TESI DOCTORAL UPF 2017

Evolution of the non-coding genome

Carla Bello

Director

Fyodor A. Kondrashov

Evolutionary Genomics Group

Center for Genomic Regulation (CRG)

Contents

Al recuerdo de mi padre

Acknowledgements

I will start this section with a brief biography about myself because I believe that who you become throughout your life is a combination of chance, necessity and self-awareness.

I grew up in a scientific environment due to both my parents being ecologists. It was because of the countless field trips we did during my childhood in Latin America and also in Spain, where I explored nature and slept under the stars, that I developed my interest in understanding the natural world.

There were many people I've met along the way that contributed to who I am. Many I left in my country, but they know they're part of me.

I wanna thank **Rory Johnson** for going to the University of Barcelona and giving a passionate talk about long non-coding RNAs (lncRNAs) while I was doing my Master and introducing me to the interesting world of lncRNAs. Without that experience I wouldn't have been so directional about my decision of going to the Center of Genomic Regulation (CRG) and going through the Bioinformatics and Genomics Group Leaders list to find the Evolutionary Genomics Lab, where my now supervisor, **Fedya**, had a fabulous picture of him in diving attire.

Like I said, I've always grew up fascinated with science and my favorite topic was evolution, that's why I was doing my Masters on Developmental Biology and Genetics, however deep inside I had a frustrated dream of being a Marine Biologist. That's why, *in part*, seeing **Fedya's** portrait gave me a nice comforting feeling and I decided to ask him to finish my Masters in his lab, to which after debating about my size and the physical space, he accepted.

I'll always be thankful for the interaction I had with **Rory** and **Fedya**, because that combination led me to study what I mainly focused on during my PhD: the evolution of lncRNAs.

I wanna thank **Fedya**, my supervisor, for introducing me to so many new experiences and of course to the concept of *epistasis*. I didn't really worked with that subject during my PhD, but I believe that thanks to his influence It is a phenomena I will never take for granted in future studies. I thank him because he always trusted I was capable to make it in the end. When I started in his lab I was so much more naive about science. I believed that we were still in the Jacob and Monod's scientific era, but through time and thanks to so many open discussions with him about the reality of science I believe I have become a much grounded person and for that I'm thankful. He has always supported me the best he can and I appreciate it very much. I am looking forward to many more scientific and personal discussions

with him throughout my life.

The members of the lab have changed a lot during my PhD and I feel that I need to thank them all. **Toby** was always there for me and answered all my questions with great knowledge and I really enjoyed all our lunches by the beach. Danke schöen for believing in me and all your advice.

Michael was the Californian surfer from the lab and he was nice to me. I remember him constantly doing multiple alignments and talking about tRNAs. Thank you for giving me the 'Learning Perl' book from O'Reilly. I'll never forget how you laughed when I asked you *how long* it would take me to be *good* at Perl.

Thank you **Romain**, for showing me how important good alignments are for constructing proper phylogenetic trees and thoughtful discussions. It has been a pleasure to meet you and I'm always happy to get news from you.

Thank you **Margarita/Margo/Margosha**, for listening to me and sharing your personal stories. I value your friendship and you were missed when you left the lab. I was sad to see your poster of *Adventure Time* disappear. I'm still keeping the postcard you sent me from Chicago and the small *matryoshka* you gave me when you left. I hope I can see you again in our future endeavors.

Inna, thank you because you seemed to care about me. You once asked me if I thought and alignment was good and I felt you trusted my input, which to me, coming from you, was an honor. Also thank you because you showed empathy in one of the worst moments of my PhD.

Thank you **Petya**, for always being so honest and funny, you showed me a lot about protein structure and how modeling interactions can help us finding targets for drug development.

Masha, thank you for your kindness, you always have a smile in your face and it makes me happy to see you. Thank you for the mushrooms, plants and homemade crafts you were always preparing. You showed me how unconventional methods can work when you are passionate about your research even against all odds. I'll never forget all the glowing worms and the fish tanks.

Onuralp, thank you for being honest and always being true to yourself. You were always trying to give me advice about the future. It was fun talking about anime and movies with you.

Mérci **Oriol**, for showing me *els valors catalans* and all about Barcelona's football. I can't believe how much you've grown since I met you! It was fun discussing with you about projects and playing around with *Awk&Sed*.

Dinara, your kindness and joyful spirit were and inspiration. Seeing you persevere and succeed in your goals makes me feel happy. Thank you for having coffees with me, listening and trusting me.

Vika, you've always were so empathetic and kind to me, when you left the lab I felt so sad but happy because you were pursuing your dreams, thank you also for listening and understanding me.

Julia, thank you for been such a free spirit, you show me that we should take things slowly, meditate and do yoga. I still haven't been able to follow through, but seeing it in you was a good influence. Thank you for all the laughs and all the funny mice videos.

Mateusz, it was great to have you in the lab. You were the one I could always talk about population genomics, ecology and the importance of protecting our ecosystems. I really enjoyed our discussions and of course all the great *Krupnik* and *Soplica* we shared. Thank you for being my friend and for your support.

Masha Tutukina, even though you came and went throughout the years I grew very fond of you and was always happy to see you coming back to Barcelona. Thank you for interesting discussions about science and being a friend.

Natalya and Dmitry, you were the first couple in our lab, thank you for being kind and caring. **Dmitry's** knowledge in protein struc-

ture and his talent in piano, guitar and singing made the lab interesting and fun. **Natalya**, thank you for being thoughtful towards me and asking me how I was throughout my PhD.

Thanks to all the newer members of the lab, **Ana** for her constant hugs, her warm madrileñan-cuban heart, and long discussions about the future. **Karen** and **Katya** for being the cool couple always thinking outside the box and for adding a modern attitude to the lab. **Katya**, it was really fun talking about art and new ways of visualizing data with you and discussing about sensitive related topics without judgment. **Karen**, it was equally nice talking about ideas and discussing the future of science, humanity and the world. **Pilar**, I'm happy that you also got passionate about lncRNAs, joining me in my project and going against the stream. Also for relating with me in different ailments related to one of our knees. **Lorena** you were always the soul of the party, having a positive attitude throughout adversities, thank you for being nice! **Louisa,** you were probably the first person I've ever consciously met from Andorra, I enjoyed taking about your fluffy cat and your mellowness around the lab.

I would like to thank all **Toni Gabaldón's lab**, including him. Thank you **Toni**, for being a friend. His lab was many times like a second lab for me.

Salva, I will always remember the first time I met you. You were so

funny and indignant about the fact that *now there were two venezuelans* in the CRG! Thank you for many coffees and helping me out when I didn't know how to solve something. I am proud that you have succeeded in becoming what you wanted.

Alex, you grew on me with time and I was sad to see you leave, I am still waiting for an invitation to Greece! **Leszek;** I always had troubles writing your last name, I want to thank you for being super fun and for helping me with scripts! **Jaime** (I should have listened to your devotion to Python and followed sooner), **Gabriela** and **Damien** were always nice to me as well. **Laia!** thank you for listening to my complaints about stuff and being patient with me. **Marina**, I've always seen you like the wisest person in your lab and your passion about Fungi is inspiring. Thanks to **Damian** for also becoming a friend, and even though he is not in that lab anymore we've kept in contact. Thank you **Cinta** and **Ewa** for listening to me and strangely **Ernst**; thank you for your input about the annotation of lncRNAs, I wish we had shared more. **Jesse**, I didn't met you very well but you seem like a nice guy. Likewise, **Irene** and all the other members for not kicking me out of the office when I go talk to **Laia**.

I wanna thank all the members of my Thesis Committee, **Tomás Marquès-Bonet, Guillaume Fillion** and **Stephan Ossowski**.

They were a source of inspiration and helped me when I was stuck and desperate because my project wasn't working out.

Cedric Notredame was always there as well, all his discussions about alignments and his philosophical conversations were very interesting to have. **Cedric Magis**, thank you for organizing the *no-paper celebration gathering*, it always reminded me that it is not a race, and that the whole point is to enjoy what we do.

Roderic Guigó, I know I should know català by now, but I promise that If I stay longer, or I ever come back I will make it a priority. I am happy to have met you and for showing your support.

Anna Vlasova, it was great to have shared so many conversations with you, about scientific and personal issues. *Spasibo* for being such a nice person.

Francisco Câmara, it was really nice having so many discussions about our personal issues and also about genome annotation.

Romina, I coudn't have made it without you, you're the best of the best! thanks to you I was able to do many things and you are the most efficient person I know. **Imma**, you were always sweet with me and very understanding, thank you for supporting me when I was down and always being there for me. **Gloria**, you gave me to sign my first contract in CRG and were always amazing as well. Mèrci.

I also want to thank the **IT department**, specially **Oscar, Isma, Gabriel, Diego** and all their team. Thanks for helping me installing libraries and fixing annoying bugs.

I thank the **CRG** as an institute for being a great environment to do science in the beautiful **PRBB** building in front of the Mediterranean sea.

There were a lot of people in **IBE** that helped me out, **Nino, Arturo, Diego, Marc, Juan, Lucas**, and many others. Thank you so much for fruitful discussions.

Of course, I wanna thank the **Spanish Ministry of Economy and Competitiveness** for giving me a grant (BES-2013-064004) and also the tax payers, because without that money I wouldn't have been able to do my PhD.

Christos, tack så-mycket for being there for me and guiding me through the whole process of becoming a PhD. You were my rock and you always listened and advice me when I felt lost. Your company has been invaluable to me. You've stayed with me through all, good and bad times, and I thank you for it.

My sister, **Alhena**. Gracias por estar para mí sin importar nada. Despite us being very different and very far away, we've kept connected and gone through tough times together. Thank you for being

my sister.

I wanna thank my uncle **Juan Luis**, because he was a great influence on me when I was growing up and many of his interesting attributes I picked up.

And last, but not least, I wanna thank **my parents**, for the inspiration they gave me. For teaching me constellations and tree names and because after all I wouldn't exist if it wasn't because of their love. They are the first humans I'm thankful for. Without them I wouldn't be the person I am now. Gracias mamá y papá, por hacerme en gran parte lo que soy y apoyarme incondicionalmente.

I've probably left people out, but I want to thank them all. Every encounter I have made has made me the person I am today.

Good and bad experiences, I am grateful for all, because that's life, and life is beautifully peachy.

Mérci Barcelona, for being an amazing city to live.

Abstract

Identifying functional elements in the non-coding genome remains a challenging task. A comparative genomic approach to this problem can help us understand how these elements evolved and give us insights into their functions. In this thesis, we focused mainly on the evolution of long non-coding RNA genes (lncRNAs) in the human lineage. These genes are known to have several key regulatory roles in development and disease making them suitable candidates for exploration. Duplication of genetic material plays a key role in the generation of novelties in genomes. Here, we hypothesized that the differences between protein-coding regions between humans and other primates might not be sufficient to explain our unique features. Therefore, using a comparative genomic approach we evaluated the contribution of lncRNA exon duplications to the human genome. We identified 62 human-specific genes that were fixed in the population and showed signs of active selection, together with tissue-specific patterns of expression. Our findings suggest that these genes might be relevant for the evolution of human-specific features and require further experimental validation. Moreover, we also studied these genes in a non-model endangered species; the Spoon-billed Sandpiper and identified 37 lncRNAs that were highly conserved in humans. Finally, we analyzed transfer RNA genes (tRNAs) between different

species of bacteria and found that there is a limit to the number of tRNA identities that can evolve due to structural and functional constraints, restricting the incorporation of new amino acids into the genetic code. Taken together, our studies focused on genes that reside in the non-coding genome and contribute to the understanding of its function.

Keywords: lncRNAs; tRNAs; duplications; evolution; humans; novelties.

Resumen

La identificación de elementos funcionales en el genoma no-codificante continúa siendo una tarea desafiante. La genómica comparativa puede ayudarnos a entender como estos elementos han evolucionado y cuáles son sus posibles funciones. En ésta tesis, nos hemos enfocado principalmente en la evolución de unos genes conocidos como *long non-coding RNAs* (lncRNAs) en el linaje humano. Éstos genes están involucrados en muchos procesos fundamentales de regulación genética e influyen múltiples procesos de desarrollo y enfermedades, lo cuál los hace candidatos idóneos de exploración. Las duplicaciones que ocurren en el genoma cumplen un rol fundamental en la generación de nuevo material genético. Aquí, hemos hipotetizado que las diferencias existentes entre las regiones codificantes de proteínas entre humanos y otros primates quizás no sea suficiente para explicar nuestras características fenotípicas únicas. Es por ello, que utilizando métodos de genómica comparativa hemos evaluado cuál es la contribución de las duplicaciones exónicas de los lncRNAs sobre nuestro genoma. De ésta manera, hemos identificado 62 genes humanos específicos que se han fijado en la población y que además muestran signos de selección activa y expresión tejido-específica. Nuestros descubrimientos sugieren que éstos genes pueden ser relevantes para la evolución de las características fenotípicas únicas de los seres humanos y requieren de

validación experimental en el futuro. Más aún, hemos estudiado éstos genes en una especie no-modelo que se encuentra en peligro de extinción; el pájaro conocido como "Correlimos Cuchareta", identificando 37 lncRNAs que se encuentran altamente conservados en humanos. Finalmente, hemos analizado RNAs de transferencia (tRNAs) entre diferentes especies de bacterias encontrando que existe un límite en el número de identidades que los tRNAs pueden evolucionar debido a una restricción estructural y funcional, lo cual impide la incorporación de nuevos aminoácidos en el código genético. En conjunto, nuestros estudios se han enfocado en genes que residen en el genoma no-codificante y contribuyen a la compresión de su funcionamiento.

Palabras clave: lncRNAs; tRNAs; duplicaciones; evolución; humanos.

Thesis overview

This thesis mainly focused on the evolution of long non-coding RNAs in the human genome through exon duplication. It addressed the question of whether these specific lncRNAs might be contributing to human-specific features by having an effect in organ development and/or disease.

Chapter 1. Is an introduction to the general importance of the non-coding genome and its effects on phenotype; it has a historical mindset and focuses mainly on lncRNAs. This first chapter gives a fundamental overview about genomic duplications and how they are known to be involved in the mechanisms of new gene generation.

Chapter 2. Presents the main objectives of the thesis.

Chapter 3. Describes how lncRNA exon duplications are involved in the origin of human-specific genes. It offers a valuable method to identify candidate long non-coding RNA genes through comparative and evolutionary approaches and by utilizing public expression datasets.

Chapter 4. Shows that the rate of nucleotide substitution is higher in species with fewer tRNA genes than in species with higher number of tRNAs because there is a saturation of recognition signals that

blocks the emergence of new tRNA identities.

Chapter 5. compares the genomes of two birds, the spoon-billed sandpiper, and its sister species, the red-necked stint; the former with an endangered population and the latter with a relatively stable population. It shows that the gradual decline of the spoon-billed sandpiper population led to an accumulation of rare deleterious alleles which reduces the fitness of the species and therefore inbreeding should be avoided in future conservation programs.

Chapter 6. Is a general discussion where I present my perspectives about the future of genomics and remark the importance of understanding the regulatory elements that control it; this understanding is fundamental for our wellbeing, as well as for the survival of other species on Earth.

Finally the **Appendix** compiles a list of studies I've participated throughout my PhD.

Chapter 1

Introduction

1.1 The non-coding genome and phenotype

From mobile elements to functional RNA molecules, the non-coding genome is defined as any region in the genome without coding potential. It comprises 98% of the total genome with the remaining 2% being protein-coding[Human Genome Sequencing Consortium, 2004], and it includes transposable elements (TEs), ribosomal RNAs (rRNAs), microRNAs (miRNAs), transfer RNAs (tRNAs), long non-coding RNAs (lncRNAs), etc. Many studies about evolution have mainly focused on protein-coding regions, and although a lot of effort

has being shifted in the last decades to understand the non-coding genome and its relevance in phenotype; there is still a large proportion of it that remains uncharacterized. With the inclusion of other relevant factors in the "Central Dogma of Molecular Biology" [Francis Crick, 1970], such as epigenetics and post-translational modifications (**Figure 1**), the importance of characterizing all the elements in the genome has become of paramount significance to many, specially to understand the nature of disease and evolution. While it is true that there might not be a need to understand every single functional element of the genome and its interactions to achieve a global understanding of disease and evolution; a goal that has been proven to be very difficult, the identification of key elements that could have a role in specific traits needs to be resolved.

Due to its self-catalytic properties, RNA is the only molecule to be known to act as a both, genotype and phenotype [Guerrier-Takada et al., 1983, Cech, 1986]. It is this unique property that gave rise to the hypothesis of the "RNA world" [Gilbert, 1986], which proposes that early life on Earth originated from self-replicating RNA molecules rather than proteins [Joyce, 1989, Cech, 2009]. This hypothesis although largely accepted, has not yet been experimentally proven. Nonetheless, there is compelling evidence supporting the "RNA world" hypothesis and it is clear that RNA has played a key

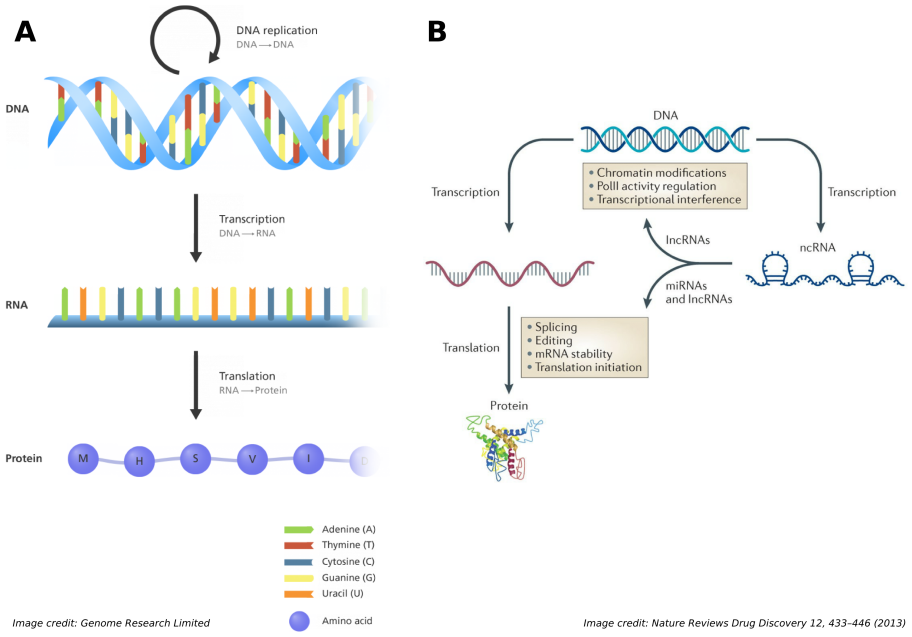

Figure 1: The evolution of the "Central Dogma of Molecular Biology". From the original one way model of "*DNA makes RNA and RNA makes protein*" in 1956 by Francis Crick **(A)** to an intricate system that goes beyond that model since 1970 **(B)** with the discovery of the reverse transcriptase by David Baltimore. The inclusion of other elements and processes in the modified version of the central dogma, such as transcriptional interference, RNA replication, RNA editing, ncRNAs and post-translational modifications are shown in **(B)**. *Image from Genome Research limited (A) and [Wahlestedt, 2013] (B).*

role in the origin of life [Robertson and Joyce, 2012] and the evolution of the species through gene regulation [Morris and Mattick, 2014] (**Figure 2**). It is therefore of uttermost interest to study how RNA interacts with all the components of the cell.

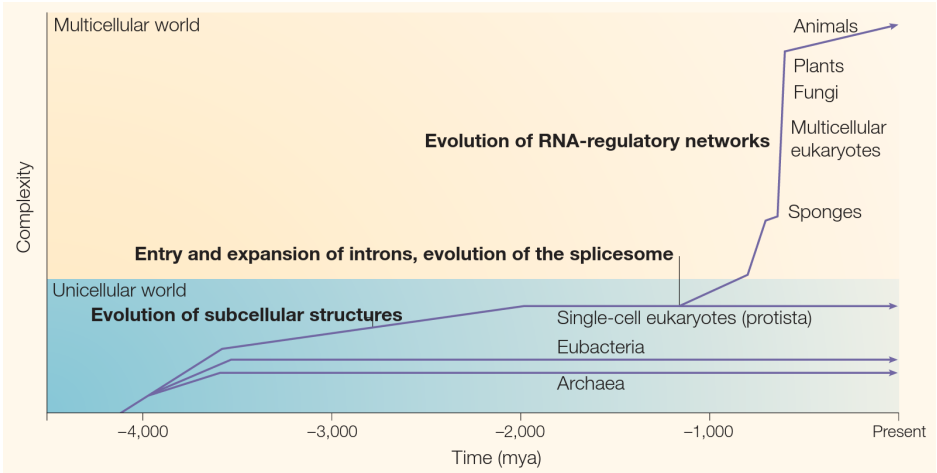

Figure 2: The evolution of life on Earth. (*simplified*) . *Image from:* [Mattick, 2004]

1.1.1 RNA evolution and function

Historically, the elucidation of the double helix of DNA [Watson and Crick, 1953, Franklin and Gosling, 1953] and the consequent discoveries of mRNA, tRNAs and the ribosomal machinery, together with the postulation of the "Central Dogma", centered the attention to proteins and the understanding of the genetic code. It wasn't until Jacob and Monod's proposed the *lac* operon model that it became apparent that gene regulation was an important mechanism governing biological systems. When evidence started showing that most of the genome was transcribed and that its composition was rich in repeated sequences [Britten and Kohne, 1968]; which was later

labeled as "junk" DNA [Ohno, 1972] it became generally accepted that most of it was non-functional. However, other key findings such as self-catalytic RNAs, the X-chromosome inactivation lncRNA *Xist*, miRNAs and siRNAs (**Figure 3**) supported the idea of RNA having a key role in the regulation of development and in the evolution of complex organisms.

After the human genome was sequenced [Venter et al., 2001, Lander et al., 2001], another project, known as The Encyclopedia of DNA Elements (ENCODE); which comprises the collective effort from many scientists around the world was formed [Material et al., 2004, Birney et al., 2007]. The project started as a pilot where the goal was to identify all the functional elements in 1% of the human genome and since then it has expanded at the whole genome level and to other model organisms (modENCODE).

One of the many key findings of the ENCODE project was the pervasive transcription of the genome; their methods allowed for the identification of many transcripts that didn't code for proteins, including long transcripts with "gene like" structure with exons and introns and alternative splicing. These transcripts had already been proposed in the 90's by Mattick, but the ENCODE project confirmed the existence of thousands of these transcripts in our genome. Today they are commonly known as long non-coding RNAs (lncRNAs).

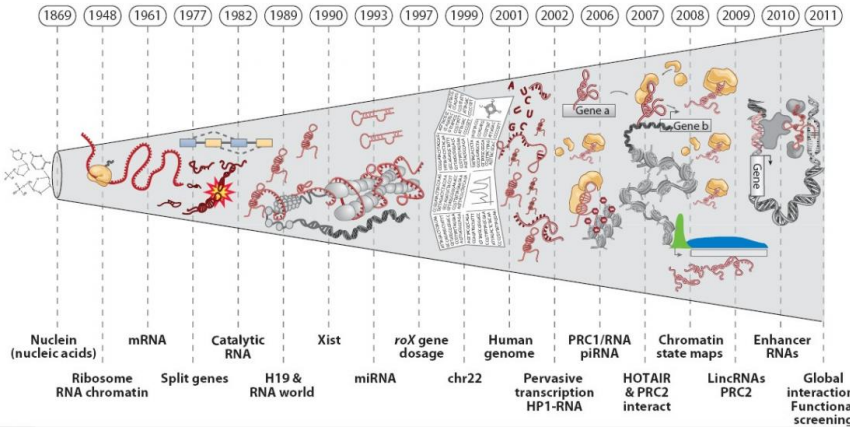

Figure 3: Time-scale of RNA discoveries through time. [*Image from:*[Rinn and Chang, 2012]

1.2 Long non-coding RNAs

Long non-coding RNAs are defined by a, seemingly arbitrary limit of at least 200 nucleotides in length and their non-coding potential. This arbitrary length was due to the nature of the experimental protocols and to differentiate lncRNAs from the average length of miRNAs, tRNAs, rRNAs and others. There are different types of lncRNAs and their classification depends on the relative position of protein-coding genes **(Figure 4)**.

RNA is a versatile molecule, and thus, lncRNAs have been found implicated in a diverse pool of regulatory mechanisms which include regulation of transcription via chromatin modifications, post-

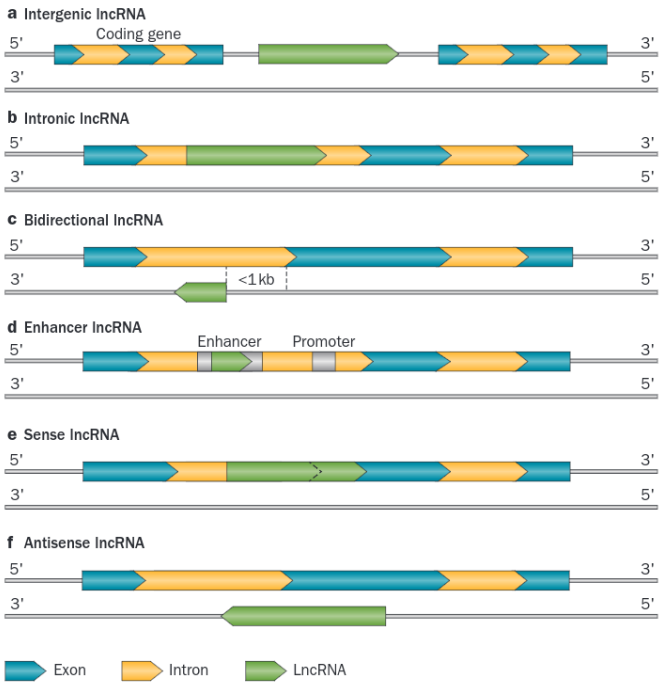

Figure 4: Classification of different lncRNA transcripts according to their genomic location. a) *Intergenic lncRNAs* (lincRNAs) are located between protein-coding genes. b) *Intronic lncRNAs* are located in an intron of a protein coding gene. c) *Bidirectional lncRNAs* are located within 1 kb of promoters in the opposite direction from the protein-coding transcript. d) *Enhancer lncRNAs* (elncRNAs) are located in enhancer regions. e) *Sense lncRNAs* are transcribed from the sense strand of protein-coding genes and overlap one or several introns and exons. f) *Antisense lncRNAs* are transcribed from the antisense strand of proteinâĂŚcoding genes and overlap one or several introns and exons of the sense sequence. [*Image and legend from:*[Devaux et al., 2015]

translational regulation of protein activity, cell-cell signaling, organization of protein complexes, allosteric regulation of proteins, as well

as recombination [Geisler and Coller, 2013].

Interestingly, the non-coding genome increases in size with the developmental complexity of the organism[Mattick, 2004]; the same scenario has been observed in lncRNAs. It has been shown that lncRNAs have a high turn-over rate [Ponting et al., 2009, Ponting et al., 2011] and that they evolve much faster than protein-coding genes which could indicate either a high plasticity of the studied genomes or an active implication of these genes in the evolution of more complex organisms.

1.2.1 From "junk" to key elements in development and disease

Although only a small proportion of lncRNAs have been well characterized, such as *Xist, H19, HOTAIR, MALAT1, NEAT1*, to name just a few; their prevalent expression and relative conservation in mammals [Guttman et al., 2009, Ponting et al., 2009, Mattick, 2010, Meader et al., 2010, Ulitsky et al., 2011, Kutter et al., 2012, Necsulea et al., 2014] has driven the hypothesis that a significant fraction might be relevant to the functional outcome of the genomes. However, even though the number of lncRNAs that have been characterized has grown in the last decade, it is still a matter of debate whether the majority of these transcripts are functional or transcriptional noise.

Mostly due to their lack of sequence conservation throughout evolution; despite that lack of sequence conservation doesn't necessarily implies lack of function [Ulitsky et al., 2011]. Nonetheless, identifying non-coding functional elements in the genome and its interactions is not a trivial task and several methods have been proposed to annotate these elements in the genome[Alexander et al., 2010], though genome wide association studies (GWAS), comparative genomics between different species, detection of signatures of natural selection [Andolfatto, 2005] and structural variants, gene expression and the *guilty by association* approach that connects the closest protein and its function to the neighboring non-coding gene or element; which works particularly well for antisense lncRNAs (NATs) and their chromatin mediated mechanism of epigenetic interaction connecting proteins to DNA [Magistri et al., 2012].

The diversity, versatility and the ability of lncRNAs to activate or repress gene expression has linked lncRNAs to many developmental processes, including senescence and aging [Grammatikakis et al., 2014, Montes et al., 2015], cardiovascular diseases [Devaux et al., 2015] and several types of cancers [Prensner and Chinnaiyan, 2011, Huarte, 2015], among others. Taken together, all these properties makes these genes interesting candidates to explore their contribution to human genome evolution.

1.3 Duplication and the origin of novel genes

The restrictive and conservative nature of natural selection in the functionally relevant parts of protein-coding genes and other genetic loci constraints the evolution of new genes. If natural selection were to be the only factor modeling evolution in the genomes it might have taken longer for complex organisms to arise, if at all [Ohno, 1970].

As the average mammalian genome mutation rate is 2.2 x 10-9 per base pair per year [Kumar and Subramanian, 2002] and surpassing the natural selection filter can have unaccounted and detrimental effects to the fitness of a species [Lynch and Conery, 2001, Kondrashov et al., 2002], evolution through mutations can be relatively slow in eukaryotes [Lynch, 1994]. Therefore, other evolutionary mechanisms have arisen in the genome by a combination of necessity and chance [Jacob, 1977] such as the duplication of genetic material, exon shuffling, retroposition, lateral gene transfer, transposable elements, gene fussion/fissions, *de novo* gene origination and often times combinations of all [Long et al., 2003].

These mechanisms have generated new genotypes and phenotypes, which later in time have been either maintained or removed by natural selection, depending of the fitness effect of the new genetic in-

formation in the population [Innan and Kondrashov, 2010a].

Therefore, there are two unresolved controversies revolving the evolution of new genotypes. First, whether natural selection or drift have played a mayor role in evolution. Second, whether single point mutations in protein-coding genes or other types of changes, such as duplications, were the drivers of phenotypic changes. These mechanisms are often times interlinked, as one of the main hypothesis for gene duplication is that the initial evolution is neutral.

Redundancy of genetic material (i.e duplication), is known to be the main source of novelties in the genome [Ohno, 1970]. Nonetheless, the most likely scenario regarding the fate of a duplicated gene in the genome is non-functionalization or gene loss (**Figure 5**) because the effects an additional copy can have in the genome are usually deleterious and therefore the copy will be removed by natural selection [Lynch, 2000, Innan and Kondrashov, 2010a]. However, if the duplicated gene or genetic sequence is maintained in the genome its fate can undergo different scenarios after fixation **(Figure 6)**, such as conservation (i.e, the copy has no detrimental effect), subfunctionalization (i.e the gene copies become complementary), specialization (i.e. the gene copies diversify into different tissues) or neofunctionalization (i.e. the gene copy gains a complete different function)[Force et al., 1997, Lynch, 2000, Long et al.,

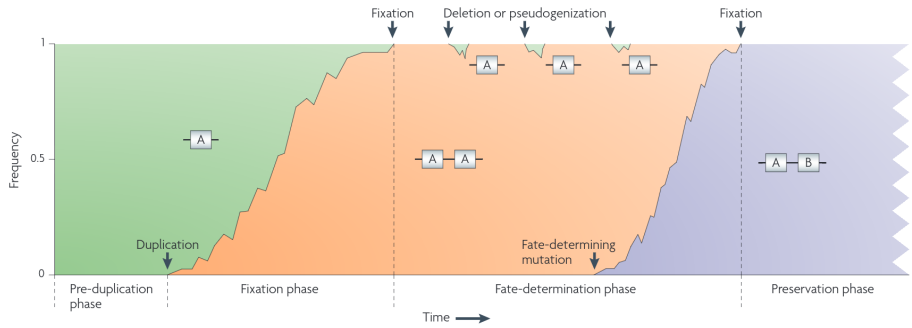

Figure 5: Phases leading to the stable preservation of a duplicated gene. After duplication, the copy is most likely to be lost to drift but can also achieve fixation. Once the genotype (A-A) is fixed, the fate-determination phase begins and continues until the fixation of a fate-determining mutation until reaching the preservation phase where the two copies are maintained by selection. Note that this figure shows the fixation and fate-determination phases separately; however, the two phases can overlap when a fate-determining mutation arises before the fixation of the duplicated copy or if the pre-existing allele works as a fate-determining mutation. If the fixation and fate-determination phases overlap, multiple selective forces can operate simultaneously. *Image and legend adapted from:*[Innan and Kondrashov, 2010b]

2003, Kaessmann, 2010]. In particular, partial duplications rather than full gene duplications play a key role in the evolution of new genes with phenotypic effects different than those of their parental counterparts[Innan and Kondrashov, 2010b, Long et al., 2003]. There are well-known cases in *Drosophila*, such as *jingwei*, the first *young gene* described to have originated from duplicated gene parts and translated into a chimeric protein with different expression than its parental gene *Adh* [Long and Langley, 1993], as well as *sphinx*, a

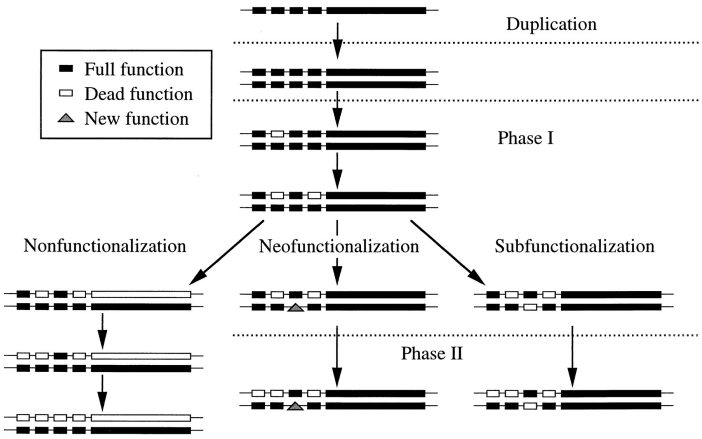

Figure 6: Fate of gene duplications with multiple regulatory regions. In the first two steps, one of the copies acquires null mutations. On the left, the next fixed mutation results in the absence of a functional product from the parental copy and becomes a non-functional pseudogene that will accumulate degenerative mutations over time. On the right, the copy acquires a null mutation in a regulatory region that is intact in the parental copy. Because both copies are now essential for complete gene expression, this third mutational event permanently preserves both members of the gene pair from future non-functionalization, however, may still eventually acquire a null mutation in one copy or the other. In the center, a regulatory region acquires a new function that preserves that copy, if it is beneficial, the duplicate copy is preserved *Image and legend adapted from:* [Force et al., 1997]

chimeric RNA gene involved in the courtship behaviour of the fruit-fly [Wang and Brunet, 2002]. In primates, *FOXP2* and *Morpheus* are also known genes that originated through duplication and have roles in speech and immune responses, respectively. There are many other genes that have originated through duplication in other organisms like rodents (*4.5Si RNA, BC1 RNA*), fishes (*Arctic AFGP, Antartic*

AFPG), plants (*Cytochrome C1, Sanguinaria rps1* and protozoans (*N-acetylneuraminate lyase*) therefore it is an interesting and compelling process to study.

1.3.1 Segmental duplications in humans and other primates

Segmental duplications (SDs) are defined as duplications of DNA sequences larger than 10 kb in the genome. Comparative genomic studies made between human and chimpanzee have shown that 33% of the segmental duplications (>94% identity) found in human are absent in the chimpanzee[Marques-Bonet et al., 2009]. Moreover, duplications have been attributed as having a greater impact (2.7%) in the genomic landscape differences between these two species than single-base pair substitutions (1.2%)[Marques-Bonet et al., 2009]. These conclusions among others [Cheng et al., 2005, Dennis et al., 2012], indicate that duplications have had a significant impact in the phenotypical differences between humans and chimpanzees.

1.3.2 Open questions about long non-coding RNAs and duplications

Despite humans and chimpanzees having 98.5% of protein-coding genome identity, we are still very different. The efforts of many scien-

tists around the world have focused on trying to identify regions that are relevant to our unique features. However, even though lncRNAs have been extensively studied in the last decade[Ulitsky, 2016, Necsulea et al., 2014, Rinn et al., 2007, Mattick, 2011, Wang et al., 2015] not much attention has been given to their possible contribution to the evolution of new phenotypical features in humans through duplications. It still an open question whether due to their high plasticity and high turnover rate lncRNAs are truly contributing to the differences we observe between humans and our closest living relatives.

Chapter 2

Objectives

1. Determine the contribution of lncRNA exon duplication to the evolution of the human genome, and their rate of duplication.
2. Correlation between alternative splicing and lncRNA exon duplication to gain insights about their connection.
3. Find unique functional elements that are responsible for the phenotypical and behavioral traits of our species.
4. Determine whether there is a correlation between the number of tRNAs of a species and the evolution of new tRNA identities.
5. *De novo* annotate lncRNA genes in a non-model endangered species; the Spoon-billed Sandpiper.

Chapter 3

Evolution of lncRNAs in the human lineage

Bello C & Kondrashov FA. Evolution of human-specific lncRNAs through exon duplication (under review)

3.1 Abstract

Whole and partial gene duplications play key roles in the evolution of novel genes and generation of new phenotypes. A large portion of the human genome is enriched in segmental duplications that are absent in other primates. However, the contribution of long non-coding RNA (lncRNA) duplications in human evolution remains unclear. Here,

we systematically addressed the rate and impact of lncRNA exon duplication in the human genome. We found that 11% of lncRNA exons had at least one highly similar copy in the genome and were significantly prevalent in alternatively spliced lncRNAs. Analysis of promoter single-nucleotide polymorphisms (SNPs) in flanking regions of lncRNAs showed evolutionary constraint indicative of a functional role of recent lncRNAs. Furthermore, the over-representation of specific classes of transposable elements (TEs) in exon flanking regions suggest a mechanism for the emergence and regulation of these genes. By integrating expression data and comparing primate genomes we identified 62 human-specific lncRNA genes that recently emerged through exon duplication, half of which were fixed in the human population. Some of these genes displayed tissue-specific expression patterns, including the brain. Overall, these results contribute to our understanding of the genomic events that have shaped the evolution of the human genome and prompt future studies of copy number variation in lncRNAs and their effects in disease and genome evolution.

3.2 Introduction

The acquisition of novel genetic elements through duplication, either by small-scale duplication or whole-genome duplication, is considered to be the main mechanism for the emergence of new genes and func-

tions [1]. When the duplicated genetic material is redundant, i.e. does not contribute to functional novelty or provides an immediate selective advantage, most of the initial gene duplications are removed through the accumulation of neutral mutations [2]. However, some duplicated genes can gain a new function, which can then be maintained and modified by selection [1,3,4]. Most empirical studies of the fates of gene duplications focus on instances of duplication of an entire protein coding gene [3]. A few studies that analyze the prevalence of duplication of non-coding RNAs have found that microRNAs (miRNAs) predominantly originated from inverted duplication of their targeted genes or by the duplication of a miRNA gene that diverged at its binding site gaining a different regulatory function [5,6]. Another class of non-coding RNA elements, piwi-interacting RNA (piRNAs), are also known to evolve through rapid turnover of gene copies [7]. Nevertheless, the duplication of non-coding sequences of long non-coding RNAs (lncRNAs) and their contribution to the origin of new genes remains generally unexplored.

A sizable fraction, 81%, of lncRNA genes are primate-specific [8,10], suggesting that their evolution may be relevant for the evolution of primate-specific features. However, the study of lncRNA evolution and their duplication effects is hampered by their lack of non-synonymous and synonymous sites, a comparison which is frequently

used in measuring the strength of selection acting on the gene [11], and by their poor functional characterization [12,13]. Nonetheless, it is known that non-coding genes have a rapid turnover rate in the genome and that they evolve faster than protein coding genes [14,15], which makes lncRNAs potential candidates for theevolution of new phenotypical traits.

The duplication of incomplete gene regions, such as exon duplications [16-18], is of particular interest because the mechanisms of their evolution are expected to differ from that of complete genes [16]. The basis for the difference is that a partial duplication may not be creating genetic redundancy, which is the principle for many of the evolutionary models of gene duplication [3]. Moreover, genomic duplications of >10 kb, known as segmental duplications (SDs), have been implicated in human evolution, with one third of SDs being absent in the chimpanzee genome [19,20]. These SDs may have contributed to the phenotypical differences between these species, with regulatory elements that modulate gene expression playing a key role in their evolution. Here, we focus on lncRNAs in the human genome for two reasons, the perceived importance of duplication in human evolution [21,22] and the relative functional obscurity of lncRNAs [23,24]. Finally, we focus on incomplete lncRNA duplications because of the lack of genetic redundancy when only a single exon is duplicated. By

studying lncRNA exon duplication in the human lineage and other great apes we aim to understand how these genes contributed to the evolution of the human genome.

3.3 Results

lncRNA exons were frequently duplicated in the human genome

To determine the contribution of exon duplication to recent evolution of lncRNAs we performed sequence similarity searches of exons coded in different lncRNA genes. Briefly, we used BLASTN [25] to compare sequences of individual annotated exons to the entire human genome sequence. Our approach limited one of the exon copies to be present in an annotated lncRNA gene. The number of lncRNA exon duplications declined with the sequence identity between the copies (**Fig 1A**), indicating either a burst of recent duplications, the removal of duplications from the genome by selection or the action of gene conversion [26,27]. Likewise, the high sequence identity of intronic sequences between the copies (**Fig 1B**), suggests that the copies appeared recently or that lncRNAs were subject to frequent gene conversion that overlapped introns. We focused on exon duplications that showed high sequence similarity between the copies (>90% sequence identity), which allowed us to study the majority of

identified duplication events while reducing the likelihood of missing duplications due to sequence divergence. To verify our results on the number of exon duplications in the human genome, we performed the blasted the human exons on other available great ape genomes; the chimpanzee, gorilla and the orangutan. We found that the distribution of the number of duplicated lncRNA exons followed an exponential decline (**Fig 1C**), with 10% of the lncRNA genes carrying at least one duplicated exon and 11% of the lncRNA exons having at least one other copy with >90% sequence identity in the human genome (**Fig 1D**). (which was expected due to the quality of the genome ssembly of other great apes. For further analysis, we focused on 562 genes harboring 877 exons with two copies in the genome (**S1 Datasheet**) because they represented half of the total instances of exon duplication **(Fig 1C,D)** and tracing the evolutionary history of multiple duplications accurately is often non-tangible.

lncRNA-lncRNA exon duplication was predominantly in antisense orientation, partial and in alternatively spliced genes

We considered whether the second exon copy was a part of an annotated non-coding or coding gene. Only 25% (211/877) of all second exons were found in another lncRNA gene, 25% (213/877) were found in a protein coding gene with the remaining half found in annotated pseudogenes (99/877), introns (77/877), and intergenic regions

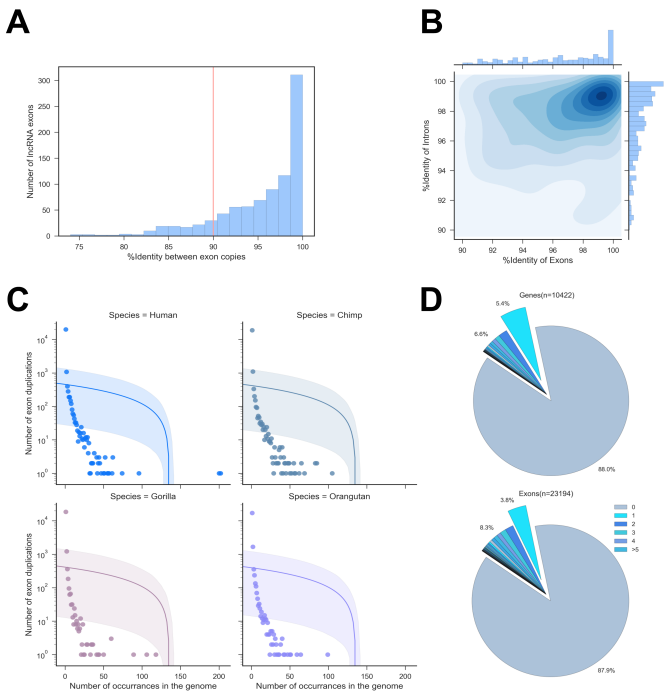

Figure 1: Duplication of lncRNA exons in the human genome and other great apes. **(A)** A histogram of the number of duplicated regions as a function of their sequence similarity. **(B)** The distribution of sequence similarity of duplicated exons and introns. **(C)** Frequency of occurrence of lncRNA exon duplications in the human, chimpanzee, gorilla and orangutan. **(D)** The distribution of the number of lncRNA genes in the human genome with different number of duplicated exons; zero duplications shown in grey with increasing number of duplications in opposite clockwise direction in the circle. **(E)** Same as **(D)**, but as a function of the number of copies of lncRNA exons.

(354/877) (Fig 2A). We found that only 12% of exon duplications could be attributed to whole gene duplication with the remaining 88% representing cases of partial gene duplications, regardless of whether

or not the recipient gene is also a lncRNA gene, a protein-coding gene or another genomic element (**Fig 2A**).

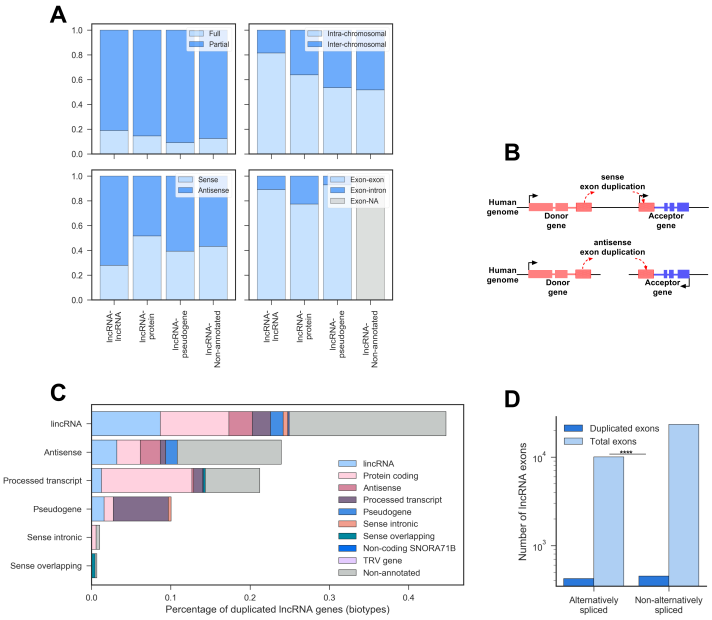

Figure 2: Exon duplications in the human genome. The number of duplicated exons as a function of different functional classes of the second exon copy. **(A)** The number of full and partial exon duplications.The number of inter- and intra-chromosomal exon duplications. The number of duplicated exons with the second copy annotated as an exon or intron. Sense and antisense duplications where the sense of the strand is relative to the transcribed annotated gene and shown in **(B)**, where the arrow shows the orientation of transcription. **(C)** Duplication patterns of different biotypes of lncRNA genes to other genomic regions. **(D)** Number of exon duplications of alternatively and constitutively spliced lncRNA exons; p-value < 0.0001 by Fisher's exact test between the two groups.

Chromosomes Y, X, 22 and 9 had the most instances of

lncRNA exon duplication (**S1 Fig**). Intra-chromosomal duplications occurred more frequently, a trend that was especially prevalent among duplications between two lncRNA genes (**Fig 2A**). The second emergent copy of many of the duplicated lncRNA exons was also annotated as an exon, however, most exons were duplicated into non-annotated regions, whereas only a minor fraction of exon duplications landed into introns of lncRNA and protein-coding genes (**Fig 2A**). Duplications with the initial lncRNA exon located in a sex chromosome were more likely to have the second copy also in a sex chromosome (**S2 Fig, S1 Datasheet**). Interestingly, the two exon copies were often found in antisense orientation when the second exon copy was found in another lncRNA gene (**Fig 2A,B**). However, the likelihood of observing sense or antisense orientation was approximately equal when the second exon was found in a protein coding gene(**Fig 2A**). The difference in the resulting orientation of lncRNA-lncRNA exon duplication relative to lncRNA-protein coding duplication suggests a difference in the functional mechanism of lncRNA-derived exons in these two classes of genes. LincRNAs (long intergenic long non-coding RNAs) were the lncRNA biotype with more instances of duplication in either a protein or another lincRNA, although about half of the duplications landed in a non-annotated region (Đś). Processed transcripts were the lncRNA biotype with more instances of duplications in protein coding genes and pseudogenes, whereas anti-

sense lncRNAs were equally likely to have an exon duplication in all of the other types of lncRNAs (**Fig 2C**).

Alternatively spliced exons may be related to exon duplication events [18,28]. Therefore, we analyzed the propensity of exon duplication events between lncRNA genes in exons known to be alternatively spliced versus constitutive exons, those in which alternative splicing has not been reported. We found that alternative exons were four times more likely to be duplicated than constitutive exons (**Fig 2D**), suggesting that exon duplication may be mechanistically connected with alternative splicing [18].

a-satellites, LTRs and other TEs were enriched in flanking regions of lncRNAs harboring duplicated exons

Transposable elements (TEs) are thought to contribute substantially to the evolution of lncRNAs [29-31]. Thus, we considered the contribution of different TEs to the sequence of recently duplicated lncRNA exons. First, we determined the frequency of various TEs and repetitive elements (REs) in duplicated lncRNA exons compared with non-duplicated lncRNA exons (**S3 Fig**). Although we used a cutoff that restricted the percentage of repeats (less than 20%, see methods) allowed in an exon to be considered a duplicated exon and not a RE, and therefore exists a bias against duplicated exons with high content of repetitive elements, we found that our set of recently du-

plicated lncRNA exons were specifically enriched in *a*-satellite DNA and low complexity sequences (**S3 Fig**), possibly reflecting a regulatory function [30]. Among other notable differences we found that our set of duplicated exons were enriched for gypsy LTRs [32] and ERV1 elements while LINE2, MIR and simple repeats were avoided (**S3 Fig**).

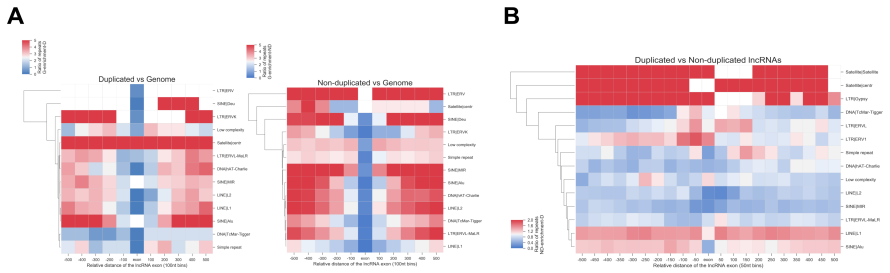

Figure 3: Repetitive elements in lncRNA duplication. **(A)** The ratios of REs in duplicated and non-duplicated lncRNA exons with their flanking regions compared to the genome. Values greater than one show an enrichment of the RE **(B)** Ratio of REs between duplicated and non-duplicated exons and flanking regions (500bp up- and downstream, 50nt bins). Enrichment of different classes of REs in duplicated exons are shown in red, and enrichment of different classes of REs in non-duplicated lncRNAs are shown in blue, white color blocks indicate missing values or values in which one of the groups was zero.

Second, we compared the prevalence of TEs and REs in lncRNA exons and their flanking regions relative to the human genome. Most TEs and REs, with the exception of *a*-satellite DNA, were found to be less prevalent in lncRNA sequences than their flanking regions

(**Fig 3A**), consistent with a possible role of duplicated lncRNAs in heterochromatin establishment [33]. Finally, we compared the prevalence of TEs and REs in the flanking regions of duplicated versus non-duplicated lncRNA exons. We found that the prevalence of specific TEs and REs within lncRNA exons generally extended to their flanking region, with the exception of Alu sequences, which were avoided in duplicated exons but were slightly more frequent in the duplicated exon flanking regions (**Fig 3B**). The prevalence of LINE1 and Alu elements in the exons and flanking regions of recently duplicated lncRNA exons suggest that their active transposition in the genome may be related to the emergence of the novel lncRNA exon copies [34,35].

The duplication rate of lncRNA exons in great apes was relatively constant over time, leading to human-specific genes

To trace the evolutionary history of the exon duplications we performed sequence searches of the duplicated regions in the genomes of all of the other great apes. We predicted the relative timeframe for the emergence of the exon duplication by comparing the number of detected copies in the great ape genomes. Overall, the rate of exon duplication appeared to be relatively constant across the great ape evolution (**Fig 4A**). However, a higher propensity of exon duplication in the orangutan and gorilla lineages was observed (**Fig 4A**),

congruent with previous observations that these lineages experienced a faster rate of segmental duplication [22]. Two exon copies in the human genome could be consistent with a recent ancestor having only one exon and the second exon emerging in a recent duplication event. Alternatively, a recent ancestor may have had several exons, with the human showing only two exons due to a recent exon loss in the human lineage. Therefore, cases of outgroup species having more than two exon copies could reflect instances of recent exon loss in human. Moreover, the observation of only one exon copy in an outgroup species provided information on the timeframe of the duplication event.

Overall, 209 new exons appeared in the human lineage since the divergence from orangutan (**Fig 4A**). For our analysis, we removed instances of apparent exon loss in the great ape lineage, where the human genome had two copies and the other great apes had more. Furthermore, we removed 89 instances of exon duplication where the human genome contained two copies and the other great apes contained zero; most of such cases were from the Y chromosome, possibly reflecting either difficulties in the assembly of the repetitive Y chromosome in other great apes or novel exons in the human genome [36]. To verify the human-specific exon duplications, we mapped the WGS reads of each primate against the human genome

and resolved 84 high-confidence human-specific exon duplications included in 62 lncRNA genes, such as the heart tissue-associated transcript 17 (HRAT17), the breast cancer associated lincRNA CYTOR (LINC00152) [37], CROCC2P, and others (**S2 Datasheet**).

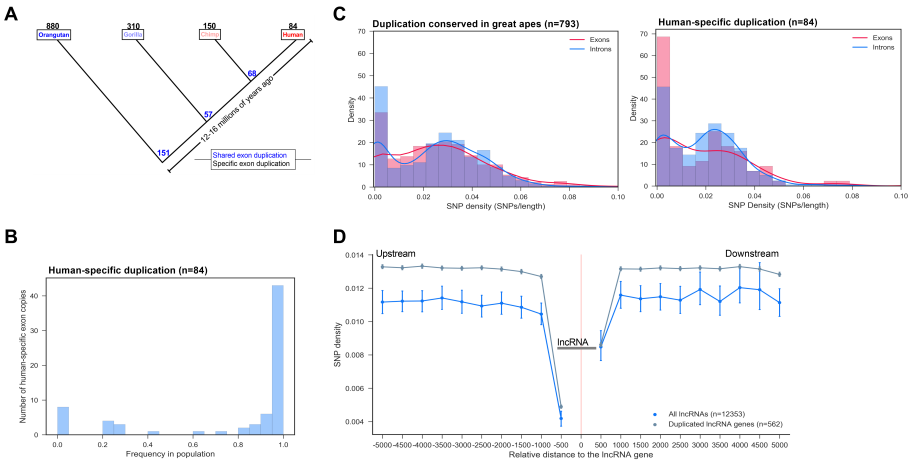

Figure 4: Evolution and frequency of exon duplications across great apes and in human populations. **(A)** Distribution of exon duplications across the great ape phylogenetic tree. The terminal numbers (black) indicate the number of identified two-copy exon duplications, while the numbers on the internal section of the tree (blue) indicate the estimated number of duplications occurring in the corresponding section of the phylogeny. **(B)** Frequency of human-specific duplicated exons in the human population; frequency values >0.6 indicate the exon was fixed in the population. **(C)** The density of SNPs in exons and introns of lncRNAs shown for exon duplications conserved in great apes and human-specific genes. The red and blue lines represent the kernel density estimate (KDE) of the lncRNA exons and introns. **(D)** The density of SNPs in the flanking regions of lncRNAs genes. A significant decrease in SNP density while approaching the promoter regions of the lncRNAs is shown; p-value < 0.0001 by Fisher's exact test.

Half of the human-specific duplicated exons were fixed in the human population

Human-specific exon duplications could be polymorphic in the human population. Therefore, to study the frequency of these duplications we focused on 22 available whole human genome sequences with high coverage [38] and utilized the normalized depth of coverage of the duplicated region relative to the average coverage of the genome as a proxy for copy number. Approximately half of the 84 human-specific exon duplications, including those in CROCCP2, CYTOR and HRAT17, were estimated to be present in all 22 individuals (**Fig 4B**), suggesting that they were fixed in the human population and indicating a possible role for these duplicated lncRNA exons in the recent evolution of the human genome. However, about one quarter of the human-specific exon duplications in lncRNAs had a frequency below 50% in the human population (**Fig 4B**), which could indicate either a role in human to human genetic variation or that the evolutionary fate of the duplicated exon has not yet been established in the population.

lncRNA genes were under negative selection for transcription in the human population

We next examined whether lncRNAs were under selective constraint by analyzing single-nucleotide polymorphisms (SNPs) in the human

population. Typical assays of selective constraint focus on the rate of accumulation of substitutions or polymorphisms in the sequence of interest relative to a selectively neutral sequence [11], such as introns. A previous study has shown that human lncRNAs shared a similar level of polymorphism between exons and introns [39]. We compared the non-duplicated exonic regions of the 562 lncRNA genes harboring the 877 recently duplicated exons with their non-duplicated intronic regions. We excluded the duplicated regions of the lncRNA genes due to the possibility of bias in the SNP-calling of those regions. In line with the previous study, we found that overall there were no significant differences in the density of SNPs in the evaluated non-duplicated regions of exons and introns, however, human-specific lncRNA genes showed a slight reduction in the SNP density in exons compared to introns, suggesting these genes might be under negative selection (**Fig 4C**). Interestingly, we observed a significant decrease (p-value, <0.0001, Fisher's test) in the density of SNPs in the promoter regions of lncRNA genes with duplicated exons (**Fig 4D**), indicating that they were under negative selection for continued expression in the human genome. Furthermore, the frequency spectra of SNPs found in the promoter region, 500bp upstream of the lncRNA genes, was not significantly different from the frequency spectra of SNPs found in the flanking regions at distances of -1000bp, +500bp and +1000bp (**S4 Fig**).

Genes harboring lncRNA exon copies had different expression profiles across several human tissues

The expression profiles of all lncRNA genes with a recent exon duplication show substantial differences between genes that shared a duplicated exon (**S5 Fig**), suggesting a diversification of the expression of these genes. However, the differences between the expression profiles of the 62 human-specific lncRNA genes appeared less pronounced (**Fig 5A**), indicating less time of divergence between the copies. Whereas several of the lncRNAs with a human-specific exon duplication were expressed ubiquitously across tissues, we identified a few cases displaying tissue-specific expression patterns, including the previously characterized human-specific duplicated neuronal gene SRGAP2C [40], which was expressed in the cerebellum (**S6A Fig, S2 Datasheet**). Likewise, we found that CROCCP2 and PDXDC2P were expressed in the cerebellum and HRAT17 in the heart (**Fig 5B and S6A Fig**), suggesting a functional role in these tissues. Moreover, the cancer related lncRNA CYTOR [37], was selectively expressed in leukemic cell lines and transformed fibroblasts (**Fig 5B and S6B Fig**). Finally, LINC00893, previously known as gene W and linked to Hunter Syndrome [41,42], was selectively expressed in the pituitary gland (**Fig 5B and S6B Fig**). We also identified lncRNA genes that had an exon duplication conserved among

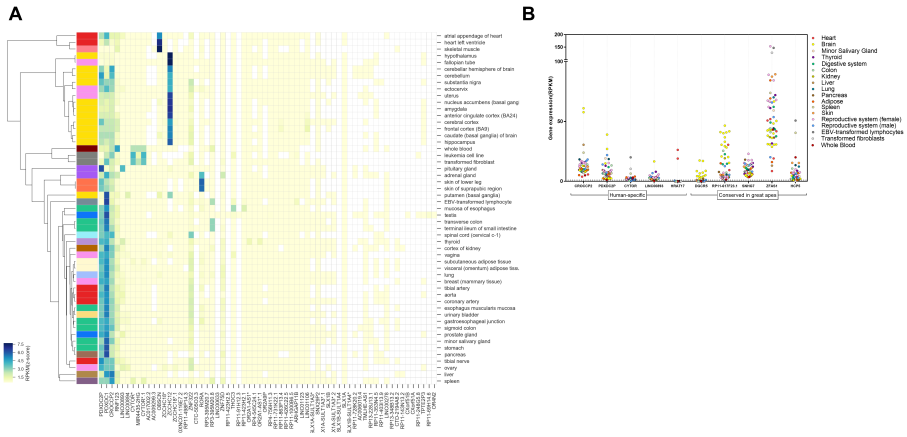

Figure 5: Expression patterns of lncRNAs harboring exon duplications across different human tissues. **(A)** Expression profiles of human-specific lncRNAs and their duplicated counterparts. Every two columns corresponds to two genes sharing a duplicated exon. The first gene of the pair is always a lncRNA, whereas the second gene could be a lncRNA, a protein coding gene or a pseudogene. The tissues have been hierarchically clustered (average linkage clustering) and the colored rows corresponds to specific tissue groups colored coded as seen in **(B)** Expression profiles of selected lncRNA genes displaying high expression in brain, heart, transformed lymphocytes and others.

all great apes, displaying tissue-specific expression patterns such as DGCR5, involved in DiGeorge Syndrome [43], and RP11-617F23.1 in several brain tissues and the oncogene ZFAS1 and breast cancer related SNHG7 in female reproductive tissues (**Fig 5B, and S6B Fig**). By utilizing a machine-learning classification framework to detect selective sweeps in human populations [44] we identified signs of positive selection in HRAT17 and CROCCP2 (**S7 Fig**), suggest-

ing that these human-specific exon duplications were relevant for the evolution of the human population.

Isoform expression preference was independent of the inclusion of a duplicated exon

Finally, we sought to ascertain whether inclusion of the duplicated exon correlated with the expression level of alternatively spliced lncRNA genes. Expression ratios between isoforms of individual genes did not reveal a general theme of expression level preference across tissues, with some genes showing preference for the isoform with and others without the duplicated exon (**S8A Fig**). However, isoforms without the duplicated exon did show a mild but significantly higher level of expression across all tissues (p-value <0.0001) (**S8B Fig**), particularly in the brain, consistent with previous studies showing that high expression level might restrict gene duplication [45].

3.4 Discussion

Duplication of genetic material is one of the staple trends of genome evolution [46], shaping genome architecture (Marques-Bonet et al. 2009), driving functional evolution [1] and possibly playing a role in ecological adaptation [47] and major evolutionary shifts [1,48,49]. Recent gene duplications are thought to have influenced recent hu-

man evolution [50,51], including the development of the human brain [40,52,53], and polymorphic copies are certain to have an impact on human pathologies [54]. Similarly, exon duplication has been implicated as a mechanism increasing diversity of protein function through the generation of novel isoforms [17,55]. Our data indicated that highly similar lncRNA exon duplications were more frequent in the great ape lineage than whole lncRNA gene duplications (**Fig 2A**). Interestingly, protein coding genes appear to show the opposite, with whole gene duplications appearing more frequently in the course of evolution [56,57] than exon duplication [16,17], mostly due to the contribution of whole genome duplication to the appearance of whole gene copies [56,57]. The fraction of exons ($\tilde{1}0\%$) emerging through exon duplication as detected by sequence similarity is similar between protein and lncRNA genes [16-18]. However, the likely stronger selection on maintaining protein sequences intact allows for the detection of older protein coding exon duplication events, suggesting that the rate of exon duplication may be higher in lncRNA genes than protein coding genes, with a relatively constant rate of evolution across the great ape lineage. The presence of active repetitive elements in the flanking regions of duplicated lncRNAs, in particular LINE1s and Alu repeats (Fig 3B), is consistent with a TE-driven mechanism of duplication linking these regions with natural genetic variation and disease [34,43,58,59]. Moreover, the overrepresentation of Îś-satellite

DNA (Fig 3B), the main component of centromeric and pericentric regions [60], indicates that recent lncRNAs could have a role in genome stability through interaction with other proteins [33,61]. Despite the general lack of selective constraint in lncRNA genes [39], we observed a significant decrease in SNP density in the promoter regions of lncRNAs indicative of negative selection (**Fig 4C**), suggesting that active transcription of lncRNA genes is important for the regulation and/or structure of the human genome. Some of the gene copies showed considerable divergence of expression after the duplication event (**S5 Fig**), suggesting either subfunctionalization or specialization of the original role of some of these genes after duplication and/or that they became a part in the landscape of transcriptional noise as raw material for evolution. Moreover, the identification of lncRNA genes with an exon duplication and a described phenotype, such as the oncogene ZFAS1 [62,63], the DiGeorge Syndrome associated DGCR5 gene [43,64] and the Hunter Syndrome associated LINC00893 gene [41] (**Fig 5B and S1 Datasheet**), suggests that some of the other uncharacterized lncRNAs genes with recently duplicated exons might be associated either with disease or development. Therefore, the identification of human-specific exon duplications in genes with specific expression patterns, such as CROCCP2 and HRAT17 , which showed recent signs of positive selection (**Fig 5**) and were fixed in human populations (**Fig 4B**) makes these genes and others (**S2 Datasheet**)

interesting candidates for follow-up functional studies that could contribute to our understanding of human evolution and disease. Taken together, the relatively rapid accumulation of lineage-specific exon duplications of lncRNAs, coupled with the evidence that the newly emerged copies were under selective constraint and fixed in human populations, suggests that structural changes of lncRNA genes may have contributed to recent great ape evolution.

3.5 Materials and methods

Identifying exon duplication events of lncRNAs in the human genome

To identify duplicated exons in the human genome, we constructed a non-redundant, non-overlapping exon dataset from 19,835 lncRNA transcripts of 12,235 lncRNA genes as annotated in GENCODE 13 [8]. Instances where two isoforms had different but overlapping exons were concatenated into a single exon sequence. A reciprocal BLASTN (version 2.2.28+) [25] using the created dataset of lncRNAs exons was performed using a cutoff of at least 70% identity and no less than 80% of the aligned length between the query and the subject, parameters that allowed the detection of old duplication events and to account for insertions and deletions. We filtered out exons with a high content of repetitive elements and kept those with $<$20% of

repetitive elements because we were interested in identifying exons and not repetitive elements throughout the genome. The resulting hits were classified as lncRNA exon duplications.

Characterization of lncRNA exons with two copies in the human genome

To describe the nature of the duplicated lncRNA exon in the human genome we divided those that only had one copy into different groups. We separated whole gene duplications from partial gene duplications by comparing the number of duplicated exons relative to the total number of exons in the lncRNA gene. If the total number of exons of the lncRNA gene was equal to the total number of exons that were duplicated for that gene, it was considered to be a whole-gene duplication, and if conversely was not, it was considered to be a partial gene duplication. The localization of the exon, or gene, copies as well as the correspondence of the copies to annotated genomic regions was performed using the same human genome assembly and annotation.

Searching for homologous exons in the genomes of great apes

We performed BLASTN (v2.2.28+) [25] with a cutoff >=70% identity and no less than 80% aligned length between query and subject, with an additional restriction for repetitive elements of less than 20%

presence in the sequence, using as query the exons of the human lncRNA genes from our non-redundant dataset and as the subject the genomes of Pantro troglodytes, Gorilla gorilla and Pongo abelii, respectively. In addition, the Batch Coordinate Conversion (liftOver) tool from the University of California, Santa Cruz [65] was used with a cutoff of 80% identity in conjunction with the BLASTN analysis to validate the homologous exons in great apes that corresponded with the duplicated exons of our human non-redundant exon dataset.

Genome mapping of great apes and coverage analysis

The Illumina Hi-Seq 2000 reads from chimpanzee (Clint and Bosco), gorilla (Banjo and Dian) and orangutan (Buschi and Babu) were downloaded from the Sequence Read Archive (SRA) (PRJNA189439, SRP018689) [66]. Paired-end reads of the non-human primates were trimmed and converted to Sanger format with FastX-Toolkit . The mapping to the reference human genome (GRCh37) was performed using BWA with the aln and sampe tools [67]. The increased edit distance parameter was n=0.04 (default). The presence of the duplicated exons in the genomes was estimated on the normalized coverage of the duplicated lncRNA exons of the sequences, including the non-human primates, mapped to the human genome.

Human-specific exon duplication as recent events

For each pair of exon duplicated genes we determined the percentage of identity of exons and introns separately using BLASTN (v2.2.28+) [25], and utilized the %identity as a proxy for the estimation of divergence between the human copies. Those that were found only in the human genome, and not in the genomes of other great apes, and had percentages of identity >=95% were considered to be human-specific candidates with the exception of 9 potential exons (S2 Table) that were included because they had support from BLASTN/liftOver and the coverage analysis.

SNPs analysis of human lncRNA exons

Based on the results of the BLASTN search, each lncRNA gene was divided into duplicated (DRs) and non-duplicated regions (non-DRs) for both exonic and intronic regions. The polymorphisms of each region for each of these regions was calculated separately using WGS data in combination with exome data from from 2,504 individuals corresponding to 26 populations from the 1000 Genomes Project Phase 3 [38]. We estimated the density of polymorphisms in the flanking regions of the lncRNA gene that had an exon duplication (5Kb upstream and downstream) in bins of 500 nucleotides using Tabix (TAB-delimited file IndeXer, v1.3) [68] and custom Perl scripts.

Allele frequency in lncRNA duplicated exons

We estimated the allele frequency of duplicated and non-duplicated lncRNA exons in human populations by using high-coverage data of exome-sequencing and whole genome sequencing (WGS) from 27 individuals of the 1000 Genomes Project by utilizing Tabix (v1.3) [68] and VCFtools (v0.1.11) [69], along with custom scripts to determine the minor allele frequency. Using the same approach we calculated the allele frequency of flanking regions (promoters) in bins of 500bp at distances of 5Kb up-and downstream the start of the lncRNA exon.

Copy number variation in human populations

To detect whether the lncRNA exon duplication has been fixed in human populations we estimated the copy number of the lncRNA exons in 22 individuals by utilizing high-coverage whole-genome-sequencing (WGS) data from the 1000 Genomes Project [38]. We estimated the copy number of each exon by using the depth of coverage approach [70]. The coverage of each lncRNA exon and its copy for each individual was determined using SAMtools (v1.3.1) [71] and normalized by the total depth of coverage of each respective human genome.

Expression of the genes harboring the lncRNA exon copies

We utilized the RNA-seq CAGE (cap Analysis of Gene Expression) data of 56 human tissues from the RIKEN FANTOM5 project (Study accession: DRP001031) [72] and analyzed the expression of lncRNA

genes that had a duplicated exon. Moreover, we compared the expression of both genes harboring the exon copies to evaluate whether the donor and acceptor had diverged expression in different tissues. Likewise, we evaluated the expression of the human-specific lncRNA genes.

Expression of different isoforms in alternatively spliced lncRNA genes The expression of different isoforms for the alternatively spliced lncRNA genes that contain a duplicated exon versus those that do not was determined by utilizing publicly available data on transcript expression of 53 tissues from 544 individuals from The Genotype-Tissue Expression project (GTEx Analysis V6, dbGaP Accession phs000424.v6.p1) [73,74]. First, we estimated the average expression per gene transcript per each tissue for the 8555 samples available. Second, we estimated the ratio of the sum of the expression value of the isoforms with a duplicated exon and those without it. We made a cutoff of expression at 0.5 RPKM to avoid outliers. Finally, we compared the mean of the expression level across all the tissues for isoforms with a duplicated exon versus the isoforms that do not have a duplicated exon by performing a paired t-test.

Availability of the data All the supporting data and materials are included in the article and scripts are available upon request.

3.6 Acknowledgements

We thank Guillaume Filion and Tomás Marqués-Bonet for extensive discussions.

Funding

The work was supported by HHMI International Early Career Scientist Program (55007424), the MINECO (BFU2012-31329, BFU2015-68723-P and BES-2013-064004), Spanish Ministry of Economy and Competitiveness Centro de Excelencia Severo Ochoa 2013-2017 grant (SEV-2012-0208), Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat's AGAUR program (2014 SGR 0974), and the European Research Council under the European Union's Seventh Framework Programme (FP7-2007-2013, ERC grant agreement 335980 EinME).

Author's contributions

1. Conceptualization: FK and CB. 2. Formal analysis:CB. 3. Funding acquisition:FK and CB. 4. Investigation:CB. 5. Project administration:FK and CB. 6. Resources:FK and CB. 7. Supervision:FK and CB. 8. Validation:CB. 9. Visualization:CB. 10. Writing - original draft:FK and CB. 11. Writing - review & editing:FK and CB.

3.7 References

1. Ohno S. Evolution by gene duplication. Berlin-Heidelberg-New-York: Springer-Verlag; 1970.

2. Lynch M. The Evolutionary Fate and Consequences of Duplicate Genes. Science (80-). 2000;290: 1151-1155. doi:10.1126/science.290.5494.1151

3. Innan H, Kondrashov F. The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet. Nature Publishing Group; 2010;11: 97-108.

4. Hahn MW. Distinguishing Among Evolutionary Models for the Maintenance of Gene Duplicates. J Hered. 2009;100: 605-617. doi:10.1093/jhered/esp047

5. Allen E, Xie Z, Gustafson AM, Sung G-H, Spatafora JW, Carrington JC. Evolution of microRNA genes by inverted duplication of target gene sequences in Arabidopsis thaliana. Nat Genet. 2004;36: 1282-1290. doi:10.1038/ng1478

6. Wang S, Adams KL. Duplicate Gene Divergence by Changes in MicroRNA Binding Sites in Arabidopsis and Brassica. Genome Biol Evol. 2015;7: 646-655. doi:10.1093/gbe/evv023

7. Assis R, Kondrashov AS. Rapid repetitive element-mediated expansion of piRNA clusters in mammalian evolution. Proc Natl Acad Sci U S A. 2009;106: 7079-82. doi:10.1073/pnas.0900523106

8. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, et al. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. Genome Res. 2012;22: 1775-89. doi:10.1101/gr.132159.111

9. Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, et al. The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature. 2014;505: 635-40. doi:10.1038/nature12943

10. Washietl S, Kellis M, Garber M. Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. Genome Res. 2014;24: 616-628. doi:10.1101/gr.165035.113

11. Li W-H. Molecular evolution. Sinauer Associates; 1997.

12. Mercer TR, Dinger ME, Mattick JS. Long non-coding RNAs: insights into functions. Nat Rev Genet. 2009;10: 155-9. doi:10.1038/nrg2521

13. Qureshi IA, Mattick JS, Mehler MF. Long non-coding RNAs in nervous system function and disease. Brain Res. 2010;1338: 20-35. doi:10.1016/j.brainres.2010.03.110

14. Ponting CP, Oliver PL, Reik W. Evolution and Functions of Long Noncoding RNAs. Cell. 2009;136: 629-641. doi:10.1016 /j.cell.2009.02.006

15. Ponting CP, Nellaker C, Meader S. Rapid Turnover of Functional Sequence in Human and Other Genomes. Annu Rev Genomics Hum Genet. 2011;12: 275-299. doi:10.1146/annurev-genom-090810-183115

16. Kondrashov F a, Koonin E V. Origin of alternative splicing by tandem exon duplication. Hum Mol Genet. 2001;10: 2661-9.

17. Letunic I, Copley RR, Bork P. Common exon duplication in animals and its role in alternative splicing. Hum Mol Genet. 2002;11: 1561-7.

18. Peng T, Li Y. Tandem exon duplication tends to propagate rather than to create de novo alternative splicing. Biochem Biophys Res Commun. 2009;383: 163-6. doi:10.1016/j.bbrc.2009.03.162

19. Jiang Z, Tang H, Ventura M, Cardone MF, Marques-Bonet T, She X, et al. Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. Nat Genet. 2007;39: 1361-1368. doi:10.1038/ng.2007.9

20. Marques-Bonet T, Kidd JM, Ventura M, Graves TA, Cheng Z,

Hillier LW, et al. A burst of segmental duplications in the genome of the African great ape ancestor. Nature. 2009;457: 877-881. doi:10.1038/nature07744

21. Marques-Bonet T, Girirajan S, Eichler EE. The origins and impact of primate segmental duplications. Trends Genet. 2009;25: 443-54. doi:10.1016/j.tig.2009.08.002

22. Lorente-Galdos B, Bleyhl J, Santpere G, Vives L, RamÃŋrez O, Hernandez J, et al. Accelerated exon evolution within primate segmental duplications. Genome Biol. 2013;14: R9. doi:10.1186/gb-2013-14-1-r9

23. Wheeler DA, Wang L, Dinger ME, Wheeler D, Wang L, Offit K, et al. From human genome to cancer genome: The first decade. Genome Res. BioMed Central; 2013;23: 1054-1062. doi:10.1101/gr.157602.113

24. Johnsson P, Lipovich L, Grandér D, Morris K V. Evolutionary conservation of long non-coding RNAs; sequence, structure, function. Biochim Biophys Acta. NIH Public Access; 2014;1840: 1063-71. doi:10.1016/j.bbagen.2013.10.035

25. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215: 403-410. doi:10.1016/S0022-2836(05)80360-2

26. Fawcett JA, Innan H. The role of gene conversion in preserving rearrangement hotspots in the human genome. Trends in Genetics. 2013. doi:10.1016/j.tig.2013.07.002

27. Innan H. Population genetic models of duplicated genes. Genetica. 2009;137: 19-37. doi:10.1007/s10709-009-9355-1

28. Abascal F, Tress ML, Valencia A. The Evolutionary Fate of Alternatively Spliced Homologous Exons after Gene Duplication. Genome Biol Evol. 2015;7: 1392-1403. doi:10.1093/gbe/evv076

29. Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, et al. Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs. Hoekstra HE, editor. PLoS Genet. 2013;9: e1003470. doi:10.1371/journal.pgen.1003470

30. Johnson R, Guigó R. The RIDL hypothesis: transposable elements as functional domains of long noncoding RNAs. RNA. Cold Spring Harbor Laboratory Press; 2014;20: 959-76. doi:10.1261/rna.044560.114

31. Kannan S, Chernikova D, Rogozin IB, Poliakov E, Managadze D, Koonin E V, et al. Transposable Element Insertions in Long Intergenic Non-Coding RNA Genes. Front Bioeng Biotechnol. Frontiers Media SA; 2015;3: 71. doi:10.3389/fbioe.2015.00071

32. Thompson PJ, Macfarlan TS, Lorincz MC. Long Terminal Repeats: From Parasitic Elements to Building Blocks of the Transcriptional Regulatory Repertoire. Mol Cell. 2016;62: 766-76. doi:10.1016/j.molcel.2016.03.029

33. Wang KC, Chang HY. Molecular Mechanisms of Long Noncoding RNAs. Mol Cell. 2011;43: 904-914. doi:10.1016/j.molcel.2011.08.018

34. Dewannieux M, Esnault C, Heidmann T. LINE-mediated retrotransposition of marked Alu sequences. Nat Genet. 2003;35: 41-48. doi:10.1038/ng1223

35. Hu S, Wang X, Shan G. Insertion of an Alu element in a lncRNA leads to primate-specific modulation of alternative splicing. Nat Struct Mol Biol. 2016;23: 1011-1019. doi:10.1038/nsmb.3302

36. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409: 860-921. doi:10.1038/35057062

37. Van Grembergen O, Bizet M, de Bony EJ, Calonne E, Putmans P, Brohee S, et al. Portraying breast cancers with long noncoding RNAs. Sci Adv. 2016;2: e1600220-e1600220. doi:10.1126/sciadv.1600220

38. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation.

Nature. Nature Research; 2015;526: 68-74. doi:10.1038/nature15393

39. Haerty W, Ponting CP. Mutations within lncRNAs are effectively selected against in fruitfly but not in human. Genome Biol. 2013;14: R49. doi:10.1186/gb-2013-14-5-r49

40. Dennis MY, Nuttle X, Sudmant PH, Antonacci F, Graves TA, Nefedov M, et al. Evolution of Human-Specific Neural SRGAP2 Genes by Incomplete Segmental Duplication. Cell. 2012;149: 912-922. doi:10.1016/j.cell.2012.03.033

41. Timms KM, Lu F, Shen Y, Pierson CA, Muzny DM, Gu Y, et al. 130 kb of DNA sequence reveals two new genes and a regional duplication distal to the human iduronate-2-sulfate sulfatase locus. Genome Res. 1995;5: 71-8.

42. Lagerstedt K, Carlberg B-M, Karimi-Nejad R, Kleijer WJ, Bondeson M-L. Analysis of a 43.6 kb deletion in a patient with Hunter syndrome (MPSII): Identification of a fusion transcript including sequences from the geneW and theIDS gene. Hum Mutat. 2000;15: 324-331.
doi:10.1002/(SICI)1098-1004(200004)15:4

43. Babcock M, Pavlicek A, Spiteri E, Kashork CD, Ioshikhes I, Shaffer LG, et al. Shuffling of Genes Within Low-Copy Repeats on 22q11 (LCR22) by Alu-Mediated Recombination Events During Evolution.

Genome Res. 2003;13: 2519-2532. doi:10.1101/gr.1549503

44. Pybus M, Luisi P, Dall'Olio GM, Uzkudun M, Laayouni H, Bertranpetit J, et al. Hierarchical boosting: a machine-learning framework to detect and classify hard selective sweeps in human populations. Bioinformatics. Oxford University Press; 2015;24: btv493. doi:10.1093/bioinformatics/btv493

45. Grishkevich V, Yanai I. Gene length and expression level shape genomic novelties. Genome Res. Cold Spring Harbor Laboratory Press; 2014;24: 1497-503. doi:10.1101/gr.169722.113

46. Magadum S, Banerjee U, Murugan P, Gangapur D, Ravikesavan R. Gene duplication as a major force in evolution. J Genet. 2013;92: 155-61.

47. Kondrashov FA. Gene duplication as a mechanism of genomic adaptation to a changing environment. Proc R Soc London B Biol Sci. 2012;279.

48. Glasauer SMK, Neuhauss SCF. Whole-genome duplication in teleost fishes and its evolutionary consequences. Mol Genet Genomics. 2014;289: 1045-1060. doi:10.1007/s00438-014-0889-2

49. Acharya D, Ghosh TC. Global analysis of human duplicated genes reveals the relative importance of whole-genome duplicates originated

in the early vertebrate evolution. BMC Genomics. 2016;17: 71. doi:10.1186/s12864-016-2392-0

50. Cheng Z, Ventura M, She X, Khaitovich P, Graves T, Osoegawa K, et al. A genome-wide comparison of recent chimpanzee and human segmental duplications. Nature. 2005;437: 88-93. doi:10.1038/nature04000

51. Ruiz-Orera J, Hernandez-Rodriguez J, Chiva C, Sabidó E, Kondova I, Bontrop R, et al. Origins of De Novo Genes in Human and Chimpanzee. Noonan J, editor. PLOS Genet. Public Library of Science; 2015;11: e1005721. doi:10.1371/journal.pgen.1005721

52. Sassa T. The Role of Human-Specific Gene Duplications During Brain Development and Evolution. J Neurogenet. 2013;27: 86-96. doi:10.3109/01677063.2013.789512

53. Charrier C, Joshi K, Coutinho-Budd J, Kim J-E, Lambert N, de Marchena J, et al. Inhibition of SRGAP2 function by its human-specific paralogs induces neoteny during spine maturation. Cell. 2012;149: 923-35. doi:10.1016/j.cell.2012.03.034

54. Kondrashov FA, Kondrashov AS. Role of selection in fixation of gene duplications. J Theor Biol. 2006;239: 141-151. doi:10.1016/j.jtbi.2005.08.033

55. Kondrashov FA, Koonin E V. Origin of alternative splicing by tandem exon duplication. Hum Mol Genet. 2001;10: 2661-9.

56. Blomme T, Vandepoele K, De Bodt S, Simillion C, Maere S, Van de Peer Y. The gain and loss of genes during 600 million years of vertebrate evolution. Genome Biol. 2006;7: R43. doi:10.1186/gb-2006-7-5-r43

57. Singh PP, Arora J, Isambert H, Peer Y Van de, Maere S, Meyer A, et al. Identification of Ohnolog Genes Originating from Whole Genome Duplication in Early Vertebrates, Based on Synteny Comparison across Multiple Genomes. Christos A, editor. PLOS Comput Biol. Public Library of Science; 2015;11: e1004394. doi:10.1371/journal.pcbi.1004394

58. Bailey JA, Eichler EE. Primate segmental duplications: crucibles of evolution, diversity and disease. Nat Rev Genet. Nature Publishing Group; 2006;7: 552-564. doi:10.1038/nrg1895

59. Liu G, NISC Comparative Sequencing Program, Zhao S, Bailey JA, Sahinalp SC, Alkan C, et al. Analysis of Primate Genomic Variation Reveals a Repeat-Driven Expansion of the Human Genome. Genome Res. 2003;13: 358-368. doi:10.1101/gr.923303

60. Aldrup-MacDonald ME, Kuo ME, Sullivan LL, Chew K, Sullivan BA. Genomic variation within alpha satellite DNA influences

centromere location on human chromosomes with metastable epialleles. Genome Res. 2016;26: 1301-1311. doi:10.1101/gr.206706.116

61. Lippman Z, Gendrel A-V, Black M, Vaughn MW, Dedhia N, Richard McCombie W, et al. Role of transposable elements in heterochromatin and epigenetic control. Nature. 2004;430: 471-476. doi:10.1038/nature02651

62. Askarian-Amiri ME, Crawford J, French JD, Smart CE, Smith MA, Clark MB, et al. SNORD-host RNA Zfas1 is a regulator of mammary development and a potential marker for breast cancer. RNA. 2011;17: 878-891. doi:10.1261/rna.2528811

63. Zhang Y, Sun L, Xuan L, Pan Z, Li K, Liu S, et al. Reciprocal Changes of Circulating Long Non-Coding RNAs ZFAS1 and CDR1AS Predict Acute Myocardial Infarction. Sci Rep. Nature Publishing Group; 2016;6: 22384. doi:10.1038/srep22384

64. Sutherland HF, Wadey R, McKie JM, Taylor C, Atif U, Johnstone KA, et al. Identification of a novel transcript disrupted by a balanced translocation associated with DiGeorge syndrome. Am J Hum Genet. 1996;59: 23-31.

65. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res. 2002;12: 996-1006. doi:10.1101/gr.229102. Article published online

before print in May 2002

66. Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, et al. Great ape genetic diversity and population history. Nature. 2013;499: 471-475. doi:10.1038/nature12228

67. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25: 1754-1760. doi:10.1093/bioinformatics/btp324

68. Li H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics. 2011;27: 718-719. doi:10.1093/bioinformatics/btq671

69. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. Oxford University Press; 2011;27: 2156-8. doi:10.1093/bioinformatics/btr330

70. Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. BMC Bioinformatics. 2013;14: S1. doi:10.1186/1471-2105-14-S11-S1

71. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinfor-

matics. 2009;25: 2078-2079. doi:10.1093/bioinformatics/btp352

72. Lizio M, Harshbarger J, Shimoji H, Severin J, Kasukawa T, Sahin S, et al. Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome Biol. 2015;16: 22. doi:10.1186/s13059-014-0560-6

73. Ardlie KG, Deluca DS, Segre A V., Sullivan TJ, Young TR, Gelfand ET, et al. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science (80-). 2015;348: 648-660. doi:10.1126/science.1262110

74. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. Nat Genet. Nature Research; 2013;45: 580-585. doi:10.1038/ng.2653

3.7.1 Supporting information

S1 Datasheet: List of lncRNA exons with two instances in the human genome with their characteristics of duplication and localization. (upon request until published) S2 Datasheet: List of candidate human-specific lncRNA exons that have two instances in the human genome, with their characteristics of duplication and localization (upon request until published).

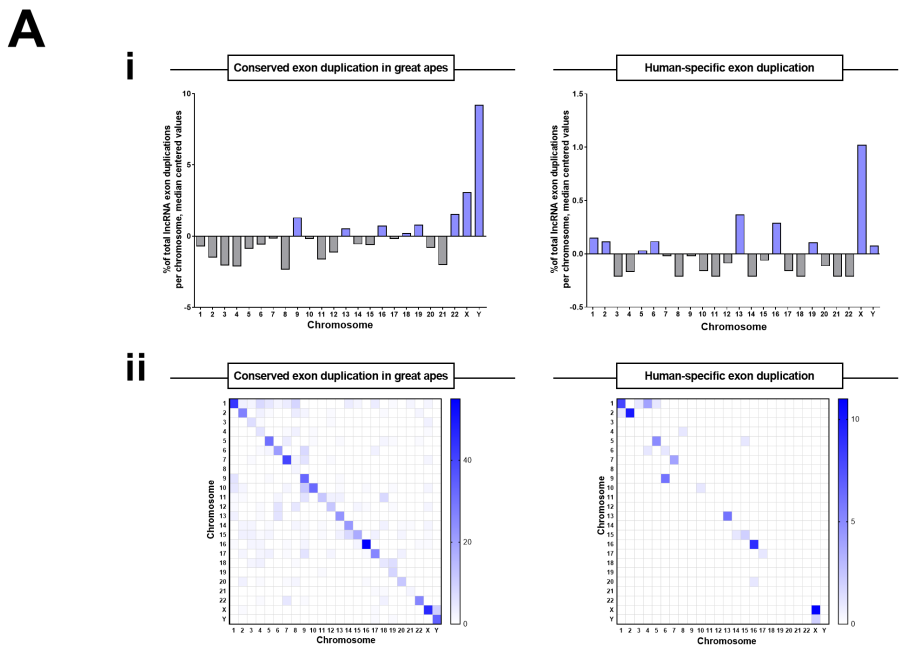

Figure 1: Localization and distribution of lncRNA exon duplications in human chromosomes. **(A)** The median centered value of the fraction of lncRNA two-copy exon duplications found across the human chromosomes for duplications conserved in great apes and human-specific. **(i)** Heatmap of the chromosomic distribution of the exon duplications conserved in great apes and human-specific **(ii)**.

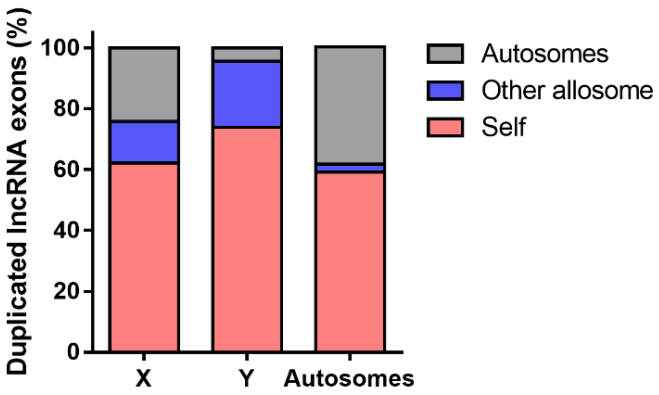

Figure 2: Frequency of duplications in the chromosomes. Sex chromosomes (allosomes) have a higher likelihood of having a duplication in another sex chromosome than autosomes to a sex chromosome. Chi-square test, Chromosome Y vs Autosome: p-value <0.0001; Chromosome X vs Autosomes: p-value <0.0001; Chromosome Y vs Chromosome X: p-value $= 0.0140$.

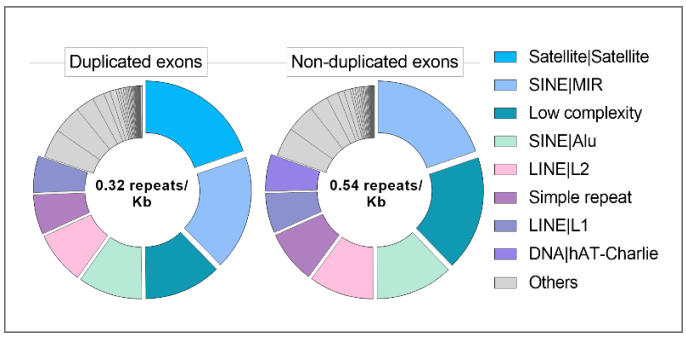

Figure 3: Frequency of repeats in duplicated and non-duplicated lncRNA exons.

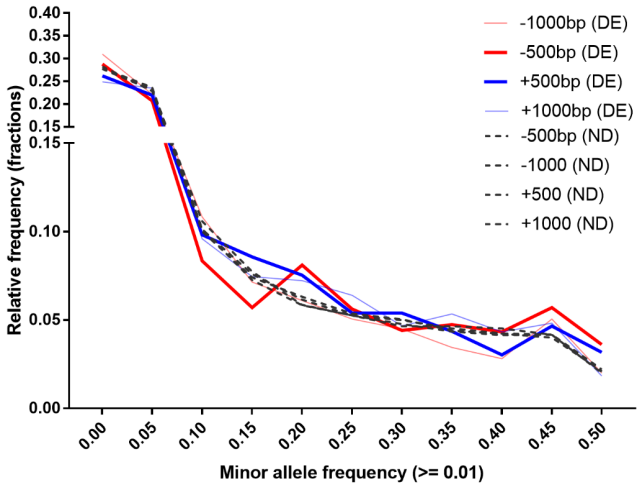

Figure 4: Frequency spectra of SNPs in -1000bp, -500bp, +500bp and +1000bp of lncRNA flanking regions, with data from duplicated lncRNA genes (blue) and data from all lncRNA genes (grey). Error bars indicate SEM. (E) Frequency of 84 human-specific lncRNA exon duplications in the human population from 22 high coverage human genomes.

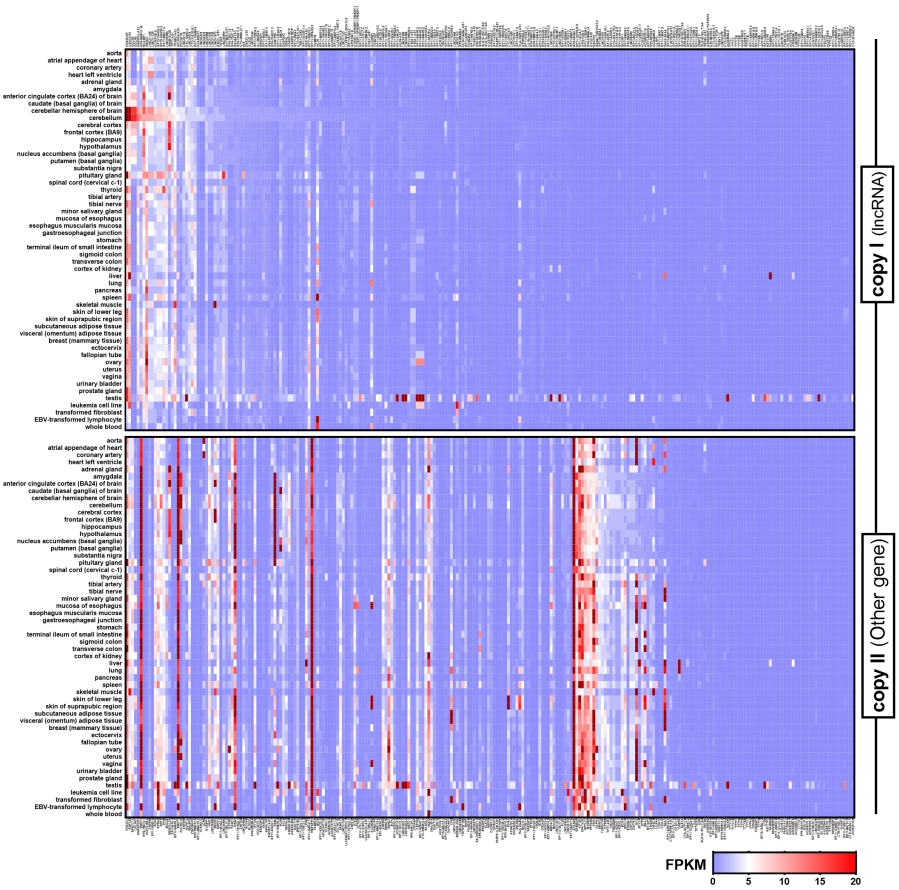

Figure 5: Expression across different human tissues. LncRNA genes sharing two-copy duplicated exons of all identified duplicated exons.

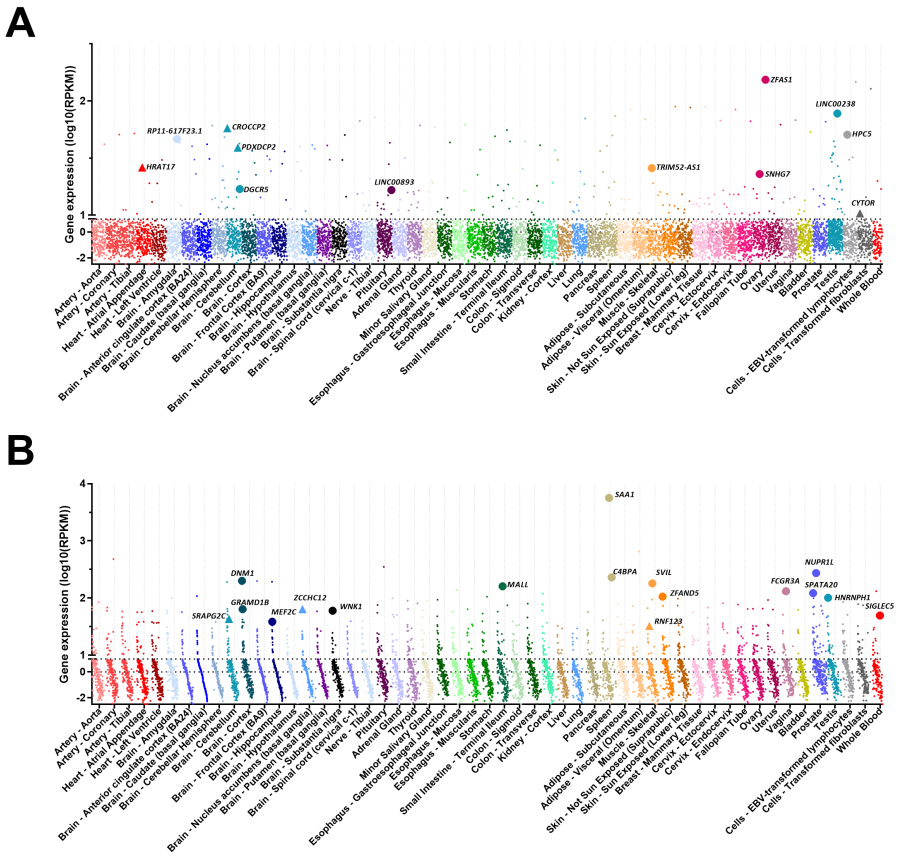

Figure 6: Expression of genes harboring an exon duplication (A) and their lncRNA counterparts (B). A few relevant cases with described phenotypes have been highlighted. Enlarged circles denote an exon duplication that was conserved among great apes, whereas an enlarged triangle denotes that the exon duplication was human-specific.

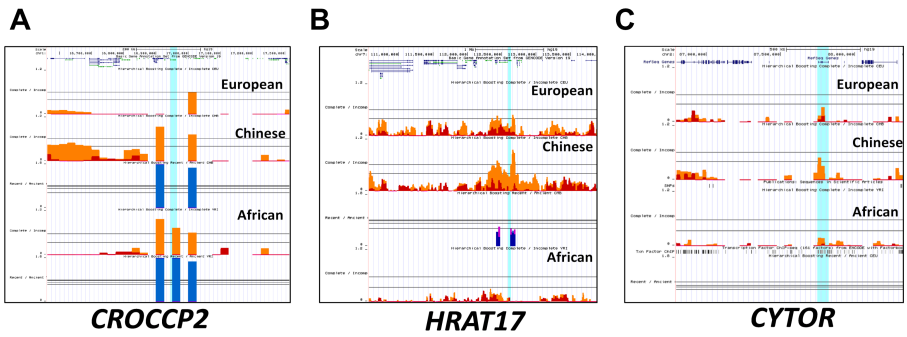

Figure 7: Examples of lncRNAs harboring human-specific exon duplications in different human populations presenting selective sweeps showing signs of positive selection. Orange and red bars show incomplete or complete selective sweeps respectively, whereas blue and pink indicate recent or ancient selective sweeps, respectively. (A) CROCCP2, (B) HRAT17 and (C) CYTOR (LINC00152).

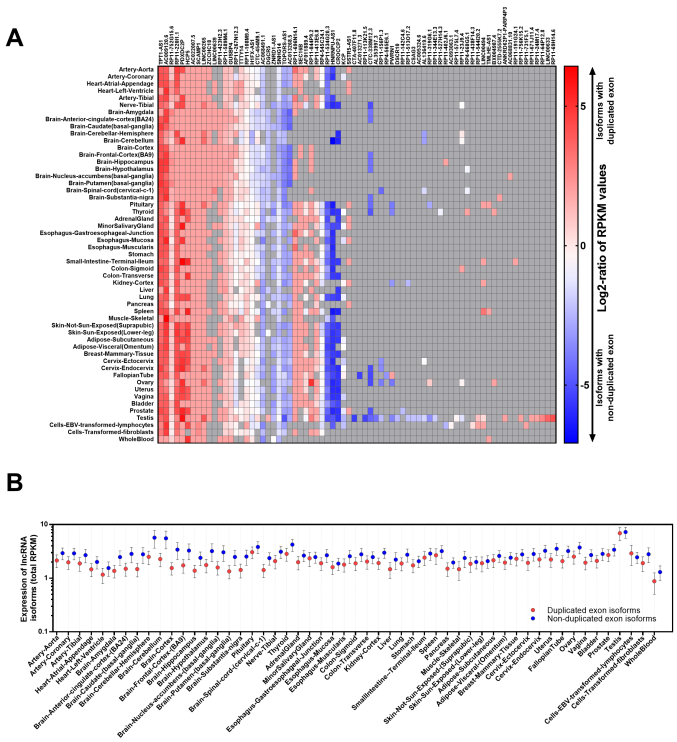

Figure 8: Differential expression levels of lncRNA isoforms with and without a duplicated exon across 53 human tissues. (A) Log2-ratio between the expression of different lncRNA isoforms with and without the duplicated exon for alternatively spliced (AS) genes. The isoforms with a duplicated exon and a higher level of expression are shown in red, isoforms without the duplicated exon and a higher level of expression are shown in blue. Grey indicates missing values or values below the cutoff of 0.5 RPKM. (B) The mean of the expression for all the isoforms of AS genes with a duplicated exon shown in red, versus the mean of the expression of isoforms of AS genes without a duplicated exon shown in blue. Isoforms with a duplicated exon have a significant trend for a lower level of expression than those without the duplicated exon in all tissues (paired t-test, p-value <0.0001). Error bars indicate SEM.

Chapter 4

Evolution of new tRNA identities

Saint-Léger A, Bello C, Dans PD, Torres AG, Novoa EM, Camacho N, et al. Saturation of recognition elements blocks evolution of new tRNA identities. Sci Adv. 2016 Apr 29;2(4):e1501860. DOI: 10.1126/sciadv.1501860

Chapter 5

The genome of the Spoon-billed Sandpiper

Mateusz Konczal, Luis Zapata, Francisco Camara, Anna Vlasova, ***Carla Bello****, Romain Derelle, Maria N. Tutukina, Maria Plyuscheva, Claudia Fontsere, Pavel S. Tomkovich, Nikolay N. Yakushev, Ivan A. Shepelev, Vladimir Yu. Arkhipov, Cristoph Zöckler, Roland Digby, Egor Y. Loktionov, Elena G. Lappo, Tomás Marqués, Roderic Guigó, Evgeny E. Syroechkovskiy, Fyodor A. Kondrashov.* ***Population genomics of the critically endangered Spoon-billed Sandpiper.*** *(in preparation)*

5.1 Abstract

Genetic factors are thought to contribute to the global decline of population size of plants and animals and among the risks to population recovery. However, comparative population genomic studies of species with declining and stable population sizes have not been performed. Here we compare the recent population genetic history of the critically endangered spoon-billed sandpiper to its sister species, the red-necked stint, until recently a species of least concern (red List). We found that spoon-billed sandpipers were most abundant 15,000-25,000 years ago during the last glacial maximum with greater availability of proper breeding habitat, and have been declining since. The red-necked stint numbers have been constant over the last 100,000 years providing a good basis for comparison. Despite 1000-fold difference in current population sizes (Red List) we found a similar level of nucleotide diversity, 1.5 x 10-3 in the spoon-billed sandpiper and 2.2 x 10-3 in the red-necked stint. However, the spoon-billed sandpiper harbors a substantially higher proportion of deleterious to neutral alleles, with 44% more nonsense and 35% more non-synonymous polymorphisms than the red-necked stint. Evidently, the prolonged decline of the spoon-billed sandpiper population over the course of $\tilde{5}000$ generations caused a reduction in efficacy of selection with enough time for the accumulation of novel delete-

rious polymorphisms. Our study suggests that species experiencing a rapid crash may have lower costs of inbreeding and higher chances of recovery compared to species with a decline of similar magnitude over a substantially longer time period. Specifically, this may affect species that, like the spoon-billed sandpiper, experienced substantial habitat loss since the last glacial maximum.

5.2 Introduction

Wildlife is experiencing a global decline in population size (Pimm et al. 2014, Tilman et al. 2017) possibly on a way to one of the massive extinction events in history (Barnosky et al. 2010). Birds are not an exception, especially migratory birds (Bairlein 2016) including populations of the East Asian Australia Flyway (EAAF). Birds along the EAAF undertake exceptionally long migration routes from the Arctic in Eastern Asia to Australia and New Zealand, with up to 90% of species with known population trends declining in number (Hua et al, 2015). A case in point is the critically endangered spoon-billed sandpiper (*Calidris pygmeus*), which breeds in the Arctic region of Chukotka and migrates along the EAAF to Southeast Asia for the winter (Figure 1). The spoon-billed sandpiper is one of the world's rarest species with an estimated 100-200 breeding pairs the wild (Zockler 2016 Bird Conservation International; Clark N, 2016

ORYX) continuing to decline in numbers (Syroechkovski, 2010 Biology Bulletin). Crucial factors contributing to the rapid recent decline of the spoon-billed sandpiper numbers include habitat loss along the migratory route and human predation (Zockler, 2010 , Wader Study Group Bulletin), which are common threat factors for many waders along the EAAF (Hua et al, 2015, Hebo, 2017, Bird Conservation International). By contrast, the genetic contribution and consequence of the decline of spoon-billed sandpiper population have not been considered. To investigate the genetic component of population decline we performed a large scale population genomic comparison of this of this flagship migratory species to its well-faring sister species, the red-necked stint (*Calidris ruficollis*).

Genetic characterization of fixed changes in the genome of a critically endangered species has the potential to reveal factors that defined its recent evolution and phenotype. Furthermore, the study of genetic variability within the population can provide insight on the population history (Schraiber and Akey 2015) and potential genetic risks, such as loss of heterozygosity and the accumulation of deleterious mutations (Spielman et al. 2004, Frankham 2005; Polishchuk 2015). The spoon-billed sandpiper is a good model species to address both of these questions, with its sister species, the red-necked stint, is abundant in the wild (citation), providing a crucial reference

for polymorphism and evolutionary analysis. Previous genome-wide analyses provided insights into population genetics of endangered species (Abscal 2016, Rogers 2017). However, these studies focused on species in which the natural variability was artificially influenced by hybridization, captive breeding or reintroduction programs and lacked a comparison with well fairing sister species. The analysis of the spoon-billed sandpiper provides an opportunity to study the population genomics of a rapidly declining population not influenced by ongoing conservation efforts, serving as a model for other endangered species.

5.3 Results and discussion

We selected samples collected in several locations of Chukotka from female individuals, including 10 spoon-billed sandpipers, nine red-necked stints and a single individual each of the red knot (*Calidris canutus*), long-toed stint (*Calidris subminuta*) and the little stint (*Calidris minuta*) species (**Figure 1**, Extended Data Table 1). We sequenced DNA extracted from one of the spoon-billed sandpipers in two pair-end libraries creating a *de novo* genome assembly. An extensive annotation effort using gene prediction software and mapping for RNA-seq data mapped the protein coding and lncRNA genes in the assembled genome (see Supplementary Methods). Overall, the

resulting genome annotation was of a similar quality to the reference chicken genome (citation), with 35x coverage, scaffold N50 = 2.8 Mb, 21,145 protein-coding genes and 5425 lncRNA transcripts (Extended Data Table 2) with 93% of the genes in the core eukaryotic genes set found in the assembled genome (Extended Data Table 3). To allow for a comparative PMSC analysis (Li and Durbin 2011) we sequenced one red-necked stint and the red knot samples to 30x coverage. The remaining 9 samples of the spoon-billed sandpiper and 8 red-necked stints were sequenced to 15x and 10x coverage, respectively. The long-toed stint and little stint were sequenced to 5x coverage. The sequences from these individuals were mapped to the reference spoon-billed sandpiper genome and identified single nucleotide polymorphisms (SNPs). We calculated the kinship coefficient between individuals of the same species, identifying two closely related pairs of spoon-billed sandpipers. No close relatives were found in the red-necked stint samples. Therefore, the final analysis included the comparison of SNPs from 7 spoon-billed sandpiper and 9 red-necked stint genomes, with $\tilde{5}$.2 million SNPs in the spoon-billed sandpiper and 9.2 million SNPs in the red-necked stint (**Table 1**).

Using the polymorphism data, we studied several aspects of the spoon-billed sandpiper population. First, we looked for evidence of population structure between individuals from different geographical

Table 1. Polymorphisms in the spoon-billed sandpiper and the red-necked stint populations.

	SNPs	SNPs (no singletons)	Indels	Indels (no singletons)
SBS	5,218,932	3,681,367	691,673	480,457
RNS	9,194,008	6,288,764	1,471,359	1,225,860

locations. We found no considerable genetic isolation between spoon-billed sandpipers from Meinopylgino and the Belyaka spit, locations 650 km apart at the edges of the modern breeding range, suggesting a lack of population structure (Fst =0.002, **Figure 1**). Second, we reconstructed the phylogeny of the sequenced *Calidris* species, verifying that the red-necked stint is the sister species to the spoon-billed sandpiper **(Figure 1A)**. Finally, we considered the possibility of genetic introgression between the spoon-billed sandpiper and the red-necked stint. We found no evidence of recent intergressions (Extended Data Figure VII.2), consistent with only a single possible hybrid over several decades of observation (Red'kin 2012).

The lack of detectable genetic introgression between the sister species allows for direct comparison of their species-specific variability. The level of nucleotide variability in the spoon-billed sandpiper population, π= 0.0022, was less than twofold smaller than in the red-necked stint population, π= 0.0015. **Figure 1A**). To compare population

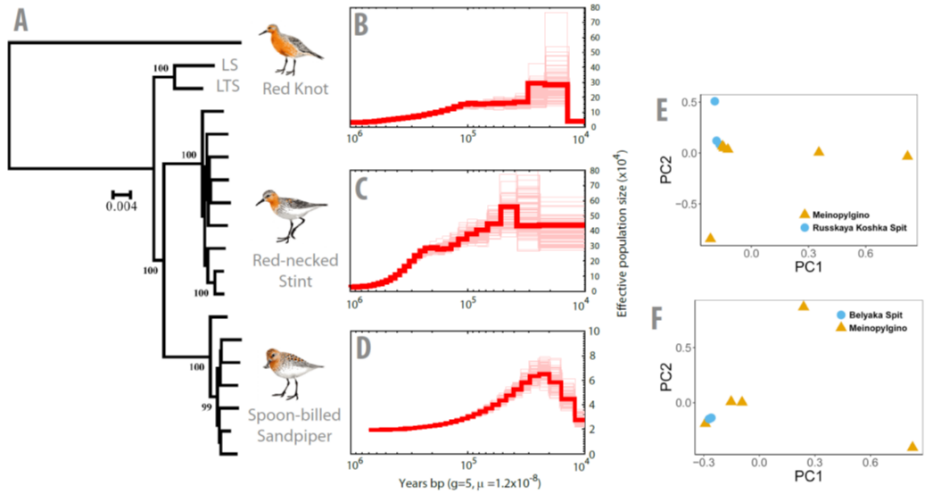

Figure 1: Polymorphisms in the spoon-billed sandpiper and the red-necked stint populations.

size dynamics of these species we performed a PMSC analysis (Li and Durbin 2011) using the three high-coverage genomes of spoon-billed sandpiper, red-necked stint and the red knot. The analysis reconstructs changes in population size derived from the expected coalescence times of the haplotype blocks in the genome. The population history of the red-necked stint was predicted to have been constant throughout the last 500,000 years (Figure 1C). By contrast, the smaller spoon-billed sandpiper population has been in continuous decline since 15-30 tya, roughly corresponding to the last glacial maximum (Clark 2009). The estimated recent decline is consistent with the relatively high levels of neutral variability. The red knot also ex-

perienced a sharp decline in population size (Piersma 2007), however, the recent establishment of red knot subpopulations (Buehler 2006) can also lead to a rapid change of haplotype structure in the population that may be interpreted by PMSC as a population decline. The latter scenario, however, is not applicable to the spoon-billed sandpiper population because it apparently lacks population structure.

For species with population size of several dozen breeding individuals, inbreeding and the concomitant loss of heterozygosity is a risk factor to long term survivability (Spielman, D., 2004; Frankham 2005; Polishchuk 2015). To characterize the degree of inbreeding in the spoon-billed sandpiper we estimated heterozygosity in 1 Mbp windows across the genome (**Figure 2**). The number of low heterozygosity stretches was higher in the spoon-billed sandpiper than the red-necked stint, however, we found no evidence of strong inbreeding, suggesting a panmictic population. Lack of long runs of homozygosity in the longest genome scaffold (Extended Data Figure III.1) is consistent with low levels of inbreeding in the spoon-billed sandpiper population. Although the decline of the spoon-billed sandpiper population did not have a strong impact on the overall level of variability, the efficacy of selection is reduced in such a declining population that can lead to the accumulation of deleterious alleles (Charlesworth 2009). We compared the strengths of selection act-

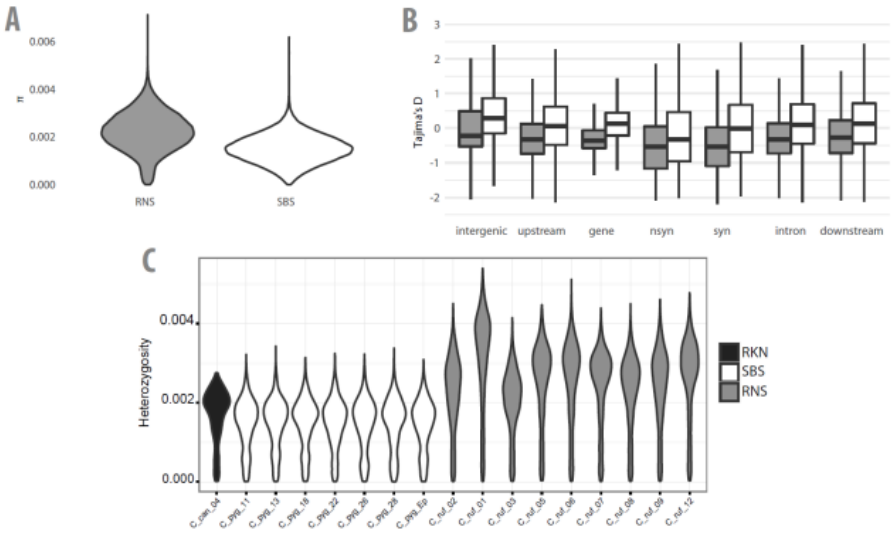

Figure 2: Patterns of genetic variation of red-necked stint (RNS) and spoon-billed sandpiper (SBS) populations. A: Distributions of nucleotide diversity values calculated in 50kb windows; B: TajimaâĂŹs D values calculated for functional classes of sites; C: Distributions of heterozygosity values calculated in 1 Mbp windows for red knot (RKN), SBS and RNS individuals.

ing in the spoon-billed sandpiper and red-necked stint population by calculating Tajima's D, a statistic that measures the relative contribution of polymorphisms of different frequencies to overall population variability (Tajima 1989), for several functional classes of sites. A negative value of Tajima's D can be caused by recent population growth or by negative selection acting against emerging polymorphisms, while a positive Tajima's D indicates the action of positive selection or population growth. The red-necked stint population has

been relatively constant over the last half a million years **(Figure 1C)**, thus, Tajima's D in that species would be mostly influenced by selection and not changes of population size. In the red-necked stint Tajima's D was negative for polymorphisms in coding regions and close to zero for polymorphisms in intergenic regions **(Figure 2B)**. These data indicate selective constraint in coding regions, while intergenic regions were mostly neutral, congruent with observations from other species (Andolfatto 2005). Tajima's D for polymorphisms in the spoon-billed sandpiper population was respectively higher for all functional classes (Figure 3B), with intergenic polymorphisms showing a substantially positive Tajima's D, an observation most consistent with a recent population decline.

We then studied the prevalence of slightly deleterious alleles in the two sister species. Conservatively assuming that intergenic polymorphisms are (mostly) neutral (**Figure 2**, Andolfatto 2005), we compared their prevalence to the prevalence of polymorphisms at other functional sites between the two species. A population bottleneck is expected to remove rare polymorphisms without affecting the density of common ones (Charlesworth 2009). Therefore, the spoon-billed sandpiper population was expected to have a lower ratio of rare deleterious to common neutral polymorphism. Surprisingly, we found a proportional increase of rare polymorphisms of all functional classes

in the spoon-billed sandpiper population, with the highest increase of nonsense and nonsynonymous sites (Table), which are expected to have the strongest effect on fitness and have the lowest frequency. Only synonymous sites had fewer polymorphisms than intergenic sites in the spoon-billed sandpiper genome.

These data can only be explained by considering two consequences of the considerable decline in effective population size of the spoon-billed sandpiper over thousands of generations (**Figure 1C**); genetic drift and relaxation of selection. Genetic drift in the declining population removed rare polymorphisms from the population (Supplementary Figure V.1), an event that over the short term influenced all polymorphisms, as is evidenced by higher Tajima's D across all types of sites (**Figure 2B**). At the same time, the reduction in the efficacy of selection increased the number of functional sites in which slightly deleterious polymorphisms could segregate, manifesting in a faster rate of deleterious polymorphisms coming into the population in the course of the population decline. This effect explains why proportional to neutral intergenic polymorphisms, nonsense and nonsynonymous polymorphisms are found at highest densities in the spoon billed sandpiper compared to the red-necked stint (**Table 2;** Supplementary Table 2). The relative lack of new synonymous variation in the spoon billed sandpiper is consistent with the action of

background selection (Charlesworth 1993) at these sites, which does not reach intron or intergenic sites due to lack of linkage disequilibrium between exons and other genomic regions (Supplementary Figure VI.1).

Table 2. Functional categories of SNPs in the spoon-billed sandpiper and red-necked stint populations. Seven out of nine individuals of red-necked stint population were selected for comparison in all possible configurations, and results were averaged across all replicates. Common polymorphisms includes only SNPs with minor allele frequency greater than 15%.

	Number of changes in the spoon-billed sandpiper genome	Number of changes in the red-necked stint genome
Nonsynonymous substitutions	14,268.1	12,741.5
Synonymous substitution	26,609.4	21,683.4
Intergenic substitutions	1,782,371.0	1,564,947.0
Nonsynonymous polymorphisms	30,107.0	38,492.3
Synonymous polymorphisms	46,866.0	86,006.0
Intergenic polymorphisms	2,708,036.0	4,348,787.0
Nonsynonymous common polymorphisms	12,873.0	12,883.1
Synonymous common polymorphisms	22,296.1	29,120.1
Intergenic common polymorphisms	1,354,796.0	1,709,662.0

To detect positive selection in the two populations we used the McDonald-

Kreitman test (McDonald and Kreitman 1991) on the fixed and segregated polymorphisms in the two populations. In the red-necked stint population we estimate that 24% of amino acid substitutions have been driven by positive selection (Table 2). No direct evidence of positive selection was detected by the McDonald-Kreitman test in the spoon-billed sandpiper lineage (alpha = -0.2, **Table 2**). However, negative alpha in the spoon-billed sandpiper population likely reflects the accumulation of deleterious non-synonymous variants in the spoon-billed sandpiper population (citation?). This hypothesis is confirmed by the observations that if only polymorphisms with a very high derived frequency are used, then the McDonald-Kreitman test, predicting fraction of amino acid substitutions that were subject to positive selection in spoon-billed sandpiper, shift towards zero (alpha = -0.07; alpha = 0.18 in the red-necked stint; Table 2). Indeed, the dn/ds ratio, reflecting fixed changes, was similar in the spoon-billed sandpiper and the red-necked stint, dn/ds = 0.18 and 0.2, respectively, indicating a similar selection pressure in evolution of the two species prior to recent population changes that affected the polymorphisms (Table 2, Supplementary Table 2).

We then searched for specific regions in the genomes of the two species that may have been under positive selection since their divergence. We calculated Z-scored values in 25kbp windows, identifying

regions that had many differences between the spoon-billed sandpiper genome relative to the orthologous red-necked stint region, and visa versa. We identified an excess of genomic regions with a high Z-scored values relative to a normal distribution (Supplementary Figure VIII.1) with 761 windows (251 regions) showing a Z-scored value >2.5 (corresponding to a p-value < 0.006), of them 129 had the excess of substitutions in the spoon-billed sandpiper lineage (620 in red-necked stint). The 128 protein coding genes found in these 114 spoon-billed sandpiper accelerated regions in were significantly enriched for various biological processes (Supplementary Table VIII.1), majority related to positive regulation of cell growth (Supplementary Figure VIII.2). Interestingly, two genes - secreted frizzled-related protein (Epyg1c017260) and E3 ubiquitin ligase SMURF1/2 (Epyg1c020968) - are involved in negative regulation of BMP signalling pathway, that might be related to beak development, as it was suggested by others. By contrast, the regions with increased accumulation of substitutions in the red-necked stint did not include genes related to beak development or to positive regulation of cell growth (Supplementary Table VIII.2; Supplementary Figure VIII.3).

Remarkably, the few genomic regions with accelerated rate of evolution in the spoon-billed sandpiper harbor the genes that have been shown in multiple studies (citations) to affect the development of the

bill - the defining morphological feature of the species with a clear ecological significance. The other genes detected in the same analysis include genes important in bone development and immunity-related genes, with the latter thought of being important for endangered species.

The consistent and gradual decline of the spoon-billed sandpiper population since the last glacial age suggests a concomitant decline in appropriate habitat. Thus, we considered possible distribution of the spoon-billed sandpiper breeding range at the time of the reconstructed population peak during the last glacial maximum 18-12 thousand years ago (tya). At the time of maximum glaciation Eurasia and North America were connected by the land bridge Beringia, due to the regression of sea levels, up to 100-130 meters lower than at present (Velichko et al., XXX; Yokoyama et al. 2000; Fairbanks 1989; Lambeck and Chappell 2001). The regression exposed vast areas of the flat continental shelf in what is now the Bering Sea, which in combination with a colder climate, provides two major factors contributing to a more widespread spoon-billed sandpiper breeding habitat during the last glacial maximum. First, the exposed continental shelf created large flat areas of rugged coastal line (Manley 2002; Ehlers and Gibbard 2004; Clark et al., 2014), including lagoons with crowberry spits, salt marshes and tundra vegetation (Erland-

son et al., 2015) likely resulting in large paleo-Anadyr and paleo-YukonâĂŞKuskokwim deltas. Second, the colder climate at the time of the last glacial maximum allowed for the presence of acceptable breeding habitat over a wider area extending far beyond the modern breeding range (*figure map in prep*). The large areas of coastal tundra and large deltas of paleo-rivers, correspond to breeding habitats of the spoon-billed sandpiper, suggesting that the peak of the spoon-billed sandpiper population may have occurred at that time due to wider availability of breeding habitat. The current critical state of the spoon-billed sandpiper population may thus be caused by it being especially sensitive to anthropogenic pressures on the flyway because of a pre-existing vulnerability associated with deglaciation-driven breeding habitat loss.

Our analysis show that the spoon-billed sandpiper experienced a persistent 5000 generation-long decline in population size since the last glacial age. The decline was, on one hand, recent enough not to affect the overall level of genetic variability, on the other, long enough to have led to a substantial accumulation of slightly deleterious alleles. Compared to the red-necked stint, the spoon-billed sandpiper carries thousands of extra rare deleterious alleles. Conversely, the high level of heterozygosity was maintained because it is mostly defined by common alleles (citation) that are not greatly affected immediately in the

course of the population decline (citation), and they are less likely to contribute to a decline in fitness with increased rate of inbreeding (citations). These data suggest that the spoon-billed sandpiper population may be particularly predisposed to loss of fitness due to inbreeding because of the long gradual decline of the spoon-billed sandpiper population that led to the accumulation of a large number of rare deleterious alleles. The recently established spoon-billed sandpiper breeding program (citation) must, therefore, be especially mindful in avoiding inbreeding among their captive individuals.

Our results also suggest that genetic studies of endangered species should not rely on the level of heterozygosity for the prediction of the risks of inbreeding but should take into account the accumulated rare deleterious variants. It is possible that the genetic risks to species survival are especially great in populations with prolonged population declines rather than those that has experienced a rapid crash. Among the possibilities of particular vulnerability of the spoon-billed sandpiper as a species is the adverse interplay of natural geological and climatic factors that initiated the initial population decline that was exacerbated by recent human activity (citation); a situation that may be pertinent to many other Arctic-dwelling species.

Conservation genetic studies focus on rapid loss of population size a risk factor in the accumulation of deleterious alleles in the endangered

population (citations). Our study suggests that a gradual decline in population size presents a greater danged to an endangered species, as more deleterious variants have the time to accumulate in a gradually declining population (citation) with the population becoming particularly vulnerable to the deleterious influence of inbreeding once the population reaches specifically low numbers, especially in species with a low geographical dispersion. The lack of inbreeding in the modern spoon-billed sandpiper is a hopeful sign for the species, however, the predicted substantial numbers of the slightly deleterious alleles in the population underscore the importance of insightful breeding regime for the recently established spoon-billed sandpiper breeding program.

This study highlights the interplay of long-term climate change and human-induced factors in defining the fate of species. The concomitant action of drift in removing pre-existing rare variants at all sites and the accumulation of new rare variants due to relaxation of selection at functionally important sites is most consistent with the higher proportion of polymorphisms at nonsense and nonsynonymous sites in the spoon billed sandpiper.

5.4 Materials and methods

Samples

Samples were collected between 2002 and 2013 at three locations (Supplementary Table 1). One sample of spoon-billed sandpiper was selected to provide the reference genome, other samples to provide information about genomic variation in the spoon-billed sandpiper and red necked stint populations. Single individuals from three other species (red knot, little stint and long-toed stint) were used for phylogenic and outgroup-based analyses. DNA from all these samples were extracted with Gentra Purgene Tissue Kit following standard protocol. We extracted also a RNA from single embryo sample, to facilitate gene prediction and genome annotation. All individuals used for population analysis were sampled prior to 2011, the year a head-starting program for the spoon-billed sandpiper was initiated. RNA was extracted from individuals collected after 2013 to aid in gene prediction.

Library preparation and sequencing

To produce a reference genome, a 0.5, 3 and 4.5 kb average insert sizes libraries were prepared from a single individual. In addition, a 450 bp fragment PCR-free library was prepared from DNA of the same bird. The 500bp fragment, and the two mate-pair libraries were sequenced on an Illumina HiSeq 2000 at the Center for Genomic Regulation (Barcelona, Spain). The 450 bp library was sequenced on an Illumina MiSeq to obtain 250bp overlapping reads. All reads were

manually inspected using FASTQC and used for genome assembling. An overview of the samples used in this study can be found in Supplementary Table 1. De novo assembly We ran DiscovarDeNovo assembler (DDN V. r52488, default parameters) using the 450bp-fragment. The scaffolds obtained covered 1.27GB of the spoon-billed sandpiper genome (1.2GB expected genome size) and had an N50 of 100kb. To improve the assembly we excluded contigs shorter than 500bp. To extend and scaffold the contigs produced by DDN we ran SSPACE-SR (v3.0) in two steps. First, we extended the contigs using the single 500bp library. Later, we scaffolded the extended assembly using both mate-pair libraries with SSPACE-SR (default parameters, -x = 0, -z = 0, -k = 5, -g = 0, -a = 0.7, -n = 15, -p = 0). The statistics for each assembly were calculated using the assemblathon pipeline (Earl et al. 2011) described in Supplementary Table 2.

Genome annotation

The reference genome was annotated with Evidence Modeler by combining different sources of evidence. Prior to the gene-finding step the assembly's complex-repeats were hard-masked using RepeatMasker (v4.0.5) with *G. gallus*-derived library of repeat elements. After that several ab-initio gene predictors were applied to the reference genome. Additionally, we aligned to the reference genome PASA-derived transcriptomic sequences, highly curated vertebrate sequences and chicken

protein models. All these sources of evidence were combined using specific weights. Functional annotation of a final set of genes was performed with a pipeline utilizing Interproscan, KEGG and Blast2GO software. Detailed information about genome annotation can be found in Supplementary Materials.

lncRNA annotation of the Spoon-billed sandpiper bird

We annotated the long non-coding RNA (lncRNA) transcripts of the Spoon-billed Sandpiper genome with a combination of tools, including the Coding Potential Calculator (CPC) and the Coding Potential Annotation Tool (CPAT). The former detects putative ORFs homologies and performs protein database parsing, whereas the latter is based on a logistic regression model. The required models for de novo prediction using CPAT were trained with 7,839 annotated long intergenic non-coding RNAs (lincRNAs) from the chicken (estimated time of divergence with the SBS, TTOL = 98.0 mya) and an equal randomly selected number of coding sequences (CDS) to avoid a biased detection model. These sequences were used as a proxy to detect the coding and non-coding potential of identified transcripts. The estimation of the optimum cutoff value (=0.59), to determine whether a transcript was either a protein coding or non-coding gene was assessed by using a ROC curve (Supplemental Fig.) using data from 8161 CDS and 7839 lincRNAs (total=16000) from the chicken

and a modification of the R script given by CPAT. The output of CPAT was fed to CPC and the remaining sequences were blasted (BLASTN) against the SBS protein-coding sequences predicted previously with EVModeller. Sequences with an overlap below 10% with a protein-coding were retained as well as those without any significant hit with a protein in the RefSeq database (latest version). Only transcripts >=200 nucleotides and with a ratio of transposable elements less than 0.4 were considered as lncRNAs. Finally, we filtered transcripts with expression levels >= 0.5 FPKM, utilizing RNA-Seq data of the previously mapped transcriptome.

Mapping and SNP calling

Samples other than used for genome assembly were sequenced with Illumina HiSeq2000 platform (2 x 125bp reads). The quality of reads was assessed with FASTQC, and low quality reads were trimmed with Trimmomatic (version 0.32) (Bolger et al. 2014). Reads were then mapped to the reference genome with bowtie2 (version 2.2.3) (Langmead and Salzberg 2012). PCR duplicates were marked, indels were realigned and sam/bam files were manipulated with SAMtools and GATK; SNPs were called with SAMtools mpileup (with additional options: -C50 -R -t DP,ADF,ADR) and filtered with bcftools filter and vcftools. Based on the empirical distributions we decided to remove sites with QUAL âĽď 30, DP âĽď 40, DP âĽě 250, MQSB

âĽď 0.001, and within 5 bp of indels. Indels were normalized and left aligned. The missing and low quality genotypes were inferred separately for each species using BEAGLE (version 4.1) (Browning and Browning 2007).

Relatedness and phylogeny reconstruction

To infer relationship between individuals within each species we used KING software (Manichaikul et al. 2010) assuming no population structure. From each group of individuals related up to first degree we randomly selected one individual for future analyses avoiding samples that were sequenced with other technology (MiSeq) or to different depth (samples sequenced for genome assembly) (Supplementary Table 1). The principal component analyses on final samples was performed with PLINK (Chang et al. 2015) and phylogeny was inferred with SNPhylo (Lee et al. 2014) software using following options -m 0.001 -l 0.1 -p 20 -M 0.2 -b -B 100 -a 30000.

Population statistics

We scanned the contigs longer than 50kb in non-overlapping 25kb windows to estimate nucleotide diversity and Tajima's D statistics within spoon-billed sandpiper and red-necked stint populations and to estimate genetic divergence ( ) between species. The VCFtools (0.1.15) software (Danecek et al. 2011) was used for these calcu-

lations and GO enrichment analyses for selected widows were performed with topGO package (Alexa and Rahnenfuhrer 2010). To calculate Tajima's D statistic per specific region we used PopGenome package within R (Pfeifer et al. 2014). To count the number of polymorphisms and species-specific substitutions we excluded tri-allelic sites and sites polymorphic in both species. Using information from an outgroup we then applied principle of parsimony to reconstruct ancestral state and count number of polymorphisms and substitutions with custom script.

Mutation rate

We used protein coding sequences from spoon-billed sandpiper and killdeer, to estimate mutation rate. The reciprocal blast was performed to identify orthologous sequences between species. We used 12,398 orthologous pairs to estimate rate of synonymous substitutions. Each pair of sequences was aligned based on the translated protein sequence. The alignment was then used for estimate rate of synonymous substitutions and number of synonymous sites with PAML (yn00) (Yang 2007). The divergence time between spoon-billed sandpiper and killdeer was assumed to be 76.0 million years, according to the www.timetree.org (Kumar et al. 2017). The calculated mutation rate was 2.40 x 10-9 as a rate of mutation per nucleotide per year and 1.20 x 10-8 as a rate of mutation per nucleotide

per generation (1 generation = 5 years).

Population history

To perform PSMC (Li and Durbin 2011) analyses, the mpileup file was generated with SAMtools, and sites were filtered to have minimum mapping quality 25, consensus quality higher than 20 and depth in a range between 10 and 2 x d, where d is average depth (Nadachowska-Brzyska et al. 2016). PSMC was then run with the following options -N50 -t5 -r5 -b -p "4+30x2+4+6+10" and output was generated with the mutation rate calculated as described above. The same procedure was performed for three species: spoon-billed sandpiper, red-necked stint and red knot using spoon-billed sandpiper genome as a reference. As a control, we performed the same analyses on the ruff genome.

5.4.1 Supplementary information about lncRNAs

We identified 5425 lncRNA transcripts (**Figure 3A**)in the SBS genome, from which only a total of 1224 transcripts had expression levels higher than 0.5 FPKM (**Figure 3B**); this could be due to the inadequate quality of the transcriptome data to detect expressed lncRNAs which are generally expressed at low levels. Only 132 transcripts had a relevant BLAST hit with another predicted ncRNA (>80% identity and length) from other bird genomes, whereas 1733 genomic

transcripts had a non-significant BLAST hit (<80% identity, <80% length) with a ncRNA from another bird. More than 50% of the total transcripts were smaller than 2000 nt, however, a few transcripts were longer than 15000nt, the content of transposable elements is shown in **Figure 3C**.

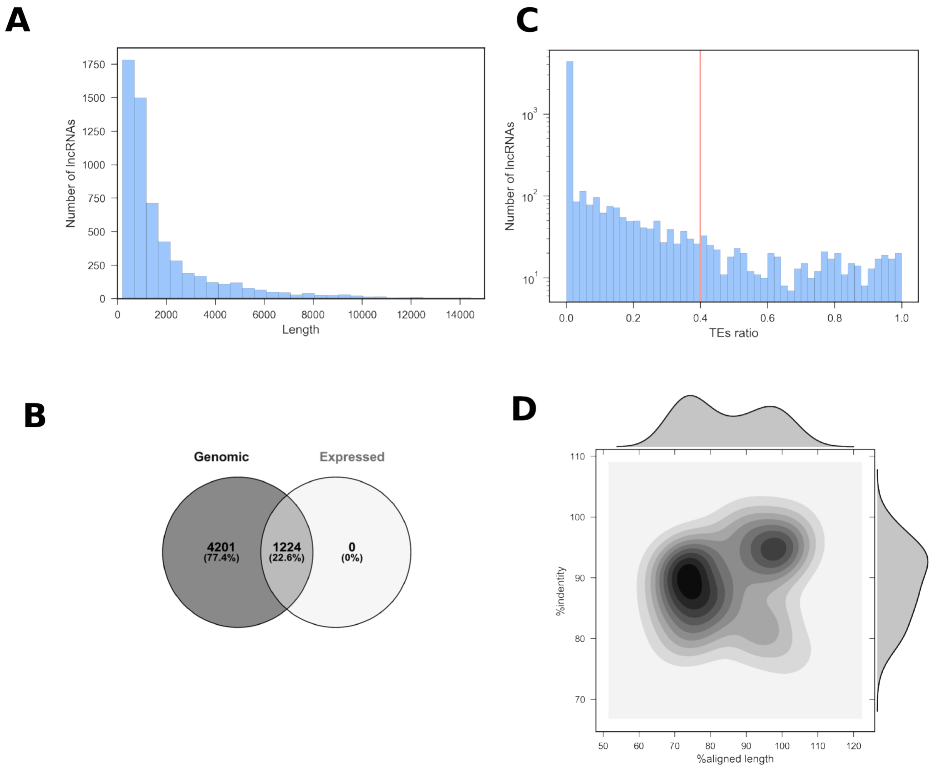

Figure 3: lncRNA annotation in the spoon-billed sandpiper. (A) Length distribution of lncRNAs. (B) Transposable elements (TE) content in identified lncRNAs. (C) Genomic versus expressed lncRNAs (>0.5FPKM). (D) Distribution of %identities and aligned lengths between the spoonbilled-sandpiper and human lncRNAs.

Moreover, we identified 37 spoon-billed sandpiper lncRNA genes with a significant Blastn hit (>70%Identity and aligned length) with an annotated human lncRNA (GENCODE v26) (**Figure 3D**), such as FOXP4-AS1-005, HTR5A-AS1, MEF2C-AS1, OIP5-AS1-001 (known as cyrano in zebrafish, citation), ZEB2-AS1 (a regulator of ZEB2 which is essential for Schwann cell differentiation, myelination and nerve repair [Quintes et al., 2016] among others. However, only 16 genes had expression levels >0.5FPKM. It is likely that these lncRNA genes are functionally relevant and it is an interesting example of conservation being an indicator of functionality throughout evolution.

Acknowledgements

We are indebted to Roman Belogorodtsev and Svetlana Belogorodtseva for invaluable on-site support of our field work. We thank Jochen Hecht and the CRG Genomics Unit for the next-gen sequencing and Yun Song for invaluable input on SMC++ application. Mateusz Konczal was supported by the Foundation for Polish Science and the Polish National Science Centre (2016/20/S/NZ8/00208). The work was supported by Royal Society for Protection of Birds (UK), Wildfowl and Wetland Trust (UK), Manfred Hermsen Foundation (Germany), NABU (Germany), Keidanren Foundation (Japan), the administration of Chukotka Autonomous District of Russian Federation, Bird Life International, Japan Ramsar Network, Heritage

Expeditions (New Zealand), HHMI International Early Career Scientist Program (55007424), the MINECO (BFU2012-31329, BES-2013-064004 and BFU2015-68723-P), Spanish Ministry of Economy and Competitiveness, Centro de Excelencia Severo Ochoa 2013-2017 grant (SEV-2012-0208), Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat's AGAUR program (2014 SGR 0974), the CERCA Programme of the Generalitat de Catalunya, and the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013, ERC grant agreement 335980-EinME).

5.5 References

Spielman, D., Brook, B. W., & Frankham, R. (2004). Most species are not driven to extinction before genetic factors impact them. Proceedings of the National Academy of Sciences of the United States of America, 101(42), 15261-15264.

Frankham, R. (2005). Genetics and extinction. Biological conservation, 126(2), 131-140.

Johnson, W. E., Onorato, D. P., Roelke, M. E., Land, E. D., Cunningham, M., Belden, R. C., ... Howard, J. (2010). Genetic restoration of the Florida panther. Science, 329(5999), 1641-1645.

Clark, P. U., Dyke, A. S., Shakun, J. D., Carlson, A. E., Clark, J., Wohlfarth, B., ... & McCabe, A. M. (2009). The last glacial maximum. science, 325(5941), 710-714.

BirdLife International. 2016. *Calidris ruficollis*. The IUCN Red List of Threatened Species 2016: e.T22693383A93401907

Pimm, S. L., Jenkins, C. N., Abell, R., Brooks, T. M., Gittleman, J. L., Joppa, L. N., ... & Sexton, J. O. (2014). The biodiversity of species and their rates of extinction, distribution, and protection. Science, 344(6187), 1246752.

Tilman, D., Clark, M., Williams, D. R., Kimmel, K., Polasky, S., & Packer, C. (2017). Future threats to biodiversity and pathways to their prevention. Nature, 546(7656), 73-81.

Barnosky, A. D., Matzke, N., Tomiya, S., Wogan, G. O., Swartz, B., Quental, T. B., ... & Mersey, B. (2011). Has the Earth's sixth mass extinction already arrived?. Nature, 471(7336), 51-57.

Hua, N., Tan, K. U. N., Chen, Y., & Ma, Z. (2015). Key research issues concerning the conservation of migratory shorebirds in the Yellow Sea region. Bird Conservation International, 25(1), 38-52.

Bairlein, F. (2016). Migratory birds under threat. Science, 354(6312), 547-548.

Zockler, C., Syroechkovskiy, E. E., & Atkinson, P. W. (2010). Rapid and continued population decline in the Spoon-billed Sandpiper *Eurynorhynchus pygmeus* indicates imminent extinction unless conservation action is taken. Bird Conservation International, 20(2), 95-111.

Clark, N. A., Anderson, G. Q., Li, J., Syroechkovskiy, E. E., Tomkovich, P. S., Zockler, C., ... & Green, R. E. (2016). First formal estimate of the world population of the Critically Endangered spoon-billed sandpiper *Calidris pygmaea*. Oryx, 1-10.

Syroechkovski, E. E., Tomkovich, P. S., Kashiwagi, M., Taldenkov, I. A., Buzin, V. A., Lappo, E. G., & Zoeckler, C. (2010). Population decline in the spoon-billed sandpiper (E*urynorhynchus pygmeus*) in northern Chukotka based on monitoring on breeding grounds. Biology bulletin, 37(9), 941-951.

Zockler C, Hla TH, Clark N, Syroechkovskiy E, Yakushev N, Daengphayon S & Robinson R. Hunting in Myanmar is probably the main cause of the decline of the Spoon-billed Sandpiper *Calidris pygmeus*. Wader Study Group Bulletin (2010) 117(1): 1-8

Hebo, Bird Conservation International, 2017. In press.

Schraiber, J. G., & Akey, J. M. (2015). Methods and models for unravelling human evolutionary history. Nature Reviews Genetics,

16(12), 727-740.

Polishchuk, L. V., Popadin, K. Y., Baranova, M. A., & Kondrashov, A. S. (2015). A genetic component of extinction risk in mammals. Oikos, 124(8), 983-993.

Li, S., Li, B., Cheng, C., Xiong, Z., Liu, Q., Lai, J., ... & Zhang, H. (2014). Genomic signatures of near-extinction and rebirth of the crested ibis and other endangered bird species. Genome biology, 15(12), 557.

Dobrynin, P., Liu, S., Tamazian, G., Xiong, Z., Yurchenko, A. A., Krasheninnikova, K., ... & Kuderna, L. F. (2015). Genomic legacy of the African cheetah, Acinonyx jubatus. Genome biology, 16(1), 277.

Robinson, J. A., [MK6] Ortega-Del Vecchyo, D., Fan, Z., Kim, B. Y., Marsden, C. D., Lohmueller, K. E., & Wayne, R. K. (2016). Genomic flatlining in the endangered island fox. Current Biology, 26(9), 1183-1189.

Abascal F et al. Extreme genomic erosion after recurrent demographic bottlenecks in the highly endangered Iberian lynx. Genome Biol. 2016 Dec 14;17(1):251.

Rogers RL, Slatkin M. Excess of genomic defects in a woolly mammoth on Wrangel island. PLoS Genet. 2017 Mar 2;13(3):e1006601.

Li H, Durbin R. Inference of human population history from individual whole-genome sequences. Nature 475.7357 (2011): 493-496.

Red'kin YA, Tomkovich PS, Zdorikov AI. Unusual specimen of the Spoon-billed Sandpiper *Eurynorhynchus pygmeus*. Wader Study Group Bulletin (2012) 119 (1): 56.

Piersma, T. Using the power of comparison to explain habitat use and migration strategies of shorebirds worldwide. JOURNAL OF ORNITHOLOGY, 148,S45-S59, 2007.

Buehler, Deborah M.; Baker, Allan J.; Piersma, Theunis (2006). Reconstructing palaeoflyways of the late Pleistocene and early Holocene Red Knot *Calidris canutus*. Ardea. 94 (3): 485-498.

Charlesworth, B. 2009. Effective population size and patterns of molecular evolution and variation. Nat. Rev. Genet. 10: 195-205.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 123 (3): 585-95

Andolfatto P. Adaptive evolution of non-coding DNA in Drosophila. Nature. 2005 Oct 20;437(7062):1149-52.

Charlesworth, B., M. T. Morgan, and D. Charlesworth. 1993. The effect of deleterious mutations on neutral molecular variation. Genetics. 134: 1289-1303.

McDonald, J. H. Kreitman, M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature 351: 652-654.

Yokoyama, Y., Lambeck, K., De Deckker, P., Johnston, P., Fifield, L.K., 2000. Timing of the Last Glacial Maximum from observed sea-level minima. Nature 406, 713-716.

Fairbanks, R.G., 1989. A 17,000-year glacio-eustatic sea level record: influence of glacial melting rates on the Younger Dryas event and deep-ocean circulation. Nature 342, 637-642.

Manley, W. F. (2002). INSTAAR. Postglacial flooding of the Bering Land Bridge: A geospatial animation: Vol. 1 . Boulder, CO: University of Colorado.

Ehlers, J., & Gibbard, P. L. (2004). Quaternary glaciations extent and chronology, Part II: North America. In J. Rose (Ed.), Developments in quaternary science: Vol. 2. CDs 1 & 2 . Amsterdam, The Netherlands: Elsevier.

Clark J., J.X.Mitrovica, J. Alder (2014). Coastal paleogeography of the Califirns-Pregon-Washington and Bering Sea continental shelves during the latest Pleistocene and Holocene: implications for archaeological records. In: Journal of Archaeological Science 52: 12-23.

Erlandson JM, Braje TJ, Gill KM, Graham MH. (2015): Ecology of

the Kelp Highway: Did Marine Resources Facilitate Human Dispersal From Northeast Asia to the Americas?, The Journal of Island and Coastal Archaeology. 10, 392-411.

Chapter 6

Discussion

6.1 Summarizing discussion

In this thesis I have presented analyses focused on the non-coding regions of the genome, specifically lncRNAs and tRNAs genes in humans and other species. By using a comparative genomic approach I was able to 1) Identify human-specific lncRNA genes that seemingly are contributing to the unique genomic landscape of the human genome. 2) Show that there is a limit in the number of tRNAs identities that can evolve using bacterial tRNA orthologs pair-wise comparisons and 3) *De novo* annotate new lncRNAs in a non-model endangered species, the Spoonbilled-sandpiper bird. Each chapter of

the thesis has its own discussion, therefore here I will present a general discussion of my research and some perspectives for the future.

"***Organism complexity arises primarily from application of new regulatory control over duplicated genes rather than by invention of new activities*". [Allen et al., 2004]** .

After the human genome sequencing project was completed, this achievement was optimistically perceived as the final step to understand our nature; it was believed that after its completion we were going to be able to solve our evolutionary history, aging and all diseases sooner rather than later. It's been a while since the first human genome was drafted and we are still far from understanding complex and subtle genetic, epigenetic, epistatic and trans-generational interactions that can influence a broad spectrum of different phenotypes. However, it is undeniable we have made great progress since and continue to do so. With the increasing number of individual human genomes and other species genomes in our databases, together with the decrease in costs per genome and technological advances, we are closer and closer to understand many diseases and developmental processes.

Moreover, it is now generally accepted that non-coding regions in the genome harbor relevant information, (i.e, most of the somatic variants

in cancer genomes occur in non-coding regions) and shifting our attention to those regions can help us understand many biological processes. An increasing amount of experimental techniques and computational tools that allows to detect and characterize interactions and non-coding variants are being utilized, such as ATAC-seq, CHIP-seq, Hi-C, Enhancer-FACS-seq, CRISPR-Cas9, OncodriveFML, just to name a few. However, it is still unclear, and a matter of debate, how much of the non-coding genome is functional, specially when referring to lncRNA genes.

I believe the concept of functionality shouldn't be treated lightly and it is true that if we measure functionality based on the degree of sequence conservation, most classes of lncRNAs will be cataloged as transcriptional noise. However, lncRNAs cannot not be studied as if they were protein-coding genes; because it has been shown that changes in their sequences don't necessarily affect their structure. The latter could explain the strong purifying selection observed in the promoter regions of all lncRNAs.

Moreover, I've found that analyses of evolutionary mechanisms, such as duplication, together with detection of selection signatures are very informative to find relevant lncRNA genes; specially in combination with gene expression data. And has lead me to consider, that many of them, at least in part, might not be just transcriptional noise, yet

we are still far from understanding all the functional elements in the genome and their role in regulation.

Finally, the approach used in this thesis could be used to analyze not only lncRNA genes, but any type of gene, to detect abnormal numbers of exon copies in diseased patients, which could lead to the identification of potential targets as biomarkers and drug development.

Chapter 7

Conclusions

- 11% of human lncRNA exons have at least one duplication and prefer partial rather than whole gene duplication.
- Our data show that lncRNAs originated by exon duplication at a constant rate throughout great ape evolution giving rise to 62 human-specific genes.
- We identified potential candidate genes that are undergoing active selection and show tissue-specific expression patterns which require further experimental validation.
- Several are expressed in the brain and testis, while others are expressed ubiquitously.

- We develop a method by using a comparative genomics approach to identify relevant small-scale duplications in the non-coding genome.

- There is an accumulation of structural and functional constraints that operate on tRNAs and acts as an operational limit that impedes the evolution of new tRNA identities and, as a result, the incorporation of new amino acids into the genetic code.

- We identified 5425 lncRNA transcripts in the SBS genome, from which only a total of 1224 transcripts had expression levels higher than 0.5 FPKM; this could be due to the inadequate quality of the transcriptome data to detect expressed lncRNAs which are generally expressed at low levels.

- 41 SBS lncRNA transcripts had a significant hit with an annotated human lncRNA, indicating these lncRNA genes are functionally relevant.

- Endangered species should be studied at the population level to develop adequate conservation programs tailored to their genetic background.

Bibliography

[Alexander et al., 2010] Alexander, R. P., Fang, G., Rozowsky, J., Snyder, M., and Gerstein, M. B. (2010). Annotating non-coding regions of the genome. *Nature Reviews Genetics*.

[Allen et al., 2004] Allen, E., Xie, Z., Gustafson, A. M., Sung, G.-H., Spatafora, J. W., and Carrington, J. C. (2004). Evolution of microRNA genes by inverted duplication of target gene sequences in Arabidopsis thaliana. *Nature Genetics*, 36(12):1282–1290.

[Andolfatto, 2005] Andolfatto, P. (2005). Adaptive evolution of non-coding DNA in Drosophila. *Nature*, 437(7062):1149–1152.

[Birney et al., 2007] Birney, E., Stamatoyannopoulos, J. a., Dutta, A., Guigó, R., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., Dermitzakis, E. T., Thurman, R. E., Kuehn, M. S., Taylor, C. M., Neph, S., Koch, C. M., Asthana, S., Malhotra, A., Adzhubei, I., Greenbaum, J. a., Andrews, R. M., Flicek, P., Boyle,

P. J., Cao, H., Carter, N. P., Clelland, G. K., Davis, S., Day, N., Dhami, P., Dillon, S. C., Dorschner, M. O., Fiegler, H., Giresi, P. G., Goldy, J., Hawrylycz, M., Haydock, A., Humbert, R., James, K. D., Johnson, B. E., Johnson, E. M., Frum, T. T., Rosenzweig, E. R., Karnani, N., Lee, K., Lefebvre, G. C., Navas, P. a., Neri, F., Parker, S. C. J., Sabo, P. J., Sandstrom, R., Shafer, A., Vetrie, D., Weaver, M., Wilcox, S., Yu, M., Collins, F. S., Dekker, J., Lieb, J. D., Tullius, T. D., Crawford, G. E., Sunyaev, S., Noble, W. S., Dunham, I., Denoeud, F., Reymond, A., Kapranov, P., Rozowsky, J., Zheng, D., Castelo, R., Frankish, A., Harrow, J., Ghosh, S., Sandelin, A., Hofacker, I. L., Baertsch, R., Keefe, D., Dike, S., Cheng, J., Hirsch, H. a., Sekinger, E. a., Lagarde, J., Abril, J. F., Shahab, A., Flamm, C., Fried, C., Hackermüller, J., Hertel, J., Lindemeyer, M., Missal, K., Tanzer, A., Washietl, S., Korbel, J., Emanuelsson, O., Pedersen, J. S., Holroyd, N., Taylor, R., Swarbreck, D., Matthews, N., Dickson, M. C., Thomas, D. J., Weirauch, M. T., Gilbert, J., Drenkow, J., Bell, I., Zhao, X., Srinivasan, K. G., Sung, W.-K., Ooi, H. S., Chiu, K. P., Foissac, S., Alioto, T., Brent, M., Pachter, L., Tress, M. L., Valencia, A., Choo, S. W., Choo, C. Y., Ucla, C., Manzano, C., Wyss, C., Cheung, E., Clark, T. G., Brown, J. B., Ganesh, M., Patel, S., Tammana, H., Chrast, J., Henrichsen, C. N., Kai, C., Kawai, J., Nagalakshmi, U., Wu, J., Lian, Z., Lian, J., Newburger, P., Zhang, X., Bickel, P., Mattick,

J. S., Carninci, P., Hayashizaki, Y., Weissman, S., Hubbard, T., Myers, R. M., Rogers, J., Stadler, P. F., Lowe, T. M., Wei, C.-L., Ruan, Y., Struhl, K., Gerstein, M., Antonarakis, S. E., Fu, Y., Green, E. D., Karaöz, U., Siepel, A., Taylor, J., Liefer, L. a., Wetterstrand, K. a., Good, P. J., Feingold, E. a., Guyer, M. S., Cooper, G. M., Asimenos, G., Dewey, C. N., Hou, M., Nikolaev, S., Montoya-Burgos, J. I., Löytynoja, A., Whelan, S., Pardi, F., Massingham, T., Huang, H., Zhang, N. R., Holmes, I., Mullikin, J. C., Ureta-Vidal, A., Paten, B., Seringhaus, M., Church, D., Rosenbloom, K., Kent, W. J., Stone, E. a., Batzoglou, S., Goldman, N., Hardison, R. C., Haussler, D., Miller, W., Sidow, A., Trinklein, N. D., Zhang, Z. D., Barrera, L., Stuart, R., King, D. C., Ameur, A., Enroth, S., Bieda, M. C., Kim, J., Bhinge, A. a., Jiang, N., Liu, J., Yao, F., Vega, V. B., Lee, C. W. H., Ng, P., Yang, A., Moqtaderi, Z., Zhu, Z., Xu, X., Squazzo, S., Oberley, M. J., Inman, D., Singer, M. a., Richmond, T. a., Munn, K. J., Rada-Iglesias, A., Wallerman, O., Komorowski, J., Fowler, J. C., Couttet, P., Bruce, A. W., Dovey, O. M., Ellis, P. D., Langford, C. F., Nix, D. a., Euskirchen, G., Hartman, S., Urban, A. E., Kraus, P., Van Calcar, S., Heintzman, N., Kim, T. H., Wang, K., Qu, C., Hon, G., Luna, R., Glass, C. K., Rosenfeld, M. G., Aldred, S. F., Cooper, S. J., Halees, A., Lin, J. M., Shulha, H. P., Zhang, X., Xu, M., Haidar, J. N. S., Yu, Y., Iyer, V. R., Green, R. D., Wadelius, C., Farnham, P. J.,

Ren, B., Harte, R. a., Hinrichs, A. S., Trumbower, H., Clawson, H., Hillman-Jackson, J., Zweig, A. S., Smith, K., Thakkapallayil, A., Barber, G., Kuhn, R. M., Karolchik, D., Armengol, L., Bird, C. P., de Bakker, P. I. W., Kern, A. D., Lopez-Bigas, N., Martin, J. D., Stranger, B. E., Woodroffe, A., Davydov, E., Dimas, A., Eyras, E., Hallgrímsdóttir, I. B., Huppert, J., Zody, M. C., Abecasis, G. R., Estivill, X., Bouffard, G. G., Guan, X., Hansen, N. F., Idol, J. R., Maduro, V. V. B., Maskeri, B., McDowell, J. C., Park, M., Thomas, P. J., Young, A. C., Blakesley, R. W., Muzny, D. M., Sodergren, E., Wheeler, D. a., Worley, K. C., Jiang, H., Weinstock, G. M., Gibbs, R. a., Graves, T., Fulton, R., Mardis, E. R., Wilson, R. K., Clamp, M., Cuff, J., Gnerre, S., Jaffe, D. B., Chang, J. L., Lindblad-Toh, K., Lander, E. S., Koriabine, M., Nefedov, M., Osoegawa, K., Yoshinaga, Y., Zhu, B., and de Jong, P. J. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816.

[Britten and Kohne, 1968] Britten, R. J. and Kohne, D. E. (1968). Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. *Science (New York, N.Y.)*, 161(3841):529–40.

[Cech, 1986] Cech, T. R. (1986). A model for the RNA-catalyzed

replication of RNA. *Proceedings of the National Academy of Sciences of the United States of America*, 83(12):4360–3.

[Cech, 2009] Cech, T. R. (2009). Crawling out of the RNA world. *Cell*, 136(4):599–602.

[Cheng et al., 2005] Cheng, Z., Ventura, M., She, X., Khaitovich, P., Graves, T., Osoegawa, K., Church, D., DeJong, P., Wilson, R. K., Pääbo, S., Rocchi, M., and Eichler, E. E. (2005). A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature*, 437(7055):88–93.

[Dennis et al., 2012] Dennis, M., Nuttle, X., Sudmant, P., Antonacci, F., Graves, T., Nefedov, M., Rosenfeld, J., Sajjadian, S., Malig, M., Kotkiewicz, H., Curry, C., Shafer, S., Shaffer, L., deÂăJong, P., Wilson, R., and Eichler, E. (2012). Evolution of Human-Specific Neural SRGAP2 Genes by Incomplete Segmental Duplication. *Cell*, 149(4):912–922.

[Devaux et al., 2015] Devaux, Y., Zangrando, J., Schroen, B., Creemers, E. E., Pedrazzini, T., Chang, C.-P., Dorn, G. W., Thum, T., and Heymans, S. (2015). Long noncoding RNAs in cardiac development and ageing. *Nature Reviews Cardiology*.

[Force et al., 1997] Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y.-L., and Postlethwait, J. (1997). Preservation of Duplicate Genes by Complementary, Degenerative Mutations.

[Francis Crick, 1970] Francis Crick (1970). Central Dogma of Molecular Biology. *Nature*.

[Franklin and Gosling, 1953] Franklin, R. E. and Gosling, R. G. (1953). Molecular Configuration in Sodium Thymonucleate. *Nature*.

[Geisler and Coller, 2013] Geisler, S. and Coller, J. (2013). RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. *Nature Reviews Molecular Cell Biology*, 14(11):699–712.

[Gilbert, 1986] Gilbert, W. (1986). Origin of life: The RNA world. *Nature*, 319(6055):618–618.

[Grammatikakis et al., 2014] Grammatikakis, I., Panda, A. C., Abdelmohsen, K., and Gorospe, M. (2014). Long noncoding RNAs (lncRNAs) and the molecular hallmarks of aging. *Aging*, 6(12):992–1009.

[Guerrier-Takada et al., 1983] Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N., and Altman, S. (1983). The RNA moiety

of ribonuclease P is the catalytic subunit of the enzyme. *Cell*, 35(3 Pt 2):849–57.

[Guttman et al., 2009] Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., Feldser, D., Huarte, M., Zuk, O., Carey, B. W., Cassady, J. P., Cabili, M. N., Jaenisch, R., Mikkelsen, T. S., Jacks, T., Hacohen, N., Bernstein, B. E., Kellis, M., Regev, A., Rinn, J. L., and Lander, E. S. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, 458(7235):223–7.

[Huarte, 2015] Huarte, M. (2015). The emerging role of lncRNAs in cancer. *Nature Medicine*, 21(11):1253–1261.

[Human Genome Sequencing Consortium, 2004] Human Genome Sequencing Consortium, I. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945.

[Innan and Kondrashov, 2010a] Innan, H. and Kondrashov, F. (2010a). The evolution of gene duplications: classifying and distinguishing between models. *Nat Rev Genet*, 11(2):97–108.

[Innan and Kondrashov, 2010b] Innan, H. and Kondrashov, F. (2010b). The evolution of gene duplications: classifying and distinguishing between models. *Nature reviews. Genetics*, 11(2):97–108.

[Jacob, 1977] Jacob, F. (1977). Evolution and tinkering. *Science (New York, N.Y.)*, 196(4295):1161–6.

[Joyce, 1989] Joyce, G. F. (1989). RNA evolution and the origins of life. *Nature*, 338(6212):217–224.

[Kaessmann, 2010] Kaessmann, H. (2010). Origins, evolution, and phenotypic impact of new genes. *Genome research*, 20(10):1313–26.

[Kondrashov et al., 2002] Kondrashov, F. A., Rogozin, I. B., Wolf, Y. I., and Koonin, E. V. (2002). Selection in the evolution of gene duplications. *Genome biology*, 3(2):RESEARCH0008.

[Kumar and Subramanian, 2002] Kumar, S. and Subramanian, S. (2002). Mutation rates in mammalian genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 99(2):803–8.

[Kutter et al., 2012] Kutter, C., Watt, S., Stefflova, K., Wilson, M. D., Goncalves, A., Ponting, C. P., Odom, D. T., and Marques, A. C. (2012). Rapid turnover of long noncoding RNAs and the evolution of gene expression. *PLoS genetics*, 8(7):e1002841.

[Lander et al., 2001] Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle,

M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.-F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada,

T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H.-C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G. R., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F. A., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S.-P., Yeh, R.-F., Collins, F., Guyer,

M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Patrinos, A., and Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

[Long et al., 2003] Long, M., Betrán, E., Thornton, K., and Wang, W. (2003). The origin of new genes: glimpses from the young and old. *Nature reviews. Genetics*, 4(11):865–75.

[Long and Langley, 1993] Long, M. and Langley, C. H. (1993). Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila. *Science (New York, N.Y.)*, 260(5104):91–5.

[Lynch, 1994] Lynch, M. (1994). Mutation Accumulation RNA Genes in Nuclear , Organelle , and Prokaryotic Transfer regions. pages 914–925.

[Lynch, 2000] Lynch, M. (2000). The Evolutionary Fate and Consequences of Duplicate Genes. *Science*, 290(5494):1151–1155.

[Lynch and Conery, 2001] Lynch, M. and Conery, J. S. (2001). The Evolutionary fvate and Consequences of Duplicate Genes. *Science*, 290(5494):1151–1155.

[Magistri et al., 2012] Magistri, M., Faghihi, M. A., St Laurent, G., Wahlestedt, C., and Wahlestedt, C. (2012). Regulation of chromatin structure by long noncoding RNAs: focus on natural antisense transcripts. *Trends in genetics : TIG*, 28(8):389–96.

[Marques-Bonet et al., 2009] Marques-Bonet, T., Girirajan, S., and Eichler, E. E. (2009). The origins and impact of primate segmental duplications. *Trends in genetics : TIG*, 25(10):443–54.

[Material et al., 2004] Material, S. O., Web, S., Press, H., York, N., and Nw, A. (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science (New York, N.Y.)*, 306(5696):636–40.

[Mattick, 2004] Mattick, J. S. (2004). Opinion: RNA regulation: a new genetics? *Nature Reviews Genetics*, 5(4):316–323.

[Mattick, 2010] Mattick, J. S. (2010). Linc-ing Long noncoding RNAs and enhancer function. *Developmental cell*, 19(4):485–6.

[Mattick, 2011] Mattick, J. S. (2011). The central role of RNA in human development and cognition. *FEBS letters*, 585(11):1600–16.

[Meader et al., 2010] Meader, S., Ponting, C. P., Lunter, G., Ponting, C., Lunter, G., Siepel, A., Bejerano, G., Pedersen, J., Hinrichs, A., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J.,

Hillier, L., Richards, S., Weinstock, G., Wilson, R., Gibbs, R., Kent, W., Miller, W., Haussler, D., Halligan, D., Keightley, P., Asthana, S., Noble, W., Kryukov, G., Grant, C., Sunyaev, S., Stamatoyannopoulos, J., Casillas, S., Barbadilla, A., Bergman, C., Katzman, S., Kern, A., Bejerano, G., Fewell, G., Fulton, L., Wilson, R., Salama, S., Haussler, D., Fu, W., O'Connor, T., Jun, G., Kang, H., Abecasis, G., Leal, S., Gabriel, S., Altshuler, D., Shendure, J., Nickerson, D., Bamshad, M., Akey, J., Carninci, P., Yasuda, J., Hayashizaki, Y., Rinn, J., Euskirchen, G., Bertone, P., Martone, R., Luscombe, N., Hartman, S., Harrison, P., Nelson, F., Miller, P., Gerstein, M., Weissman, S., Snyder, M., Ponting, C., Oliver, P., Reik, W., Ponting, C., Belgard, T., Young, R., Marques, A., Tibbit, C., Haerty, W., Bassett, A., Liu, J., Ponting, C., Brown, C., Ballabio, A., Rupert, J., Lafreniere, R., Grompe, M., Tonlorenzi, R., Willard, H., Franke, A., Baker, B., Sleutels, F., Zwart, R., Barlow, D., Rinn, J., Kertesz, M., Wang, J., Squazzo, S., Xu, X., Brugmann, S., Goodnough, L., Helms, J., Farnham, P., Segal, E., Chang, H., Young, T., Matsuda, T., Cepko, C., Bernard, D., Prasanth, K., Tripathi, V., Colasse, S., Nakamura, T., Xuan, Z., Zhang, M., Sedel, F., Jourdren, L., Coulpier, F., Triller, A., Spector, D., Bessis, A., Tripathi, V., Ellis, J., Shen, Z., Song, D., Pan, Q., Watt, A., Freier, S., Bennett, C., Sharma, A., Bubulya, P., Blencowe, B., Prasanth, S., Prasanth, K., Orom, U., Derrien,

T., Beringer, M., Gumireddy, K., Gardini, A., Bussotti, G., Lai, F., Zytnicki, M., Notredame, C., Huang, Q., Guigo, R., Shiekhattar, R., Wang, K., Chang, H., Guttman, M., Rinn, J., Mohamed, J. S., Gaughwin, P., Lim, B., Robson, P., Lipovich, L., Guttman, M., Donaghey, J., Carey, B., Garber, M., Grenier, J., Munson, G., Young, G., Lucas, A., Ach, R., Bruhn, L., Yang, X., Amit, I., Meissner, A., Regev, A., Rinn, J., Root, D., Lander, E., Huarte, M., Rinn, J., Wapinski, O., Chang, H., Ponjavic, J., Ponting, C., Lunter, G., Guttman, M., Amit, I., Garber, M., French, C., Lin, M., Feldser, D., Huarte, M., Zuk, O., Carey, B., Cassady, J., Cabili, M., Jaenisch, R., Mikkelsen, T., Jacks, T., Hacohen, N., Bernstein, B., Kellis, M., Regev, A., Rinn, J., Lander, E., Marques, A., Ponting, C., Chodroff, R., Goodstadt, L., Sirey, T., Oliver, P., Davies, K., Green, E., Molnar, Z., Ponting, C., Ulitsky, I., Shkumatava, A., Jan, C., Sive, H., Bartel, D., Kutter, C., Watt, S., Stefflova, K., Wilson, M., Goncalves, A., Ponting, C., Odom, D., Marques, A., Schorderet, P., Duboule, D., Nielsen, R., Ward, L., Kellis, M., Abecasis, G., auton, A., Brooks, L., DePristo, M., Durbin, R., Handsaker, R., Kang, H., Marth, G., McVean, G., Charlesworth, B., Mackay, T., Richards, S., Stone, E., Barbadilla, A., Ayroles, J., Zhu, D., Casillas, S., Han, Y., Magwire, M., Cridland, J., Richardson, M., Anholt, R., Barrón, M., Bess, C., Blankenburg, K., Carbone, M., Castellano, D., Chaboub, L., Duncan, L., Harris,

Z., Javaid, M., Jayaseelan, J., Jhangiani, S., Jordan, K., Lara, F., Lawrence, F., Lee, S., Librado, P., Linheiro, R., Lyman, R., Cabili, M., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., Rinn, J., Yoo, E., Cooke, N., Liebhaber, S., Latos, P., Pauler, F., Koerner, M., Senergin, H., Hudson, Q., Stocsits, R., Allhoff, W., Stricker, S., Klement, R., Warczok, K., aumayr, K., Pasierbek, P., Barlow, D., Ponjavic, J., Oliver, P., Lunter, G., Ponting, C., Haddrill, P., Charlesworth, B., Halligan, D., Andolfatto, P., Lunter, G., Ponting, C., Hein, J., Parsch, J., Novozhilov, S., Saminadin-Peter, S., Wong, K., Andolfatto, P., Watterson, G., Tajima, F., Tajima, F., Jukes, T., Cantor, C., Drake, J., Bird, C., Nemesh, J., Thomas, D., Newton-Cheh, C., Reymond, A., Excoffier, L., Attar, H., Antonarakis, S., Dermitzakis, E., Hirschhorn, J., Keightley, P., Eyre-Walker, A., Boyko, A., Williamson, S., Indap, A., Degenhardt, J., Hernandez, R., Lohmueller, K., Adams, M., Schmidt, S., Sninsky, J., Sunyaev, S., White, T., Nielsen, R., Clark, A., Bustamante, C., Williamson, S., Hernandez, R., Fledel-Alon, A., Zhu, L., Nielsen, R., Bustamante, C., Keightley, P., Eyre-Walker, A., Torgerson, D., Boyko, A., Hernandez, R., Indap, A., Hu, X., White, T., Sninsky, J., Cargill, M., Adams, M., Bustamante, C., Clark, A., Kryukov, G., Schmidt, S., Sunyaev, S., Chen, C., Wang, J., Cohen, B., Artieri, C., Haerty, W., Singh, R., Charlesworth, B., Ohta, T., Ohta, T., Gillespie, J., Eyre-Walker, A., Keightley, P.,

Eyre-Walker, A., Keightley, P., Smith, N., Gaffney, D., Tenesa, A., Navarro, P., Hayes, B., Duffy, D., Clarke, G., Goddard, M., Visscher, P., Li, H., Durbin, R., Keightley, P., Lercher, M., Eyre-Walker, A., Pang, K., Frith, M., Mattick, J., Blow, M., McCulley, D., Li, Z., Zhang, T., Akiyama, J., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., Afzal, V., Bristow, J., Ren, B., Black, B., Rubin, E., Visel, A., Pennacchio, L., Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D., Lagarde, J., Veeravalli, L., Ruan, X., Ruan, Y., Lassmann, T., Carninci, P., Brown, J., Lipovich, L., Gonzalez, J., Thomas, M., Davis, C., Shiekhattar, R., Gingeras, T., Hubbard, T., Notredame, C., Harrow, J., Guigò, R., Kondrashov, A., Chamary, J., Parmley, J., Hurst, L., Nakagawa, S., Naganuma, T., Shioi, G., Hirose, T., Eißmann, M., Gutschner, T., Hämmerle, M., Günther, S., Caudron-Herger, M., Groß, M., Schirmacher, P., Rippe, K., Braun, T., Zörnig, M., Diederichs, S., Nakagawa, S., Ip, J., Shioi, G., Tripathi, V., Zong, X., Hirose, T., Prasanth, K., Zhang, B., Arun, G., Mao, Y., Lazar, Z., Hung, G., Bhattacharjee, G., Xiao, X., Booth, C., Wu, J., Zhang, C., Spector, D., Bond, A., Vangompel, M., Sametsky, E., Clark, M., Savage, J., Disterhoft, J., Kohtz, J., Lewejohann, L., Skryabin, B., Sachser, N., Prehn, C., Heiduschka, P., Thanos, S., Jordan, U., Dell'Omo, G., Vyssotski, A., Pleskacheva, M., Lipp, H., Tiedge, H., Brosius,

J., Prior, H., Zhong, J., Chuang, S., Bianchi, R., Zhao, W., Lee, H., Fenton, A., Wong, R., Tiedge, H., Hillenmeyer, M., Fung, E., Wildenhain, J., Pierce, S., Hoon, S., Lee, W., Proctor, M., Onge, R. S., Tyers, M., Koller, D., Altman, R., Davis, R., Nislow, C., Giaever, G., Mercer, T., Dinger, M., Mattick, J., Dinger, M., Amaral, P., Mercer, T., Mattick, J., Bonferroni, C., Ladoukakis, E., Pereira, V., Magny, E., Eyre-Walker, A., Couso, J., Belgard, T., Marques, A., Oliver, P., Abaan, H., Sirey, T., Hoerder-Suabedissen, A., García-Moreno, F., Molnár, Z., Margulies, E., Ponting, C., Amberger, J., Bocchini, C., Scott, A., Hamosh, A., Stone, E., Hutter, S., Vilella, A., Rozas, J., Andolfatto, P., McDonald, J., Kreitman, M., Jenkins, D., Ortori, C., Brookfield, J., Ludwig, M., Kreitman, M., Eyre-Walker, A., Keightley, P., and Edgar, R. (2010). Massive turnover of functional sequence in human and other mammalian genomes. *Genome Research*, 20(10):1335–1343.

[Montes et al., 2015] Montes, M., Nielsen, M. M., Maglieri, G., Jacobsen, A., Højfeldt, J., Agrawal-Singh, S., Hansen, K., Helin, K., van de Werken, H. J. G., Pedersen, J. S., and Lund, A. H. (2015). The lncRNA MIR31HG regulates p16(INK4A) expression to modulate senescence. *Nature communications*, 6:6967.

[Morris and Mattick, 2014] Morris, K. V. and Mattick, J. S. (2014). The rise of regulatory RNA. *Nature Reviews Genetics*.

[Necsulea et al., 2014] Necsulea, A., Soumillon, M., Warnefors, M., Liechti, A., Daish, T., Zeller, U., Baker, J. C., Grützner, F., and Kaessmann, H. (2014). The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature*, 505(7485):635–40.

[Ohno, 1970] Ohno, S. (1970). *Evolution by gene duplication.* Springer-Verlag, Berlin-Heidelberg-New-York.

[Ohno, 1972] Ohno, S. (1972). So much "junk" DNA in our genome. *Brookhaven symposia in biology*, 23:366–70.

[Ponting et al., 2011] Ponting, C. P., Nellåker, C., and Meader, S. (2011). Rapid Turnover of Functional Sequence in Human and Other Genomes. *Annual Review of Genomics and Human Genetics*, 12(1):275–299.

[Ponting et al., 2009] Ponting, C. P., Oliver, P. L., and Reik, W. (2009). Evolution and functions of long noncoding RNAs. *Cell*, 136(4):629–41.

[Prensner and Chinnaiyan, 2011] Prensner, J. R. and Chinnaiyan, A. M. (2011). The emergence of lncRNAs in cancer biology. *Cancer discovery*, 1(5):391–407.

[Quintes et al., 2016] Quintes, S., Brinkmann, B. G., Ebert, M., Fröb, F., Kungl, T., Arlt, F. A., Tarabykin, V., Huylebroeck, D.,

Meijer, D., Suter, U., Wegner, M., Sereda, M. W., and Nave, K.-A. (2016). Zeb2 is essential for Schwann cell differentiation, myelination and nerve repair. *Nature Neuroscience*, 19(8):1050–1059.

[Rinn and Chang, 2012] Rinn, J. L. and Chang, H. Y. (2012). Genome regulation by long noncoding RNAs. *Annual review of biochemistry*, 81:145–66.

[Rinn et al., 2007] Rinn, J. L., Kertesz, M., Wang, J. K., Squazzo, S. L., Xu, X., Brugmann, S. A., Goodnough, L. H., Helms, J. A., Farnham, P. J., Segal, E., and Chang, H. Y. (2007). Functional Demarcation of Active and Silent Chromatin Domains in Human HOX Loci by Noncoding RNAs. *Cell*, 129(7):1311–1323.

[Robertson and Joyce, 2012] Robertson, M. P. and Joyce, G. F. (2012). The origins of the RNA world. *Cold Spring Harbor perspectives in biology*, 4(5).

[Ulitsky, 2016] Ulitsky, I. (2016). Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. *Nat Rev Genet*, 17(10).

[Ulitsky et al., 2011] Ulitsky, I., Shkumatava, A., Jan, C. H., Sive, H., and Bartel, D. P. (2011). Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell*, 147(7):1537–50.

[Venter et al., 2001] Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. a., Holt, R. a., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, a. G., Nadeau, J., McKusick, V. a., Zinder, N., Levine, a. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, a., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, a., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, a. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. a., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, a. K., Narayan, V. a., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, a., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, a., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C.,

Cravchik, a., Woodage, T., Ali, F., An, H., Awe, a., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, a., Center, a., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, a., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, a., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigó, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, a., Mi, H., Lazareva, B., Hatton, T., Narechania, a., Diemer, K., Muruganujan, a., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, a., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, a., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, a., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J.,

Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, a., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, a., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, a., Zandieh, a., and Zhu, X. (2001). The sequence of the human genome. *Science (New York, N.Y.)*, 291(5507):1304–51.

[Wahlestedt, 2013] Wahlestedt, C. (2013). Targeting long non-coding RNA to therapeutically upregulate gene expression. *Nature Reviews Drug Discovery*.

[Wang et al., 2015] Wang, M., Yuan, D., Tu, L., Gao, W., He, Y., Hu, H., Wang, P., Liu, N., Lindsey, K., and Zhang, X. (2015). Long noncoding RNAs and their proposed functions in fibre development of cotton (Gossypium spp.). *The New phytologist*, 207(4):1181–97.

[Wang and Brunet, 2002] Wang, W. and Brunet, F. (2002). Origin of sphinx, a young chimeric RNA gene in Drosophila melanogaster. *Proceedings of the ...*, 99(7):4448–53.

[Watson and Crick, 1953] Watson, J. D. and Crick, F. H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–8.

Appendix A

List of publications

1. Saint-Léger A., **Bello C.,** Dans PD., Torres AG., Novoa EM., Camacho N., Orozco M, Kondrashov F.A., and Ribas de Pouplana L.(2016). Saturation of recognition elements blocks evolution of new tRNA identities. Science Advances, 2(4).

2. Bello C and Kondrashov FA. Evolution of human-specific lncRNAs through exon duplication (*under review*).

3. Mateusz Konczal, Luis Zapata, Francisco Camara, Anna Vlasova, **Carla Bello**, Romain Derelle, Maria N. Tutukina, Maria Plyus cheva, Claudia Fontsere, Pavel S. Tomkovich, Nikolay N. Yakushev, Ivan A. Shepelev, Vladimir Yu. Arkhipov, Cristoph Zockler, Roland Digby, Egor Y. Loktionov, Elena G. Lappo., Tomás

Marqués, Roderic Guigó, Evgeny E. Syroechkovskiy, Fyodor A. Kondrashov. **Population genomics of the critically endangered Spoon-billed Sandpiper.** (*in preparation*)