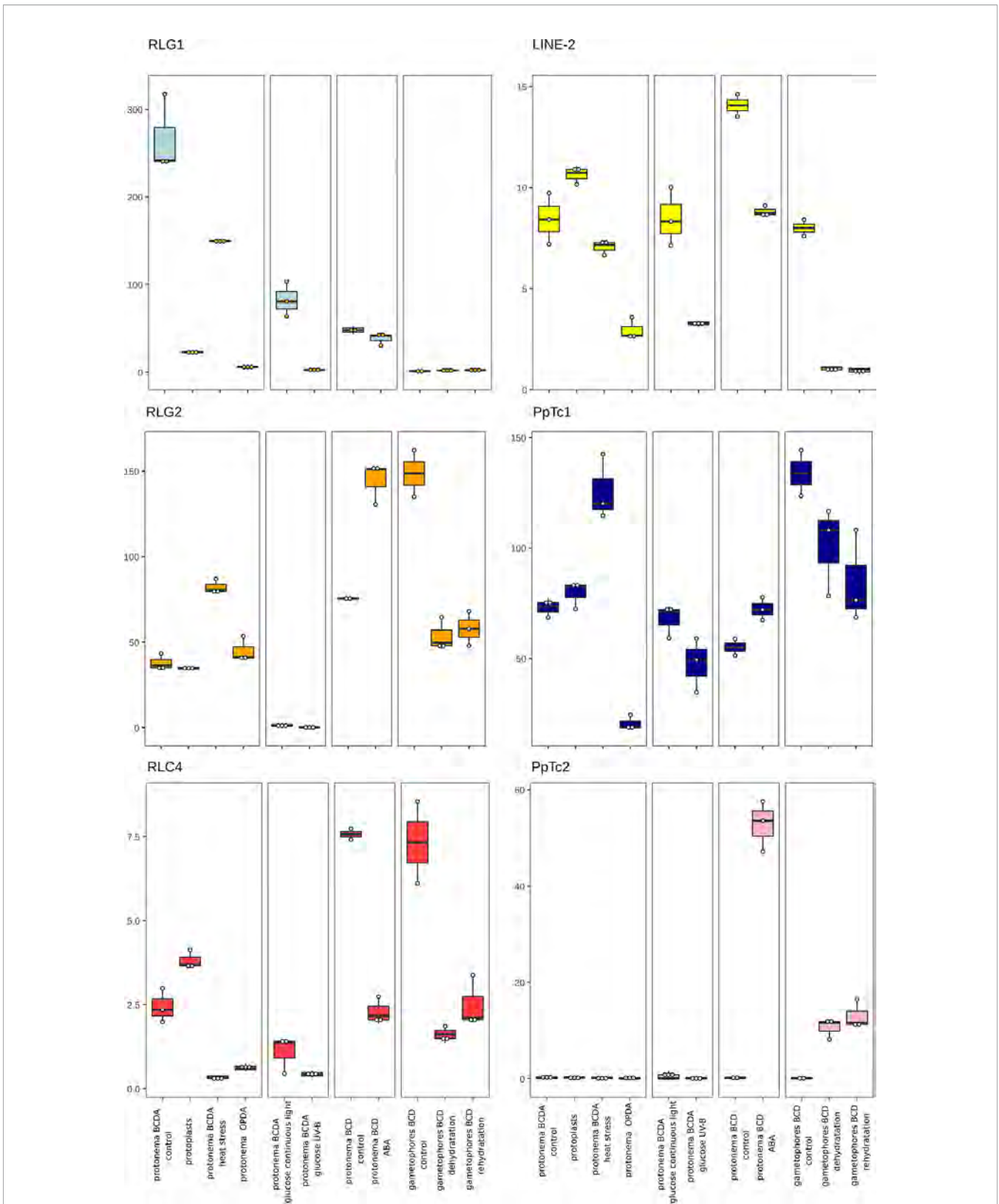


FIGURE 4 | *P. patens* TE expression under stress conditions. Normalized TE expression (see methods) under different stress conditions selected from the *P. patens* Gene Atlas library (Perroud et al., 2018).

TABLE 1 | Total number of complete elements (total), number of elements showing 80% identity over 80% of the length of the corresponding RNA assembly (80/80) for each indicated TE family.

	Total	80/80
RLG1	5092	3636
RLG2	529	25
RLC4	96	75
tRLC5	332	88
PpTc2	22	22

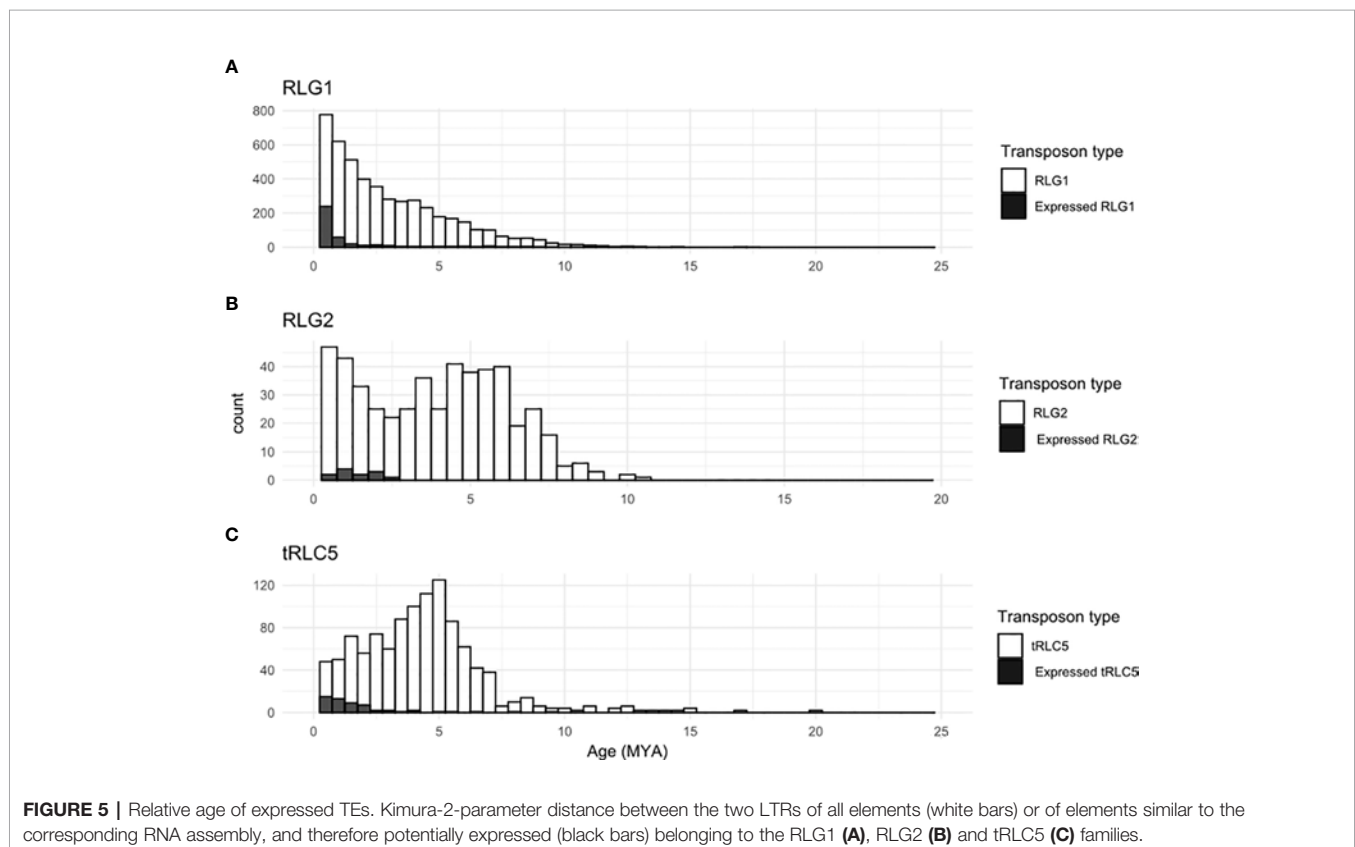
elements are located close to a gene, and the analysis of the patterns of expression of both genes provides information on the possible expression of the two Mariner-like elements. In the case of *PpTc1* the TE is only expressed in the conditions where the gene (Pp3c20_23510V3.1) is expressed (**Supplementary Figure 13**), which suggests that the expression detected for *PpTc1* could be the result of read-through transcription from the neighboring gene. On the contrary, the expression of *PpTc2* and the gene located nearby (Pp3c9_17220V3.1) do not overlap. Indeed, only *PpTc2*, and not the gene located nearby, is expressed in gametophores submitted to dehydration and rehydration and its expression is strongly induced in protonema treated with ABA which is not the case for the close by gene (**Supplementary Figure 14**). We confirmed the induction of *PpTc2* by ABA by qRT-PCR (**Supplementary Figure 15**). Therefore, whereas we cannot rule out the possibility that *PpTc1* expression could be the result of a readthrough expression from a neighboring gene, the transcript corresponding to *PpTc2* seems to be the result of a

genuine TE transcription. Moreover, the sequence variability of the RNA-Seq reads corresponding to *PpTc2*, suggests that other *PpTc2* elements may also be expressed. Indeed, although the *PpTc2* copy located in the vicinity of the Pp3c9_17220V3.1 gene is almost identical to the RNA assembly (99.4%), other *PpTc2* copies also show high similarity to the assembly (**Table 1**) and may also be expressed.

TE Mobility During Recent *P. patens* Evolution

The transcription of a copy of the TE in case of retrotransposons, and/or of the proteins necessary to mobilize the element, is the first and obligatory step of TE transposition. Therefore, the transcription of the different TEs reported here suggests that different TEs may have recently moved during *P. patens* evolution. We have already reported that this is indeed the case for RLG1, as RLG1 elements are polymorphic between the Gransden and Villersexel accessions. Here we decided to expand the analysis for possible insertion polymorphisms to all *P. patens* TEs using data from 4 different *P. patens* accessions, Reute, Kaskaskia, Villersexel, and the one from which the reference genome has been obtained, Gransden. To this end we used PopoolationTE2 to look for TE polymorphisms among these accessions using paired-end short-read resequencing data from Reute, Kaskaskia, Villersexel, that we mapped to the Gransden reference genome.

We found an important number of RLG1 polymorphisms in the three analyzed accessions with respect to Gransden (**Table 2**).



The number of polymorphisms in Reute was much smaller than in the two other accessions, which is in accordance with the close genetic relationship between Gransden and Reute (Hiss et al., 2017). Interestingly, in addition to polymorphisms related to RLG1 elements, we also detected polymorphic insertions of RLG2, RLG3, tRLC5/RLC5 and *PpTc1* (Table 2). In general, the number of polymorphisms is higher in Villersexel and smaller in Reute, as seen for RLG1.

The number of polymorphic insertions was particularly high for RLG3 and tRLC5/RLC5. In order to start analyzing the potential impact of the polymorphic insertions described here in the phenotypic differences between the four *P. patens* ecotypes, we analyzed the locations of the polymorphic TE insertions (Supplementary Table 3) and found that 20% of them are located close to genes, with potential consequences on their coding capacity or expression (Table 3).

DISCUSSION

The Challenging Analysis of TE Transcription

Different programs to measure TE transcription from NGS data exist (Jin et al., 2015; Lerat et al., 2017). These programs usually rely on mapping RNA-Seq reads to a TE annotation or a consensus of a TE family. Although these programs can be very useful for certain genomes and particular TE families, they may not be adequate in others. Indeed, most eukaryote genomes, and in particular those of plants, contain an important number of fragmented or degenerated TE copies in addition to full copies of TEs. As the TE fragments can also be included in transcripts, and outnumber the complete copies (Hoen et al., 2015; Bennetzen and Park, 2018), an estimation of the expression of TEs that would not discriminate between transcripts corresponding to TE fragments or to complete elements will overestimate the expression of certain families and will lead to erroneous results. This is what we came across when starting to analyze the expression of *P. patens* TEs. As an example, as already explained, among the short assemblies discarded there was one (TRINITY_DN331_c0_g1) corresponding to a RLC5 solo-LTR. An analysis of the RNA-Seq reads matching this assembly showed their specific accumulation in ABA-treated protonema cells and in gametophores under dehydration/rehydration stress. The results presented here show that the RLC5 solo-LTR is expressed as the result of read-through transcription from the ABA-induced *Pp3c4_32070* gene located just upstream of it. An analysis of RLC5 expression

based solely on mapping RNA-Seq reads to the TE annotation would have led to the wrong conclusion that RLC5 is induced by ABA and drought stresses. On the contrary, the approach described here, which is similar to the one previously described for the analysis of the expression of human TEs (Guffanti et al., 2018), allows for the assessment of the expression of RNAs corresponding to complete elements potentially resulting from genuine TE transcription.

Different Retrotransposon and DNA Transposon Families Are Transcribed in *P. patens*

The results presented here show that at least four LTR-RTs (RLG1, RLG2, RLC4 and tRLC5) and one DNA transposon (*PpTc2*) are expressed in *P. patens*. Among those, RLG1 and RLG2 are highly expressed during normal *P. patens* development, RLG1 being expressed mainly in protonema tissues whereas the expression of RLG2 is increased in gametophores. RLC4 seems also to be expressed in gametophores, albeit at a low level, and tRLC5 is expressed in young sporophytes. Therefore, during *P. patens* development, there is an important expression of different transposons. In addition, although RLG1 seems to be repressed by most stresses, different TEs are activated by stress. RLG2 is overexpressed under heat shock and ABA treatment, and *PpTc2* is induced by ABA and by dehydration and rehydration treatments. Mosses are known to be tolerant to dehydration and rehydration (Cuming et al., 2007; Cui et al., 2012), which, together with the associated changes of temperature, are part of their natural lifestyle. The dehydration/rehydration stresses and the ABA treatment, known to mediate the responses to those stresses (Cuming et al., 2007), and to some extent heat stress, could thus be considered as part of the normal development of *P. patens* or, at least, frequent stresses *P. patens* has to face.

Recent Mobilization of *P. patens* TEs

The expression of different TEs in normal *P. patens* growing conditions could allow the mobilization of TEs and the

TABLE 2 | TE polymorphisms in the different *P. patens* accessions.

Accession	RLG1	RLG2	RLG3	RLC4	tRLC5/RLC5	LINE-1	LINE-2	PpTc1	PpTc2	Total
Reute	18	0	4	0	5	0	0	0	0	27
Kaskaskia	147	0	15	0	17	0	0	0	0	179
Villersexel	229	2	48	2	21	0	0	1	0	303
Total	394	2	67	2	43	0	0	1	0	509

TABLE 3 | Distance of polymorphic TE insertions to genes.

Accession	Inside Genes	< 1 kb closest gene	> 1 kb closest gene	Total
Reute	4	4	19	27
Kaskaskia	12	27	140	179
Villersexel	22	34	247	303
Total	38	65	406	509

generation of genetic variability that could potentially affect gene expression/function in this haploid species. The analysis presented here shows that many TE insertions are polymorphic between different *P. patens* accessions. Indeed, we have detected an important number of polymorphic insertions of RLG1, RLG3 and tRLC5/RLC5 elements. The high number of polymorphisms related to RLG3 is intriguing as we did not detect expression. RLG3 may therefore be expressed under different environmental conditions not tested here. Alternatively, RLG3 may have lost the ability to transcribe and transpose recently during evolution. In all cases, the highest number of polymorphisms with respect to the Gransden accession is found in Villersexel and the lowest in Reute, which is in accordance with the number of SNPs these accessions show with respect to the Gransden reference genome (Lang et al., 2018). We have also found a small number of polymorphic insertions of RLG2, RLC4 and *PpTc1*. The number of detected TE polymorphisms with respect to the Gransden reference genome in these accessions is probably underestimated, as none of the programs available to look for TE polymorphisms, including the one used here, can detect polymorphic TE insertion sitting in repetitive regions (Vendrell-Mir et al., 2019). In any case, the polymorphisms detected here illustrate the potential of TEs to generate genetic variability in *P. patens*. Moreover, an important fraction of the polymorphisms detected are within or close (less than 1 Kb) to a gene, which suggests that TE movement may have impacted gene coding or gene regulation, and therefore may have contributed to the phenotypic variability of *P. patens*.

The Heterochromatic tRLC5 Elements Are Transcribed in Sporophytes

In addition to generate new alleles or new gene regulations, TEs are also involved in chromosome structure and function. In plants, TEs have been shown to provide origins of replication in heterochromatic regions (Sequeira-Mendes et al., 2019), and are frequently part of centromeres (Lermontova et al., 2015). Different retrotransposon have been found to specifically accumulate in the centromeres of the green algae *Coccomyxa subellipsoidea* (Blanc et al., 2012) or the liverwort *M. polymorpha* (Diop et al., 2020) were they could support centromere function. Interestingly, tRLC5 was previously proposed to mark the centromere and participate to the centromere function in *P. patens* (Lang et al., 2018). We show here that tRLC5 is transcribed in *P. patens*. In spite of its heterochromatic nature, centromere sequences have been shown to be transcribed in yeast, animals and plants and this transcription seems vital for the maintenance of the centromere chromatin identity and in several aspects of centromere function (Chan and Wong, 2012; Perea-Resa and Blower, 2018). Young sporophytes are a key developmental stage of *P. patens* where meiosis takes place (Charlot et al., 2014). We show here that most meiosis-specific genes (Mercier et al., 2015) are highly induced in green sporophytes (**Supplementary Figure 16**), the developmental stage where tRLC5 is expressed. It has been proposed that demethylation of centromeric DNA during meiosis may allow the transcription of centromeric sequences, which could serve as markers recognized by other factors and allow centromere assembly (Liu et al., 2015). The expression of tRLC5 in

the centromere, at the moment meiosis takes place, could thus play a role in centromere assembly and function during this key process. On the other hand, the transcription pattern of tRLC5, specifically activated in young sporophytes, is reminiscent of the expression of the Athila retrotransposon of Arabidopsis, which also concentrates in the centromere and is expressed in the pollen grain (Keith Slotkin, 2010). It has been proposed that TE expression in the vegetative nurse cells of the pollen may allow re-establishing its silencing in the sperm cells (Martínez et al., 2016). The expression of tRLC5 in the sporophyte could also fulfill a similar role. Further experimental work will be required to explore any of these two non-exclusive hypotheses.

Are Some of the *P. patens* TE Families the Result of a Horizontal Transfer from Fungal Species?

In addition to the characterization of the transcriptional activity of *P. patens* TEs, the work presented also allowed us to better characterize two Mariner-like elements. These *P. patens* elements, that are transcribed and mobile, are more closely related to fungal elements than to any Mariner-like element found in plants, suggesting that they may have been horizontally transmitted from fungi. Interestingly, another *P. patens* Mariner-like element already described was also shown to be closely similar to fungal TEs (Castanera et al., 2016), which suggest that the horizontal transfer of Mariner-like elements from fungi to *P. patens* may have been a frequent event. The Mariner TE family is ubiquitous in the genomes of virtually all extant eukaryotic species and seem to be particularly prone to horizontal transfer, probably because they contain a transcriptionally promiscuous “blurry” promoter (Palazzo et al., 2019). Early land plants were aided by mutualistic interactions with fungi and these symbiotic interactions with fungi have been maintained in some bryophytes such as *M. polymorpha* (Humphreys et al., 2010). Surprisingly, although *P. patens* contains the strigolactone signaling pathway, which induce mycorrhizal signaling, it has not been shown to establish mycorrhizal interactions (Delaux et al., 2013; Field and Pressel, 2018; Rensing, 2018). The potential horizontal transfer of Mariner-like elements could be a remnant of this lost interaction, although an ulterior close contact between *P. patens* and different fungi may have also be at the origin of these horizontal transfers. It is interesting to note that *P. patens* is the only plant that shares with fungi the traces of past infections of giant virus relatives (Maumus et al., 2014), which also highlights the close relationship with fungi that *P. patens* has maintained during its recent evolution.

CONCLUSION

In summary, the results presented here show that TEs have an important activity in *P. patens*, with the transcriptional activation of different TE families in normal *P. patens* growing conditions, suggesting that TEs may have shaped *P. patens* genome and may continue to contribute to its function, including adaptation to stresses and the intraspecific genetic variability.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

JC and FN conceived the project. PV-M performed all the experiments. ML-O obtained the RNA from green sporophytes. PV-M, FN, and JC drafted the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by grants from the Ministerio de Economía y Competitividad (AGL2016-78992-R) to JC and from the Investissement d'Avenir program of the French National Agency of Research for the project GENIUS (ANR-11-BTBR-0001_GENIUS) to FN. The work at CRAG is

REFERENCES

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2
- Anderson, S. N., Stitzer, M. C., Zhou, P., Ross-Ibarra, J., Hirsch, C. D., and Springer, N. M. (2019). Dynamic patterns of transcript abundance of transposable element families in maize. *G3 Genes Genomes Genet.* 9, 3673–3682. doi: 10.1534/g3.119.400431
- Bennetzen, J. L., and Park, M. (2018). Distinguishing friends, foes, and freeloaders in giant genomes. *Curr. Opin. Genet. Dev.* 49, 49–55. doi: 10.1016/j.gde.2018.02.013
- Blanc, G., Agarkova, I., Grimwood, J., Kuo, A., Bruggeman, A., Dunigan, D. D., et al. (2012). The genome of the polar eukaryotic microalga *Coccomyxa subellipsoidea* reveals traits of cold adaptation. *Genome Biol.* 13. doi: 10.1186/gb-2012-13-5-r39
- Bowen, N. J., and McDonald, J. F. (2001). Drosophila euchromatic LTR retrotransposons are much younger than the host species in which they reside. *Genome Res.* 11, 1527–1540. doi: 10.1101/gr.164201
- Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009). TrimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25, 1972–1973.
- Castanera, R., López-Varas, L., Borgognone, A., LaButti, K., Lapidus, A., Schmutz, J., et al. (2016). Transposable Elements versus the Fungal Genome: Impact on Whole-Genome Architecture and Transcriptional Profiles. *PLoS Genet.* 12. doi: 10.1371/journal.pgen.1006108
- Castanera, R., Ruggieri, V., Pujol, M., Garcia-Mas, J., and Casacuberta, J. M. (2020). An Improved Melon Reference Genome With Single-Molecule Sequencing Uncovers a Recent Burst of Transposable Elements With Potential Impact on Genes. *Front. Plant Sci.* 10, 1815. doi: 10.3389/fpls.2019.01815
- Chan, L. F., and Wong, L. H. (2012). Transcription in the maintenance of centromere chromatin identity. *Nucleic Acids Res.* 40, 11178–11188. doi: 10.1093/nar/gks921
- Charlot, F., Chelysheva, L., Kamisugi, Y., Vrielynck, N., Guyon, A., Epert, A., et al. (2014). RAD51B plays an essential role during somatic and meiotic recombination in *Physcomitrella*. *Nucleic Acids Res.* 42, 11965–11978. doi: 10.1093/nar/gku890
- Cui, S., Hu, J., Guo, S., Wang, J., Cheng, Y., Dang, X., et al. (2012). Proteome analysis of *Physcomitrella patens* exposed to progressive dehydration and rehydration. *J. Exp. Bot.* 64, 711–726. doi: 10.1093/jxb/err296

supported by a Spanish Ministry of Economy and Competitiveness grant for the Center of Excellence Severo Ochoa 2016–2019 (SEV-2015-0533); the IJPB benefits from the support of Saclay Plant Sciences-SPS (ANR-17-EUR-0007). PV-M holds a FPI (Formación de Personal Investigador) fellowship from the Spanish Ministerio de Economía y Competitividad.

ACKNOWLEDGMENTS

We thank all the members of JC's lab for their critical reading of the manuscript. ML-O would like to thank Eva Sundberg for their support during his post-doctoral stay at SLU. We also wish to thank Moaine Elbaidouri (U. Perpignan) for helpful discussions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpls.2020.01274/full#supplementary-material>

- Cuming, A. C., Cho, S. H., Kamisugi, Y., Graham, H., and Quatrano, R. S. (2007). Microarray analysis of transcriptional responses to abscisic acid and osmotic, salt, and drought stress in the moss, *Physcomitrella patens*. *New Phytol.* 176, 275–287. doi: 10.1111/j.1469-8137.2007.02187.x
- Delaux, P. M., Séjalon-Delmas, N., Bécard, G., and Ané, J. M. (2013). Evolution of the plant-microbe symbiotic “toolkit”. *Trends Plant Sci.* 18, 298–304. doi: 10.1016/j.tplants.2013.01.008
- Diop, S.II, Subotic, O., Giraldo-Fonseca, A., Waller, M., Kirbis, A., Neubauer, A., et al. (2020). A pseudomolecule-scale genome assembly of the liverwort *Marchantia polymorpha*. *Plant J.* 101, 1378–1396. doi: 10.1111/tpj.14602
- Field, K. J., and Pressel, S. (2018). Unity in diversity: structural and functional insights into the ancient partnerships between plants and fungi. *New Phytol.* 220, 996–1011. doi: 10.1111/nph.15158
- Flutre, T., Duprat, E., Feuillet, C., and Quesneville, H. (2011). Considering transposable element diversification in de novo annotation approaches. *PLoS One* 6, 1. doi: 10.1371/journal.pone.0016526
- Gao, X., Hou, Y., Ebina, H., Levin, H. L., and Voytas, D. F. (2008). Chromodomains direct integration of retrotransposons to heterochromatin. *Genome Res.* 18, 359–369. doi: 10.1101/gr.7146408
- Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883
- Guffanti, G., Bartlett, A., Klengel, T., Klengel, C., Hunter, R., Glinsky, G., et al. (2018). Novel Bioinformatics Approach Identifies Transcriptional Profiles of Lineage-Specific Transposable Elements at Distinct Loci in the Human Dorsolateral Prefrontal Cortex. *Mol. Biol. Evol.* 35, 2435–2453. doi: 10.1093/molbev/msy143
- Hiss, M., Meyberg, R., Westermann, J., Haas, F. B., Schneider, L., Schallenberg-Rüdinger, M., et al. (2017). Sexual reproduction, sporophyte development and molecular variation in the model moss *Physcomitrella patens*: introducing the ecotype Reute. *Plant J.* 90, 606–620. doi: 10.1111/tpj.13501
- Hoen, D. R., Hickey, G., Bourque, G., Casacuberta, J., Cordaux, R., Feschotte, C., et al. (2015). A call for benchmarking transposable element annotation methods. *Mob. DNA* 6, 13. doi: 10.1186/s13100-015-0044-6
- Humphreys, C. P., Franks, P. J., Rees, M., Bidartondo, M.II, Leake, J. R., and Beerling, D. J. (2010). Mutualistic mycorrhiza-like symbiosis in the most ancient group of land plants. *Nat. Commun.* 1 (1), 1–7. doi: 10.1038/ncomms1105
- Jin, Y., Tam, O. H., Paniagua, E., and Hammell, M. (2015). Tetrascripts: A package for including transposable elements in differential expression analysis

- of RNA-seq datasets. *Bioinformatics* 31 (22), 3593–3599. doi: 10.1093/bioinformatics/btv422
- Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* 30 (4), 772–780. doi: 10.1093/molbev/mst010
- Keith Slotkin, R. (2010). The epigenetic control of the athila family of retrotransposons in Arabidopsis. *Epigenetics* 5, 483–490. doi: 10.4161/epi.5.6.12119
- Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16, 111–120.
- Kofler, R., Gómez-Sánchez, D., and Schlötterer, C. (2016). PoPoolationTE2: Comparative Population Genomics of Transposable Elements Using Pool-Seq. *Mol. Biol. Evol.* 33, 2759–2764. doi: 10.1093/molbev/msw137
- Lang, D., Ullrich, K. K., Murat, F., Fuchs, J., Jenkins, J., Haas, F. B., et al. (2018). The Physcomitrella patens chromosome-scale assembly reveals moss genome structure and evolution. *Plant J.* 93. doi: 10.1111/tjp.13801
- Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9 (4), 357.
- Leebens-Mack, J. H., Barker, M. S., Carpenter, E. J., Deyholos, M. K., Gitzendanner, M. A., Graham, S. W., et al. (2019). One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 574, 679. doi: 10.1038/s41586-019-1693-2
- Lerat, E., Fablet, M., Modolo, L., Lopez-Maestre, H., and Vieira, C. (2017). TEtools facilitates big data expression analysis of transposable elements and reveals an antagonism between their activity and that of piRNA genes. *Nucleic Acids Res.* 45, 17. doi: 10.1093/nar/gkw953
- Lermontova, I., Sandmann, M., Mascher, M., Schmit, A. C., and Chabouté, M. E. (2015). Centromeric chromatin and its dynamics in plants. *Plant J.* 83, 4–17. doi: 10.1111/tjp.12875
- Letunic, I., and Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* 47, 256–259. doi: 10.1093/nar/gkz239
- Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352
- Liu, Y., and Yang, G. (2014). Tc1-like transposable elements in plant genomes. *Mob. DNA* 5. doi: 10.1186/1759-8753-5-17
- Liu, Y., Su, H., Zhang, J., Liu, Y., Han, F., and Birchler, J. A. (2015). Dynamic epigenetic states of maize centromeres. *Front. Plant Sci.* 6, 904. doi: 10.3389/fpls.2015.00904
- Marchler-Bauer, A., Derbyshire, M. K., Gonzales, N. R., Lu, S., Chitsaz, F., Geer, L. Y., et al. (2015). CDD: NCBI's conserved domain database. *Nucleic Acids Res.* 43, 222–226. doi: 10.1093/nar/gku1221
- Martínez, G., Panda, K., Köhler, C., and Slotkin, R. K. (2016). Silencing in sperm cells is directed by RNA movement from the surrounding nurse cell. *Nat. Plants* 2 (4), 1–8. 16030. doi: 10.1038/nplants.2016.30
- Maumus, F., Epert, A., Nogué, F., and Blanc, G. (2014). Plant genomes enclose footprints of past infections by giant virus relatives. *Nat. Commun.* 5, 4268. doi: 10.1038/ncomms5268
- Mercier, R., Mézard, C., Jenczewski, E., Macaisne, N., and Grelon, M. (2015). The Molecular Biology of Meiosis in Plants. *Annu. Rev. Plant Biol.* 66, 297–327. doi: 10.1146/annurev-arplant-050213-035923
- Ou, S., Su, W., Liao, Y., Chougule, K., Ware, D., Peterson, T., et al. (2019). Benchmarking Transposable Element Annotation Methods for Creation of a Streamlined, Comprehensive Pipeline. *Genome Biol.* 20, 275. doi: 10.1186/s13059-019-1905-y
- Palazzo, A., Lorusso, P., Miskey, C., Walisko, O., Gerbino, A., Marobbio, C. M. T., et al. (2019). Transcriptionally promiscuous “blurry” promoters in Tc1/mariner transposons allow transcription in distantly related genomes. *Mob. DNA* 10:13. doi: 10.1186/s13100-019-0155-6
- Perea-Resa, C., and Blower, M. D. (2018). Centromere Biology: Transcription Goes on Stage. *Mol. Cell. Biol.* 38. doi: 10.1128/mcb.00263-18
- Perroud, P. F., Haas, F. B., Hiss, M., Ullrich, K. K., Alboresi, A., Amirebrahimi, M., et al. (2018). The Physcomitrella patens gene atlas project: large-scale RNA-seq based expression data. *Plant J.* 95, 168–182. doi: 10.1111/tjp.13940
- Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS One* 5. doi: 10.1371/journal.pone.0009490
- Quinlan, A. R., and Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi: 10.1093/bioinformatics/btq033
- Rensing, S. A., Ick, J., Fawcett, J. A., Lang, D., Zimmer, A., Van De Peer, Y., et al. (2007). An ancient genome duplication contributed to the abundance of metabolic genes in the moss Physcomitrella patens. *BMC Evol. Biol.* 7, 130. doi: 10.1186/1471-2148-7-130
- Rensing, S. A., Lang, D., Zimmer, A. D., Terry, A., Salamov, A., Shapiro, H., et al. (2008). The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. *Science* 319, 64–69. doi: 10.1126/science.1150646
- Rensing, S. A. (2018). Great moments in evolution: the conquest of land by plants. *Curr. Opin. Plant Biol.* 42, 49–54. doi: 10.1016/j.pbi.2018.02.006
- Schaefer, D. G., and Zrýd, J. P. (2001). The moss Physcomitrella patens, now and then. *Plant Physiol.* 127, 1430–1438. doi: 10.1104/pp.010786
- Schaefer, D. G., Delacote, F., Charlot, F., Vrielynck, N., Guyon-Debast, A., Le Guin, S., et al. (2010). RAD51 loss of function abolishes gene targeting and de-represses illegitimate integration in the moss Physcomitrella patens. *DNA Repair (Amst)*. 9, 526–533. doi: 10.1016/j.dnarep.2010.02.001
- Sequeira-Mendes, J., Vergara, Z., Peiró, R., Morata, J., Aragüez, I., Costas, C., et al. (2019). Differences in firing efficiency, chromatin, and transcription underlie the developmental plasticity of the Arabidopsis DNA replication origins. *Genome Res.* 29, 784–797. doi: 10.1101/gr.240986.118
- Tenaillon, M.II, Hollister, J. D., and Gaut, B. S. (2010). A triptych of the evolution of plant transposable elements. *Trends Plant Sci.* 15, 471–478. doi: 10.1016/j.tplants.2010.05.003
- Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., et al. (2012). Primer3-new capabilities and interfaces. *Nucleic Acids Res.* 40, 115. doi: 10.1093/nar/gks596
- Vendrell-Mir, P., Barteri, F., Merenciano, M., González, J., Casacuberta, J. M., and Castanera, R. (2019). A benchmark of transposon insertion detection tools using real data. *Mob. DNA* 10, 53. doi: 10.1186/s13100-019-0197-9
- Vives, C., Charlot, F., Mhiri, C., Contreras, B., Daniel, J., Epert, A., et al. (2016). Highly efficient gene tagging in the bryophyte Physcomitrella patens using the tobacco (Nicotiana tabacum) Tnt1 retrotransposon. *New Phytol.* 212, 759–769. doi: 10.1111/nph.14152

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Vendrell-Mir, López-Obando, Nogué and Casacuberta. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Chapter 2.4: Complementary results and discussion

This analysis has confirmed that RLG1, as was observed in the previous data (Lang et al., 2018; Vives et al., 2016), is expressed in protonemata and that the DNA TEs *PpTc1* and *PpTc2* are expressed in *P. patens* as was previously described (Liu & Yang, 2014). This analysis has also identified the transcription of other families that were not characterized as expressed, such as the RLC4 and RLG2 families. With this study we also characterized under which conditions some TE families can be induced. It should be noted that the gene atlas contained RNAseq data for only two accessions, Gransden for most of the conditions except for sporophyte development where the Reute accession was used, detecting expression of the RLC5 family. Since the publication of this article, we have analyzed the expression of RLG1 in different samples of protonemata of other accessions (Kaskaskia and Reute) confirming its expression also in these accessions.

Interestingly, we found TE families that are expressed under development conditions and are repressed under stresses such as RLG1, that is repressed under all tested stresses and is only expressed in protonemata tissue. We also identified other TEs that are only induced under specific developmental stages, such as RLC5, that is induced in the sporophytic phase of the development where meiosis takes places. Plant TEs expression has usually been attributed to stress conditions (Grandbastien, 1998), but as observed here, this can change widely between different TE families. However, we also observed some TEs that are induced under specific stresses such as RLG2 and *PpTc2* that are induced under ABA treatment.

Regarding the transposition landscape, we detected some families that are highly expressed but for which only a few polymorphisms among all accessions were detected, such as the RLG2 family. On the contrary, we also found several polymorphic TE insertions for the RLG3 family, when this family was not detected as expressed. As this study is limited to a certain number of conditions, it may not fully reflect all the possible conditions or stresses that the plant may undergo in nature and under which the TEs may be expressed and transpose. In the same way, it is also possible that some of these families are no longer able to transpose in the sequenced individuals and remain active in other *P. patens* individuals.

To test if some of these families that we detected as transcriptionally active are able to transpose on the moss maintained in the laboratory, we took advantage that the group of Fabien Nogu . our collaborators from IJPB-INRA had sequenced, using Illumina paired-end short-reads, genomic DNA of the same clone maintained in the lab for the past 20 years at four different time points: 2007, 2011, 2016 and 2018 (Bessoltane et al., 2022). These samples were mostly maintained by propagating them asexually during all this time.

We used a combination of the results of PopoolationTE2 and Jitterbug to maximize the chance to detect new transposition events that occurred during this period. However, we could not detect any new transposition event. There are several reasons that could explain this result. First, that one of the main limitations of the strategies to detect TE polymorphisms from resequencing data is that they are unable to detect transpositions events in highly repetitive sequences (Vendrell-Mir et al., 2019). Therefore, if the TEs are preferentially targeting these regions, we may miss the transposition events that may have occurred. Moreover, as *P. patens* has been propagated mostly asexually, it could be possible that there have been some somatic transposition events that have not been maintained in the consecutive propagation of the moss or that we could not detect as they were only present in a small part of the total cell population. Finally, there could be silencing mechanisms, such as the presence of small RNAs targeting the RLG1 and RLG3 elements (Coruh et al., 2015) that could inhibit their transposition and, for this reason, we were not able to identify any transposition event from this data.

CHAPTER 3: IMPACT OF TEs IN
PHYSCOMITRIUM PATENS GENOME

CHAPTER 3: IMPACT OF TEs IN *PHYSCOMITRIUM PATENS* GENOME

Chapter 3.1: Introduction

In this third chapter we will focus on the impact of TEs in the structure of the genome and the genes of *P. patens*. As already explained, the heterochromatic regions and the genes of *P. patens* are distributed evenly among the chromosomes forming heterochromatic regions interspersed with euchromatic regions. One of the main components of these heterochromatic islands are RLG1 elements. In the first part of this chapter, we will focus on the impact of this LTR-RT family in the structure of the genome, particularly in the heterochromatin by removing heterochromatic regions mainly formed by RLG1 elements.

The second part of this chapter focuses on the impact of TEs on gene expression. As we observe in our study of the polymorphic TEs (Vendrell-Mir et al., 2020) and as it was described in the publication of the reference genome (Lang et al., 2018), although most of the RLG1 elements are located in these RLG1 islands, it is also possible to find RLG1 elements in the vicinities of genes. These RLG1 elements could have an impact on the expression of these genes. To study the impact of these TE insertion polymorphisms in gene expression we have selected a few cases of TEs close to genes.

Chapter 3.2: Objectives:

- Study the potential impact of Transposable Elements in the structure of the genome of *Physcomitrium patens*.

- Study the potential impact of Transposable Elements in the genic regions in *Physcomitrium patens*.

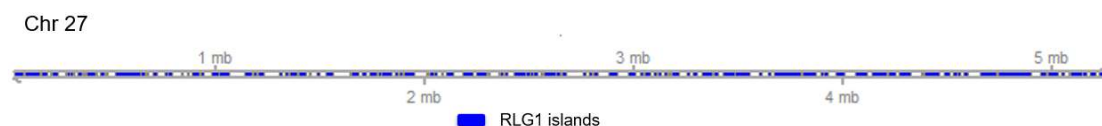
Chapter 3.3: Impact of Transposable Elements in the structure of *Physcomitrium patens* genome

Introduction

As presented before, the genome of *P. patens* has heterochromatic regions distributed all along the chromosome arms. These heterochromatic regions are mostly formed of Gypsy elements, with the RLG1 LTR-RT being the main component, representing almost half of the TE content of the genome and a 25% of the sequenced genome (Lang et al., 2018). This percentage could be even higher as most of these RLG1 elements are forming nested structures, generating highly fragmented structures that are difficult to annotate. In this work we refer to all the heterochromatic regions that contain RLG1 elements as RLG1 islands.

We found these islands distributed homogeneously among the chromosomes (Figure 18), ranging in size from 6.5 Kbp to up to 372 Kbp.

Figure 18: Example of the distribution of the RLG1 islands in a chromosome, in this case in the chromosome 27 of *P. patens*.



To try to understand what the impact of these elements in the maintenance of the structure of the genome is, we have designed different strategies to eliminate these RLG1 islands using genome editing technologies, in this case the CRISPR/Cas9 system.

The development of CRISPR/ Cas9 strategies has allowed the production of targeted double strand breaks allowing gene editing of the genomes of several plant species in the last decade (Bortesi & Fischer, 2015), including *P. patens* (Collonnier et al., 2017). This tool has been widely used to produce targeted mutagenesis including deletions of several Kbp in plant genomes, for example in rice (Zhou et al., 2014). Therefore, we decided to

study the impact of removing heterochromatic RLG1 elements using the CRISPR/Cas9 system.

We have designed two strategies, both based on the use of CRISPR/Cas9, to produce targeted deletions of RLG1 elements in the *P.patens* genome. As a first strategy we targeted the LTRs of the RLG1 elements to produce as many eliminations of these elements as possible. The second strategy was to delete specific RLG1 islands of the smallest chromosome of *P. patens*, targeting the unique sequences flanking these regions.

Most of the *P. patens* transformations needed were performed by Florence Charlot from the group of Fabien Nogu  at the IJPB, and the phenotyping of the obtained clones was also performed in the Lab of Fabien Nogu . I performed the design of the constructs in collaboration with Nogue's group. I also performed the cloning of the different constructs and the genotyping of the clones, including the PCRs and the qPCRs. I received the help of the undergraduate student Am lia Morat  who worked under my supervision to genotype part of the clones and to do some cloning steps for the constructs used over this chapter. Jordi Morata from our group also helped me develop Perl script to find gRNA sequences over the TEs of the genome.

Results

Non-selective elimination of RLG1 elements using CRISPR/Cas9:

Selection of a guide RNA to target and eliminate RLG1 elements from *P. patens* using the CRISPR/Cas9 system

The annotation of the published *P. patens* genome (Lang et al., 2018) has over 77987 RLG1 fragments and at least 5092 complete elements. The number of elements in the genome is probably higher as there are parts of it that are still lacking and have not been assembled to the reference genome.

To produce double strand cuts over the RLG1 elements we used the *Streptococcus pyogenes* CRISPR/Cas9 (SpCas9). The guide sequence of the SpCas9 requires a given sequence of 20 nucleotides that matches the specific sequence of interest and the Protospacer Adjacent Motive (PAM) NGG.

To look over all the RLG1 sequence and find which was the guide sequence that will maximize the number of cuts produced inside the RLG1 elements, Jordi Morata, a postdoc in our group during the time that the project was designed, developed a script to look for all these possible sequences matching these criteria and classify them considering the maximum number of cuts produced.

Filtering with a minimum of 2500 putative hits in the genome, we ended up with 27 gRNAs. Using these gRNAs we could produce targeted cuts in the RLG1 elements that could affect between 1792 to 6293 RLG1 elements, considering that a 100% of the targeted sequenced is cut (Table 6).

Table 6: Sequences selected to produce elimination of RLG1 elements all over the genome filtered with a minimum of 2500 cuts over the genome. In grey in the table the selected gRNA to proceed with the elimination.

guide RNA sequence	Hits genome	Hits RLG1 elements	Hits in other TEs	Hits in genic regions	% hits in RLG1 of the genome	% hits in heterochromatic regions affected
CTATCTAACTAGGGGCTACGTGG	2516	1887	213	21	2,420%	16,31%
TTTGCCATCAGTTTGAGGATGG	2517	2505	183	14	3,212%	20,90%
TCTCTAGTTTTCTTGCTTTTGG	2571	2046	265	53	2,624%	17,40%
TAATAATTACATATAGAAATAGG	2644	1792	223	44	2,298%	15,08%
ACCATGGTCTCTGTTTCTATGG	2761	2750	176	23	3,526%	22,34%
TCCATAGAAACAGAAGACCATGG	2763	2752	11	23	3,529%	22,34%
ATTATTAGCTTACATGATTATGG	2970	1998	972	49	2,562%	16,53%
AGGCAAGTTTTTCTACGTGTGGG	3029	2306	723	43	2,957%	19,01%
AAGGCAAGTTTTTCTACGTGTGG	3181	2415	301	33	3,097%	19,67%
AACTAGACTAGAAGCAAGAAAGG	3335	2399	229	37	3,076%	19,68%
TGCTTCTAGTCTAGTTTCAAGGG	3390	2398	239	36	3,075%	19,62%
ATATATAGAGACAAGAGTGAAGG	3592	2455	302	43	3,148%	19,25%
TATATATACTATCTAACTAGGGG	3677	2793	369	43	3,581%	22,48%
TTGCTTCTAGTCTAGTTTCAAGG	3734	2644	285	42	3,390%	21,34%
TGAAGATCAAGCTAACTATGTGG	3766	2723	318	47	3,492%	21,42%
CCAATATGTGTTGACTTGTAGG	3859	3250	383	61	4,167%	24,85%
CTATATATACTATCTAACTAGGG	3884	2949	390	42	3,781%	23,36%
TAGAAACGTGGTCAACAAGTTGG	4101	3082	389	44	3,952%	23,88%
GACTTTTTGGATTAGAAACGTGG	4113	3056	405	46	3,919%	23,82%
GACACATGCCTTGACTTACAAGG	4319	3272	408	52	4,196%	24,95%
TCTATATATACTATCTAACTAGG	4528	3480	471	51	4,462%	26,61%
ACAAAGTCAACACATATTGGAGG	4759	3624	444	76	4,647%	27,13%
CTAACAACATGTGTGGCACATGG	4824	3653	463	62	4,684%	27,35%
GGTGTCTTAAATTGACTTTTTGG	4875	3681	452	75	4,720%	27,88%
AAAACCTGCCTTGTAAGTCAAGG	4927	3751	461	65	4,810%	27,77%
CTTCAAGCTAACAACATGTGTGG	5102	3881	488	64	4,976%	28,55%
AGGAGTTGACAAGAGTGAAGAGG	6775	6293	464	18	8,069%	32,27%

With the goal of minimizing the effect on genic regions and avoiding direct cuts on genes we looked for overlaps of the guide sequences with genic regions, discarding all these gRNA sequences that have targets in genes. We ended up selecting the gRNA that produce 6775 cuts in the genome. We look manually to all the possible cuts produced inside genes for this gRNA observing that these sequences corresponded to genes that were poorly predicted in the gene annotation and were actually LTR-RT sequences. Moreover,

this sequence matched the LTRs of the RLG1 elements that, after the cleavage, could facilitate the recombination between the two LTRs of a single element or between LTRs of different RLG1 of an RLG1 islands, removing part of these heterochromatic regions of the genome.

To predict the efficiency of the CRISPR/Cas9 system using this gRNA we used the CRISPOR software (Concordet & Haeussler, 2018). This online tool provides information about the specificity of the gRNA (number off-targets in a given genome) and predicts the efficiency of the cleavage (Table 7).

Table 7: Prediction by CRISPOR of the efficiency of the designed guide sequence.

Position/ Strand	Guide Sequence + PAM + Restriction Enzymes <input type="checkbox"/> Only G- <input type="checkbox"/> Only GG- <input type="checkbox"/> Only A-	MIT Specificity Score	CFD Spec. score	Predicted Efficiency		Outcome	Off-targets for 0-1-2-3-4 mismatches + next to PAM	Genome Browser links to matches sorted by CFD off-target score
				Show all scores Doench TIG Mor-Mateos		Out-of-Frame Lindel		No exons. <input type="checkbox"/> Chr10 only
21 / fw	AGGAGTTGACAAGAGTGAAG AGG Enzymes: MbolI Cloning / PCR primers	0	2	63	57	60 81	6775 - 4043 - 3762 - 3251 - 1054 6775 - 2562 - 1763 - 956 - 259 18885 off-targets	0:Chr10 13.01 Mbp 0:scaffold_1126 2.40 Kbp 0:Chr01 24.83 Mbp show all...

As expected, the specificity scores were low as the tool is designed for single cuts in the genome. Despite that, the predicted efficiency scores were high enough for efficient cleavage. This tool also gave us an estimation of all the possible additional cleavage sites. With a minimum of 6775 without any mismatch and a maximum of 18885 cuts considering 4 mismatches in the guide sequence.

Production of a *P. patens* line containing a stable integration of a CRISPR/Cas9 targeting the RLG1 element

In order to increase the efficiency of the approach, we decided to obtain first a transgenic line expressing the CRISPR/Cas9 and the guide RNA against RLG1 elements. The transgene that we designed contained two homology arms at each side of the construct for integration at a specific location (between two highly expressed genes in chromosome 18, see Figure 19) by homologous recombination, it also contains the gRNA targeting the RLG1 elements, the CRISPR/Cas9 protein and a nptII selective marker to be able to select the clones that have integrated the complete locus using the antibiotic G418.

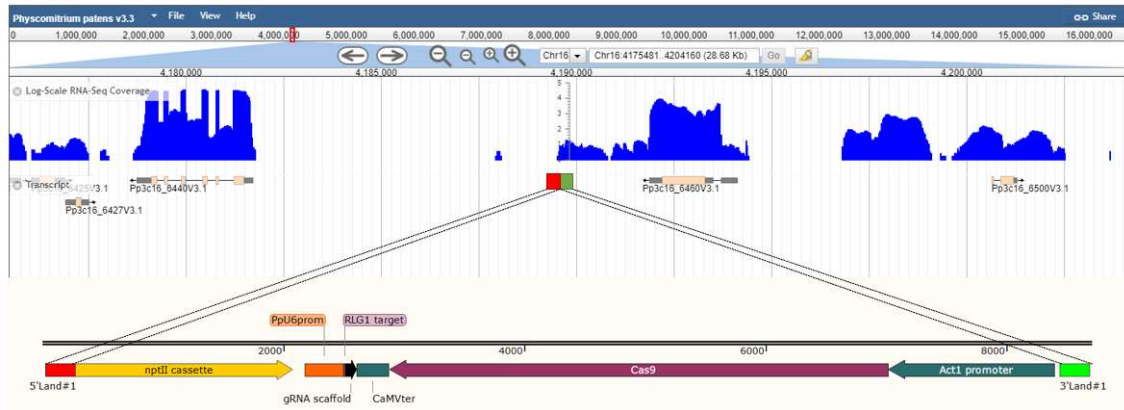


Figure 19: Introduction of the cassette containing the *nptII* gene, the gRNA with the guide targeting the RLG1 elements and the Cas9 using two DNA homologous sequences to the 3' of the gene *Pp3c16_6420V3.1* and at the 5' of the gene *Pp3c166440V3.1*. Both genes are highly expressed. The same strategy has been used previously by our collaborators to introduce transgenes using the same recognition sequence without any problem.

We performed two independent transformations using this construct obtaining a total of 39 clones that were transiently resistant to G418 (21 from the first transformation and 18 from the second transformation). However, we did not obtain any clone stably resistant to G418. This could suggest a detrimental effect of the expression of the insert although it may also suggest that the strategy was not working as expected due to unknown reasons.

When compared to previous transformations with other constructs, the number of transiently resistant clones was also low. In the transformation that we obtained the maximum number of transiently resistant clones this was the 0,007% from the total of transformed protoplasts, while in the previous transformations performed in the laboratory we obtained an efficiency of transiently resistant clones between a 0,52% and a 1.1% (Vives et al., 2016).

Although the obtained result suggested that no integration of the locus happened in the clones that were transiently resistant to G418, it would be possible that the CRISPR/Cas9 system worked in these clones and had produced cuts over the RLG1 elements above the genome. For this reason, we selected 22 clones that were able to regenerate, and we extracted DNA to check if there had been changes on the population of RLG1 elements in the genome.

As a first strategy, we selected 5 complete RLG1 elements flanked by unique sequences and that had the gRNA recognition sequences in both LTRs to check for a deletion of the RLG1 element. These RLG1 elements are located at: Chr01:5840024-5840024,

Chr08:8353455-8356012, Chr11: 6970182-6976693, Chr15:1092145-1098677 and Chr22:8577002-864284.

Although in all the positive controls (Gransden Wt) we could amplify the TEs, in most of the CRISPR/Cas9 transformed clones we could not amplify them. However, in most cases we could not amplify either a band corresponding to the deletion. Interestingly, in a single clone (Clone 18) we could detect a band that could correspond to the deletion for the RLG1 element located at Chr15:1092145-1098677 (Figure **20**).

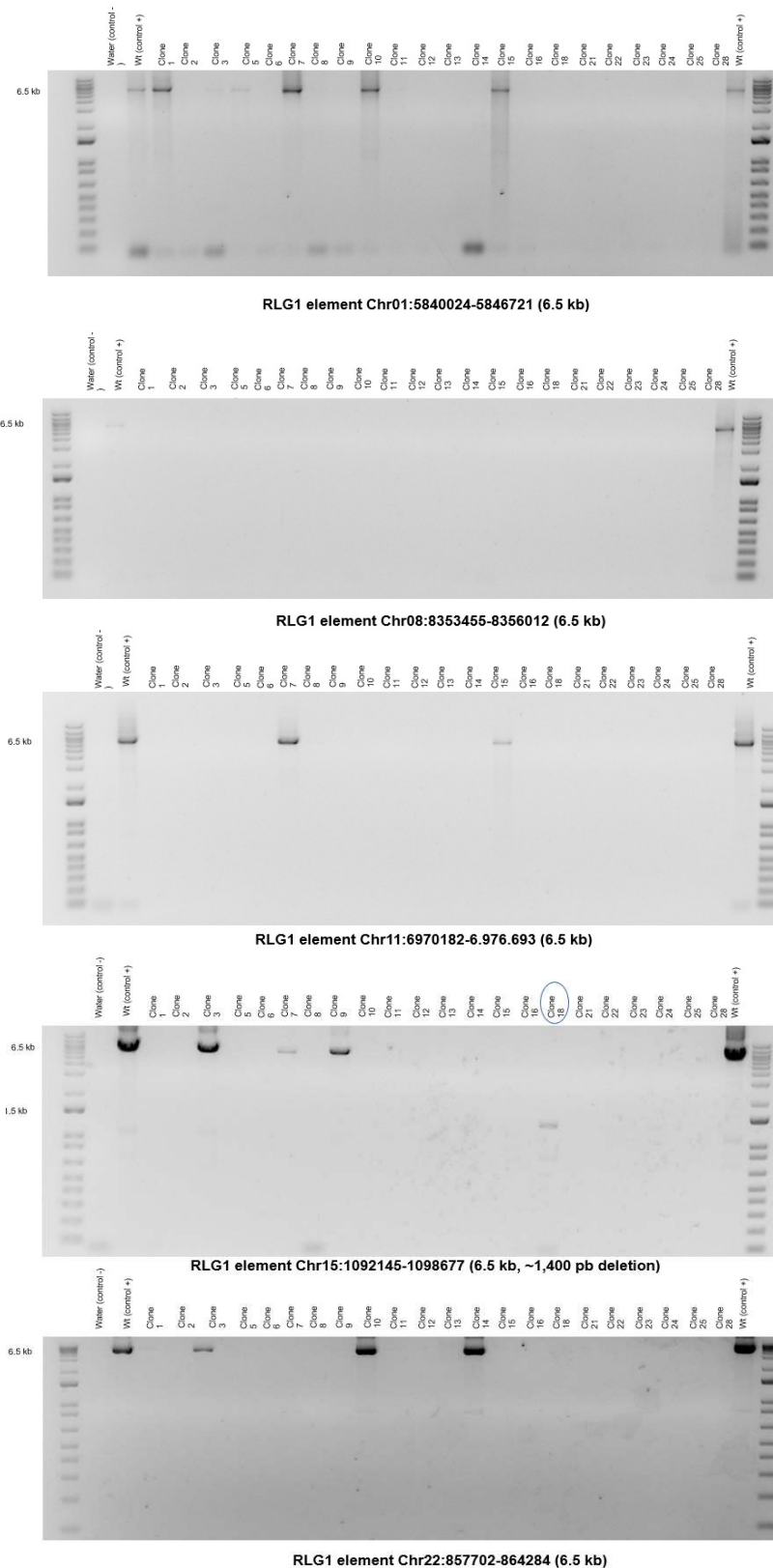


Figure 20: Agarose gel of the PCR products of 5 different RLG1 elements located at 5 different chromosomes. In Gransden Wt, a band of 6.5 kb is expected while a smaller band is expected in the different clones that have been transformed with the CRISPR/Cas9 targeting the RLG1 elements. Only a smaller band has been observed in clone 18 for the RLG1 element located at chromosome 15 (blue circle) while in all the other clones no band was amplified, or it was amplified a band corresponding to the Wt locus.

To test if the presence of the CRISPR/Cas9 targeting the RLG1 had an effect decreasing the total number of RLG1 elements, we selected two lines (Clone 12 that we could never amplify none of the TEs selected and Clone 18 that seemed to have at least a deletion) to perform a qPCR to compare the number of RLG1 elements to the single copy gene APT (Figure 21).

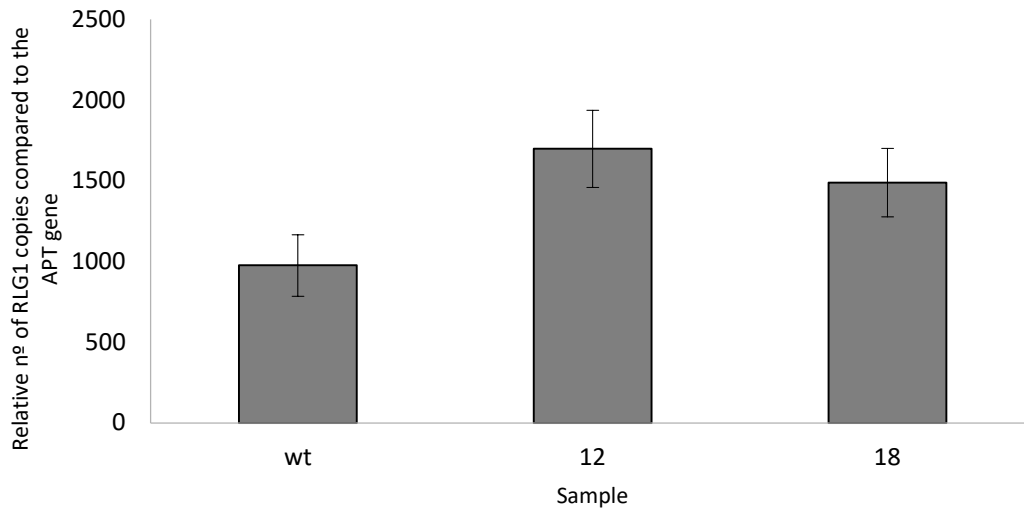


Figure 21: RLG1 quantification compared to the single copy gene APT for Gransden Wt compared to two clones that have been transiently transformed with the CRISPR/Cas9 targeting the RLG1 elements.

Surprisingly, the number of RLG1 clones detected was slightly higher in these two clones, and in any case, it was not lower.

Summarizing all the obtained results from these first transformations, we obtained a low number of regenerated clones when transformed with the CRISPR/Cas9 system targeting the RLG1 elements, none of them had stably integrated the CRISPR/cas9 system targeting the RLG1 elements and in the clones that we performed a qPCR we did not observe a decrease in the total number of RLG1 elements in the genome.

All this data suggests that producing double strand breaks over the RLG1 elements could be detrimental for the cells and that the only clones that were able to regenerate were the ones with a limited number of cuts or were the CRISPR/Cas9 system was not active.

As a way to select rare deletions, we designed a DNA template containing two homology arms of the edited LTRs without the recognition sequence flanking a Hygromycin resistance gene (Figure 22).

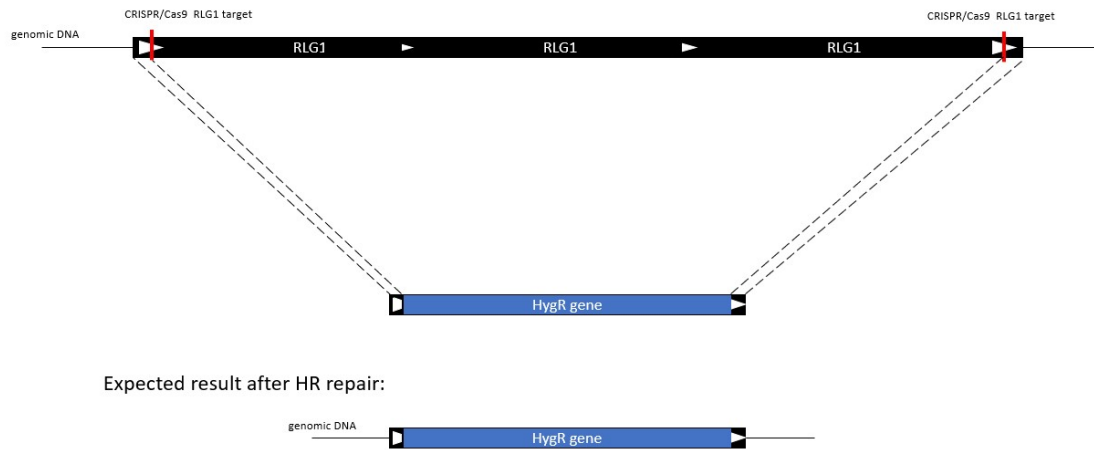


Figure 22: Use of a DNA template to replace the RLG1 islands that have been cut by the CRISPR/Cas9 system for a donor template containing homology to the LTRs to be integrated by homologous recombination after the cut took place and a HygR gene to select those clones that have been integrated the construct.

With the help of Amalia Morató, a former undergraduated student in the laboratory, we performed a *P. patens* transformation with this recombination template together with the CRISPR/Cas9 and the gRNA to target the RLG1 elements but unfortunately, we could not obtain any clone stably resistant to Hygromycin.

Analyzing the impact of the production of targeted DSB over the RLG1 elements

To ensure that the CRISPR/Cas9 was producing the predicted double strand breaks into the RLG1 elements and evaluate its possible detrimental effect we designed a new experiment. In this case we compared the effect of targeting a single copy gene with a gRNA that was used previously by our collaborators and was known to efficiently cut at the APT gene. We then compared the regeneration of the protoplasts where this single copy gene was targeted and protoplasts where this was combined with the targeting of the RLG1 elements.

We decided to compare to the targeting of the APT gene as it encodes for the APRT protein than in the normal function of a cell converts adenine to adenosine monophosphate, but it can also convert adenine analogous compounds to toxic components that will result in the dead of the cell, such as in the case of the use of 2-Fluoroadenine (2FA) that will be converted to the toxic compound 2-FluoroAMP (Schaff, 1994). If the APT gene is mutated it would not be able to metabolize the 2FA and convert

it to the toxic compound 2-FluoroAMP, being able to survive in a medium with 2FA. This system has widely been used by our collaborators, Dr. Nogu  Lab. (Collonnier et al., 2017; Guyon-Debast et al., 2021; P.-F. Perroud et al., 2022). The limitation of this system is that the APT mutated clones are sterile and have a strong phenotype (less gametophores) when compared to Wt Gransden.

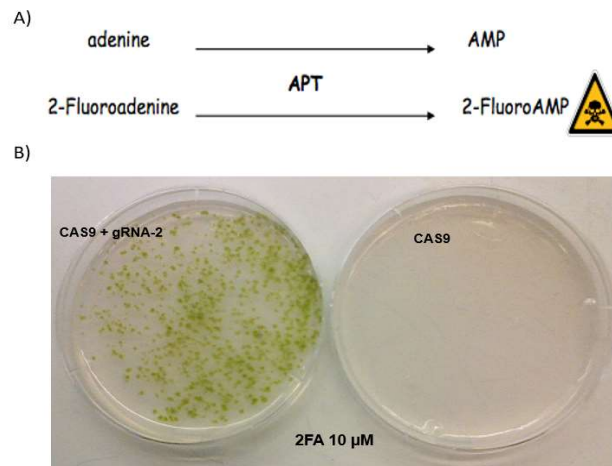


Figure 23:A) Under normal conditions Adenine is metabolized by APRT protein to AMP but when 2-Fluoroadenine (2FA) is present the APRT metabolize this compound converting it to the toxic compound 2-FluoroAMP that is lethal for the organism. B) Comparison of the regeneration after transforming with the CRISPR/Cas9 system and a gRNA targeting the APT gene (left) compared to a transformation only transformed using the CRISPR/Cas9 system without any gRNA (right) both under a BCDA medium supplemented with 2FA. Only the clones that have the APT gene can survive (left) while without the edition of the APT gene all the clones die (right). Figure provided by Dr. Fabien Nogu .

Florence Charlot from Fabien Nogu  group performed two independent transformations. From these transformations we confirmed that targeting the RLG1 elements has a huge detrimental effect for the cells (Table 8).

Table 8: Results of the transformations comparing the regeneration transforming with the CRISPR/Cas9 targeting the APT gene compared to transforming with the CRISPR/Cas9 targeting the APT and the RLG1 elements at the same time.

	Transformation	Regenerating protoplasts after 1 week	Clones resistant to 2FA	% Of regeneration compared to the total number of regenerating protoplasts
1 st Repetition	CRISPR/Cas9 +gRNA APT	28400	1350	4,75%
	CRISPR/Cas9 +gRNA APT+gRNA RLG1	15050	6	0,04%
2 nd Repetition	CRISPR/Cas9 +gRNA APT	14001	759	5,42%
	CRISPR/Cas9 +gRNA APT+gRNA RLG1	17113	18	0,11%

We compared the development of the clones that were resistant to APT and where we also targeted the RLG1 elements with those where we only targeted the APT gene (G2.4). In three clones (K2, K6 and K15) we observed less elongation of the protonema when compared to the G2.4 clone under the medium BCD (medium not supplemented with ammonium widely used to phenotype in *P. patens*). This phenotype was particularly strong in the clone K6 where the protonema was drastically reduced with less presence of caulonema compared to the line G2.4 (Figure 24).

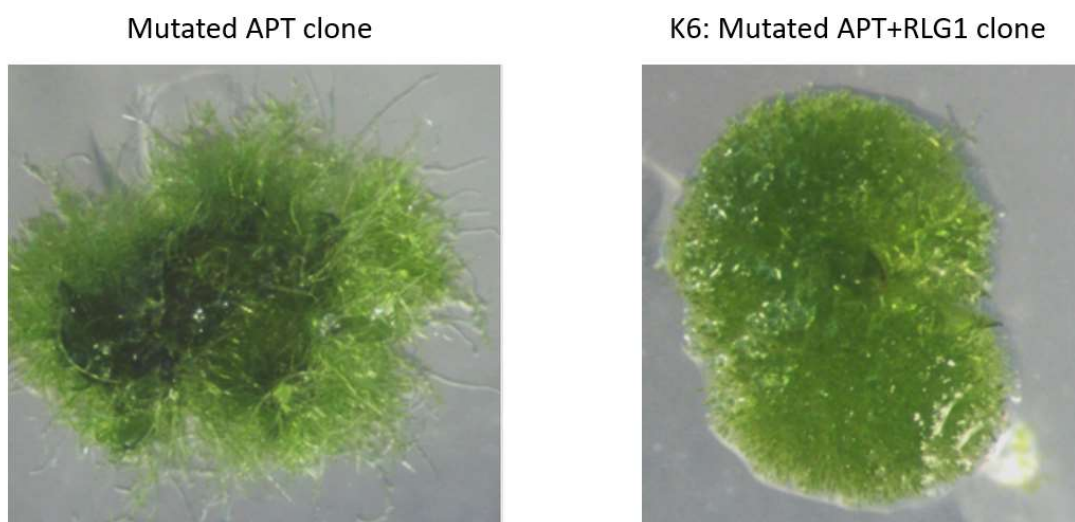


Figure 24: Phenotypic differences between the clone 6 transformed with the CRISPR/Cas9 system targeting the APT+ the RLG1 elements (right) compared to the clone G2.4 that has only been edited the APT gene both grown under BCD medium (left). A lesser presence of caulonema is observed in the clone 6 with a reduced growth compared to the clone G2.4.

We managed to extract DNA of 12 different clones where we targeted both the APT and the RLG1 elements. To check if there were major changes in the RLG1 content we performed a quantification of the RLG1 elements by qPCR used the single copy gene Pp3c10_20470V3.1 as an internal control. In general, we did not observe a substantial reduction of the number of RLG1 elements in any clone (Figure 25).

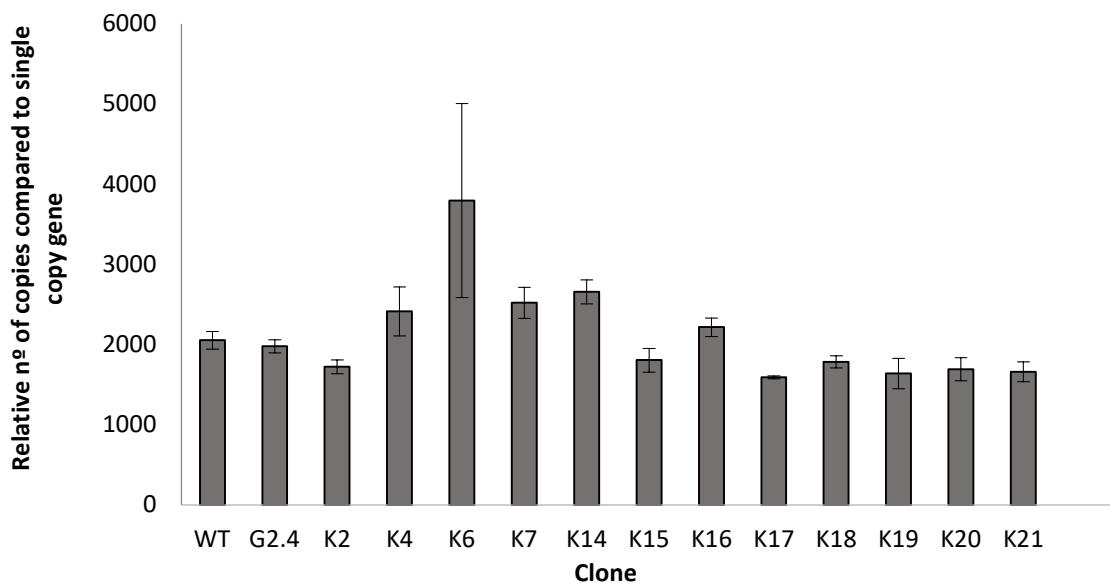


Figure 25: Relative quantification of the Number of RLG1 elements compared to the Pp3c10_20470V3.1 gene for the different samples that have been transformed with the CRISPR/Cas9 targeting the RLG1 and the APT gene compared to Gransden Wt and to G2.4 (APT KO).

There is a slight decrease in the number of RLG1 copies in some clones, such as K2 or K17, but this result could also be explained by the variability of the qPCR technique. Interestingly, the line that exhibit a strong phenotype K6 it seems to have a significant increase of the number of RLG1 copies. This increase could be due to an induction of the RLG1 elements transposition or due to the repair of the deletions using RLG1 islands of the genomes that are much bigger than the ones eliminated, increasing the number of RLG1 elements. To analyze whether of the two hypothesis is true we will need to sequence these lines using long reads approaches and compare it with the not treated samples.

Selective elimination of Chromosome 27 RLG1 islands

Although our results suggest that producing targeted cuts to many RLG1 elements in the genome is detrimental for the host survival, we could not analyze in detail the effect of these eliminations. For this reason, we designed an alternative approach to study the impact of eliminating a selected number of RLG1 islands from a single chromosome and study the effect of these eliminations over the maintenance of the structure of this chromosome. To do that, we decided to eliminate RLG1 islands from the smallest chromosome of *P. patens*, chromosome 27.

We first defined the RLG1 islands of the chromosome 27 by searching all the RLG1-rich repetitive regions that contained at least 2 RLG1 elements and that did not contain any gene inside, obtaining a total of 110 RLG1 islands, representing a total of a 62% of the Chromosome 27 (Figure 26A).

From these 110 RLG1 islands we selected the islands that contain the higher number of RLG1 elements, resulting in 15 islands that contain from 14 RLG1 elements to 30 RLG1 elements ranging in size from 40 kb to 160 kb (Table 9). The elimination of these 15 islands from the genome would result into the elimination of a 22% of the chromosome and a 35% of all the RLG1 islands of the chromosome (Figure 26B).

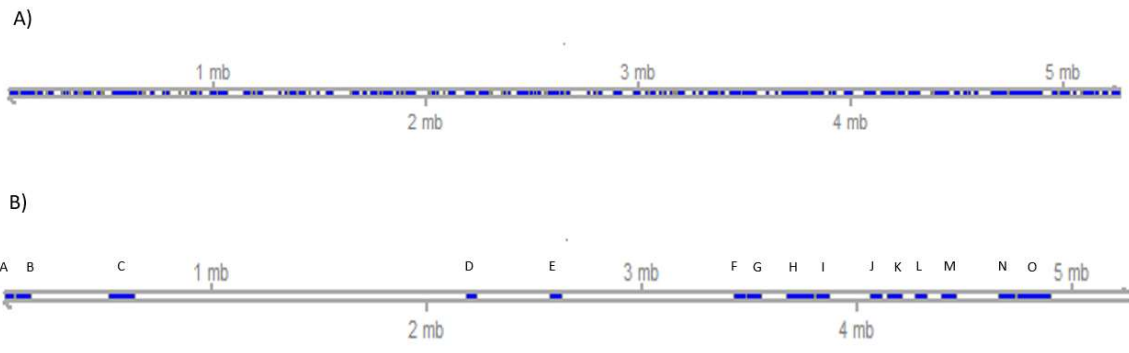


Figure 26:A) Chromosome 27 with all the RLG1 islands, heterochromatic regions in blue. B) In blue 15 biggest RLG1 islands of chromosome 27 ranging in size from 40 kb to 160 kb.

Table 9: Selected RLG1 islands from chromosome 27 with the position, number of RLG1 elements, bp length and the name given to the island. In grey the RLG1 islands selected to produce the first deletions.

Chromosome	Start position	End position	N° RLG1 Elements	Length (bp)	Name
Chr27	40521	80600	14	40079	iRLG1-A
Chr27	93408	162197	18	68789	iRLG1-B
Chr27	518269	640367	20	122098	iRLG1-C
Chr27	2182199	2231160	15	48961	iRLG1-D
Chr27	2567799	2629289	26	61490	iRLG1-E
Chr27	3424985	3481993	20	57008	iRLG1-F
Chr27	3486530	3556539	15	70009	iRLG1-G
Chr27	3669583	3800246	20	130663	iRLG1-H
Chr27	3805377	3872512	30	67135	iRLG1-I
Chr27	4056396	4115748	17	59352	iRLG1-J
Chr27	4135867	4212064	17	76197	iRLG1-K
Chr27	4268903	4323290	20	54387	iRLG1-L
Chr27	4389190	4463172	21	73982	iRLG1-M
Chr27	4659508	4738773	19	79265	iRLG1-N
Chr27	4741208	4900930	25	159722	iRLG1-O

Our goal in this case was to perform various transformations eliminating in each transformation two of these islands and check the effect of these eliminations at the phenotypic level and on the expression of the flanking genes.

From the 15 islands we decided to start from the biggest (Island O on Figure 26.B) with a total of 25 RLG1 elements and a genomic size of ~160kbp, and one of the smallest., The smallest one is island A, but it is located at the extreme of the chromosome and the elimination could affect the telomeric regions of the chromosome. Therefore, we decided to focus on island D that contains 15 RLG1 elements with a genomic size of 49 Kbp.

With the help of Amàlia Morató, a former undergraduate student in the laboratory, we performed a transformation using the 4 gRNAs targeting the flanking sequence of the two RLG1 islands, the CRISPR/Cas9 and a plasmid containing a NptII gene to transiently select the clones that have been transformed. We obtained a total of 211 clones that were transiently resistant to G418 (50mg/L).

To analyze the potential elimination of the selected islands, we designed primers at the unique sequences flanking the RLG1 islands to amplify the empty locus for both islands. We checked the clones for the amplification of a deletion, but we failed to amplify a band in any of those clones.

As we could not amplify the deletion from any of the analyzed clones, we designed a template to replace the RLG1 islands by an antibiotic resistant gene through homologous recombination which would facilitate the selection of the clones where the RLG1 islands had been eliminated (Figure 27).

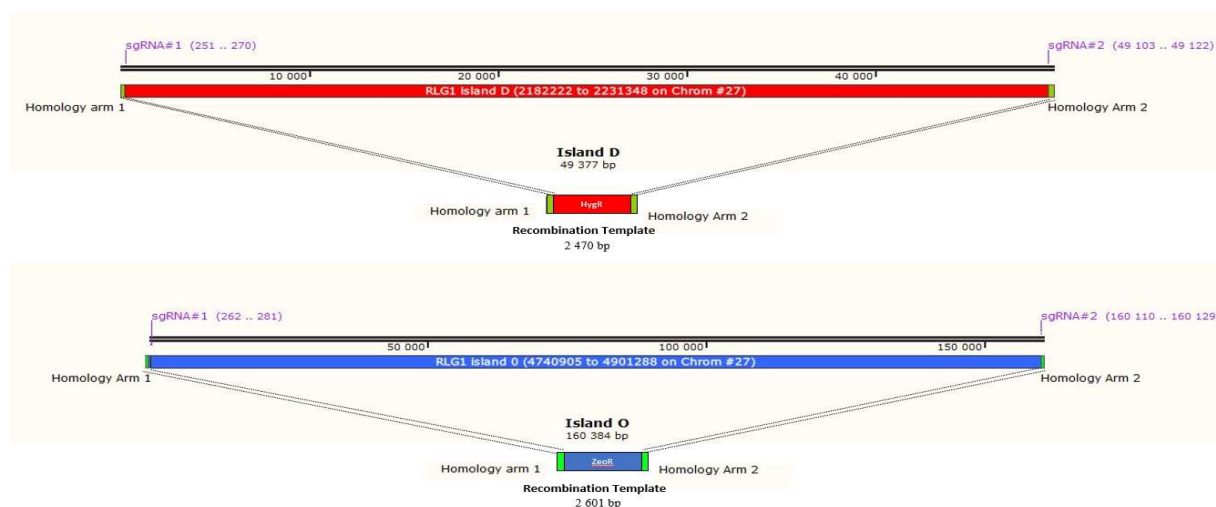


Figure 27: Replacement of Island D and Island O for two recombination templates containing a Hygromycin resistant gene cassette in the case of island D and a Zeocin resistant gene cassette in the case of Island O. Both recombination templates are flanked by 250 bp of homologous sequence to the gRNA cutting sides for their respective RLG1 islands.

Using these recombination templates Florence Charlot from Fabien Nogué group performed three independent transformations of *P. patens*: A first transformation with the gRNAs to eliminate the island D, the CRISPR/Cas9 system and the DNA template to replace this island for the HygR gene cassette; a second transformation with the gRNAs to eliminate the island O, the CRISPR/Cas9 system and the DNA template to replace this island for the ZeoR gene cassette; and a third transformation with the 4 gRNAs to replace both RLG1 islands, the CRISPR/Cas9 system and both recombination templates (Table 10).

Table 10: Results of the transformations done to replace the Island D, island O or both islands at the same time. At the first column the percentage of protoplast that regenerated after one week being similar for all the transformations done, the second column describe the antibiotic used to select the clones that have stably integrated the construct into their genome and the last column describe the number of clones that are stably resistant to the antibiotics in each case.

	% Regeneration After 1 week	Selection	Resistant Clones
CRISPR/CAS9 + gRNAs targeting Island D+ DNA template to replace Island D	19.03%	HygB 25 mg/L	92
CRISPR/CAS9 + gRNAs targeting Island O+ DNA template to replace Island O	19.86%	Zeo 100 mg/L	26
CRISPR/CAS9 + gRNAs targeting Island D and Island O+ DNA templates to replace both	20.00%	HygB 25 mg/L Zeo 100 mg/L	55

We genotyped the clones obtained by PCR using primers in the unique sequence flanking the RLG1 islands and internal primers inside the recombination templates.

We obtained 6 clones where we could amplify both flanking regions of the replacement for Island D from the 50 clones genotyped and 4 clones for Island O over the 26 clones obtained. However, we were not able to amplify any of the four locus corresponding to a replacement of both islands from the 55 clones resistant to both antibiotics.

We sequenced the PCR product for both flanks of the replacements of 3 independent clones of island D (clones 2,3 and 6) and 3 independent clones of island O (5,10 and 22) (Figure 28), which confirmed that the replacement took place in both extremes as expected in these clones.

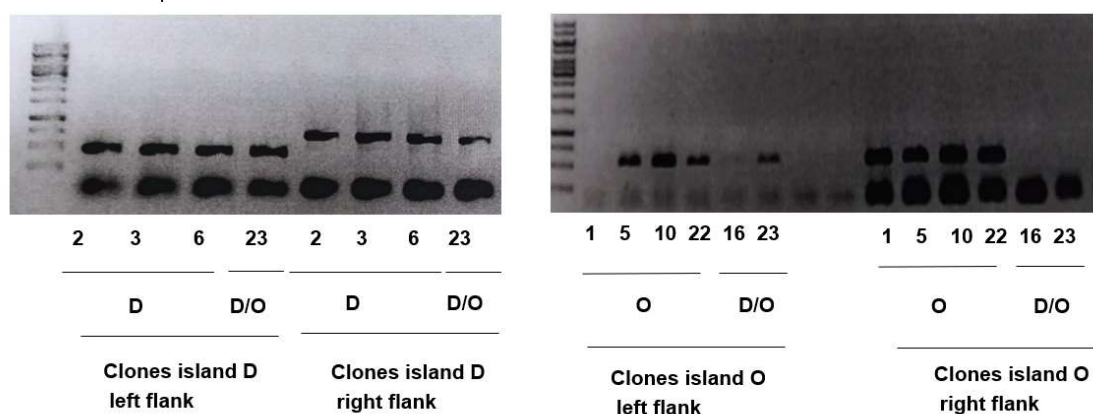


Figure 28: Selected clones of the Island D (D) and island O (O) and both replacements (D/O) selected for sanger sequencing. We observe than in the case of the islands selected for both replacements they always lack one of the extremes. We sequenced the clones 2,3 and 6 of island D and the clones 5,10 and 22 of island O,

However, we could not amplify in any of the clones the complete replaced region. To try to understand what happened in these we used the primers designed inside the recombination template to discard multiple integrations of the recombination template. To our surprise, in several clones we obtained bands of 3.5 kbp for both replacements (Figure 29). After sequencing the PCR product, we observed that these bands corresponded to the complete plasmids that contained the recombination template (the HygR cassette for island D and the ZeoR cassette for island O). The multiple integration of the complete plasmid at the replaced locus could explain the lack of amplification of a single replacement at the locus D and O.

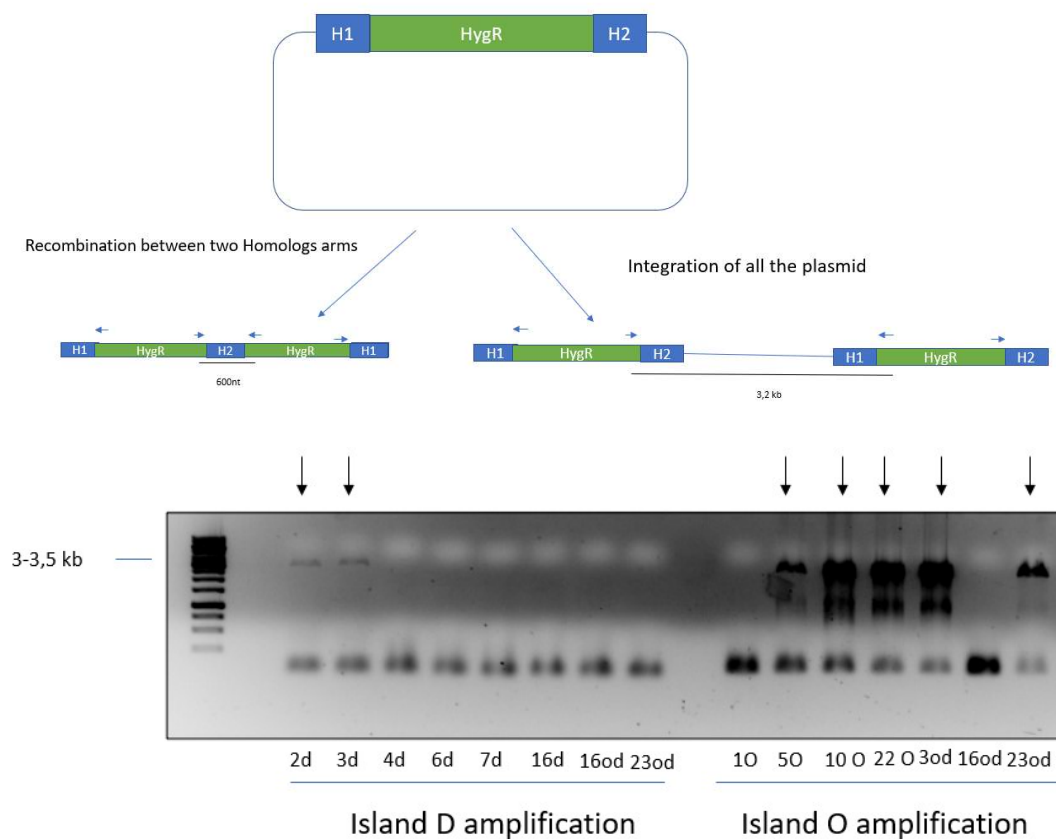


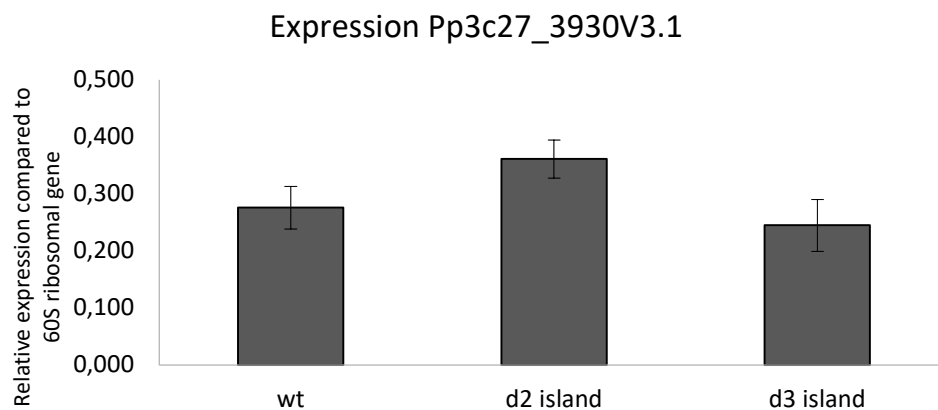
Figure 29: In top a scheme the two main hypothesis of integration of the constructs into the locus D. If multiple replacements have taken place of only the recombination place, we expect amplification of a band of 600bp while if there are multiple integrations of the plasmids, we expected bands of ~3.2Kb. At the bottom PCRs using internal primers to verify the presence of the constructs observing bands of around 3.5 Kbp corresponding to the plasmids after sequencing of the PCR product.

The sequence of the junctions of clones 2d, 3d, 5o and 10o, were as expected for a recombination event. Therefore, we decided to proceed with the analysis for these 4 clones, in spite of the possible complex integration of the plasmid containing the template. In none of these 4 clones we observed any obvious developmental phenotype.

Impact of the RLG1 islands replacement in their flanking genes

We then decided to analyze whether the elimination of the heterochromatic regions had an impact on the expression of the surrounding genes. To analyze their expression, we first looked under which conditions these genes are expressed. In the case of the genes flanking island O, the gene Pp3c27_7919V3.1 was detected as lowly expressed in the gene atlas and in the PEATmoss database (Fernandez-Pozo et al., 2020; P. F. Perroud et al., 2018) it was only expressed in imbibed spores, a tissue that is difficult to generate.. Similarly, the other gene flanking island O, the gene Pp3c27_8050V3.1 is also expressed at very low level in most of the development condition of *P. patens*.

For this reason, we decided to analyze the expression of the genes flanking the RLG1 island D, Pp3c27_3930V3.1 and Pp3c27_3970V3.1 as both genes were reported as expressed in protonema in the databases. We analyzed their expression in protonema after 7 days of growth (Figure 30).



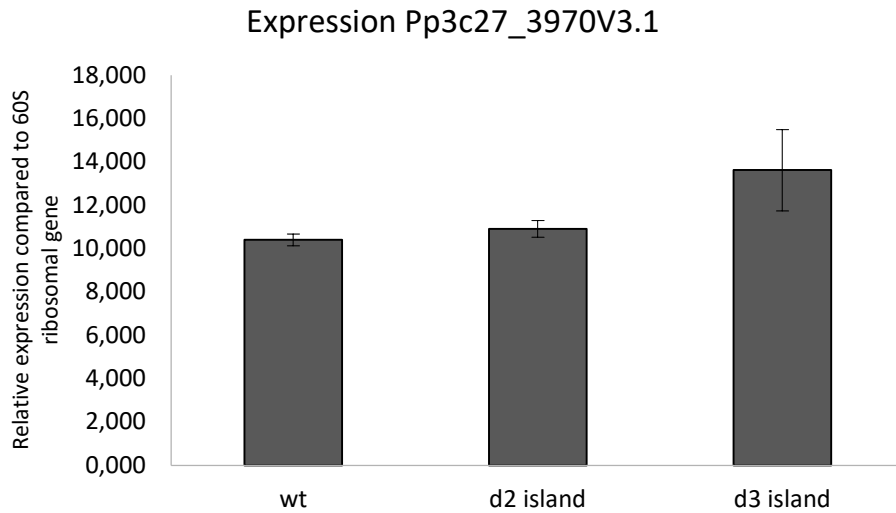


Figure 30: Relative expression of the genes Pp3c27_3930V3.1 and Pp3c27_3970V3.1 compared to the 60S ribosomal protein expression in protonemata of 7 days old. Both genes are flanked by the RLG1 island D of chromosome 27 that we have replaced by the recombination template in the clones d2 and d3, as a control we used Gransden wild type material.

In both clones and for both genes we saw a small effect on the expression when compared to the Gransden Wt. In the case of the expression of the gene Pp3c27_3930V3.1, the replacement seems to increase the expression for the clone d2 although in the clone d3 we did not observe any difference. While in the other flanking gene (Pp3c27_3970V3.1) there seems to be more expression in both clones when compared to Gransden Wt.

Discussion

In this chapter we described different experiments that aim at studying the impact of heterochromatic RLG1 elements on the genome of *P. patens*. We show that the non-selective elimination of RLG1 elements by producing targeted cuts at the LTRs of these elements has a strong deleterious effect for the host survival. The high number of cuts produced could have a lethality effect not directly related to the elimination of these TEs, but could be a direct effect of the high number of DSB introduced, similarly to what has been observed after treating the moss with mutagenic agents that cause random DSB, such as bleomycin (Holá et al., 2013). In addition, introducing DSB at repetitive RLG1 elements could also induce recombination and genome rearrangements or eliminations of essential genes which could explain its strong deleterious effect.

However, we obtained clones that were able to survive after targeting the RLG1 elements. Those clones could be the ones that had a mild to weak effect, and probably those that had less DSBs in the genome. When we checked the phenotype of these clones most of them did not have a particular developmental phenotype. Moreover, we did not detect a significant reduction of the number of RLG1 elements in the genome. The only exception was clone K6, where we did not observe a reduction but an increase of the number of RLG1 copies in the genome. Two possible scenarios could account for an increase in RLG1 elements. First, this could be due to an induction of RLG1 transposition by the DSBs. It has been shown that retrotransposon sequences can be used to repair DSBs (Ono et al., 2015) and it can be envisaged that DSBs induce transposition of some elements. Alternatively, the DNA repair of the deletions by homologous recombination could involve as a template RLG1 islands bigger than the one deleted leading to an overall increase of the number of RLG1 copies.

An analysis of these clones using long reads may clarify this. However, although these technologies are getting cheaper, it is still an expensive experiment. More important, to be able to analyze the results we would need a reference genome with a good assembly of the heterochromatic regions. The reference genome of *P. patens* (Lang et al., 2018) has a chromosome level assembly but is still far in terms of quality of genomes such as the

last genomes published of *A. thaliana* (B. Wang et al., 2021) or other plant species published during the recent years (Sun et al., 2021)

These heterochromatic regions are prone to be misassembled and to accumulate assembly errors. To properly compare the obtained results, we will need to probably produce a better reference genome to be able to detect these events. Even in the case that we resequence some of these lines we will depend on the need of high-quality reads with enough read length in order to be able to characterize these events in the case that they may have occurred.

Another aspect that we could analyze is to what extent the possible alteration of the structure of the chromosomes may have altered the fertility and the ability to undergo meiosis. The APT mutants are sterile so we cannot check the production of sporophytes on the clones that we have produced DSB at the same time over the APT and the RLG1 copies. To do that we should probably repeat the first experiment done transforming Gransden Wt and let the clones regenerate, perform the qPCR again and select a number of these clones to generate sporophytes. However, the production of sporophytes in Gransden is low (5-6% of the clones generate sporophytes while other accessions such as Reute more than 75% of the gametophores produce sporophytes (Hiss et al., 2017). This fact could difficult even more the analysis of the effect of targeting RLG1 elements.

It has recently been reported that targeting centromeric TEs in *Cryptococcus neoformans* results in chromosome shuffling, an increase of the number of chromosomes and formation of new telomeres and in most of the affected clones the strains were unable to undergo meiosis (Yadav et al., 2020). Also, the targeting with CRISPR/Cas9 of the HERV-W, LINE1 and Alu TEs lead to the production of non-viable human cell lines (Velasco, 2019)

Replacing specific RLG1 islands of the smallest chromosome a feasible but difficult approach

To understand the effect over the structure of the genome of the RLG1 families we designed an alternative approach to remove the biggest RLG1 islands of the chromosome 27 and study then the effect that this may have in the development and to the global capacity to undergo mitosis or meiosis, or over the stability of the smallest chromosome.

We thought that this more conservative approach would allow us to control and understand better the impact of the elimination of these heterochromatic islands. Unfortunately, we obtained clones where we only replaced one of the 15 selected RLG1 islands. In this case we replaced in some clones the Island D of 47 Kbp and in other clones the Island O of 159 Kbp, but we could not eliminate both islands at the same time.

Moreover, in the clones that we manage to amplify the flanking sequence of the replacement of the Island D or Island O for the recombination template we always failed to amplify a clean replacement, and all the data point to multiple integrations of the plasmids in the locus. These multiple integrations could also have an effect in the structure of the genome as the introduced transgenic locus contains the antibiotic resistant genes expressed under the control of a 35S promoter that can be targeted by the silencing machinery of the cell and may not be comparable to a clean replacement of the RLG1 islands.

We could not observe major changes in the development of the cells in the clones that we replaced the RLG1 regions. Although we observed slight changes in expression of the flanking genes of the RLG1 islands between the clones that we replaced the island D.

To further study the effect of elimination of these RLG1 islands we will need to replace several of these islands. With the approach used in this study although it may be feasible, in terms of time and genotyping is not the optimal approach as it will require several rounds of transformations, genotyping, and phenotyping to analyze the effect of these replacements that may require several years of work to eliminate all these regions.

We expect that some of the limitations that we have found when trying to remove these heterochromatic regions will be overpassed in the near future with the development of new technologies such as the paired prime editing (Choi et al., 2022) to produce precision genomic deletions. For the moment has been proved to produce targeted deletions of the human cell lines of up to 10 kb. We expected that in the near future similar systems that produce targeted deletions of bigger size will arise and will be adapted to plants. These strategies could be used to perform the targeted deletions of these RLG1 regions to then understand much better the impact of these eliminations and the role of these heterochromatic regions over the genome structure.

Chapter 3.4: Impact of Transposable Elements in genic regions

Introduction

In chapter 2 we reported the identification of several polymorphic TEs insertion between the different accessions of *P. patens*, most of them belonging to the RLG1 family, followed by the RLG3 and the RLC5 families. Most of these TIPs were located far from the genic regions, as expected for their global distribution. Despite that, a small fraction of these TIPs are located at the vicinity of genes, or even inside genes, and could have a direct impact on their expression. These changes could lead to phenotypic differences between the different accessions.

TE insertions can modify the expression of neighboring genes leading to phenotypical changes (an extensive list of TEs associated with plant phenotypes can be found in the review: Wei & Cao, 2016). Here we analyzed whether polymorphic TE insertions located close to genes lead to changes in gene expression between different accessions. To do that, we have selected different TIPs located close to genes that could lead to changes into the expression of the genes.

Results

In crops, such as rice, maize or tomato, that are of agricultural interest there is a high number of varieties with phenotypical data available (such as seed number, seed length or plant weight) and that have been resequenced (Alexandrov et al., 2015; Sauvage et al., 2014; Y. Xiao et al., 2017). In these plants it is possible to perform Genome Wide Association analysis (or GWAS) to make the link between genotype and phenotype. This strategy has been used to characterize mutations, that may or may not be causal, that are associated with different traits. This has been typically done using SNPs as genetic information (Spindel et al., 2016). However, these approaches have also been done using structural variants (Fuentes et al., 2019) and in the most recent years using TIPs (Carpentier et al., 2019; Domínguez et al., 2020). In the lab we have also used this approach to perform GWAS using TIPs in rice (Castanera et al., 2021).

In the case of *P. patens* the situation is diametral the opposite. There are only a few varieties that had been sequenced and their phenotypic information, regarding their differences in terms of development, for example, is less detailed than in crops. Therefore, it is much more difficult to try to link genotypic differences to changes in the phenotype. In chapter 2 we have identified several polymorphic TEs between the four accessions that had resequencing datasets available. Belonging to samples collected from the following locations: Gransden (United Kingdom), Villersexel (France), Reute (Germany) and Kaskaskia (United States of America) (Figure 31).



Figure 31: Distribution of the different *P. patens* accessions that have been resequenced using short-reads.

To maximize the chance to detect a TIPs in the different accessions that may have a phenotypical impact, we decided to increase the number of TIP candidates and perform an additional TIP search. To do that, we followed the strategies explained in the first chapter of this dissertation and that were published in Mobile DNA (Vendrell-Mir et al., 2019), which shows that the combination of PopoolationTE2 and Jitterbug increases the sensitivity to detect non-reference insertions. We therefore combined these two programs to detect non-reference insertions and we also used Pindel (Ye et al., 2009) to detect the possible absence in the different accessions of the TE insertions that are present in the reference genome. Finally, we filtered out the possible insertions that were supported with less than a 70% of the reads in the region because, as *P. patens* is haploid during most of the life cycle, we do not expect heterozygous insertions. We obtained 281 TIPs located at less than 1kb from a gene (Table 11).

Table 11: Number of Genes with a TIPs in the three accessions Reute, Kaskaskia and Villersexel as compared to the Gransden Reference genome.

	Genes with TIPs inside their gene body	Genes with TIPs at less than 1 kb from their gene body
Reute	10	44
Kaskaskia	19	111
Villersexel	37	178
TOTAL	58	281

Selection of TIPs potentially affecting gene expression

We manually curated the 281 TIPs detected close to genes or inside genes by checking individually the alignments of the reads to the reference genome. We discarded those cases that may correspond to misassembled regions of the genome or other structural variants.

We then prioritized the remaining genes based on the following criteria. First if the genes were single copy in *P. patens*. We expected that in the case of multicopy genes it would be much difficult to assess the impact of the transposon into the expression. Second, the presence of potentially homologous genes in other species. This could help in the study of the function of the genes by producing KO in other species or to use the previous knowledge in other species to understand the role that the gene may have in *P. patens*. Finally, the location of the TIP compared to the gene. We prioritized polymorphisms located inside the gene or located close to the 5' of the gene. Ending up with the list described in Table 12:

Table 12: Selected TE Polymorphisms insertions between the different accessions selected based on the previous criteria.

TE information					Gene Information					
Chr	Start	End	Accession	TE family	Chr	Start	End	Name	Distance TE/Gene (bp)	Gene Information
Chr04	16620254	16620520	Villersexel	RLG1	Chr04	16620254	16623417	Pp3c4_24710V3.1	0	AP2-TF;regulation of transcription,sequence-specific DNA binding transcription factor activity,Ethylene-responsive transcription factor
Chr04	17640810	17640993	Reute	RLC5	Chr04	17640435	17642026	Pp3c4_26220V3.1	0	Putative glucan endo-1,3-beta-glucosidase 3 precursor ((1->3)-beta-glucan endohydrolase) ((1->3)-beta-glucanase) (Beta-1,3-endoglucanase) (Beta-1,3-glucanase)
Chr17	3098590	3098774	Reute	RLG1, RLC5,RLG3	Chr17	3095057	3102736	Pp3c17_3870V3.1	0	Tetratricopeptide repeat protein 39C
Chr03	3480652	3481152	Villersexel	RLG1	Chr03	3481117	3482434	Pp3c3_5290V3.1	0	Encodes an alpha-tubulin isoform required for right handed helical growth, only expressed under dehydration/rehydration in <i>P. Patens</i>
Chr07	13111540	13112040	Villersexel	RLG3	Chr07	13108373	13111845	Pp3c7_19020V3.1	0	L-serine biosynthetic process:oxidation-reduction process NAD binding: phosphoglycerate dehydrogenase activity:amino acid binding, D-3-phosphoglycerate dehydrogenase
Chr07	5987636	5988136	Villersexel	RLC5	Chr07	5987968	5992059	Pp3c7_9190V3.1	0	Methyltransferase PMT10-Related,dehydration-responsive family protein
Chr11	11232822	11233322	Villersexel	RLG3	Chr11	11232455	11233651	Pp3c11_17080V3.1	0	DNA-dependent RNA polymerase II largest subunit
Chr14	1661147	1661647	Villersexel	tRLC5	Chr14	1659137	1662150	Pp3c14_2320V3.1	0	Unknown function
Chr14	1661147	1661647	Villersexel	tRLC5	Chr14	1661432	1662150	Pp3c14_2330V3.1	0	ODF3A_XENLA Outer dense fiber protein 3
Chr24	1027369	1027869	Villersexel	RLG1	Chr24	1025270	1028140	Pp3c24_1520V3.1	0	F14G6.8; expressed protein
Chr11	1179003	1179503	Kaskaskia	RLG1	Chr11	1178933	1179696	Pp3c11_2040V3.1	0	CASP-like protein UU2
Chr11	13931628	13932128	Kaskaskia	tRLC5	Chr11	13932110	13935805	Pp3c11_20830V3.1	0	deSI-like protein,phosphorylation ATP binding:protein serine/threonine kinase activity
Chr12	16260712	16261212	Kaskaskia	RLG1	Chr12	16258343	16260729	Pp3c12_24800V3.1	0	Encodes PIRL3, a member of the Plant Intracellular Ras-group-related LRRs (Leucine rich repeat proteins)
Chr14	5797944	5798444	Kaskaskia	RLG1	Chr14	5795415	5797978	Pp3c14_9040V3.1	221	Zinc-finger of C2H2 type (zf-met),histone-lysine N-methyltransferase SETD1B isoform X1
Chr16	8159332	8159832	Kaskaskia	RLG1	Chr16	8159217	8159391	Pp3c16_12810V3.1	0	ribulose bisphosphate carboxylase small chain c ribulose bisphosphate carboxylase small chain
Chr18	2013933	2014433	Kaskaskia	RLG1	Chr18	2014361	2016806	Pp3c18_2470V3.1	0	polyubiquitin containing 7 ubiquitin monomers:tetraubiquitin protein:polyubiquitin
Chr25	6965680	6966180	Kaskaskia	RLC5	Chr25	6962577	6966547	Pp3c25_9800V3.1	0	Probable LRR receptor-like serine/threonine-protein kinase RKF3-like
Chr04	7218652	7218662	Kaskaskia	RLG1	Chr04	7216339	7218898	Pp3c4_10010V3.1	0	Late Embryogenesis abundant hydroxyproline-rich glycoprotein
Chr07	3018411	3018421	Kaskaskia	RLG2	Chr07	3017696	3021410	Pp3c7_4770V3.1	0	Mitogen-Activated protein kinase 1-related

Chr10	16972477	16972487	Kaskaskia	RLG3	Chr10	16968202	16972671	Pp3c10_25170V3.1	0	F-actin capping protein, alpha subunit
Chr10	9609655	9609665	Villersexel	RLG3	Chr10	9609219	9620058	Pp3c10_14060V3.1	0	transmembrane transport voltage-gated chloride channel activity integral component of membrane
Chr17	5559790	5559800	Kaskaskia	RLG1	Chr17	5558207	5565909	Pp3c17_7260V3.1	0	tRNA dihydrouridine synthesis:oxidation-reduction process flavin adenine dinucleotide binding:tRNA dihydrouridine synthase activity
Chr24	1027599	1027609	Villersexel	RLG1	Chr24	1025270	1028140	Pp3c24_1520V3.1	0	F14G6.8; expressed protein
Chr24	10582473	10582483	Villersexel	RLG3	Chr24	10582205	10583638	Pp3c24_16070V3.1	0	Potassium inward rectifier (KIR)-Like channel 3 related
Chr25	1895694	1895704	Villersexel	RLC5	Chr25	1895437	1900638	Pp3c25_2840V3.1	0	probable arabinosyltransferase ARAD2
Chr06	2183479	2183489	Reute	Unknown LTR-RT	Chr06	2181880	2183441	Pp3c6_3920V3.1	39	dephosphorylation:trehalose biosynthetic process trehalose- phosphatase activity
Chr07	701728	702228	Villersexel	RLG1	Chr07	700011	701683	Pp3c7_1110V3.1	46	Unknown function
Chr10	9837955	9837965	Villersexel	RLG1	Chr10	9835527	9837878	Pp3c10_14440V3.1	78	signal peptidase complex subunit 1 (SPCS1)
Chr01	18218091	18218101	Villersexel	RLG1	Chr01	18208996	18218010	Pp3c1_25070V3.1	82	pleiotropic drug resistance 3 (PDR3)
Chr20	14912741	14913241	Villersexel	tRLC5	Chr20	14913422	14916551	Pp3c20_22950V3.1	182	Protein folding:response to high light intensity:response to endoplasmic reticulum stress:response to hydrogen peroxide:response to hea
Chr11	10397162	10397662	Reute	RLG3	Chr11	10397868	10400759	Pp3c11_15720V3.1	207	RNA polymerase II second largest subunit
Chr02	1185244	1185744	Kaskaskia	RLG1	Chr02	1178938	1185019	Pp3c2_1770V3.1	226	tubulin-specific chaperone c-related tubulin-specific chaperone c- related
Chr13	5436794	5437294	Kaskaskia	RLG1	Chr13	5433680	5436490	Pp3c13_8060V3.1	305	charged multivesicular body protein 2a
Chr05	3329091	3329591	Villersexel	RLG1	Chr05	3320271	3328744	Pp3c5_4990V3.1	348	RAD5_ASPFU DNA repair protein
Chr04	17688741	17689241	Villersexel	RLG1	Chr04	17684478	17688289	Pp3c4_26240	453	ABC transporter I family member 17 gi

From the previous list, we selected three potentially interesting cases based on the described criteria: the gene Pp3c17_3870V3.1, Pp3c4_24710V3.1 and Pp3c14_9040V3.1.

Analysis of the potential effect of the TIP in the gene Pp3c17_3870V3.1

We identified a polymorphic insertion of a TE inside the eighth exon of the gene Pp3c17_3870V3.1 (Figure 32). This polymorphism was only detected as present in the accession Reute and absent in all the other accessions.

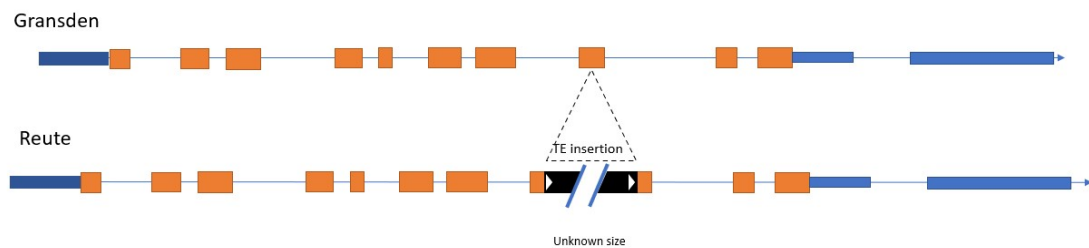


Figure 32: Polymorphic TE insertion predicted in Reute inside the eighth exon of the gene Pp3c17_3870V3.1 compared to the gene structure in Gransden. In blue boxes, the 5' and 3'UTR sequences, in orange, boxes the exons. The arrow indicates the orientation of the gene.

The gene Pp3c17_3870V3.1 encodes for a gene similar to the human gene TTC39C or the Tetratricopeptide repeat protein 39C. This gene is conserved in most vertebrates, has been also found in invertebrates, although less conserved in structure, and has also been detected in the fungi. Regarding the plant kingdom this gene has not been found in seed plants but after a search done for this study we identified genes encoding similar proteins in other Bryophytes, green algae's and in Ferns. This gene has an unknown function in humans but it has been hypothesized that it may be involved in the control of the anaphase during the cell cycle by interacting with chaperones (Blatch & Lässle, 1999).

In the moss the gene is highly expressed specially in the archegonia and during the early formation of the sporophytes according to the transcriptome atlas (Ortiz-Ramírez et al., 2016) (Figure 33).

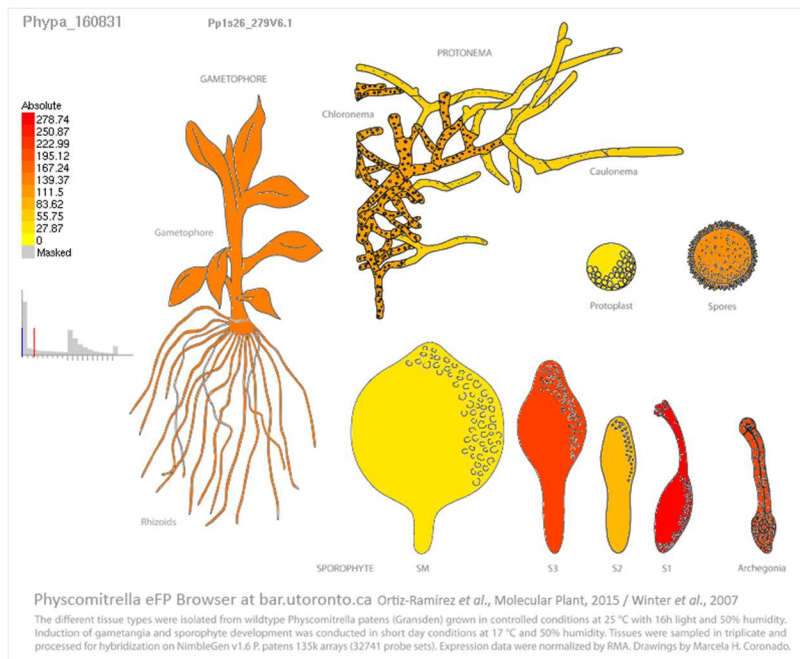


Figure 33: Expression of the gene Pp3c17_3870V3.1 during the development of *P. patens* image extracted from the transcriptome atlas (Ortiz-Ramírez *et al.*, 2016) that can be accessed through the following webpage http://bar.utoronto.ca/efp_physcomitrella/cgi-bin/efpWeb.cgi.

To check if the polymorphism may have an impact in the gene expression we looked at the public libraries available through PEATMOSS (Fernandez-Pozo *et al.*, 2020). We confirmed that the gene is expressed in Gransden and in Reute (the only accessions for which there is expression data in PEATMOSS). Despite this, after aligning reads of Gransden and Reute RNA libraries to the transcript sequence we observed that the transcript in Reute is truncated at the eighth exon (Figure 34) and no reads were observed at the 3'CDS after the insertion when compared to Gransden.

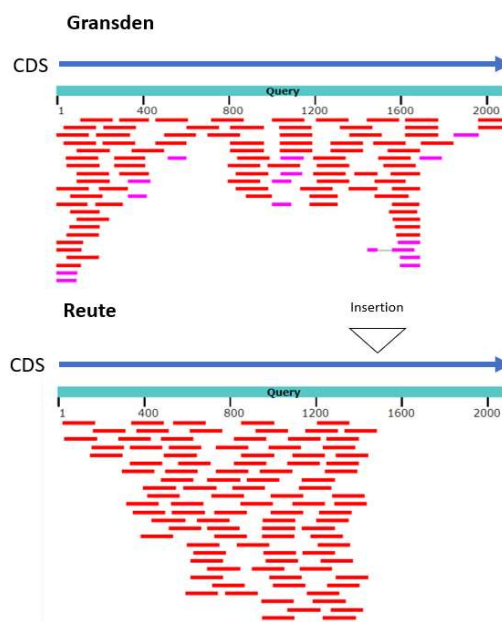


Figure 34: Alignment of RNA-seq short reads to the gene in Gransden (upper part) and Reute (bottom part). We observe the absence of reads at the end of the 3'CDS in reute while in Gransden we detect reads covering all the transcript.

The group of Fabien Nogu  performed a KO of the gene in Gransden by base editing (Guyon-Debast et al., 2021). Unfortunately, the clones obtained did not have any development difference when compared to the Gransden Wt lines.

The insertion predicted by Jitterbug did not point out to a single family of transposons, instead, there were reads assigned to three different families RLG1, RLG3 and RLC5. To confirm the presence of the polymorphism we designed primers flanking the polymorphic side to amplify the transposon insertion. We could amplify the Gransden Wt band, but we could not amplify any PCR product in Reute. All of this suggest that the predicted insertion could be a structural variation involving multiple integrations of TEs or a genome rearrangement.

Due to the lack of a development phenotype in the lines that had the gene truncated compared to the Wt samples and the complex nature of the TIP we decided to focus on the other genes potentially affected by a TE polymorphism.

Analysis of the potential effect of the TIP in the 3' UTR of the gene Pp3c4_24710V3.1

This TE insertion polymorphism was only detected as present in the accession Villersexel and absent in all the other accessions. This TIP consists in an insertion of a an RLG1 element at the end of the 3' UTR of the gene (Figure 35).

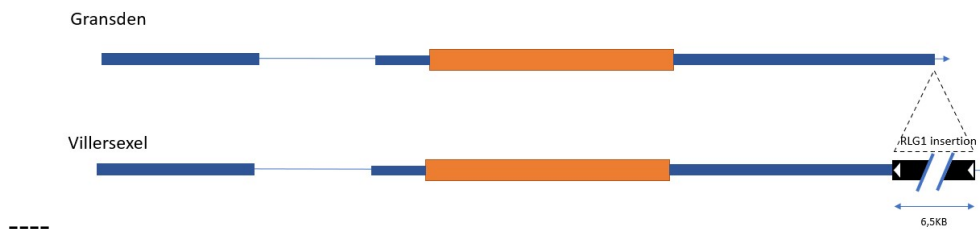


Figure 35: Polymorphic TE insertion predicted in Villersexel at the end of the 3' UTR of the gene Pp3c4_24710V3.1 compared to the gene structure in Gransden- In blue boxes, the 5' and 3'UTR sequences, in orange boxes, the exons. The arrow indicates the orientation of the gene.

The gene Pp3c4_24710V3.1 encodes an AP2 transcription factor, homolog of the gene RAP2.1 in *A. thaliana*. Mutants of this gene in *A. thaliana* showed an improved tolerance to drought and to cold stress (Dong & Liu, 2010). In *P. patens* this gene has been detected as differentially expressed under several stresses such as UV-B radiation (Wolf et al., 2010), cold stress (Tan et al., 2017) and dehydration stress (Arif et al., 2019).

During the development of the moss this gene is highly expressed in the protonema, especially in the caulonema tissue and in the rhizoids of the plant, not being expressed in the other development conditions (Ortiz-Ramírez et al., 2016) (Figure 36).

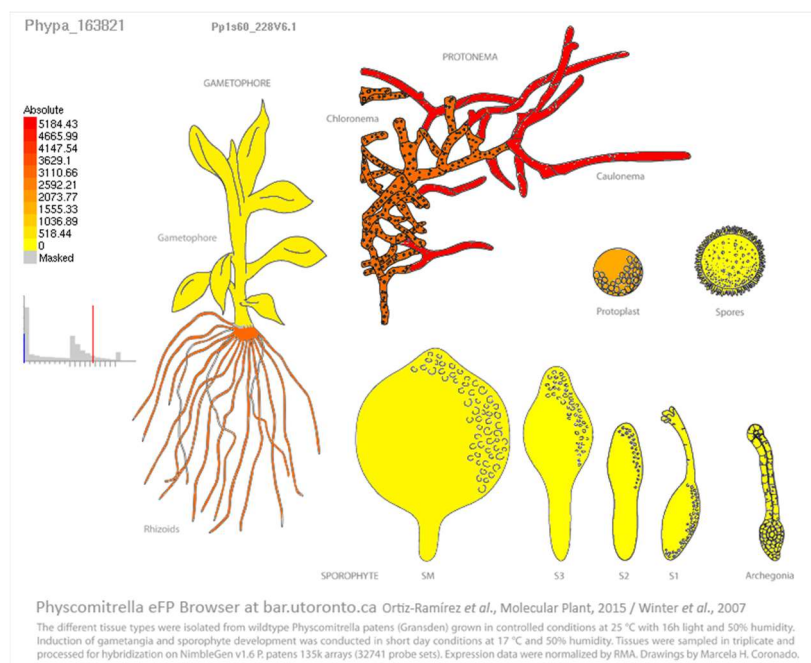


Figure 36: Expression of the gene Pp3c4_24710V3.1 during the development of *P. patens*. Observing a high expression in protonema, especially in caulonema, the rhizoids and in the protoplast. Image extracted from the transcriptome atlas (Ortiz-Ramírez et al., 2016) that can be accessed through the following webpage http://bar.utoronto.ca/efp_physcomitrella/cgi-bin/efpWeb.cgi.

From the recently published gene atlas and through the PEATMOSS database (Fernandez-Pozo et al., 2020; Perroud et al., 2018) we observed a similar profile of expression, detecting the gene differentially expressed under the same development conditions, under ABA treatment and UV-B treatment.

To confirm the presence of the insertion we designed primers to amplify by PCR the locus both in Gransden and Villersexel. We expected a band of ~300bp in Gransden and a band

of ~6.8 Kbp in Villersexel if it corresponded to a RLG1 insertion. We could confirm by PCR (Figure 37) and sequencing of the PCR products that an RLG1 element is inserted at the end of the 3' UTR in opposite transcriptional orientation with respect to the gene Pp3c4_24710V3.1.

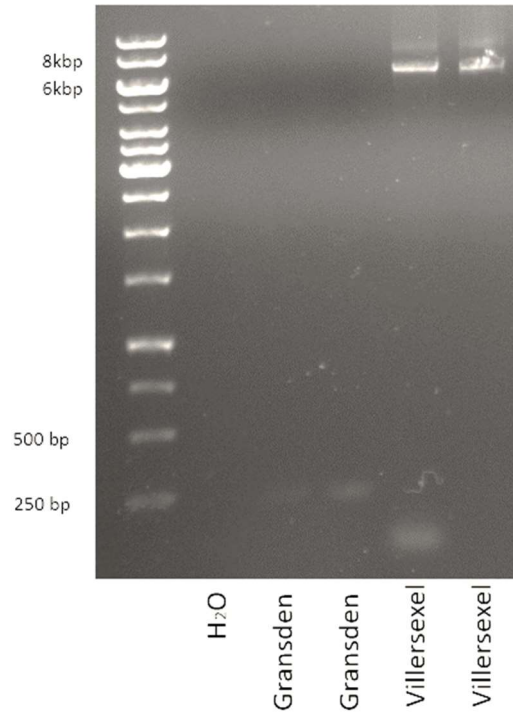


Figure 37: Verification of the RLG1 insertion in the Pp3c7_24710V3.1 by PCR amplification, observing a band corresponding to the empty locus in Gransden (~300bp) and a band corresponding to a RLG1 insertion in Villersexel (between 6 and 8 Kbp).

We checked whether the gene is differentially expressed between Gransden and Villersexel by performing a qRT PCR on cDNA of protonemata of 7 days old and of gametophores. The gene is more expressed in Gransden when compared to Villersexel and is only expressed in protonemata (Figure 38).

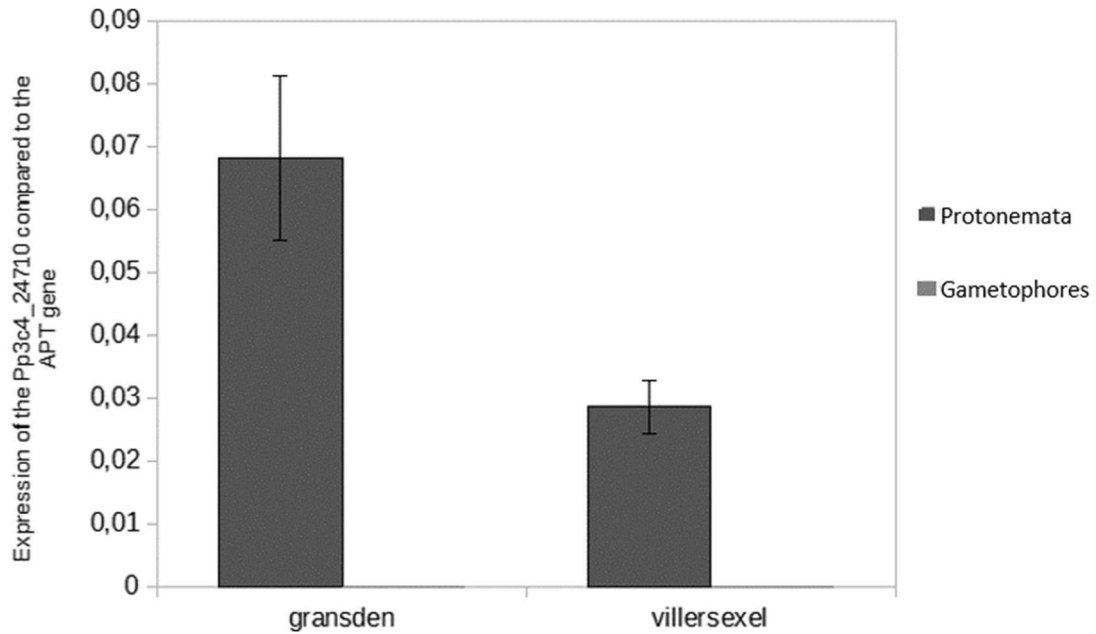


Figure 38: Relative expression of the gene Pp3c4_24710V3.1 between Gransden and Villersexel compared to the housekeeping gene APT in protonemata and in gametophores.

As Gransden and Villersexel are the accessions that are more genetically distant according to the number of SNPs and TEs (Lang et al., 2018; Vendrell-Mir et al., 2020) and that both accessions have huge development differences, we decided to perform a comparison in protoplasts, which should be a more comparable tissue. Also, in this case the gene is more expressed in Gransden than in Villersexel (Figure 39).

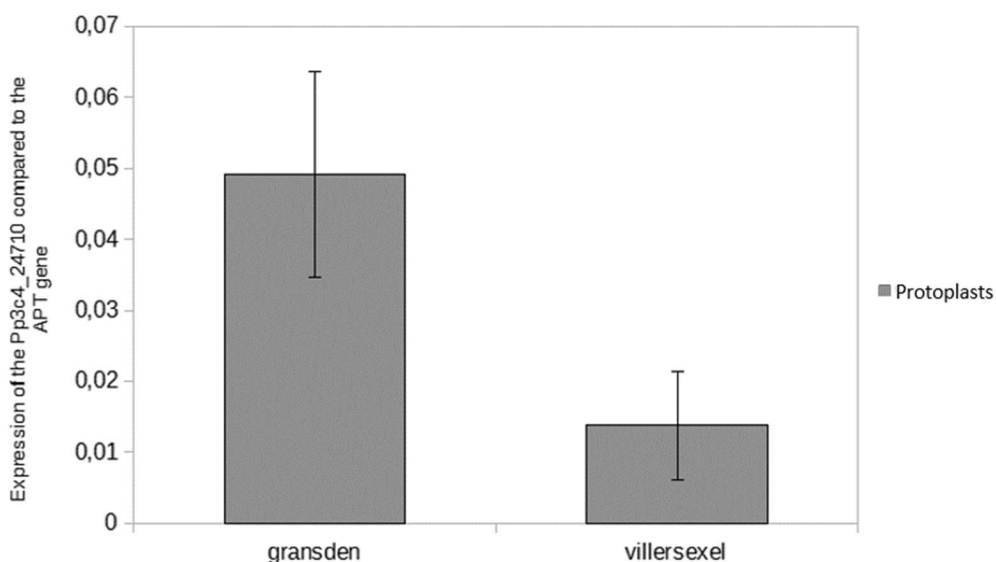


Figure 39: Relative expression of the gene Pp3c4_24710V3.1 between Gransden and Villersexel compared to the housekeeping gene APT in protoplasts.

For this reason, we performed a KO of the gene targeting the ATG of the coding sequence and the end of the 3' UTR to produce a deletion of the gene by CRISPR/Cas9. After performing the transformation and analyzing over 50 clones we obtained three independent KO of the gene that had small indels interrupting the ATG starting codon.

We performed a first phenotyping by growing the clones for 30 days in the medium BCD, but after this period we could not observe any difference in terms of phenotype between Gransden and the clones that had a KO of the gene, although there were clear differences between Gransden and Villersexel (which contains the RLG1 insertion) in terms of production of rhizoids (Figure 40).

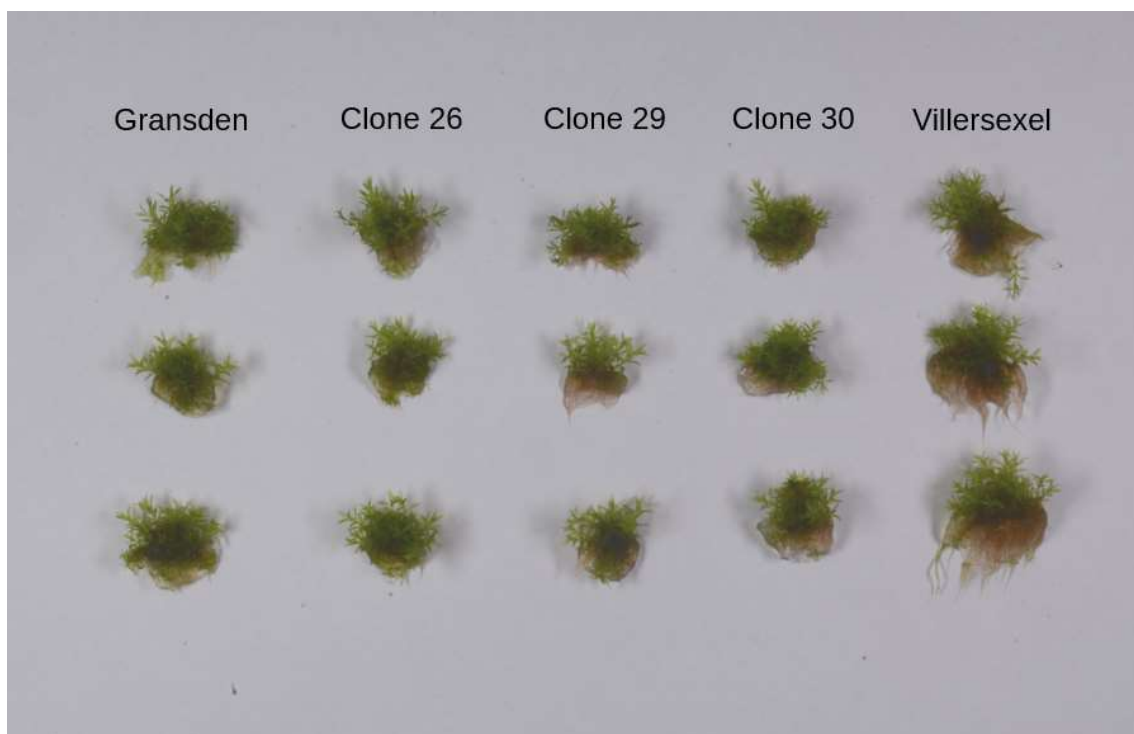


Figure 40: Independent colonies of Gransden, Pp3c4_24710V3.1 (Mutants lines 26, 29 and 30) and Villersexel after 30 days of growth in BCD.

Moreover, Florence Charlot from the group of Fabien Nogué performed a phenotyping experiment of the clones under different concentrations of ABA. No obvious differences were seen between the different mutated samples and the Wt Gransden clones, although differences were observed between Gransden and Villersexel under these conditions (Figure 41).



Figure 41: Growth differences between Gransden Wt and Villersexel Wt in BCD medium under different concentrations of ABA (0 µg/L, 1 µg/L and 2 µg/L).

Due to the lack of phenotypical differences between Gransden and the KO clones we decided to focus our attention on the last selected gene from the three selected genes affected by a TIP between the different accessions.

Analysis of the potential effect of the TIP in the promoter region of the essential gene Pp3c14_9040V3.1

The last analyzed gene that has a TIP in their neighboring region is the gene Pp3c14_9040V3.1, at 229 bp of the 5' UTR and at 400bp of the CDS. This TIP was only detected in the accession Kaskaskia and absent in the other accessions including Gransden. The TIP predicted corresponded to the family RLG1 (Figure 42).

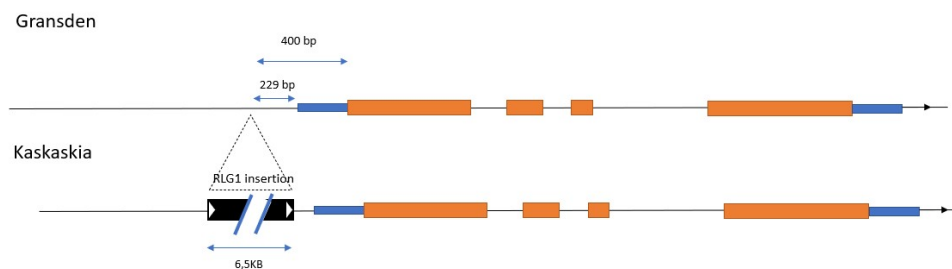


Figure 42: Polymorphic RLG1 insertion predicted in Kaskaskia 229 bp of the 5' UTR of the gene Pp3c14_9040V3.1 and at 400 bp of the CDS sequence, compared to the gene structure in Gransden. In blue boxes, the 5' and 3'UTR sequences, in orange boxes, the exons. The arrow indicates the orientation of the gene.

The gene Pp3c14_9040V3.1 encodes for a protein that contains two conserved domains: a Zinc finger of C2H2 type and a DNA topoisomerase 2-like protein. This gene is highly conserved in the plant kingdom (Figure 43), although its function remains unknown.

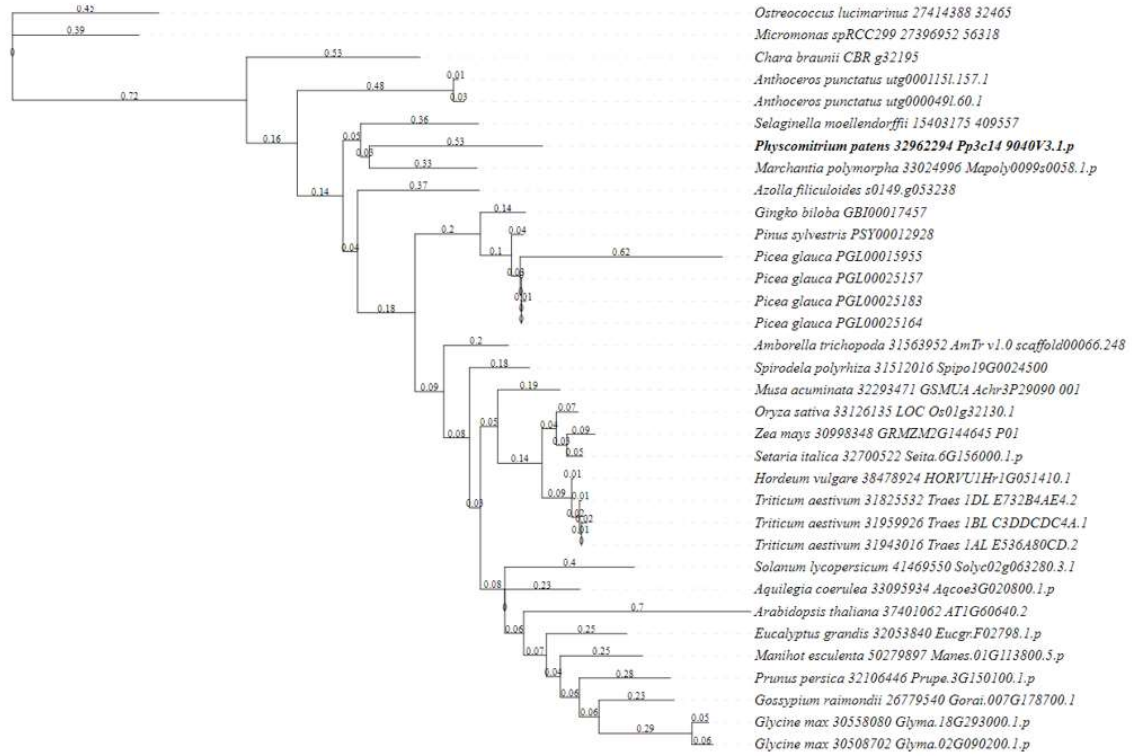


Figure 43: Phylogenetic tree of the representative sequences of representative sequences of all the plant kingdom. Representative sequence obtained from shoot.bio (Emms & Kelly, 2022). The obtained sequence were aligned using MAFFT (Katoh & Standley, 2013) and a phylogenetic tree was built using iqtree (Nguyen et al., 2015).

The gene is highly expressed in spores (Ortiz-Ramírez et al., 2016), although it could be also detected as expressed in most of the tissues (Figure 44).

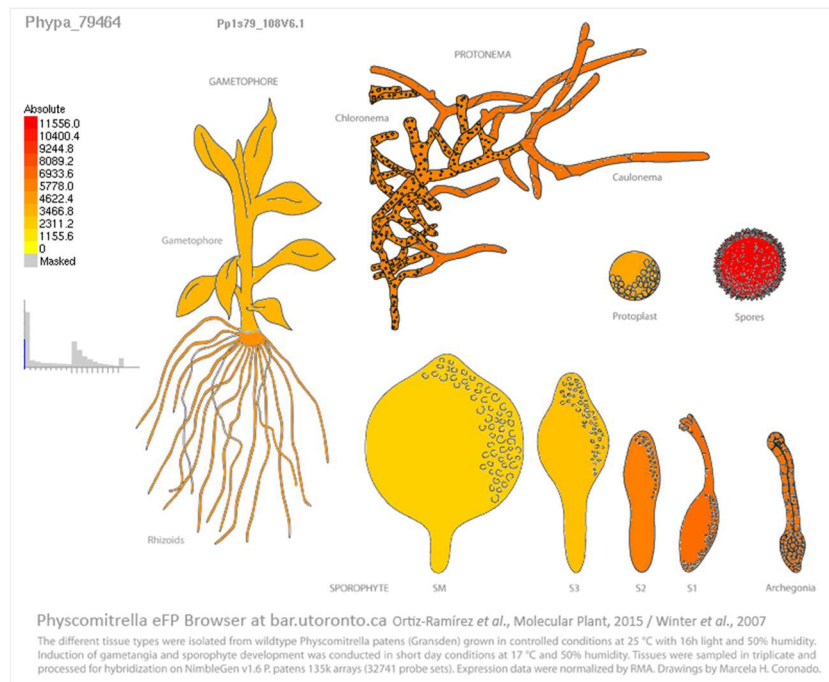


Figure 44: Expression of the gene Pp3c14_9040V3.1 during the development of *P. patens*. Observing a high expression in the spores. Image extracted from the transcriptome atlas (Ortiz-Ramírez et al., 2016) that can be accessed through the following webpage http://bar.utoronto.ca/efp_physcomitrella/cgi-bin/efpWeb.cgi.

A similar result was obtained when looking at the development database deposited in PEATMOSS (Fernandez-Pozo et al., 2020), where the highest expression was detected in imbibed spores. Moreover, the gene is repressed under heat stress and under different light conditions such as in darkness or under red or far-red light.

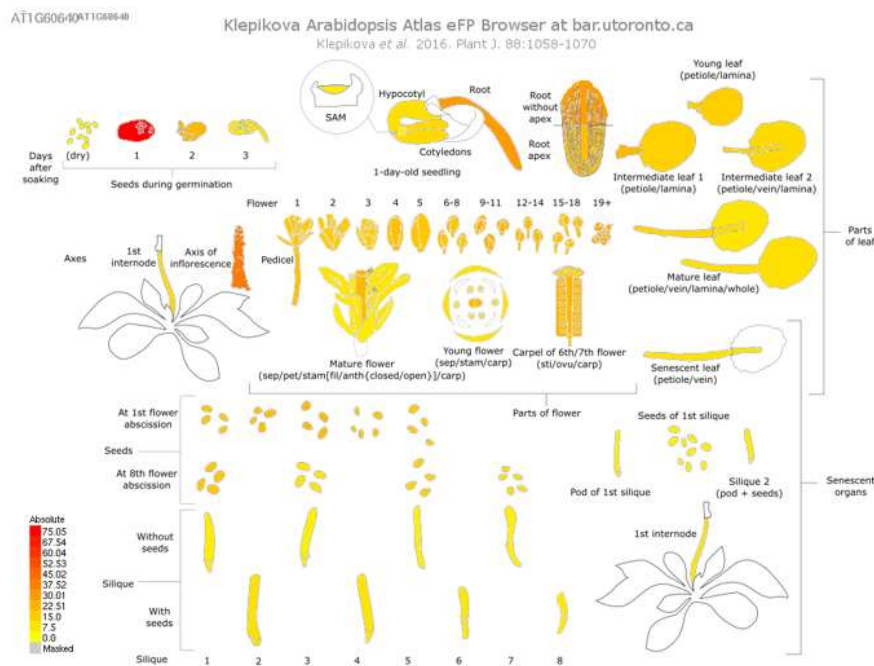
Five genes are coexpressed in the same development conditions with a correlation value of a 97%.: Pp3c1_3590V3.1, Pp3c18_14790V3.1, Pp3c21_1450V3.1, Pp3c2_9490V3.1 and the gene Pp3c16_20170V3.1

Pp3c1_3590V3.1 encodes the BOP1 ribosome biogenesis protein. When the miRNA controlling this gene are knocked out, and this gene is overexpressed, there is an early transition from protonemata to gametophores (Saleh et al., 2011). However, a KO of this gene has no phenotype (Hata et al., 2019). The gene Pp3c18_14790V3.1 also encodes for a Ribosome Biogenesis protein that is homolog to other BOP genes in other plants. The genes Pp3c21_1450V3.1 and Pp3c2_9490 encode for a RIO1 kinase homologs protein. RIO1 in yeast encodes for a cytoplasmic non-ribosomal protein that is required for the processing of the 20S pre ribosomal RNA to the 18S ribosomal RNA in the pre-40S

ribosomal subunit (Vanrobays et al., 2001). Although their function is not known in plants a KO of this gene in yeast is lethal (Vanrobays et al., 2001). Finally, the gene Pp3c16_20170V3.1 encodes for a putative transcription factor containing a Zinc finger CCHC domain. The homolog *Arabidopsis thaliana* gene encodes for the gene AT5G52380, a transcription factor of unknown function.

Most of these genes are involved in the development control of the plants being expressed in *A.thaliana* in the seeds shortly after imbibition during the germination (Klepikova et al., 2016). Similarly, we detected that the homolog gene of Pp3c14_9040V3.1 in *Arabidopsis thaliana* (AT1G60640), is highly expressed during the germination of the seed shortly after imbibition (Klepikova et al., 2016), a similar pattern was observed for the homologous gene in *Medicago truncatula* (Benedito et al., 2008) (Figure 45).

Arabidopsis thaliana



Medicago truncatula

Plant eFP: Medtr7g082850

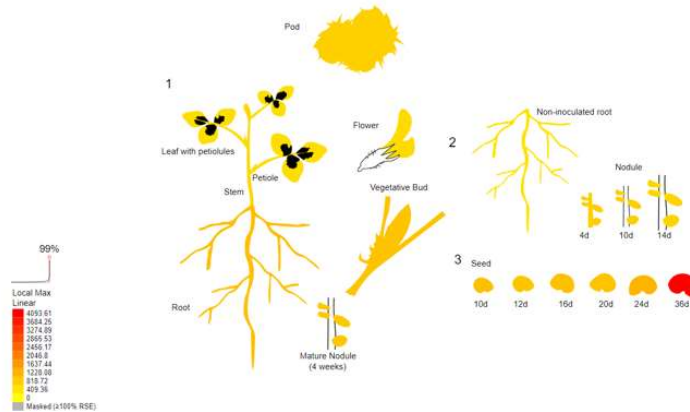


Figure 45: Gene expression over the development of *Arabidopsis thaliana* (Klepikova et al., 2016) and *Medicago truncatula* (Benedito et al., 2008) homolog genes of Pp3c14_9040V3.1, observing a high expression during the germination of the seed in both species.

All this data suggests that the gene Pp3c14_9040V3.1 may be a transcription factor involved in the process of germination of the spores in *P. patens*, that may play a role in cell cycle regulation and/or cell differentiation.

We confirmed the presence of the polymorphism in Kaskaskia and the absence in Gransden by designing primers flanking the TIP predicted side and amplifying by PCR the region, confirming the presence of an RLG1 insertion at the expected place in Kaskaskia (Figure 46).

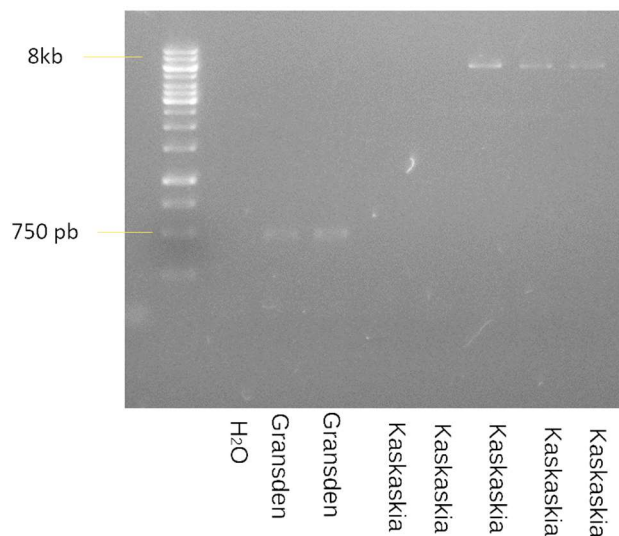


Figure 46: PCR amplification of the locus corresponding to the polymorphism in Gransden and in Kaskaskia. If it corresponds to Gransden, we expect a band of 760 bp while in Kaskaskia as there the insertion of a RLG1 LTR-RT we expect a band of 7kbp. The lanes correspond to independent DNA extractions of Gransden Wt and Kaskaskia Wt. In some Kaskakia Wt extractions we could not amplify the locus.

We sequenced the PCR amplification products of Gransden and Kaskaskia, confirming the presence of an RLG1 insertion in Kaskaskia, in the same orientation than the gene and located at 229 bp of the 5' UTR and at 400bp of the start of the CDS sequence.

As a first approach to study if the polymorphism could have an impact on gene expression, we performed qRT PCRs during the development of protonemata to study if the gene is differentially expressed. We quantified the expression of the gene in protonemata tissue at the days 4, 6, 8 and 10 of freshly generated protonemata, observing the gene is more expressed in Kaskaskia (Figure 47).

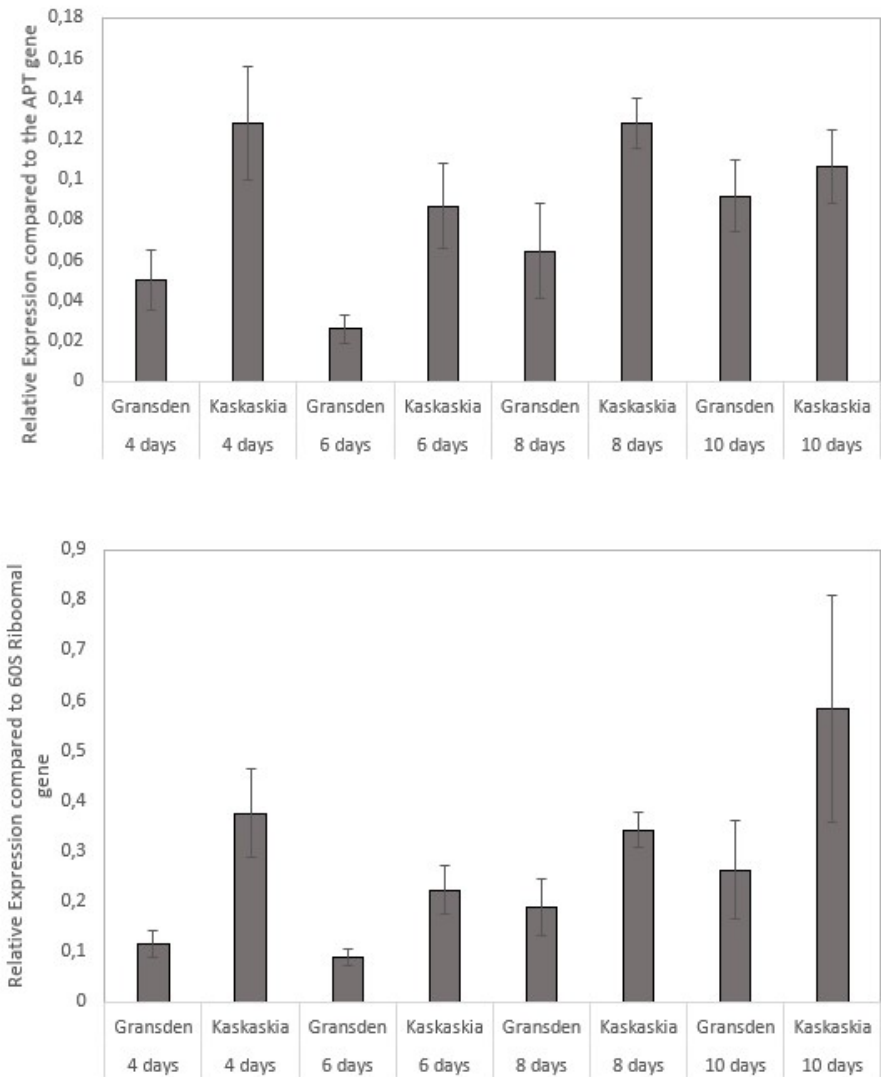


Figure 47: Expression of the gene Pp3c14_9040V3.1 when compared to the housekeeping APT (Top) and the 60S ribosomal protein (Bottom).

On the other hand, the group of Fabien Nogu  (IJPB-INRAE) designed gRNAs to target the start codon of the CDS to produce a KO of the gene using a base editing approach, and to produce deletions using CRISPR/Cas9.

Among the base edited clones obtained 77% had no mutations in the coding region whereas a 23% had C to T or C to G changes within the coding sequence but, in all cases, the ORF was conserved without being interrupted by a stop codon (data can be accessed at Guyon-Debast et al., 2021).

Among the 39 clones obtained in the CRISPR/Cas9 approach, 9 of them did not have any change in their sequence, 19 clones had a deletion that in all cases was a multiple of 3 nucleotides without generating a stop codon, 4 had a deletion multiple of 3 nucleotides and an nucleotide substitution that did not generate a premature stop codon, 5 clones had nucleotide substitutions that did not generate a premature stop codon and finally two clones had an insertion multiple of three nucleotides preserving the ORF (Table 13).

Table 13: Type of mutation for the 39 sequenced clones of the CRISPR/Cas9 transformation targeting the Pp3c14_9040V3.1 gene. In the third and sixth column, length of the nucleotide insertion or deletion and/or sequence that has been substituted.

clone	type	Length ins/del & Sequence	Clone	type	Length ins/del & Sequence
2	deletion	12 nt	28	substitution	AT to GA
3	deletion	3 nt	29	No change	
6	deletion	12 nt	30	substitution	TCGATC to CGGACA
7	No change		31	deletion	12 nt
8	No change		32	No change	
11	No change		33	deletion	48 nt
13	substitution	A to C	34	deletion	6 nt
14	deletion	9 nt	37	substitution	T to A
15	substitution	TTCGT to CAATG	41	deletion+ substitution	12 +A/G
16	deletion	9 nt	43	No change	
17	deletion+ substitution	12 nt + T to C	44	deletion+ substitution	3 nt +T to C substitution
18	deletion	12 nt	45	deletion	9 nt
19	deletion	9 nt	46	deletion	9 nt
20	deletion	12 nt	47	insertion	3 nt
21	deletion+ substitution	15 nt + A to C	48	deletion	12 nt
22	deletion	12 nt	50	No change	
24	insertion	6 nt	51	No change	
25	deletion	15 nt	52	No change	
26	deletion	21 nt	55	deletion	12 nt
27	deletion	18 nt			

In summary, none of the clones obtained using the two different approaches was a KO of the gene. The fact that all the mutations are multiple of three and are not generating a premature stop codon in the coding region suggests that a KO of this gene in *P. patens* is lethal. According to our collaborators, the clones that they manage to introduce small indels had a strong development phenotype.

In order to study the potential impact of the RLG1 insertion on the expression of this gene, we designed a strategy based on CRISPR/Cas9 and homologous recombination to swap the gene upstream regions of Gransden and Kaskaskia (which contains the RLG1 insertion). We used the CRISPR/Cpf1 tool instead of the CRISPR/Cas9 classical system as the promoter region had a low GC content which make it difficult to define guide RNAs sequences inside this region. The CRISPR/Cpf1 system track RNAs, named cRNAs, have a protospacer adjacent motif that is much common in low GC regions such as 5'-TTTN-3' being possible for us to design the cRNAs in this region.

Therefore, we designed cRNAs to cut in the regions flanking the RLG1 element and used as templates to repair in Kaskaskia a DNA containing the sequence of Gransden (without the transposon), and in Gransden a DNA containing the RLG1 transposon of Kaskaskia flanked by two homologous arms matching the sequencing flanking the transposon (see Figure 48).

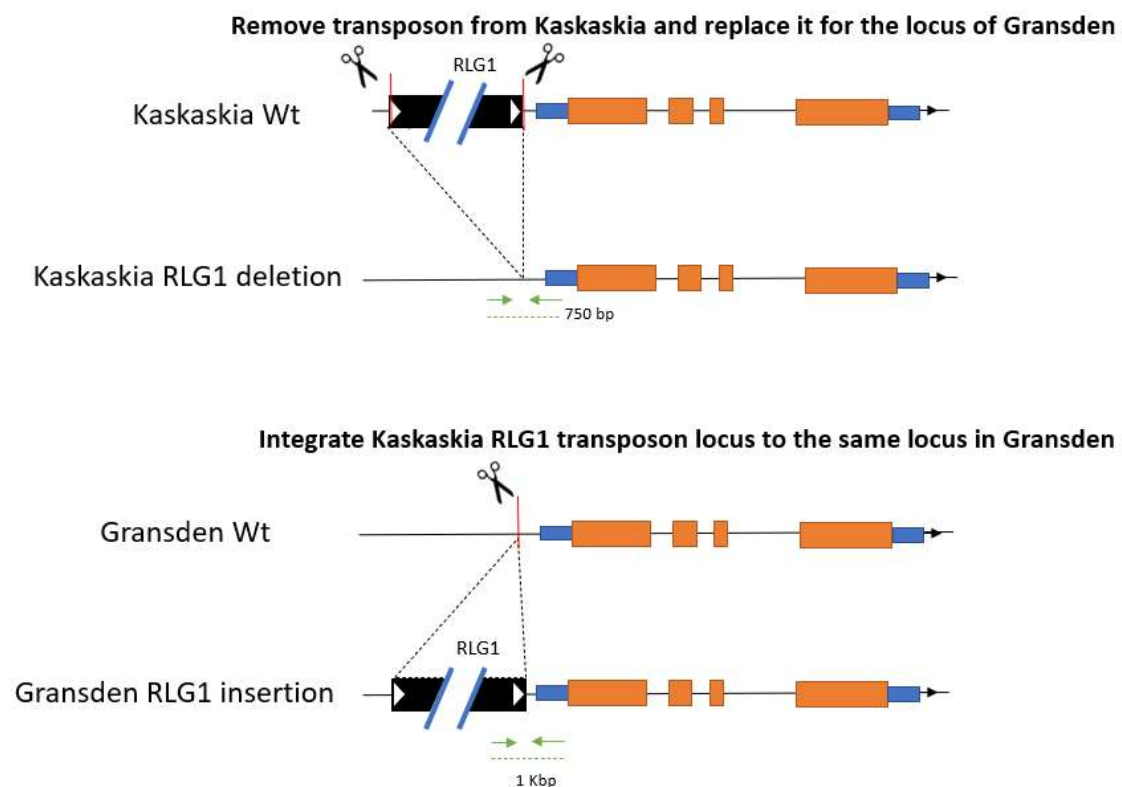


Figure 48: Scheme of the replacement strategy used to swap the RLG1 TE between Gransden and Kaskaskia. In red and accompanied by a scissors, the expected cut sides of the CRISPR/Cpf1 system, in black, the TE, and in orange and blue, the gene Pp3c14_9040V3.1 In green, a representation of the primers used to genotype the replacement after producing the deletion we expect a product amplification of 750 bp and in Gransden if we manage to integrate the TE an integration of 1 Kbp.

After performing the transformation of Gransden and Kaskaskia and selecting transiently for the presence of the plasmids in the transformation, we genotyped by PCR using an internal primer of the RLG1 and a unique primer flanking the RLG1 insertion to check for the presence of the transposon in Gransden and two primers flanking the transposon in Kaskaskia to amplify the absence of the transposon in Kaskaskia.

We genotyped 80 clones of each transformation and obtained 5 clones in Gransden that putatively contained the RLG1 element and 8 clones in Kaskaskia probably corresponding to the deletion of the RLG1 element. In Kaskaskia, we obtained 34 clones that had a smaller size that could correspond to a removal of the transposon but not to the replacement for the expected Gransden locus.

Over the 5 clones that amplified a band corresponding to the RLG1 insertion in Gransden, we repeated the PCRs using primers flanking the RLG1 insertion to amplify the complete RLG1 element, and in two cases we confirmed the presence of a complete RLG1 (Figure 49).

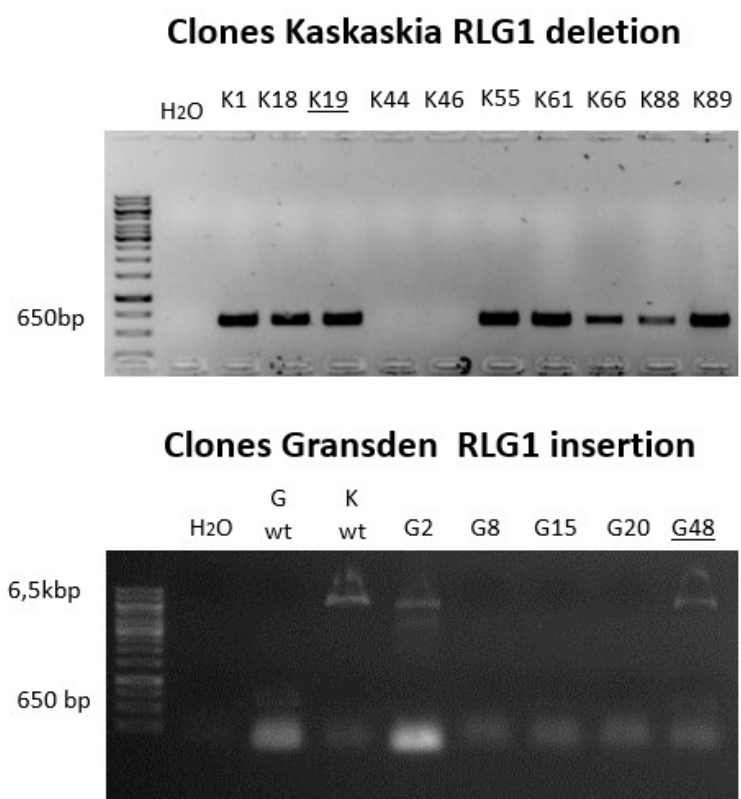


Figure 49: Verification of the RLG1 replacement in the 7 obtained clones of Kaskaskia (top) and the insertion of the TE at the expected locus in transformed clones of Gransden (bottom). Underlined the clones that we selected for analyzing the impact of the replacement.

The sequencing of the PCR products confirmed the expected sequence for an insertion of the RLG1 element at the locus. However, although we could amplify the complete RLG1 insertion we failed to amplify the region using primers located outside the recombination template.

As we have observed in the previous part of the chapter, when inserting sequences using CRISPR/Cas9 there is the possibility to integrate multiple copies of the template. To analyze if this had happen we performed a PCR to check the presence of multiple integrations by designing primers inside the recombination template but in opposite direction to the RLG1 insertion (primers oxPV129 - oxPV130 in the previous graph) that should result in an amplification product in case multiple copies of the RLG1 template have integrated. When using this approach, we amplified a band in the clones G2, G8, G15, G20, but nothing in Gransden Wt, Kaskaskia Wt and in the clones G48 (Figure 50). For this reason, we selected the clone G48 for sequencing and named as Gransden RLG1 insertion (G ins) for the posterior analysis.

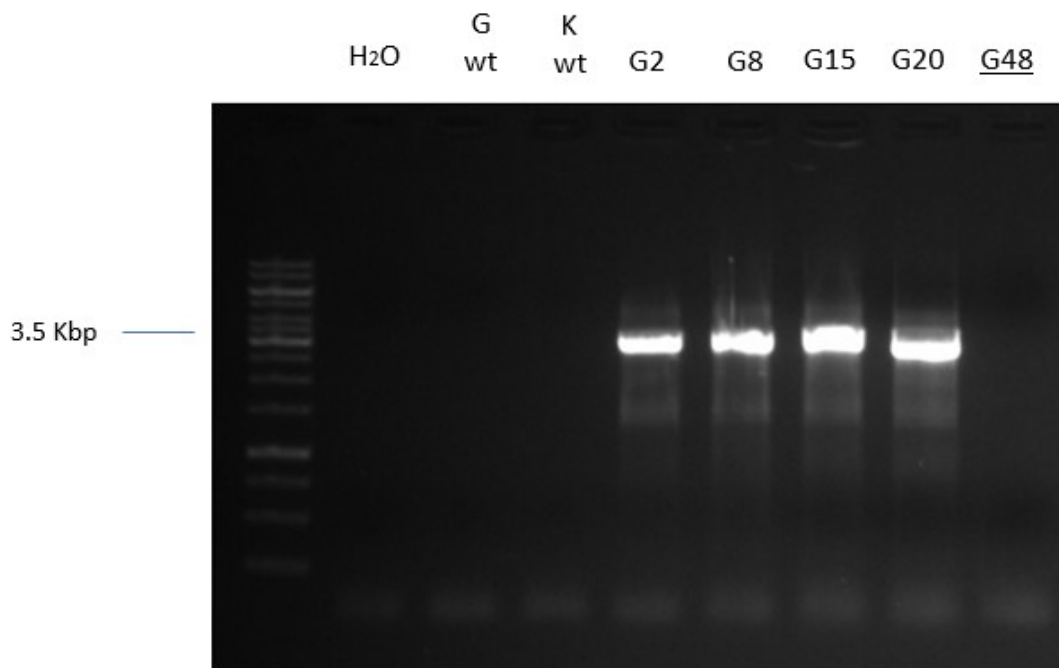
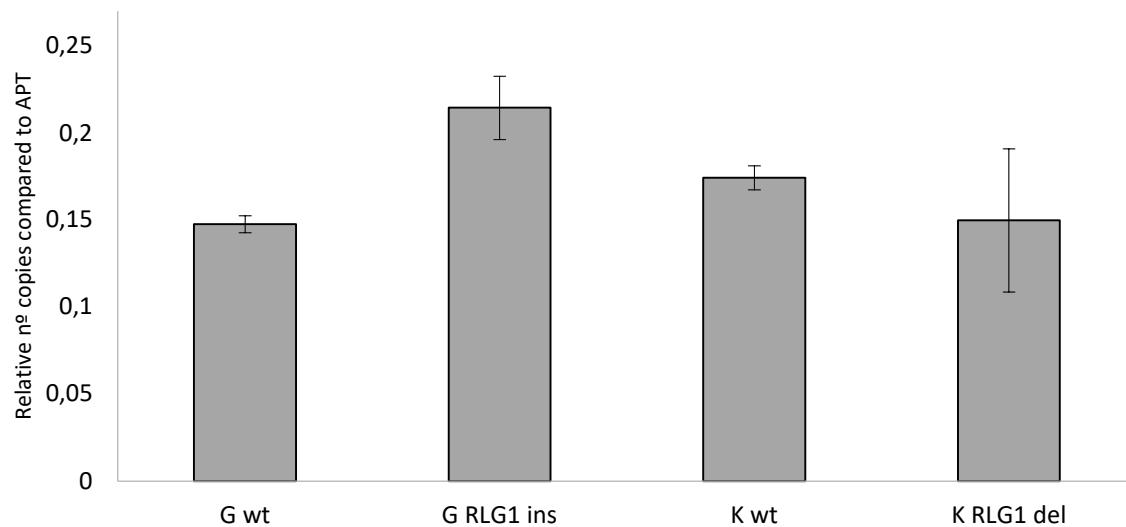


Figure 50: Verification by PCR of the presence of integration of multiple copies by using the combination of primers oxPV129 – oxPV130. The observed bands of around 3.5 Kbp have the exact size expected if it amplifies part of the plasmid that contains the RLG1 template used to integrate it in the genome, being an indication than in the clones G2, G8,G 15 and G20 there are integrations of the plasmid in the genomic DNA. Underlined the cloned used for the posterior analysis.

We sequenced the PCR products of three clones of *Kaskaskia* RLG1 deletion (K1, K18 and K19) and selected K19 to perform the next analysis as this clone has the predicted sequence of the expected replacement, while the other had some SNPs not found either in Gransden or in *Kaskaskia*.

The clones G48 and K19, together with Gransden and *Kaskaskia* wt, were grown in parallel and we performed first a qRT-PCR using three biological replicates of protonemata grown for 7 days in the medium BCDA. This analysis showed that the presence of the RLG1 element seem to induce the expression of Pp3c14_9040V3.1 in Gransden. However, the presence of the RLG1 element did not induce a significant change in the expression in *Kaskaskia*. There was a lot of variation between the different biological replicates in the *Kaskaskia* samples that could explain the obtained result for these samples (Figure 51).



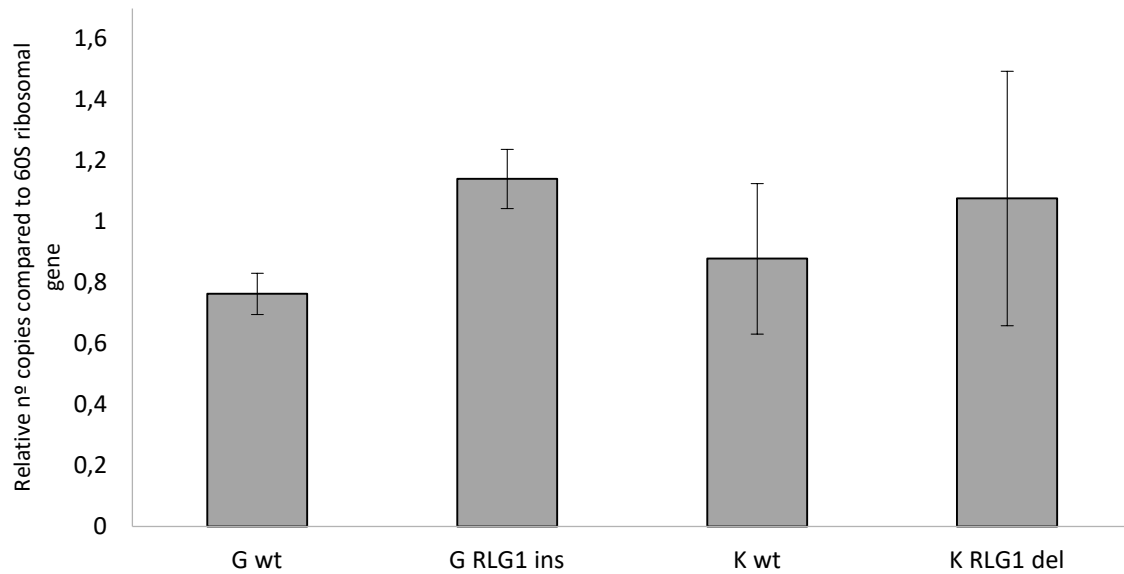


Figure 51: Relative expression of the gene Pp3c14_9040V3.1 compared to the housekeeping genes APT (top) and 60S Ribosomal Protein (bottom) of protonemata 7 days old.

To check whether the polymorphic RLG1 insertion could influence the expression during the development of the plant, we performed a time course over 14 days of development of the protonemata taking samples at the day 4, 7, 10 and 14. We performed this experiment in the lab with the help of Marc Pulido, a student in the laboratory, and Svitlana Sushko, undergraduate student under my supervision. I performed the protonemata manipulation while Marc and Svitlana performed the RNA extractions and the qRT PCRs. At the time that this manuscript was written, only one biological replicate had been analyzed. We observed that, an effect of the presence of the RLG1 insertion in both Gransden and Kaskaskia, although the effect does not seem to be consistent among samples and accessions (Figure 52).

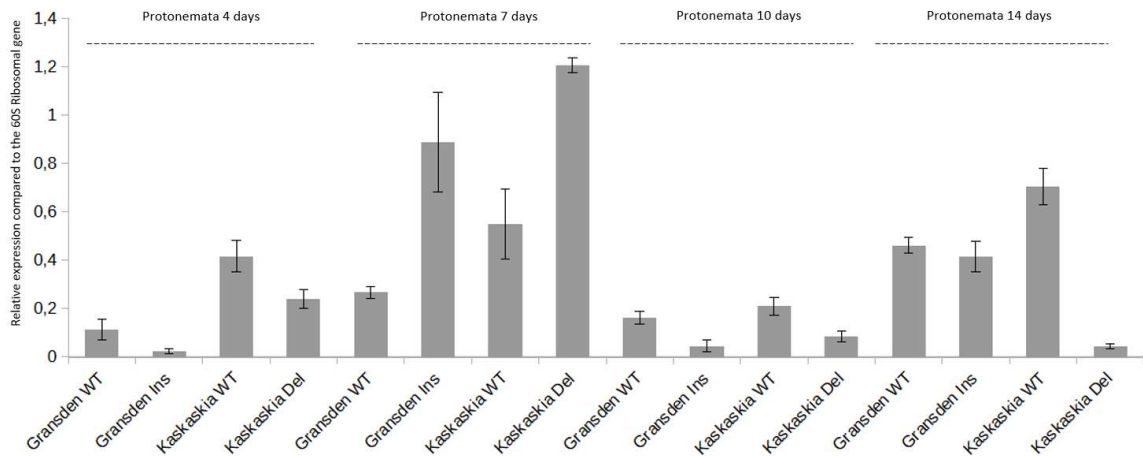


Figure 52: Relative expression of the Pp3c14_9040V3.1 gene when compared to the housekeeping gene 60S Ribosomal Protein, during 14 days of development of the protonemata,

We have also started to analyze the possible effect of the presence of the RLG1 insertion on the phenotype, which is not straightforward. There are obvious differences in terms of development between Gransden and Kaskaskia. While Gransden generates much more gametophores during the development, Kaskaskia generates less gametophores and much more protonemata (Figure 53).

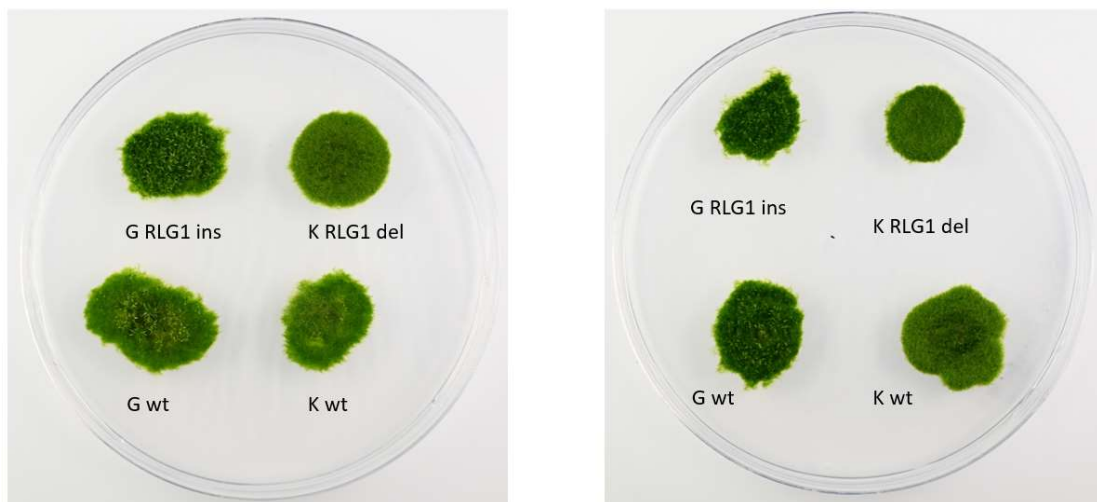


Figure 53: Images of the different clones after 21 days of growth in BCDA medium (ammonium rich medium).

There seem to be slight differences between the Gransden and Kaskaskia wt clones and the clones where we performed the replacement. In a first analysis done by Florence Charlot, from Fabien Nogué group, it was observed an increase of the apical growth of the protonemata in the Kaskaskia without the RLG1 insertion as compared to Kaskaskia wt

under BCD medium (medium lacking ammonium). A deeper analysis will be required to conclude what can be the impact of the replacement and the role of the gene in the development of *P. patens*.

Discussion

Analyzing the impact of TE insertions on genes is not straightforward. The vast majority of the cases described in the literature started with the detection of a clear phenotype which, after a molecular analysis turned out to be caused by a TE insertion. In this work we attempted the reverse approach (e.g. looking for the possible phenotypic impact of a characterized TE insertion) and selected three cases of genes that could potentially be affected by polymorphic TE insertions.

In the first case, the gene Pp3c17_3870V3.1, we observed a clear impact of a RLG1 insertion on its expression. The insertion of one or several TEs into a gene resulted in a truncated transcript. However, this insertion was complex, involving several TE elements or chromosomal rearrangements.

In the two other selected genes (Pp3c4_24710V3.1 and Pp3c14_9040V3.1) we could confirm the presence of a polymorphic RLG1 insertion, and in both cases, we detected differences of expression in the two selected genes between the different accessions, that could potentially be caused by the polymorphic RLG1 insertion. However, in both cases we could not detect a clear phenotype linked to the presence of the RLG1 insertion.

This was particularly striking for Pp3c14_9040V3.1, as the different approaches to generate mutants, which only generated weak alleles but no knockouts, suggested an essential function for the gene.

The approach of swapping a transposon at the exact same place between different individuals, up to our knowledge, has not been previously done in multicellular eukaryotes. This is a very powerful approach to study the possible impact of the insertion. Here, by exchanging the RLG1 transposon between the different accessions we showed that the presence of the transposon lead to changes in the expression of the Pp3c14_9040V3.1 gene both in Gransden, when integrating the transposon, and in Kaskaskia, when eliminating the transposon. Despite this, when integrating the transposon in Gransden does not lead to a similar expression pattern to the one observed in Kaskaskia. In the same way, when removing the transposon from Kaskaskia does not lead to a similar expression than in Gransden. We should consider that there are other genes that could potentially regulate the network controlling the expression of the gene

and that there could be differences in these genes between Kaskaskia and Gransden that could explain the differences observed. There could also be epigenetic changes in the affected locus between Gransden and Kaskaskia accessions that are not recovered after doing the swapping in these samples, such as methylation or the chromatin status.

As said, at the moment that this manuscript is being written, we had not managed to observe obvious phenotypical differences between the swapped locus clones compared to their respective Wt lines. To try to clarify if there are any phenotypical differences the members of the group of Fabien Nogué: Pierre-François Perroud and Florence Charlot are growing the different clones under different mediums to determine whether these clones have phenotypical differences when compared to the Wt lines or not.

Moreover, they have also generated spores of the different clones to try to determine if there is a difference in the germination of the spores, as the gene is highly expressed during the process of germination of the spores (Fernandez-Pozo et al., 2020; Ortiz-Ramírez et al., 2016). We could also observe there if there are differences in the development of the progeny when compared to the paternal lines, as there could be epigenetic changes after the generation of the progeny over the RLG1 polymorphic locus.

Julie Calbry another member of the group of Fabien Nogué is performing the KO of the homologous gene in *A. thaliana* which could help us understand what the role of this transcription factor is.

On the other hand, although the RLG1 insertion does not have an obvious phenotypical effect in the moss grown in the laboratory it could have an impact on nature important for the adaptation of the different individuals to the environment.

CHAPTER 4: AN ENDOGENOUS
VIRUS IN THE MOSS
PHYSCOMITRIUM PATENS

CHAPTER 4: AN ENDOGENOUS VIRUS IN THE MOSS *PHYSCOMITRIUM PATENS*

Chapter 4.1: Introduction

Mobile genetic elements can be found in virtually all organisms. Viruses are a particular type of mobile genetic elements that, as opposed to TEs, can move from cell to cell and from organism to organism. They are described as infective agents with small genomes that can only complete their life cycle within a living host cell.

Thus, viruses are mainly propagated through horizontal transmission, whereas TEs are essentially transmitted vertically from the parent to the progeny. Despite that, some viruses are also transmitted vertically (Martin et al., 2011), and transposons can also be transmitted horizontally between individuals, as it has been shown, for example for the recent invasion of the *Drosophila* species by the P-element (Kelleher, 2016).

From a structural and mechanistic point of view, some TEs and some viruses are extremely similar. In fact, the LTR-RT *copia* transposons have been classified as *Pseudoviridae* viruses and the LTR-RT *gypsy* transposons have been classified as the *Metaviridae* virus. Similarly, Retroviruses have also been classified as a superfamily of Retrotransposons in TE classifications (Wicker et al., 2009).

Both Retroviruses and LTR-RTs share a common ancestor, although their origin has not been resolved as it is possible that Retroviruses originated from a LTR-RTs that captured an envelope protein, or that LTR-RTs evolved from Retroviruses that lost the envelope protein (Hayward, 2017). Probably both processes may have occurred multiple times during the evolutionary processes.. This phenomenon has not only been observed between LTR-RTs and Retroviruses. For example, *Caulomiviridae* are closely related to LTR-RTs but are double stranded DNA viruses that do not need to integrate into the genome to replicate (Krupovic & Koonin, 2017) .

On the other hand, genomes contain integrated sequences that are remnants of past viral infections, such as the Endogenous Pararetrovirus sequences of plant genomes, although in this case it has been suggested that they may be active viruses that use the genome as a reservoir (Chabannes & Iskra-Caruana, 2013). As we can see, the limits between what we can define as a transposon and a virus may be somehow blurry.

As we have introduced in the general introduction, the study of transposons and their impact and dynamics in plant genomes has been mostly developed in angiosperms, especially in crop species and the model species *Arabidopsis thaliana*, with much lesser knowledge in other species such as bryophytes. A similar situation can be found for the study of plant viruses, where most of the studies are done in crops, mostly due to their economical relevance.

In the last recent years there have been efforts to expand our knowledge on viruses infecting any organisms including all plants (Wu et al., 2022). Despite that, there are still few known viruses identified that are naturally infecting non-seed plants, such as algae and bryophytes.

In the case of *Physcomitrium patens*, previous studies proved that, as expected, the plant can be infected in the lab by viruses (Hühns et al., 2003; Šola et al., 2022). Moreover, traces of past viral infections on the genome have been detected, in particular of nucleocytoplasmic large DNA virus relatives (Maumus et al., 2014), which can also be transcribed (Lang et al., 2018). However, no viruses naturally infecting *P. patens* have been described to date.

During the time that we were developing the tools used in chapter one to detect the expression of *P. patens* TEs we decided to check for the presence of viral RNA in the RNAseq data already available that could reflect a viral infection of *P. patens*. We used an approach similar to the one used by Gilbert et al., 2019. To do that we assembled the reads from the RNAseq gene atlas database (Perroud et al., 2018) that did not map to the *P. patens* genome, and we then analyzed the contigs for similarities to RNA dependent RNA polymerases using HMM profiles. This allowed us to describe the first virus to naturally infect the moss, that we named *Physcomitrium patens Amalgavirus 1* (PHPAV1). This was a collaborative project between our group, Fabien Nogué's group in France and Stefan Rensing's group in Germany, and was published last year in Plant Journal (Vendrell-Mir et al., 2021).

From this work I performed the viral detection from the RNAseq gene atlas database, all the qRT-PCRs described in the paper and the different phylogenetic analysis and comparisons to other families, whereas Pierre-François Perroud, who works now at the laboratory of Fabien Nogué and previously in the laboratory of Stefan Rensing, performed the crosses described in the paper. The different accessions, and the different transgenic lines described in the paper were provided by Fabien Nogué and Stefan Rensing and were cultured and developed by Florence Charlot and Pierre-François Perroud from Fabien Nogué's group and the members of Stefan Rensing's group. Part of the experiments of the growth of the moss were done in our lab, such as the time course during all the development of the moss described in the paper and the initial detection of the virus in the lab. Stefan Haas, Rabea Meyberg and Stefan Rensing provided several RNAseq libraries and helped us with the analysis.

After the objectives section and as the main body of this chapter, a copy of the published article is included. All the supplementary material cited in the article can be access through the following DOI:

10.1111/tpj.15545

Chapter 4.2: Objectives

The main objective of this Chapter is:

- Detect and validate the possible presence of viruses infecting *Physcomitrium patens*.

Chapter 4.3: A vertically transmitted amalgavirus is present in certain accessions of the bryophyte *Physcomitrium patens*

A vertically transmitted amalgavirus is present in certain accessions of the bryophyte *Physcomitrium patens*

Pol Vendrell-Mir^{1,†} , Pierre-François Perroud^{2,†} , Fabian B. Haas³ , Rabea Meyberg³ , Florence Charlot² , Stefan A. Rensing^{3,4} , Fabien Nogue^{2,*}  and Josep M. Casacuberta^{1,*} 

¹Centre for Research in Agricultural Genomics CSIC-IRTA-UAB-UB, Campus UAB, Edifici CRAG, Bellaterra, Barcelona 08193, Spain,

²Institut Jean-Pierre Bourgin, INRAE, AgroParisTech, Université Paris-Saclay, Versailles 78000, France,

³Plant Cell Biology, Department of Biology, University of Marburg, Marburg, Germany, and

⁴BIOSS Centre for Biological Signalling Studies, University of Freiburg, Freiburg, Germany

Received 10 August 2021; accepted 16 October 2021; published online 5 November 2021.

*For correspondence (e-mail josep.casacuberta@cragenomica.es [JMC]; fabien.nogue@inrae.fr [FN]).

†These authors contributed equally to this work.

SUMMARY

In the last few years, next-generation sequencing techniques have started to be used to identify new viruses infecting plants. This has allowed to rapidly increase our knowledge on viruses other than those causing symptoms in economically important crops. Here we used this approach to identify a virus infecting *Physcomitrium patens* that has the typical structure of the double-stranded RNA endogenous viruses of the Amalgaviridae family, which we named *Physcomitrium patens* amalgavirus 1, or PHPAV1. PHPAV1 is present only in certain accessions of *P. patens*, where its RNA can be detected throughout the cell cycle of the plant. Our analysis demonstrates that PHPAV1 can be vertically transmitted through both paternal and maternal germlines, in crosses between accessions that contain the virus with accessions that do not contain it. This work suggests that PHPAV1 can replicate in genomic backgrounds different from those that actually contain the virus and opens the door for future studies on virus–host coevolution.

Keywords: *Physcomitrium patens*, *Physcomitrella patens*, amalgavirus, vertical transmission, ribosomal frameshift.

INTRODUCTION

Since the seminal works of Martinus Beijerinck and Dmitri Ivanovsky in the late 19th century that allowed the characterization of tobacco mosaic virus (Scholthof, 2004), thousands of viruses have been discovered in organisms across the three domains of life. Both DNA and RNA viruses infect plants. However, whereas unicellular chlorophyte algae seem to be infected mainly by large DNA viruses, RNA viruses and small DNA viruses seem to be the major classes of viruses infecting angiosperms (Mushegian et al., 2016).

Although most plant virus infections are asymptomatic in the wild (Roossinck, 2015), the vast majority of viruses have been discovered analyzing the lesions or symptoms they induce in economically important crops. Only recently, thanks to the development of next-generation sequencing (NGS) techniques, this has been complemented by a more systematic screen for the presence of viruses in wild and cultivated plants (Villamor et al., 2019).

As a consequence, there is still a clear bias in the databases towards viruses infecting angiosperms, with the only exception of the unicellular algae Chlorophyta, which have a relatively well-studied virome (Mushegian et al., 2016; Yamada et al., 2006). In particular, the knowledge on viruses infecting bryophytes, which include liverworts, mosses, and hornworts, is very limited, with only few viral related RNA-dependent RNA polymerase (RdRP) sequences described in some species (Mushegian et al., 2016).

Physcomitrium patens, formerly known as *Physcomitrella patens*, has been widely used as a model species to study plant evolution and development (Rensing et al., 2020). However, in spite of the wide range of tools and information available for this species, no viruses naturally infecting *P. patens* have been described to date. The only viral-related sequences described in *P. patens* are sequences likely acquired horizontally from nucleocytoplasmic large DNA virus relatives (NCLDVs) (Maumus

et al., 2014). These sequences, which are remnants of past infections, are transcribed in *P. patens* and it has been proposed that they might protect gametes from viral infection via small interfering RNA-mediated silencing (Lang et al., 2018). Moreover, although no virus has been described to infect *P. patens* in the wild, it has been shown that *P. patens* can be infected in the laboratory with viruses such as the tomato spotted wilt virus (Hühns et al., 2003). Here we used the NGS-based approach described in (Gilbert et al., 2019) to look for viral sequences in available *P. patens* RNA sequencing (RNA-Seq) libraries (Haas et al., 2020; Kamisugi et al., 2016; Perroud et al., 2018) and we describe the first *P. patens*-associated virus, *P. patens* amalgavirus 1 (PHPAV1).

RESULTS

Searching for a *P. patens* virus

As most viruses infecting plants are either single-stranded RNA (ssRNA) or double-stranded RNA (dsRNA) viruses we decided to take advantage of the large amount of RNA-Seq data available for *P. patens* (Perroud et al., 2018) to screen for the possible presence of viral sequences that may correspond to *P. patens*-infecting viruses. We screened the RNA-Seq libraries for sequences not mapping to the *P. patens* reference genome. These sequences were then assembled into a total of 184 184 contigs, corresponding to 135 766 independent sequences, as some of them represented different isoforms of the same sequences. The length of these contigs ranged from 200 nucleotides (nts) to 12 470 nts. We then looked for sequence similarities of these contigs to known viruses using different approaches. We first used blastx to look for similarities between the proteins potentially encoded by the RNA contigs and those of the reference viral database (Refseq) from NCBI (on 23 November, 2018). Six contigs gave significant similarity with annotated viruses: four to the same isolate of the tobamovirus pepper mild mottle virus, one to another tobamovirus (tropical soda apple mosaic virus [TSAMV]), and one to RNA segment 3 of the bromovirus brome mosaic virus (Table S1). However, in five of these cases, the *P. patens* contigs were short (close to the lower threshold of the contig size, i.e., 200 nt) and matched a small fraction of the corresponding viral sequences (Table S1). Moreover, in all cases the sequence obtained from *P. patens* RNA-Seq data was almost identical to that of the virus isolate reported (Table S1), even when these viruses infect Solanaceae plants or cereals, which are phylogenetically very distant from *P. patens*. All this suggested that the presence of these viral sequences was due to a possible contamination during library preparation or sequencing processes rather than to the presence of viruses infecting some of the *P. patens* samples. The contig showing similarity to TSAMV had a size similar to that of the virus (6318

nt). The sequence of this contig was also almost identical (99.8% identity) to that of the original virus isolated from *Solanum viarum* (Adkins et al., 2007), suggesting that the presence of the corresponding sequences in *P. patens* RNA-Seq libraries may also be the result of a laboratory contamination. Consistent with this possibility the reads corresponding to the contig showing similarity to TSAMV were present in only one out of the three replicates of the two positive samples, which were obtained from protonema of the Gransden accession (Table S2). The presence of spurious reads in NGS samples used for viral detection is a known problem (Cantalupo and Pipas, 2019). The sources of contamination can be manifold, including viruses contaminating the laboratory reagents (Naccache et al., 2013) and human viruses present in the laboratory personnel. In this respect, it is interesting to note that plant-infecting viruses, including TSAMV, are frequently found in the human oropharynx and gut as a result of plant consumption (Aguado-García et al., 2020; Balique et al., 2015), which may suggest a possible source of contamination.

The second approach followed was to look for similarities of the obtained RNA contigs to viral RdRPs using HMM profiles, a strategy recently used with success to identify new RNA viruses in fungi (Gilbert et al., 2019). Two contigs showed significant similarity to RdRP. One of them corresponded to the contig showing sequence similarity to TSAMV, which was already discarded as a possible contamination (see above). The second contig showing similarity to RdRPs (TRINITY_DN26323_c4_g7_i1), which is 3597 nts long, contains a sequence giving significant similarity to an RdRP HMM profile. A search for sequence similarities showed that the nucleotide sequence of this contig was not significantly similar to any sequence deposited in public databases, including the *P. patens* genome, but that the potentially encoded polypeptide did show significant similarity with RdRPs from amalgaviruses and partitiviruses (Figure 1a). The absence of nucleotide sequence similarity to sequences deposited in the databases suggested that the corresponding RNA-Seq reads present in the *P. patens* libraries were not the result of a contamination with a known virus and that they may correspond to a previously undetected *P. patens* virus. Moreover, the distribution of the RNA-Seq reads corresponding to this contig among the *P. patens* samples, with reads present in all replicates of only few samples of similar tissues (Table S2), also suggested that the presence of the reads was not the result of a contamination and was consistent with the presence of a virus in the specimen being sequenced.

Characterization of the first *P. patens* amalgavirus

The viruses of the Amalgaviridae family are small dsRNA viruses (around 3400 bp) containing two partially overlapping ORFs (Figure 2a). ORF1 encodes a product whose

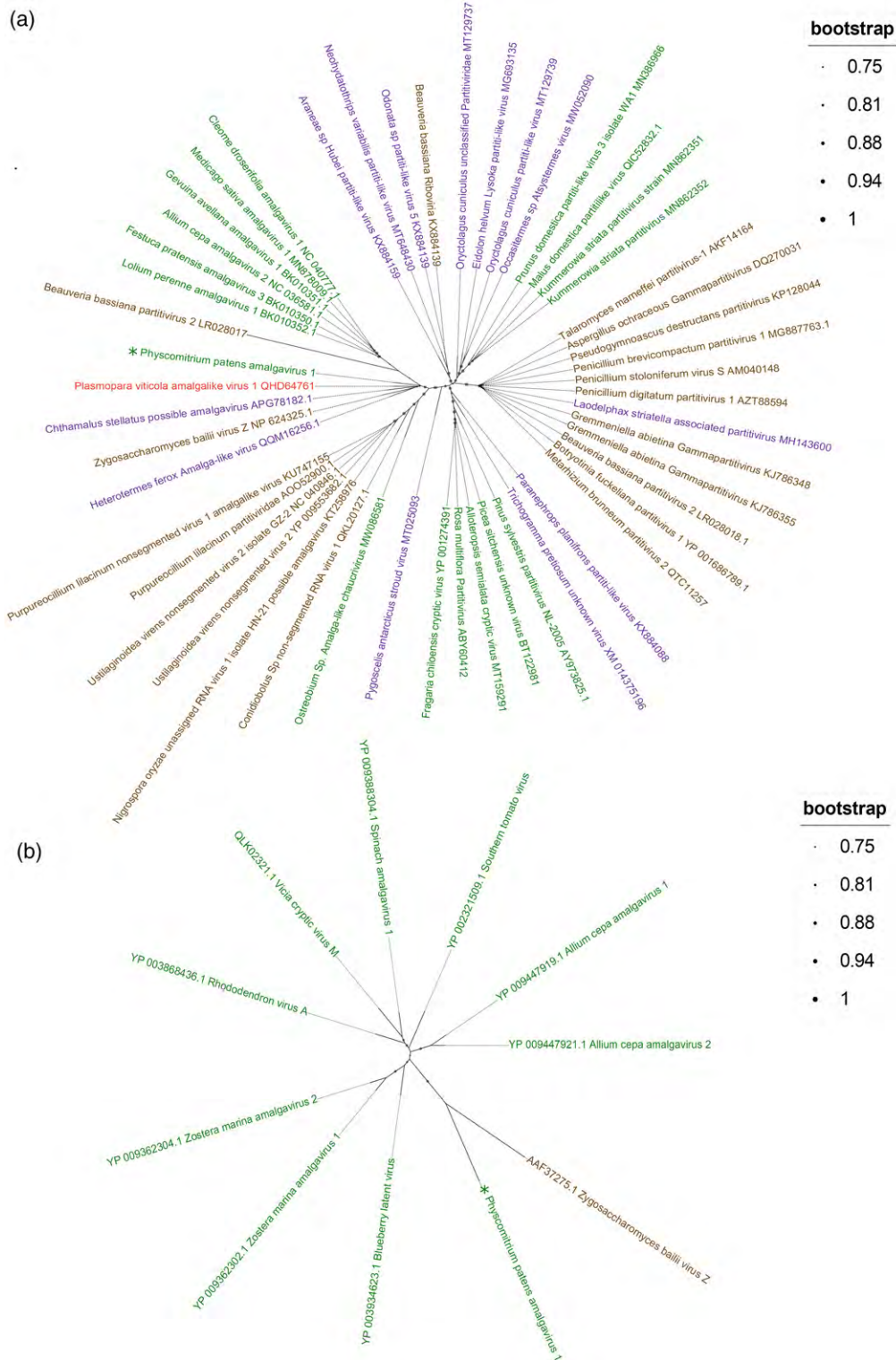
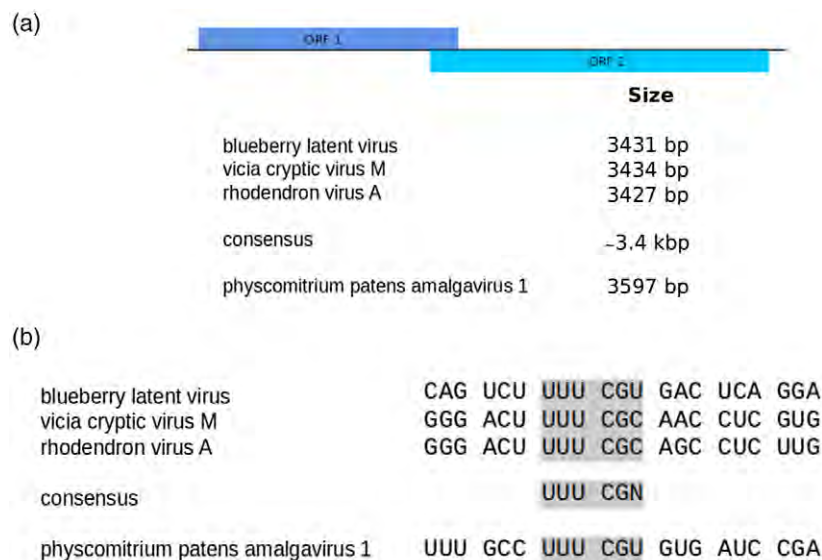


Figure 1. (a) Neighbor-joining phylogenetic unrooted tree of the RdRP encoded by PHPAV1 and the most similar sequences deposited in databases, including RdRPs from amalgaviruses and partitiviruses. The RdRP putative sequence of the *P. patens* PHPAV1 virus is labeled with an asterisk. RdRP sequences obtained from plant, fungus, oomycete, and animal viruses are shown in green, brown, red, and violet, respectively. (b) Neighbor-joining phylogenetic unrooted tree of the RdRP proteins encoded by PHPAV1 and representative viruses of the nine amalgavirus species and the single zybavirus species using the same color code used in (a).

Figure 2. (a) Schematic representation of the ORFs encoded by PHPAV1 compared with those of already characterized amalgaviruses. (b) Comparison of the putative ribosomal frameshift sequence of PHPAV1 with the consensus translation frameshift motif of amalgaviruses.



function has not been established yet, although it has been proposed to function as a nucleocapsid protein (Krupovic et al., 2015) or a replication factory matrix-like protein (Isogai et al., 2011). ORF2 is encoded in the +1 frame with respect to ORF1 and it is supposed to be translated as part of an ORF1-ORF2 fusion protein by means of a specific ribosome frameshift (Martin et al., 2011; Nibert et al., 2016). The ORF2-encoded part of the ORF1-ORF2 fusion protein shows sequence similarity to partitivirus RdRPs. The Amalgaviridae family has been divided into two genera: amalgaviruses, which include nine viral species, all infecting plants, and zybaviruses, with only one viral species, infecting budding yeast. Apart from the different types of host, the viruses belonging to the two genera also differ in their genome size, which is smaller for the zybavirus (3.1 kbp) than for the amalgaviruses (3.4–3.5 kbp) (Tzanetakis et al., 2021).

The sequence of contig TRINITY_DN26323_c4_g7_i1 is 3597 bp long (Data S1) and contains two overlapping ORFs, ORF1 and ORF2, with ORF2 being encoded in the +1 frame with respect to ORF1 (Figure 2a). The two ORFs are 1119 and 2601 nts long, respectively. They are flanked by a 5' UTR of 8 nts and a 3' UTR of 15 nts, and the two ORFs overlaps for 176 nts (positions 951 to 1127 in the sequence). Interestingly, the ORF1-ORF2 overlapping region contains the sequence UUU CGU that fits the consensus of the +1 ribosomal frameshift motif described for amalgaviruses (UUU CGN) (Nibert et al., 2016) (Figure 2b). As already mentioned, ORF2 shows high sequence similarity to RdRPs of amalgaviruses and partitiviruses, whereas ORF1 does not show significant similarities to proteins in public databases. Because of all the above, we propose that the sequence of contig TRINITY_DN26323_c4_g7_i1 corresponds to the complete RNA of a new virus of the

Amalgaviridae family. As this virus infects a plant and has a genome size similar to that of characterized amalgaviruses and bigger than that of the only zybavirus characterized (Tzanetakis et al., 2021), we named this virus *P. patens* amalgavirus 1 (PHPAV1). Nevertheless, a phylogenetic analysis of the RdRP sequence of PHPAV1 and those of the nine amalgavirus species and the single zybavirus species shows that PHPAV1 RdRP seems more closely related to that of the single zybavirus described than to that of the characterized amalgaviruses (Figure 1b). A better sampling of the viruses of the Amalgaviridae family infecting plants and fungi will probably be needed to clarify this apparent contradiction. In any case, PHPAV1 constitutes not only the first virus of the Amalgaviridae family described in *P. patens* but also the first virus described infecting a bryophyte.

Presence of PHPAV1 in different *P. patens* accessions

Analysis of the reads assembled corresponding to PHPAV1 showed that they were all from the sporophyte libraries analyzed. These libraries were the only ones obtained from the *P. patens* Reute accession (Table S2). This suggested that the virus may accumulate in sporophytes or that it may be present in the Reute accession but not in Gransden, which was the accession all the other libraries were obtained from. In order to analyze the pattern of accumulation of PHPAV1 in the Reute ecotype we performed qRT-PCR analyses using RNAs obtained from different tissues and organs of the Reute accession. As shown in Figure 3, the viral RNA was detected in all tissues analyzed and not only in sporophytes.

In order to analyze the presence of PHPAV1 in *P. patens* accessions we performed qRT-PCR analysis on RNA obtained from gametophores of eight different *P. patens*

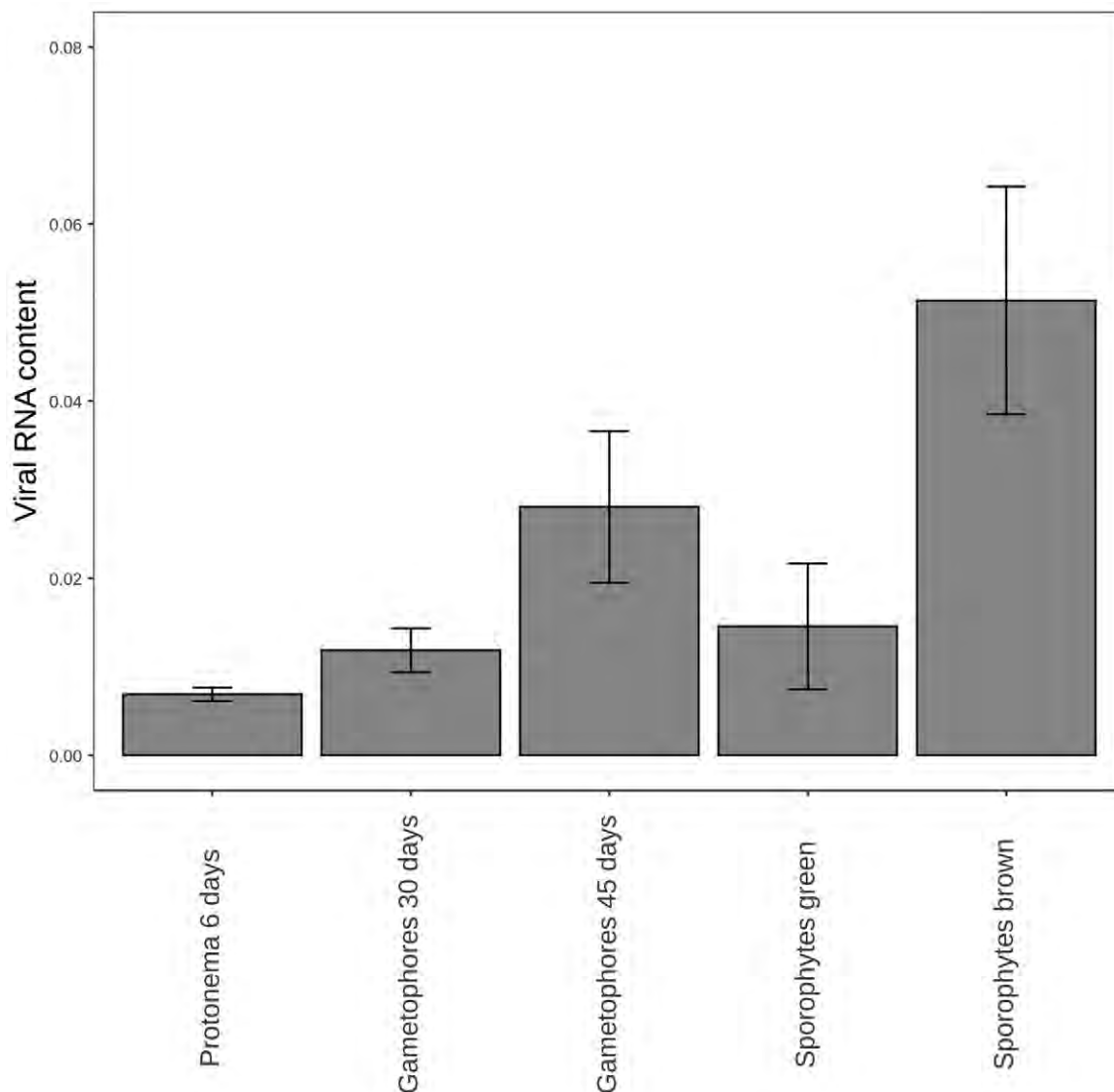


Figure 3. Quantification of PHPAV1 RNA in different tissues of *P. patens* Reute. The levels of viral RNA are given relative to the mRNA levels of the gene encoding the 60S ribosomal subunit.

accessions: Gransden, Reute, Villersexel, Kaskaskia, Wisconsin, Uppsala, Trondheim, and Lviv, which have different geographical origins (Figure 4a). We confirmed the presence of the viral RNA in Reute and its absence in Gransden, as the analysis of the RNA-Seq data already suggested. Moreover, we detected the virus in two additional accessions, Kaskaskia and Lviv, whereas the virus was not detected in the remaining four accessions, Villersexel, Wisconsin, Uppsala, and Trondheim (Figure 4b). The presence of the virus does not correlate with any symptom, although the phenotypic differences between the accessions make it difficult to rule out minor phenotypic effects. Sequencing of the 308-bp PCR products indicated that the sequence of the virus present in different accessions is not identical. Indeed, we detected three different

base pairs in the sequence obtained from Kaskaskia and one different base pair in the sequence obtained from Lviv with respect to the sequence of the virus present in Reute. RNA-Seq data from *P. patens* Kaskaskia and Villersexel have been recently reported (Haas et al., 2020; Kamisugi et al., 2016). We did not detect reads corresponding to PHPAV1 in the Villersexel libraries, in line with the qRT-PCR results discussed above, which suggests that PHPAV1 is not present in this accession. On the other hand, the RNA-Seq data from the Kaskaskia accession gave us the opportunity to deduce the complete sequence of the virus present in this accession. Here we detected a total of 19 different base pairs over the entire 3597 bp of the virus sequence compared to the sequence obtained from Reute samples, including the one already detected based on the

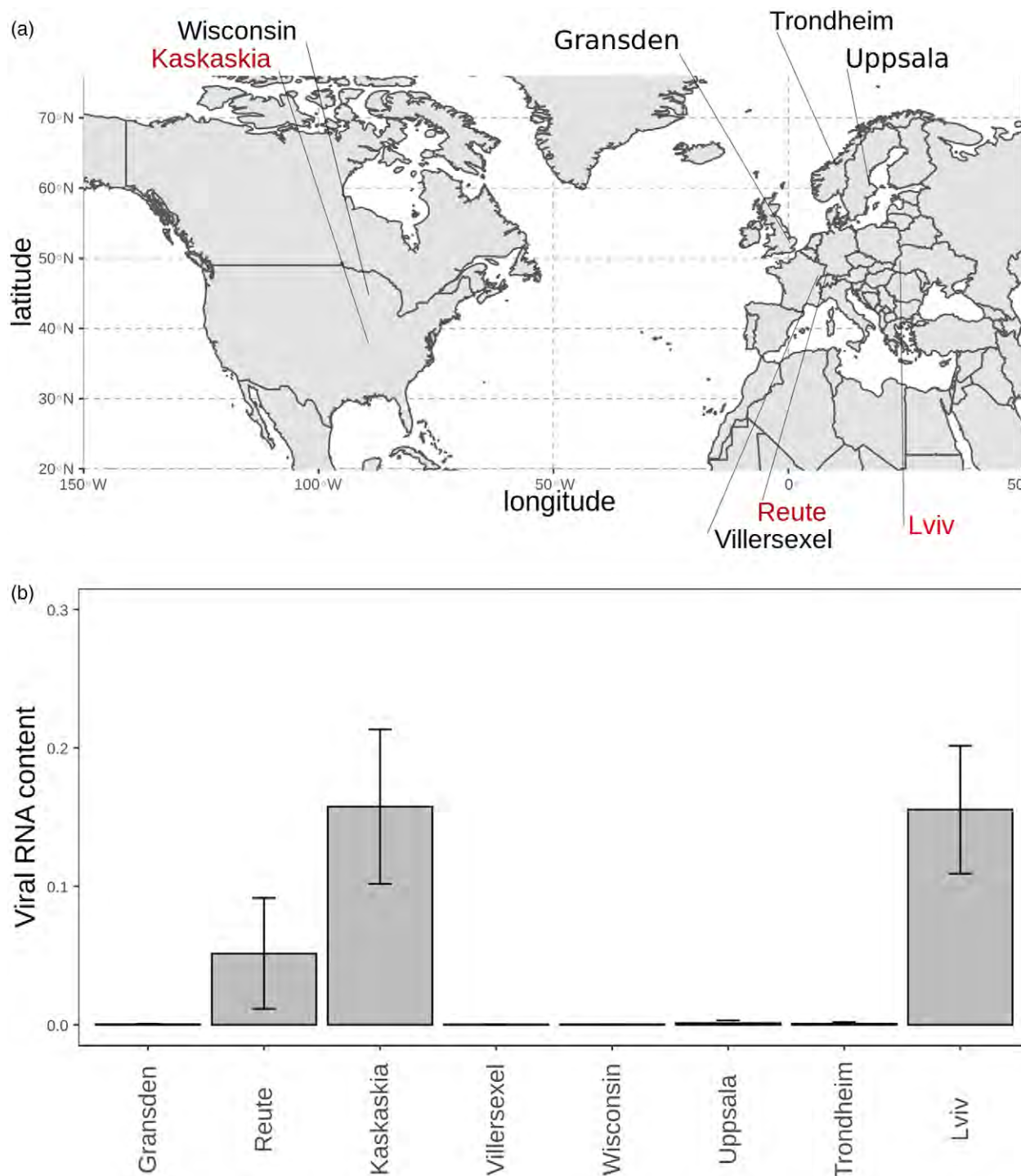


Figure 4. Presence of PHPAV1 in different *P. patens* accessions. (a) Geographical distribution of the *P. patens* accessions analyzed. The names of the accessions containing the virus are shown in red and those not containing the virus are shown in black. (b) Quantification of the PHPAV1 levels in gametophores of different *P. patens* accessions. The viral RNA levels are given relative to the mRNA levels of the gene encoding the 60S ribosomal subunit.

PCR analysis mentioned above (Table S3). These nucleotide variations are outside the motif potentially linked to the translation frameshift allowing the translation of the ORF1-ORF2 fusion protein and would result in only five

amino acid changes in the encoded polypeptides, two in ORF1 and three in ORF2 (Table S3). Interestingly, sequencing of the PCR products from the three accessions that contain the virus, as well as of the Reute and Kaskaskia

RNA-Seq reads available, did not reveal any variability of the virus sequence within each of these ecotypes.

PHPAV1 vertical transmission

Amalgaviruses are endogenous viruses thought to be transmitted vertically (Martin et al., 2011). We analyzed the transmission of the virus by means of crosses between accessions containing or lacking the virus. In order to analyze its possible transmission through parental and maternal gametes, we performed two different crosses. For the first one, designed to assess the potential transmission of the virus through the male germline, we took advantage of the reduced male fertility of the Gransden accession (Meyberg et al., 2020) to cross it with the virus-containing Kaskaskia accession. We used a Kaskaskia line expressing a YFP (*Ka-YFP*; Methods S1) to ensure selecting sporophytes resulting from the cross. In the second cross, a Reute male sterile *Ppccdc39* strain (Meyberg et al., 2020), containing the virus, was crossed with Villersexel-mCherry (*Vx-red*) which does not contain the virus, acting as male progenitor. As shown in Figure 5, the virus could be detected in a pool of 10 individual plants resulting from a single sporophyte from each of the two crosses, indicating that the

virus can be transmitted by both male (Figure 5a) and female (Figure 5b) reproductive organs, antheridia and archegonia. The sequence of the PCR products allowed identifying the characteristic nucleotide differences of the PHPAV1 sequence from Reute and Kaskaskia accessions, confirming the origin of the identified sequences. Interestingly, whereas the 10 individual plants resulting from the Villersexel × Reute cross all showed the presence of the virus (Figure 5e), only a fraction (6/10) of the plants resulting from the Kaskaskia × Gransden cross seems to contain PHPAV1 (Figure 5d). As all the 10 plants come from the same fertilization event, differences in the presence of the virus should be due to a post-fertilization loss of the virus in some plants. Whether this loss is a stochastic event during cell division or it is due to genetic differences between Kaskaskia and Gransden accessions affecting viral maintenance remains to be analyzed. Moreover, as the virus present in the Kaskaskia accession is not identical to the one present in the Reute accession, the differences in the prevalence of the virus in the offspring could also be due to differences in the persistence of the virus itself. In order to start investigating this aspect, we designed a cross between Kaskaskia (*Ka-YFP*) acting as paternal progenitor

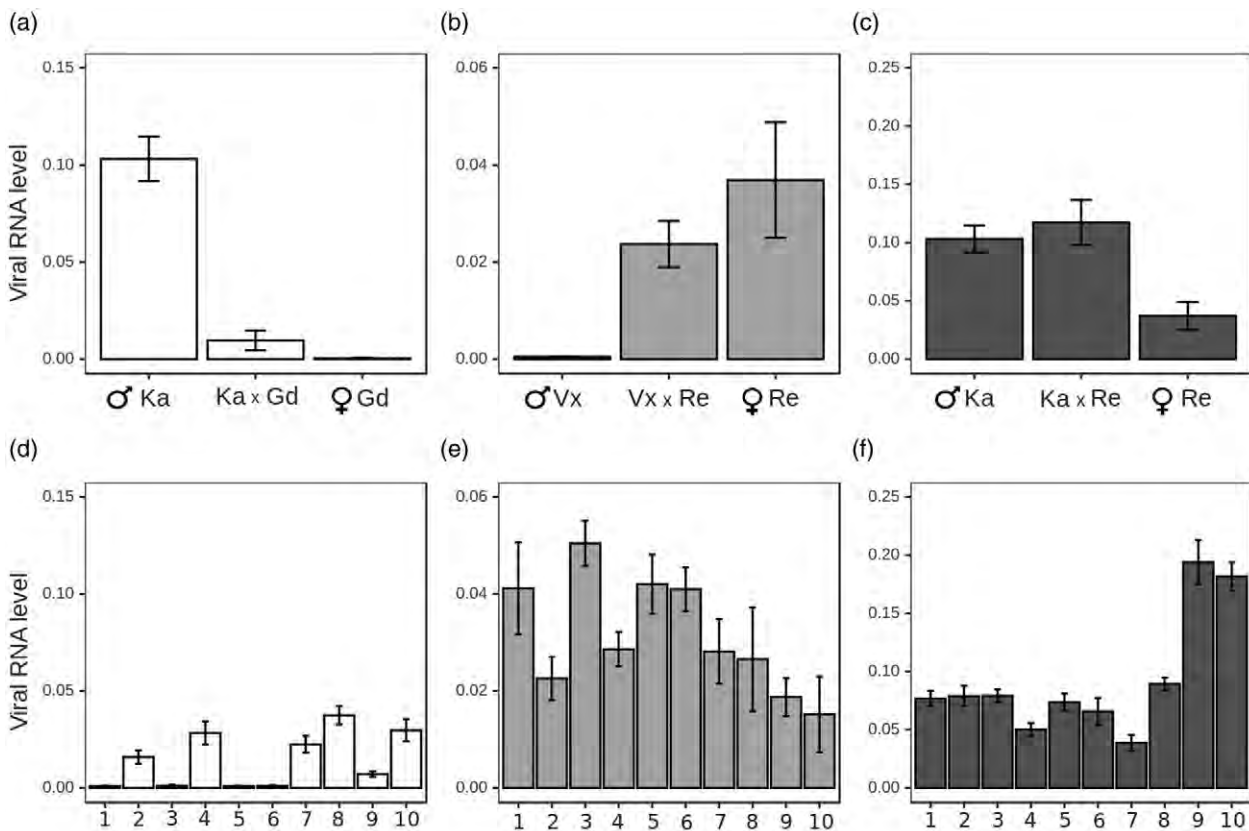


Figure 5. Detection of the PHPAV1 RNA in crosses of different accessions of *P. patens*. (a and d) Crosses between Kaskaskia acting as male and Gransden as female. (b and e) Crosses between Kaskaskia acting as male and Reute as female. (c and f) Crosses between Villersexel acting as male and Reute as female. (a, b, and c) Quantification of the RNA of the virus in a pool of 10 plants resulting from a single sporophyte obtained from each cross. (d, e, and f) Quantification of the virus in each of the 10 independent plants. The viral RNA levels are given relative to the mRNA levels of the gene encoding the 60S ribosomal subunit.

and Reute (*Ppccdc39*) acting as maternal progenitor. As shown in Figure 5c, the virus could be detected in a pool of 10 individual plants resulting from a single sporophyte from the cross, and also in each of the 10 individual plants resulting from this cross (Figure 5f). The analysis of the sequence of the virus present in these 10 individual plants showed that only the Reute virus was detected.

DISCUSSION

Amalgaviridae is a recently reported family of dsRNA viruses, described as an amalgam between viruses of the Totiviridae and Partitiviridae families, with a structure more related to the former but encoding proteins phylogenetically related to the latter (Martin et al., 2011). The Amalgaviridae family has been recently divided in two genera: amalgaviruses, including nine viral species infecting plants, and zybaviruses, including only one species that infects budding yeast (Tzanetakis et al., 2021). Whereas most amalgaviruses reported are plant viruses (Goh et al., 2018; Martin et al., 2011; Park et al., 2018; Sabanadzovic et al., 2009; Zhan et al., 2019), viruses described as amalgaviruses and present in fungi have also been reported (Koloniuk et al., 2015), and different sequences potentially related to amalgaviruses and zybaviruses obtained from plant, fungi, oomycetes, and animals have been deposited in public databases. The host species and the size of PHPAV1, which is more similar to that of the viruses classified as amalgaviruses, prompted us to tentatively classify PHPAV1 as an amalgavirus. However, the RdRP sequence of PHPAV1 seems more closely related to that of the fungus-infecting virus that defines the zybavirus genus than to that of the nine plant-infecting viruses of the amalgavirus genus. Moreover, phylogenetic analysis of the PHPAV1 sequence together with the most similar sequences deposited in databases shows that PHPAV1 falls into a not well-defined group of sequences of viruses from fungi, oomycetes, and animals outside the two main groups, of which one contains most plant viral sequences and the other contains most fungal viral sequences. The apparent lack of consistency of this phylogeny with that of the species these viruses infect could simply be the result of a partial sampling. Amalgaviruses have only been described recently and the fact that in most cases these endogenous viruses do not seem to lead to lesions or symptoms makes their discovery more difficult. Therefore, it is very plausible that the picture we have at present is incomplete and this could lead to apparently incongruent phylogenetic relationships when comparing these sequences. Indeed, whereas no amalgaviruses have been described in bryophytes, sequences related to partivirus RdRPs, which are close relatives of amalgaviruses, have been detected in different bryophytes (Mushegian et al., 2016). Alternatively, although amalgaviruses are thought to be transmitted vertically, the close proximity in

nature of many of the species these viruses infect could support their horizontal transfer. A more in depth sampling and careful analysis would be required to answer this open question.

The viruses of the Amalgaviridae family are vertically transmitted endogenous viruses. As they are dsRNA virus and they do not integrate into the genome of their hosts, the dsRNA needs to be present throughout the life cycle of the host. We show here that PHPAV1 is present in all tissues tested, which are representative of the *P. patens* life-cycle, including in the tissues and organs that will give rise to the germline. In addition, we show that PHPAV1 is transmitted through protoplasts, as the transgenic lines *Ka-YFP* and *Ppccdc39* were obtained by regenerating single transgenic protoplasts obtained from *Kaskaskia* and Reute plants, respectively.

However, PHPAV1 is only present in certain *P. patens* accessions. Among the eight accessions tested, PHPAV1 appears to be present in only three, Reute, *Kaskaskia*, and Lviv, and to be absent from the other five, Gransden, Villersexel, Wisconsin, Uppsala, and Trondheim. This distribution of the virus does not fit the geographical distribution of *P. patens* accessions, with the virus present in geographically distant accessions and absent from closer ones. For example, the virus is present in one of the two accessions from the USA, *Kaskaskia*, and is present in Reute but not in Villersexel in spite of these two accessions being from close geographical origins in central Europe. This distribution could be the result of recent independent infections of the accessions that contain the virus or the recent loss in some accessions of the virus, which could have infected a precursor individual of all *P. patens* accessions.

The results presented here show that the virus can be transmitted through both paternal and maternal germ lines and that the virus can be detected in the offspring of crosses between accessions, even when only one of the parental accessions (either the parental or the maternal progenitor) contains the virus. This suggests that although the virus is only present in certain accessions, it can replicate and be maintained in other *P. patens* genetic backgrounds. However, it is interesting to note that in the cross between *Kaskaskia* acting as the male strain and Gransden acting as the female strain, only six out of the 10 plants resulting from a single sporophyte showed detectable levels of PHPAV1, suggesting that the virus was not maintained at a detectable level in all of them. This absence of PHPAV1 could simply be the result of a stochastic loss of the virus due to an uneven distribution of the virus present at low level in the progenitor cell. But this result could also indicate that certain genetic backgrounds, resulting from the recombination of Gransden and *Kaskaskia* genomes, may not allow the maintenance of the virus. In other words, this result may suggest that one or several genetic factors, variable among *P. patens* accessions, may restrict

the proliferation of the virus. This would explain why the virus can be present in a particular *P. patens* accession and absent from another accession located geographically close and potentially interfertile.

Finally, our results also show that in a cross between two accessions containing the virus, the offspring does not necessarily inherit the virus from both parents. Indeed, in the cross analyzed here between Kaskaskia acting as paternal progenitor and Reute acting as maternal progenitor, the 10 plants resulting from a single sporophyte did contain the virus, but in all cases the virus was inherited from the maternal progenitor. This result could suggest that viral transfer from the maternal and from the paternal progenitor is not equally efficient, but it may also point to an effect on viral persistence of viral variants, as the sequence of the virus present in Kaskaskia and Reute accessions is not identical. It is interesting to note that, whereas the sequence of the virus from different accessions is different, we have not detected any variability of the virus within each of the accessions. RNA viruses are often characterized by a high sequence variability known as quasispecies-like structure (Holland et al., 1982). However, it has been suggested that endogenous persistent viruses may be much less variable (Safari and Roossinck, 2014). Although there is almost no population study of persistent viruses in plants, a recent report of two persistent endogenous viruses from spinach (*Spinacia oleracea*), a partitivirus and an amalgavirus, showed that they present very low sequence diversity (Samarfard et al., 2020). Similarly, the analysis of archeological samples has shown that a maize (*Zea mays*) endogenous persistent virus belonging to the family Chrysoviridae has undergone only about 3% divergence after 1000 years of maize cultivation (Peyambari et al., 2018). The reasons for the low variability of endogenous persistent viruses are not fully understood, but may be related to their stamped machine mode of replication, which they share with most dsRNA viruses, and their particular lifestyle (Safari and Roossinck, 2014). Indeed, their long coexistence with their hosts and not being targeted by the host silencing mechanisms makes a quasispecies-like structure much less advantageous. The absence of sequence variability of the viral sequences within a single accession, together with the sequence differences found in PHPAV1 from different accessions, may suggest an accession-specific evolution and adaptation.

Most infections of amalgaviruses in plants are asymptomatic. Although in some cases the infected plants did show symptoms, the correlation between the symptoms and the presence of the virus could not be confirmed (Zhan et al., 2019). In any case, the absence of symptoms does not imply that the infections have no effect on the host. It has been shown recently that the infection of tomato (*Solanum lycopersicum*) by Southern tomato virus, which belongs to the amalgavirus family, results in changes of

expression of micro-RNAs that could modify the response of the host to different stresses (Elvira-González et al., 2020). In fact, different reports point to beneficial effects of persistent viral infections in plants, including protection against acute infections of more harmful viruses, regulation of nodulation, better tolerance to stress, and increased plant height and fruit production (Takahashi et al., 2019).

Eukaryotic, and in particular, plant genomes contain integrated sequences from endogenous viruses. For example, plant genomes frequently contain sequences of caulimoviruses (Geering et al., 2014), and it has been shown that the *P. patens* genome contains NCLDV sequences (Lang et al., 2018; Maumus et al., 2014). These integrated viral sequences may protect the plant against viral infections, as proposed for the integrated NCLDV in *P. patens* (Lang et al., 2018), and may also contribute to the evolution of the genome acting as sources of novel genetic material (Geering et al., 2014). As amalgaviruses do not integrate in the genome of their hosts, a direct role in host genome evolution is less obvious to imagine. However, their potential beneficial effects remain a very attractive field for future study. The report here of an amalgavirus infecting only some *P. patens* accessions provides an interesting model to study the potential function of these viruses and their adaptation to their hosts.

EXPERIMENTAL PROCEDURES

Libraries used

We checked for the presence of PHPAV1 in Gransden and Reute accessions using the RNA-Seq data available in the first release of the *P. patens* gene atlas (Lang et al., 2018) and in Kaskaskia and Villersexel accessions using RNA-Seq data deposited in NCBI SRA in BioProjects PRJNA601618 and PRJNA602303 (Haas et al., 2020) and libraries SRX031156 and SRX031155 (Kamisugi et al., 2016).

Contig formation and virus identification

All reads were quality-trimmed using bbdduk (Bushnell, 2014). We mapped all the reads to the *P. patens* genome (including the plastid genomes) (Lang et al., 2018) using STAR (Dobin et al., 2013) with the following flag to select the unmapped reads: '-outSAMunmapped'. Then, we searched for all these reads that did not align to the reference genome using the samtools-1.12 package (Li et al., 2009) with the flag (-f 4). The selected reads from all libraries were pooled to be assembled using Trinity (Grabherr et al., 2011) with default parameters.

Two different approaches were followed to identify viral sequences in the assembled sequences. First, we used blastx to search for similarities between the putatively encoded polypeptides and viral proteins deposited at the reference viral database from NCBI Refseq (Brister et al., 2015) with an e-value cutoff of 0.05 (done on 23 November, 2018). Second, we searched for RdRP signatures in all the polypeptides potentially encoded by the assembled sequences using HMMscan (version 3.1b2), based on hidden Markov models (HMMs), in an approach similar to the one described by Gilbert et al. and using the same e-value cutoff of 10 (Gilbert et al., 2019). We used the HMM profiles from the

following PFAM accessions: PF00680, PF00978, PF02123, PF05183, PF07925, and PF17501.

Phylogenetic analyses

We took the best 50 hits from the blastx search done (19 April, 2021) using the nucleotide sequence of PHPAV1 against the non-redundant protein sequences from the NCBI database to build the phylogenetic tree. All these sequences were aligned using Mafft (Kato and Standley, 2013) (with --auto and -adjust_direction parameters) and trimmed with TrimAL (Capella-Gutiérrez et al., 2009) (with -automated1 mode). We inferred an approximately maximum-likelihood phylogenetic tree using the FastTree program (Price et al., 2010) with default parameters and visualized it using iTol (Letunic and Bork, 2019).

Virus variability among different accessions

The reads corresponding to PHPAV1 from different accessions were aligned to the Reute PHPAV1 sequence using bowtie2 (Langmead and Salzberg, 2013) (with: --threads 8 --local --no-unal -k 2).

Plant material and culture

We used wild-type *P. patens* accessions of different origins: Grandsden (Gd) pedigree Gd_JP_St Louis (Haas et al., 2020) and Reute-K1 (Re) (Hiss et al., 2017), which are routinely used in laboratory settings, and Villersexel-K3 (Vx), Kaskaskia (Ka), Lviv (Lv), Trondheim-K2 (Td), and Uppsala-K1 (Up), which were obtained from the International Moss Stock Center (<https://www.moss-stock-center.org/>). In addition, we used the Wisconsin (Wi) accession, which has been recently isolated from the wild (Haas et al., 2020). In order to facilitate the isolation of crossed plants, we used transgenic *P. patens* expressing a fluorescent protein. *Vx-red* has been established in a Villersexel-K3 background and accumulates mCherry (Perroud et al., 2011). *Re-mCherry* is a Reute-K1 transformant that accumulates mCherry (Perroud et al., 2019). *Ka-YFP* has been established in the Kaskaskia ecotype and accumulates eYFP (Methods S1). For crossing analyses, a male sterile Re-mutant, *Ppccdc39*, was used (Meyberg et al., 2020).

If not mentioned otherwise, mosses were grown at 24°C with a cycle of 16 h of light (quantum irradiance of 60–80 $\mu\text{mol m}^{-2} \text{s}^{-1}$ light) and 8 h of darkness on plates containing BCDA medium (Cove et al., 2009). Harvested samples were systematically flash-frozen in liquid N₂ and stored at –80°C until further processing. For all comparisons between accessions as well as for the progeny analyses, gametophores were harvested from plants grown for 1 month.

Protonema Reute samples were harvested from a 6-day-old entrained culture grown on solid BCDA medium overlaid with cellophane. Reute juvenile (asexual) gametophore samples were harvested from a 1-month-old culture grown on BCD medium (Cove et al., 2009). Adult Reute gametophore (bearing gametangia) samples were harvested from culture grown for 1 month on BCD medium and subsequently transferred to 15°C with a cycle of 8 h of light (30 $\mu\text{mol m}^{-2} \text{s}^{-1}$) and 16 h of darkness for 2 weeks (Hohe et al., 2002). The presence of gametangia was visually confirmed prior to harvesting. The sporophyte samples were obtained from culture grown initially under the adult gametophore regime but watered after the 2 weeks at 15°C to facilitate fertilization. After subsequent growth in the same conditions, green and brown sporophyte samples (Hiss et al., 2017) were collected after 5–7 weeks.

Crossing experiments were performed by coculture of two parental strains as previously described (Perroud et al., 2011). Mature sporophytes (brown) displaying fluorescence developed on a gametophore devoid of fluorescence were manually isolated and

stored at least for a week in the dark at 4°C before performing the germination assay. For each tested sporophyte, 300 to 400 spores were germinated (Perroud et al., 2019) and the predicted 1:1 segregation of the fluorescent marker was confirmed to ensure crossing prior to further work. Then, 10 sporelings were isolated per crossed sporophyte and their tissues were amplified to perform subsequent analyses.

RNA extraction and cDNA library preparation

RNA extraction and DNase treatment were done using the Maxwell[®] RSC Plant Kit (Promega, Ref #AS1500) following the manufacturer's instructions. RNA was resuspended in 30 μl of sterile water. We used 200 ng of total RNA for first-strand cDNA synthesis using a 5'-modified oligo-dT primer (Casacuberta et al., 1995) and a specific primer complementary to the PHPAV1 sequence (αPV63 : 5'-CCCACCTCCCTTACAGACGATC-3') with SuperScript[™] III reverse transcriptase (Thermo Fisher). As a control we performed a reaction with the same amount of RNA (200 ng) without adding the SuperScript[™] III reverse transcriptase.

qRT-PCR

Quantitative real-time PCRs were performed in 96-well plates using a Roche LightCycler II instrument. We used SYBR Green I Master Mix (Roche Applied Science) with primers (αPV64 : 5'-CAAAAGCCTATTCCTCTGCATGG-3', αPV65 : 5'-CTGAGAGAACTGTCCCGTAACT-3') at 1 μM and 40 ng of cDNA obtained from the reverse transcription to conduct the qRT-PCR analysis. Each sample was run in triplicate with negative reverse transcriptase for each sample and negative controls (H₂O). The following PCR conditions were used: 95°C for 5 min, followed by 95°C for 10 sec, 56°C for 10 sec, and 72°C for 10 sec. For normalization we used the relatively highly expressed gene encoding the 60S ribosomal subunit as previously described (Le Bail et al., 2013). The PCR products were purified using the Macherey-Nagel Nucleospin[®] Gel and PCR cleanup kit (Ref #740609.50) and sequenced.

ACKNOWLEDGEMENTS

This work was supported by grants from the Spanish Ministerio de Economía, Industria y Competitividad (AGL2016-78992-R/FEDER) and Ministerio de Ciencia e Innovación (PID2019-106374RB-I00/AEI/10.13039/501100011033) to JMC. The IJPB benefits from the support of the LabEx Saclay Plant Sciences (SPS) (ANR-10-LABX-0040-SPS). We thank Prof. Mitsuyasu Hasebe for the generous gift of the plasmid p35S-loxP-BSD. We are grateful to Leonie Verhage for providing us with the *P. patens* ecotypes Lviv, Trondheim-K2, and Uppsala-K1. We thank Jean-Luc Gallois and Juan José López-Moya for helpful discussions.

CONFLICT OF INTEREST

The authors have no conflict of interest to declare.

DATA AVAILABILITY STATEMENT

All relevant data can be found within the manuscript and its supporting materials. The complete sequence of PHPAV1 can be found at GenBank under accession number BK059361.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Table S1. Sequence similarity of the selected contigs with the corresponding viruses.

Table S2. Reads mapping to the contigs corresponding to PHPAV1 and TSAMV in a subset of the analyzed gene atlas (Perroud et al., 2018) samples.

Table S3. Single-nucleotide polymorphisms detected between the reference PHPAV1 sequence from Reute and the viral reads detected in Kaskaskia.

Methods S1. Information regarding the development of the *Ka-YFP* line.

Data S1. Sequence of contig TRINITY_DN26323_c4_g7_i1, which corresponds to PHPAV1.

REFERENCES

- Adkins, S., Kamenova, I., Rosskopf, E. & Lewandowski, D. (2007) Identification and characterization of a novel tobamovirus from tropical soda apple in Florida. *Plant Disease*, **91**, 287–293.
- Agudo-García, Y., Taboada, B., Morán, P., Rivera-Gutiérrez, X., Serrano-Vázquez, A., Iba, P. et al. (2020) Tobamoviruses can be frequently present in the oropharynx and gut of infants during their first year of life. *Scientific Reports*, **12**, 13595.
- Bail, A.L., Scholz, S. & Kost, B. (2013) Evaluation of reference genes for RT-qPCR analyses of structure-specific and hormone regulated gene expression in *Physcomitrella patens* gametophytes. *PLoS One*, **8**, e70998.
- Balique, F., Lecoq, H., Raoult, H. & Colson, P. (2015) Can plant viruses cross the kingdom border and be pathogenic to humans? *Viruses*, **7**, 2074–2098.
- Brister, J., Ako-Adjei, D., Bao, Y. & Blinkova, O. (2015) NCBI viral genomes resource. *Nucleic Acids Research*, **43**, D571–D577.
- Bushnell, B. (2014) *BBMap: a fast, accurate, splice-aware aligner*. No. LBNL-7065E. Berkeley, CA: Ernest Orlando Lawrence Berkeley National Laboratory.
- Cantalupo, P. & Pipas, J. (2019) Detecting viral sequences in NGS data. *Current Opinion in Virology*, **39**, 41–48.
- Capella-Gutiérrez, S., Silla-Martínez, J.M. & Gabaldón, T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, **25**, 1972–1973.
- Casacuberta, J.M., Vernhettes, S. & Grandbastien, M.A. (1995) Sequence variability within the tobacco retrotransposon Tnt1 population. *EMBO Journal*, **14**, 2670–2678.
- Cove, D.J., Perroud, P.F., Charron, A.J., McDaniel, S.F., Khandelwal, A. & Quatrano, R.S. (2009) Culturing the moss *Physcomitrella patens*. *Cold Spring Harbor Protocols*, **4**, pdb.prot5136.
- Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S. et al. (2013) STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
- Elvira-González, L., Medina, V., Rubio, L. & Galipienso, L. (2020) The persistent southern tomato virus modifies miRNA expression without inducing symptoms and cell ultra-structural changes. *European Journal of Plant Pathology*, **156**, 615–622.
- Geering, A., Maumus, F., Copetti, D., Choisine, N., Zwickl, D. J., Zytnicki, M. et al. (2014) Endogenous florendoviruses are major components of plant genomes and hallmarks of virus evolution. *Nature Communications*, **5**, 5269.
- Gilbert, K.B., Holcomb, E.E., Allscheid, R.L. & Carrington, J.C. (2019) Hiding in plain sight: new virus genomes discovered via a systematic analysis of fungal public transcriptomes. *PLoS One*, **14**, e0219207.
- Goh, C., Park, D., Lee, J., Sebastiani, F. & Hahn, Y. (2018) Identification of a novel plant amalgavirus (Amalgavirus, Amalgaviridae) genome sequence in *Cistus incanus*. *Acta Virologica*, **62**, 122–128.
- Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I. et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, **29**, 644–652.
- Haas, F.B., Fernandez-Pozo, N., Meyberg, R., Perroud, P.F., Göttig, M., Stingl, N. et al. (2020) Single nucleotide polymorphism charting of *P. patens* reveals accumulation of somatic mutations during in vitro culture on the scale of natural variation by selfing. *Frontiers in Plant Science*, **11**, 813.
- Hiss, M., Meyberg, R., Westermann, J., Haas, F.B., Schneider, L., Schallenberg-Rüdinger, M. et al. (2017) Sexual reproduction, sporophyte development and molecular variation in the model moss *Physcomitrella patens*: introducing the ecotype Reute. *The Plant Journal*, **90**, 606–620.
- Hohe, A., Rensing, S., Mildner, M., Lang, D. & Reski, R. (2002) Day length and temperature strongly influence sexual reproduction and expression of a novel MADS-box gene in the Moss *Physcomitrella patens*. *Plant Biology*, **4**, 595–602.
- Holland, J., Spindler, K., Horodyski, F., Grabau, E., Nichol, S. & VandePol, S. (1982) Rapid evolution of RNA genomes. *Science*, **215**, 1577–1585.
- Hühns, S., Bauer, C., Buhmann, S., Heinze, C., von Bargaen, S., Paape, M. et al. (2003) Tomato spotted wilt virus (TSWV) infection of *Physcomitrella patens* gametophores. *Plant Cell, Tissue and Organ Culture*, **75**, 183–187.
- Isogai, M., Nakamura, T., Ishii, K., Watanabe, M., Yamagishi, N. & Yoshikawa, N. (2011) Histochemical detection of Blueberry latent virus in high-bush blueberry plant. *Journal of General Plant Pathology*, **77**, 304–306.
- Kamisugi, Y., Whitaker, J.W. & Cuming, A.C. (2016) The transcriptional response to DNA-double-strand breaks in *Physcomitrella patens*. *PLoS One*, **11**(8), e0161204.
- Katoh, K. & Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, **30**, 772–780.
- Koloniuk, I., Hrabáková, L. & Petřík, K. (2015) Molecular characterization of a novel amalgavirus from the entomopathogenic fungus *Beauveria bassiana*. *Archives of Virology*, **160**, 1585–1588.
- Krupovic, M., Dolja, V.V. & Koonin, E.V. (2015) Plant viruses of the Amalgaviridae family evolved via recombination between viruses with double-stranded and negative-strand RNA genomes. *Biology Direct*, **10**, 1–7.
- Lang, D., Ullrich, K.K., Murat, F., Fuchs, J., Jenkins, J., Haas, F. B. et al. (2018) The *Physcomitrella patens* chromosome-scale assembly reveals moss genome structure and evolution. *The Plant Journal*, **93**, 515–533.
- Langmead, B. & Salzberg, S. (2013) Bowtie2. *Nature Methods*, **9**, 357–359.
- Letunic, I. & Bork, P. (2019) Interactive tree of life (iTOL) v4: recent updates and new developments. *Nucleic Acids Research*, **47**, W256–W259.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N. et al. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
- Martin, R.R., Zhou, J. & Tzanetakis, I.E. (2011) Blueberry latent virus: an amalgam of the Partitiviridae and Totiviridae. *Virus Research*, **155**, 175–180.
- Maumus, F., Epert, A., Nogué, F. & Blanc, G. (2014) Plant genomes enclose footprints of past infections by giant virus relatives. *Nature Communications*, **5**, 4268.
- Meyberg, R., Perroud, P.F., Haas, F.B., Schneider, L., Heimerl, T., Renzaglia, K.S. et al. (2020) Characterisation of evolutionarily conserved key players affecting eukaryotic flagellar motility and fertility using a moss model. *New Phytologist*, **227**, 440–454.
- Mushegian, A., Shipunov, A. & Elena, S. (2016) Changes in the composition of the RNA virome mark evolutionary transitions in green plants. *BMC Biology*, **14**, 68.
- Naccache, S., Greninger, A., Lee, D., Coffey, L., Phan, T., Rein-Weston, A. et al. (2013) The perils of pathogen discovery: origin of a novel parvovirus-like hybrid genome traced to nucleic acid extraction spin columns. *Journal of Virology*, **87**, 11966–11977.
- Nibert, M.L., Pyle, J.D. & Firth, A.E. (2016) A +1 ribosomal frameshifting motif prevalent among plant amalgaviruses. *Virology*, **498**, 201–208.
- Park, D., Goh, C., Kim, H. & Hahn, Y. (2018) Identification of two novel Amalgaviruses in the common Eelgrass (*Zostera marina*) and in silico analysis of the Amalgavirus +1 programmed ribosomal frameshifting sites. *The Plant Pathology Journal*, **34**, 150–156.
- Perroud, P.F., Cove, D.J., Quatrano, R.S. & McDaniel, S.F. (2011) An experimental method to facilitate the identification of hybrid sporophytes in the moss *Physcomitrella patens* using fluorescent tagged lines. *New Phytologist*, **191**, 301–306.
- Perroud, P.F., Haas, F.B., Hiss, M., Ullrich, K. K., Alboresi, A., Amirebrahimi, M. et al. (2018) The *Physcomitrella patens* gene atlas project: large-scale RNA-seq based expression data. *The Plant Journal*, **95**, 168–182.
- Perroud, P.F., Meyberg, R. & Rensing, S.A. (2019) *Physcomitrella patens* Reute mCherry as a tool for efficient crossing within and between ecotypes. *Plant Biology*, **21**, 143–149.
- Peyambari, M., Warner, S., Stoler, N., Rainer, D. & Roossinck, M. (2018) A 1,000-year-old RNA virus. *Journal of Virology*, **93**, e01188–e1218.

- Price, M.N., Dehal, P.S. & Arkin, A.P.** (2010) FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.
- Rensing, S.A., Goffinet, B., Meyberg, R., Wu, S.Z. & Bezanilla, M.** (2020) The moss *Physcomitrium (Physcomitrella) patens*: a model organism for non-seed plants. *Plant Cell*, **32**, 1361–1376.
- Roossinck, M.** (2015) Plants, viruses and the environment: ecology and mutualism. *Virology*, **479–480**, 271–277.
- Sabanadzovic, S., Valverde, R., Brown, J., Martin, R. & Tzanetakis, I.** (2009) Southern tomato virus: the link between the families Totiviridae and Partitiviridae. *Virus Research*, **140**, 130–137.
- Safari, M. & Roossinck, M.** (2014) How does the genome structure and life-style of a virus affect its population variation? *Current Opinion in Virology*, **9**, 39–44.
- Samarfard, S., McTaggart, A., Sharman, M., Bejerman, N. & Dietzgen, R.** (2020) Viromes of ten alfalfa plants in Australia reveal diverse known viruses and a novel RNA virus. *Pathogens*, **9**, 214.
- Scholthof, K.-B.** (2004) Tobacco mosaic virus: a model system for plant biology. *Annual Review of Phytopathology*, **42**, 13–34.
- Takahashi, H., Fukuhara, T., Kitazawa, H. & Kormelink, R.** (2019) Virus latency and the impact on plants. *Frontiers in Microbiology*, **10**, 2764.
- Tzanetakis, I., Sabanadzovic, S. & Valverde, R.** (2021) Amalgaviruses (Amalgaviridae). *Encyclopedia of Virology*, **3**, 154–157.
- Villamor, D.E.V., Ho, T., Al Rwahnih, M., Martin, R.R. & Tzanetakis, I.E.** (2019) High throughput sequencing for plant virus detection and discovery. *Phytopathology*, **109**, 716–725.
- Yamada, T., Onimatsu, H. & Etten, J.L.V.** (2006) Chlorella viruses. *Advances in Virus Research*, **66**, 293–336.
- Zhan, B., Cao, M., Wang, K., Wang, X. & Zhou, X.** (2019) Detection and characterization of Cucumis melo Cryptic Virus, Cucumis melo Amalgavirus 1, and Melon Necrotic Spot Virus in Cucumis melo. *Viruses*, **11**, 81.

Chapter 4.4: Complementary results and discussion

The work presented here has allowed us to detect and characterize the first virus that infects in the wild the moss *P. patens*, the amalgavirus *PHPAVI*. We could demonstrate that the virus is transmitted vertically through both the paternal and maternal lineages. A similar result, has been observed in the master dissertation of Sevgi Coskan on the transmission of the Amalgavirus Southern Tomato Virus (Coskan et al., 2016), where they observed that the virus could also be transmitted through the paternal and maternal lineages to the progeny.

As discussed on the article, in one of the crosses that we established between Gransden and Kaskaskia we observed an uneven distribution of the virus in the progeny, where in some cases the virus was not present. All the progeny was originated from different spores that were formed in a single sporophyte and therefore from a single zygote. The virus consequently had to be present in the early stages of the embryo and be lost in some cells during the formation of the spores. As said, this could be due to stochastic reasons, such as the low concentration of the virus in the first cells after the cross, leading to an heterogenous distribution in the progeny, or that in some genetic backgrounds from the cross the virus may not be allowed to be maintained. To check which of the two hypothesis is true we would need to grow more spores from single capsules and check the presence of the virus in the progeny. If the second hypothesis is true and there is a genetic background that does not allow the maintenance of the virus, the study of the different crosses could allow the identification of key components involved in the maintenance or silencing of the virus.

It should be noted that this virus is a dsRNA virus and that dsRNAs are one of the main targets of the silencing machinery (Béclin et al., 2002). Therefore, the study of the backcrosses of these lines could allow the identification of new mechanisms involved in gene silencing and allow the understanding of how the Amalgaviruses are maintained without being silenced in the progeny.

Finally, we also tried to explore if we could generate an infective clone of the virus in *P. patens*. With this goal we designed two different constructs: a fusion of the virus with a

Green Fluorescence Protein (GFP) to try to track the virus during the development of the moss and a virus with slight changes in the nucleotide sequence that does not alter the aminoacid sequence of the encoded proteins. In both constructs we used a 35S promoter prior to the sequence of the virus in order to induce the transcription (Figure 54).

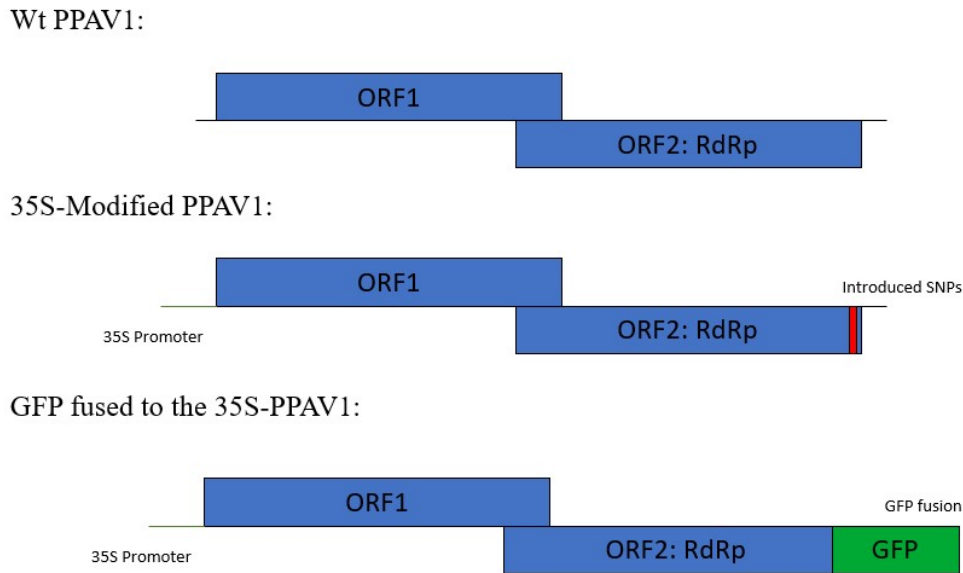


Figure 54: At the top structure of the PPAV1 wt virus, where the two blue boxes represent the two Open Reading Frames (ORF) of the virus. After this, the two different constructs that were developed in the lab. In the mid of the figure, the construct that we modified that had slight changes in the nucleotide sequence located at the second ORF that does not lead to any aminoacid change (marked in a red box). To induce the expression, we fused the sequence of the virus to a 35S promoter (in green). At the bottom the second construct that we developed with a fusion of a GFP at the end of the second ORF in the same frame.

We transformed the two constructs independently to the Gransden accession and although we could get clones that were transiently resistant to the antibiotic present in the plasmid, we failed to detect the presence of any of the modified virus in these lines. We could not observe the presence of the constructs for any of the two strategies, neither by fluorescence nor by PCR. As said previously, from the crosses between Gransden and Kaskaskia we observed that in some progenies of the cross the virus was absent and that there could be the possibility that the virus cannot be propagated in the Gransden accession. We repeated the transformation in Reute obtaining the same result without any transformant line that contained the modified virus but observing the presence of the Wt virus. One hypothesis that could explain the observed result is that the introduced

mutations affect the stability of the virus, not being able to replicate although the construct could be expressed. Another possibility is that the main path of transcription of these viruses is not through the RNA polymerase II and therefore the induction using the 35S promoter may not lead to a viable transcript.

As there were RNAseq libraries produced from *P. patens* Reute Wt samples using a polyA purification (therefore the transcripts were produced by the RNA polymerase II) and for the same samples they also produced RNAseq libraries using a depletion of the total RNA present for the samples, we could compare the presence of the virus for the two different RNA purifications procedures (Table 14).

Table 14: Read counts and Transcripts per Million (TPM) normalized counts of the different RNAseq libraries mapped to the Amalgavirus sequence of juvenile gametophores with libraries produced with a polyA purification and a total RNA purification (Total RNA).

Sample	Tissue	Replicate	n° Reads	TPM
BBTWW	juvenile gametophore (polyA library)	Rep 1	6.00	0.35
BBTWY	juvenile gametophore (polyA library)	Rep 2	10.00	0.63
BBTWX	juvenile gametophore (polyA library)	Rep 3	5.00	0.28
BXHHX	juvenile gametophore (Total RNA)	Rep 1	1574.00	396.38
BXHHZ	juvenile gametophore (Total RNA)	Rep 2	1898.00	472.52
BXHHY	juvenile gametophore (Total RNA)	Rep 3	1491.00	243.37

We observed that the virus is detected in high quantities in the Total RNA libraries but lowly expressed or not detected in the polyA libraries. This could explain why we did not detect the modified virus in the transformed lines with the virus constructs as we used a 35S promoter that is transcribed by the polymerase II and probably is not the best promoter based on the above data to transcribe and introduce the modified virus.

To conclude, through this work we have identified the virus PHPAV1, the first known virus known to infect *P. patens*, that apparently does not cause any symptom to the infected mosses. Moreover, we could get insights into the distribution of the virus across the population of the moss collected around the world, that it can be found during all the life cycle of the moss and that the virus can be transmitted vertically through the maternal and the paternal lines.

GENERAL DISCUSSION AND
FUTURE PRESPECTIVES

GENERAL DISCUSSION AND FUTURE PRESPECTIVES

General discussion

Mobile genetic elements, and in particular TEs, are an important source of genetic variability, shaping eukaryotic genome structures. In this thesis we have developed different approaches to study the impact and dynamics of mobile genetic elements in the model organism *Physcomitrium patens*.

In the first chapter we identified the main problems that may arise when working with short-read libraries to identify TIPs from DNA-resequencing data and to determine the transcription of TEs from short reads RNA-seq data. We have evaluated and developed tools and methods that allowed us to analyze the transcriptional and transpositional landscape in different organisms. In this thesis we have combined the two approaches to identify the transcriptional and transpositional landscape of *P. patens* (Vendrell-Mir et al., 2020). These methods are also being successfully used to perform other projects in our group in other plant species. such as *Oryza sativa*, *Arabidopsis thaliana* or *Prunus dulcis* and *Prunus persica* (de Tomás et al., 2022).

Through the work done in *P. patens* we have identified four families of LTR-RT (RLG1, RLG2, RLC4 and RLC5), a LINE and two DNA transposon families that are transcriptionally active. The copies more similar to the detected transcripts were young TEs and in some cases were polymorphic in the population of resequenced samples of *P. patens*, proving that there has been recent TE activity in the moss for these families. Although we could not detect transposition events during the last period of 20 years of the moss maintained in the lab.

To try to understand the impact that the TEs may have in the maintenance of structure of the genome of *P. patens*, especially the role of the LTR-RT family RLG1. In this thesis we have developed two different approaches to eliminate RLG1 rich regions by using the CRISPR/Cas9 system. We could not remove a substantial part of these heterochromatic regions, but we could prove that producing nonselective cuts at RLG1 elements have a strong deleterious effect in *P. patens*. This could be due to the removal of these sequences or to the consequences of introducing a high number of DSBs in the genome. Despite that, some clones were able to regenerate. In these clones we observed only a slight decrease, or even an increase in the number of RLG1 elements. In some cases, we also observed development phenotypes in these clones when compared to the untreated

samples. To try to dig deeper into the effect of these eliminations across all the genome we would need to sequence the genomes of these clones and compared it to the untreated genome samples. However, our results suggest that this is probably not the best strategy to study the potential role of the heterochromatic RLG1 elements in the genome of *P. patens*.

As an alternative, we developed a second strategy that consisted in the removal of specific RLG1 islands from a single chromosome. We were able to delete two of these islands, island D of 49 kb and island O of 160 kb, although the replacement of the deleted region with a selective marker did not result in the simple insertion of the marker sequence and probably consists of multiple insertions of the marker and the plasmid sequences. Surprisingly, in spite of removing an heterochromatic region of an important length, they had a small effect on the expression of the neighboring genes.

At the moment that this study was conceived there was only the chromosome scale of this bryophyte. The publication of more genome assemblies of bryophytes, and in particular of mosses or even other strains of *P. patens* could allow in a near future a comparison of the genome architecture to study indirectly the impact over these TEs heterochromatic regions on the global genome structure. We also expect that the improvement of the genome editing technologies will result on an increase of the efficiency of the production of targeted deletions. This will facilitate the study of these heterochromatic islands using a similar approach to the ones described on this dissertation.

Concerning the impact of the RLG1 elements on genic regions, we selected three polymorphisms located close to genes of potential interest. Of the three, the one potentially most interesting was the RLG1 TIP between Gransden and Kaskaskia located at the promoter region of the gene Pp3c14_9040V3.1, a transcription factor. Our mutagenic analysis of Pp3c14_9040V3.1 suggests that it is an essential gene, as the host organism cannot survive when a KO is produced. We could also confirm that the presence of the RLG1 TIP in the promoter region alters the expression of the gene. The preliminary analyses performed within the framework of this thesis did not allow us to determine what could be the phenotypic consequence of this change in expression. More work will be needed to phenotype the different clones in the lab. However, it is possible that the RLG1 polymorphic insertion has evolutionary consequences on the fitness and survival of the different accessions in their environments which may not be easy to replicate in the laboratory.

In the course of this thesis, I have studied the impact that heterochromatic and euchromatic insertions of the RLG1 element, the most prevalent TE in *P. patens*, may have in this genome. Surprisingly, although we managed to eliminate RLG1 heterochromatic islands of as much as 160 kb, and we eliminated and introduced RLG1 elements sitting very close to an essential gene, we did not detect a major phenotypic consequence of these insertions. As already discussed, most of the work that has showed the important phenotypic consequences of TE insertions started from the analysis of an obvious phenotype (e.g. fruit color in different crops), and not from the characterization of a TE or a polymorphic insertion. It may well be that most TE insertions, even those located within or very close to genes, may have a weak effect on gene expression and the phenotype, which only manifests in particular environmental conditions. In fact, this will ensure the maintenance of active TEs and the possibility to adapt to rare environmental conditions.

I had also the chance to try to identify other mobile genetic elements, a part from TEs, that could be active in *P. patens*. In particular, I looked for the presence of viruses on the moss, identifying the first virus naturally infecting *P. patens*. This virus, PHPAV1, belongs to the *Amalgaviridae* group of viruses, which are endogenous viruses that, as transposons, are transmitted vertically. The analysis of this virus and its transmission through the parental and maternal lineages, opens several interesting questions. We identified the virus in 3 different accessions, Reute (Germany), Kaskaskia (USA), and Lviv (Ukraine), but not in other accessions located geographically close, suggesting that there may be host genetic determinants for its maintenance. This was also the conclusion we reached when analyzing the transmission of the virus in crosses between accessions that contain the virus with others that do not. Analyzing these crosses in more detail may also allow the study of how these endogenous viruses manage to scape host control and be maintained in the germ line.

Future perspectives

Studying the preference of Integration of LTR-Retrotransposons

As outlined in the introduction, the distribution of TEs in the genome is the result of a balance between the preference of insertion of the some TEs and the posterior process of selection of these insertions (Sultana et al., 2017). In general, in plants the LTR-RTs of the Copia superfamily are found close to genes while the transposons of the Gypsy superfamily are found in heterochromatic regions. A typical example of the former is the Copia LTR-RT of tobacco, Tnt1, that has been identified to target genic regions in its host and when introduced in different species (Courtial et al., 2001; Vives et al., 2016; Kwon et al., 2019). A typical example of the later are the centromeric retrotransposons of grasses (Sharma & Presting, 2014). However, the preference of insertion may vary extensively between different families of transposons within the two superfamilies. For example, the Tall Copia LTR-RT from *Arabidopsis lyrata* targets the centromeric repeats (Tsukahara et al., 2012) while the most similar transposon in *A. thaliana* ATCOPIA93 targets genic regions, in particular regions enriched in the histone mark H2A.Z (Quadrana et al., 2019).

Through the work done during this thesis we have studied two different LTR-RT from *P. patens* that have completely different distributions among the *P. patens* genome. RLG1 is found mostly accumulated in the heterochromatic regions while RLC5 is found mostly in a single position that are the putative centromeric regions. One of the questions our group would like to study is how these, and other, LTR-RTs are target specific regions of the genome for integration. A close inspection of the sequence of RLG1 showed that at the end of the 3' integrase of the RLG1 there is a chromodomain. These chromodomains in other species have been identified as the responsible of the targeting of the Gypsy LTR-RT to the heterochromatic regions (Gao et al., 2008). In the case of the RLG1 elements, to try to identify if the transposons are preferentially integrating into these heterochromatic regions, Cristina Vives, a former PhD student in the lab, and the former master student Pedro Pastor designed a non-autonomous version of the RLG1 elements. This strategy was based on the approach previously developed by the group to elaborate a non-autonomous elements from the tobacco Tnt1 LTR-RT (Vives et al., 2016). They build a non-autonomous copy, named miniRLG1 containing a nptII gene in opposite direction to the transcription of the transposon, containing an intron oriented in the same direction than the transcription of the transposon, truncating the nptII gene. The nptII gene

confers resistance to Kanamycin or Geneticin. In the case that there is a transposition event of the construct the intron will be eliminated in the new copies, and the clones could be selected by the presence of Geneticin Resistant clones. Moreover, a 7 nt changes were introduced at the 5' LTR at the U5 region to be able to differentiate the new copies. We expected that as the transposon is expressed during the development of *P. patens* the expressed copies of the genome could deliver the machinery necessary for the mobilization of the non-autonomous copy. During the thesis, we performed three independent transformations using this construct, but we could never identify new transpositions of this miniRLG1 construct. Although this could be due to the non-autonomous copy that we build was not functional, as seen in chapter 2 we could not either identify new transpositions events in samples that have been maintained in the laboratory for the last 20 years.

Due that we could not identify new transpositions events to solve this question we designed an alternative approach. To try to identify if the RLG1 are targeting the heterochromatic regions and if this pattern of integration is maintained in other species, together with the former master student Marc Pulido we identified different RLG1 copies that have recently been active, according to their transcription, the identity of their LTRs, and their polymorphic TE absence in the closest accession Reute. We have cloned one of these RLG1 elements and replaced part of the 5' LTR with a 35S enhancer to induce the expression in other plant species. A similar strategy has been previously done to mobilize the transposons Tnt1 of tobacco (Mhiri et al., 1997) and Ty1 of yeast (Curcio & Garfinkel, 1991). This strategy, apart from inducing the expression of the transposon, allow the identification by PCR of new transposition events, as the 5' LTR in the new copies will be fully reconstructed, being possible to identify these events through a PCR from the 5' LTR to the coding region of the transposon. We designed and built the constructs to transform *A. thaliana*. Plant transformations have been done using these constructs but at the time that this dissertation is being finalized we have still not verified if the transposon has been mobilized or not. A similar approach is being designed to mobilize the *P. patens* RLC5 transposon into *A. thaliana* and check whether it goes to the centromeric regions in this species or not.

In our group we are also interested into the study of the mechanisms related to the preference of integration of LTR-RT of other species, such as the Tnt1 LTR-RT Copia element from *Nicotiana tabacum*, which mechanisms of transposition have been studied

in the lab for a long time (Beguiristain et al., 2001; Gonzalvo, 2006; Hernández-Pinzón et al., 2009, 2012; Vives et al., 2016) or the LTR-RTs of *Oryza sativa*, work that has been recently started in our group.

In the case of the Tnt1 LTR-RT the former PhD. Student Cristina Vives with the help of the former Master student Pedro Pastor and myself, during the time that I did my bachelor practices in the lab, performed two transformations of the miniTnt1 system (Vives et al., 2016) into *A. thaliana*. Transforming in one case with the miniTnt1 system complemented with the Wt proteins of Tnt1 and as a control with the miniTnt1 system with a modified proteins of the Tnt1 transposon that had mutations in the three aminoacids that forms the catalytic domain of the integrase, responsible of producing the DSB into the genomic DNA to integrate the transposon. Interestingly, when compared to *P. patens* (Vives et al., 2016), in *A. thaliana* we observed that we had a few integrations when using the miniTnt1 + the proteins with the mutated catalytic domain of the integrase that were never observed in the case of *P. patens* (Table 15).

Table 15: Number of clones obtained by transforming with the miniTnt1 system combined with the Wt proteins of Tnt1 or the proteins with the mutations over the catalytic domain of the integrase of Tnt1 in *A. thaliana* and on *P. patens*.

Constructs	N° lines with transpositions <i>Arabidopsis thaliana</i>	N° lines with transpositions <i>Physcomitrium patens</i>
Tnt1 Wt proteins + miniTnt1	130	163
Tnt1 mutated catalytic domain integrase + miniTnt1	10	0

We thought about two hypothesis that could explained the observed results: Or the introduced mutations at the catalytic domain were not producing a completely defective protein, and there was still some catalytic activity, or the miniTnt1 elements were able to integrate into DSBs present in the genome. This phenomenon has been observed in other species such as in yeast (Moore & Haber, 1996) or in mammals (Ono et al., 2015).

To test this second hypothesis we performed transformations in *P. patens* using the miniTnt1 with the two constructs of proteins (Wt and Mutated integrase) in lines mutants for RAD51 (Schaefer et al., 2010) and in Wt lines. Our collaborators, the group of Fabien Nogué, observed that these lines had more double strand breaks when compared to the wt lines. After performing the transformation, we observed for the first time clones that had

the transposon integrated in *P. patens* when using the Tnt1 proteins that had the catalytic domain mutated, but only on the mutated RAD51 background (Table 16).

Table 16: Number of clones obtained by transforming with the miniTnt1 system combined with the Wt proteins of Tnt1 or the proteins with the mutations over the catalytic domain of the integrase of Tnt1 in Wt *P. patens* and in Δ Rad51 *P. patens*.

Constructs	Wt <i>P. patens</i> Line	ΔRad51 <i>P. patens</i> Line
Tnt1 Wt proteins + miniTnt1	12	32
Tnt1 mutated catalytic domain integrase + miniTnt1	0	6

This suggest that the transposon could indeed integrate into the DSBs without the need of a functional catalytic domain. Despite that, the number of obtained clones was relatively low. Further transformations and further characterizations of the integration sides (such as the presence of target sides duplications) will be required to confirm the hypothesis and conclude how frequent is this phenomenon and which are the mechanisms behind these transposition events.

During these years I have been involved in different projects dealing with the dynamics and impact of TEs in plants. In particular, I have been involved in the study of the dynamics and impact of TEs on 1059 varieties of rice (Castanera et al., 2021). This work gave us the possibility to compare the different integration pattern of the different LTR-RT of the copia family that we annotated over these 1059 varieties. We also aligned and compared the different LTR-RT copia families using the aminoacid sequence of the Integrase and the Reverse transcriptase. We observed that there are families of LTR-RT that are closely related that have different TIPs distributions among the genome (Figure 55).

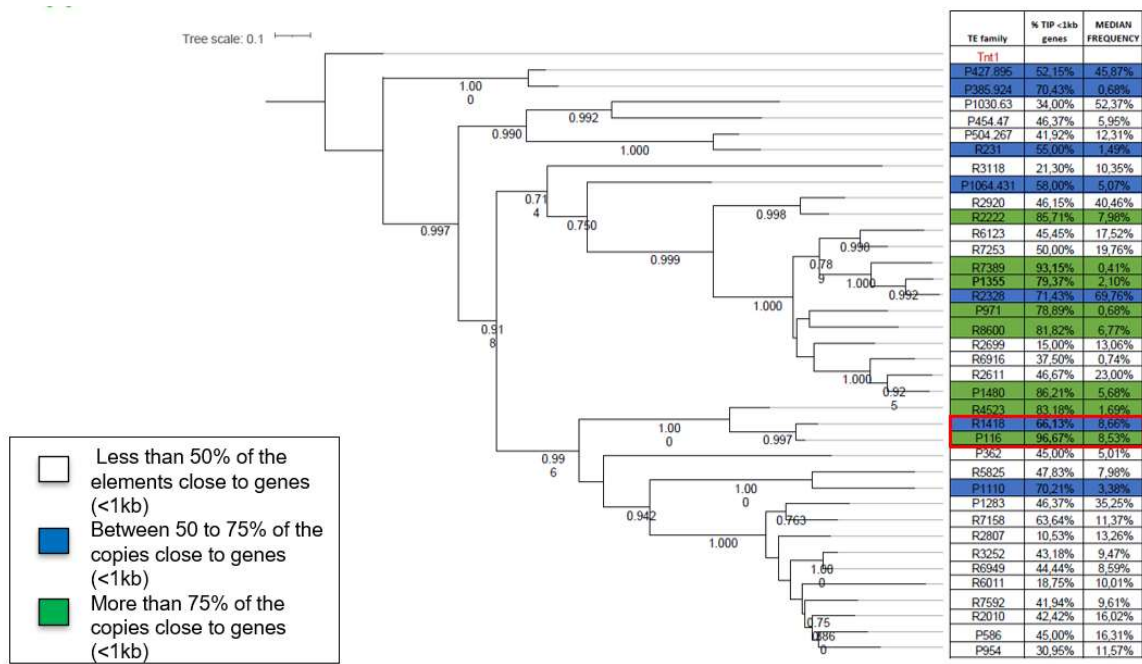


Figure 55: Phylogenetic tree of the different LTR-RT of the copia family using the alignment of the Integrase and the reverse transcriptase of the different LTR-RT Copia families using the Tnt1 TE as an outgroup. In the table as a first column the name of the TEs, the second column the percentage of TIPs found at less than 1 kbp from the genic regions and the third column the median frequency of the TIPs detected in the population. Marked in red in the figure an example of two TE families that have similar TIP frequencies in the population but different patterns of integration.

These differences could be due to differences in the activity of the transposons over time, or due to differences among the preferences of insertions of the transposons. For this reason, we compared transposons that had similar TIP frequencies in the population. We observed that even among closely related families with similar TIP frequencies there are differences in distribution. After comparing the aminoacid sequence of the recent copies of these families, we observed that the differences between these families were in a variable sequence at the end of the integrase protein (Figure 56).

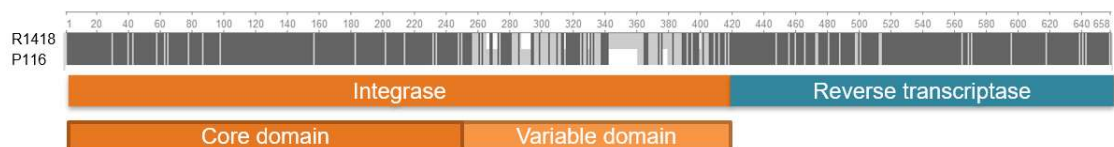


Figure 56: Protein alignment of the Integrase and Reverse transcriptase of two Copia families of LTR-RTs (R1418 and P116) that are closely related (marked in a red box in the previous figure) and have different pattern of integrations in the population. At the top, in dark grey the sequences that are highly conserved in the alignment, in light grey the sequences that are poorly aligned between the sequences. In white the GAPS in the alignment. Under this, represented in Orange the sequence belonging to the Integrase sequence, and

in blue the sequence corresponding to the Reverse Transcriptase. We observe that the Core domain of the integrase and the reverse transcriptase are highly conserved, while the end of the integrase is not conserved between this two Copia LTR-RT families.

In *Saccharomyces cerevisiae* it has been proved that a sequence located at the end of the integrase of the Ty1 TE interacts with the AC40 protein of the RNA polymerase III complex, targeting the insertion of the TE to the promoter region of the genes transcribed by these complex (Bonnet & Lesage, 2021). They proved that by eliminating these sequence of the Ty1 TE it loose the specificity of integration and after integrating this sequence at the end of the integrase sequence in the Ty5 TE, they observed that this TE acquired the same pattern of integration than the Ty1 TE (Asif-Laidin et al., 2020).

We designed a similar approach by introducing the variable sequences detected between the different LTR-RT Copia families of *Oryza sativa* to the end of the Tnt1 integrase sequence, using the miniTnt1 system to check whether the introduction of these sequences changes the pattern of integration of the miniTnt1 TE.

We started by replacing the Tnt1 3'Integrase sequence by the integrase variable sequence of the Copia transposon R7389 of rice that is one of the most active copies in the population of rice. We also designed a Tnt1 that had a deletion at the end of the 3' Integrase with the goal to check if the removal of this sequence alters the insertion pattern of the miniTnt1 system. At the moment that this dissertation is being finished we managed to mobilize the miniTnt1 combined with the use of the wt Tnt1 proteins, the Tnt1 proteins with the deletion at the 3' integrase sequence and, with less efficiency, the Tnt1 with the R7389 rice Int 3' sequence. The current efforts of the laboratory are into increasing the efficiency of transposition of the miniTnt1 system to have enough insertions to compare if there are changes in the integration pattern.

The study of the mechanism's differences between closely related LTR-RT families could help to understand how LTR-RT contributes to shape the different genomes architectures between the population of a given species, being a key contributor to the evolution of the genome's structures. Moreover, a better understanding of the interaction between the transposons proteins with the host proteins has also a potentially biotechnological interest. For example, these interactors could be fused to a gene editing system (such as the CRISPR/Cas9) to target the insertion of a desired DNA sequence to a given place of the

genome, using LTR-RT as vectors. The engineering of LTR-Retrotransposons could also be transmitted to the closely related LTR-Retroviruses to produce LTR-Retroviruses vectors to deliver DNA to specific regions of the genome, avoiding the mutagenic effects over genic regions.

CONCLUSIONS

CONCLUSIONS

- The detection of polymorphic TEs insertions and TEs transcription through next generation short-read sequencing approaches requires dedicated strategies to accurately assess both processes.
- We have developed tools and approaches that allow for a reliable detection of TIPs and an accurate measurement of TE expression.
- There are different families of retrotransposons (RLG1, RLG2, RLC4, RLC5 and the LINE-2 family) and DNA transposons (PpTc1 and PpTc2) that are transcriptionally and transpositionally active on *Physcomitrium patens*.
- The RLG1 LTR-RT family of *Physcomitrium patens* may have important impact in the maintenance of the structure of the genome and although is mainly found in heterochromatic regions it can alter the expression of genes through their movement.
- *Physcomitrium patens Amalgavirus 1* is an endogenous virus that infects different accessions of the moss *Physcomitrium patens* around the globe, being transmitted vertically through both maternal and paternal lineages to the progeny.

MATERIALS AND METHODS

MATERIALS AND METHODS

Chapter 1: Development of bioinformatic tools to identify transposable elements mobilization and transcription

Chapter 1.3: Comparison of different available tools to detect Transposon Insertion polymorphisms using short-read data. Supplementary results:

LorTE: Long reads TIP detection

We have run LorTE (Disdero & Filée, 2017) (20X coverage), a PacBio structural variation caller specifically developed to identify TE presence-absence polymorphisms on the PacBio reads that were used to assemble MH63 genome (SRA:SRR5456657). We have run LorTE with a E-value for the BLASTn and Megablast used by the tool of 1e-40 and a minimum read coverage of 10. The results were intersected with the curated library of polymorphic LTR-RT that we generated between MH63 and Nipponbare.

Chapter 1.4: Detection of Transposable Element transcription from short-read data

Libraries used for the analysis:

The RNAseq data used for this study was obtained from the *P. patens* Gene Atlas library (Perroud et al., 2018). For the analysis done using Tetranscripts and TEtools we used the libraries from the Heat shock experiment (SRA: SRX712882, SRX712881, SRX712747, SRX712763, SRX712750).

The libraries used for the TE transcriptome assembly approach were a selection of different development conditions and stresses from the *P. patens* Gene atlas approach. (SRX712744, SRX712739, SRX712754, SRX712856, SRX712865, SRX712884, SRX713928, SRX713945, SRX713938, SRX713921, SRX713947, SRX713922, SRX712755, SRX712749, SRX712748, SRX712855, SRX712857, SRX712859, SRX712878, SRX712862, SRX712863, SRX712864, SRX712877, SRX712867, SRX712868, SRX712858, SRX712747, SRX712763, SRX712750, SRX712736, SRX712756, SRX712737, SRX712743, SRX712753, SRX712882, SRX712881, SRX712879, SRX713924, SRX713933, SRX713927, SRX713907, SRX713936, SRX713943, SRX712760, SRX712765, SRX712758, SRX713937, SRX713935,

SRX713942, SRX713920, SRX713919, SRX713911, SRX712880, SRX712874, SRX712876, SRX712873, SRX712883, SRX712877, SRX712866, SRX712870, SRX712871).

All the raw reads used in this study were quality trimmed using bbduk (<https://sourceforge.net/projects/bbmap/>). Before mapping the reads using the three different approaches, we discarded all the reads mapping to the plastid genome (chloroplast or mitochondria) and the rRNA.

TE clustering

All the analysis was done based on the TE annotation described in Lang et al., 2018. LTR-RTs families of the annotation were further classified in clusters. To do that we classified all the LTR-RT to clusters using Silix (Miele et al. 2011) with an 80% of homology over 80% of the length of the sequence. We consider as a cluster each group that had at least three TE copies. LTR-RT consensus sequence of each cluster was obtained by aligning all the sequence of each cluster using Mafft (Kato & Standley, 2013) with the default parameters. The poorly aligned sequences were trimmed using Trimmal (Bushnell, 2015) and a consensus of each sequence was build using the cons tool of the EMBOSSpackage (Rice et al. 2000).

A phylogenetic tree was build based on the alignment of the reverse transcriptase sequence of each consensus separated between Copia and Gypsy TEs using as an outgroup the *Oryza sativa* LTR-RTs Tos17 as a Copia outgroup and the LTR-RTs CRR as a Gypsy outgroup.

TETranscripts pipeline:

To launch TETranscripts, RNAseq data was aligned to the V3.3 *P. patens* genome (Lang et al., 2018) using STAR (Dobin et al., 2013) with the default parameters except for the parameter `--outFilterMismatchNoverLmax` that was set to 0.04.

TETranscripts (Jin et al., 2015) was launched with the default parameters and using different TE annotation files, depending on if it was launched with the classifications done in Lang et al.,2018, if it was done using the custom LTR-RT clusters.

The command line used was the following one:

```
Tetranscripts -tc -t Ppatens_heat_stress_rep1.out.bam Ppatens_heat_stress_rep2.out.bam Ppatens_heat_stress_rep3.out.bam -c Ppatens_control_rep1.out.bam Ppatens_control_rep2.out.bam Ppatens_control_rep3.out.bam --GTF $gene_annotation.gtf --TE $TE_annotation.gtf
```

TETools pipeline

TEtools was used following the instructions provided by the TEtools manuscript (Lerat et al., 2017) using the copies of the genome classified into the clusters that we previously defined.

To count the reads belonging to each cluster, the raw reads were aligned to the copies using Bowtie2 (Langmead and Steven L Salzberg, 2013) with the following command line:

```
bowtie2 -p 4 --time --very-sensitive -x TE_sequence.fasta.index2 --dovetail -X 273 -1 Ppatens_seq_1.fastq -2 Ppatens_seq_1.fastq -S Ppatens_seq.sam;
```

The counting was done using the first module of TEtools: TEcount with a custom rosetta file to analyze the transposable element using the copies of the previously defined clusters provided for *P. patens*.

To check the differential expressed TE families we used the second module of TEtools, TEDiff that relies on the DESeq2 package (Love et al., 2014) to estimate the differentially expressed TEs.

TE transcripts assembly

To assemble the TE transcripts, we first mapped all the reads from all the RNAseq libraries to the TE annotation provided in the *P. patens* V3.1 genome using Bowtie2, with the previous provided command line. All the reads that mapped were extracted using samtools (H. Li et al., 2009) samtofasta. All the remaining reads were combined and assembled using Trinity (Grabherr et al., 2011) using the following command line:

```
Trinity --seqType fa --single TE_reads.fa --CPU 8 --max_memory 60G
```

To identify to which copy or TE family each assembly transcript corresponded all the assembled contigs were aligned to the TE copies using BLASTn with a e-value cutoff of 10^{-5} . For LTR-RT transcripts we kept all the alignments with a minimum length of 1000 bp, we then manually curated all the LTR-RT transcripts that were longer than 1000 bp

and all the other TE family transcripts. To curate them we looked for TE coding domains using CD-search (Marchler-Bauer et al., 2015). Finally, we curated the remaining contigs that had TE coding domains and that were not chimeric or truncated copies.

To estimate the expression of the different selected TE assembled contigs, RNA-seq data were mapped to the assemblies using Bowtie2. To count the number of mapped reads we counted only those reads that belonged to sense expression. Expression data was normalized by counting the total number of mapped reads to each transcript normalized by the length of the transcript (Kbp) and the total number of raw reads per each library before mapping to the genome and the differentially expressed clusters were obtained using Deseq2 (Love et al., 2014) following the same pipeline used by TEtools (Lerat et al., 2017).

Chapter 2: TEs dynamics in *Physcomitrium patens*

2.3: Complementary results and discussion

To assess if there has been recently mobilization of TEs we used Illumina short-reads of samples maintained in the lab that were stored during the years 2007, 2011, 2016 and 2018 (Bessoltane et al., 2022) facilitated by Fabien Nogué group.

We mapped the short-read data Paired-end reads to the reference genome using BWA SW (H. Li & Durbin, 2010). TE insertions were detected using PoPoolationTE2 (Kofler et al., 2016) using the separate mode with the default parameters. We filtered for a frequency of the insertions in the population of 0.3 from the output of PopoolationTE2 and discarded all these insertions that were intersecting with a TE in the reference genome annotation (Lang et al., 2018). The remaining results were filtered by manually inspection of the alignments to the reference genome using IGV (J. T. Robinson et al., 2011).

Chapter 3: Impact of TEs in *Physcomitrium patens* genome

Chapter 3.3: Impact of Transposable Elements into the structure of *Physcomitrium patens* genome

Non-selective elimination of RLG1 elements using CRISPR/Cas9

gRNA design to target RLG1 elements

A fasta file with all the annotated RLG1 elements in the genome was extracted using bedtools(Quinlan & Hall, 2010) bedtofasta from the Main TE annotation (Lang et al., 2018). To look for all the possible gRNAs targeting the RLG1 elements a script was designed by Jordi Morata looking for all the possible 20 nucleotide+ NGG sequence in the set of RLG1 and ordering them by the number of times that appeared over all the list of RLG1 elements. We looked for all the possible off targets by using BLASTn-short (BLAST, 2009) using all the different and using bedtools intersect (Quinlan & Hall, 2010) to look for intersected positions between the possible gRNAs matching sequences to the positions of to the genes and other TE families. To predict the efficiency of the CRISPR/Cas9 system using these gRNAs we used the CRISPOR software (Concordet & Haeussler, 2018).

Plasmids cloning

The sequence of the gRNA containing the *P. patens* U6 promoter the selected RLG1 targeting gRNA sequence and the gRNA scaffold was synthesized by IDT with the attB1 and attB2 flanking sequences. This fragment was cloned to the plasmid pDONR207 using the Gateway™ BP reaction in a plasmid named as pENTRY-PpRLG1#1. Finally, the construct was cloned to the plasmid pLand#1-Sp-Cas9 by digesting the plasmid pENTRY-PpRLG1#1 using the enzymes BsiWI and XbaI and the destination plasmid with XbaI and Aac65I, to obtain the final plasmid pLand#1-PpRLG1#1-Cas9 that contains the gRNA cassette targeting the RLG1 locus, the CRISPR/Cas9 and a nptII R gene cassette.

To replace the RLG1 elements affected by the CRISPR/Cas9 targeting. A template was synthesized by Twist bioscience and cloned to a plasmid named as pAM1 containing a

homologous sequence of part of the LTR of the RLG1 elements and containing two unique restriction sites (Sall and NotI) to clone the HygR gene from the plasmid BZRf. We cloned the HygR gene to pAM1 by digesting with Sall and NotI the plasmid pAM1 and ligating to the product of the restriction of the plasmid BZRf with Sall and NotI to generate the final plasmid pAM2.

The other plasmids used to perform the transformations targeting the APT gene were facilitated by Fabien Nogué group, using the plasmids p164 containing the gRNA that targets the APT gene and the p165 containing the CRISPR/Cas9 system.

Plant material

To perform all the experiments, we used the *P. patens* Gransden accession. The plants were grown at 24°C with a cycle of 16 h of light (quantum irradiance of 60–80 $\mu\text{mol m}^{-2} \text{s}^{-1}$ light) and 8 h of darkness on plates containing BCDA medium for propagation and BCD or BCDA medium for phenotyping (Cove et al., 2009).

Plant transformation

Moss transformation were transformed with 20 μg of the pLand#1-PpRLG1#1-Cas9 plasmid following the protocol described in (Charlot et al., 2022). Explained briefly Gransden was grown over BCDA medium overlaid with a cellophane and fragmented and regenerated in a new medium every 6-7 days. After two rounds of fragmentation, the protonemata tissue was treated with DriselaseTM (Ref:85186-71-6) for 30 minutes to digest the cell wall and we purified the protoplasts by filtering and transformed 20 μg of the plasmid by PEG transformation. protoplasts dissolved in a solution of alginate were plated in a BCDA medium supplemented with mannitol and glucose and enriched in calcium overlaid in a cellophane. After one week of growth the plants were selected for the presence of the plasmid transiently in the cells by transferring the cellophane to a new BCDA medium containing 50mg/L of G418. The clones that were able to grow after one week of growth were individualized and checked for the presence of stably resistant clones to G418 and grown at the same time in BCDA without any antibiotic.

The transformation done targeting the RLG1 elements and trying to replace the RLG1 elements by the DNA template with the HygR gene was performed using the same protocol described previously but using instead the plasmids p165 (CRISPR/Cas9, pAM2 (HygR RLG1 recombination template) and the plasmid pENTRY-PpRLG1#1. The samples were selected transiently for the resistance to Hygromycin (at 20mg/L) and to look for stably resistant clones, the clones were moved permanently to BCDA medium containing the same antibiotic at the same concentration.

The last described transformations were performed by doing two times two transformations in parallel following the same protocol but using in one case the plasmids p164 (gRNA targeting the APT gene) and p165 (CRISPR/Cas9) in one transformation and in the other the same plasmids adding the plasmid pENTRY-PpRLG1#1 that contains the gRNA targeting the RLG1 elements. After one week of regeneration in BCDA supplemented with mannitol and glucose the cellophane was transferred to BCDA medium with 2FA at a concentration of 10 μ M to select the clones that had the gene APT edited, estimating in each transformation the number of regenerated protoplasts after one week of growth and the number of resistant clones to 2FA in each transformation.

DNA extraction:

All DNA extractions were performed from tissue of protonemata or gametophores frozen in liquid nitrogen and stored at -80°C. We followed an adapted protocol of DNA extraction based on Doyle & Doyle, 1987. The extraction buffer was prepared using 0.16 g CTAB, 2.24 ml 5M NaCl, 800 μ l 1M Tris pH8, 320 μ l 0.5 M EDTA and 16 μ l β -mercaptoethanol dissolved in 4.624 ml of double distilled water. To extract DNA of each sample, the buffer was heated at 60°C before using it. 50 to 100 mg of tissue frozen sample was powdered in a mortar and transferred to a 1.5 ml tube. 600 μ L of the buffer were immediately added and the tubes were kept at 60°C for 40 minutes mixing gently every 10 minutes. After this the samples were treated with RNaseH for 30 minutes at 37°C. 600 μ l of Chloroform-isoamyl alcohol (24:1, v:v) were added to the tube, mixed and centrifugated for 15 minutes at room temperature at 13.000 rpm. The supernatant was transferred to a new 1.5 ml tube containing 600 μ l of cold isopropanol. The tube was mix by inversion and centrifugated at 13.000 rpm at room temperature for 30 minutes. The

supernatant was discarded and 200 μ L of 70% of Ethanol was added, the tubes were centrifuged at the same speed for 10 minutes. All the supernatant was discarded. The remaining pellet was dried at room temperature and 50 μ L of TE buffer was used to resuspend the DNA. The DNA was resuspended overnight in the fridge at 4°C, quantified using the Thermo Scientific™ Nanodrop™ and stored at -20°C.

PCR over RLG1 elements

The PCR was performed using the NEB™ polymerase LongAMP® Polymerase (New England BioLabs™, REF #M0323S) according to manufacturer's instructions with a Tm of 52°C and an elongation time of 8 minutes. The PCRs were done in a total volume of 20 μ l using 20 μ M of each primer, 0.25 μ M of dNTPs and 1 μ L of the polymerase.

The primers used were the following ones (Table 17):

Table 17: Primers used to genotype different RLG1.

Name	Primer sequence	Description
oPV45	CTAATTTGGCTCTCTACTTGGTGAA	Check locus RLG1 Chr01
oPV46	GTCAAATCAAACCTCCAGCTACTA	
oPV47	CCACTTGCTCTCAAATCTCCTAATA	Check locus RLG1 Chr08
oPV48	CAAATCCCATAGAGCTAATGTCAC	
oPV49	GTATGTGGTGAGAAGAAAGGAGCTA	Check locus RLG1 Chr11
oPV50	GTTCTCTCACTCTCAAATCACTCA	
oPV51	GTTGATAAGATGTTGTTAGGCAAGG	Check locus RLG1 Chr15
oPV52	CAAGTATGAATTAAGTCGTCCAAGC	
oPV53	GTTGACTGTAATCAACGTAGAGCAA	Check locus RLG1 Chr22
oPV54	GGCTATCTATATCAGCACGGCAATA	

qPCRs quantifications of RLG1 elements:

Quantitative real-time PCR were done in 96-well plates using the Roche LightCycler II instrument. SYBER Green I Master Mix (Roche™ Applied Science, REF # 03003230001), primers at a concentration of 1 μ m and 10 ng of the DNA obtained were used for the qPCR. Each sample was run per triplicate with non-template controls. The amplification conditions used were 95°C for 5 min, followed by 95°C for 10 s, 56°C for 10 s, and 72°C for 10 s, ending with the melting curve to check the specificity of the qPCR. We used first the housekeeping gene adenine phosphoribosyl transferase (APT) (Schaefer et al., 2010) to relativize the number of copies of the RLG1 elements to a single copy gene of the genome. The primers over the RLG1 elements used were oCV5 and oCV6 (Vives et al., 2016).

The samples obtained that had the gene APT edited we followed the same protocol but using the primers over the single copy gene BRCA#2, using the couple of primers oBRCA5 and oBCRA6, to relativize the number of RLG1 copies.

Selective elimination of RLG1 elements using CRISPR/Cas9

Selecting RLG1 islands from chromosome 27

We used the published Main TE annotation (Lang et al., 2018) and bedtools coverage (Quinlan & Hall, 2010) to obtain the regions with higher RLG1 TE content of the chromosome 27 located at the intergenic regions. We selected the regions that had the highest number of RLG1 elements and extracted a fasta sequence of each of these intergenic regions including the flanking genes with bedtools bedtofasta.

gRNAs and plasmids design

We searched for all the possible gRNAs using the intergenic region and unique sequence flanking the selected RLG1 islands of chromosome 27. We used the software CRISPOR (Concordet & Haeussler, 2018) to select unique gRNAs inside these sequences that do not have off-targets in other places of the genome with a high cut efficiency predicted by the software (Table 18).

Table 18: gRNAs used to eliminate RLG1 islands of chromosome 27.

Name	gRNA Sequence	Description
D#1	AAATCCTGTAGATCACAACA AGG	gRNA used to eliminate RLG1 island D of chromosome 27
D#2	ACCGGATTACTGGCTACGGG CGG	
O#1	ATGTCAACTACATGTCAAAG TGG	gRNA used to eliminate RLG1 island O of chromosome 27
O#2	CTGTCAAATAAAGAGGCC AGG	

The gRNA sequences were introduced by mutagenic PCRs to the plasmid p312 (pEntPp-gRNA-APT#23). Explained briefly, two primers were designed containing 20 bp with homology to the p312 plasmid flanking the old gRNA sequence and at the 3' of each primer the new gRNA was introduced (Figure 57).

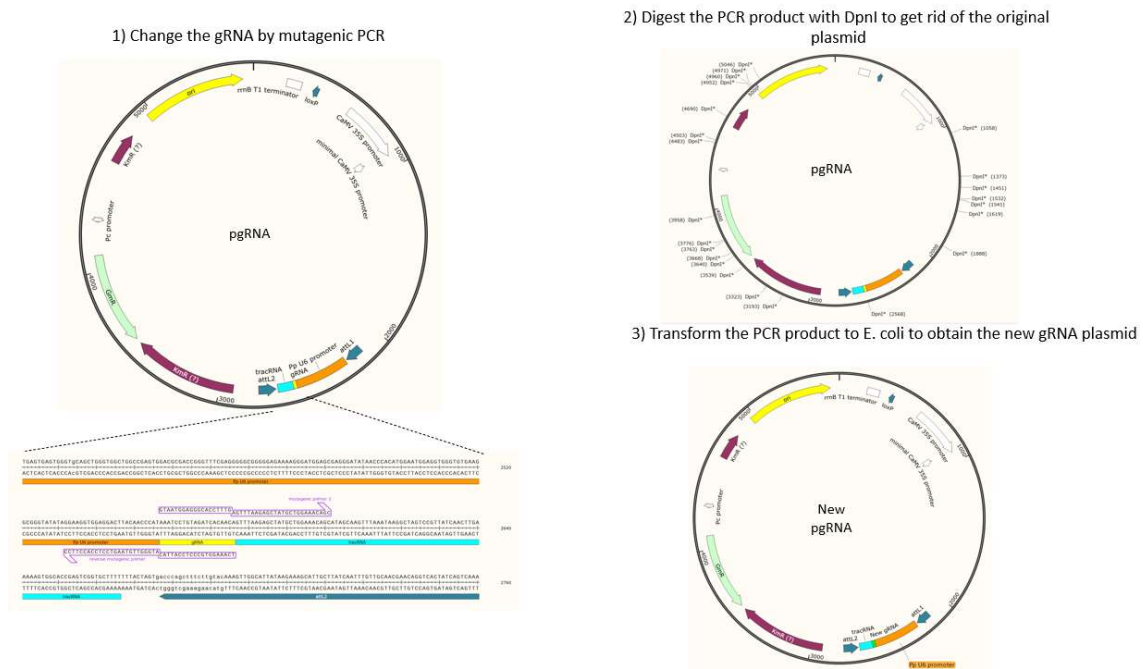


Figure 57: Replacement of the gRNA sequence by using two mutagenic primers. In 1) We designed two gRNAs that had homology to the flanking sequence of the gRNA (20 bp) per flank and that had at the 3' sequence the gRNA of interest. A PCR was done over the plasmids with 15 cycles. 2) The product of the PCR was digested using DpnI. DpnI only cuts methylated sides, getting rid of the original plasmid and keeping the PCR product that is not methylated. 3) *E. coli* was transformed with the PCR product obtaining the new plasmid with the gRNA of interest.

Then the following mix was prepared for each mutagenic primers: 31.5 μ l H₂O, 5x Phusion™ Buffer HF (Thermofisher #F-530XL) 10 μ l, 10mM dNTP 1 μ l, mutagenic Fw primer (10 pM) 2 μ l, mutagenic Rev primer (10 pM) 2 μ l, Phusion DNA polymerase 0.5 μ l (Thermofisher #F-530XL) and 1 μ l of p312 (10 ng/ μ l). Then the following PCR was done. A first step at 98°C for 30 seconds then the next three steps were repeated 15 times: 98°C for 10 seconds, 54°C for 30 seconds and 72°C for 3 minutes and finally a last step at 72°C for 5 minutes was done. Following this step 1 μ l of DpnI was added to the mix and incubated at 37°C to digest all the remaining plasmid. The DpnI was heat inactivated at 80°C for 20 minutes. Finally, 2 μ l of the mix were used to transform Top10 *E. coli* cells transforming by chemical transformation and plated in LB+Gentamycin. The colonies that were able to grow were checked by colony PCR for the presence of the expected gRNA and sanger sequencing.

The plasmids developed to replace the RLG1 islands D and O by a DNA template containing an antibiotic resistance gene were synthesized by Twist bioscience with the names pIsland D and pIsland O. Each of them had two homologous arms of 250 nt with

homology to the unique sequences flanking the RLG1 islands. Between the two homologous arms we cloned the HygR cassette from BHRf to build the plasmid pPV27 and the ZeoR cassette from the plasmid BZRf to build the plasmid pPV28.

Plant material

To perform all the experiments, we used the *P. patens* Gransden accession. The growth conditions were the same than the explained in the plant material of the previous section.

Plant transformation

P. patens were transformed as described in the previous section but using instead the plasmids pBNRf that contains a Neomycin resistant gene to select for the presence of the plasmids transiently, p165 that contains the CRISPR/Cas9 system and the plasmids that contains the gRNAs targeting the RLG1 islands; p321 (RLG1#D1) and pRLG1#D2 to target RLG1 island D or pRLG1#O1 and pRLG1#O2 to target RLG1 island O. We also transform using a combination of all the gRNAs targeting the 2 islands at the same time.

To replace the RLG1 islands D and O for the DNA designed templates containing the desired antibiotics (HygR cassette for RLG1 island D and ZeoR cassette for RLG1 island O) another round of transformations were performed using three different combinations at equimolar concentrations to up to 20µg of DNA per transformation:

- The plasmids p165, p321, pRLG1#D2 and the plasmid pPV27 to replace RLG1 island D for the HygR cassette, selecting after one week using BCDAT medium containing Hygromycin 25mg/L
- The plasmids p165, pRLG1#O1, pRLG1#O2 and the plasmid pPV28 to replace RLG1 island O for the ZeoR cassette, selecting after one week using BCDAT medium containing zeocin 100mg/L
- The plasmids p165, p321, pRLG1#D2, pRLG1#O1, pRLG1#O2, pPV27 and pPV28 to replace both RLG1 islands at the same time. Selecting after one week using BCDA medium containing Hygromycin 25mg/L and a second round of selection using BCDA medium containing Zeocin 100 mg/L.

Plant genotyping

DNA extractions were performed using the protocol described in the non-selective elimination of TEs.

Plants were genotyped for the presence of changes in the islands D and O using a combination of the following primers (Table 19):

Table 19: Primers used to check the modifications over the selected RLG1 islands of chromosome 27.

Name	Primer sequence	Description
d1 island fw	GAGAGTGTAAGCTTGAGAGATAAGT	Primers used to check the modification over the RLG1 island D, left flank
d1 island rev	GTTGACATCCACTAAACAGAAG	
d2 island fw	CATTAAAGATGGAGGTGATGTC	Primers used to check the modification over the RLG1 island D, right flank
d2 island rev	CAGGATGAATCAGTTCAGAAG	
o1 island fw	ACTTAAGTCGCTACGCTTAGTAG	Primers used to check the modification over the RLG1 island O, left flank
o1 island rev	AGCAGTTCTAACACTCCAAGAT	
o2 island fw	GGTCACAACATAACTTTTGTCTATC	Primers used to check the modification over the RLG1 island O, right flank
o2 island rev	AGAACTCTCTCTTTGGGTTAAG	
oxPV89	GAGAGTGTAAGCTTGAGAGATAAGT	check insertion homology arm1 island D (HygR)
oxPV90	GTTCCCAGATAAGGGAATTAGGGTT	
oxPV91	ACCACGTAGCCTGATACCTCT	check insertion homology arm2 island D (HygR)
oxPV92	CCCTTTGGTCTTCTGAGACTGT	
oxPV93	ACTTAAGTCGCTACGCTTAGTAG	check insertion homology arm1 island O (ZeoR)
oxPV94	CTCGGTACCATAACTTCGTATAGCA	
oxPV95	CCCCACTCATGATTTTATAGGGTCT	check insertion homology arm2 island O (ZeoR)
oxPV96	CGCTTAAAAATTGGTATCAGAGCCA	

All the genotyping's were done by PCR using the DreamTaq™ polymerase (Thermoscientific: REF #EP0703) using the manufacturers protocol adapting the Tm and the elongation time for each PCR. The amplification conditions were the following ones: 2 min at 95 °C, 35 cycles of 30 seconds at 95 °C, 30 seconds at the annealing temperature and 1 min per kbp at 72 °C with a final step of 5 minutes at 72 °C. When needed the PCRs products were sanger sequenced by CRAGs facilities.

RNA extraction and cDNA production

Plant material for RNA extraction was collected from protonemata grown for 1 week in BCDA medium overlaid with a cellophane. The plants were grown at 24°C with a cycle of 16 h of light (quantum irradiance of 60–80 $\mu\text{mol m}^{-2} \text{s}^{-1}$ light) and 8 h of darkness on

plates. The plant material was collected 4 hours after the lights turn on and immediately frozen in liquid N₂.

To extract RNA 20 to 100 mg of *P. patens* material were grinded using a mortar. We followed the manufacturer's instruction described in the Maxwell® RSC Plant Kit (Promega, Ref #AS1500) to perform the RNA extraction and the DNase treatment. The RNA was resuspended in 50 µl of sterile water, the RNA was quantified using the Nanodrop™ device and visualized the integrity in an agarose gel electrophoresis.

1 µg of total RNA was used to perform the first-strand cDNA synthesis using a modified oligo-dT primer (Casacuberta et al., 1995), using the SuperScript™ III reverse transcriptase (Thermo Fisher, REF #18080093). As a control we performed the same reaction with the same amount of RNA without adding the Reverse transcriptase.

qRT- PCR

Quantitative real-time PCRs were performed in 96-well plates using a Roche LightCycler II instrument. We used SYBR Green I Master Mix (Roche Applied Science) with primers over the genes Pp3c27_3930V3.1 and Pp3c27_3970V3.1 (Table 20) to estimate their relative expression at a final concentration of 1 µM and 50 ng of cDNA obtained from the reverse transcription to conduct the qRT-PCR analysis.

Table 20: Primers used to do the qRT PCRs over the genes Pp3c27_3930V3.1 and Pp3c27_3970V3.1.

Name	Primer Sequence	Description
oxPV118	TTGGAGCCTGGGCTATGAAC	Primers qRT PCR over the gene flanking island d Pp3c27_3930V3.1
oxPV119	TGGATGTGTTGGACACCAGG	
oxPV120	GGTTCTGCGGCTATGGATGA	Primers qRT PCR over the gene flanking island d Pp3c27_3970V3.1
oxPV121	GGGACACCTTCCCTCAGTTG	

Each sample was run in triplicate with negative reverse transcriptase for each sample and negative controls (H₂O). The following PCR conditions were used: 95°C for 5 min, followed by 95°C for 10 sec, 56°C for 10 sec, and 72°C for 10 sec. For normalization we used the relatively highly expressed genes. Encoding for the 60S ribosomal subunit and the APRT, using the primers previously described in Le Bail et al., 2013.

Chapter 3.4: Impact of Transposable Elements into *the genic regions*

TIP detection

We used the publicly available DNA-seq resequencing data of the three accessions of *P. patens* (Kaskaskia, SRX2234698; Reute, SRX1528135 and Villersexel, SRX030894) to detect the TIP polymorphisms. We combined the TIP annotation that we performed in Vendrell-Mir et al., 2020 with a new annotation. For this new annotation we ran Jitterbug (Hénaff et al., 2015) to improve the number of non-reference insertion polymorphisms and Pindel (Ye et al., 2009) to detect the polymorphic reference TE insertions.

To run both tools the reads were aligned to the reference genome using BWA-Aln (H. Li & Durbin, 2010), using the following script:

```
bwa aln -t 12 -n 4 -o 1 -e 3 -f $1.sai name_sample fastq_1
bwa aln -t 12 -n 4 -o 1 -e 3 -f $2.sai name_sample fastq_2
bwa sampe name_sample fastq_1.sai fastq_2.sai fastq_1 fastq_2 >
name_sample.sam;
samtools view -bS name_sample.sam > name_sample.bam;
samtools sort name_sample.bam > name_sample_sorted.bam;
```

Pindel was run with the following command:

```
pindel -f ppatens_genome.fa -i pindelconfig.tab -T 12 -c ALL -x 5 -r false -t
false -A 35 -o ppatens_sample/
```

The deletions detected by Pindel were filtered with a minimum length of 100 bp and a maximum length of 25000 bp keeping all the ones that were overlapping with annotated TEs with the goal to detect the polymorphic TE insertions.

Jitterbug was run with the following command:

```
jitterbug.py --psorted name_sample_sorted.bam -t TE_annotation.gff3 -l
sample_name -n TE_family -q 15 -o sample_TE_jitterbug
```

All the results were combined to a final gff3 that was filtered for a minimum TIP zygosity of 70%.

In silico expression analysis:

The profiles of expression in different development conditions and stresses of *P. patens* were analyzed through the microarrays over the development published on Ortiz-Ramírez et al., 2016 and the development conditions and stresses published on Fernandez-Pozo et al., 2020; Perroud et al., 2018. The RNAseq libraries from Reute and Gransden when available from Fernandez-Pozo et al., 2020 and Perroud et al., 2018 were aligned to the genes using SRA BLASTn (BLAST, 2009).

Search of homologous genes in other plant species

We looked for homologous genes using the protein sequence of the different genes over the Phytozome database (Goodstein et al., 2012) and the Refseq non-redundant protein database of NCBI using a reciprocal Blastp. We performed the phylogenetic analysis over the gene Pp3c14_9040V3.1 using the online tool SHOOT (Emms & Kelly, 2022).

Material used

To perform all the experiments, we used the *P. patens* Gransden, Villersexel-K3 (Vx), Reute-K1 (Re) and Kaskaskia (Ka) accessions. The plants were grown at 24°C with a cycle of 16 h of light (quantum irradiance of 60–80 $\mu\text{mol m}^{-2} \text{s}^{-1}$ light) and 8 h of darkness on plates containing BCDA (Cove et al., 2009) overlaid with a cellophane.

To genotype the insertions protonemata of 7 days old were fragmented of each accession and used to extract DNA following the protocol described in the Materials and methods of chapter 3.3: non-selective elimination of RLG1 elements using CRISPR/Cas9, DNA extraction.

To analyze the expression in protonemata for the different plants, we fragmented protonemata of the moss for the different accessions every 7 days letting it grow during this time in BCDA medium overlaid with a cellophane. After two rounds of fragmentation, we let the moss grow in BCDA medium overlaid in a cellophane and collected the samples at the time point of interest, collecting them always at 12 am and immediately keeping the sample in liquid Nitrogen. To collect gametophores, we let a

small fragment of protonemata grow in BCDA medium without a cellophane, after 21 days, gametophores were collected. Protoplasts were collected following the same protocol used to transform (Charlot et al., 2022), after digesting with Driselase™ and filtering the protoplasts the protoplasts were moved to a 1.5ml tube and immediately frozen in liquid nitrogen.

Polymorphism verification

Primers (Table 21) flanking the selected TE polymorphisms were designed and checked between Gransden and the accession where the TE insertion was detected. The following mix was prepared for each couple of primers: 30.5 µl H₂O, 5x Phusion™ Buffer HF (Thermofisher #F-530XL) 10 µl, 10mM dNTP 1 µl, Fw primer (10 pM) 2 µl, Rev primer (10 pM) 2 µl, Phusion DNA polymerase 0.5 µl (Thermofisher #F-530XL) and 2 µl of genomic DNA (50 ng/ul). Then the following PCR was done A first step at 98°C for 30 seconds then the next three steps were repeated 15 times 98°C for 10 seconds, the annealing temperature of the couple of primers for 30 seconds and 72° C for 8 minutes and finally a last step at 72°C for 16 minutes.

Table 21: Primers used to confirm the insertion over the different polymorphic TEs Pp3c14_9040V3.1 and Pp3c7_24710V3.1 and to genotype the KOs and the TE replacement in the different *P. patens* accessions.

Name	Primer Sequence	Description
oPV84	ACCTCTTCATTCTCCTCATCAGAGT	
oPV85	GGTAATGGAATGTTTCAATTGTTTG	
oxPV104	TCCGCAACTTCTAATGCGCT	Primers used to genotype the polymorphic RLG1 insertion over the Pp3c14_9040V3.1 promoter region and confirm the replacement
oxPV105	AGTATATCGTGTGACTGACAATGC	
oxPV106	GAGAGGCGCTCAAAGCTCTA	
oxPV107	GGTCTTCTCGCCCTGGAATG	
oxPV129	TTATTTACACACACACACATGTATG	
oxPV130	GACAATAATAGCCTTAAATAACAGTG	
oPV66	ATTGAGCCCATCCTTGAGGT	Primers used to genotype the polymorphic RLG1 insertion at the 3' UTR of the gene Pp3c4_24710V3.1 and confirm the editing of the gene
oPV67	TCCCAAACCTTCAGTCTTCAG	
oPV68	GTGGATTTTGAATGGATTGC	
oPV69	AGACCCTGAATGGAGTGGTG	
oPV70	GAGAGAGTGATTGTGGATTTTGAAT	
oPV71	AGAAAGATACTCGACCCAGAAAGAT	

In the cases that we could amplify both locus we sanger sequenced the PCR products. The PCR amplification corresponding to the RLG1 insertion in Villersexel located at the end of the 3' gene Pp3c4_27410V3.1 was cloned using the pGEM®-T easy vector system (Promega #A1360). The PCR amplification corresponding to the RLG1 insertion in Kaskaskia located at the promoter region of the gene Pp3c14_9040V3.1 was cloned to a plasmid using the pENTR™/D-TOPO™ system (Thermofisher #K240020). After that both fragments were sanger sequenced using the primers M13 fw and M13 rev flanking the integration of the PCR product and using primers over the RLG1 elements.

qRT-PCR Analysis

Quantitative real-time PCRs were performed in 96-well plates using a Roche LightCycler II instrument. We used SYBR Green I Master Mix (Roche Applied Science) with primers over the genes Pp3c4_24710V3.1 and Pp3c14_9040V3.1 (Table 22) to estimate their relative expression at a final concentration of 1 µM and 50 ng of cDNA obtained from the reverse transcription to conduct the qRT-PCR analysis. Each sample was run in triplicate with negative reverse transcriptase for each sample and negative controls (H₂O). The following PCR conditions were used: 95°C for 5 min, followed by 95°C for 10 sec, 56°C for 10 sec, and 72°C for 10 sec. For normalization we used the relatively highly

expressed genes. Encoding for the 60S ribosomal subunit and the APRT, using the primers previously described in Le Bail et al., 2013.

Table 22: Primers used to perform the qRT-PCRs over the genes Pp3c14_9040V3.1 and Pp3c7_24710V3.1.

Name	Primer Sequence	Description
oqPV15	CAGTTTTGGAGCCGTTAGGA	Primers qRT PCR over the gene Pp3c4_24710V3.1
oqPV16	ATAACCCACGACGTGAAACC	
oqPV28	GAGGAGTGGAGTGCTTTTCG	Primers qRT PCR over the gene flanking island d Pp3c14_9040V3.1
oqPV29	GCCGCTCAGAGTGAGTTTCT	

KOs of the genes and TIPs replacement

Knock Outs of the genes were produced using CRISPR/Cas9 over *P. patens* Gransden accession transforming the plants using the protocol described in Charlot et al., 2022. Using the plasmids p164 (CRISPR/Cas9 system) and the gRNAs plasmids, using the BNRf plasmid to select transiently for the presence of the plasmids. The obtained clones were genotyped by PCR using primers flanking the targeted deletions.

To replace the TE over the Pp3c14_9040V3.1 we used the CRISPR/Cpf1 system due to the promoter where we wanted to produce the replacement had a high A/T enrichment that did not allowed to find matching guide RNAs on these sequences. The cRNA sequence were introduced to a plasmid backbone using the mutagenic PCR technique previously described. The polymorphic side of Gransden without the TE was synthesized by Twist Bioscience ® introducing the SNPs found in Kaskaskia to the Gransden empty locus, the synthetic product was delivered by Twist bioscience ® as a plasmid named pGransden_locus. The locus of the insertion that was used was directly the cloned PCR product cloned to the the pENTR™/D-TOPO™ system named pKaskaskia_locus. The gRNAs and cRNAs sequence used to target the edition of the different genes were the ones described in Table 23.

Table 23: gRNAs and cDNAs used to perform the KOs and the RLG1 replacement in the different accessions.

Name	gRNA Sequence or cRNA Sequence	Description
gPp3c4#1	GATAGAAATGAAGGTATGAG	gRNA used to target the Pp3c4_24710V3.1
gPp3c4#1	CGGTAAATATGCCTGACTTG	
gPp3c17#1	AAAACCTCCGAAAGACACCAGT	gRNA used to target the Pp3c17_3870V3.1 gene
gPp3c17#2	TAGTTTCAACCCATTCAGCAC	
gPp3c14#1	CTACCGAAGCCTCAGCCGACT	gRNA used to target the Pp3c14_9040V3.1 gene
gPp3c14#2	ATGGAGGTTTTAGAGCTTTGA	
cPp3c14#1	ATTCTATTGGTAATGGAATGTTT	cRNA to eliminate the RLG1 polymorphic TE from Pp3c14_9040V3.1 in Kaskaskia
cPp3c14#2	ACATCCACTACTCACGAATATTT	
cPp3c14#3	TAATTGACAAAATACTTGTTGAC	cRNA to introduce the RLG1 polymorphic TE from Pp3c14_9040V3.1 in Gransden

P. patens Gransden accession was transformed using the protocol described in (Charlot et al., 2022) using a plasmid containing the CRISPR/Cpf1 system, a plasmid containing a cRNA targeting the promoter region where we wanted to introduce the RLG1 element, the plasmid pKaskaskia_locus and the plasmid BNRf to transiently select the clones that had been transformed after one week of regeneration using BCDA medium containing G418 (50 mg/L).

P. patens Kaskaskia accession was also transformed using the same protocol using the same plasmids (CRISPR/Cpf1 and BNRf) except for the cRNAs. We used two cRNAs flanking the RLG1 insertion and the pGransden_locus. Plants were also selected after one week of regeneration using BCDA medium containing G418 (50mg/L).

The obtained clones were genotyped using the primers described in the Table 21 selecting the clones of Gransden that had the RLG1 insertion and the ones of Kaskaskia that had the expected deletion. The selected PCR products were sanger sequenced.

Phenotypical analysis of the obtained KO clones and of the TE replacement clones

The KOs of Pp3c4_24710V3.1 were grown in BCD and BCDA to check the presence of any difference between Gransden Wt and Kaskaskia Wt. We also checked the phenotype in BCD supplemented with different concentrations of ABA (0 µg/L, 1 µg/L and 2 µg/L).

The clones where we performed the TE replacement and the control lines were grown on BCD and BCDA medium to check if there was any development difference between

them. We induced the production of sporophytes to check whatever they had a difference in the development over this stage growing initially the different lines under BCD medium 1 month on BCD medium and subsequently transferred them to 15°C with a cycle of 8 h of light ($30 \mu\text{mol m}^{-2} \text{s}^{-1}$) and 16 h of darkness after 2 weeks at 15°C the plants were watered to facilitate the generation of sporophytes (Hohe et al., 2002).

We collected protonemata samples of the different TE replacement clones, Gransden Wt and Kaskaskia Wt after two rounds of fragmentation and grown over BCDA with a cellophane during the days 4,7,10 and 14 to perform the quantification of the expression of the Pp3c14_9040V3.1 gene at the different time points using the protocol and primers previously described.

Chapter 4: An endogenous virus in the moss *Physcomitrium patens*

Chapter 4.4: Complementary results

Plasmids construction and transformation

The sequence of the virus was synthesized by Twist Bioscience®. As the limit of sequence that can be synthesized are 5 Kbp. We divided the virus on two different constructs; The construct pTwist-Amalga-5' that contains the 35S promoter, half of the PPAV1 fused to the GFP and the construct pTwist-Amalga-3' that contains the 35S terminator and the 3' of the PPAV1. A version containing only the Amalgavirus fused to the GFP was done with a ligation of the Amalga-3' digested with the enzymes Bgl2 and NcoI to the Amalga-5' digested with the enzymes PciI and BclI. A version containing only the modified Amalgavirus was done with a ligation Amalga-3' digested with Bgl2 and NcoI to the Amalga-5' digested with PciI and Bgl2.

The plasmids were transformed to Gransden and Reute accessions using the protocol described in Charlot et al., 2022 using the plasmid BNRf to select transiently for the clones that had been transformed selecting for 1 week in BCDA containing G418 (50 mg/L). After that it was checked for the presence of the constructs by RNA extraction and cDNA production as described on the manuscript of Vendrell-Mir et al., 2021 and checked by PCR using primers flanking the modified product.

RNAseq libraries analysis

The PPAV1 sequence was concatenated to the *P. patens* transcriptome sequence in FASTA file format. We used the SRA libraries BBTWW (SRX3364034), BBTWY (SRX3364009), BBTWX (SRX3364005), BXHHX (SRX3364085), BXHHZ (SRX3364083), BXHHY (SRX3364082) and mapped them to the modified transcriptome using Bowtie2 (Langmead and Steven L Salzberg, 2013), we estimated the normalized expression for the different genes and the PPAV1 for each library using RSEM (B. Li & Dewey, 2011).

BIBLIOGRAPHY

BIBLIOGRAPHY

- Alexandrov, N., Tai, S., Wang, W., Mansueto, L., Palis, K., Fuentes, R. R., Ulat, V. J., Chebotarov, D., Zhang, G., Li, Z., Mauleon, R., Hamilton, R. S., & McNally, K. L. (2015). SNP-Seek database of SNPs derived from 3000 rice genomes. *Nucleic Acids Research*, 43(Database issue), D1023-7. <https://doi.org/10.1093/nar/gku1039>
- Alioto, T., Alexiou, K. G., Bardil, A., Barteri, F., Castanera, R., Cruz, F., Dhingra, A., Duval, H., Fernández I Martí, Á., Frias, L., Galán, B., García, J. L., Howad, W., Gómez-Garrido, J., Gut, M., Julca, I., Morata, J., Puigdomènech, P., Ribeca, P., ... Arús, P. (2020). Transposons played a major role in the diversification between the closely related almond and peach genomes: results from the almond genome sequence. *The Plant Journal : For Cell and Molecular Biology*, 101(2), 455–472. <https://doi.org/10.1111/tpj.14538>
- Arif, M. A., Hiss, M., Tomek, M., Busch, H., Meyberg, R., Tintelnot, S., Reski, R., Rensing, S. A., & Frank, W. (2019). ABA-Induced Vegetative Diaspore Formation in *Physcomitrella patens*. *Frontiers in Plant Science*, 10, 315. <https://doi.org/10.3389/fpls.2019.00315>
- Asif-Laidin, A., Conesa, C., Bonnet, A., Grison, C., Adhya, I., Menouni, R., Fayol, H., Palmic, N., Acker, J., & Lesage, P. (2020). A small targeting domain in Tyl1 integrase is sufficient to direct retrotransposon integration upstream of tRNA genes. *The EMBO Journal*, 39(17), e104337–e104337. <https://doi.org/10.15252/embj.2019104337>
- Aurélié, K., Alexander, S., & Cédric, F. (2017). Dynamics of genome size evolution in birds and mammals. *Proceedings of the National Academy of Sciences*, 114(8), E1460–E1469. <https://doi.org/10.1073/pnas.1616702114>
- Babaian, A., Thompson, I. R., Lever, J., Gagnier, L., Karimi, M. M., & Mager, D. L. (2019). LIONS: Analysis suite for detecting and quantifying transposable element initiated transcription from RNA-seq. *Bioinformatics*, 35(19), 3839–3841. <https://doi.org/10.1093/bioinformatics/btz130>
- Béclin, C., Boutet, S., Waterhouse, P., & Vaucheret, H. (2002). A Branched Pathway for Transgene-Induced RNA Silencing in Plants. *Current Biology*, 12(8), 684–688. [https://doi.org/https://doi.org/10.1016/S0960-9822\(02\)00792-3](https://doi.org/https://doi.org/10.1016/S0960-9822(02)00792-3)
- Beguiristain, T., Grandbastien, M.-A., Puigdomènech, P., & Casacuberta, J. M. (2001). Three Tnt1 Subfamilies Show Different Stress-Associated Patterns of Expression in Tobacco. Consequences for Retrotransposon Control and Evolution in Plants. *Plant Physiology*, 127(1), 212–221. <https://doi.org/10.1104/pp.127.1.212>
- Benedito, V. A., Torres-Jerez, I., Murray, J. D., Andrianakaja, A., Allen, S., Kakar, K., Wandrey, M., Verdier, J., Zuber, H., Ott, T., Moreau, S., Niebel, A., Frickey, T., Weiller, G., He, J., Dai, X., Zhao, P. X., Tang, Y., & Udvardi, M. K. (2008). A gene expression atlas of the model legume *Medicago truncatula*. *The Plant Journal : For Cell and Molecular Biology*, 55(3), 504–513. <https://doi.org/10.1111/j.1365-313X.2008.03519.x>
- Bessoltane, N., Charlot, F., Guyon-Debast, A., Charif, D., Mara, K., Collonnier, C., Perroud, P.-F., Tepfer, M., & Nogué, F. (2022). Genome-wide specificity of plant genome editing by both CRISPR–Cas9 and TALEN. *Scientific Reports*, 12(1),

9330. <https://doi.org/10.1038/s41598-022-13034-2>

- BLAST. (2009). *Nucleotide BLAST: Search nucleotide databases using a nucleotide query*. Basic Local Alignment Search Tool.
- Blatch, G. L., & Lässle, M. (1999). The tetratricopeptide repeat: a structural motif mediating protein-protein interactions. *BioEssays*, *21*(11), 932–939. [https://doi.org/https://doi.org/10.1002/\(SICI\)1521-1878\(199911\)21:11<932::AID-BIES5>3.0.CO;2-N](https://doi.org/https://doi.org/10.1002/(SICI)1521-1878(199911)21:11<932::AID-BIES5>3.0.CO;2-N)
- Boeke, J. D., Garfinkel, D. J., Styles, C. A., & Fink, G. R. (1985). Ty elements transpose through an RNA intermediate. *Cell*, *40*(3), 491–500. [https://doi.org/10.1016/0092-8674\(85\)90197-7](https://doi.org/10.1016/0092-8674(85)90197-7)
- Bogaerts-Marquez, M., Barron, M. G., Fiston-Lavier, A. S., Vendrell-Mir, P., Castanera, R., Casacuberta, J. M., & Gonzalez, J. (2020). T-lex3: An accurate tool to genotype and estimate population frequencies of transposable elements using the latest short-read whole genome sequencing data. *Bioinformatics*, *36*(4), 1191–1197. <https://doi.org/10.1093/bioinformatics/btz727>
- Bonnet, A., & Lesage, P. (2021). Light and shadow on the mechanisms of integration site selection in yeast Ty retrotransposon families. *Current Genetics*, *67*(3), 347–357. <https://doi.org/10.1007/s00294-021-01154-7>
- Bortesi, L., & Fischer, R. (2015). The CRISPR/Cas9 system for plant genome editing and beyond. *Biotechnology Advances*, *33*(1), 41–52. <https://doi.org/10.1016/j.biotechadv.2014.12.006>
- Bundock, P., & Hooykaas, P. (2005). An Arabidopsis hAT-like transposase is essential for plant development. *Nature*, *436*(7048), 282–284. <https://doi.org/10.1038/nature03667>
- Bushnell, B. (2015). *BBMap*. <https://Sourceforge.Net/Projects/Bbmap/>.
- Carpentier, M.-C., Manfroi, E., Wei, F.-J., Wu, H.-P., Lasserre, E., Llauro, C., Debladis, E., Akakpo, R., Hsing, Y.-I., & Panaud, O. (2019). Retrotranspositional landscape of Asian rice revealed by 3000 genomes. *Nature Communications*, *10*(1), 24. <https://doi.org/10.1038/s41467-018-07974-5>
- Casacuberta, J. M., & Santiago, N. (2003). Plant LTR-retrotransposons and MITEs: Control of transposition and impact on the evolution of plant genes and genomes. *Gene*, *311*(1–2), 1–11. [https://doi.org/10.1016/S0378-1119\(03\)00557-2](https://doi.org/10.1016/S0378-1119(03)00557-2)
- Castanera, R., Ruggieri, V., Pujol, M., Garcia-Mas, J., & Casacuberta, J. M. (2020). An Improved Melon Reference Genome With Single-Molecule Sequencing Uncovers a Recent Burst of Transposable Elements With Potential Impact on Genes. *Frontiers in Plant Science*, *10*. <https://doi.org/10.3389/fpls.2019.01815>
- Castanera, R., Vendrell-Mir, P., Bardil, A., Carpentier, M. C., Panaud, O., & Casacuberta, J. M. (2021). Amplification dynamics of miniature inverted-repeat transposable elements and their impact on rice trait variability. In *Plant Journal* (Vol. 107, Issue 1). <https://doi.org/10.1111/tpj.15277>
- Cavrak, V. V., Lettner, N., Jamge, S., Kosarewicz, A., Bayer, L. M., & Mittelsten Scheid, O. (2014). How a Retrotransposon Exploits the Plant's Heat Stress Response for Its Activation. *PLoS Genetics*, *10*(1).

<https://doi.org/10.1371/journal.pgen.1004115>

- Chabannes, M., & Iskra-Caruana, M. L. (2013). Endogenous pararetroviruses - A reservoir of virus infection in plants. *Current Opinion in Virology*, 3(6), 615–620. <https://doi.org/10.1016/j.coviro.2013.08.012>
- Charlot, F., Goudounet, G., Nogué, F., & Perroud, P.-F. (2022). *Physcomitrium patens Protoplasting and Protoplast Transfection BT - Protoplast Technology: Methods and Protocols* (K. Wang & F. Zhang (eds.); pp. 3–19). Springer US. https://doi.org/10.1007/978-1-0716-2164-6_1
- Chen, J., Hu, Q., Zhang, Y., Lu, C., & Kuang, H. (2014). P-MITE: a database for plant miniature inverted-repeat transposable elements. *Nucleic Acids Research*, 42(D1), D1176–D1181. <https://doi.org/10.1093/nar/gkt1000>
- Choi, J., Chen, W., Suiter, C. C., Lee, C., Chardon, F. M., Yang, W., Leith, A., Daza, R. M., Martin, B., & Shendure, J. (2022). Precise genomic deletions using paired prime editing. *Nature Biotechnology*, 40(2), 218–226. <https://doi.org/10.1038/s41587-021-01025-z>
- Civáň, P., Švec, M., & Hauptvogel, P. (2011). On the Coevolution of Transposable Elements and Plant Genomes. *Journal of Botany*, 2011, 1–9. <https://doi.org/10.1155/2011/893546>
- Coletta, R. Della, Qiu, Y., Ou, S., Hufford, M. B., & Hirsch, C. N. (2021). *How the pan-genome is changing crop genomics and improvement*. 1–19.
- Collonnier, C., Epert, A., Mara, K., Maclot, F., Guyon-Debast, A., Charlot, F., White, C., Schaefer, D. G., & Nogué, F. (2017). CRISPR-Cas9-mediated efficient directed mutagenesis and RAD51-dependent and RAD51-independent gene targeting in the moss *Physcomitrella patens*. *Plant Biotechnology Journal*, 15(1), 122–131. <https://doi.org/10.1111/pbi.12596>
- Concordet, J.-P., & Haeussler, M. (2018). CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. *Nucleic Acids Research*, 46(W1), W242–W245. <https://doi.org/10.1093/nar/gky354>
- Coruh, C., Cho, S. H., Shahida, S., Liu, Q., Wierzbicki, A., & Axtell, M. J. (2015). Comprehensive annotation of *Physcomitrella patens* small RNA loci reveals that the heterochromatic short interfering rna pathway is largely conserved in land plants. *Plant Cell*, 27(8), 2148–2162. <https://doi.org/10.1105/tpc.15.00228>
- Coskan, S., Alcalá-Briseo, R., & Polston, J. (2016). Distribution and vertical transmission of Southern tomato virus in tomato. *PHYTOPATHOLOGY*, 106(12), 119–120.
- Courtial, B., Feuerbach, F., Eberhard, S., Rohmer, L., Chiapello, H., Camilleri, C., & Lucas, H. (2001). Tnt1 transposition events are induced by in vitro transformation of *Arabidopsis thaliana*, and transposed copies integrate into genes. *Molecular and General Genetics*, 265(1), 32–42. <https://doi.org/10.1007/s004380000387>
- Cove, D. J., Perroud, P.-F., Charron, A. J., McDaniel, S. F., Khandelwal, A., & Quatrano, R. S. (2009). Culturing the moss *Physcomitrella patens*. *Cold Spring Harbor Protocols*, 2009(2), pdb-prot5136.
- Curcio, M. J., & Garfinkel, D. J. (1991). Single-step selection for Ty1 element

- retrotransposition. *Proceedings of the National Academy of Sciences*, 88(3), 936–940. <https://doi.org/10.1073/pnas.88.3.936>
- Danilevicz, M. F., Tay Fernandez, C. G., Marsh, J. I., Bayer, P. E., & Edwards, D. (2020). Plant pangenomics: approaches, applications and advancements. *Current Opinion in Plant Biology*, 54, 18–25. <https://doi.org/10.1016/j.pbi.2019.12.005>
- de Tomás, C., Bardil, A., Castanera, R., Casacuberta, J. M., & Vicient, C. M. (2022). Absence of major epigenetic and transcriptomic changes accompanying the interspecific cross between peach and almond. *Horticulture Research*.
- Diop, S. I., Subotic, O., Giraldo-Fonseca, A., Waller, M., Kirbis, A., Neubauer, A., Potente, G., Murray-Watson, R., Boskovic, F., Bont, Z., Hock, Z., Payton, A. C., Duijsings, D., Pirovano, W., Conti, E., Grossniklaus, U., McDaniel, S. F., & Szövényi, P. (2020). A pseudomolecule-scale genome assembly of the liverwort *Marchantia polymorpha*. *The Plant Journal*, 101(6), 1378–1396. <https://doi.org/https://doi.org/10.1111/tpj.14602>
- Disdero, E., & Filée, J. (2017). LoRTE: Detecting transposon-induced genomic variants using low coverage PacBio long read sequences. *Mobile DNA*, 8, 5. <https://doi.org/10.1186/s13100-017-0088-x>
- Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1), 15–21. <https://doi.org/10.1093/bioinformatics/bts635>
- Domb, K., Katz, A., Harris, K. D., Yaari, R., Kaisler, E., Nguyen, V. H., Hong, U. V. T., Griess, O., Heskiau, K. G., Ohad, N., & Zemach, A. (2020). DNA methylation mutants in *Physcomitrella patens* elucidate individual roles of CG and non-CG methylation in genome regulation. *Proceedings of the National Academy of Sciences of the United States of America*, 117(52), 33700–33710. <https://doi.org/10.1073/pnas.2011361117>
- Domínguez, M., Dugas, E., Benchouaia, M., Leduque, B., Jiménez-Gómez, J. M., Colot, V., & Quadrana, L. (2020). The impact of transposable elements on tomato diversity. *Nature Communications*, 11(1), 4058. <https://doi.org/10.1038/s41467-020-17874-2>
- Dong, C.-J., & Liu, J.-Y. (2010). The Arabidopsis EAR-motif-containing protein RAP2.1 functions as an active transcriptional repressor to keep stress responses under tight control. *BMC Plant Biology*, 10(1), 47. <https://doi.org/10.1186/1471-2229-10-47>
- Doyle, J. J., & Doyle, J. L. (1987). *A rapid DNA isolation procedure for small quantities of fresh leaf tissue*.
- Elliott, T. A., Heitkam, T., Hubley, R., Quesneville, H., Suh, A., Wheeler, T. J., Anselem, J., Berrens, R. V., Gonzalez, J., Goubert, C., Lesica, G., Rosen, J., Smit, A. F., Storer, J. M., & Schaack, S. (2021). TE Hub: A community-oriented space for sharing and connecting tools, data, resources, and methods for transposable element annotation. *Mobile DNA*, 12(1), 1–5. <https://doi.org/10.1186/s13100-021-00244-0>
- Emms, D. M., & Kelly, S. (2022). SHOOT: phylogenetic gene search and ortholog

- inference. *Genome Biology*, 23(1), 85. <https://doi.org/10.1186/s13059-022-02652-8>
- Fernandez-Pozo, N., Haas, F. B., Meyberg, R., Ullrich, K. K., Hiss, M., Perroud, P.-F., Hanke, S., Kratz, V., Powell, A. F., Vesty, E. F., Daum, C. G., Zane, M., Lipzen, A., Sreedasyam, A., Grimwood, J., Coates, J. C., Barry, K., Schmutz, J., Mueller, L. A., & Rensing, S. A. (2020a). PEATmoss (Physcomitrella Expression Atlas Tool): a unified gene expression atlas for the model plant *Physcomitrella patens*. *The Plant Journal*, 102(1), 165–177. <https://doi.org/https://doi.org/10.1111/tpj.14607>
- Feschotte, C., Swamy, L., & Wessler, S. R. (2003). Genome-wide analysis of mariner-like transposable elements in rice reveals complex relationships with stowaway miniature inverted repeat transposable elements (MITEs). *Genetics*, 163(2), 747–758. <https://doi.org/10.1093/genetics/163.2.747>
- Finnegan, D. J. (1997). Transposable elements: How non-LTR retrotransposons do it. *Current Biology*, 7(4), R245–R248. [https://doi.org/https://doi.org/10.1016/S0960-9822\(06\)00112-6](https://doi.org/https://doi.org/10.1016/S0960-9822(06)00112-6)
- Finnegan DJ(1989).Eukaryotic transposable elements and genome evolution. *Trends Genet. Apr*;5(4):103-7. doi: 10.1016/0168-9525(89)90039-5.
- Fricker, A. D., & Peters, J. E. (2014). Vulnerabilities on the Lagging-Strand Template: Opportunities for Mobile Elements. *Annual Review of Genetics*, 48(1), 167–186. <https://doi.org/10.1146/annurev-genet-120213-092046>
- Fuentes, R. R., Chebotarov, D., Duitama, J., Smith, S., De la Hoz, J. F., Mohiyuddin, M., Wing, R. A., McNally, K. L., Tatarinova, T., & Grigoriev, A. (2019). Structural variants in 3000 rice genomes. *Genome Research*, 29(5), 870–880.
- Fultz, D., Choudury, S. G., & Slotkin, R. K. (2015). Silencing of active transposable elements in plants. *Current Opinion in Plant Biology*, 27, 67–76. <https://doi.org/10.1016/j.pbi.2015.05.027>
- Gao, X., Hou, Y., Ebina, H., Levin, H. L., & Voytas, D. F. (2008). Chromodomains direct integration of retrotransposons to heterochromatin. *Genome Research*, 18(3), 359–369. <https://doi.org/10.1101/gr.7146408>
- Gilbert, K. B., Holcomb, E. E., Allscheid, R. L., & Carrington, J. C. (2019). Hiding in plain sight: New virus genomes discovered via a systematic analysis of fungal public transcriptomes. *PLoS One*, 14(7), e0219207.
- Gonzalvo, N. S. (2006). *Control de la transposición e impacto de los elementos transponibles en genomas vegetales: análisis del retrotransposón Tnt1 de tabaco y del Mite Emigrant de Arabidopsis*. Universitat Autònoma de Barcelona.
- Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., & Rokhsar, D. S. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research*, 40(Database issue), D1178–D1186. <https://doi.org/10.1093/nar/gkr944>
- Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., Di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., ... Regev, A. (2011). Full-length transcriptome assembly from

RNA-Seq data without a reference genome. *Nature Biotechnology*.
<https://doi.org/10.1038/nbt.1883>

- Grandbastien, M. A. (1998). Activation of plant retrotransposons under stress conditions. *Trends in Plant Science*, 3(5), 181–187. [https://doi.org/10.1016/S1360-1385\(98\)01232-1](https://doi.org/10.1016/S1360-1385(98)01232-1)
- Guffanti, G., Bartlett, A., Klengel, T., Klengel, C., Hunter, R., Glinsky, G., & Macciardi, F. (2018). Novel Bioinformatics Approach Identifies Transcriptional Profiles of Lineage-Specific Transposable Elements at Distinct Loci in the Human Dorsolateral Prefrontal Cortex. *Molecular Biology and Evolution*, 35(10), 2435–2453. <https://doi.org/10.1093/molbev/msy143>
- Guyon-Debast, A., Alboresi, A., Terret, Z., Charlot, F., Berthier, F., Vendrell-Mir, P., Casacuberta, J. M., Veillet, F., Morosinotto, T., Gallois, J. L., & Nogu e, F. (2021). A blueprint for gene function analysis through Base Editing in the model plant *Physcomitrium* (*Physcomitrella*) *patens*. *New Phytologist*, 230(3), 1258–1272. <https://doi.org/10.1111/nph.17171>
- Haas, F. B., Fernandez-Pozo, N., Meyberg, R., Perroud, P. F., G ottig, M., Stingl, N., Saint-Marcoux, D., Langdale, J. A., & Rensing, S. A. (2020). Single Nucleotide Polymorphism Charting of *P. patens* Reveals Accumulation of Somatic Mutations During *in vitro* Culture on the Scale of Natural Variation by Selfing. *Frontiers in Plant Science*, 11(July), 1–18. <https://doi.org/10.3389/fpls.2020.00813>
- Hata, Y., Naramoto, S., & Kyojuka, J. (2019). BLADE - ON - PETIOLE genes are not involved in the transition from protonema to gametophore in the moss *Physcomitrella patens*. *Journal of Plant Research*, 132(5), 617–627. <https://doi.org/10.1007/s10265-019-01132-8>
- Havecker, E. R., Gao, X., & Voytas, D. F. (2004). The diversity of LTR retrotransposons. *Genome Biology*, 5(6), 225. <https://doi.org/10.1186/gb-2004-5-6-225>
- Hayashi, K., & Yoshida, H. (2009). Refunctionalization of the ancient rice blast disease resistance gene *Pit* by the recruitment of a retrotransposon as a promoter. *The Plant Journal*, 57(3), 413–425. <https://doi.org/https://doi.org/10.1111/j.1365-313X.2008.03694.x>
- Hayward, A. (2017). Origin of the retroviruses: when, where, and how? *Current Opinion in Virology*, 25, 23–27. <https://doi.org/10.1016/j.coviro.2017.06.006>
- Hedges, D. J., & Deininger, P. L. (2007). Inviting instability: Transposable elements, double-strand breaks, and the maintenance of genome integrity. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*, 616(1–2), 46–59. <https://doi.org/10.1016/j.mrfmmm.2006.11.021>
- H enaff, E., Vives, C., Desvoves, B., Chaurasia, A., Payet, J., Gutierrez, C., & Casacuberta, J. M. (2014). Extensive amplification of the E2F transcription factor binding sites by transposons during evolution of Brassica species. *The Plant Journal : For Cell and Molecular Biology*, 77(6), 852–862. <https://doi.org/10.1111/tbj.12434>
- H enaff, E., Zapata, L., Casacuberta, J. M., & Ossowski, S. (2015). Jitterbug: somatic and germline transposon insertion detection at single-nucleotide resolution. *BMC*

Genomics, 16(1), 768. <https://doi.org/10.1186/s12864-015-1975-5>

- Hernández-Pinzón, I., Cifuentes, M., Hénaff, E., Santiago, N., Espinás, M. L., & Casacuberta, J. M. (2012). The Tnt1 retrotransposon escapes silencing in tobacco, its natural host. *PLoS One*, 7(3), e33816.
- Hernández-Pinzón, I., de Jesús, E., Santiago, N., & Casacuberta, J. M. (2009). The frequent transcriptional readthrough of the tobacco Tnt1 retrotransposon and its possible implications for the control of resistance genes. *Journal of Molecular Evolution*, 68(3), 269–278.
- Hirochika, H. (2001). Contribution of the Tos17 retrotransposon to rice functional genomics. *Current Opinion in Plant Biology*, 4(2), 118–122.
- Hiss, M., Meyberg, R., Westermann, J., Haas, F. B., Schneider, L., Schallenberg-Rüdinger, M., Ullrich, K. K., & Rensing, S. A. (2017). Sexual reproduction, sporophyte development and molecular variation in the model moss *Physcomitrella patens*: introducing the ecotype Reute. *Plant Journal*, 90(3), 606–620. <https://doi.org/10.1111/tj.13501>
- Hohe, A., Rensing, S. A., Mildner, M., Lang, D., & Reski, R. (2002). Day length and temperature strongly influence sexual reproduction and expression of a novel MADS-box gene in the moss *Physcomitrella patens*. *Plant Biology*, 4(05), 595–602.
- Holá, M., Kozák, J., Vágnerová, R., & Angelis, K. J. (2013). Genotoxin induced mutagenesis in the model plant *Physcomitrella patens*. *BioMed Research International*, 2013, 535049. <https://doi.org/10.1155/2013/535049>
- Hoyt, S. J., Storer, J. M., Hartley, G. A., Grady, P. G. S., Gershman, A., Lima, L. G. de, Limouse, C., Halabian, R., Wojenski, L., Rodriguez, M., Altemose, N., Core, L., Gerton, J. L., Makalowski, W., Olson, D., Rosen, J., Smit, A. F. A., Straight, A. F., Vollger, M. R., ... O'Neill, R. J. (2021). From telomere to telomere: the transcriptional and epigenetic state of human repeat elements. *BioRxiv*, 3112, 2021.07.12.451456. <https://doi.org/10.1126/science.abk3112>
- Hühns, S., Bauer, C., Buhlmann, S., Heinze, C., von Barga, S., Paape, M., & Kellmann, J.-W. (2003). Tomato spotted wilt virus (TSWV) infection of *Physcomitrella patens* gametophores. *Plant Cell, Tissue and Organ Culture*, 75(2), 183–187. <https://doi.org/10.1023/A:1025070722420>
- Izsvák, Z., Ivics, Z., Shimoda, N., Mohn, D., Okamoto, H., & Hackett, P. B. (1999). Short Inverted-Repeat Transposable Elements in Teleost Fish and Implications for a Mechanism of Their Amplification. *Journal of Molecular Evolution*, 48(1), 13–21. <https://doi.org/10.1007/PL00006440>
- Jangam, D., Feschotte, C., & Betrán, E. (2017). Transposable Element Domestication As an Adaptation to Evolutionary Conflicts. *Trends in Genetics*, 33(11), 817–831. <https://doi.org/10.1016/j.tig.2017.07.011>
- Jeong, H. H., Yalamanchili, H. K., Guo, C., Shulman, J. M., & Liu, Z. (2018). An ultra-fast and scalable quantification pipeline for transposable elements from next generation sequencing data. *Pacific Symposium on Biocomputing*, 0(212669), 168–179. https://doi.org/10.1142/9789813235533_0016

- Jiang, N., Bao, Z., Zhang, X., Hirochika, H., Eddy, S. R., McCouch, S. R., & Wessler, S. R. (2003). An active DNA transposon family in rice. *Nature*, *421*(6919), 163–167. <https://doi.org/10.1038/nature01214>
- Jin, Y., Tam, O. H., Paniagua, E., & Hammell, M. (2015). Tetranscripts: A package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinformatics*, *31*(22), 3593–3599. <https://doi.org/10.1093/bioinformatics/btv422>
- Kapitonov, V. V., & Jurka, J. (2007). Helitrons on a roll: eukaryotic rolling-circle transposons. *Trends in Genetics : TIG*, *23*(10), 521–529. <https://doi.org/10.1016/j.tig.2007.08.004>
- Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*. <https://doi.org/10.1093/molbev/mst010>
- Kaul, S., Koo, H. L., Jenkins, J., Rizzo, M., Rooney, T., Tallon, L. J., Feldblyum, T., Nierman, W., Benito, M.-I., & Lin, X. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, *408*(6814), 796–815.
- Kawahara, Y., de la Bastide, M., Hamilton, J. P., Kanamori, H., McCombie, W. R., Ouyang, S., Schwartz, D. C., Tanaka, T., Wu, J., Zhou, S., Childs, K. L., Davidson, R. M., Lin, H., Quesada-Ocampo, L., Vaillancourt, B., Sakai, H., Lee, S. S., Kim, J., Numa, H., ... Matsumoto, T. (2013). Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice (New York, N.Y.)*, *6*(1), 4. <https://doi.org/10.1186/1939-8433-6-4>
- Keane, T. M., Wong, K., & Adams, D. J. (2013). RetroSeq: transposable element discovery from next-generation sequencing data. *Bioinformatics (Oxford, England)*, *29*(3), 389–390. <https://doi.org/10.1093/bioinformatics/bts697>
- Kelleher, E. S. (2016). Reexamining the P-Element Invasion of *Drosophila melanogaster* Through the Lens of piRNA Silencing. *Genetics*, *203*(4), 1513–1531. <https://doi.org/10.1534/genetics.115.184119>
- Kidwell, M. G. (2002). Transposable elements and the evolution of genome size in eukaryotes. *Genetica*, *115*(1), 49–63. <https://doi.org/10.1023/A:1016072014259>
- Kidwell, M. G., & Lisch, D. R. (2000). Transposable elements and host genome evolution. *Trends in Ecology and Evolution*, *15*(3), 95–99. [https://doi.org/10.1016/S0169-5347\(99\)01817-0](https://doi.org/10.1016/S0169-5347(99)01817-0)
- Klepikova, A. V, Kasianov, A. S., Gerasimov, E. S., Logacheva, M. D., & Penin, A. A. (2016). A high resolution map of the *Arabidopsis thaliana* developmental transcriptome based on RNA-seq profiling. *The Plant Journal*, *88*(6), 1058–1070. <https://doi.org/https://doi.org/10.1111/tpj.13312>
- Kofler, R., Gómez-Sánchez, D., & Schlötterer, C. (2016). PoPoolationTE2: Comparative Population Genomics of Transposable Elements Using Pool-Seq. *Molecular Biology and Evolution*. <https://doi.org/10.1093/molbev/msw137>
- Komatsu, M., Shimamoto, K., & Kyojuka, J. (2003). Two-step regulation and continuous retrotransposition of the rice LINE-type retrotransposon Karma. *The Plant Cell*, *15*(8), 1934–1944. <https://doi.org/10.1105/tpc.011809>

- Kress, W. J., Soltis, D. E., & Kersey, P. J. (2022). *Green plant genomes : What we know in an era of rapidly expanding opportunities*. 119(4), 1–9.
<https://doi.org/10.1073/pnas.2115640118/-/DCSupplemental.Published>
- Krupovic, M., & Koonin, E. V. (2017). Homologous capsid proteins testify to the common ancestry of retroviruses, caulimoviruses, pseudoviruses, and metaviruses. *Journal of Virology*, 91(12), e00210-17.
- Kumar, S., Stecher, G., Suleski, M., & Hedges, S. B. (2017). TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Molecular Biology and Evolution*, 34(7), 1812–1819. <https://doi.org/10.1093/molbev/msx116>
- Kwon, S., Mehta, P., Dickstein, R., Gill, U. S., & Nandety, R. S. (2019). *Genome-wide analysis of flanking sequences reveals that Tnt1 insertion is positively correlated with gene methylation in Medicago truncatula*. 1106–1119.
<https://doi.org/10.1111/tpj.14291>
- Lanciano, S., Zhang, P., Llauro, C., & Mirouze, M. (2021). Identification of Extrachromosomal Circular Forms of Active Transposable Elements Using Mobilome-Seq. *Methods in Molecular Biology (Clifton, N.J.)*, 2250, 87–93.
https://doi.org/10.1007/978-1-0716-1134-0_7
- Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., ... Trust:, T. W. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822), 860–921. <https://doi.org/10.1038/35057062>
- Lang, D., Ullrich, K. K., Murat, F., Fuchs, J., Jenkins, J., Haas, F. B., Piednoel, M., Gundlach, H., Van Bel, M., Meyberg, R., Vives, C., Morata, J., Symeonidi, A., Hiss, M., Muchero, W., Kamisugi, Y., Saleh, O., Blanc, G., Decker, E. L., ... Rensing, S. A. (2018). The *Physcomitrella patens* chromosome-scale assembly reveals moss genome structure and evolution. *Plant Journal*, 93(3), 515–533.
<https://doi.org/10.1111/tpj.13801>
- Langmead and Steven L Salzberg. (2013). Bowtie2. *Nature Methods*.
<https://doi.org/10.1038/nmeth.1923>.Fast
- Le Bail, A., Scholz, S., & Kost, B. (2013). Evaluation of reference genes for RT qPCR analyses of structure-specific and hormone regulated gene expression in *Physcomitrella patens* gametophytes. *PloS One*, 8(8), e70998–e70998.
<https://doi.org/10.1371/journal.pone.0070998>
- Lerat, E., Casacuberta, J., Chaparro, C., & Vieira, C. (2019). On the importance to acknowledge transposable elements in epigenomic analyses. *Genes*, 10(4).
<https://doi.org/10.3390/genes10040258>
- Lerat, E., Fablet, M., Modolo, L., Lopez-Maestre, H., & Vieira, C. (2017). TETools facilitates big data expression analysis of transposable elements and reveals an antagonism between their activity and that of piRNA genes. *Nucleic Acids Research*, 45(4), e17. <https://doi.org/10.1093/nar/gkw953>
- Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A., Durbin, R., Edwards, S. V., Forest, F., Gilbert, M. T. P., Goldstein, M. M., Grigoriev, I. V., Hackett, K. J., Haussler, D., Jarvis, E. D., Johnson, W. E.,

- Patrinos, A., Richards, S., Castilla-Rubio, J. C., ... Zhang, G. (2018). Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(17), 4325–4333. <https://doi.org/10.1073/pnas.1720115115>
- Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, *12*(1), 323. <https://doi.org/10.1186/1471-2105-12-323>
- Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btp698>
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btp352>
- Lisch, D. (2013). How important are transposons for plant evolution? *Nature Reviews Genetics*, *14*(1), 49–61. <https://doi.org/10.1038/nrg3374>
- Liu, Y., & Yang, G. (2014). Tc 1-like transposable elements in plant genomes. *Mobile DNA*, *5*(1), 17. <https://doi.org/10.1186/1759-8753-5-17>
- Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 1–21. <https://doi.org/10.1186/s13059-014-0550-8>
- Lynch, M., & Walsh, B. (2007). *The origins of genome architecture* (Vol. 98). Sinauer Associates Sunderland, MA.
- Martin, R. R., Zhou, J., & Tzanetakis, I. E. (2011). Blueberry latent virus: An amalgam of the Partitiviridae and Totiviridae. *Virus Research*, *155*(1), 175–180. <https://doi.org/10.1016/j.virusres.2010.09.020>
- Maumus, F., Epert, A., Nogué, F., & Blanc, G. (2014). Plant genomes enclose footprints of past infections by giant virus relatives. *Nature Communications*, *5*(May). <https://doi.org/10.1038/ncomms5268>
- Mcclintock, B. (1940). *The stability of broken ends of chromosomes in Zea mays*. 3.
- Mcclintock, B. (1950). The origin and behavior of mutable loci in maize. *PNAS*, *36*(6), 344–355.
- Medina, R., Johnson, M. G., Liu, Y., Wickett, N. J., Shaw, A. J., & Goffinet, B. (2019). Phylogenomic delineation of Physcomitrium (Bryophyta: Funariaceae) based on targeted sequencing of nuclear exons and their flanking regions rejects the retention of Physcomitrella, Physcomitridium and Aphanorrhagma. *Journal of Systematics and Evolution*, *57*(4), 404–417.
- Mhiri, C., Morel, J.-B., Vernhettes, S., Casacuberta, J. M., Lucas, H., & Grandbastien, M.-A. (1997). The promoter of the tobacco Tnt1 retrotransposon is induced by wounding and by abiotic stress. *Plant Molecular Biology*, *33*(2), 257–266.
- Michael, T. P., & Vanburen, R. (2020). ScienceDirect Building near-complete plant genomes. *Current Opinion in Plant Biology*, *54*, 26–33. <https://doi.org/10.1016/j.pbi.2019.12.009>
- Moore, J. K., & Haber, J. E. (1996). Capture of retrotransposon DNA at the sites of

- chromosomal double-strand breaks. *Nature*, 383(6601), 644–646.
<https://doi.org/10.1038/383644a0>
- Naish, M., Alonge, M., Wlodzimierz, P., Tock, A. J., Abramson, B. W., Schümcker, A., Mandáková, T., Jamge, B., Lambing, C., & Kuo, P. (2021). The genetic and epigenetic landscape of the Arabidopsis centromeres. *Science*, 374(6569), eabi7489.
- Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution*, 32(1), 268–274.
<https://doi.org/10.1093/molbev/msu300>
- Ono, R., Ishii, M., Fujihara, Y., Kitazawa, M., Usami, T., Kaneko-Ishino, T., Kanno, J., Ikawa, M., & Ishino, F. (2015a). Double strand break repair by capture of retrotransposon sequences and reverse-transcribed spliced mRNA sequences in mouse zygotes. *Scientific Reports*, 5(July), 1–5. <https://doi.org/10.1038/srep12281>
- Ortiz-Ramírez, C., Hernandez-Coronado, M., Thamm, A., Catarino, B., Wang, M., Dolan, L., Feijó, J. A., & Becker, J. D. (2016). A Transcriptome Atlas of *Physcomitrella patens* Provides Insights into the Evolution and Development of Land Plants. *Molecular Plant*, 9(2), 205–220.
<https://doi.org/https://doi.org/10.1016/j.molp.2015.12.002>
- Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J. R. A., Hellinga, A. J., Lugo, C. S. B., Elliott, T. A., Ware, D., Peterson, T., Jiang, N., Hirsch, C. N., & Hufford, M. B. (2019). Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biology*, 20(1), 1–18.
<https://doi.org/10.1186/s13059-019-1905-y>
- Perroud, P.-F., Guyon-Debast, A., Veillet, F., Kermarrec, M.-P., Chauvin, L., Chauvin, J.-E., Gallois, J.-L., & Nogué, F. (2022). Prime Editing in the model plant *Physcomitrium patens* and its potential in the tetraploid potato. *Plant Science : An International Journal of Experimental Plant Biology*, 316, 111162.
<https://doi.org/10.1016/j.plantsci.2021.111162>
- Perroud, P. F., Haas, F. B., Hiss, M., Ullrich, K. K., Alboresi, A., Amirebrahimi, M., Barry, K., Bassi, R., Bonhomme, S., Chen, H., Coates, J. C., Fujita, T., Guyon-Debast, A., Lang, D., Lin, J., Lipzen, A., Nogué, F., Oliver, M. J., Ponce de León, I., ... Rensing, S. A. (2018). The *Physcomitrella patens* gene atlas project: large-scale RNA-seq based expression data. *Plant Journal*, 95(1), 168–182.
<https://doi.org/10.1111/tpj.13940>
- Piegu, B., Guyot, R., Picault, N., Roulin, A., Sanyal, A., Kim, H., Collura, K., Brar, D. S., Jackson, S., Wing, R. A., & Panaud, O. (2006). Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Research*, 16(10), 1262–1269.
<https://doi.org/10.1101/gr.5290206>
- Quadrana, L., Etcheverry, M., Gilly, A., Caillieux, E., Madoui, M.-A., Guy, J., Bortolini Silveira, A., Engelen, S., Baillet, V., Wincker, P., Aury, J.-M., & Colot, V. (2019). Transposition favors the generation of large effect mutations that may facilitate rapid adaption. *Nature Communications*, 10(1), 3421.
<https://doi.org/10.1038/s41467-019-11385-5>

- Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*.
<https://doi.org/10.1093/bioinformatics/btq033>
- Rensing, S. A., Gof, B., Meyberg, R., Wu, S., & Bezanilla, M. (2020). *The Moss Physcomitrium (Physcomitrella) patens : A Model Organism for Non-Seed Plants OPEN*. 32(May), 1361–1376. <https://doi.org/10.1105/tpc.19.00828>
- Rensing, S. A., Lang, D., Zimmer, A. D., Terry, A., Salamov, A., Shapiro, H., Nishiyama, T., Perroud, P.-F., Lindquist, E. A., Kamisugi, Y., & others. (2008). The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. *Science*, 319(5859), 64–69.
- Rensing, S. A., Rombauts, S., de Peer, Y., & Reski, R. (2002). Moss transcriptome and beyond. *Trends in Plant Science*, 7(12), 535–538.
- Rishishwar, L., Mariño-Ramírez, L., & Jordan, I. K. (2017). Benchmarking computational tools for polymorphic transposable element detection. *Briefings in Bioinformatics*, 18(6), 908–918. <https://doi.org/10.1093/bib/bbw072>
- Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, 29(1), 24–26. <https://doi.org/10.1038/nbt.1754>
- Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2009). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1), 139–140. <https://doi.org/10.1093/bioinformatics/btp616>
- Saleh, O., Issman, N., Seumel, G. I., Stav, R., Samach, A., Reski, R., & Frank, W. (2011). *MicroRNA534a control of BLADE-ON-PETIOLE 1 and 2 mediates juvenile-to-adult gametophyte transition in Physcomitrella patens*. 1, 661–674. <https://doi.org/10.1111/j.1365-313X.2010.04451.x>
- Santiago, N., Herráiz, C., Goñi, J. R., Messeguer, X., & Casacuberta, J. M. (2002). Genome-wide Analysis of the Emigrant Family of MITEs of Arabidopsis thaliana. *Molecular Biology and Evolution*, 19(12), 2285–2293. <https://doi.org/10.1093/oxfordjournals.molbev.a004052>
- Satheesh, V., Fan, W., Chu, J., & Cho, J. (2021). Recent advancement of NGS technologies to detect active transposable elements in plants. *Genes and Genomics*, 43(3), 289–294. <https://doi.org/10.1007/s13258-021-01040-z>
- Sauvage, C., Segura, V., Bauchet, G., Stevens, R., Do, P. T., Nikoloski, Z., Fernie, A. R., & Causse, M. (2014). Genome-Wide Association in Tomato Reveals 44 Candidate Loci for Fruit Metabolic Traits. *Plant Physiology*, 165(3), 1120–1132. <https://doi.org/10.1104/pp.114.241521>
- Saze, H., & Kakutani, T. (2007). Heritable epigenetic mutation of a transposon-flanked Arabidopsis gene due to lack of the chromatin-remodeling factor DDM1. *The EMBO Journal*, 26(15), 3641–3652. <https://doi.org/10.1038/sj.emboj.7601788>
- Schaefer, D. G., Delacote, F., Charlot, F., Vrielynck, N., Guyon-Debast, A., Le Guin, S., Neuhaus, J. M., Doutriaux, M. P., & Nogué, F. (2010). RAD51 loss of function abolishes gene targeting and de-represses illegitimate integration in the moss Physcomitrella patens. *DNA Repair*. <https://doi.org/10.1016/j.dnarep.2010.02.001>

- Schaff, D. A. (1994). The adenine phosphoribosyltransferase (APRT) selectable marker system. *Plant Science*, *101*(1), 3–9. [https://doi.org/https://doi.org/10.1016/0168-9452\(94\)90159-7](https://doi.org/https://doi.org/10.1016/0168-9452(94)90159-7)
- Schwarz, R., Koch, P., Wilbrandt, J., & Hoffmann, S. (2022). Locus-specific expression analysis of transposable elements. *Briefings in Bioinformatics*, *23*(1), 1–10. <https://doi.org/10.1093/bib/bbab417>
- Shahid, S., & Slotkin, R. K. (2020). The current revolution in transposable element biology enabled by long reads. *Current Opinion in Plant Biology*, *54*, 49–56. <https://doi.org/10.1016/j.pbi.2019.12.012>
- Sharma, A., & Presting, G. G. (2014). Evolution of centromeric retrotransposons in grasses. *Genome Biology and Evolution*, *6*(6), 1335–1352. <https://doi.org/10.1093/gbe/evu096>
- Šola, I., Rusak, G., & Ludwig-Müller, J. (2022). Response of the moss *Physcomitrium patens* to satellite-associated cucumber mosaic virus infection on the level of salicylic acid, quercetin and indole-3-acetic acid. *European Journal of Plant Pathology*, 441–452. <https://doi.org/10.1007/s10658-022-02487-w>
- Spindel, J. E., Begum, H., Akdemir, D., Collard, B., Redoña, E., Jannink, J. L., & McCouch, S. (2016). Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. *Heredity*, *116*(4), 395–408.
- Strange, A., Li, P., Lister, C., Anderson, J., Warthmann, N., Shindo, C., Irwin, J., Nordborg, M., & Dean, C. (2011). Major-effect alleles at relatively few loci underlie distinct vernalization and flowering variation in *Arabidopsis* accessions. *PloS One*, *6*(5), e19949. <https://doi.org/10.1371/journal.pone.0019949>
- Sultana, T., Zamborlini, A., Cristofari, G., & Lesage, P. (2017). Integration site selection by retroviruses and transposable elements in eukaryotes. *Nature Reviews Genetics*, *18*(5), 292–308. <https://doi.org/10.1038/nrg.2017.7>
- Sun, Y., Shang, L., Zhu, Q. H., Fan, L., & Guo, L. (2021). Twenty years of plant genome sequencing: achievements and challenges. *Trends in Plant Science*, 1–11. <https://doi.org/10.1016/j.tplants.2021.10.006>
- Szövényi, P., Gunadi, A., & Li, F.-W. (2021). Charting the genomic landscape of seed-free plants. *Nature Plants*, *7*(5), 554–565. <https://doi.org/10.1038/s41477-021-00888-z>
- Tan, T., Sun, Y., Peng, X., Wu, G., Bao, F., He, Y., Zhou, H., & Lin, H. (2017). ABSCISIC ACID INSENSITIVE3 Is Involved in Cold Response and Freezing Tolerance Regulation in *Physcomitrella patens*. *Frontiers in Plant Science*, *8*. <https://doi.org/10.3389/fpls.2017.01599>
- Tsukahara, S., Kawabe, A., Kobayashi, A., Ito, T., Aizu, T., Shin-i, T., Toyoda, A., Fujiyama, A., Tarutani, Y., & Kakutani, T. (2012). Centromere-targeted de novo integrations of an LTR retrotransposon of *Arabidopsis lyrata*. *Genes & Development*, *26*(7), 705–713. <https://doi.org/10.1101/gad.183871.111>
- VanBuren, R., Wai, C. M., Ou, S., Pardo, J., Bryant, D., Jiang, N., Mockler, T. C., Edger, P., & Michael, T. P. (2018). Extreme haplotype variation in the desiccation-

- tolerant clubmoss *Selaginella lepidophylla*. *Nature Communications*, 9(1), 13. <https://doi.org/10.1038/s41467-017-02546-5>
- Vanrobays, E., Gleizes, P. E., Bousquet-Antonelli, C., Noaillac-Depeyre, J., Caizergues-Ferrer, M., & Gélugne, J. P. (2001). Processing of 20S pre-rRNA to 18S ribosomal RNA in yeast requires Rrp10p, an essential non-ribosomal cytoplasmic protein. *The EMBO Journal*, 20(15), 4204–4213. <https://doi.org/10.1093/emboj/20.15.4204>
- Velasco, O. C. (2019). *Targeting the transposable elements of the genome to enable large-scale genome editing and bio-containment technologies*. Université Paris Saclay (COMUE).
- Vendrell-Mir, P., Barteri, F., Merenciano, M., González, J., Casacuberta, J. M., & Castanera, R. (2019). A benchmark of transposon insertion detection tools using real data. *Mobile DNA*, 10(1), 1–19. <https://doi.org/10.1186/s13100-019-0197-9>
- Vendrell-Mir, P., López-Obando, M., Nogué, F., & Casacuberta, J. M. (2020). Different Families of Retrotransposons and DNA Transposons Are Actively Transcribed and May Have Transposed Recently in *Physcomitrium* (*Physcomitrella*) patens. *Frontiers in Plant Science*, 11(August), 1–13. <https://doi.org/10.3389/fpls.2020.01274>
- Vendrell-Mir, P., Perroud, P. F., Haas, F. B., Meyberg, R., Charlot, F., Rensing, S. A., Nogué, F., & Casacuberta, J. M. (2021). A vertically transmitted amalgavirus is present in certain accessions of the bryophyte *Physcomitrium patens*. *Plant Journal*, 108(6), 1786–1797. <https://doi.org/10.1111/tpj.15545>
- Vicient, C. M., & Casacuberta, J. M. (2017). Impact of transposable elements on polyploid plant genomes. *Annals of Botany*, 120(2), 195–207. <https://doi.org/10.1093/aob/mcx078>
- Vives, C., Charlot, F., Mhiri, C., Contreras, B., Daniel, J., Epert, A., Voytas, D. F., Grandbastien, M.-A., Nogué, F., & Casacuberta, J. M. (2016). Highly efficient gene tagging in the bryophyte *Physcomitrella patens* using the tobacco (*Nicotiana tabacum*) Tnt1 retrotransposon. *New Phytologist*, 212(3), 759–769. <https://doi.org/https://doi.org/10.1111/nph.14152>
- Wang, B., Yang, X., Jia, Y., Xu, Y., Jia, P., Dang, N., Wang, S., Xu, T., Zhao, X., Gao, S., Dong, Q., & Ye, K. (2021). High-quality *Arabidopsis thaliana* Genome Assembly with Nanopore and HiFi Long Reads. *Genomics, Proteomics & Bioinformatics*. <https://doi.org/https://doi.org/10.1016/j.gpb.2021.08.003>
- Wei, L., & Cao, X. (2016). The effect of transposable elements on phenotypic variation: insights from plants to humans. *Science China Life Sciences*, 59(1), 24–37. <https://doi.org/10.1007/s11427-015-4993-2>
- Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., & Schulman, A. H. (2009). Reply: A unified classification system for eukaryotic transposable elements should reflect their phylogeny. *Nature Reviews Genetics*, 10(4), 276. <https://doi.org/10.1038/nrg2165-c4>
- Widiez, T., Symeonidi, A., Luo, C., Lam, E., Lawton, M., & Rensing, S. A. (2014). The chromatin landscape of the moss *Physcomitrella patens* and its dynamics during

- development and drought stress. *Plant Journal*, 79(1), 67–81.
<https://doi.org/10.1111/tpj.12542>
- Wolf, L., Rizzini, L., Stracke, R., Ulm, R., & Rensing, S. A. (2010). The Molecular and Physiological Responses of *Physcomitrella patens* to Ultraviolet-B Radiation . *Plant Physiology*, 153(3), 1123–1134. <https://doi.org/10.1104/pp.110.154658>
- Wu, H., Fu, P., Fu, Q., Zhang, Z., Zheng, H., Mao, L., Li, X., Yu, F., & Peng, Y. (2022). Plant Virus Database: a resource for exploring the diversity of plant viruses and their interactions with hosts. *BioRxiv*.
- Xiao, H., Jiang, N., Schaffner, E., Stockinger, E. J., & van der Knaap, E. (2008). A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. *Science (New York, N.Y.)*, 319(5869), 1527–1530.
<https://doi.org/10.1126/science.1153040>
- Xiao, Y., Liu, H., Wu, L., Warburton, M., & Yan, J. (2017). Genome-wide Association Studies in Maize: Praise and Stargaze. *Molecular Plant*, 10(3), 359–374.
<https://doi.org/https://doi.org/10.1016/j.molp.2016.12.008>
- Xu, Z.-S., Yang, Q.-Q., Feng, K., & Xiong, A.-S. (2019). Changing carrot color: insertions in DcMYB7 alter the regulation of anthocyanin biosynthesis and modification. *Plant Physiology*, 181(1), 195–207.
- Yadav, V., Sun, S., Coelho, M. A., & Heitman, J. (2020). *Centromere scission drives chromosome shuffling and reproductive isolation*. 1–12.
<https://doi.org/10.1073/pnas.1918659117>
- Ye, K., Schulz, M. H., Long, Q., Apweiler, R., & Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics (Oxford, England)*, 25(21), 2865–2871. <https://doi.org/10.1093/bioinformatics/btp394>
- Zhang, L., Hu, J., Han, X., Li, J., Gao, Y., Richards, C. M., Zhang, C., Tian, Y., Liu, G., Gul, H., Wang, D., Tian, Y., Yang, C., Meng, M., Yuan, G., Kang, G., Wu, Y., Wang, K., Zhang, H., ... Cong, P. (2019). A high-quality apple genome assembly reveals the association of a retrotransposon and red fruit colour. *Nature Communications*, 10(1), 1494. <https://doi.org/10.1038/s41467-019-09518-x>
- Zhang, Y., Cheng, T. C., Huang, G., Lu, Q., Surleac, M. D., Mandell, J. D., Pontarotti, P., Petrescu, A. J., Xu, A., Xiong, Y., & Schatz, D. G. (2019). Transposon molecular domestication and the evolution of the RAG recombinase. *Nature*, 569(7754), 79–84. <https://doi.org/10.1038/s41586-019-1093-7>
- Zhou, H., Liu, B., Weeks, D. P., Spalding, M. H., & Yang, B. (2014). Large chromosomal deletions and heritable small genetic changes induced by CRISPR/Cas9 in rice. *Nucleic Acids Research*, 42(17), 10903–10914.
<https://doi.org/10.1093/nar/gku806>

ACKNOWLEDGMENTS

ACKNOWLEDGMENTS

La llista de persones a qui he d'agrair és llarga i espero no deixar-me a ningú.

M'agradaria començar pel Pep, per la confiança dipositada en mi durant tot aquest temps, la llibertat que m'ha deixat per explorar, tot fent ciència, pel mentoratge i tot l'aprenentatge que m'ha transmès sobre el fascinant món dels transposons i sobretot per la paciència que ha tingut durant tot aquest temps.

També a la resta de membres del meu grup, començant per les dues persones que em van introduir al món dels transposons i la molssa i que han sigut part fonamental en que hagi acabat fent el doctorat. La Cristina i la Bea. Continuant per els dos primers bioinformàtics amb qui vaig tenir ocasió d'interaccionar i que em van ensenyar els bàsics per sobreviure en el món de la bioinformàtica el Jordi i el Marc. Al Carlos Vicent per tot el suport científic rebut durant els lab meetings. Al Fabio, la Fàtima, el Pedro, l'Amelie i el Miguel per tota l'ajuda rebuda i les bones estones al laboratori. També he d'agrair tota l'ajuda i suport rebut durant els darrers anys per el Carlos, la Noemia i l'Andrea tant al laboratori com fora d'ell. I al Raúl pel suport, guiatge, les seves meravelloses idees i les seves constants ganes d'ajudar que han permès molt sovint desencallar o tirar endavant projectes que d'altra manera no haurien sigut possibles. Finalment, vull agrair a tots aquells estudiants que he tingut la sort de compartir estones al laboratori i intentar ensenyar-los encara que crec que he après jo més d'ells. Moltes gràcies Amàlia, Lana i Marc. Amb vosaltres tot ha sigut molt fàcil.

Em cal agrair també tot el suport rebut per part del grup de la Soraya. A la pròpia Soraya, a la Sandra, la Jonice, les Andrees, la Poonam, el Luis, la Fátima, el Juanjo, el Carlos Cu, l'Unai i sobretot a la Michela per tot el suport i aprenentatge que em va donar en els temps que estàvem sols al laboratori i anava bastant perdut per allà. També al Nacho que ha aguantat totes les meves constants preguntes i paranoies estoicament durant tots aquests anys.

Je tiens à remercier Fabien Nogué et toute son équipe, en particulier Florence, Anoucka, Pierre-François, Julie, Louanne et Guillaume. Pour toute l'aide fournie pendant mes deux séjours, pour avoir rendu tout si facile pour moi pendant que j'étais là et pour m'avoir supporté pendant ces périodes ainsi que pour l'aide constante fournie pendant tout le doctorat. Je suis très reconnaissant d'avoir pu vous rencontrer. Je tiens également à remercier les personnes du bâtiment social et toutes les autres personnes du bâtiment 7

pour les bons moments que j'ai passés à l'INRA-Versailles. Quiero agradecer también a Leandro y a todo su grupo por las discusiones tenidas durante mi visita a Paris.

A tot el suport rebut per tot el personal del CRAG, sigui del servei que sigui, d'hivernacles, genòmica, bioinformàtica, informàtica, administració, seqüenciació així com la resta d'habitants d'aquest edifici. Tot ha sigut molt fàcil amb vosaltres.

Per acabar, vull agrair tot el suport rebut per tota la família i amistats. A la Xènia per la portada. A les companyes de pis, les Laures, a les colles del poble i als companys de carrera que estan tots dispersos pel món. I sobretot als meus pares, el meu germà i els avis. No crec que hagués començat ni acabat mai el doctorat sinó hagués sigut per vosaltres. I a l'Eva que mitjançant el teu suport has fet que aquest viatge meravellós que ha sigut la tesis hagi resultat molt més fàcil.

ANNEXES

ANNEXES

Other articles published during the thesis in collaboration with other projects:

- Bogaerts-Marquez, M., Barron, M. G., Fiston-Lavier, A. S., Vendrell-Mir, P., Castanera, R., Casacuberta, J. M., & Gonzalez, J. (2020). T-lex3: An accurate tool to genotype and estimate population frequencies of transposable elements using the latest short-read whole genome sequencing data. *Bioinformatics*, 36(4), 1191–1197. <https://doi.org/10.1093/bioinformatics/btz727>

- Castanera, R., Vendrell-Mir, P., Bardil, A., Carpentier, M. C., Panaud, O., & Casacuberta, J. M. (2021). Amplification dynamics of miniature inverted-repeat transposable elements and their impact on rice trait variability. In *Plant Journal* (Vol. 107, Issue 1). <https://doi.org/10.1111/tpj.15277>

- Guyon-Debast, A., Alboresi, A., Terret, Z., Charlot, F., Berthier, F., Vendrell-Mir, P., Casacuberta, J. M., Veillet, F., Morosinotto, T., Gallois, J. L., & Nogué, F. (2021). A blueprint for gene function analysis through Base Editing in the model plant *Physcomitrium* (*Physcomitrella*) patens. *New Phytologist*, 230(3), 1258–1272. <https://doi.org/10.1111/nph.17171>

Genetics and population analysis

***T-lex3*: an accurate tool to genotype and estimate population frequencies of transposable elements using the latest short-read whole genome sequencing data**

María Bogaerts-Márquez¹, Maite G. Barrón¹, Anna-Sophie Fiston-Lavier², Pol Vendrell-Mir³, Raúl Castanera³, Josep M. Casacuberta³, and Josefa González^{1,*}

¹Institute of Evolutionary Biology (CSIC-Universitat Pompeu Fabra), ²Institut des Sciences de l'Evolution de Montpellier (UMR 5554, CNRS-UM-IRD-EPHE), ¹¹ Université de Montpellier, Place Eugène Bataillon, Montpellier, France, ³ Center for Research in Agricultural Genomics, CRAG (CSIC-IRTA-UAB-UB), Campus UAB, Cerdanyola del Vallès, Barcelona, Spain

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Transposable elements (TEs) constitute a significant proportion of the majority of genomes sequenced to date. TEs are responsible for a considerable fraction of the genetic variation within and among species. Accurate genotyping of TEs in genomes is therefore crucial for a complete identification of the genetic differences among individuals, populations, and species.

Results: In this work, we present a new version of *T-lex*, a computational pipeline that accurately genotypes and estimates the population frequencies of reference TE insertions using short-read high-throughput sequencing data. In this new version, we have re-designed the *T-lex* algorithm to integrate the BWA-MEM short-read aligner, which is one of the most accurate short-read mappers and can be launched on longer short-reads (e.g. reads >150 bp). We have added new filtering steps to increase the accuracy of the genotyping, and new parameters that allow the user to control both the minimum and maximum number of reads, and the minimum number of strains to genotype a TE insertion. We also showed for the first time that *T-lex3* provides accurate TE calls in a plant genome.

Availability: To test the accuracy of *T-lex3*, we called 1,630 individual TE insertions in *Drosophila melanogaster*, 1,600 individual TE insertions in humans, and 3,067 individual TE insertions in the rice genome. We showed that this new version of *T-lex* is a broadly applicable and accurate tool for genotyping and estimating TE frequencies in organisms with different genome sizes and different TE contents. *T-lex3* is available at Github: <https://github.com/GonzalezLab/T-lex3>

Contact: josefa.gonzalez@ibe.upf-csic.es

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Amplification dynamics of miniature inverted-repeat transposable elements and their impact on rice trait variability

Raúl Castanera^{1,*} , Pol Vendrell-Mir¹ , Amélie Bardil¹ , Marie-Christine Carpentier² , Olivier Panaud²  and Josep M. Casacuberta^{1,*} 

¹Centre for Research in Agricultural Genomics CSIC-IRTA-UAB-UB, Campus UAB, Edifici CRAG, Bellaterra, Barcelona 08193, Spain, and

²Laboratoire Génome et Développement des Plantes, UMR CNRS/UPVD 5096, Université de Perpignan Via Domitia, 52 Avenue Paul Alduy, Perpignan Cedex 66860, France

revised 31 March 2021; accepted 8 April 2021; published online 18 April 2021.

*For correspondence (e-mail raul.castanera@cragenomica.es; josep.casacuberta@cragenomica.es).

SUMMARY

Transposable elements (TEs) are a rich source of genetic variability. Among TEs, miniature inverted-repeat TEs (MITEs) are of particular interest as they are present in high copy numbers in plant genomes and are closely associated with genes. MITEs are deletion derivatives of class II transposons, and can be mobilized by the transposases encoded by the latter through a typical cut-and-paste mechanism. However, MITEs are typically present at much higher copy numbers than class II transposons. We present here an analysis of 103 109 transposon insertion polymorphisms (TIPs) in 738 *Oryza sativa* genomes representing the main rice population groups. We show that an important fraction of MITE insertions has been fixed in rice concomitantly with its domestication. However, another fraction of MITE insertions is present at low frequencies. We performed MITE TIP-genome-wide association studies (TIP-GWAS) to study the impact of these elements on agronomically important traits and found that these elements uncover more trait associations than single nucleotide polymorphisms (SNPs) on important phenotypes such as grain width. Finally, using SNP-GWAS and TIP-GWAS we provide evidence of the replicative amplification of MITEs.

Keywords: miniature inverted-repeat transposable element, transposable elements, genome-wide association studies, traits, rice, transposition, genetic factor.

Methods

A blueprint for gene function analysis through Base Editing in the model plant *Physcomitrium (Physcomitrella) patens*

Anouchka Guyon-Debast¹ , Alessandro Alboresi² , Zoé Terret³ , Florence Charlot¹ , Floriane Berthier¹ , Pol Vendrell-Mir⁴ , Josep M. Casacuberta⁴ , Florian Veillet⁵ , Tomas Morosinotto² , Jean-Luc Gallois³  and Fabien Nogué¹ 

¹Institut Jean-Pierre Bourgin, INRAE, AgroParisTech, Université Paris-Saclay, Versailles 78000, France; ²Department of Biology, University of Padova, Padova 35121, Italy; ³INRAE, GAFL, Montfavet 84143, France; ⁴Centre for Research in Agricultural Genomics CSIC-IRTA-UAB-UB, Campus UAB, Edifici CRAG, Bellaterra, Barcelona 08193, Spain; ⁵IGEPP, INRAE, Institut Agro, Université de Rennes, Ploudaniel 29260, France

Author for correspondence:
Fabien Nogué
Email: Fabien.nogue@inrae.fr

Received: 5 October 2020
Accepted: 21 December 2020

New Phytologist (2021) 230: 1258–1272
doi: 10.1111/nph.17171

Key words: adenine deaminase, APRT, base editing, Cas9, CRISPR, cytosine deaminase, *Physcomitrella patens*, *Physcomitrium patens*.

Summary

- CRISPR-Cas9 has proven to be highly valuable for genome editing in plants, including the model plant *Physcomitrium patens*. However, the fact that most of the editing events produced using the native Cas9 nuclease correspond to small insertions and deletions is a limitation.
- CRISPR-Cas9 base editors enable targeted mutation of single nucleotides in eukaryotic genomes and therefore overcome this limitation. Here, we report two programmable base-editing systems to induce precise cytosine or adenine conversions in *P. patens*.
- Using cytosine or adenine base editors, site-specific single-base mutations can be achieved with an efficiency up to 55%, without off-target mutations. Using the *APT* gene as a reporter of editing, we could show that both base editors can be used in simplex or multiplex, allowing for the production of protein variants with multiple amino-acid changes. Finally, we set up a co-editing selection system, named selecting modification of APRT to report gene targeting (SMART), allowing up to 90% efficiency site-specific base editing in *P. patens*.
- These two base editors will facilitate gene functional analysis in *P. patens*, allowing for site-specific editing of a given base through single sgRNA base editing or for *in planta* evolution of a given gene through the production of randomly mutagenised variants using multiple sgRNA base editing.