NGS applications in genome evolution and adaptation

A reproducible approach to NGS data analysis and integration

Pablo Prieto Barja

TESI DOCTORAL UPF / 2016

DIRECTOR DE LA TESI

Dr Cedric Notredame

BIOINFORMATICS AND GENOMICS PROGRAMME AT CRG

CENTRE FOR GENOMIC REGULATION

Acknowledgments

During the last years I had the invaluable opportunity of undergoing a PhD at the Comparative Bioinformatics lab with Cedric Notredame in the CRG. But it was thanks to Dr Xavier Messeguer, with whom I had the opportunity to collaborate and work tightly, who recommended and pointed me to the Msc in Bioinformatics for Health Sciences and PhD program at the UPF. Without his help and guidance I wouldn't probably have entered the program and for that I will be always grateful.

This time couldn't had been better at the CRG, the institute that provided a marvellous scientific environment in which to look upon many outstanding scientists and so much invaluable research experiences motivating me and making me enjoying even further science.

It would not have been possible without the chance to take my PhD with Dr. Cedric Notredame and receive his guidance and support. My personal experience in his laboratory and under his tutoring has been very enjoyable and extraordinary. It has been specially refreshing to see through his eyes another view on how to envision and approach new scientific ideas with enthusiasm and without hesitation. Both his scientific advice and challenging views had a strong impact on my views and way of enjoying research.

I would like to thank all scientists with whom I had the opportunity to interact, especially Ionas Erb with whom I collaborated in many projects and enjoyed our time dreaming about Mexico, dinosaurs and cats. Also to all my colleagues in Cedric's group, people with whom I had the opportunity to share time working, sharing knowledge and having fruitful discussions. Even outside of the working place I found some of the best friends for the last ongoing years of my life.

I would like to thank Dr. Roderic Guigo for taking me into consideration and giving me the possibility to participate in outstanding projects and collaborations from which I have seen up to which extent large efforts and amazing science can be set up, and how important it is to work in large community efforts that can end up transforming and shaping scientific research in the next years to come.

Finally, I would like to thank Dr Gerald Spaeth and Cedric Notredame for the opportunity to carry and get deeply involved in projects that have made my passion and excitement for science go beyond my limits, which has left a mark in me, both scientifically and personally, shaping my career and my future as a scientist.

Abstract

In this PhD I have used NGS technologies in different organisms and scenarios such as in ENCODE, comparing the conservation and evolution of long non-coding RNA sequences between human and mouse, using experimental evidences from genome, transcriptome and chromatin. A similar approach was followed in other organisms such as the mesoamerican common bean and in chicken. Other analysis carried with NGS data involved the well known parasite, Leishmania Donovani, the causative agent of Leishmaniasis. I used NGS data obtained from genome and transcriptome to study the fate of its genome in survival strategies for adaptation and long term evolution. All this work was approached while working in tools and strategies to efficiently design and implement the bioinformatics analysis also known as pipelines or workflows, in order to make them easy to use, easily deployable, accessible and highly performing. This work has provided several strategies in order to avoid lack of reproducibility and inconsistency in scientific research with real biological applications towards sequence analysis and genome evolution.

Resum

En aquest doctorat he utilitzat tecnologies NGS en diferents organismes i projectes com l'ENCODE, comparant la conservació i evolució de seqüències de RNA llargs no codificant entre el ratolí i l'humà, utilitzant evidències experimentals del genoma, transcriptoma i cromatina. He seguit una estratègia similar en altres organismes com són la mongeta mesoamericana i el pollastre. En altres anàlisis he hagut d'utilitzar dades NGS en l'estudi del conegut paràsit leishmània Donovani, l'agent causatiu de la malaltia Leishmaniosis. Utilitzant dades NGS obtingudes del genoma i transcriptoma he estudiat les conseqüències del genoma en estratègies d'adaptació i evolució a llarg termini. Aquest treball es va realitzar mentre treballava en eines i estratègies per dissenyar eficientment i implementar els anàlisis bioinformàtics coneguts com a diagrames de treball, per tal de fer-los fàcils d'utilitzar, fàcilment realitzables, accessibles i amb un alt rendiment. Aquest treball present diverses estratègies per tal d'evitar la falta de reproductibilitat i consistència en la investigació científica amb aplicacions reals a la biologia de l'anàlisi de seqüències i evolució de genomes.

Preface

At the beginning of my PhD I found myself entering into areas of NGS applications which were used in a wide range of topics and questions with an increasing number of research, knowledge, methods, tools and data being made publicly available in a very rapid way. One of my concerns was on how to make use of that knowledge and data while trying to follow the same protocols that had been already published and described. To my surprise that presented to be a much more complicated problem than expected, and my main concern towards sharing my scientific research. While other analysis were being published in similar areas of my research and studies, I was finding myself quite often having difficulties to find all the information published and reproduce the strategies presented in order to include them into my own analysis, or to just validate other approaches and results. This year's and the experience presented here have driven me through my main biological questions on sequence evolution and allowed me to satisfactorily address interesting questions. The work presented here shows some guidelines and effective strategies applied to real biological questions to avoid a downward spiral of irreproducibility.

Keywords

Reproducibility, evolution, genome, parasite, long non-coding RNA

Index

1. Introduction

1.1. Towards high throughput

Sequences are the core of bioinformatics and computational biology. But before reaching to that point, before the sequence and sequencing era, biology was not data driven at all. This was in part due to the data scarcity and the lack of computational power and dominion. Bioinformatics was coined as the convergence and merging process between pre-existing but yet a set of unmatured techniques and theories. The integration and reduction of other sciences attempting to explain biological processes such as physics, chemistry and mathematics were instrumentalized in order to fortify scientific research in biology. This process of integration lead to the use of methods and techniques developed for the computational analysis of sequences in biology. But before the settlement of bioinformatics through history we could see science paving the way to current approaches in the field to gather, transform, integrate and process biological data. During last 70 years researchers experienced and witnessed key events in research, which conceived and transformed computational biology.

Although nowadays sometimes in the sequencing era bioinformatics origin is pointed towards the origins of DNA sequencing, indeed its foundations and roots can be traced further away than that. Before sequencing capabilities were achieved in the 40's researched was focus in studying proteins, which are the first sequences determined, and not DNA and RNA which due to the smaller alphabet made them more difficult to distinguished at the beginning. In that decade new chromatography techniques were presented (Martin and Synge 1941) and further improved (Sanger 1945; Edman 1949) which allowed describing single amino acid composition. Most of the efforts carried in identifying proteins on that decade were related to hormone proteins, which were easier to purify, and with a size and amount practical to work with. With the techniques to read amino acid sequences it became more evident the importance of sequence content and changes within them, and the relation to the structure and function, as it is the case of a described single mutation in hemoglobin producing sickle cell (Pauling and Itano 1949). But it wasn't until later in the 60's that this process was automated (Edman and Begg 1967) allowing faster rate of sequence identification. The number of protein identified and its publications containing its sequences started to increase. Some efforts were made in order to increase the number of these sequences publishing the first computer programs in FORTRAN able to predict the sequence and structure composition (Dayhoff 1965). In order to make available all sequences, a collection of protein sequences was published every year (Dayhoff and National Biomedical Resea...).

Although the number of sequences remained low, around 100, it settled the precedent on how to start organizing collections of sequences. The atlas of amino acid sequences contained as well tables of frequencies for amino acid substitutions, PAM (Dayhoff and Foundation 1979) from where others were influenced and established future computation resources like BLOSUM (Henikoff and Henikoff 1992). The availability of sequences to analyze allowed biology to be more quantitative, which allowed to apply statistical and mathematical approaches in the study of biological molecules and its processes. With sequences available scientists were able to compare them, one of the first analysis derived from new sequences available was applied to evolutionary biology with the first algorithm described for tree phylogenetic reconstruction in the late 60's (Fitch and Margoliash 1967). During the late 60's and the 70's followed years of development an application of new algorithms to molecular biology which are the foundations of today's analysis, such as the case in protein structures (Lee and Richards 1971), RNA structure prediction (Delisi and Crothers 1971), population genetics (Kimura 1969; Kimura and Ohta 1971) among others . Although computational biology started to settle down as a major field of research in biology, it wasn't until the middle of the 80's and beginning of the 90's that become more widely adopted with the creation of more resources in databases like GenBank (Bilofsky et al. 1986) and EMBL data library (Hamm and Cameron 1986). Across the next years more algorithms were developed and specially tools were made available which allowed more researchers to apply the techniques in their own analysis, as in the case of gene prediction (Fickett 1982; Shepherd 1981), RNA folding (Dumas and Ninio 1982) or multiple sequence alignments (Lipman et al. 1989; Higgins and Sharp 1988) to show some examples.

The exponential data increase that happened decades ago forged and pushed bioinformatics forward and made it an essential toolkit driving sequence analysis in computational biology becoming a standard in many topics such as gene prediction, homology, protein structure and phylogeny. And during the last decade it has geared and turned towards approaching massive sequence problems. In the new scenario data has moved from sets of sequences to an expanded to extensive amounts of high throughput datasets increasing the magnitude and impact of computational and data driven research, previously never thought. This new approach of tackling large data sets and collections of data has proven how the path opened up by bioinformatics allows to gain new insights of biological processes and systems much more complex using multilevel data integration approaches in a feasible way previously unexpected because of the complexity seen. But that raises the interesting question on how to scale up reasonably to be able to cope and manage all genomic data and the complexity that comes together with it.

1.2. One-to-many applications

Across the large bibliography of scientific publications, bioinformatics has taken its fair place among the highest ranked and cited papers of the era (Van Noorden et al. 2014). Among the first 10 most cited scientific papers we can find ClustalW (Thompson et al. 1994) a multiple sequence aligner program that allows comparing many sequences similarities. On the next most cited papers we can find other bioinformatics methods papers such as two versions of BLAST alignment tool already introduced (Altschul et al. 1990; Altschul et al. 1997). The high ranking of those methodological articles and its standing against other essential biological tools such as crystal structure determination (Sheldrick 2008) points to how important and prevalent bioinformatics analysis have become in molecular biology and its wide adoption in such a short period of time. In the case of BLAST the fact of having multiple versions of the highly cited tools (States and Gish 1994; Morgulis et al. 2008) for different applications shows how extensive can be the use and applications of similar methods such as sequence search for genomic vs protein spaces. In a wet lab exist different techniques and variations of a protocol in order to approach a similar problem, but in the case of bioinformatics one can see how this becomes a large issue with the ability of researchers to develop easily their own methods. Each method has its own flavors and specialties addressing problems in different ways and leading to (probably) different results. In order to evaluate and review tools for the same purpose every year reviews and benchmark papers are published as a way to keep track in the literature of all the advances and sheer some light into the best application for each case. Even though there has been for a long time some scientific journals devoted to the publication of literature reviews, nowadays it is very common to see practical bioinformatics reviews making it into these journals (Stein 2001; Lee et al. 2007; Li and Homer 2010). The most typical applications used in bioinformatics related to sequences are sequence search tools like BLAST, often benchmarked against new search methods (Ye et al. 2011; Hauser et al. 2016). Some of these tools do even have more than one publication of the same method including new updates (Zhao et al. 2012; Kim et al. 2013), which includes the improvements and add new approaches to tackle the increase in the amount of sequence data available during the last decades. In other publications scenarios, researchers from different groups are invited to take part in a contest where an effort is made in order to evaluate methods in a more neutral and fair environment in a more or less regular basis (Steijger et al. 2013; Earl et al. 2014; Kryshtafovych et al. 2009). In contrast, in other reviews authors gather themselves different datasets published in the literature and use them to test different tools whether or not they are the authors (Conesa et al. 2016). Benchmarking is full of caveats as not all the tools are comparable at the same scale, as they convey differences in formats, inputs, parameters and computational resources. All these issues have to be taken into account when benchmarking and trying to reproduce computations in order to

establish a fair and objective analysis, which will help take the decision upon which method to use for each purpose. As bioinformatics keep evolving, analysis and protocols keep changing and becoming sometimes more and more complex. Usually this means it will require various steps and various sources of data to produce results and finalize the analysis. The compilation of a series of steps, tools and computation is commonly known as workflow or pipeline . In a workflow each step and piece of data is required and connected to another step of the analysis being carried. Given the diverse nature of information, formats and applications, in bioinformatics there are tools mostly developed for specific steps or analysis, and pipelines end up being an integration of different tools, data and steps that need to be properly connected, thus rising the complexity of the protocol being carried.

1.3. Data big bang

High throughput data has been accumulating and the rate of generated keeps increasing (Baker 2010; Stephens et al. 2015). As an example of data with the pace of generation increasing rapidly, we have the case of yeast genomics. In the case of a single species, with a very well known and studied genome (Dujon 2010) with a long track of scientific research published and also studying different similar genomes species (Blackwell et al. 2016), an increase of 2000 genomes and tens of thousands of gene families has been seen in the last three years; together with an extreme increase in the number of bases per genome available as a consequence of the improvement in sequencing technologies and capabilities. Further away from a single species, the use of sequencing for metagenomics analysis in our environments (Tringe and Rubin 2005; Afshinnekoo et al. 2015) or within organisms (Zarowiecki 2012; Krishnan et al. 2014) is increasing the number as well, both the number of applications and sequences obtained using very similar technologies with variations in the protocols. The relationship between data generation and technology typically changes after a short period of time that takes for new technologies to be adopted. After that period the technology starts generating an exponential increase in the amount of data to analyze. Since method development usually goes behind or after the appearance of new technologies, and it only further develops once the technology is starting to be adopted, those methods typically will not be able to cope with an exponential rate of data generation thus lagging behind the ability to generate fast enough results out of such amounts of data. Further problems include how to store, adapt, scale and handle results in a more reliable way, since the drop of sequencing cost had a steady decrease, the computation and storage becomes a real issue to take into account, even in the cost estimation (Muir et al. 2016). As a consequence of that increase in sequencing capabilities and massive parallel sequencing reaching out to sequencing centers there has been a turn in the estimation of the overall

cost. As the technology keeps developing, the price for the technology has been lowering faster than Moore's law, but what was not taken into account is the need to maintain an infrastructure resilient and capable of both receiving and keeping massive amounts of data while retaining the capability to process it. If there is no focus on improving computing capabilities, data will become a bottleneck where it will be generated much faster than it can be processed. As an illustration for this scenario, the rate of data generation for the last generation of sequencers from Illumina the X-Ten, are made up of ten HiSeq X sequencers which end up being able of produce 320 samples per week, the equivalent to 18TB of data waiting to be processed in time. In order to process all the data without creating bottlenecks, it is important to develop solutions with the clear ideas of the limitations in mind. In the case of human genome sequencing in cancer and other disease related projects (Hudson (Chairperson) et al. 2010; The Cancer Genome Atlas Research Netw...; Telenti et al. 2016), sequencing its already gathering and promising to keep generating hundreds of thousands of genomes alongside and coupled to other sources of sequencing information, not only the genome. The accumulation and availability of large datasets present other problems related to the reutilization of data, how to access the source and raw data, how to organize it and anonymize it in order to allow proper sharing, replication and inclusion in new analysis (Eisenstein 2015). There are some directions on how to proceed on this matters in order to tackle the situation, but it is certainly a technological challenge how to organize, structure and access data in a practical way (Bourne et al. 2015; Global Alliance for Genomics and Heal...).

Molecular biology is facing new data growing at paramount speed, which has turn in the Big Data problem. Researchers might be facing with the issue of having to store dozens of petabytes of data, with large analysis and projects having to face new challenges involving technical aspects from Big Data. Although it is a new scenario in this research field, other fields once had to face the same problems and started developing solutions.

1.4. Reproducibility

1.4.1. Motivation

Replication of data and its results is considered the cornerstone of scientific research (Helene Richter et al. 2010). The central dogma relies on being able to generate scientific knowledge for which the procedure and specifics of the analysis should be described accurately and let any other researcher be able to reproduce and lead to the same results. The main problem with reproducibility that has drawn a lot of attention in the last years (Ioannidis 2005; Prinz et al. 2011; Announcement: Reducing our

irreproduc...; Further confirmation needed 2012; Error prone 2012; Must try harder 2012; Christakis and Zimmerman 2013; Gunn 2014; Freedman and Inglese 2014; Freedman et al. 2015; Roth and Cox 2015) is the fact that scientists expect that the content in published scientific journals to be clearly understood and used to replicate whenever is feasible due to the accessibility of proper resources (Roquet et al. 2014; Baud et al. 2014). In computational biology, since typically these resources can be acquired easily it would be expected that by reading a publication and with the data available whenever necessary, one could understand how is the computation done and be able to reproduce again the same results. One would believe that in computational biology, reproducibility would be easier as typically the results of it tends to be more quantitative and be precisely described and annotated (Davison 2012); whereas experimental cases results tending to be more qualitative described and with specific requirements to run experiments (Freedman et al. 2015; Cyranoski 2016). Contrary to the expectations the replication of results in computational biology is not as easy as would be expected, and is rarely the case (Ioannidis et al. 2001; Ioannidis et al. 2009; Firtina and Alkan 2016). As computational biology keeps advancing both computers and tools used to analyze the data become more and more complex, leading to the development of large protocols also known as pipelines or workflows. The level of skills and effort needed not only to perform analysis but also to re-analyze and integrate data already published into meta-analysis has increased during the last years (Nalls et al. 2014; Cheng et al. 2015; Wang et al. 2016). Although it may seem far easier to reproduce computational analysis than experimental work, due to the ever changing nature of environments and the fast development of both software and technology, which have to carry on with the pace of fast track publication records, reproducibility has become much more difficult than expected.

Scientific journals are well aware of the problem, there have been cases where, upon reproducibility issues, papers have been modified or even worse, calling for retraction because of the inability to reproduce the results claimed in publications (Gallego Llorente et al. 2015; Rhinn et al. 2013; Decullier et al. 2013). Academy and government institutions are also aware of the issue as they have started to implement rules towards how scientific results and data should be kept and reported for the sake of accessibility and reproducibility (Morin et al. 2012; Collins and Tabak 2014; Stodden et al. 2013; Announcement: Reducing our irreproduc...). New guidelines and recommendations have been described to solve certain issues around this topic for such issues as improving the descriptions and accessibility to data (Le Novère et al. 2005; Waltemath et al. 2011; Mack et al. 2015), but typically those guidelines present light requirements that improve accessibility to the research but do not help in order to replicate results, not being a clear requirement to publication in a journal. From the beginning of the publication process, this situation becomes a waste of time for the editorial journal,

which cannot replicate and confirm what an *in silico* analysis should facilitate. This situation is as well very harmful for the scientific community as it diminishes its credibility and reliability leading to profound and terrible consequences at both personal and organizational level. Altogether reproducibility concerns are becoming an increasing concern to the scientific community as it is impossible for researchers to validate the results, and even worse, not even reviewers are expected to go through code or have experts to verify its validity and its related statements.

In this situation, even though reproducibility is a well-known issue, it does not receive enough attention as to make it a first priority, which makes the whole scientific development potentially and inevitably untrusting from the very beginning, as there is no need to be committed towards reproducibility. Reproducibility not being a top priority will set its root already in the design and setup of an analysis, and then will become more difficult over time to try to overcome and reshape computation after an inefficient setup has been developed. One of the reasons for reproducibility not being top priority is that scientists are often recognized for the results they obtain, but it is not so well recognized the effort of designing analysis and protocols reproducible in computational biology, those are way less recognized than the development of precise tools or methods. In terms of scientific productivity and output, the consequence coming out of these issues around irreproducibility is the duplicity of efforts to solve the same issue over and over again by different researchers. The situation ends up arising multiple publications of similar analysis with more or less the same applications and methods, because of the inability of sharing properly the knowledge previously and not being able of reusing and reaching to some level of agreement as a community on how to define standardized procedures. Although analysis are carried out following a very specific procedure, basically written and coded in a computer, there are some limitations in how to describe precisely the computations within an article if only described in text as used to be done with experiments in the literature (Gil et al. 2007; Bourne 2010). That becomes a problem downstream of scientific publications when in order to understand and replicate published results one has to start a journey of discovery and even a new project (Garijo et al. 2013) in order to achieve the same results whenever it is even possible to do it. But that can be a long process of trial and error with a lot of effort spent to elucidate how the descriptions published can be really transcribed and reproduced, before even being able to try with different scenarios and datasets. The main approach has been to describe and enumerate the steps and elements such as parameters, tools and configurations that have been carried out to generate the results, but without providing an explicit protocol of the procedure to validate the results themselves. This strategy has obvious drawbacks, among them is the completeness and quality of descriptions given to set up the analysis, apart from leading to a loss of time and effort to put it into practice and validate, leading to a duplication of code prone to

errors and lacking details, which does not guarantee that afterward the effort it will end up working correctly and with the same results. The reason for the failure in this approach to explain the procedure is the lack of enough resolution to deliver all the specific information with details that computational analysis need in order to be replicated (Hoffman 2016; Piccolo and Frampton 2016). A very simple and clear example of this situations is when differences in the computational platform and systems behind data and the code used to generate results are obliterated and not explicitly highlighted, leading to possible architectural and environmental differences, which will cause another researcher to not replicate the same results because its own system has differences not highlighted as a requirement in the original publication which will lead to differences in numerical calculations and therefore in results.

Another strategy used to allow other researchers to obtain the same results and try the same analysis upon publication is to make available the procedure used to generate the results only integrating it into a web service or online web server that will allow others to replicate results. As it might sound a step ahead towards a better solution to achieve reproducibility, this option can lead to obscurity, as there is often a lack of open source distribution when no source is available and only the web service is able to replicate results, which in fact are a black box (Morin et al. 2012; Boettiger 2014). These options lead to the inability of the scientific community to review and validate results, specially for journal reviewers who are not able to validate themselves in hand the reproducibility of what it is being published which in turn do not promote transparency in the publication of distribution of scientific knowledge. It also do not provide any kind of control under what is inside the black box, as there is no possibility if anything changes over time and why.

1.4.2. Barriers

We can already find out before the appearance of new tools and technologies to tackle reproducibility issues, some hints and strategies being discussed on how to provide a better organization and structured way to prepare data, code and results in order to get one step closer to reproducibility. In order to explore the collection of tools available that try to guarantee reproducible research and computation in an efficient manner, one has to be able to identify key elements and actors involved in the problem. Here are some of them described and used to evaluate the needs that need to be fulfilled to compare different options available.

Infrastructure: As sometimes bioinformatics workflows require of different types of computations (memory, CPU, I/O, network) is important to be able to change the type

of computational resources on demand, but being able to support this kind of infrastructures for the analysis it also requires a deep knowledge and expertise on it. HPC-based systems or cloud resources are very useful when needed to have infrastructure at point time with high availability and scaling easily for huge loads of computation. When having to deploy computation in parallel and communication between parallel executions, the network can also become a bottleneck because of similar reasons as before, and without enough network throughput, either as a an interface for communication between computation or with the storage system, it can delay and hurt performance. Most of NGS analyses are prone to generate a lot of I/O operations, creating and copying lots of files that can affect the performance of the computation specially if being done at large scale in a shared file system.

Deploy: Trying to generate results deploying computation in a different environment where it was initially developed usually is a big challenge. Different infrastructures have different capacity and resources but are also managed and structured in a different way, it can be challenging to have to develop a piece of software that takes into account all possible environments and its differences within, to allow you deploy the computations without having to modify the code. In the worst case scenario an already designed analysis won't be able to run and it will need to be rewritten from scratch. That problem used to be one of the reasons or excuses to avoid having to share and make reproducible computation, specially when there was no simple way to define and translate workflows to be run in different environments.

Variability: Another challenge is to be able to deploy computation with all software made available for different environments with different resources. The high number of bioinformatics tools used within NGS workflows can be overwhelming to setup and configure in many systems, especially if the users do not have all permissions to self-manage those systems as in shared clusters or HPCs. An approach that works remarkably well is the use of virtualization software, which provides a layer of abstraction from the hosting system being able to have above it another system setup up and tailored for the workflow. Lately other types of virtualization software have been developed and becoming widely adopted, this so called lightweight virtualization like Docker also allows to containerize more specific applications with a smaller impact in the system and network allowing to set up multiple virtualized applications for different steps of the workflow, which in turn can take care of the extra layer of choosing parameters and environment management to deploy all the computation in different environments, with different resources and configuration at very high level without further need to integrate this layer of information inside the workflow itself.

Learning: Although there are many different solutions to those problems, such as guidelines and workflow management software, the newer solutions, which have recently appeared and are geared towards resolving more of this general issues while adding features to enhance and make easier the integration with big data and large computational analysis, usually have the disadvantage of offering another learning step towards a different language or technology for which biologists are not used to. The adoption of such tool will take time and will still remain uncertain as for which tool to select whenever there is not enough information on which is the best tool for every scenario and type of analysis. Also without any standard and new tools coming out, it is difficult to see which one is going to mature enough and not end up dying, while having enough information and adoption on the field.

Publication: But before the annotation and indications become accessible, sometimes not even the code used to generate results becomes available. There are no policies, or they are not clear enough, regarding the fact of making available the specific pieces used to obtain the results in the main scientific journals such as Nature, Science, PNAS. Typically those journals ask for the availability of the software that ensures the results published, but that is clearly not enough, as there are ways of making available black boxes which does not provide any way for science to review and validate the inner specifics of the methods used when they are not transparent and are opaque. That situation opens the possibility to make changes within the black box, which will lead to differences in the behavior between before and after the publication process, or even in the future. Those changes cannot be tracked or evaluated transparently and do not let anybody know about any the level of consistency or any situation happening behind.

Flexibility and plug-and-play: In computational biology the way to design and implement a workflow is usually quite tangled and relies on a trial and error strategy where for one step of the computation, multiple tools might be available to be used, although with some differences in the results. Usually a biologist will have to develop and try to plug in and out multiple tools at different stages of the development in order to evaluate which one gives the better result. Another need related to flexibility, is the ability of using different parameters or configurations for the same step that might change the behavior and the outcome of that piece of computation. It has to be easily changeable without requiring too much effort in the development of the pipeline, even if the entire results might end up changing.

Checkpoints and testing: Important features related to reproducibility are testing for integrity and consistency all the way down the pipeline. In order to be able to test for integrity, as an aid during the development, tests can be design in order to define what are the outcomes of a pipeline in any given environment, just a default execution

predefined that should not give any troubles if the pipeline is properly executed, that allows to track integrity during development and modifications with light executions. Another point is consistency checks, because the workflow needs to be able to detect when there is any inconsistency in what is expected to get out of it as a result. This means that within the workflow some rules and safeguards need to be checked from time to time in order to validate the output of the computation not only completely after the completion of the workflow but also partially, with certain level of granularity. In this way it becomes trivial to detect errors that can arise at any part of the workflow implementation, which can be attributed to such actions as changing configuration parameters or the dataset used along the pipeline, which might not be fitting or in a wrong format for one way or another (data-config-tool). Such features are extremely helpful when trying new configurations or ways of deploying computation even in different environments. It then becomes very useful to have the feature of error-detection together with a caching feature as so to be able to resume computation from where it was left as soon as the errors are fixed in order to avoid recomputation and duplication of efforts if there is no change affecting it.

Documentation: Another big trouble, that goes against the philosophy in scientific research, is the lack of proper or even documentation at all of the code used. As in a typical experiment, it is of critical importance in research to be able to keep track and record all the way down a protocol, understand what is every single step and all the elements involved and how they are involved in the analysis. As it also is in computational biology, and all the computation should also be considered an experiment with the same importance as a wet-lab experiment. All the steps need to be understood, as well as the data and configurations used. The fact that code is available is not enough. Without proper description and documentation it can become very tedious to realize how the workflow goes if one is not familiar with the code. Therefore, is of extremely importance to be able to fill in proper descriptions of all the elements involved, such as which configurations are used by default and if not, which one is used and why; the tools and proper version used, and the aim and procedure and the expected output of the results. The lack of annotation in the code written and the procedure used to generate results halts other researchers than the authors from reproducing or reaching to the same conclusions, often giving an advantage position and closing in the range of action over a topic. Even if there is any annotation given, if the quality and specificity it is not enough, the publication of that code and that analysis, can only lead to the replication of the analysis using the very same dataset published,, without being able to understand the analysis and, therefore, use it further. It is also important to also be able to record and keep track of all the intermediate steps in computations, both for data, such as intermediate files, but also the code with variable parameters that have been used. In this way it becomes much easier to debug and backtrack when facing any issue, and it is

also easier to take snapshots, which can be later on shared and deployed in other environments to continue computations.

Accessibility and collaboration: As many large efforts on computational biology, teams are organized, usually across institutions, to in a collaboration among them put an effort in data generation, collection and analysis. In this scenario it is necessary to be able to organize during the effort a system that will ensure the accessibility to a repository where all raw data can be obtained and shared openly. It is also mandatory to be able to record and share all the steps of the workflow during the development of the analysis in a shared repository which will allow easy collaboration and integration of all the committed work of all the members involved. The importance of keeping track of every single change: when, who made it and the repercussions and being able to see differences and understand them is also important; usually this feature is included in any version control system which could also be integrated into the framework. The setup of these features before submission for publication makes the process of evaluation, replication and review straightforward and can ensure the quality and reproducibility of all the work.

All these issues need to be address in order to be able to tackle the big bio data problem (Marx 2013), which together with large-scale can allow and lead to big computation. When this is achieved infrastructure and resources can be better exploited and more performances will lead to faster results (which can be a 20 fold reduction in time), in order to avoid big computation bottlenecks and underused resources. For that purpose many solutions are being explored that will allow to use data and deploy computational in highly parallel systems.

1.5. Approaches to reproducibility

Reproducibility is an interesting topic that has drawn attention not only in biosciences, but also in computer science (Gent 2013; Sussman). It has lead to the development of interesting tools to aid in making computation reproducible, without depending on the hardware where it is running. Here are reviewed the applications currently used and proposed as the best solutions to face reproducibility issues.

1.5.1. Cloud

Cloud computing can be of great help to mitigate the issues of reproducibility in science (Fusaro et al. 2011; Stein 2010; Kasson 2013; Baker 2010; Balazinska et al.). Issues

specially related to resources where using the cloud can make easier to set them up. Reproducibility can be further improved and made easier upon publication by the authors when putting together a virtualization software and a cloud environment compatible between them. With this combo, reproducibility is not strongly attached to the physical resources used originally where computation took place, instead it offers the possibility to be distributed and deployed in any machine that supports virtualization software. This one can provide the same snapshot, an exact replica of the setup that was used on the system to generate the results, and can be deployed in any other environment available at any time in the cloud to replicate the results. That solution allows forgetting about issues related to tool, package and library versions; but also on the configuration parameters used to set up the environment. The virtualized environment can carry all the elements needed software-wise and data-dependent as well ensuring the consistency. Now, the only concern is whether there is enough infrastructure to virtualize and simulate the infrastructure once used to deploy the computation in the scenario of a large analysis and big bio-data. For this purpose cloud computing offers the perfect solution as it allows users to have on demand all the resources needed seted up and managed by a provider without having to maintain them while they are not being used. Regarding to the main challenges of this solutions, sometimes it has to be decided to commit to a specific solution, not all cloud and virtualization providers are compatible, although in the last years there is pretty much standardization and it is not so problematic to work with major technologies and providers such as VMware, Virtualbox, Docker and Amazon EC2, Google Cloud, Open Stack. The cost is the main concern under this scenario, where large computation requiring huge resources can be too expensive, although prices are lowering down during the last years, and possibilities such as using instances with reduced price whenever they are underused such as in Amazon Spot Instances are becoming easier to use, specially for non critical work, or even for critical work using tools relying in machine learning to predict how to better deploy your computation at lower risk with lower price. Another issue could be related to administrative control and cloud security, as depending on the project it might require to fulfill certain requirements in terms of data protection and access.

Cloud environments also provide solutions for another big challenge associated with big bio data, as on many research fields the need to store and make available all the data to share. But in the case of biodata the main difference is that there is a wide range of data types and formats used either to store the raw data or processed files and results, and the ability to generate and connect all data in order to be useful and be able to use all of its relevant knowledge that can be extracted from it becomes an essential and critical piece. For this use, several attempts are being made to have connected cloud environments to relevant data repositories and biobanks with the tools necessary to access it, close to

them, so it makes easier to use all of the resources for analysis. This approach broadens the access to both data and the downstream computational part and can help with data geo-availability.

The tendency on large consortia and institutional driven research groups is to build cloud based infrastructures, sometimes closely related to big technological companies such as Google or Amazon (DNAnexus). That interesting paradigm changes the way scientific research was receiving funding, as there is no more need to dedicate budget to spend on acquisition and maintenance of hardware, but rather on consumption of it. Another issue due to the stage of transition from typical commodity based or HPC dedicated hardware for computation into a cloud-based environment is the replication of datasets and not a full-reliance over cloud storage. Even though data can be stored, backed up and maintained in the cloud, most institutions still rely on local copies rather than on an updated distributed repository of all of their data and code. This creates the problem and the need to be able to access, move and deploy easily both the data and the code used to compute. This situation is still not fully developed and the direction is steering towards data availability from day 0 (generation) in connected clouds where researches will be able to have deployed both data and code, as otherwise it will become a bottleneck due the increase of data generation and transfer therefore.

With the need to easily deploy data in the cloud and access to computational resources large enough to handle big data, the software dependency started to crack. As for the data there was a need for accessibility and connectivity; for the code, there is a need for easiness of deployment into any infrastructure that it might be supported in the cloud for computation. This means that for these tools to be used, a biologist needs to be able to deploy them correctly and it has to be able to scale up easily with large computational resources, without the need to adapt or develop new solutions only suitable for new systems/clouds. Cloud based environments can be used in order to facilitate the computational requirements. In a cloud environment it becomes easier to distribute large computations in a unified and scalable environment where accessibility/durability and performance can be ensured whenever is needed without having to engage into the implementation and management of an entire cluster or HPC local system.

Another problematic thing related to code deployment is that the availability of cloud to analyze large amounts of data allows defining and establishing more heavy and powerful ways to obtain results out of the data. By combining all sources of big data available nowadays, large computations can be deployed and can be intersected to cross out more results. But that only leads to another layer of complexity, where not only the data and the layers must be easily accessible, but it also needs to be easily connected in a way that computation can be largely streamlined. But the many to many relationship

of data and codes results into easily troubles and crashes, which also calls for a solution in a model that allows computations to be easily tracked and saved without trashing everything and start over.

1.5.2. Virtualization

The very first technological approach to face reproducibility has been already put in practice with some researchers facilitating openly the access to snapshots of their environments used to replicate the data (Angiuoli et al. 2011; Nocq et al. 2013; Dahlö et al. 2015). These snapshots contain should contain all the code and data necessary to only achieve the same results. The solution through virtualization has received major attention as it offers the possibility to share an entire package that combines system, code and data. Some of the virtualization technologies are quite popular and widespread and can be deployed in HPC and the cloud computing environments easily, as these technologies are commonly well supported in any modern computer and system. Even though this technology solves some of the issues related to reproducibility, there are still other obstacles beyond this point not directly related to the technology and therefore left unsolved. Even with the ability to mimic and deploy computation if the code and data are not openly distributed without restriction, as only presented as binaries for example, there is no possibility to fully acknowledge reproducibility, just the replication of some results. It might also not be possible to deploy the same virtualization technology in all the systems for computation. The setup of the environment still needs to be configured if the computation is too large to run in commodity hardware.

1.5.2.1. Docker

Virtualization technologies can be further used and integrated inside bioinformatics analysis in different ways than typically used to do. When properly integrated these technologies can avoid the drawbacks of typical virtualized environments. In order to avoid the extreme growth on the number of virtual environments needed to execute and finalize a workflow, if the virtualization technology can be separated and integrated, the workflow manager could take care in each part of the computation to select what needs to be virtualized and with what. Such an approach allows further compartmentalizing and a better reuse of pieces of workflows, allowing for a more flexible plug and play approach. The best solution that became highly available as a technology and has become widely adopted is Docker. It is an open source software that offers an alternative to the typical virtualization hypervisor-based software. The main difference between these two strategies for virtualization is that the later offers an extra layer of

hardware virtualization, not allowing the virtual operating system to access directly the underlying hardware. This difference between the technologies allows a much more lightweight version of virtualization software, as it requires less from the container-based approach software like Docker to run on the host system. It uses a well known technology; LXC containers available in Linux systems for a long time, which allow simple abstraction and virtualization over a shared operating system kernel. The difference in the approach is also reflected in performance (Containers), as hypervisors are typically a heavier solution for virtualization while container based even though being lighter had some issues regarding security and system compatibility in non-Linux operating systems.

Other advantages of container technology from Docker are related to some of the features provided by the tools that precisely help to attack some of the issues related to reproducibility. This technology offers a better solution to document and integrate the information needed and related to the environment and software that is going to be used within the container to virtualize execution. It includes features to annotate the version of the base operating system used, all the commands used to install, and to set up the specific software, environment and additional checks to look for consistency during the build process. The beauty of it is the format; all those instructions are contained in a single text file. The simple format facilitates the recording and versioning of the containers, which can be on demand committed to a repository where can also be shared and accessible to any researcher. The content of the containers is up to the researches, as it can contain all the specified software, but it is not only restricted to the code. A container can include instructions to fetch datasets and databases, or setup connection to services or file systems where it can access all the data. Whenever it is needed all can be packaged into a heavier object, a snapshot, containing all the code and data after the build process and the container is completely built. Due to the recipe based configuration and organization of the build process in a text file, one can commit to repositories and have ready as many containers as needed which wasn't a feasible option with typical virtualization tools, as the number of images grow the size and, therefore, data transfer would increase becoming impractical to deploy, fast and easily, all the virtual instances. All these possibilities have lead to this piece of technology to start being widely adopted and to think of ways on how to organize biological applications and workflows in a folder of shareable and publishable resources useful for the communities, with an increasing recognition (Moreews et al. 2015; Sallou and Monjeaud ; Boettiger 2014; Belmann et al. 2015).

1.5.3. Workflow management software

Nowadays in bioinformatics there is a term used to define the procedure designed and used to carry on analysis on data. It is called a workflow, and it mainly describes a group of steps that involve tasks such as data gathering and processing or transformation and analysis of data generated from it. The group of steps involved in a workflow can include several bioinformatics tools and languages that allow the field to integrate several sources of information and different methods in order to produce results that allow us to backup our hypothesis. In the era of high throughput sequencing, bioinformatics tend to be made of pipelines or workflows to process the sequencing data. These workflows involve different data files and or databases, which are connected through a series of steps that typically involve long computations, which will end up passing through data transformations and filtering steps. One workflow starting from the genome sequence of a species can end up encompassing tens of different steps before generating the final results (Steinbiss et al. 2016).

Before sequence analysis became so complex with the advent of NGS, bioinformatics already had some pipelines typically developed in scripting languages that made a fast learning curve in order to work in a similar way as used to do with a terminal. Such languages provide basic features such as variables, compartmentalization and access to system tools and file system quite easily. Some others grew and became better and gave the possibility to use further functions more specialized through the use of libraries like Perl or Python. Solutions in order to automate computation in bioinformatics have greatly advanced, getting away from just scripting to process data and having all the steps joined into a script that simply execute everything serially and handle the minimal requirements for it to complete. That could be done either in a more primitive language such as BASH or using Make (Stallman et al. 2004) to construct recipe like scripts, rule based instructions to set the dependencies to run the analysis. Make is the most widely known build automation tools, it became one of the first attempts to automate the execution of more complex analysis, although having being developed for building other tools, it provided a simple and very specific syntax with one main purpose: follow rules to execute commands and generate results by taking care of the dependencies between commands and resolving them automatically. It also has some other features, such as not being limited to a specific language and it is also able to keep track of which elements have been already processed in order to avoid re-computation. These features are part of a list of requirements in order to generate more complex workflows, but it also carried some limitations on how complex the rules and dependencies can be defined. The downsides of these options and the main reason why it started to decay in

popularity is because the code becomes unusable if the environment changes or if it updates due to the entanglement of the scripts and the underlying infrastructure where the computation is run. The growth in complexity and loss of readability becomes more and more apparent if the workflow grows substantially in both data dependencies and parallel computation, which is also not so straightforward to manage across the whole workflow when that happens. However, Makefiles and simple scripting in some cases it is still proven enough, as the time to set up a small workflow with experience in such a way does not pose a big effort and is usually well handled.

In order to overcome such limitations a new set of tools categorized as workflow managers have been specifically design to tackle the issue of the management of complex computational pipelines (Leipzig 2016). They are becoming an essential piece to develop pipelines and complex analysis, as evidenced from the publication records. Nowadays new scientific workflow tools provide new ways of designing and preparing large computations allowing to focus on the main steps that are carried to process data and run analysis without having to put so much effort on other tasks that although are very important, can be abstracted. That level of abstraction means that the manager will take care of deploying and adapting the runtime into different systems without the need to hardcode it in the workflow, leaving a pipeline with less dependencies that can halt from replicating results in different environments. Although the challenge is to remove all these blocking elements for reproducibility, different workflow managers do it in different ways and using different strategies, which means that it can also become impractical to port easily one workflow into another manager, having to probably rewrite it from scratch. But these should not be a major trouble, as for programming languages has been the same issue, and as long as the workflow manager of election fits the needs of the analysis and is broad enough to be run across different platforms it would be fit enough for its purpose. Especially important is to consider the use of certain features like virtualization or cloud support. In the case of not owning an infrastructure that allows running the workflow, it can be seamlessly deployed by running in a virtual environment and in the cloud, without having to care about ownership and maintenance.

The suitability and features of a workflow manager is an interesting topic to discuss, as it affects how well a workflow manager can fit for the analysis. Depending on the level of abstraction that a workflow manager offers it will be critical to decide whether it is more or less suited for one case or another. In the case of HPC and clusters although equipment are generally running in similar systems, the management of parallel computation can differ greatly. In one case it can be that one cluster is running on a GridEngine or any other batch queue system to distribute and execute computation, while in another system computation could be managed by Hadoop. In order to be able

to run a workflow in different systems there is a need for high level abstraction that will allow to take and run computation without any change just by selecting which is the type of the underlying parallel system that is going to be used.

Probably the most important step affecting the selection and usage of a framework is the main purpose of the analysis. Depending on the features of the computation, and its application results, it might be better to select one tool or another. One has to be able to identify what are the limitations, the resources, and the tasks that need to be done in order to design the workflow. Let's say that the analysis is a block of a serial and monolithic applications, then, it won't make much sense to try to design a pipeline that has a split design with low granularity, that is meant to take better advantage on distributed computations, instead a more simple and faster application is a better approach. Whereas if the application is related to a web interface it will be better to select a tool that is better adapted to access databases and generate results easily plug into a website or web server.

Currently there are many options to choose from when looking into workflow management tools and frameworks. Tools can be very different and classified with different features both because of the features but also because of the design patterns they have followed to solve the issues of reproducibility. Depending on the situation one might suit better than the others for specific application, although there are others that hold a more general audience. There are other options more oriented towards more specific applications, like the case of RAP an RNAseq analysis pipeline (D'Antonio et al. 2015), Kepler (Wang and Altintas 2012), Chipster (Kallio et al. 2011) and even now some are available from biotech companies such as the BaseSpace from Illumina, DNAnexus, SevenBridges.

Previously mentioned workflow managers and other like Taverna(Wolstencroft et al. 2013), Pegasus (Deelman et al. 2005) or Galaxy (Goecks et al. 2010) are based on graphical interfaces. Most of them are web interfaces of online services, which are well suited to run the analysis with an interactive session for biologists who do not have the technical skills to code and prepare their own workflows from scratch. Instead, in the bioinformatics field, researchers have common skills well suited to prepare and adapt custom workflows without the need to interact with any interface, typically running into scripting or domain specific languages of preference in order to manage all the computation. For this reason, some tools have been developed in order to assists in design and management of workflows in a lower level. These workflow managers have been developed in order to aid bioinformaticians with a number of certain elements involved in workflow development, which are common and can be automated and abstracted from the process. The main purpose of those are to let the bioinformatician

focus in the development of all the steps for the analysis to obtain the results, without having to take care of other more technical and lower level issues such as handling file system, or orchestrating the execution in parallel environments. Which in turn can be a major advantage, as typically researchers have to develop their workflows, that have to run in different infrastructures like personal computer, workstation, cluster or the cloud depending on the environment they are sitting, which sometimes can be multiple. With the later parallelization, it becomes a critical operation when needed to use large computational resources to deploy and run huge computations. With the abstraction layers that workflow management provide, it becomes easier to shifts the balance of time and effort spent in more technicalities towards focusing on the analysis and results on which scientific research is built upon.

Workflow managers have been around for some time already, although all of the frameworks aim for reproducibility; many have different approaches and features that makes the difference. There are some workflow managers worth to mention which have been widely adopted in many institutions because of the ease of use towards non-bioinformatician users who still want to analyze NGS data but do not know how to prepare their own workflows from scratch and code them. Such frameworks like Taverna (Wolstencroft et al. 2013; Missier et al. 2010; Sroka et al. 2010) were developed in order to allow the setup and distribution of web services, which can be deployed in machines and large infrastructures to process NGS data. It can be done through an interface that allows the users to either select a pre existing workflow from a large repository or design one using its interface that contains a large amount of tools to be used within. That service also provides other third party, the web services, from which it can interact and retrieve updated data from databases. A similar approach is the one used by Galaxy (Goecks et al. 2010), a widely adopted workflow manager web-based which also allows to run, share and design workflows to process NGS data. It provides a web application interface that can be either used through the public servers or installed on dedicated and private resources. Galaxy holds a broad range of applications available from scratch (Blankenberg et al. 2014) by having a repository of tools compatible, both as predefined workflows but also as core elements that can be used to design your own workflow step by step with all the parameters and configuration easily adapted by the user. Through this interface all the options are available to manage the data, tools and workflows used to run in the infrastructure, it also allows to manage accounts and the computation within it. It currently also has some support to be able to deploy in clouds (Afgan et al. 2012), like Amazon EC2. The results obtained by Galaxy has been cited and published in several publications (Goecks et al. 2015; Davidson et al. 2016). It is able to track the metadata related to all the processes of the workflow and allows the users to annotate and document their own executions. It has also great capabilities to reuse components and pieces of data and results into other steps and other

pipelines and great capabilities to allow users to import and export histories, and share results and workflows directly.

Other workflow managers have been developed more oriented towards bioinformaticians. One of the first managers developed for this purpose is Ruffus. Ruffus (Goodstadt 2010) is a Python library to build computational pipelines offering a series of markup helpers and functions using a syntax similar to Python, and was meant to ease the development taking into account tracking of files, parallelization and running in different cluster grid engines as well as visualization of the pipeline graph. Bpipe (Sadedin et al. 2012) came after as a next step in evolution towards reproducibility while it was trying to solve the same problem in a similar way, it was based in Groovy Java's scripting language, but relying in a more flexible approach where command lines could be introduced directly and use environment variables. It was more targeted towards the general audience, used to build pipelines around bash scripting. It also offered a collection of helping functions in order to do split/gather operations over data between the steps of the workflow. Those features provide a much easier way to implement in a workflow from previous experiences, using a much more modular, re-usable, not so cumbersome and less complex code. It also allows to keep track of the execution, and resume it after an error by caching previously completed steps. A different approach from the previous one, was developed in SnakeMake (Köster and Rahmann 2012), which was following the trace of widely, used Make. But this time it was developed in Python and offers several extensions to make it more flexible than common Makefiles. Most of the features included were already in previously commented software like automatic parallelization, dependency detection, checkpoints and resuming. But it provided a different type of designing workflows declaring input, output, configuration parameters and the commands as rules that were put together using dependencies with a very similar approach to Makefiles. Another workflow manager was developed by Spotify written in Python and has been mainly used in web-streaming data processing and therefore was adapted for scientific computation, called SciLuigi (Lampa et al. 2015). It added compatibility with typical cluster engines and added some capabilities to log and track down the all the execution steps. Luigi offers also a web interface to follow the graph and the computation on real time to debug. Other frameworks similar in Python have been developed with a more cumbersome definition of the dependencies in Leaf (Napolitano et al. 2013), where the definition of the steps is detached from the dependency on graph definition, which is defined separately using characters in a string. That definition is afterwards processed by the library to generate the execution graph and connect it to all the defined steps. More recently another framework was developed again in Python, COSMOS (Gafni et al. 2014; Souilmi et al. 2015), this system offered new capabilities like the ability of tracking in real time the computation, showing statistics and storing all the information

related to execution in a database which could also be accessed through a web server. It also added support to starCluter and GlusterFS file system in order to be able to run in the cloud as Amazon AWS. Although it has to be noted that the configuration and setup of the environment requires manual intervention and it is not directly handled by the framework, but through auxiliary scripts. A novel approach was taken compared to the previous tools by BigDdataScript (Cingolani et al. 2015). Instead of relying on another language to build their manager upon like typically Python or Java, it has its own syntax, including its own parser and debugger that uses a very simple and reduced grammar to allow for conditional and control structures, and also the inherent workflow grammar to define the pieces of itself. Although it initially tries to abstract many of the elements already discussed here, there are some other features, which are not detected by default as with the case of parallelization that has to be specified, but overall supports different environments and cloud as well.

There has been also further development in workflows management tools for very focused and specific applications such as Queue (https://software.broadinstitute.org/gatk/download/queue) framework from GATK (McKenna et al. 2010; DePristo et al. 2011; Van der Auwera et al. 2002), which is only focused towards its variant calling framework. Since variant calling can have multiple steps and usually processes large amounts of data when dealing with cohorts of samples, Queue was developed to better design workflows that could handle jobs in parallel together with GATK. Other similar cases of so-called frameworks have risen to take care of other typical NGS applications like in the case of bcbio-gen (https://github.com/chapmanb/bcbio-nextgen). Another development in the field worth mentioning is the Common Workflow Language (CWL) **(Peter et al. 2016)**. The purpose of CWL is to become a standard and common format for scientific data workflows. It is made up of a contribution from multiple organizations. Its language defines all the elements required to build up in a workflow being able to define with syntax similar to YAML, with specifications of inputs/outputs, runtime environment, metadata and executables. The main drawback of this model is that to become a standard, it needs to be widely adopted by developers of workflow managers, and yet not so many have done it. If it was adopted one of the best features would be the possibility of having large repositories of workflows that could be easily imported into any workflow manager. Recently CWL has added support for Docker containers within the workflows designs. Just a few tools from the most well known collection of workflow managers, like Galaxy or Taverna, have already implemented some level of support for CWL within them, as well as bcbio-gen.

2. Reproducibility

Di Tommaso P, Palumbo E, Chatzou M, Prieto P, Heuer ML, Notredame C. The impact of Docker containers on the performance of genomic pipelines. PeerJ. 2015 Sep 24;3:e1273. DOI: 10.7717/peerj.1273

3. LncRNA evolution

3.1. A comparative encyclopedia of DNA elements in the mouse genome

Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, et al. A comparative encyclopedia of DNA elements in the mouse genome. Nature. 2014 Nov 20;515(7527):355–64. DOI: 10.1038/nature13992

3.2. Enhanced transcriptome maps from multiple mouse tissues reveal evolutionary constraint in gene expression

Pervouchine DD, Djebali S, Breschi A, Davis CA, Barja PP, Dobin A, et al. Enhanced transcriptome maps from multiple mouse tissues reveal evolutionary constraint in gene expression. Nat Commun. 2015 Dec 13;6(1):5903. DOI: 10.1038/ncomms6903

3.3. Third Report on Chicken Genes and Chromosomes

Schmid M, Smith J, Burt DW, Aken BL, Antin PB, Archibald AL, et al. Third Report on Chicken Genes and Chromosomes 2015. Cytogenet Genome Res. 2015 Jul 14;145(2):78–179. DOI: 10.1159/000430927

3.4. Genome and transcriptome analysis of the Mesoamerican common bean and the role of gene duplications in establishing tissue and temporal specialization of genes

Vlasova A, Capella-Gutiérrez S, Rendón-Anaya M, Hernández-Oñate M, Minoche AE, Erb I, et al. Genome and transcriptome analysis of the Mesoamerican common bean and the role of gene duplications in establishing tissue and temporal specialization of genes. Genome Biol. 2016 Feb 25;17(1):32. DOI: 10.1186/s13059-016-0883-6

4. Leishmania genome adaptation and evolution

Asexual maintenance of genetic diversity in the protozoan pathogen ***Leishmania donovani***

P. Prieto Barja[1,7,$], P. Pescher[2,$], G. Bussotti[3], F. Dumetz[4], H. Imamura[4], D. Kedra[1], V. Chaumeau[5], H. Himmelbauer[1,7], P. Bastien[5], Y. Sterkers[5], J.C. Dujardin[4], C. Notredame[1,6,*], and G. F. Späth[2,*]

[1]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology**,** Barcelona, Spain; [2]Institut Pasteur, INSERM U1201, Unité de Parasitologie moléculaire et Signalisation, Paris, France; [3]HUB de Bioinformatique et Biostatistiques, Centre de Bioinformatique, Biostatistique et Biologie Intégrative (C3BI), Institut Pasteur, Paris, France; [4]Institute of Tropical Medicine, Molecular Parasitology Unit, Antwerpen, Belgium; [5]University Montpellier 1, Faculty of Medicine, Laboratory of Parasitology–Mycology, CNRS 5290, IRD 224, University Montpellier 1&2 (UMR "MiVEGEC") and Centre Hospitalier Universitaire, Montpellier, France; [6]Universitat Pompeu Fabra (UPF), Barcelona, Spain; [7]Institute of Biotechnology Muthgasse 18 1190 Vienna

Running title: *L. donovani* haplotype selection

[*] **Corresponding authors:**

G. F. Späth, Tel: 01.40.61.38.58; E-mail: gerald.spaeth@pasteur.fr

C. Notredame, Tel: +34 933160271; Email: cedric.notredame@crg.eu

[$]these authors contributed equally to this work

Keywords: *Leishmania donovani*, aneuploidy, genome instability, dosage compensation, fitness, haplotypes

ABSTRACT

Leishmania donovani causes visceral leishmaniasis, a fatal disease when left untreated. The process through which the parasite adapts to environmental change remains largely unknown. Here we show that aneuploidy is an integral part of parasites adaptation and that karyotypic fluctuations allow for selection of beneficial haplotypes, with important impact on parasites phenotype, including transcriptomic output, proliferation and infectivity. To avoid loss of diversity resulting from karyotype and haplotype selection, *L. donovani* takes advantage of two mechanisms: (i) polyclonal selection of beneficial haplotypes resulting in co-existing subpopulations that preserve the original diversity, and (ii) generation of new diversity as a result of higher mutation rates tolerated by aneuploidy-prone chromosomes. Our results uncover high aneuploidy turnover and

haplotype selection as a new mechanism of *L. donovani* evolutionary adaptation that preserves genetic diversity under strong selection. This process may be of broad significance to other human diseases, including fungal infection and cancer.

INTRODUCTION

Rapid pathogen adaptation to novel environments is a major threat to human health. In parasites evolving under constant host selection, fitness gains often come along with increased pathogenicity[1]. Aneuploidy has recently been reported to be an important driver for evolutionary adaptation in fungal and protist pathogens. Variations in chromosome copy number have been shown to cause phenotypic variation resulting from changes in both transcriptomic and protein output. Aneuploidy has recently emerged as an important driver in evolutionary adaptation of fungal and protist pathogens, with chromosome copy number variation inducing phenotypic change through transcriptomic and protein expression modulation[2]. Beyond its effect on gene count, aneuploidy can also impact cellular phenotypes through the selection of beneficial alleles. While this phenomenon has received considerable attention in human genetics[3] and cancer genome analysis[4,5], the role of aneuploidy and allelic selection in genome evolution and adaptation of pathogenic eukaryotes remains to be elucidated.

We have addressed this question in the protozoan parasite *Leishmania donovani*, an important human pathogen that causes fatal visceral leishmaniasis [6]. During its life cycle *Leishmania* undergoes a major developmental transition from insect-stage promastigotes to mammalian-stage amastigotes, which adapts these parasites for extra- and intracellular survival, respectively [7]. In addition to environmentally induced stage differentiation, *Leishmania* can adapt to a variety of unpredictable fluctuations inside its human host, notably pharmacological intervention. Such environment-genotype interactions likely select parasites for higher fitness and have important consequences on disease outcome as demonstrated by the emergence of drug resistant clinical isolates [8]. The underlying genetic mechanisms that drive short-term *Leishmania* evolution remain, however, largely unknown. In the absence of classical transcriptional gene regulation, these early-branching eukaryotes often control protein abundance via gene and chromosome amplification [9,10]. *Leishmania* thus represents an ideal non-conventional system to investigate how karyotype variations and haplotype selection drive parasite fitness gains and how transient and stable aneuploidies impact the parasite's long-term evolutionary trajectory.

RESULTS AND DISCUSSION

Karyotype and haplotype selection in field isolates

We first explored the impact of aneuploidy on genetic diversity in *Leishmania* by taking advantage of the recently published sequencing data of 204 *L. donovani* field isolates [11]. These populations originate from independent evolutionary radiations. They have taken place after a bottleneck caused by DDT vector control campaigns in the 1960s and thus constitute a convenient benchmark to study genetic diversity dynamics. Read-depth sequencing analyses were carried out to explore ploidy variations across culture adapted isolates (Fig. 1A). Within most isolates full chromosome amplification (as opposed to local episomal amplifications) was evidenced by predominantly uniform variations of read-depth across individual chromosomes (Supplementary Fig. 1). These same analyses also suggest a dominance of population-wide - rather than mosaic - aneuploidies within each isolates. The most frequent variations are consistent with previous reports [12,13] and involve trisomies (chr 5, 8, 9, 11-16, 22, 23) and tetrasomies (chr 8, 23 and 31). Co-occurrence analysis shows a higher prevalence for some aneuploidy combinations such as chr 5 and 26 or Chr 19, 25 and 34 (Supplementary Fig. 2). Principal component analysis (PCA) based on karyotype profiles (Fig. 1B) does not reveal any strong founding effect but rather a very heterogeneous collection of samples in which the most common karyotypic combination - all chromosomes diploid except 31 - makes up about 10% of the 204 isolates.

The apparent discrepancy between karyotypic diversity and the previously reported genetic structure of that same isolate collection[11] suggests that chromosome copy number fluctuations may be very dynamic with frequent, reversible and independent transitions between polysomic states. Two mutually exclusive models may account for such diversity; the first one - *monoclonal* - involves the rapid expansion of a single individual parasite carrying a beneficial set of driving polysomies, while under the second scenario - *polyclonal* - each culture adapted isolate results from the expansion of mixed subpopulations sharing the same independently generated driving polysomies. We used allele frequency analysis to discriminate between these two hypotheses. Merged profiles were produced by stacking the individual allele frequency distributions of each chromosome from isolates having strictly identical polysomy levels. The presence of well-defined peaks in most chromosome profiles suggests that the alleles are balanced the same way across isolates (Supplementary Fig. 3A). In disomic chromosomes most of these profiles are compatible with those expected for diploid asexual populations in which alleles would diffuse under very weak or non-existing selection. When dominated by alleles at near-neutral equilibrium this process results in unimodal allele frequency distributions centered on 50% as observed for most disomic chromosomes (Fig. 1C - left panel, and Supplementary Fig. 3A - left column).

The situation is more complex in trisomic chromosomes for which profile analysis reveals a mixture of unimodal and bimodal distributions (Fig. 1C - middle panel, Supplementary Fig. 3A - middle column). Within a given isolate, bimodal

distributions reflects homogenous chromosomal duplications across the entire population, i.e. when one chromosome is duplicated, the alleles it carries see their frequency increased to 66% with respect to the non-duplicated chromosome, whose relative allele frequencies decreases from 50% to 33%. In contrast, unimodal distribution reflects a perfect equilibrium between the two possible outcomes of a transition from disomy to trisomy, i.e. equal likelihood of either chromosome within a given pair to be duplicated. Unimodal distributions are perfectly compatible with a polyclonal origin but bimodality is best explained by a monoclonal process, in which a single original duplication would have founded the whole population. The coexistence of these two types of profiles within the same field isolates, e.g. trisomic bimodal chr 5 with trisomic unimodal chr 12 (Supplementary Fig 4A), and trisomic bimodal chr 6 with trisomic unimodal chr 12 (Supplementary Fig 4B), is an apparent paradox that most likely reflects haplotype selection acting as a confounding factor. Indeed, polyclonal origin may also result in a bimodal profile provided one of the two possible trisomic haplotype combinations gets selected over the other. Under this scenario, highly selected haplotypes result in bimodal distributions centered on 33 and 66% (e.g. chr 5), while lower levels of selection result in closer peaks (e.g. chr 6) - up to coalescence at 50% in the absence of any selection (e.g. chr 12) (Fig. 1C - middle panel). Strong selection of the same haplotype across isolates should therefore result allele frequencies being highly conserved across these isolates (i.e. any given allele should have similar frequencies when comparing isolates). This is exactly what we observed for chr 5, whose allele frequencies varies much less across isolates than similar frequencies measured on chr 6 or 12 (Fig. 1C - right panel). In this analysis, the high variation of allele frequencies measured for chr 12 alleles is in perfect agreement with the frequent unimodal allele profile of this chromosome (Supplementary Fig 4A and B) that implies the possible coexistence of alternative trisomic haplotypes across isolates resulting from limited haplotype selection. While chr 5 and 12 represent the two extremes of the selection spectrum, the intermediate profiles of chr 6 (Fig. 1C, middle and left panels, Supplementary Figure 4B) further confirms that haplotypes have the capacity to diffuse under selective pressures of various intensities.

In vivo *karyotype fluctuations*

The polyclonal hypothesis implies the pre-existence of a population-wide karyotypic variability. The maintenance of such a variability would require frequent polysomic fluctuations. We measured this effect by monitoring chromosomal copy number when re-passaging an individual aneuploidic field isolate through hamsters (Methods). The rapid shift towards a predominant disomic karyotype confirms our hypothesis of rapid aneuploidy fluctuations (Fig. 2A). It is unclear, however, if the aneuploidies observed during culture adaptation result from the expansion of existing subpopulations or constitute a reversible de novo phenomenon. This issue cannot be resolved by HiSeq that merely provides an integrative measurement that may not be sensitive enough to

reflect low frequency mosaicism. We therefore addressed this question by applying single cell DNA-FISH analysis to liver and spleen isolated *L. donovani* amastigotes. Our analysis confirmed important polysomic differences *in situ* in liver and spleen with mosaic aneuploidies observed for chr 5, 17, 22, and 27 (Fig. 2B and C). It is this pre-existing diversity that most likely contributes to the emergence of population-wide aneuploidies observed during culture adaptation of the field isolates. This finding is perfectly in line with the polyclonal model that implies the capacity of simultaneously selecting multiple individuals from co-existing subpopulations.

***In vitro* karyotype and haplotype selection**

While the field isolate and the *in vivo* analyses suggest that parasite adaptation relies on a finely tuned mechanism able to cope with frequent aneuploidies, these observations are merely static snapshots of a process that appears to be highly dynamic. In order to elucidate this phenomenon further we turned towards the experimental Sudanese *L. donovani* strain LD1S. In contrast to *L. donovani* clinical isolates from the Indian sub-continent, this strain contains a large number of heterozygous sites that make it an ideal model for polysomy longitudinal monitoring. We used HiSeq to follow hamster-derived LD1S amastigotes during adaptation to *in vitro* culture. Read depth analyses showed that parasites rapidly establish stable trisomies for chr 5, 9, 23, and 26 between *in vitro* passages p2 (20 generations) and p10 (100 generations)(Fig. 3A and Supplementary Fig. 5A). These karyotypic trajectories were highly reproducible across two independent experiments and matched the most common variations observed in the field isolates (Supplementary Fig. 2). All aneuploidies do not appear to be stable and homogenous. For instance, chr 20 underwent a transient trisomy between passages p2 and p20 (190 generations), while mosaic aneuploidies were established for chr 14 and 15 at p10 and maintained thereafter as judged by their intermediate read-depth.

We next measured variations across single individuals by applying HiSeq on 8 sub-clones derived from p20 parasites. This approach provides an alternative to single cell sequencing that is not currently technically feasible in *Leishmania*. Systematic comparison between p20 and the 8 individual clones made it possible to model the original population complexity (Fig. 3B and Supplementary Fig. 5B). Haplotype and karyotype comparisons suggest that the eight clones may have arisen from at least three independent founding individuals. Indeed, while karyotypic variation may be explained by rapid aneuploidy turnover, direct haplotype comparisons (Fig. 4A, Supplementary Fig. 6) clearly show how the nature of chr 5 and 9 trisomies sets clones 1 and 8 apart from the rest. For instance, the systematic differences of dominant alleles on chr 5 between clones 1 and 8 as opposed to the rest of the clones indicates that the two trisomies corresponding to the two groups of clones were established through two duplications of different chr 5 autosomes. While a complex scenario of chained duplications/reversions may account for these differences, the most parsimonious

reconstruction merely requires two trisomies of independent origins and is perfectly compatible with a high aneuploidy turnover. This very same explanation applies to chromosome 9 and further supports the common independent origin of clones 1 and 8. Similar haplotype comparisons also indicated some more genomic heterogeneity in the larger group of clones, especially with respect to chr 15, whose variations are compatible with an independent origin for clone 4 (Fig. 4A, Supplementary Figs.7 and 8). Of course, the sequencing of 8 individuals has limited statistical power, and one should not ignore that our speculation of three original founders relies on a parsimonious hypothesis. Yet, the existence of clearly distinct subpopulation featuring similar aneuploidies brings further supporting evidence to the polyclonal hypothesis, under which individual strains represent complex mixtures of stable subpopulations.

Polyclonality is further supported by clear indications of convergent karyotype and haplotype selection taking place during culture adaptation. At the karyotypic level, the most obvious traces are observed for chr 5, 9 and 26, whose trisomies are shared across all the clones despite their likely polyclonal origin. Selection also appears to occur at the level of haplotype, the most obvious signal being the one associated with chr 26. The haplotype map of this chromosome shows a consistent selection of the same allele combination across clones of different origins (Fig 4A). Unfortunately, the simultaneous haplotype/karyotype selection taking place on this chromosome makes it impossible to discriminate between purifying (lethality of one of the two possible trisomies) or positive (higher fitness of one trisomy) selection pressure being imposed on its haplotype. Chromosome 20 gives, by contrast, a very clear evidence for positive haplotype selection. This chromosome is heterozygous at p2, but becomes trisomic at p10 before reverting back to disomy at p20 (Figs. 3A and 4B, supplementary Fig. 8). After reversion the allele frequency profiles shows, however, a near-perfect homozygote disomy in 6 out of 8 clones (i.e. relatively flat allele frequency profile, Supplementary Fig. 7). It therefore appears that this transient trisomy provides an intermediate step towards the establishment of a homozygote disomy, whose independent selection in at least two subpopulations is consistent with positive fitness contribution.

Impact of aneuploidies on phenotypic variations and long term genetic diversity

The high aneuploidy prevalence coupled with the haplotype selection we observed during parasite culture adaptation points towards a functional role for these genomic variations. We therefore took advantage of the clones' high karyotypic variability to measure the impact of polysomy on transcript output. Since *Leishmania* largely lacks classical gene regulation mechanisms, it has long been expected that polysomy should result in proportional transcriptional variation and that aneuploidy mediated gene expression variations could be a trait under selection. We used HiSeq analysis to compare read-depth variations between genomic and transcriptomic levels (Fig. 5A, Supplementary Fig. 9) and found a very high correlation for most chromosomes

(r=0.72), with the notable exception of chr 31. In this chromosome - mostly tetrasomic across various *Leishmania* strains - transcriptomic output appears to be halved thus canceling the polysomic effect. We also found karyotypic fluctuations resulting from culture adaptation to come along with rapidly increasing *in vitro* fitness as judged by the decreasing generation time (Fig. 5B), and decreased *in vivo* fitness indicated by lower infectivity (Fig. 5C). These observations confirm the physiological consequences of karyotype and haplotypic variation, but they do not make it possible to determine if this process may play a role during the parasite life cycle. We addressed this question by separately sequencing the same isolate obtained from spleen and liver of one individual infected hamster. We found significantly distinct allele profile variations for chr 20, whose allelic diversity is three times lower in spleen as compared to liver (Fig. 5D, Supplementary Fig. 11), while read depth analysis indicates a disomic state in both tissues (Supplementary Fig. 10). These observations recapitulate almost perfectly the results obtained *in vitro* analysing parasite clones and indicate that *in vivo* transient polysomies may play a role in parasite adaptation. The fate of chr 20 clearly shows how the combination of frequent aneuploidies and haplotype selection may lead to a rapid loss of heterozygosity. This process could seriously compromise the parasite adaptation capacity in the longer term and raises the important question of its capacity to avoid an evolutionary dead-end - especially given the absence of sexual reproduction of these parasites in the mammalian host [14]. Maintaining diversity under such circumstances poses a dilemma for all microbial pathogens as it requires a mechanism that would allow some compromise between immediate survival and longer term adaptation. We therefore searched for traces of increased genetic diversity associated with frequent aneuploidies and found a clear signal in the 204 field isolates (Fig. 6). Integrating genetic variation across each individual isolate shows that polysomy prone chromosomes exhibit a significantly higher level of heterozygous sites than their more stable counterparts. This trend is especially strong for chr 31 - the most stable, most frequent and highest order aneuploidy. This important observation indicates that *L. donovani* takes advantage of aneuploidy in order to accumulate mutations and increase its diversity. It confirms that even though transient *in vivo* aneuploidies are difficult to detect and quantify, they are nonetheless frequent enough to leave their mark on the parasite genome thus shaping its genetic diversity.

CONCLUSIONS

Drawing from the sequenced genomes of 204 *L. donovani* field isolates and conducting evolutionary experiments we have uncovered highly dynamic karyotype changes during parasite growth *in vitro* and *in vivo***.** This allows the emergence and selection of new alleles to foster the parasites' genetic diversity during asexual growth within its mammalian host. We have demonstrated that karyotypic variation modulates transcript

abundance and generates considerable phenotypic variability, with fitness gains *in situ* associated with tissue-specific haplotype selection. The genomic landscape defined by an exhaustive combination of all possible trisomies, disomies and monosomies along with their haplotype variations is enormous. Aneuploidy turnover therefore provides the parasite with a genetic potential for adaptation comparable to the one that may be achieved through sexual reproduction.

Rapid aneuploidy turnover combined with haplotype selection allows for fast adaptation, but it comes along with a heavy genetic cost: relative loss of heterozygosity. The transient trisomy of chr 20 and its tissue-specific haplotypic diversity we observed *in vivo* illustrates especially well the parasites' dilemma between over-adaptation to any given environment - that may involve irreversible genetic tradeoffs - and the maintenance of enough genetic diversity for future adaptation. In this chromosome, the loss of heterozygosity seems to contribute to the parasite adaptive capacity across conditions and tissues, but one cannot ignore that this same process may also push the parasite into an evolutionary dead end.

Yet, the very existence of the parasite testifies of the capacity it has evolved to establish a fine-tuned balance between the conflicting requirements of short- and long-term adaptation. The parasite long-term survival relies on two complementary mechanisms for the generation and maintenance of genetic diversity. The first one relates to the polyclonal origin of selected populations. We have shown that culture adapted isolates do not have a single founding parent but result from the simultaneous selection of several individuals. This process results in the maintenance of a high level of genetic diversity. It benefits from the rapid aneuploidy turnover that allows selected polysomies to occur frequently and independently in genetically distinct individuals. The second one is a direct consequence of the relaxed selection reported to occur after gene duplication [15]. Our findings are in perfect agreement with these models and confirm that aneuploidy prone chromosomes have a significantly higher mutation rate than their more stable counterparts, thus making aneuploidy one of the drivers of genetic diversity in *L. donovani*.

Leishmania, like all known infectious agents, provides us with some of the best examples of genetic selection driven survival. But it does so in a very unusual way and while in most parasites survival is mediated by high mutation rates, genetic material exchange through sexual - or analogous - mechanisms, Leishmania appears to have evolved around a different, yet unreported, strategy. Our many findings are all consistent with the notion that Leishmania adaptation relies on genome instability to enhance the parasite evolvability. This strategy bears direct consequences for current protocols in *Leishmania* drug and biomarker discovery. First, *Leishmania* genome instability limits all current and future drugs that directly target the parasite biology and

are therefore bound to select for resistance phenotypes[16,12,8]. New strategies for anti-leishmanial drug discovery are thus needed to avoid direct parasite selection, for example by targeting the parasites' dependence on the host cell metabolism. Second, the massive genomic changes we observed during *Leishmania* culture adaptation call into question current protocols of biomarker discovery, which all rely on *in vitro* expansion of clinical isolates. Alternative, culture-independent approaches need to be established, for example by propagating field isolates in experimental animal models or applying direct tissue sequencing. Regardless of how these new approaches get implemented, it is clear that genome instability needs to be considered when investigating medically relevant aspects of the parasite's phenotype such as tissue tropism, drug susceptibility and pathogenicity. Our findings clearly set the stage for the future discovery of haplotypes with diagnostic and prognostic value.

Material and Methods

L. donovani isolates, culture and axenic amastigote differentiation

L. donovani 204 field isolates clinical samples from the ISC were used in this analyses to track evolutionary diversity maintained across populations [11]. Infectious *L. donovani* strain 1S2D (MHOM/SD/62/1S-CL2D) was obtained from Henry Murray, Weill Cornell Medical College, New York, USA. Axenic *L. donovani* 1S2D, clone LdB, was cultured as described [17,18]. Briefly, axenic promastigotes were grown at 26 °C in M199 media supplemented with 10 % FCS, 25 mM HEPES pH 6.9, 4 mM $NaHCO_3$, 1 mM glutamine, 1 x RPMI 1640 vitamin mix, 0.2 µM folic acid, 100 µM adenine, 7.6 mM hemin, 8 µM biopterin, 50 U/ml of penicillin, and 50 µg/ml of streptomycin. Axenic amastigotes were cultured at 37°C with 5 % CO_2 in RPMI 1640 supplemented with 1 mM glutamine, 1 x RPMI 1640 vitamin mix, 0.2 µM folic acid, 100 µM adenine, 20 % FCS, 1 x RPMI amino acid mix, 50 U/ml of penicillin, and 50 µg/ml of streptomycin and 28 mM MES. Four days after induction of differentiation by pH and temperature shift as previously described [17], axenic amastigotes were centrifuged for 10 min at 4°C and 1500 g. Supernatants were removed, cells washed twice in PBS, and adjusted to 2 x 10^8 parasites per tube for RNA or DNA extraction.

Hamster infection and isolation of infectious amastigotes

Anesthetized hamsters were inoculated by intra-cardiac injection with 5 x 10^7 infectious amastigotes obtained from infected hamster spleens or livers. Hamster weight was monitored and animals were euthanized with CO_2 after four months of infection. Spleens and livers were collected, weighed, and homogenized in PBS supplemented with 2.5 mg/ml saponine using the gentleMACS homogenizer with gentleMACS M tubes from Miltenyi. Parasite burden was determined by limiting dilution as described [19]. For amastigote purification, suspensions were adjusted to 25 ml with PBS and

cleared by centrifugation at 130*g* for 5 min at room temperature (RT). The supernatants were collected, 1 ml of saponine (25 mg/ml) was added under gentle agitation, and parasites were harvested 5 min later by centrifugation at 1800*g* for 10 min at RT. After two washing steps with PBS, remaining host cell contaminants were removed by Percoll centrifugation. Parasite were resuspended in 3 ml of 45 % Percoll, and layered above a cushion of 2 ml of 90 % Percoll in a 15 ml falcon tube. After 30 min of centrifugation at 3500*g*, 15°C, amastigotes were recovered from the interface of the gradient and washed 3 times in medium or PBS (1800*g*, 10min, 15°C). Tissue-derived amastigotes were adjusted to 2 x 10^8 parasites per tube for RNA or DNA extraction or inoculated in culture medium for differentiation into promastigotes and further culture.

Promastigotes derived from splenic amastigotes

1 x 10^7 amastigotes purified from hamster spleens were inoculated in 5 ml of promastigote culture medium for differentiation and culture. Promastigotes were then maintained in culture by dilution in fresh medium once they reached stationary phase. At passages 2, 10 and 20 corresponding to approximately 20, 80 and 190 generations respectively, parasites in exponential growth phase were collected and adjusted to 2 x 10^8 parasites per tube for DNA extraction. After 20 *in vitro* passages, serial dilutions of promastigotes were plated on M199 Agar plates for cloning and 8 clones were selected and amplified in promastigote medium.

DNA FISH analysis

DNA probes for chromosomes 5, 17, 22, and 27 were prepared as previously described [20]. Amastigotes were purified from infected hamster livers and spleens as described above, immobilized on glass slides, fixed in 4% paraformaldehyde and 4% acetic acid. Fluorescence *in situ* hybridization was performed according to Sterkers [20]. *Leishmania* cells were viewed by phase contrast, and fluorescence was visualized using appropriate filters on a ZeissAxioplan 2 microscope with a 100 x objective. Digital images were captured using a Photometrics CoolSnap CDD camera (Roper Scientific) and processed with MetaView (Universal Imaging). Z-Stack image acquisitions (15 planes of 0.25 mm) were systematically performed for each cell analysed using a Piezo controller, allowing to view the nucleus in all planes and to count the total number of labelled chromosomes. The ploidy was estimated on 150 to 300 labelled cells.

Genomic sequencing

All *Leishmania* samples, except the repassage shown on Figure 2a the liver/splenic differential analysis shown on Figure 5c were performed using DNeasy blood and tissue kits from Qiagen according to manufacturer instructions. Acid nucleic concentrations were measured with a NanoDrop® spectrometer. Between 2 to 5 µg were used for sequencing. DNA sequencing was performed using a Illumina Hiseq 2000 platform and

TruSeq v3 kits. 1 µg of nucleic acids were used for DNAseq while the rest of the material was used for quality control. In the passage experiments (sp-ama, P2, P10 and P20) two paired end libraries were constructed with insert sizes of 100 and 160 bp respectively. The libraries used for the clones only included an insert size of 160 bp. All libraries were sequenced in 125 bp reads.

All Sequencing reads have been submitted to the European Nucleotide Archive (ENA) and are available under the Accession Number PRJEB15282.

Transcriptomic sequencing

RNAs were performed using RNeasy mini plus blood and tissue kits from Qiagen according to manufacturer instructions. Acid nucleic concentrations were measured with a NanoDrop® spectrometer. Between 2 to 5 µg were used for sequencing. RNA sequencing was performed using an Illumina Hiseq 2000 platform and TruSeq v3 kits. 4 µg of nucleic acids were used for RNAseq, respectively, and the rest of the material was used for the quality control. RNA libraries were sequenced on a single flow cell single stranded 51 bp read. Sequencing reads have been submitted to the European Nucleotide Archive (ENA) and are available under the submission number PRJEB15282.

Reads mapping and read depth analysis

RNA and DNA sequences determined from all samples were aligned to the same reference genome (*L. donovani* BPK282A1, Accession number: PRJEA61817). Mapping and post processing was carried out using a combination of BWA mem (v0.7.8) [21] along with Samtools (v1.3)[22] in order to refine read information and clean mapped sequences. Alignments were further refined using GATK (v2.8) IndelRealigner and MarkDuplicates[23]. In order to estimate chromosome somy levels for every sample, Samtools was used to estimate read-depth for every base, so as to determine median read-depth. This measure was calibrated to determine somy, under the assumption that the most frequent somy levels correspond to diploidy. These calibrated values were used to select subsets of chromosomes having comparable aneuploidy levels. Uncalibrated values were used to measure chromosome covariation of somy levels across all field samples and to run a Principle Component Analysis (PCA) using the R software.

Relative expression estimation

Gene RNA read counts corresponding to annotated ORFs were estimated using reference gene annotation from TriTrypDBv7. The actual read-counts were measured with BedTools (v2.19)[24] using the BAM files obtained after mapping. Gene copy numbers were estimated using a similar protocol. In order to normalize for gene length, read counts were estimated in Reads Per Kilobase Mapped (RPKM)[25] across all genes

and samples. These RPKM measurements were then used to compare expression levels across samples.

Variant calling and allele frequency analysis

Pileups were generated across the reference genome for all the samples using Samtools mpileup and its multiallelic caller to detect variants in all samples. Variants were filtered using a quality threshold of 15 using a Read Position Bias (RPB) filter of 0.1 and retaining alleles with frequencies comprised between 0.1 and 0.9 included. In order to filter out sites potentially affected by episomal amplifications, regions with a read-depth higher than 1.6 x the median depth were excluded from the analysis. The resulting sites were then used to draw allele frequency statistics, and produce allele frequency profiles. *Allele frequency profiles* were produced by measuring the frequency of every nucleotide at every position and by compiling the resulting numbers were compiled so as to generate a distribution plot. *Allele dispersion plots* were produced by measuring for each nucleotide at each position the standard deviation of frequencies across a set samples, compiling the resulting values and by plotting their distribution.

Funding

Plan Nacional [BFU2011-28575 to C.N., P.P.B., D.K.]; Center for Genomic Regulation (CRG); ; Spanish Ministry of Economy and Competitiveness, 'Centro de Excelencia Severo Ochoa 2013–2017' [SEV-2012–0208]; Center for Genomic Regulation (CRG) (C.N, P.P.B., D.K). This work was supported by a grant from the Institut Pasteur strategic fund to the LeiSHield consortium.

Authors Contribution

P.P.B., P.P, C.N. and G.F.S worked on all aspects of work, contributed to the design of the project and wrote the article, G.B contributed to *in silico* analysis, F.D. performed hamster infection experiment with field isolate, H.I. and J.C.D. contributed to the field isolate analysis, D.K. helped analyzing the sequencing data, H.H. was responsible for the genomic sequencing of the *in vitro* clones, V.C, P.B. and Y.S. contributed the DNA-FISH analysis,

Biography

1. Pallen, M. J. & Wren, B. W. Bacterial pathogenomics. *Nature* **449,** 835–842 (2007).
2. Selmecki, A. M., Dulmage, K., Cowen, L. E., Anderson, J. B. & Berman, J. Acquisition of aneuploidy provides increased fitness during the evolution of antifungal drug resistance. *PLoS Genet.* **5,** e1000705 (2009).
3. Sabeti, P. C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419,** 832–837 (2002).
4. Staaf, J. *et al.* Landscape of somatic allelic imbalances and copy number alterations in human lung carcinoma. *Int. J. Cancer* **132,** 2020–2031 (2013).
5. Scheinfeldt, L. B. & Tishkoff, S. A. Recent human adaptation: genomic approaches, interpretation and insights. *Nat. Rev. Genet.* **14,** 692–702 (2013).
6. Alvar, J. *et al.* Leishmaniasis worldwide and global estimates of its incidence. *PLoS One* **7,** e35671 (2012).
7. Zilberstein, D. & Shapira, M. The role of pH and temperature in the development of Leishmania parasites. *Annu. Rev. Microbiol.* **48,** 449–470 (1994).
8. Leprohon, P., Fernandez-Prada, C., Gazanion, É., Monte-Neto, R. & Ouellette, M. Drug resistance analysis by next generation sequencing in Leishmania. *Int. J. Parasitol. Drugs Drug Resist.* **5,** 26–35 (2015).
9. Dujardin, J.-C., Mannaert, A., Durrant, C. & Cotton, J. A. Mosaic aneuploidy in Leishmania: the perspective of whole genome sequencing. *Trends Parasitol.* **30,** 554–555 (2014).
10. Sterkers, Y., Crobu, L., Lachaud, L., Pagès, M. & Bastien, P. Parasexuality and mosaic aneuploidy in Leishmania: alternative genetics. *Trends Parasitol.* **30,** 429–435 (2014).
11. Imamura, H. *et al.* Evolutionary genomics of epidemic visceral leishmaniasis in the Indian subcontinent. *Elife* **5,** (2016).
12. Downing, T. *et al.* Whole genome sequencing of multiple Leishmania donovani clinical isolates provides insights into population structure and mechanisms of drug resistance. *Genome Res.* **21,** 2143–2156 (2011).
13. Rogers, M. B. *et al.* Chromosome and gene copy number variation allow major structural change between species and strains of Leishmania. *Genome Res.* **21,** 2129–2142 (2011).
14. Rougeron, V., De Meeûs, T., Kako Ouraga, S., Hide, M. & Bañuls, A.-L. 'Everything you always wanted to know about sex (but were afraid to ask)' in Leishmania after two decades of laboratory and field analyses. *PLoS Pathog.* **6,** e1001004 (2010).
15. Lynch, M. & Conery, J. S. The evolutionary fate and consequences of duplicate genes. *Science* **290,** 1151–1155 (2000).
16. Gazanion, É., Fernández-Prada, C., Papadopoulou, B., Leprohon, P. & Ouellette, M. Cos-Seq for high-throughput identification of drug target and resistance mechanisms in the protozoan parasite Leishmania. *Proc. Natl. Acad. Sci. U. S. A.* **113,** E3012–21 (2016).
17. Goyard, S. *et al.* An in vitro system for developmental and genetic studies of Leishmania donovani phosphoglycans. *Mol. Biochem. Parasitol.* **130,** 31–42 (2003).
18. Saar, Y. *et al.* Characterization of developmentally-regulated activities in axenic amastigotes of Leishmania donovani. *Mol. Biochem. Parasitol.* **95,** 9–20 (1998).
19. Pescher, P., Blisnick, T., Bastin, P. & Späth, G. F. Quantitative proteome profiling informs on phenotypic traits that

adapt Leishmania donovani for axenic and intracellular proliferation. *Cell. Microbiol.* **13,** 978–991 (2011).

20. Sterkers, Y., Lachaud, L., Crobu, L., Bastien, P. & Pagès, M. FISH analysis reveals aneuploidy and continual generation of chromosomal mosaicism in Leishmania major. *Cell. Microbiol.* **13,** 274–283 (2011).
21. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
22. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
23. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).
24. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
25. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5,** 621–628 (2008).

Figures

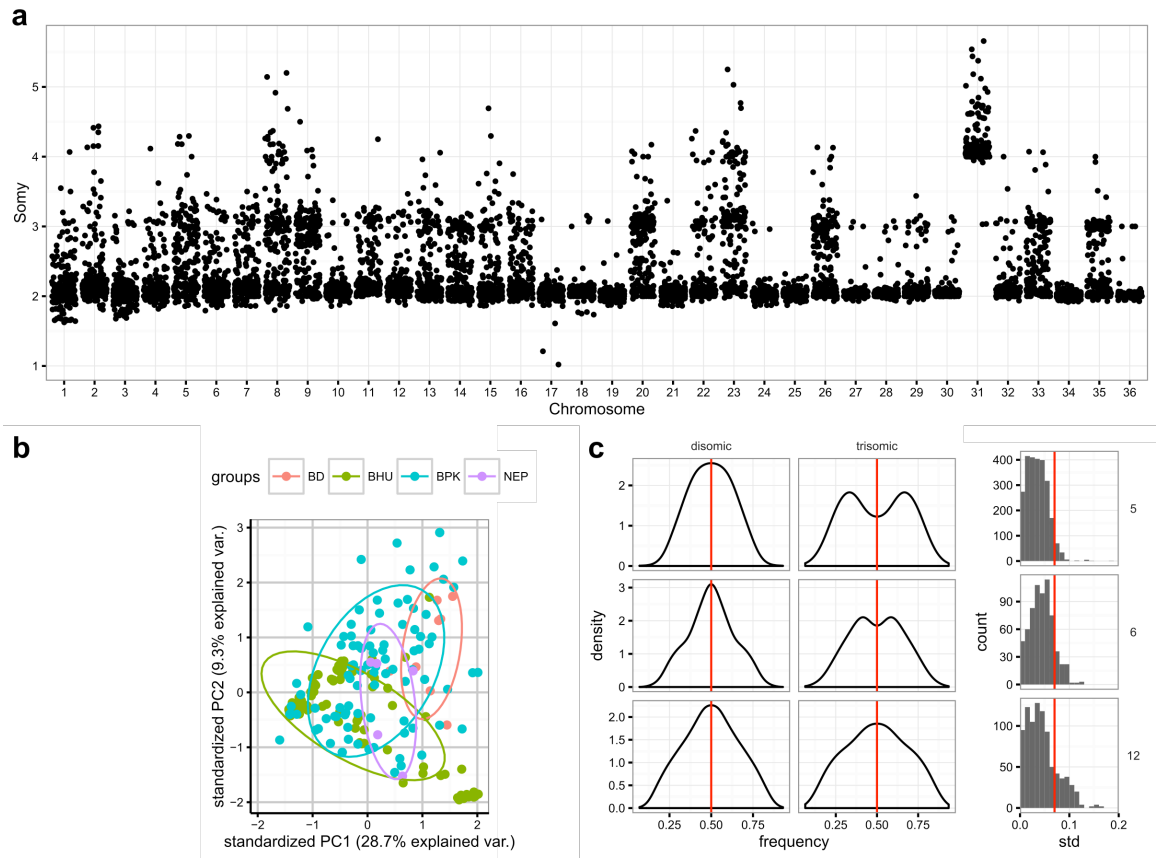


Figure 1: Genome instability and haplotype selection in 204 *L. donovani* field isolates. Polysomy analysis based on read-depth comparison and derived allele frequency analysis on specific aneuploidies is shown. *(A) Somy analysis.* Chromosome copy number was estimated by read-depth analysis and plotted for each chromosome and field isolate. *(B) PCA analysis.* Each isolate somy profile was used to run a PCA analysis. Isolates are colored by genetic proximity. *(C) Allele frequency analysis.* The number of alleles (count) was plotted against their frequency. Distributions of allele frequencies are shown for three selected chromosomes (5, 6 and 12) (see full panel in Supplementary Figure 3) and are displayed for disomic (left panel) or trisomic (middle panel) isolates. The red line corresponds to a frequency of 50% (0.50). The left panel shows allele dispersion plots obtained by measuring distribution of allele frequency standard deviations (Std dev) across the corresponding isolates. Haplotype selection is indicated by a low Std as observed for trisomic chromosome 5. The red line was added as a visual aid to materialize the differences in the number of highly variable alleles across these three chromosomes.

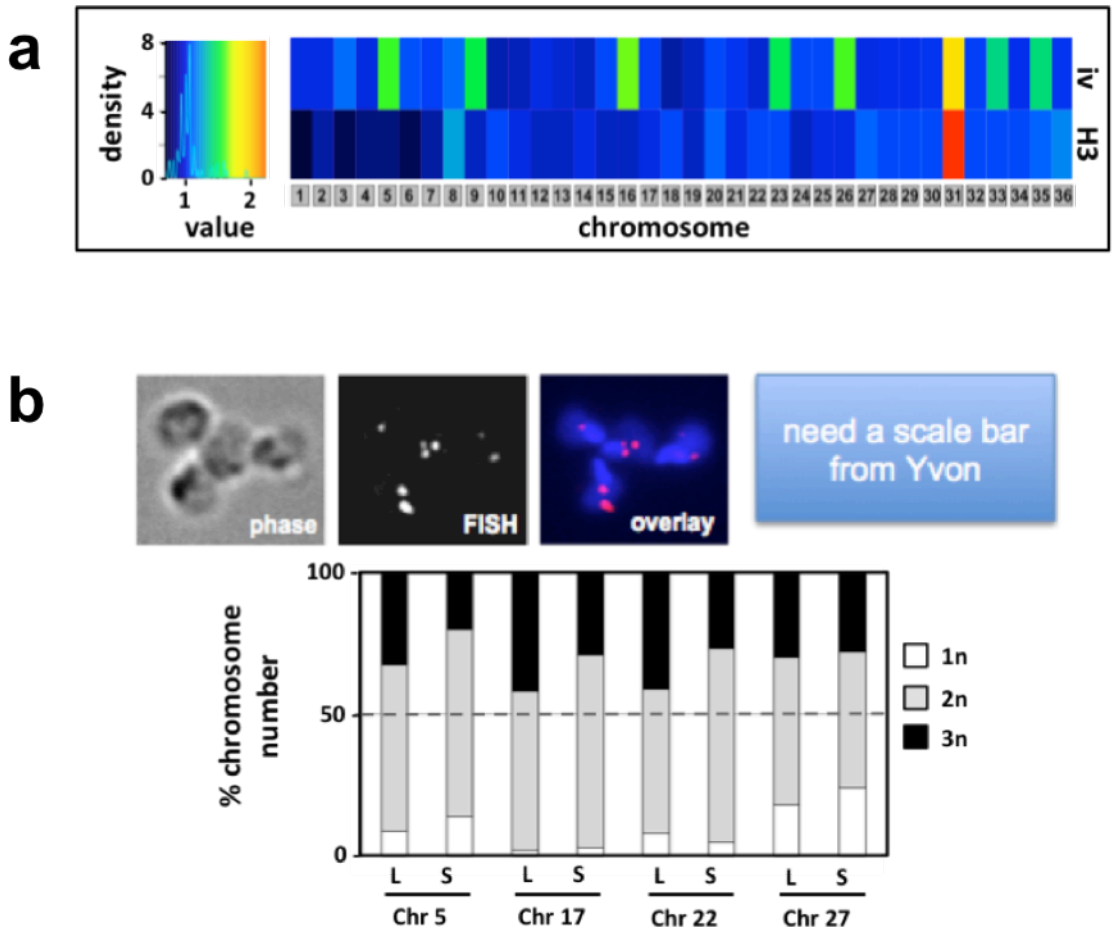


Figure 2: In vivo aneuploidy dynamics. The three panels document the reversible nature and the mosaicism of *L. donovani* aneuploidy *in situ* during hamster infection. *(A) Read depth analysis.* Heat map representing the polysomy level as calculated by read density of the *L. donovani* field isolate XXX after over 10 passages in culture (iv) and following three passages in the hamster (H3). *(B) DNA-FISH analysis. L. donovani* strain LD1S amastigotes purified from infected hamster liver (L) and spleen (S) were analyzed with fluorescent labeled probes specific for chromosomes (Chr) 5, 17, 22, and 27 and signals were analyzed by microscopy. The signal for chr 5 in spleen-derived amastigotes is shown as an example (phase, phase contrast; FISH, fluorescent signal from DNA-FISH analysis; overlay, merged image of DNA-FISH and nuclear signal obtained with DAPI stain). The bar corresponds to XXX. The % chromosome number was calculated counting 100-300 individual cells per condition (lower panel). The level of somy is indicated by the bar filling, with white for monosomy (1n), gray for disomy (2n) and black for trisomy (3n).

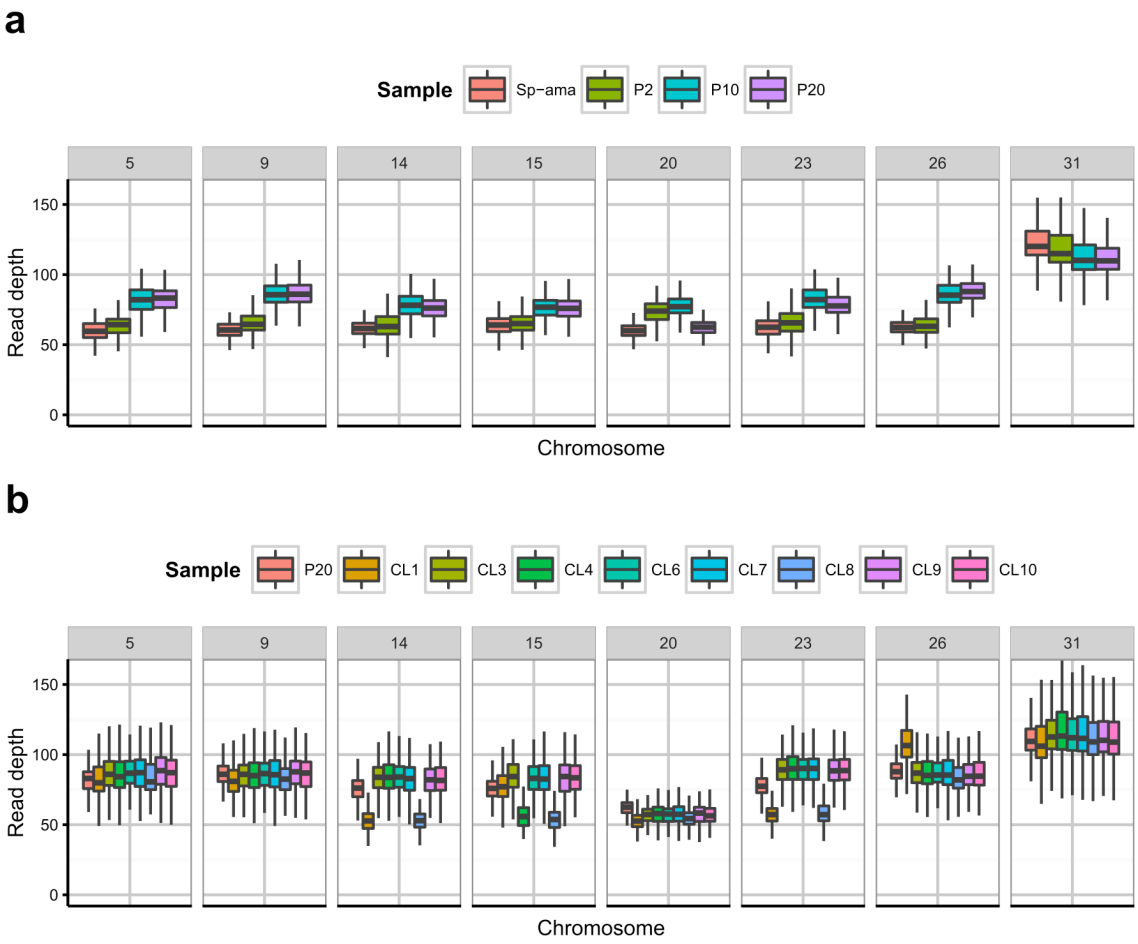
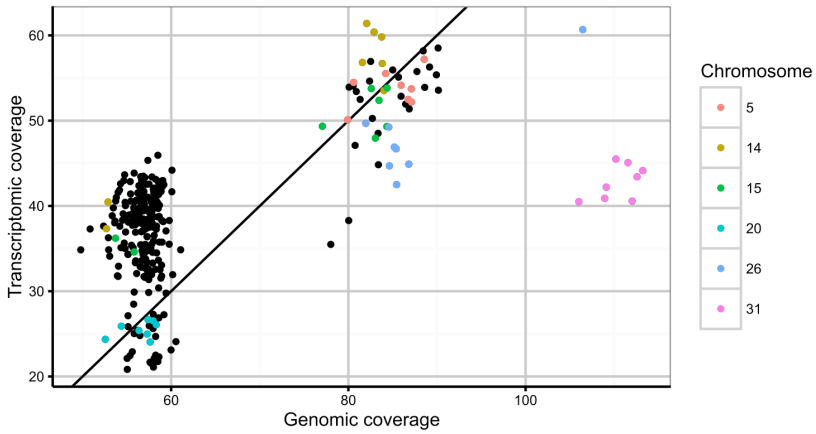


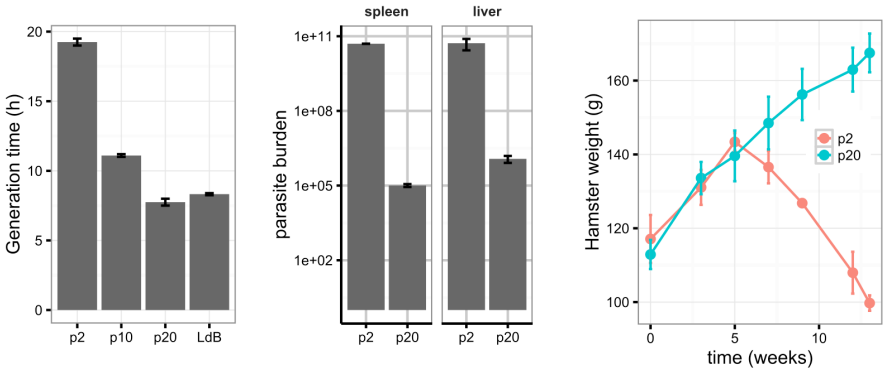
Figure 3. *In vitro* aneuploidy dynamics. Read-depth variations in *L. donovani* LD1S of individual chromosomes at different passages during culture adaptation and in sub-clones derived from the parasite population at passage 20. *(A) Read depth variation during culture adaptation.* Variations in read-depth for *L. donovani* LD1S amastigotes isolated from infected hamster spleen (sp-ama) and derived promastigotes at passage 2, 10, and 20 (p2, p10, p20). Only aneuploidic chromosomes are shown (see full panel in Supplementary Figure 5A). Read-depth is displayed using a standard box-plot representation with the central line materializing the median of the distribution. *(B) Read depth variation on individual sub-clones.* Variations on these same chromosomes as shown in panel A are shown for 8 individual parasite cultures subcloned from p20 parasites (sub-clones are indicated by CL).

Figure 4: Fluctuations of allele frequency during culture adaptation. These panels document haplotype selection by following the allele frequency distribution in the sub-clones and the original p20 parasite population. *(A) Haplotype selection in clones.* Variable sites in chromosomes undergoing amplification were painted according to their dominant alleles to visualize haplotype variability across the 8 sub-clones derived from p20 parasites. Each line represents a chromosome from the original p20 population (first line) or the sub-clones. Allele frequency distributions are shown in the profiles on the right of each panel. The color code corresponds to the read depth, with disomic-to-trisomic transitions indicated by a continuum from orange to green color. *(B) Haplotype selection during culture adaptation.* Allele frequency distributions corresponding to hamster-derived splenic amastigotes (sp-ama) and derived promastigotes during culture adaptation at passaged (p) 2, 10 and 20 is shown for selected chromosomes (for full panel see Supplementary Figure 6).

a


Transcriptomic coverage
Genomic coverage
Chromosome
5
14
15
20
26
31

b


Generation time (h)
p2
p10
p20
LdB
spleen
liver
parasite burden
Hamster weight (g)
time (weeks)
p2
p20

c

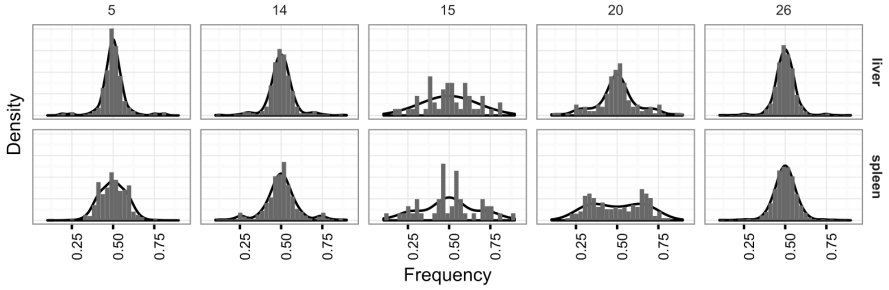

5
14
15
20
26
Density
liver
spleen
Frequency

Figure 5: Aneuploidies and haplotype selection influence phenotype and fitness. This figure shows a quantification of aneuploidy and haplotype selection on transcriptomic output, virulence, growth rate and tissue-specific parasite adaptation. *(A) Correlation of aneuploidy and transcript output.* The median transcriptomic output of each chromosome was plotted against the median read-depth of genomic coverage for each of the eight sub-clones derived. *(B) Parasite phenotype in vitro and in vivo.* LD1S promastigotes during culture adaptation at passages p2, p10, and p20 were assessed for *in vitro* growth to determine the generation time (left panel), *in vivo* growth in infected hamster spleen and liver (middle panel), and pathogenicity by monitoring hamster weight as a function of time (right panel). *(D) Tissue-specific haplotype selection.* Amastigotes were isolated from infected hamster liver and spleen, purified genomic DNA was subjected to HTseq analysis, and allele profiles were established by plotting allele density versus frequency.

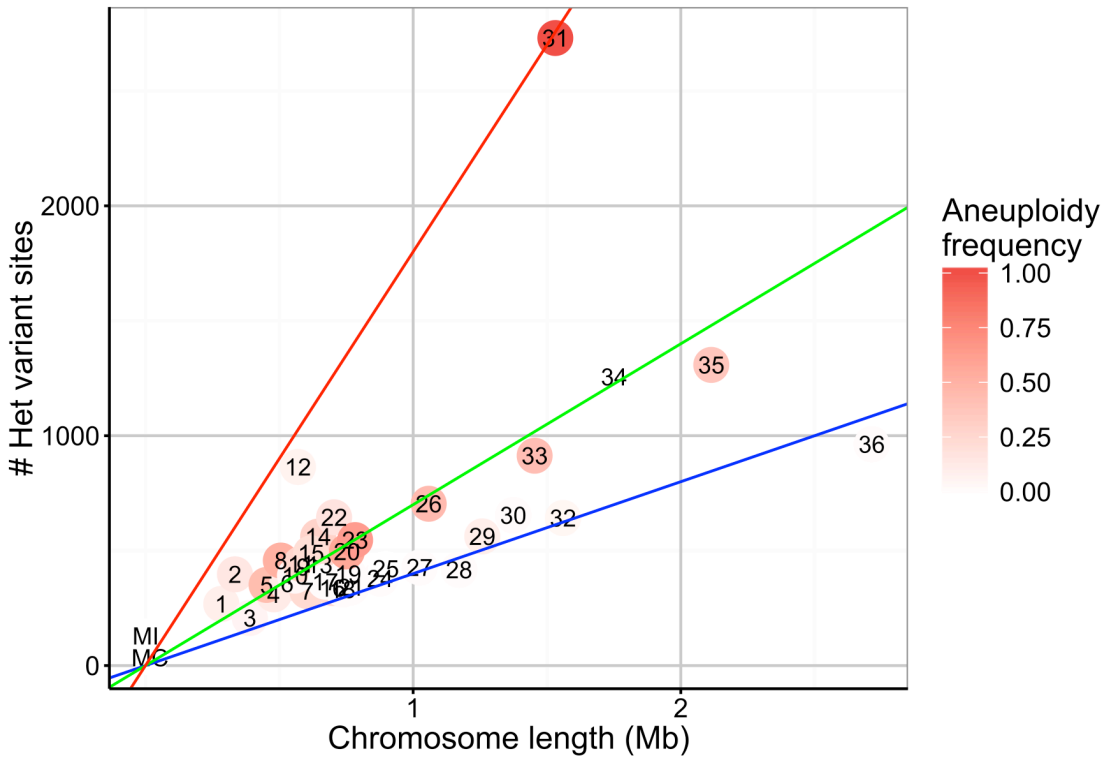
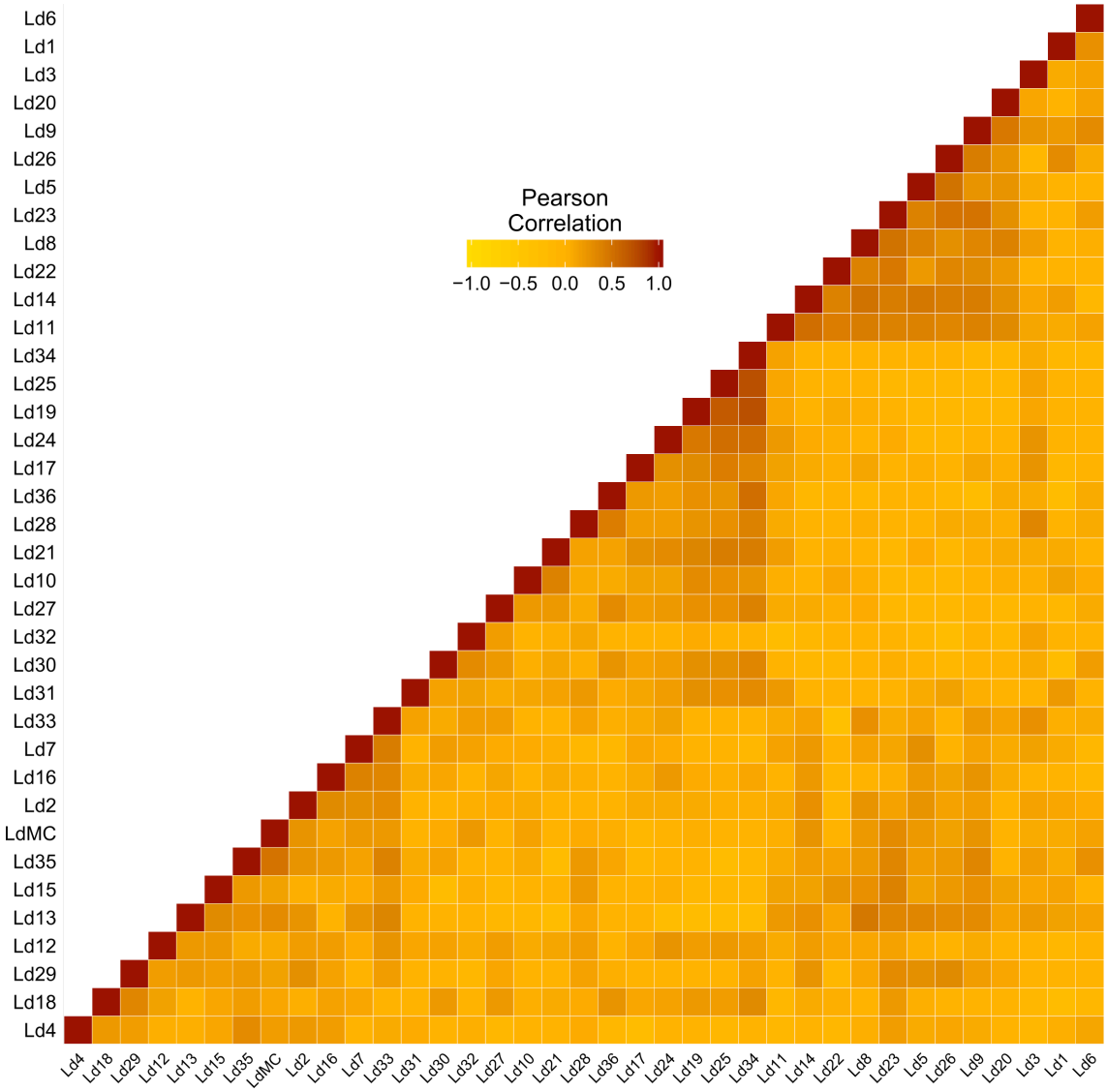


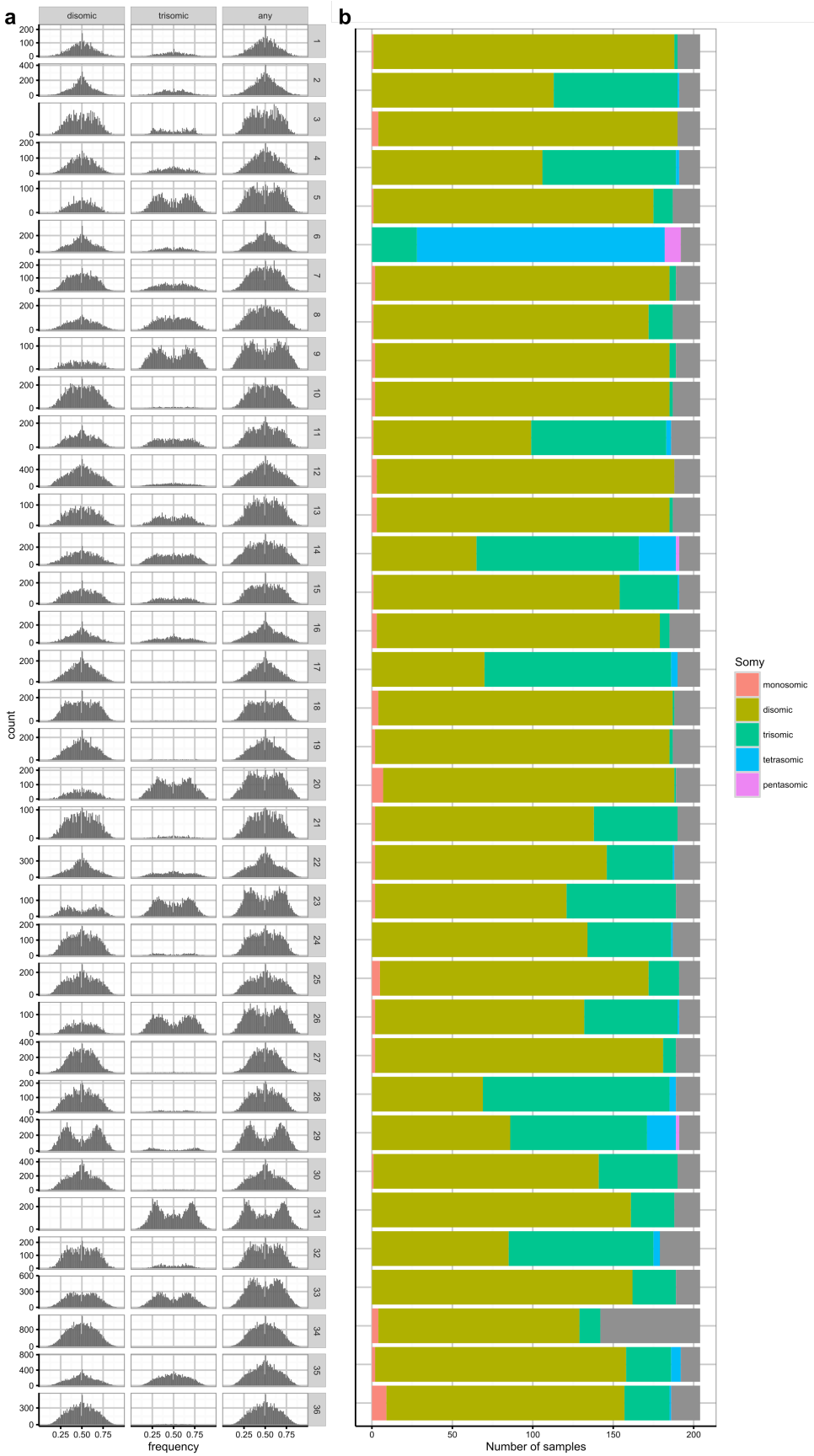
Figure 6: Aneuploidy influence on heterozygosity. Each chromosome is represented by its index on the graph and colored according to aneuploidy frequency after culture adaptation, as measured in the 204 field isolates . The three colored lines materialize three apparent regimen of mutation rates.

Supplementary Information

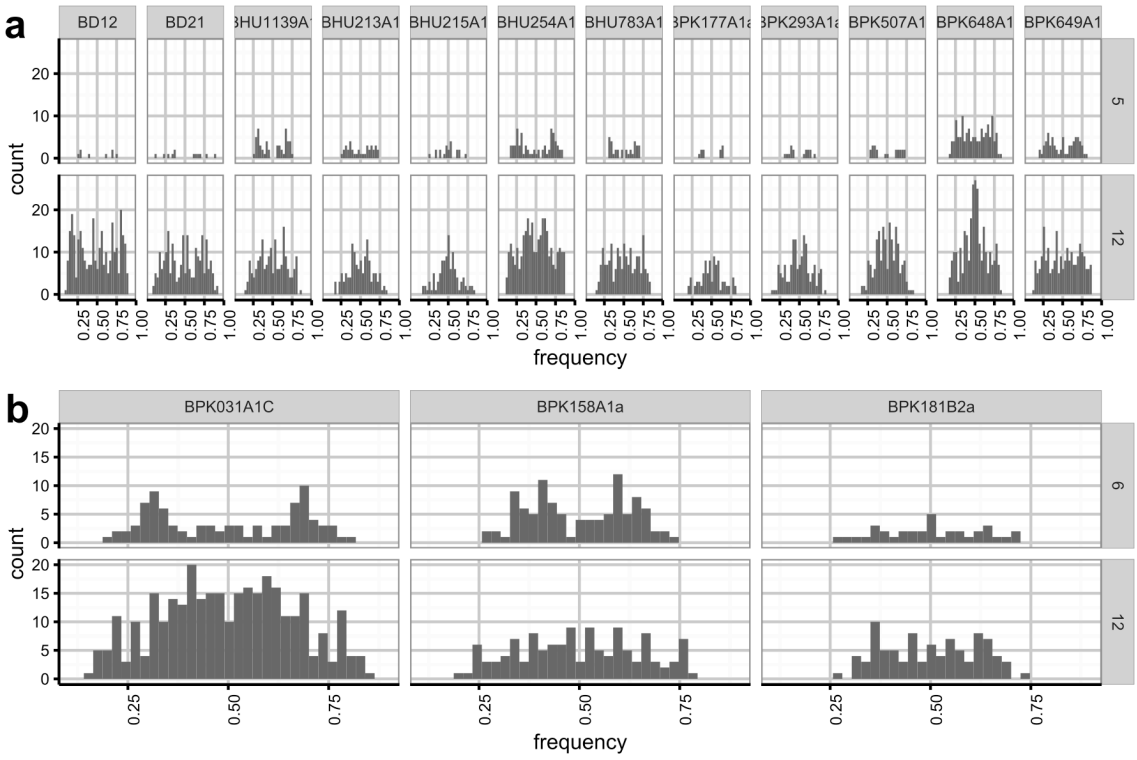
Supplementary Figure 1 Colored rendering of normalized read-depth variations across the 204 field isolates genomes with light colors representing disomic levels and darker color higher somy level up to blue for tetrasomic. Each chromosome is divided in 10 bins and colored according to the median normalized read-depth measured on this bin so as to reflect partial amplification. The relative homogenous coloring suggests the prevalence of chromosome wide amplifications rather than partial amplifications.



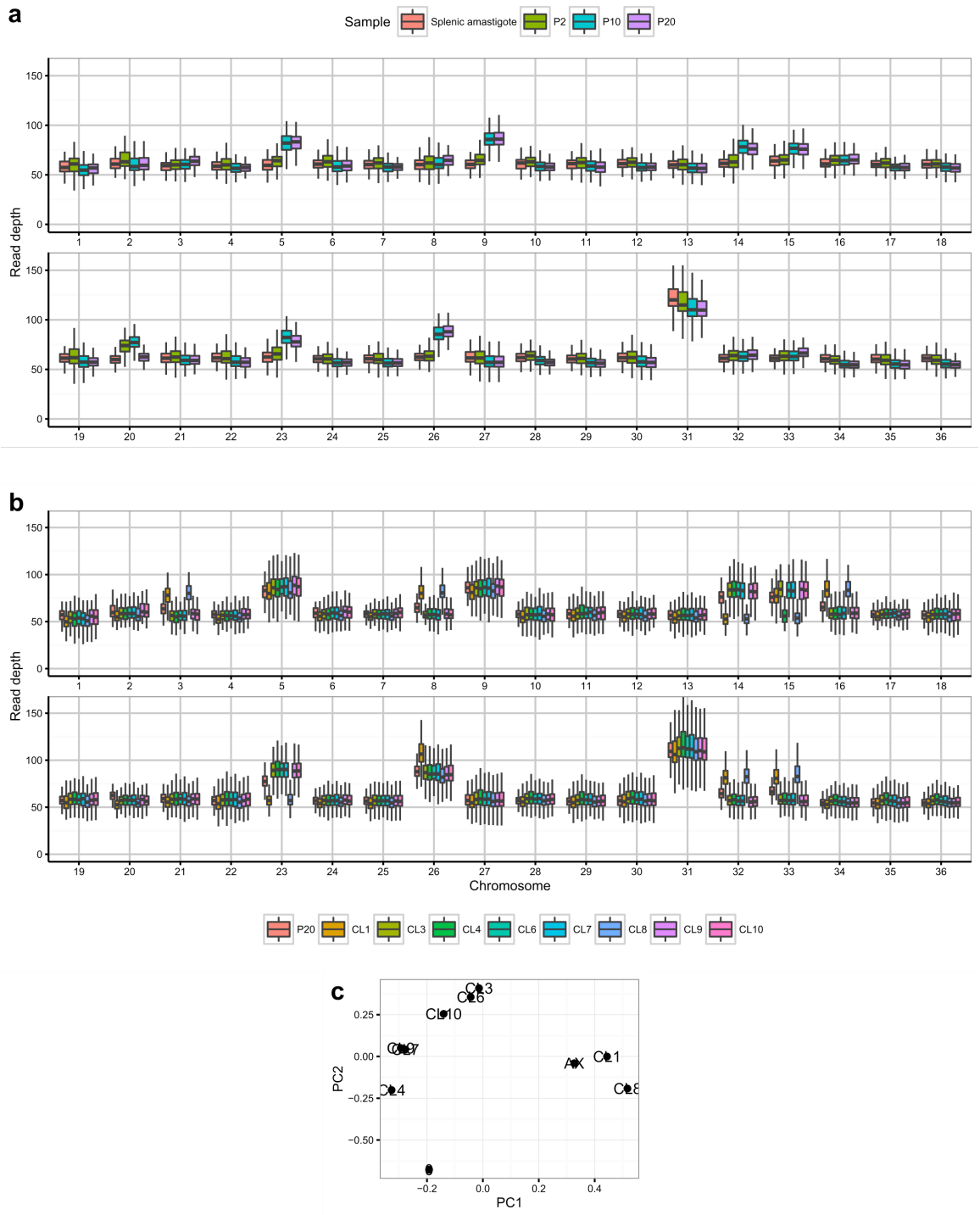
Supplementary Figure 2 Pearson correlation measured for each pair of chromosomes while considering the median read depth measured in every field isolate. Two main clusters appear one containing typically trisomic chromosomes and another minor cluster with chromosomes that remain at disomic level for almost all the samples.


a
disomic
trisomic
any
count
frequency
0.25 0.50 0.75
b
Number of samples
0
50
100
150
200
Somy
monosomic
disomic
trisomic
tetrasomic
pentasomic

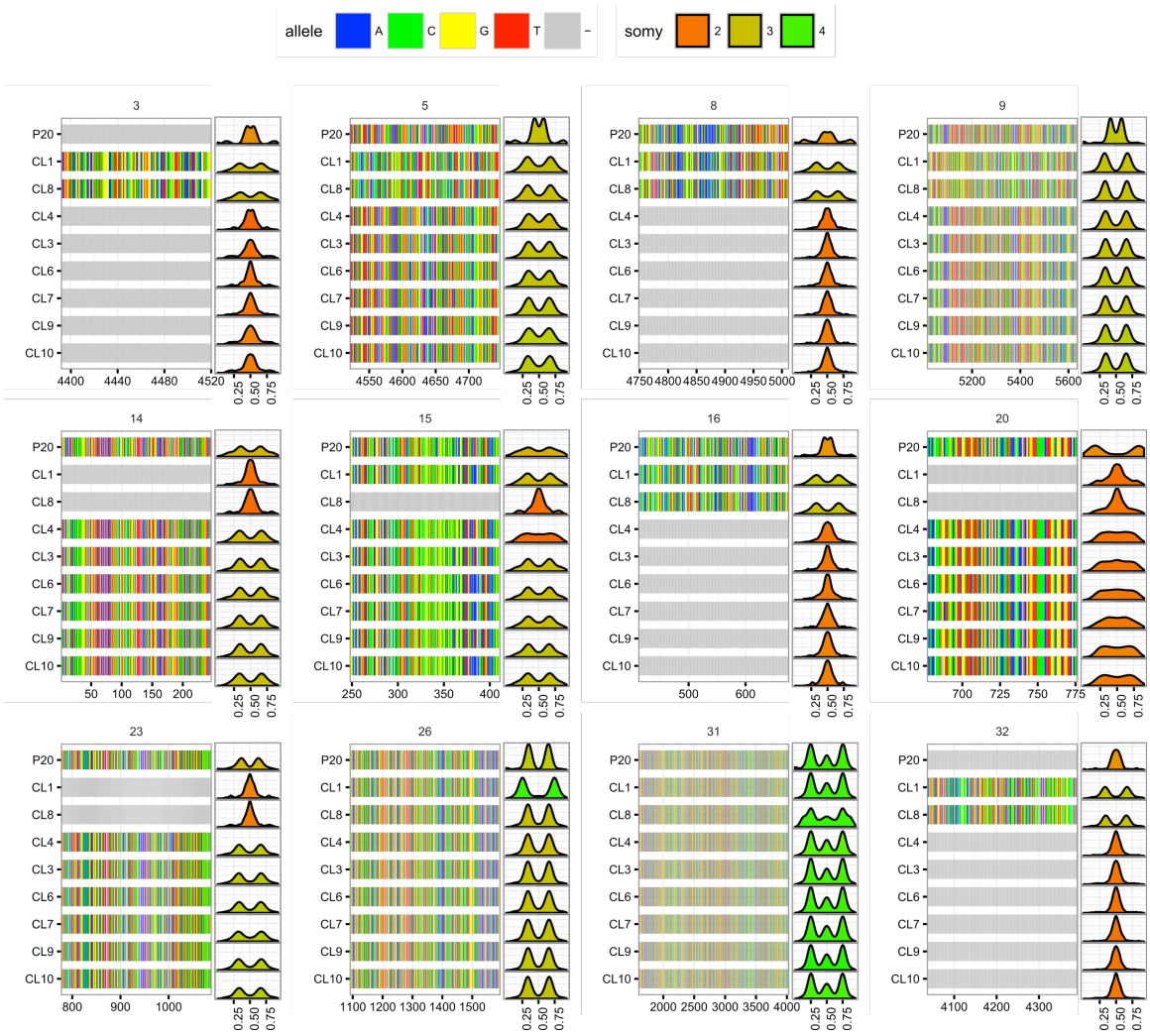
Supplementary Figure 3 Allele frequency combined distribution profiles on strictly disomic, trisomic and tetrasomic chromosomes in the 204 field isolates (A). Frequency plots were obtained by individually estimating a distribution in each chromosome of each isolate and by stacking up the profiles corresponding to chromosomes with similar somy level. This procedure made it possible to give the same contribution to each isolate thus avoiding the averaging effect that would have resulted from piling up the allele profiles. The number of chromosomes corresponding to each category is compiled in the bar graph (B).



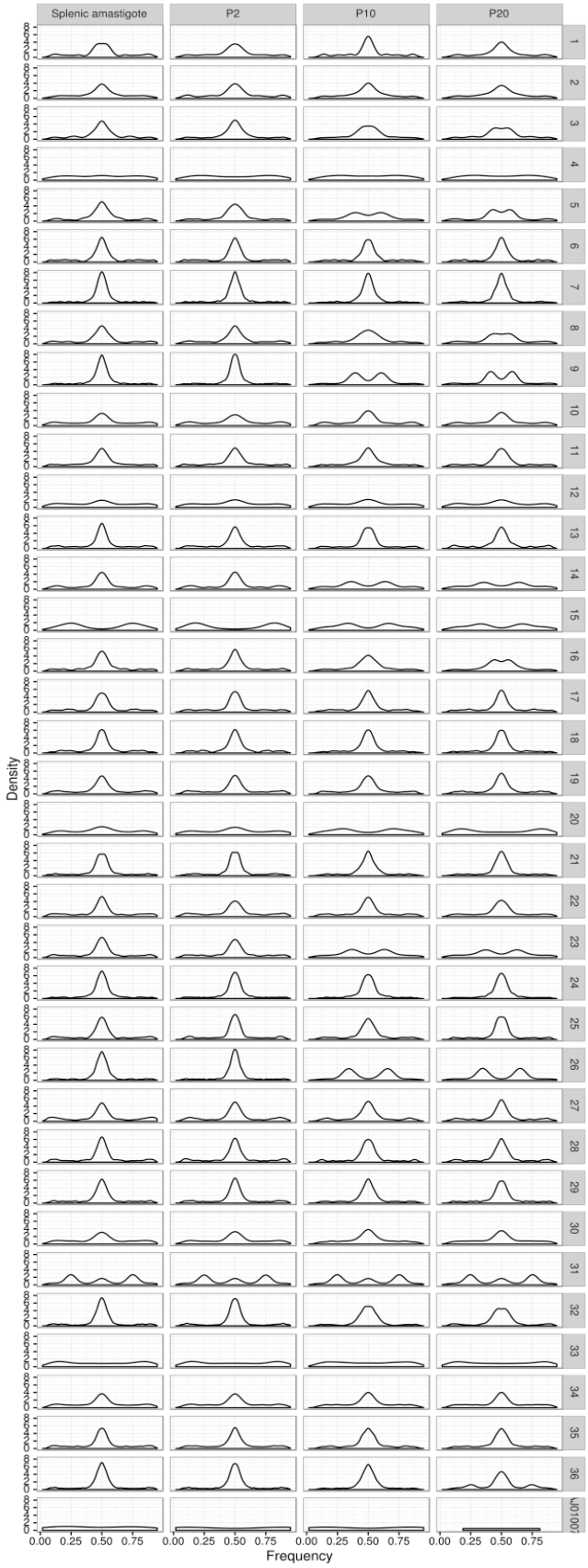
Supplementary Figure 4 co-occurrence of trisomies with predominantly bimodal (chr 5) and unimodal (chr 12) allele frequency distributions in field isolates. The isolates having a trisomy for both chr 12 and either 5 (A) or 6 (B) were selected and the allele profiles of the considered chromosomes were plotted. Note that no isolates was found to be simultaneously trisomic for 5, 6 and 12. The profiles show that in several isolates the allele profile shape are different between 12 and 5 or 6, with 12 (bottom line in A and B) being predominantly unimodal and 5 or 6 mostly bimodal.



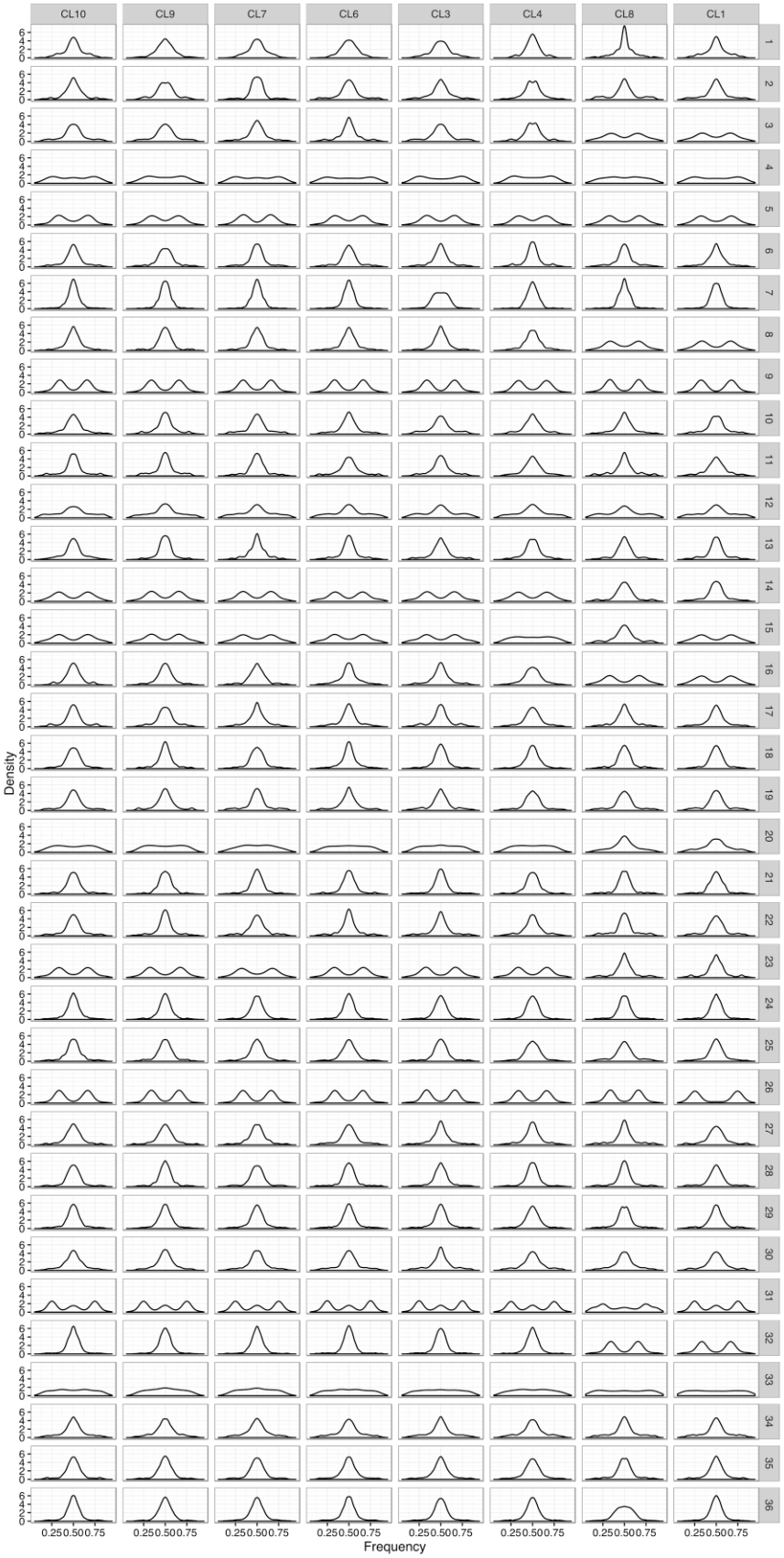
Supplementary Figure 5. Full karyotype variation of the LD1S adaptation time course. The original splenic amastigote, the P2, P10 and p20 passage as well as 8 subclones derived from P20 were deep sequenced, the reads were mapped against the same reference genome and the median RPKM was estimated for each chromosome of the time course (A). Eight subclones were then derived from P20 and sequenced in a similar way (B). The median read-depth profile (i.e. one value per chromosome) were then used to run a PCA analysis on the clones, thus revealing well defined subgroups.



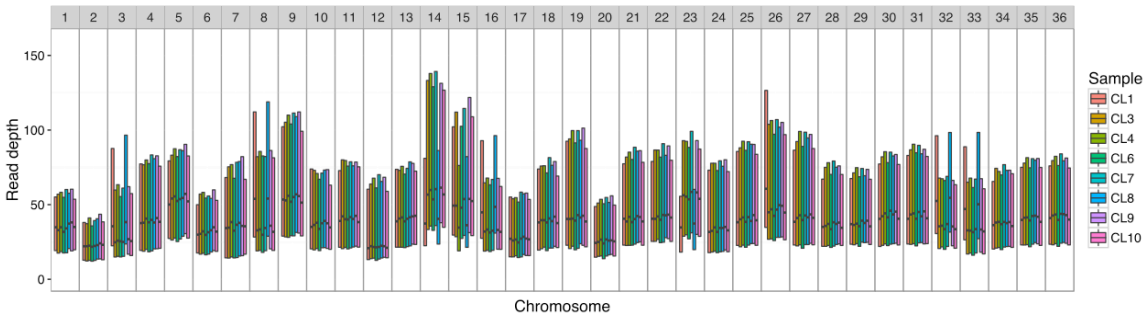
Supplementary Figure 6. Full Haplotype and karyotype variations maps of the 8 clones and p20. The chromosomes having no copy number variation are not shown. Each stripe corresponds to the dominant nucleotide. Sites in which the difference of frequency between the top allele and the second highest frequency allele is less than 10% are shaded in gray. The plot next to the painted chromosomes is the distribution of allele frequency colored according to the median read-depth of the chromosome.


Splenic amastigote
P2
P10
P20
Density
Frequency
0.00 0.25 0.50 0.75

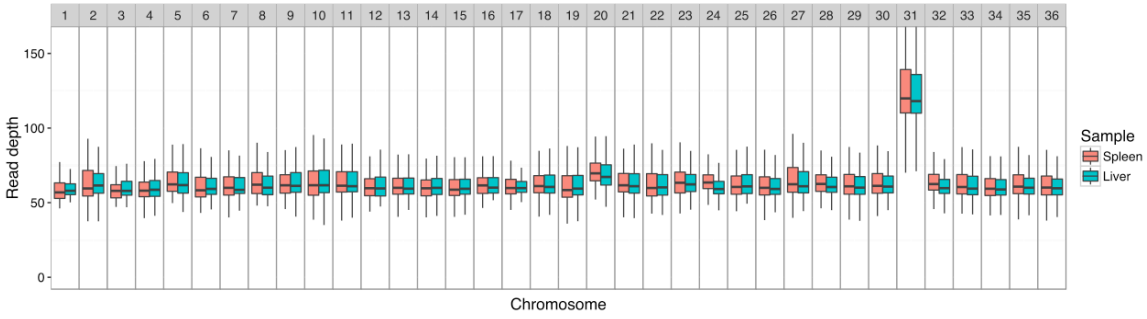
Supplementary Figure 7. Passage 1 to 20 allele profiles. This figure displays the individual allele frequency profiles of each chromosome for the splenic amastigote and the adapted cultures P2, P10 and P20.


CL10
CL9
CL7
CL6
CL3
CL4
CL8
CL1
Density
Frequency
0.250.500.75

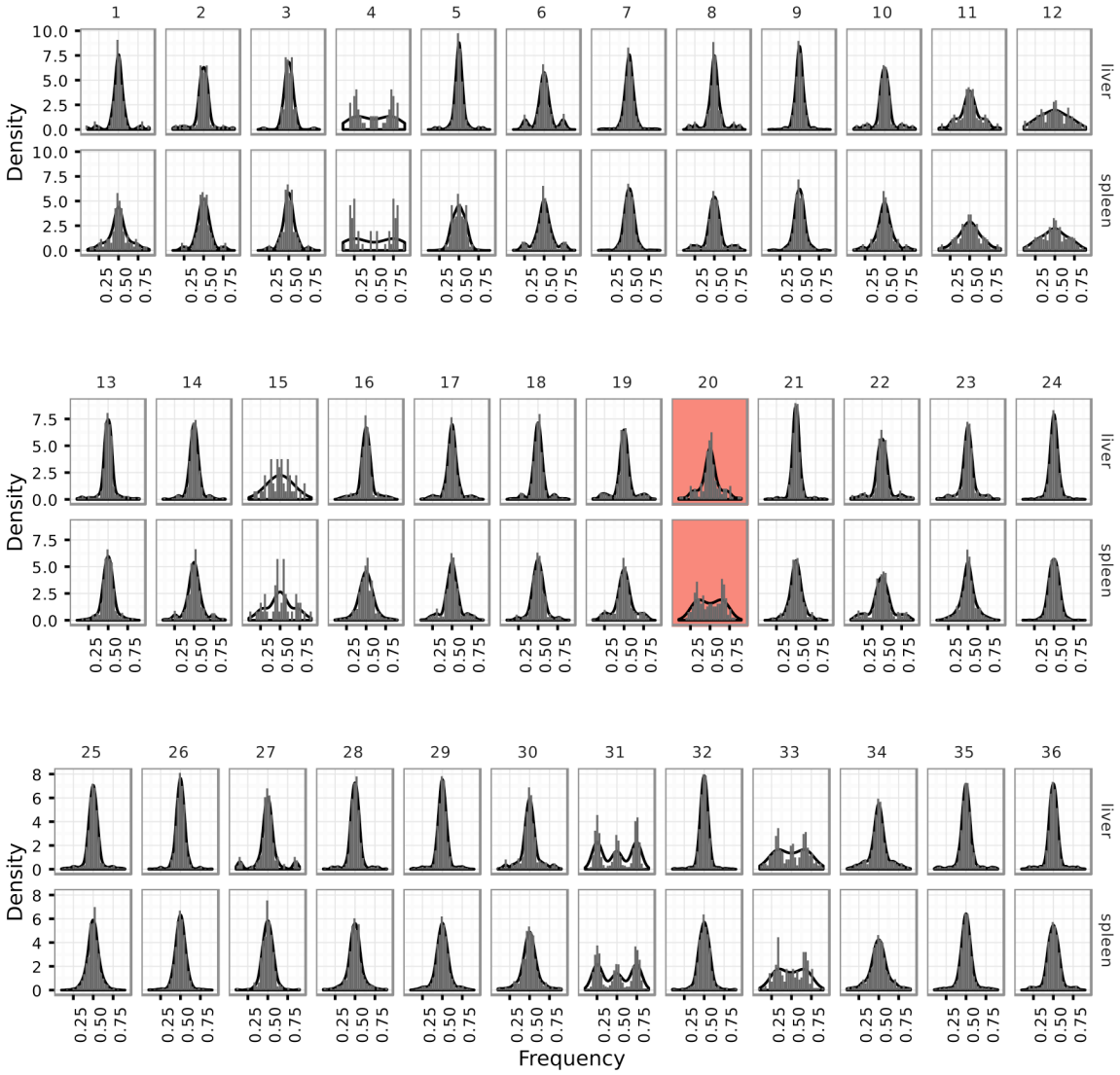
Supplementary Figure 8. Clones allele profiles. Similar analysis on the 8 sub-clones derived from P20.



Supplementary Figure 9. Distribution of chromosome transcriptomic output (in RPKM) in the 8 subclones derived from P20. The median values were used to populate Figure 5A.



Supplementary Figure 10. Spleen and Liver Boxplot. Distribution of genomic RPKM in the liver and spleen samples obtained after infecting a hamster with a P20 sample.



Supplementary Figure 11. Allelic frequency distributions in spleen and liver after re-infection. All profiles are statistically identical with the exception of chromosome 20 (highlighted in red).

4. Discussion

4.1. Challenges of reproducibility

Reproducibility in scientific research is a troubling matter, which has been growing concern in the scientific community over time. Sometimes even up to the extent as to describe current state in science as a crisis of reproducibility, after the collection of surveys with information about scientist's experiences (Baker 2016). The heart of scientific publication relies on the expectation of scientific research being knowledgeable and reproducible. And yet across years, many examples have been found and solutions have been proposed in order to identify, quantify and overcome this issues (Stodden et al. 2013; Firtina and Alkan 2016). More specifically in computational biology and bioinformatics, where there are inherent sources of variability while trying to convey any kind of analysis that relies on numerical calculations. These sources lead to differences in results, and therefore a certain level of instability than can be appreciated when executing the same analysis in different environments using the same methods, data and analysis. The solution used by Nextflow (Nextflow - A DSL for parallel and sca...; Di Tommaso) has been proposed previously (Boettiger 2014), but the lack of standardization and agreement on final solutions to tackle these issues has not lead to the adoption and integration of these in bioinformatics. This technology was new to bioinformatics and needed a deeper study and evaluation on the cost and the consequences upon introduction within bioinformatics analysis and its common runtime environments. Results in benchmarks of ENCODE (The ENCODE Project Consortium 2012) data sets and typical genomics applications used in large-scale projects proved this approach to be more than convenient and without a noticeable overhead or penalty in performance. This approach has received attention and has been proposed in guidelines on how to approach container technologies and take advantage for reproducible purposes following the same line proposed in our work (Moreews et al. 2015; Belmann et al. 2015; Pabinger et al. 2016; Byron et al. 2016).

Nowadays there has not been major further development in the area of workflow management tools. Although there are some tools available that have focused on the development and improvement towards achieving a better support for HPC or cloud environments (Afgan et al. 2012; Liu et al. 2014), there has not been recent development to tackle the main reproducibility barriers. With that in mind, Nextflow includes major support for different platforms and computational environments as well as shown in benchmark, but it has also focused on its development towards including features to manage seamless integration with publishing resources in order to ensure not only computational operability but also to embrace code sharing policies and guidelines long

proposed (Barnes 2010; Code share 2014; Easterbrook 2014), altogether integrated within the same umbrella.

The ultimate goal towards reproducibility of software management tools trying to address these issues is to harness all elements involved or related to computational analysis that might introduce any degree of instability. Although much is being done to try to solve these issues, not so much effort has been done in a real evaluation of the effect on real life applications. Therefore with applications existing such as Nextflow to ensure computational reproducibility, it is of imperial need to try to evaluate what is the extent of variability that needs to be acknowledged before trying to bring applications into the clinics.

So far, many aspects related to computation and numerical stability have been explored by workflow managers, but not so many aspects and efforts have been put in the data side. With large amounts of sequencing data accumulating and its raw input being archived in repositories like NCBI SRA's or EBI ENA (Leinonen et al. 2011; Leinonen et al. 2011), it is important to pay attention on how the data is being stored with all its related metadata, and allow a reference and standardized API in order to access and process all of it. This issue has been raised previously and some solutions have been proposed in order to structure and make a more meaningful retrieval of the data, but yet it hasn't been addressed and integrated at the same level of computation.

4.2. LncRNA evolution

Although sequence conservation in lncRNAs it's not as high as in protein coding genes, it's nonetheless higher than other elements in the genome such as ancient repeats (Ponjavic et al. 2007). Therefore, it is still useful to use phylogenetic depth, as the number of species in a tree up to which lncRNAs are conserved, together with the homology relationship and sequence level conservation in order to catalog and to prioritize analysis (Pervouchine et al. 2015) those lncRNAs whose enrichment and correlation with experimental data is high. Although there is some degree of conservation at the expression level between species, the lack of sequence conservation makes functional characterization challenging. In order to address this issue other studies have tried to correlate lower sequence conservation with RNA secondary structure conservation (Johnsson et al. 2014). Moreover, the data also suggested that not only sequence conservation could be constrained, but also transcriptional levels and its regulation as shown in the comparison of expression levels between mouse and human, enriched as well with epigenetic marks. Annotation of lncRNAs in curated datasets such as GENCODE for human and mouse genomes (Harrow et al. 2012; Mudge and Harrow

2015) had an increased step in the number of annotated gene and transcript models released year after year. Although a large number of genes have been included in reference annotations and in experimentally derived annotations, our results concerning a small fraction of the lncRNAs being highly constrained and conserved across mammalian annotation still holds and has been replicated by other following similar approaches recently in the literature (Chen et al. 2016; Gardner et al. 2015; Hezroni et al. 2015). Our approach to identify and estimate the level of sequence of conservation is stringent enough to call homology and orthology between species, as it will probably lack models which are fast-evolving and with a turnover rate much higher than typically seen in protein coding genes (Freyhult et al. 2007; Roshan et al. 2008). There are other approaches that have been explored, although using similar but less extensive data sets of experimental data (Washietl et al. 2014; Necsulea et al. 2014), but using RNA-seq read data mapping from one species to another as a proxy to estimate conservation is not a good model to establish homology relationship, but rather a way to find incomplete transcript processivity similarities, specially when transcript reconstruction and homology is not properly evaluated altogether from these analysis. It is often difficult to cross datasets from different sources specially when in different approaches have been used with different filters in order to try to classify lncRNA. This can be extrapolated from our comparison crossing recently published datasets of lncRNA, which did not provide neither a good support in terms of experimental evidences of transcription and full transcript model definition, neither good positional overlaps (Ravasi et al. 2010; FANTOM Consortium and the RIKEN PMI a...). Other approaches related to functional characterization through secondary structure have also raised critics and concerns as to what it can be detected when analyzing lncRNA structures (Rivas and Eddy 2000). Even when structures have conserved motifs might be arising due to sequence bias composition from specific regions or even degenerated elements, which might be confounding sequence and structure similarity with functional conservation.

Publication of new datasets and annotations have been accumulating through years on different species including ours (Li et al. 2015; Bateman et al. 2011; Paytuví Gallart et al. 2016; Amaral et al. 2011; Xuan et al. 2015), allowing to retrieve more information and annotations which can later be used in order to perform more integrative analysis. Annotation catalogs and databases on lncRNAs have become a powerful resource for research on this topic. Although sometimes, the information is quite heterogeneous and needs some manual curation and it also raises the question on which is the best dataset or how to merge all the information. Proper evaluation and curation will therefore be needed to meet quality standards as in the GENCODE resources.

Future work will benefit from the integrative approach followed in this work, helping to establishing a comparative map between species such as human and mouse that

provides a catalogue and a reference that guides and makes easier when comparing and trying to translate research between species. Comparative analysis and comparative genomics have provided the best strategies to gain insights into lncRNA evolution and functional characterization. Integrating all the tools and methods available has proven to be the best aim to tackle multiple questions, though many are still unanswered. In order to fully understand functional relevance of lncRNAs, after accumulation of experimental and evolutionary evidence further steps are needed to elucidate its functions *in vitro* and *in vivo* with experiments such as loss of function, transcription modulate or genome modification with CRISPR systems (Goff and Rinn 2015; Han et al. 2014). Other work carrying knockouts have shed some light into sponge-like mechanisms of lncRNAs involved in disease and cancer (Du et al. 2016), while other studies in mouse knockouts have shown the importance and impact of some of these players during developmental processes (Fatica and Bozzoni 2014).

4.3. Leishmania genomics and evolution

Leishmania is among the most deadliest pathogens in the world. L. donovani species when left untreated generates visceral leishmaniasis, provoking between 200,000 and 300,000 cases of the disease every year (Alvar et al. 2012). In the old world most of the cases of Leishmaniasis are coming from indian subcontinents, from where clinical genomic efforts have been carried to characterize the epidemiology and specially the rise of drug resistance (Imamura et al. 2016). In our work we revisited the 204 clinical isolates from indian subcontinents and found traces of recurrent episodes of aneuploidy in nearly the same chromosomes. Some of these events have also been seen in other species recently sequenced (Llanes et al. 2015). We continued these analysis with a hybrid strain in an *in vitro* system to monitor aneuploidy through time.

Aneuploidy previously detected and described in the literature is the event where chromosomes have more copies than expected, this is two copies of the typically diploid genome of Leishmania species (Sterkers et al. 2011; Sterkers et al. 2012). Due to our results in the 204 isolates we expect the genome of these parasites to be extremely dynamic for us to be able to capture aneuploidic events so often. We sought to interrogate the dynamics of these events by passaging a field isolate in an *in vivo* system, in hamster, where aneuploidy rapidly goes back towards disomic state. We show with FISH that these dynamic transitions of aneuploidies can still be observed in the hamster, with different organs showing different patterns of chromosome aneuploidies in a lower proportion as compared to cultured parasites, pointing towards the preexistence of aneuploidic subpopulations. It is interesting to point out that

different organs might reflect different environments and strongholds of the immune system that the parasite needs to adapt differently (Stanley and Engwerda 2007).

From our genomic analysis on the field isolates, traces of convergence towards the same haplotypes being favoured were observed on the chromosome copies that were amplified across the 204 field isolates. We therefore hypothesize on the opportunistic and mechanistic purpose of these events, happening more often than expected (Mannaert et al. 2012), and being more than a culture effect and rather a tool of the parasite.

In our results we do not find any trace of recombination, consistent with previous work discussing about the origins and mechanisms of leishmania in the intracellular stage, for which no sexual event has been observed (Victoir and Dujardin 2002; Rougeron et al. 2010). Although there is some controversy and findings related to parasexual and hybridization events with possible cases of automixis happening while the parasite is in the insect guts (Inbar et al. 2013; Akopyants et al. 2009; Sadlova et al. 2011). Despite of these cases can occur and the parasite is able to exchange genetic material and generic hybrids, it does not seem to occur often, and so the parasite is limited during the transition facing a radical change in the environment between insect and mammalian host, in which it may need to adapt very fast (Kaye and Scott 2011).

In our *in vitro* system we show how the aneuploid events have a high rate of turnover, showing even some cases of reversion towards disomic state with a transient trisomic event. This is consistent with the low heterozygosity shown in studies from clinical isolates, which would prevent chromosomes from accumulating many deleterious mutations (Rougeron et al. 2015). But, by using our hybridized strain with a high amount of mutations we are able to follow the fate of karyotypes and haplotypes at a single base resolution across the genome.

Aneuploidy has been studied in several species to understand what is the downstream effect and its level of tolerance in several species (Pavelka et al. 2010; Mulla et al. 2014; Gasch et al. 2016). But it is specially interesting in leishmania, as kinetoplastids lack any complex transcriptional regulation mechanisms, and therefore the differences in copy number impacts severely on the levels of expression and translation in the cell (Rogers et al. 2011). Overall we demonstrate recurrent whole chromosome aneuploidy events, with homogeneous increase in genomic depth, which is typically reflected in the transcriptome, except for some special cases as in chromosome 31. Our haplotype based analysis allows us to resolve which is the pattern of expression on RNA-seq data, and to see how many of these haplotypes are being expressed and at what rate. We also demonstrate a direct relationship between the number of copies per haplotype and its

relative levels of expression, not observing any distinction at the population level on which are the copies transcribed, reflecting the relative number of copies per chromosome both at the DNA and RNA levels.

Our results provide further insights into subpopulation structure without using single cell genome sequencing, obtaining clones after passaging parasites in our *in vitro* system. This method allows to perform deep sequencing of separate individuals from a large population while allowing to grow them after isolation and see its level of homogeneity, which has not been previously studied. Combining this strategy with our highly heterozygous hybrid has allowed us to elucidate the polyclonal origin buried from the beginning in infective populations. Such a complex structure of infectious population of parasites has been shown to evolve with environmental fitness adaptation and its consequences to virulence previously uncovered in Leishmania.

The importance of aneuploidy and the mechanism of fast adaptation to new environments is an interesting model system to study aneuploidy in cancer genetics as well. Genomic instability is a hallmark of cancer (Negrini et al. 2010), it has been observed in many cases how through aneuploidy cancer cells are able to become more resistant and escape drug target mechanisms or therapies (Gordon et al. 2012; Giam and Rancati 2015; Burrell et al. 2013) by either increasing cell to cell diversity by harvesting new mutations and/or having different number of copies per cell, which can lead to differences in dosage dependent expression and translation in the cell.

Future research in line with the results presented is aiming towards achieving a deeper level of understanding of this mechanism. Following our *in vitro* setup, to elucidate the subpopulation structures with a much finer grain strategy, single cell genome sequencing is already ongoing. Together with the *in vitro* system, it will allow to select subpopulation of parasites and evaluate them *in vivo* to detect changes in fitness related to differences in combinations of karyotypes and haplotypes between individuals and its pathogenic outcome in different environments. This line of research is going to provide some light into parasite population structure and how it impacts the infection capabilities of parasites. It is also of extreme importance to approach drug research and new therapies in the light of the work presented here. Typically clonal lines are used to develop and test treatments without taking into account how easy will be for another parasite through this mechanism to develop resistance. It is also very important to integrate this level of analysis in clinical research in order to screen parasite infections before treating them, to detect differences of subpopulations of parasites that could present difficulties and run into multidrug resistance problems.

Conclusion

From chapter Reproducibility:

- Nextflow is a new workflow management system developed and presented to address the main reproducibility issues that have been found during last years when dealing with complex computational pipelines. It provides features to avoid numerical instabilities between different environments, which leads to different results from the analysis, and also the possibility to automatically track, share and deploy published computational pipelines from repositories.
- The approach followed by Nextflow to develop pipelines allows to fast-prototype, its Dataflow model implicitly allows automatic parallelization in a large number of different environments, having support for all major platforms in HPC/cluster environments like Gridengine, Torque, PBS, Slurm, also in the cloud.
- The software provides a large set of pipelines already developed and tested publicly that are available and tested for several bioinformatics applications which can be used and deployed from day 0. Compared to other available workflow softwares we have one of the largest community support, and it is the only one providing all the features described.
- In our paper, container technology is benchmarked in several computational pipelines such as long non-coding RNA identification, variant calling and RNA quantification on the ENCODE datasets. This evaluation on the effects of introducing containers in our pipelines when running in HPC environments show how feasible and reliable this approach is. It was demonstrated that the overhead introduced in the pipelines is minimal, and even in some cases it presents an increase in performance.
- Pipelines developed and benchmarked in Nextflow allow seamless integration of containers. It simplifies the process and takes care of deploying computation with containers and allows pipelines to share and execute computational pipelines in a reproducible way. This features relieve authors from taking care of the technicalities related to virtualization and allows them to focus on the development of the analysis, while still having all the benefits in reproducibility and performance.

From chapter LncRNA evolution:

- Published an extended set of mouse gene and transcript set by mapping 17,547 human long non-coding RNA transcripts models annotated in Gencode to the mouse genome using a pipeline to find and map developed long non-coding

RNA mapping pipeline. From the homolog set 2,327 (13.26%) human lncRNA transcripts (corresponding to 1,679, or 15.48%, of the lncRNA genes) were homologous to 5,067 putative mouse transcripts (corresponding to 3,887 putative genes).

- Using one-to-one sequence homology and genome maps, we found and described a total set of 851 orthologs between human and mouse lncRNA genes. Of these, only 189 overlap with the set of 2,736 one-to-one human–mouse lncRNAs recently described, reflecting the yet incomplete characterization of mammalian lncRNAs, and of lncRNA orthologs.
- Expression levels of orthologous lncRNAs correlate weakly with phylogenetic depth (that is, the number of mammalian species in which a given lncRNA can be detected
- Although lncRNAs show distinct tissue- and species-specific expression patterns, we identified 12 lncRNAs expressed in at least 50% of the samples in each species. This small set of conserved broadly expressed genes may play important functions in mammalian cells. We found these genes to be highly enriched among nuclear lncRNAs
- Reviewed and expanded the collection of lncRNA in chicken genome using a combination of ab initio lncRNA set derived from RNAseq data from 20 different tissues and merged them with homology based predictions from human GENCODE LncRNAs set. Level of conservation was measured using a set of 42 newly sequenced avian genomes finding two subsets, 5,058 conserved in more than 10 genomes and another subset of 1251 conserved in more than 40 avian genomes.
- In the genome and transcriptome of Mesoamerican common bean we published a set of 1033 P.vulgaris lncRNA genes. The gene set was derived from a set of A. Thaliana lncRNAs found in the literature from which 38 were mapped into P.Vul genome using our pipeline, while the rest of genes were ab initio derived from RNAseq from 7 organs. The total set of LncRNAs were mapped onto 12 other plant genomes to evaluate the level of conservation in the plant kingdom and found 94% genes conserved in other bean genomes, and 526 bean specific lncRNA genes.

From chapter Leishmania:

- Part of this work presents the reanalysis of 204 L.Donovani clinical field isolates sequenced genomes from Indian subcontinents. Measurements of karyotype changes on the genomes and tracking of its evolutionary diversity accumulated on different parasite populations showing clear signals of a trend to generate more diversity in the chromosomes that often are evidenced by read depth analysis on the samples to undergo aneuploidy episodes.

- We show further results using a L. Donovani strain in a in vitro system and using high throughput sequencing how to detect and quantify chromosomal amplifications and track its fate across time in an in vitro system. Using further sequencing and subcloning of populations we demonstrate the existence of different subpopulations and using allele frequency we are able to differentiate between haplotype changes across time and across subpopulations.
- We manifest how many different combinations of chromosome copies with different haplotypes arise during the experiments, which allow the parasite to create differences among parasite subpopulations, which allows to explore a vast genotypic space. Together with changes in phenotypic features as growth and infectivity we see a correlation between haplotype changes and its response to the environment, proposing the usage of aneuploidy and haplotype selection as a mechanism of fast environmental adaptation in absence of sexual reproduction, while aneuploidy alone itself accounting for a mechanism to keep generating diversity due to the lack of heterozygosity.

Bibliography

Afgan, Enis, Brad Chapman, and James Taylor. 2012. "CloudMan as a Platform for Tool, Data, and Analysis Distribution." BMC Bioinformatics 13 (November): 315.

Afshinnekoo, Ebrahim, Cem Meydan, Shanin Chowdhury, Dyala Jaroudi, Collin Boyer, Nick Bernstein, Julia M. Maritz, et al. 2015. "Geospatial Resolution of Human and Bacterial Diversity with City-Scale Metagenomics." Cell Systems 1 (1): 72–87.

Akopyants, Natalia S., Nicola Kimblin, Nagila Secundino, Rachel Patrick, Nathan Peters, Phillip Lawyer, Deborah E. Dobson, Stephen M. Beverley, and David L. Sacks. 2009. "Demonstration of Genetic Exchange during Cyclical Development of Leishmania in the Sand Fly Vector." Science 324 (5924): 265–68.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. "Basic Local Alignment Search Tool." Journal of Molecular Biology 215 (3): 403–10.

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. "Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs." Nucleic Acids Research 25 (17): 3389–3402.

Alvar, Jorge, Iván D. Vélez, Caryn Bern, Mercé Herrero, Philippe Desjeux, Jorge Cano, Jean Jannin, Margriet den Boer, and WHO Leishmaniasis Control Team. 2012. "Leishmaniasis Worldwide and Global Estimates of Its Incidence." PloS One 7 (5): e35671.

Amaral, Paulo P., Michael B. Clark, Dennis K. Gascoigne, Marcel E. Dinger, and John S. Mattick. 2011. "lncRNAdb: A Reference Database for Long Noncoding RNAs." Nucleic Acids Research 39 (Database issue): D146–51.

Angiuoli, Samuel V., Malcolm Matalka, Aaron Gussman, Kevin Galens, Mahesh Vangala, David R. Riley, Cesar Arze, James R. White, Owen White, and W. Florian Fricke. 2011. "CloVR: A Virtual Machine for Automated and Portable Sequence Analysis from the Desktop Using Cloud Computing." BMC Bioinformatics 12 (August): 356.

"Announcement: Reducing Our Irreproducibility." 2013. Nature News 496 (7446): 398.

Baker, Monya. 2010. "Next-Generation Sequencing: Adjusting to Data Overload." Nature Methods 7 (7). Nature Publishing Group: 495–99.

———. 2016. "1,500 Scientists Lift the Lid on Reproducibility." Nature 533 (7604): 452–54.

Balazinska, Magdalena, Bill Howe, and Dan Suciu. n.d. "Data Markets in the Cloud: An Opportunity for the Database Community." http://www.vldb.org/pvldb/vol4/p1482-balazinska.pdf.

Barnes, Nick. 2010. "Publish Your Computer Code: It Is Good Enough." Nature 467 (7317): 753.

Bateman, Alex, Shipra Agrawal, Ewan Birney, Elspeth A. Bruford, Janusz M. Bujnicki, Guy Cochrane, James R. Cole, et al. 2011. "RNAcentral: A Vision for an International Database of RNA Sequences." RNA 17 (11): 1941–46.

Baud, Amelie, Victor Guryev, Oliver Hummel, Martina Johannesson, Rat Genome Sequencing and Mapping Consortium, and Jonathan Flint. 2014. "Genomes and Phenomes of a Population of Outbred Rats and Its Progenitors." Scientific Data 1 (June): 140011.

Belmann, Peter, Johannes Dröge, Andreas Bremges, Alice C. McHardy, Alexander Sczyrba, and Michael D. Barton. 2015. "Bioboxes: Standardised Containers for Interchangeable Bioinformatics Software." GigaScience 4 (October): 47.

Bilofsky, H. S., C. Burks, J. W. Fickett, W. B. Goad, F. I. Lewitter, W. P. Rindone, C. D. Swindell, and C. S. Tung. 1986. "The GenBank Genetic Sequence Databank." Nucleic Acids Research 14 (1): 1–4.

Blackwell, M., I. V. Grigoriev, and T. W. Jeffries. 2016. "Comparative Genomics of Biotechnologically Important Yeasts." Proceedings of the. National Acad Sciences. http://www.pnas.org/content/early/2016/08/16/1603941113.short.

Blankenberg, Daniel, Gregory Von Kuster, Emil Bouvier, Dannon Baker, Enis Afgan, Nicholas Stoler, Galaxy Team, James Taylor, and Anton Nekrutenko. 2014. "Dissemination of Scientific Software with Galaxy ToolShed." Genome Biology 15 (2): 403.

Boettiger, Carl. 2014. "An Introduction to Docker for Reproducible Research, with Examples from the R Environment." arXiv [cs.SE]. arXiv. http://arxiv.org/abs/1410.0846.

Bourne, Philip E. 2010. "What Do I Want from the Publisher of the Future?" PLoS Computational Biology 6 (5): e1000787.

Bourne, Philip E., Jon R. Lorsch, and Eric D. Green. 2015. "Perspective: Sustaining the Big-Data Ecosystem." Nature 527 (7576): S16–17.

Burrell, Rebecca A., Sarah E. McClelland, David Endesfelder, Petra Groth, Marie-Christine Weller, Nadeem Shaikh, Enric Domingo, et al. 2013. "Replication Stress Links Structural and Numerical Cancer Chromosomal Instability." Nature 494 (7438): 492–96.

Byron, Sara A., Kendall R. Van Keuren-Jensen, David M. Engelthaler, John D. Carpten, and David W. Craig. 2016. "Translating RNA Sequencing into Clinical Diagnostics: Opportunities and Challenges." Nature Reviews. Genetics 17 (5): 257–71.

Cheng, Timothy H. T., Deborah Thompson, Jodie Painter, Tracy O'Mara, Maggie Gorman, Lynn Martin, Claire Palles, et al. 2015. "Meta-Analysis of Genome-Wide Association Studies Identifies Common Susceptibility Polymorphisms for Colorectal and Endometrial Cancer near SH2B3 and TSHZ1." Scientific Reports 5 (December). Nature Publishing Group: 17369.

Chen, Jenny, Alexander A. Shishkin, Xiaopeng Zhu, Sabah Kadri, Itay Maza, Mitchell Guttman, Jacob H. Hanna, Aviv Regev, and Manuel Garber. 2016. "Evolutionary Analysis across Mammals Reveals Distinct Classes of Long Non-Coding RNAs." Genome Biology 17 (February): 19.

Christakis, Dimitri A., and Frederick J. Zimmerman. 2013. "Rethinking Reanalysis." JAMA: The Journal of the American Medical Association 310 (23): 2499–2500.

Cingolani, Pablo, Rob Sladek, and Mathieu Blanchette. 2015. "BigDataScript: A Scripting Language for Data Pipelines." Bioinformatics 31 (1): 10–16.

"Code Share." 2014. Nature 514 (7524): 536.

Collins, Francis S., and Lawrence A. Tabak. 2014. "Policy: NIH Plans to Enhance Reproducibility." Nature 505 (7485): 612–13.

Conesa, Ana, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, et al. 2016. "A Survey of Best Practices for RNA-Seq Data Analysis." Genome Biology 17 (January): 13.

Containers, Linux. n.d. "IBM Research Report." https://domino.research.ibm.com/library/cyberdig.nsf/papers/0929052195DD819C85257D2300681E7B/$File/rc25482.pdf.

Cyranoski, David. 2016. "Replications, Ridicule and a Recluse: The Controversy over NgAgo Gene-Editing Intensifies." Nature 536 (7615): 136–37.

Dahlö, Martin, Frédéric Haziza, Aleksi Kallio, Eija Korpelainen, Erik Bongcam-Rudloff, and Ola Spjuth. 2015. "BioImg.org: A Catalog of Virtual Machine Images for the Life Sciences." Bioinformatics and Biology Insights 9 (September): 125–28.

D'Antonio, Mattia, Paolo D'Onorio De Meo, Matteo Pallocca, Ernesto Picardi, Anna Maria D'Erchia, Raffaele A. Calogero, Tiziana Castrignanò, and Graziano Pesole. 2015. "RAP: RNA-Seq Analysis Pipeline, a New Cloud-Based NGS Web Application." BMC Genomics 16 (June): S3.

Davidson, Robert L., Ralf J. M. Weber, Haoyu Liu, Archana Sharma-Oates, and Mark R. Viant. 2016. "Galaxy-M: A Galaxy Workflow for Processing and Analyzing Direct Infusion and Liquid Chromatography Mass Spectrometry-Based Metabolomics Data." GigaScience 5 (February): 10.

Davison, A. 2012. "Automated Capture of Experiment Context for Easier Reproducibility in Computational Research." Computing in Science Engineering 14 (4): 48–56.

Dayhoff, Margaret Oakley, and National Biomedical Research Foundation. 1969. Atlas of Protein Sequence and Structure. [Vol. 1], [Vol. 1],. Silver Spring [Md.]: National Biomedical Research Foundation.

Dayhoff, M. O. 1965. "Computer Aids to Protein Sequence Determination." Journal of Theoretical Biology 8 (1): 97–112.

Dayhoff, M. O., and National Biomedical Research Foundation. 1979. Atlas of Protein Sequence and Structure. Atlas of Protein Sequence and Structure. National Biomedical Research Foundation.

Decullier, Evelyne, Laure Huot, Géraldine Samson, and Hervé Maisonneuve. 2013. "Visibility of Retractions: A Cross-Sectional One-Year Study." BMC Research Notes 6 (June): 238.

Deelman, Ewa, Gurmeet Singh, Mei-Hui Su, James Blythe, Yolanda Gil, Carl Kesselman, Gaurang Mehta, et al. 2005. "Pegasus: A Framework for Mapping Complex Scientific Workflows onto Distributed Systems." Scientific Programming 13 (3). Hindawi Publishing Corporation: 219–37.

Delisi, C., and D. M. Crothers. 1971. "Prediction of RNA Secondary Structure." Proceedings of the National Academy of Sciences of the United States of America 68 (11): 2682–85.

DePristo, Mark A., Eric Banks, Ryan Poplin, Kiran V. Garimella, Jared R. Maguire, Christopher Hartl, Anthony A. Philippakis, et al. 2011. "A Framework for Variation Discovery and Genotyping Using next-Generation DNA Sequencing Data." Nature Genetics 43 (5): 491–98.

Di Tommaso, Paolo. 2016. "A Curated List of Nextflow Pipelines." Accessed May 18. https://github.com/nextflow-io/awesome-nextflow/.

Dujon, Bernard. 2010. "Yeast Evolutionary Genomics." Nature Reviews. Genetics 11 (7). Nature Publishing Group: 512–24.

Dumas, J. P., and J. Ninio. 1982. "Efficient Algorithms for Folding and Comparing Nucleic Acid Sequences." Nucleic Acids Research 10 (1): 197–206.

Du, Zhou, Tong Sun, Ezgi Hacisuleyman, Teng Fei, Xiaodong Wang, Myles Brown, John L. Rinn, et al. 2016. "Integrative Analyses Reveal a Long Noncoding RNA-Mediated Sponge Regulatory Network in Prostate Cancer." Nature Communications 7 (March): 10982.

Earl, Dent, Ngan Nguyen, Glenn Hickey, Robert S. Harris, Stephen Fitzgerald, Kathryn Beal, Igor Seledtsov, et al. 2014. "Alignathon: A Competitive Assessment of Whole-Genome Alignment Methods." Genome Research 24 (12): 2077–89.

Easterbrook, Steve M. 2014. "Open Code for Open Science?" Nature Geoscience 7 (11). Nature Research: 779–81.

Edman, P. 1949. "A Method for the Determination of Amino Acid Sequence in Peptides." Archives of Biochemistry 22 (3): 475.

Edman, P., and G. Begg. 1967. "A Protein Sequenator." European Journal of Biochemistry / FEBS 1 (1): 80–91.

Eisenstein, Michael. 2015. "Big Data: The Power of Petabytes." Nature 527 (7576): S2–4.

"Error Prone." 2012. Nature 487 (7408): 406.

FANTOM Consortium and the RIKEN PMI and CLST (DGT), Alistair R. R. Forrest, Hideya Kawaji, Michael Rehli, J. Kenneth Baillie, Michiel J. L. de Hoon, Vanja Haberle, et al. 2014. "A Promoter-Level Mammalian Expression Atlas." Nature 507 (7493): 462–70.

Fatica, Alessandro, and Irene Bozzoni. 2014. "Long Non-Coding RNAs: New Players in Cell Differentiation and Development." Nature Reviews. Genetics 15 (1): 7–21.

Fickett, J. W. 1982. "Recognition of Protein Coding Regions in DNA Sequences." Nucleic Acids Research 10 (17): 5303–18.

Firtina, Can, and Can Alkan. 2016. "On Genomic Repeats and Reproducibility." Bioinformatics 32 (15): 2243–47.

Fitch, W. M., and E. Margoliash. 1967. "Construction of Phylogenetic Trees." Science 155 (3760): 279–84.

Freedman, Leonard P., Iain M. Cockburn, and Timothy S. Simcoe. 2015. "The Economics of Reproducibility in Preclinical Research." PLoS Biology 13 (6): e1002165.

Freedman, Leonard P., Mark C. Gibson, Stephen P. Ethier, Howard R. Soule, Richard M. Neve, and Yvonne A. Reid. 2015. "Reproducibility: Changing the Policies and Culture of Cell Line Authentication." Nature Methods 12 (6): 493–97.

Freedman, Leonard P., and James Inglese. 2014. "The Increasing Urgency for Standards in Basic Biologic Research." Cancer Research 74 (15): 4024–29.

Freyhult, Eva K., Jonathan P. Bollback, and Paul P. Gardner. 2007. "Exploring Genomic Dark Matter: A Critical Assessment of the Performance of Homology Search Methods on Noncoding RNA." Genome Research 17 (1): 117–25.

"Further Confirmation Needed." 2012. Nature Biotechnology 30 (9): 806.

Fusaro, Vincent A., Prasad Patil, Erik Gafni, Dennis P. Wall, and Peter J. Tonellato. 2011. "Biomedical Cloud Computing with Amazon Web Services." PLoS Computational Biology 7 (8): e1002147.

Gafni, Erik, Lovelace J. Luquette, Alex K. Lancaster, Jared B. Hawkins, Jae-Yoon Jung, Yassine Souilmi, Dennis P. Wall, and Peter J. Tonellato. 2014. "COSMOS: Python Library for Massively Parallel Workflows." Bioinformatics 30 (20): 2956–58.

Gallego Llorente, M., E. R. Jones, A. Eriksson, V. Siska, K. W. Arthur, J. W. Arthur, M. C. Curtis, et al. 2015. "Ancient Ethiopian Genome Reveals Extensive Eurasian Admixture throughout the African Continent." Science 350 (6262): 820–22.

Gardner, Paul P., Mario Fasold, Sarah W. Burge, Maria Ninova, Jana Hertel, Stephanie Kehr, Tammy E. Steeves, Sam Griffiths-Jones, and Peter F. Stadler. 2015. "Conservation and Losses of Non-Coding RNAs in Avian Genomes." PloS One 10 (3): e0121797.

Garijo, Daniel, Sarah Kinnings, Li Xie, Lei Xie, Yinliang Zhang, Philip E. Bourne, and Yolanda Gil. 2013. "Quantifying Reproducibility in Computational Biology: The Case of the Tuberculosis Drugome." PloS One 8 (11): e80278.

Gasch, Audrey P., James Hose, Michael A. Newton, Maria Sardi, Mun Yong, Zhishi Wang, and Duncan T. Odom. 2016. "Further Support for Aneuploidy Tolerance in Wild Yeast and Effects of Dosage Compensation on Gene Copy-Number Evolution." eLife 5 (March). eLife Sciences Publications Limited: e14409.

Gent, Ian P. 2013. "The Recomputation Manifesto." arXiv [cs.GL]. arXiv. http://arxiv.org/abs/1304.3674.

Giam, Maybelline, and Giulia Rancati. 2015. "Aneuploidy and Chromosomal Instability in Cancer: A Jackpot to Chaos." Cell Division 10 (1): 1–12.

Gil, Y., E. Deelman, M. Ellisman, T. Fahringer, G. Fox, D. Gannon, C. Goble, M. Livny, L. Moreau, and J. Myers. 2007. "Examining the Challenges of Scientific Workflows." Computer 40 (12): 24–32.

Global Alliance for Genomics and Health. 2016. "GENOMICS. A Federated Ecosystem for Sharing Genomic, Clinical Data." Science 352 (6291): 1278–80.

Goecks, Jeremy, Bassel F. El-Rayes, Shishir K. Maithel, H. Jean Khoury, James Taylor, and Michael R. Rossi. 2015. "Open Pipelines for Integrated Tumor Genome Profiles Reveal Differences between Pancreatic Cancer Tumors and Cell Lines." Cancer Medicine 4 (3): 392–403.

Goecks, Jeremy, Anton Nekrutenko, James Taylor, and Galaxy Team. 2010. "Galaxy: A Comprehensive Approach for Supporting Accessible, Reproducible, and Transparent Computational Research in the Life Sciences." Genome Biology 11 (8): R86.

Goff, Loyal A., and John L. Rinn. 2015. "Linking RNA Biology to lncRNAs." Genome Research 25 (10): 1456–65.

Goodstadt, Leo. 2010. "Ruffus: A Lightweight Python Library for Computational Pipelines." Bioinformatics 26 (21): 2778–79.

Gordon, David J., Benjamin Resio, and David Pellman. 2012. "Causes and Consequences of Aneuploidy in Cancer." Nature Reviews. Genetics 13 (3): 189–203.

Gunn, William. 2014. "Reproducibility: Fraud Is Not the Big Problem." Nature 505 (7484): 483.

Hamm, G. H., and G. N. Cameron. 1986. "The EMBL Data Library." Nucleic Acids Research 14 (1): 5–9.

Han, Jinxiong, Jun Zhang, Li Chen, Bin Shen, Jiankui Zhou, Bian Hu, Yinan Du, Peri H. Tate, Xingxu Huang, and Wensheng Zhang. 2014. "Efficient in Vivo Deletion of a Large Imprinted lncRNA by CRISPR/Cas9." RNA Biology 11 (7): 829–35.

Harrow, Jennifer, Adam Frankish, Jose M. Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L. Aken, et al. 2012. "GENCODE: The Reference Human Genome Annotation for The ENCODE Project." Genome Research 22 (9): 1760–74.

Hauser, Maria, Martin Steinegger, and Johannes Söding. 2016. "MMseqs Software Suite for Fast and Deep Clustering and Searching of Large Protein Sequence Sets." Bioinformatics 32 (9): 1323–30.

Helene Richter, S., Joseph P. Garner, Corinna Auer, Joachim Kunert, and Hanno Würbel. 2010. "Systematic Variation Improves Reproducibility of Animal Experiments." Nature Methods 7 (3). Nature Publishing Group: 167–68.

Henikoff, S., and J. G. Henikoff. 1992. "Amino Acid Substitution Matrices from Protein Blocks." Proceedings of the National Academy of Sciences of the United States of America 89 (22): 10915–19.

Hezroni, Hadas, David Koppstein, Matthew G. Schwartz, Alexandra Avrutin, David P. Bartel, and Igor Ulitsky. 2015. "Principles of Long Noncoding RNA Evolution Derived from Direct Comparison of Transcriptomes in 17 Species." Cell Reports 11 (7): 1110–22.

Higgins, D. G., and P. M. Sharp. 1988. "CLUSTAL: A Package for Performing Multiple Sequence Alignment on a Microcomputer." Gene 73 (1): 237–44.

Hoffman, Joseph I. 2016. "Reproducibility: Archive Computer Code with Raw Data." Nature 534 (7607): 326.

Hudson (Chairperson), Thomas J., Warwick Anderson, Axel Aretz, Anna D. Barker, Cindy Bell, Rosa R. Bernabé, M. K. Bhan, et al. 2010. "International Network of Cancer Genome Projects." Nature 464 (7291). Nature Publishing Group: 993–98.

Imamura, Hideo, Tim Downing, Frederik Van den Broeck, Mandy J. Sanders, Suman Rijal, Shyam Sundar, An Mannaert, et al. 2016. "Evolutionary Genomics of Epidemic

Visceral Leishmaniasis in the Indian Subcontinent." eLife 5 (March). doi:10.7554/eLife.12613.

Inbar, Ehud, Natalia S. Akopyants, Melanie Charmoy, Audrey Romano, Phillip Lawyer, Dia-Eldin A. Elnaiem, Florence Kauffmann, et al. 2013. "The Mating Competence of Geographically Diverse Leishmania Major Strains in Their Natural and Unnatural Sand Fly Vectors." PLoS Genetics 9 (7): e1003672.

Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." PLoS Medicine 2 (8): e124.

Ioannidis, John P. A., David B. Allison, Catherine A. Ball, Issa Coulibaly, Xiangqin Cui, Aedín C. Culhane, Mario Falchi, et al. 2009. "Repeatability of Published Microarray Gene Expression Analyses." Nature Genetics 41 (2): 149–55.

Ioannidis, J. P., E. E. Ntzani, T. A. Trikalinos, and D. G. Contopoulos-Ioannidis. 2001. "Replication Validity of Genetic Association Studies." Nature Genetics 29 (3): 306–9.

Johnsson, Per, Leonard Lipovich, Dan Grandér, and Kevin V. Morris. 2014. "Evolutionary Conservation of Long Non-Coding RNAs; Sequence, Structure, Function." Biochimica et Biophysica Acta 1840 (3): 1063–71.

Kallio, M. Aleksi, Jarno T. Tuimala, Taavi Hupponen, Petri Klemelä, Massimiliano Gentile, Ilari Scheinin, Mikko Koski, Janne Käki, and Eija I. Korpelainen. 2011. "Chipster: User-Friendly Analysis Software for Microarray and Other High-Throughput Data." BMC Genomics 12 (1): 1–14.

Kasson, Peter M. 2013. "Computational Biology in the Cloud: Methods and New Insights from Computing at Scale." Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, 451–53.

Kaye, Paul, and Phillip Scott. 2011. "Leishmaniasis: Complexity at the Host–pathogen Interface." Nature Reviews. Microbiology 9 (8). Nature Publishing Group: 604–15.

Kim, Daehwan, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L. Salzberg. 2013. "TopHat2: Accurate Alignment of Transcriptomes in the Presence of Insertions, Deletions and Gene Fusions." Genome Biology 14 (4): R36.

Kimura, M. 1969. "The Number of Heterozygous Nucleotide Sites Maintained in a Finite Population due to Steady Flux of Mutations." Genetics 61 (4): 893–903.

Kimura, Motoo, and Tomoko Ohta. 1971. "Protein Polymorphism as a Phase of Molecular Evolution." Nature 229 (5285). Nature Publishing Group: 467–69.

Köster, Johannes, and Sven Rahmann. 2012. "Snakemake--a Scalable Bioinformatics Workflow Engine." Bioinformatics 28 (19): 2520–22.

Krishnan, Muthukalingan, Chinnapandi Bharathiraja, Jeyaraj Pandiarajan, Vimalanathan Arun Prasanna, Jeyaprakash Rajendhran, and Paramasamy Gunasekaran. 2014. "Insect Gut Microbiome - An Unexploited Reserve for

Biotechnological Application." Asian Pacific Journal of Tropical Biomedicine 4 (Suppl 1): S16–21.

Kryshtafovych, Andriy, Krzysztof Fidelis, and John Moult. 2009. "CASP8 Results in Context of Previous Experiments." Proteins 77 Suppl 9: 217–28.

Lampa, Samuel, Jonathan Alvarsson, and Ola Spjuth. 2015. "Poster: Using Workflow Tools to Streamline Bioinformatics Computing on High-Performance E-Infrastructures," October. doi:10.13140/RG.2.1.1143.6246.

Lee, B., and F. M. Richards. 1971. "The Interpretation of Protein Structures: Estimation of Static Accessibility." Journal of Molecular Biology 55 (3): 379–400.

Lee, David, Oliver Redfern, and Christine Orengo. 2007. "Predicting Protein Function from Sequence and Structure." Nature Reviews. Molecular Cell Biology 8 (12): 995–1005.

Leinonen, Rasko, Ruth Akhtar, Ewan Birney, Lawrence Bower, Ana Cerdeno-Tárraga, Ying Cheng, Iain Cleland, et al. 2011. "The European Nucleotide Archive." Nucleic Acids Research 39 (Database issue): D28–31.

Leinonen, Rasko, Hideaki Sugawara, Martin Shumway, and International Nucleotide Sequence Database Collaboration. 2011. "The Sequence Read Archive." Nucleic Acids Research 39 (Database issue): D19–21.

Leipzig, Jeremy. 2016. "A Review of Bioinformatic Pipeline Frameworks." Briefings in Bioinformatics, March. doi:10.1093/bib/bbw020.

Le Novère, Nicolas, Andrew Finney, Michael Hucka, Upinder S. Bhalla, Fabien Campagne, Julio Collado-Vides, Edmund J. Crampin, et al. 2005. "Minimum Information Requested in the Annotation of Biochemical Models (MIRIAM)." Nature Biotechnology 23 (12): 1509–15.

Li, Aimin, Junying Zhang, Zhongyin Zhou, Lei Wang, Yujuan Liu, and Yajun Liu. 2015. "ALDB: A Domestic-Animal Long Noncoding RNA Database." PloS One 10 (4): e0124003.

Li, Heng, and Nils Homer. 2010. "A Survey of Sequence Alignment Algorithms for next-Generation Sequencing." Briefings in Bioinformatics 11 (5): 473–83.

Lipman, D. J., S. F. Altschul, and J. D. Kececioglu. 1989. "A Tool for Multiple Sequence Alignment." Proceedings of the National Academy of Sciences of the United States of America 86 (12): 4412–15.

Liu, Bo, Ravi K. Madduri, Borja Sotomayor, Kyle Chard, Lukasz Lacinski, Utpal J. Dave, Jianqiang Li, Chunchen Liu, and Ian T. Foster. 2014. "Cloud-Based Bioinformatics Workflow Platform for Large-Scale next-Generation Sequencing Analyses." Journal of Biomedical Informatics 49 (June): 119–33.

Llanes, Alejandro, Carlos Mario Restrepo, Gina Del Vecchio, Franklin José Anguizola, and Ricardo Lleonart. 2015. "The Genome of Leishmania Panamensis: Insights into Genomics of the L. (Viannia) Subgenus." Scientific Reports 5 (February): 8550.

Mack, Steven J., Robert P. Milius, Benjamin D. Gifford, Jürgen Sauter, Jan Hofmann, Kazutoyo Osoegawa, James Robinson, et al. 2015. "Minimum Information for Reporting next Generation Sequence Genotyping (MIRING): Guidelines for Reporting HLA and KIR Genotyping via next Generation Sequencing." Human Immunology 76 (12): 954–62.

Mannaert, An, Tim Downing, Hideo Imamura, and Jean-Claude Dujardin. 2012. "Adaptive Mechanisms in Pathogens: Universal Aneuploidy in Leishmania." Trends in Parasitology 28 (9): 370–76.

Martin, A. J., and R. L. Synge. 1941. "A New Form of Chromatogram Employing Two Liquid Phases: A Theory of Chromatography. 2. Application to the Micro-Determination of the Higher Monoamino-Acids in Proteins." Biochemical Journal 35 (12): 1358–68.

Marx, Vivien. 2013. "Biology: The Big Challenges of Big Data." Nature 498 (7453): 255–60.

McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, et al. 2010. "The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data." Genome Research 20 (9): 1297–1303.

Missier, Paolo, Stian Soiland-Reyes, Stuart Owen, Wei Tan, Alexandra Nenadic, Ian Dunlop, Alan Williams, Tom Oinn, and Carole Goble. 2010. "Taverna, Reloaded." In Scientific and Statistical Database Management, edited by Michael Gertz and Bertram Ludäscher, 471–81. Lecture Notes in Computer Science 6187. Springer Berlin Heidelberg.

Moreews, François, Olivier Sallou, Hervé Ménager, Yvan Le Bras, Cyril Monjeaud, Christophe Blanchet, and Olivier Collin. 2015. "BioShaDock: A Community Driven Bioinformatics Shared Docker-Based Tools Registry." F1000Research 4 (December): 1443.

Morgulis, Aleksandr, George Coulouris, Yan Raytselis, Thomas L. Madden, Richa Agarwala, and Alejandro A. Schäffer. 2008. "Database Indexing for Production MegaBLAST Searches." Bioinformatics 24 (16): 1757–64.

Morin, A., J. Urban, P. D. Adams, I. Foster, A. Sali, D. Baker, and P. Sliz. 2012. "Research Priorities. Shining Light into Black Boxes." Science 336 (6078): 159–60.

Mudge, Jonathan M., and Jennifer Harrow. 2015. "Creating Reference Gene Annotation for the Mouse C57BL6/J Genome Assembly." Mammalian Genome: Official Journal of the International Mammalian Genome Society 26 (9-10): 366–78.

Muir, Paul, Shantao Li, Shaoke Lou, Daifeng Wang, Daniel J. Spakowicz, Leonidas Salichos, Jing Zhang, et al. 2016. "The Real Cost of Sequencing: Scaling Computation to Keep Pace with Data Generation." Genome Biology 17 (March): 53.

Mulla, Wahid, Jin Zhu, and Rong Li. 2014. "Yeast: A Simple Model System to Study Complex Phenomena of Aneuploidy." FEMS Microbiology Reviews 38 (2): 201–12.

"Must Try Harder." 2012. Nature 483 (7391): 509.

Nalls, Mike A., Nathan Pankratz, Christina M. Lill, Chuong B. Do, Dena G. Hernandez, Mohamad Saad, Anita L. DeStefano, et al. 2014. "Large-Scale Meta-Analysis of Genome-Wide Association Data Identifies Six New Risk Loci for Parkinson's Disease." Nature Genetics 46 (9): 989–93.

Napolitano, Francesco, Renato Mariani-Costantini, and Roberto Tagliaferri. 2013. "Bioinformatic Pipelines in Python with Leaf." BMC Bioinformatics 14 (June): 201.

Necsulea, Anamaria, Magali Soumillon, Maria Warnefors, Angélica Liechti, Tasman Daish, Ulrich Zeller, Julie C. Baker, Frank Grützner, and Henrik Kaessmann. 2014. "The Evolution of lncRNA Repertoires and Expression Patterns in Tetrapods." Nature 505 (7485): 635–40.

Negrini, Simona, Vassilis G. Gorgoulis, and Thanos D. Halazonetis. 2010. "Genomic Instability--an Evolving Hallmark of Cancer." Nature Reviews. Molecular Cell Biology 11 (3): 220–28.

"Nextflow - A DSL for Parallel and Scalable Computational Pipelines." 2016. Accessed September 21. https://www.nextflow.io/.

Nocq, Julie, Magalie Celton, Patrick Gendron, Sebastien Lemieux, and Brian T. Wilhelm. 2013. "Harnessing Virtual Machines to Simplify next-Generation DNA Sequencing Analysis." Bioinformatics 29 (17): 2075–83.

Pabinger, Stephan, Karina Ernst, Walter Pulverer, Rainer Kallmeyer, Ana M. Valdes, Sarah Metrustry, Denis Katic, et al. 2016. "Analysis and Visualization Tool for Targeted Amplicon Bisulfite Sequencing on Ion Torrent Sequencers." PloS One 11 (7). journals.plos.org: e0160227.

Pauling, L., and H. A. Itano. 1949. "Sickle Cell Anemia a Molecular Disease." Science 110 (2865): 543–48.

Pavelka, Norman, Giulia Rancati, Jin Zhu, William D. Bradford, Anita Saraf, Laurence Florens, Brian W. Sanderson, Gaye L. Hattem, and Rong Li. 2010. "Aneuploidy Confers Quantitative Proteome Changes and Phenotypic Variation in Budding Yeast." Nature 468 (7321). Nature Research: 321–25.

Paytuví Gallart, Andreu, Antonio Hermoso Pulido, Irantzu Anzar Martínez de Lagrán, Walter Sanseverino, and Riccardo Aiese Cigliano. 2016. "GREENC: A Wiki-Based Database of Plant lncRNAs." Nucleic Acids Research 44 (D1): D1161–66.

Pervouchine, Dmitri D., Sarah Djebali, Alessandra Breschi, Carrie A. Davis, Pablo Prieto Barja, Alex Dobin, Andrea Tanzer, et al. 2015. "Enhanced Transcriptome Maps from Multiple Mouse Tissues Reveal Evolutionary Constraint in Gene Expression." Nature Communications 6 (January): 5903.

Peter, Amstutz, Crusoe Michael R., Tijanić Nebojša, Chapman Brad, Chilton John, Heuer Michael, Kartashov Andrey, et al. 2016. "Common Workflow Language, v1.0." Figshare, July. doi:10.6084/m9.figshare.3115156.v2.

Piccolo, Stephen R., and Michael B. Frampton. 2016. "Tools and Techniques for Computational Reproducibility." GigaScience 5 (1): 30.

Ponjavic, Jasmina, Chris P. Ponting, and Gerton Lunter. 2007. "Functionality or Transcriptional Noise? Evidence for Selection within Long Noncoding RNAs." Genome Research 17 (5): 556–65.

Prinz, Florian, Thomas Schlange, and Khusru Asadullah. 2011. "Believe It or Not: How Much Can We Rely on Published Data on Potential Drug Targets?" Nature Reviews. Drug Discovery 10 (9): 712.

Ravasi, Timothy, Harukazu Suzuki, Carlo Vittorio Cannistraci, Shintaro Katayama, Vladimir B. Bajic, Kai Tan, Altuna Akalin, et al. 2010. "An Atlas of Combinatorial Transcriptional Regulation in Mouse and Man." Cell 140 (5): 744–52.

Rhinn, Herve, Ryousuke Fujita, Liang Qiang, Rong Cheng, Joseph H. Lee, and Asa Abeliovich. 2013. "Integrative Genomics Identifies APOE ε4 Effectors in Alzheimer's Disease." Nature 500 (7460): 45–50.

Rivas, E., and S. R. Eddy. 2000. "Secondary Structure Alone Is Generally Not Statistically Significant for the Detection of Noncoding RNAs." Bioinformatics 16 (7): 583–605.

Rogers, Matthew B., James D. Hilley, Nicholas J. Dickens, Jon Wilkes, Paul A. Bates, Daniel P. Depledge, David Harris, et al. 2011. "Chromosome and Gene Copy Number Variation Allow Major Structural Change between Species and Strains of Leishmania." Genome Research 21 (12): 2129–42.

Roquet, Fabien, Guy Williams, Mark A. Hindell, Rob Harcourt, Clive McMahon, Christophe Guinet, Jean-Benoit Charrassin, et al. 2014. "A Southern Indian Ocean Database of Hydrographic Profiles Obtained with Instrumented Elephant Seals." Scientific Data 1 (September): 140028.

Roshan, Usman, Satish Chikkagoudar, and Dennis R. Livesay. 2008. "Searching for Evolutionary Distant RNA Homologs within Genomic Sequences Using Partition Function Posterior Probabilities." BMC Bioinformatics 9 (January): 61.

Roth, Kevin A., and Audra E. Cox. 2015. "Science Isn't Science If It Isn't Reproducible." The American Journal of Pathology 185 (1): 2–3.

Rougeron, V., T. De Meeûs, and A-L Bañuls. 2015. "A Primer for Leishmania Population Genetic Studies." Trends in Parasitology 31 (2): 52–59.

Rougeron, Virginie, Thierry De Meeûs, Sandrine Kako Ouraga, Mallorie Hide, and Anne-Laure Bañuls. 2010. "'Everything You Always Wanted to Know about Sex (but Were Afraid to Ask)' in Leishmania after Two Decades of Laboratory and Field Analyses." PLoS Pathogens 6 (8). Public Library of Science: e1001004.

Sadedin, Simon P., Bernard Pope, and Alicia Oshlack. 2012. "Bpipe: A Tool for Running and Managing Bioinformatics Pipelines." Bioinformatics 28 (11): 1525–26.

Sadlova, Jovana, Matthew Yeo, Veronika Seblova, Michael D. Lewis, Isabel Mauricio, Petr Volf, and Michael A. Miles. 2011. "Visualisation of Leishmania Donovani Fluorescent Hybrids during Early Stage Development in the Sand Fly Vector." PloS One 6 (5). Public Library of Science: e19851.

Sallou, Olivier, and Cyril Monjeaud. n.d. "GO-Docker: A Batch Scheduling System with Docker Containers." In 2015 IEEE International Conference on Cluster Computing, 514–15. IEEE.

Sanger, F. 1945. "The Free Amino Groups of Insulin." Biochemical Journal 39 (5): 507–15.

Sheldrick, George M. 2008. "A Short History of ıt SHELX." Acta Crystallographica. Section A, Foundations of Crystallography 64 (1): 112–22.

Shepherd, J. C. 1981. "Method to Determine the Reading Frame of a Protein from the Purine/pyrimidine Genome Sequence and Its Possible Evolutionary Justification." Proceedings of the National Academy of Sciences of the United States of America 78 (3): 1596–1600.

Souilmi, Yassine, Jae-Yoon Jung, Alex Lancaster, Erik Gafni, Saaid Amzazi, Hassan Ghazal, Dennis Wall, and Peter Tonellato. 2015. "COSMOS: Cloud Enabled NGS Analysis." BMC Bioinformatics 16 (2): 1–1.

Sroka, Jacek, Jan Hidders, Paolo Missier, and Carole Goble. 2010. "A Formal Semantics for the Taverna 2 Workflow Model." Journal of Computer and System Sciences 76 (6): 490–508.

Stallman, Richard M., Roland McGrath, and Paul D. Smith. 2004. GNU Make: A Program for Directing Recompilation, for Version 3.81. Free Software Foundation.

Stanley, Amanda C., and Christian R. Engwerda. 2007. "Balancing Immunity and Pathology in Visceral Leishmaniasis." Immunology and Cell Biology 85 (2): 138–47.

States, D. J., and W. Gish. 1994. "Combined Use of Sequence Similarity and Codon Bias for Coding Region Identification." Journal of Computational Biology: A Journal of Computational Molecular Cell Biology 1 (1): 39–50.

Steijger, Tamara, Josep F. Abril, Pär G. Engström, Felix Kokocinski, The RGASP Consortium, Tim J. Hubbard, Roderic Guigó, Jennifer Harrow, and Paul Bertone. 2013. "Assessment of Transcript Reconstruction Methods for RNA-Seq." Nature Methods 10 (12). Nature Research: 1177–84.

Steinbiss, Sascha, Fatima Silva-Franco, Brian Brunk, Bernardo Foth, Christiane Hertz-Fowler, Matthew Berriman, and Thomas D. Otto. 2016. "Companion: A Web Server for Annotation and Analysis of Parasite Genomes." Nucleic Acids Research, April. doi:10.1093/nar/gkw292.

Stein, L. 2001. "Genome Annotation: From Sequence to Biology." Nature Reviews. Genetics 2 (7): 493–503.

Stein, Lincoln D. 2010. "The Case for Cloud Computing in Genome Informatics." Genome Biology 11 (5): 1–7.

Stephens, Zachary D., Skylar Y. Lee, Faraz Faghri, Roy H. Campbell, Chengxiang Zhai, Miles J. Efron, Ravishankar Iyer, Michael C. Schatz, Saurabh Sinha, and Gene E. Robinson. 2015. "Big Data: Astronomical or Genomical?" PLoS Biology 13 (7): e1002195.

Sterkers, Yvon, Laurence Lachaud, Nathalie Bourgeois, Lucien Crobu, Patrick Bastien, and Michel Pagès. 2012. "Novel Insights into Genome Plasticity in Eukaryotes: Mosaic Aneuploidy in Leishmania." Molecular Microbiology 86 (1): 15–23.

Sterkers, Yvon, Laurence Lachaud, Lucien Crobu, Patrick Bastien, and Michel Pagès. 2011. "FISH Analysis Reveals Aneuploidy and Continual Generation of Chromosomal Mosaicism in Leishmania Major." Cellular Microbiology 13 (2): 274–83.

Stodden, Victoria, Peixuan Guo, and Zhaokun Ma. 2013. "Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals." PloS One 8 (6): e67111.

Sussman, Gerald Jay. n.d. "Building Robust Systems an Essay." https://groups.csail.mit.edu/mac/users/gjs/6.945/readings/robust-systems.pdf.

Telenti, Amalio, Levi T. Pierce, William H. Biggs, Julia di Iulio, Emily H. M. Wong, Martin M. Fabani, Ewen F. Kirkness, et al. 2016. "Deep Sequencing of 10,000 Human Genomes." bioRxiv. doi:10.1101/061663.

The Cancer Genome Atlas Research Network, John N. Weinstein, Eric A. Collisson, Gordon B. Mills, Kenna R. Mills Shaw, Brad A. Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M. Stuart. 2013. "The Cancer Genome Atlas Pan-Cancer Analysis Project." Nature Genetics 45 (10). Nature Research: 1113–20.

The ENCODE Project Consortium. 2012. "An Integrated Encyclopedia of DNA Elements in the Human Genome." Nature 489 (7414). Nature Research: 57–74.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. "CLUSTAL W: Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting,

Position-Specific Gap Penalties and Weight Matrix Choice." Nucleic Acids Research 22 (22): 4673–80.

Tringe, Susannah Green, and Edward M. Rubin. 2005. "Metagenomics: DNA Sequencing of Environmental Samples." Nature Reviews. Genetics 6 (11). Nature Publishing Group: 805–14.

Van der Auwera, Geraldine A., Mauricio O. Carneiro, Christopher Hartl, Ryan Poplin, Guillermo del Angel, Ami Levy-Moonshine, Tadeusz Jordan, et al. 2002. "From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline." In Current Protocols in Bioinformatics. John Wiley & Sons, Inc.

Van Noorden, Richard, Brendan Maher, and Regina Nuzzo. 2014. "The Top 100 Papers." Nature 514 (7524): 550–53.

Victoir, Kathleen, and Jean-Claude Dujardin. 2002. "How to Succeed in Parasitic Life without Sex? Asking Leishmania." Trends in Parasitology 18 (2): 81–85.

Waltemath, Dagmar, Richard Adams, Daniel A. Beard, Frank T. Bergmann, Upinder S. Bhalla, Randall Britten, Vijayalakshmi Chelliah, et al. 2011. "Minimum Information About a Simulation Experiment (MIASE)." PLoS Computational Biology 7 (4): e1001122.

Wang, Jianwu, and Ilkay Altintas. 2012. "Early Cloud Experiences with the Kepler Scientific Workflow System." Procedia Computer Science 9 (January): 1630–34.

Wang, Zhaoming, Wei Jie Seow, Kouya Shiraishi, Chao A. Hsiung, Keitaro Matsuo, Jie Liu, Kexin Chen, et al. 2016. "Meta-Analysis of Genome-Wide Association Studies Identifies Multiple Lung Cancer Susceptibility Loci in Never-Smoking Asian Women." Human Molecular Genetics 25 (3): 620–29.

Washietl, Stefan, Manolis Kellis, and Manuel Garber. 2014. "Evolutionary Dynamics and Tissue Specificity of Human Long Noncoding RNAs in Six Mammals." Genome Research, January. doi:10.1101/gr.165035.113.

Wolstencroft, Katherine, Robert Haines, Donal Fellows, Alan Williams, David Withers, Stuart Owen, Stian Soiland-Reyes, et al. 2013. "The Taverna Workflow Suite: Designing and Executing Workflows of Web Services on the Desktop, Web or in the Cloud." Nucleic Acids Research 41 (Web Server issue): W557–61.

Xuan, Hongdong, Linzhong Zhang, Xueshi Liu, Guomin Han, Juan Li, Xin Li, Aiguo Liu, Mingzhi Liao, and Shihua Zhang. 2015. "PLNlncRbase: A Resource for Experimentally Identified lncRNAs in Plants." Gene 573 (2): 328–32.

Ye, Yuzhen, Jeong-Hyeon Choi, and Haixu Tang. 2011. "RAPSearch: A Fast Protein Similarity Search Tool for Short Reads." BMC Bioinformatics 12 (1): 1–10.

Zarowiecki, Magdalena. 2012. "Metagenomics with Guts." Nature Reviews. Microbiology 10 (10). Nature Publishing Group: 674–674.

Zhao, Yongan, Haixu Tang, and Yuzhen Ye. 2012. "RAPSearch2: A Fast and Memory-Efficient Protein Similarity Search Tool for next-Generation Sequencing Data." Bioinformatics 28 (1): 125–26.