Understanding the mechanisms of *de novo* gene evolution using transcriptomics data

Jorge Ruiz Orera

TESI DOCTORAL UPF / 2016

DIRECTORA DE LA TESI

Dra. Maria del Mar Albà Soler

Evolutionary Genomics Group

Research Programme on Biomedical Informatics (GRIB)

FIMIM (Fundació Hospital del Mar Medical Research Institute)

Universitat Pompeu Fabra

DEPARTMENT OF EXPERIMENTAL AND HEALTH SCIENCES

Universitat Pompeu Fabra Barcelona

A todos los que me han acompañado en este camino,
y especialmente a mi abuelo.

Acknowledgments

"Somewhere, something incredible is waiting to be known."
Carl Sagan

Llegado a este apartado, uno se da cuenta de cuánta gente me ha acompañado en estos últimos cuatro años de tesis. No puedo evitar sonreír al recordar esta etapa en la que numerosas personas han colaborado, ya sea de una forma científica u otra diferente. Por eso, pienso que toda esta gente se merece esta sección ya que, después de todo, han sido una parte indispensable para que yo haya llegado a este punto. De una forma más personal o global, intentaré no olvidarme de nadie. En cualquier caso, tengo una sincera palabra para todos vosotros: ¡Gracias!

Mar, moltes gràcies per haver-me donat aquesta oportunitat, per la teva confiança i per tot el que m'has ajudat durant aquests 4 anys com a supervisora del meu doctorat. Aquesta és possiblement l'etapa en la que més he après, no només sobre biologia evolutiva sinó sobre el que implica ser un investigador, una cosa que sempre he volgut conèixer i viure.

También quiero agradecer a todas las personas que, antes de empezar la tesis, me mostraron este apasionante campo de la Bioinformática. Especial mención a Javier Forment, mi primer profesor de Bioinformática en la carrera, así como a Ana Conesa, que fue mi supervisora en mi primer proyecto bioinformático y me ayudó a decidirme por este campo. Igualmente, gracias a mis profesores del Máster de Bioinformática de la UPF que me ayudaron a aprender mucho en esta materia.

Gracias a los compañeros y colaboradores de nuestro grupo de EG, me habéis ayudado muchísimo, especialmente en los comienzos. Magda, Núria, Steve, Cinta, Antonio, Isa, J.A. Subirana, Xavier, Alejandro… Gracias sobretodo a José Luis y Will, que han estado conmigo durante gran parte del doctorado y con los que he podido asistir a numerosos congresos. Vuestros comentarios y correcciones han sido muy valiosos para mis artículos y toda mi tesis.

Muchísimas gracias a toda la gente, más allá de nuestro grupo, que me ha acompañado en el despacho 486 o en el PRBB durante estos años. Muchas personas han ido y venido durante este tiempo, personas muy diferentes que han contribuido mucho a que me sintiera como en casa y con muchos de los cuales tengo una gran amistad: Juanra, Isaac, Caterina, Emma, Sabari, Carlota, Davide, Joan, Inés, Jaume, Jesse, Daniel, Elk, Bernat, Vicky, Amadís, Janet, Pau, JC, Héctor (mucha suerte en París), Miriam (gracias por la visita al frio Norte de Alemania), Juanlu, Babita… 'Vollywood' volley team, I hope you win a cup in 'Cracks' league next year! Y al equipo Champis/Boletus, muchas gracias por tanta diversión en los partidos, hemos mejorado mucho y, ¡tuvimos recompensa!

A Miguel y Alfons, nuestros IT guys, muchas gracias por toda la ayuda recibida y por tanta paciencia con las continuas sobrecargas del cluster.

I want to thank Diethard for providing me with the opportunity of doing an internship in the Max Planck Institute in Plön. I had the opportunity of working in a very beautiful town with an excellent research institute and

amazing people, thank you all. Rafik, muchas gracias por ser un inmejorable anfitrión y por mostrarme el trabajo tan interesante que habéis hecho allí. Mucha suerte en New York!

Por suerte, he podido conocer gente estupenda no sólo en el PRBB sino fuera de él. Muchísimas gracias a todos mis compañeros de máster y a la gente fantástica que he conocido durante mis años en Barcelona, especialmente a grandes amigos como Edu, Gael, Francisco, Miguel, Carlos, Chrissi, Isa, Guillaume, Celsa y Noelia, que siempre han estado cuando se les necesitaba, incluso a pesar de la distancia. Vielen Dank an die `Deutsche Vita` Freunden, ihr seid Super! Y sobretodo, muchas gracias Jesús, tengo la gran suerte de que mi 'hermano' viva en la misma ciudad que yo.

Muchas gracias a todos mis amigos de Teruel que conozco desde hace incontables años. Siempre ha sido especial volver a casa y encontrarme con ellos, o realizando escapadas por Europa con anécdotas de todo tipo.

Gracias a mis padres, a mi hermana, a mi abuela. Ellos me han apoyado desde que, hace ya 10 años, elegí el camino de estudiar fuera en algo tan desconocido y novedoso como era la Biotecnología. Si he llegado aquí es gracias a todo lo que me han transmitido ellos a lo largo de mi vida. Gracias a toda mi familia, y en especial gracias a mi abuelo, que no ha podido completar este camino conmigo pero que hubiera estado muy orgulloso de lo que he conseguido.

Thank you all so much!

Abstract

Genes contain essential information for the correct functioning and adaptation of the organism. The differences in the behavior or physiology of closely related species are highly connected to differences in their gene content. But, how do novel genes arise? For years, the major mechanism for gene birth was gene duplication and subsequent sequence divergence. However, recent comparative genomics studies have shown that some genes are originated *de novo* from previously non-functional genomic sequences, although the mechanisms involved are not yet fully understood. This thesis investigates the mechanisms for *de novo* gene origination and evolution using high-throughput sequencing of complete transcripts and ribosome-protected fragments. We have identified thousands of *de novo* genes in human and chimpanzee and obtained evidence that these genes are mostly expressed from recently arisen promoters. We have shown that a large number of poorly conserved genes, including genes previously believed to be non-coding, are translated. In addition, we have found a link between the capacity of a sequence to be translated and its nucleotide sequence composition. The analysis of polymorphic variants has revealed that many non-conserved peptides evolve neutrally and thus could be precursors of new genes. Taken together, the results show that there is abundant raw material for *de novo* birth of new functional proteins.

Resum

Els gens contenen informació que és essencial pel correcte funcionament i adaptació de l'organisme. Les diferències en el comportament i la fisiologia d'espècies properes estan molt lligades a les diferències en el contingut de gens. Però, com s'originen els gens? Durant anys, la duplicació gènica i la posterior divergència de seqüència va ser el mecanisme principal. Però estudis recents basats en la genòmica comparativa han posat de manifest que existeixen gens que s'originen *de novo* a partir de seqüències genòmiques no funcionals, encara que els mecanismes no es comprenen del tot bé. Aquesta tesi investiga els mecanismes de formació i evolució de gens *de novo* utilitzant seqüenciació massiva de transcrits complerts i de fragments protegits per ribosomes. Hem identificat milers de gens *de novo* en humans i ximpanzé i obtingut evidència de que aquests gens s'expressen majoritàriament a partir de promotors que han aparegut recentment en l'evolució. Hem demostrat que molts dels gens poc conservats, incloent gens que prèviament es creia que no eren codificants, es tradueixen. A més, hem observat que existeix una relació entre la capacitat que té una seqüència per ser traduïda i la seva composició nucleotídica. L'anàlisi de les variants polimòrfiques ha revelat que molts dels pèptids de ratolí no conservats en humans evolucionen de forma neutra i que, per tant, podrien ser precursors de nous gens. En conjunt, el material de partida per la formació de noves proteïnes funcionals és abundant en el genoma.

Preface

Science is fascinating. This sentence has been in my mind for most of my life, since I started to explore the world around me trying to understand and give explanations to everything, and realizing that every time more and more questions raised in my head. Thirst for knowledge made me become very interested in very different areas of natural science, to the extent that I turned my attention to the sky. I spent several years dedicating a part of my time to a very interesting hobby: Astronomy. It turned me to be more and more overwhelmed about the inmensity of the Universe. I cannot forget the moment in which I saw that photo of the Earth taken by the Voyager I, described as a Pale Blue Dot by Carl Sagan in one of his more illustrious quotes. Such ‘mote of dot suspended in a sunbeam’ had survived for thousands of millions of years in a hostile environment defined by the aggressive nature of the Universe, and it had given rise to an incredible variety of life forms that, so far, are unique in our vast Solar System. As many people, I was very enthusiast with the possibility of finding the answers about how life emerges and evolves, but I understood that the answer had to be in our own planet.

Such thoughts, that I kept with me during the next years, were fundamental to arrive to this point. Darwin's theory of Evolution opened a new world to me; I discovered that the possibility of tracing and understanding the history of the life in the Earth was in our hand. Moreover, bioinformatics revolutionized the field of the molecular evolution and made us possible to work with a high amount of data of

inestimable worth. I knew that my time to add my little grain of sand to the human knowledge had arrived, and I hope that this thesis starts to accomplish that purpose. It was not an easy task but I realized how hard work, learning from mistakes and, most importantly, enjoying the daily work are essential to be proud of the result. Working in *de novo* gene evolution has been a special challenge since it is a rather recent field that was previously unknown or underrated, but that nowadays it is bringing a new layer of information that might 'mind the gap' of the knowledge in the evolution of specific species and lineages. I will be always indebted to science and this is only the beginning of a path as uncertain as exciting.

Jorge Ruiz Orera
Barcelona, August 2016

Table of contents

Abbreviations

aa: Amino acid

bp: Base pair

cDNA: Complementary DNA

CDS: Annotated coding sequence

codRNA: Protein-coding RNA

FPKM: Fragments Per Kilobase per total Million mapped reads

lincRNA: Long intergenic non-coding RNA

lncRNA: Long non-coding RNA

NAT: Natural antisense transcript

NMD: Nonsense mediated decay

nt: Nucleotide

ORF: Open reading frame

PAS: Polyadenylated site

PN: Number of non-synonymous altering polymorphisms

PS: Number of synonymous altering polymorphisms

Ribo-seq: Ribosome profiling sequencing

RNA-seq: RNA sequencing

smORF: Small ORF (< 100 amino acid)

SNP: Single nucleotide polymorphism

TE: Translational efficiency

TEs: Transposable elements

TSS: Transcriptional start site

uORF: Upstream open reading frame

UTR: Untranslated region

1 INTRODUCTION

"Our own genomes carry the story of evolution, written in DNA, the language of molecular genetics, and the narrative is unmistakable."
Kenneth. R. Miller

1.1. Brief history of evolution and genetics

The word *evolution*, which today denotes the most widely accepted theory to explain how life proliferated and diversified on our planet, comes from the Latin words *evolvere* and *evolutio,* that describe the unrolling of a scroll (Bowler, 1989). In the 17th century, the English word was commonly used to define the process of unrolling, opening out, or revealing. Nevertheless, modern evolution as understood today did not start until 1859, when Charles Darwin published the book 'On the Origin of the Species', introducing for the first time the definition of 'natural selection' - the differential survival and reproduction of individuals due to differences in phenotype. Such explanation uncovered the mechanisms that drive evolution and that initiated in a single common ancestor (C. Darwin, 1859). The importance of the concept was noted simultaneously by Alfred Wallace (Wallace, 1858). By the end of the 18th century, *evolution* became the general term for a process of development[1],

1 Darwin refused to link his theory with the thought that the history of life was a simple chronological unrolling of a predetermined creative plan. However, such caution in avoiding this idea was futile.

especially when involving a gradual change from a simple to a more complex state.

During Darwin's time it was not known how traits were inherited in subsequent generations. An important advance was made when Gregor Mendel determined the laws of inheritance by selectively breeding pea plants; he tracked the segregation of traits and how they appeared in the offspring, and he recognized the mathematical patterns of inheritance in subsequent generations (Mendel, 1866)[2].

These discoveries brought about the concept of the *gene.* The etymology of 'gene' begins with the Greek word *genesis* or *genos*. This concept has since become one of the central themes in biology. Over centuries, humans have crossbreed animals and plants to select for advantageous traits, but the word 'gene' was not coined until 1909 by Wilhelm Johannsen (Johannsen, 1909). Twenty years earlier, Hugo De Vries defined the word *pangene* as the unit associated with the inheritance of specific traits (Vries, 1889), following Darwin's hypothetical mechanism of heredity, pangenesis (Charles Darwin, 1868).

Mendelian genetics and natural selection were fully consistent and derived in the modern evolutionary synthesis, which led to a new consensus to describe evolution: *"Evolution is change in the properties of populations of organisms over time, being population the unit of evolution"* (Mayr, 2002).

2 While Mendel is now recognized as the father of the genetics, his work was only appreciated once his study was rediscovered at the early 19th century.

1.1. BRIEF HISTORY OF EVOLUTION AND GENETICS

Even so, it was only in the 1940s that the gene was considered to be a blueprint for a protein. This was a result of the discovery of gene mutations that cause defects in steps of metabolic pathways (Beadle & Tatum, 1941). George Beadle and Edward Tatum created the “one gene, one enzyme” hypothesis, later referred as “one gene, one polypeptide”.

In the following decade, the discovery of the three-dimensional structure of DNA (Watson & Crick, 1953) revolutionized the field of molecular biology. The central dogma described how DNA is transcribed to a RNA molecule, denominated transcript, and how such transcripts are translated into proteins[3] (Crick, 1958). This new view of the gene was based on the sequence of the gene itself rather than on the physical locus responsible for a phenotype. Friedrich Vogel published his 'preliminary estimate' of the number of genes in the human genome in 1964 (Vogel, 1964), based on the number of amino acids in the alpha- and beta-chains of hemoglobin (141 and 146, respectively). He estimated that such proteins had an average size, and that the whole sequence of every chromosome was protein-coding. Vogel calculated the molecular weight of the human chromosomes and estimated a total number of 6.7 million human genes, a wildly exaggerated number which was based on a series of incorrect assumptions that were at that time reasonable.

Ten years later, the first gene from the bacteriophage MS2 was sequenced (Fiers et al., 1971, 1976) and it was discovered that a gene is spliced into exons that are joined and introns that are removed, and that a gene can use

3 Some exceptions were found, as ribosomal (rRNA) or transfer (tRNA) small RNAs that do not produce any proteins.

different exon-intron combinations to express different transcripts in a mechanism called alternative splicing (Berget, Moore, & Sharp, 1977). This mechanism was later found to be ubiquitous (Black, 2003).

The first genome was sequenced in 1977 by Fred Sanger yielding to a new genomics era. Such breakthrough felt on the Sanger sequencing method, which initially sequenced the ΦX174 bacteriophage containing 5,386 nucleotides (Sanger et al., 1977). Several species were sequenced over the next decades, such as *S. cerevisiae* (12.1 Mb, 1996), *E. coli* (4.6 Mb, 1997), *C. elegans* (97 Mb, 1998), and *D. melanogaster* (1.65 Gb, 2000).

In 1990, before the human genome was sequenced, the Human Genome Project reported an estimate of 100,000 human genes, based on a very rough calculation that human genes are 30,000 bases long, and that genes covered the entire 3 GB genome. Another historic main breakthrough in genomics took place in 2001 with the complete sequencing of the human genome (Lander et al., 2001; Venter et al., 2001). The number of genes decreased to 30,000-35,000, but once again these numbers were based on estimates that kept varying throughout the following years. Until 2008, several definitions were proposed requiring the translation of the protein as a key aspect of the gene (Lewin, 2007; Pearson, 2006; Wain et al., 2002). Graziano Pesole updated this definition to include all the RNA categories that are not translated into proteins (Pesole, 2008). Nowadays, a gene is defined as a DNA locus that encodes a functional RNA or protein product, which is the molecular unit of heredity (Slack, 2014).

1.1. BRIEF HISTORY OF EVOLUTION AND GENETICS

Four different mechanisms driving species evolution have been defined: mutation, selection, genetic drift, and migration. Mutations are the result of the misincorporation of nucleotides during DNA replication or the result of unrepaired DNA damage. This yields the required standing variation for natural selection to act. Because the genetic code is redundant, mutations in coding sequences might be classified as synonymous or non-synonymous, depending on whether the mutation alters the amino acid sequence. Other mutations might affect gene regulation by affecting promoter activity, exon-intron structure, or unbalancing the expression of different gene isoforms. Mendelian diseases are those caused by one or a few mutations and show relatively simple patterns of inheritance. In recent years, important efforts have also been made to identify mutations involved in complex diseases such as cancer (Sebestyén et al., 2016; Supek et al., 2014).

Several methods have been developed to look for signatures of selection at the molecular level. These methods can be based on the measurement of the divergence between homologous sequences (substitution-based methods) or between individuals of the same species (polymorphism-based methods) (Biswas & Akey, 2006; Jensen, Wong, & Aquadro, 2007). Substitution is the term generally employed to denote nucleotide differences in homologous sequences from the same or different species. By estimating the number of substitutions in coding sequences, it is possible to calculate the dN/dS ratio, which is the number of non-synonymous substitutions per non-synonymous site (dN) divided by the number of synonymous substitutions per synonymous site (dS). Such ratio can be used to examine if specific protein-coding sequences are subjected

to selection. A dN/dS of 1 corresponds to neutrality, while a ratio smaller than 1 indicates purifying selection or negative selection. Negative selection reflects functional constrains and hence dN/dS ratio are typically below 1 for most protein-coding genes. A dN/dS ratio over 1 indicates positive selection, advantageous variants increase in frequency until they fix in a population.

In population-based analyses the ratio PN/PS defines the relative abundance of non-synonymous (PN) and synonymous (PS) polymorphisms. Given the nature of the genetic code, there are more possible non-synonymous mutations than synonymous mutations. Under neutrality, the PN/PS ratio is expected to be approximately 2.89 (Nei & Gojobori, 1986), while lower ratios indicate purifying selection. Values of such ratios vary depending on the species, for example due to differences in the number of slightly deleterious mutations segregating in the population (Eyre-Walker, Woolfit, & Phelps, 2006).

"Volumes of history written in the ancient alphabet of G and C, A and T."
Sy Montgomery

1.2. Transcriptomics: from sequence to function

The transcriptome is defined as the set and quantity of transcripts for a specific developmental stage or physiological condition in a cell. Characterizing the transcriptome is essential to identify the functional elements of the genome and understand how they are regulated.

1.2.1. The advent of transcriptomics

Various technologies have been developed with the aim of cataloging and quantifying the transcriptome. Hybridization-based approaches are based on the incubation of fluorescently labeled complementary DNA (cDNA) in microarrays. Although inexpensive, they have important limitations: They rely on previous knowledge about the genes and the genome, cross-hybridization may produce high background signal levels (Okoniewski & Miller, 2006; Royce, Rozowsky, & Gerstein, 2007), and they have a limited dynamic range of detection. In contrast, sequence-based approaches can directly identify the cDNA sequence. The first technology was the Sanger sequencing of cDNA or expressed sequence tag (EST) libraries (Boguski, Tolstoshev, & Bassett, 1994; Gerhard, 2004). Later, tag-based methods as serial analysis of gene expression (SAGE) (Harbers & Carninci, 2005; Velculescu, Zhang, Vogelstein, & Kinzler, 1995), cap analysis of gene expression (CAGE) (Kodzius, 2006; Nakamura & Carninci, 2004; Shiraki, 2003), and massively parallel signature sequencing (MPSS) (Brenner, 2000; Peiffer, 2008; Reinartz, 2002) provided precise and high throughput quantitative data. However,

these were expensive technologies and a high fraction of the tags were not long enough to be uniquely mapped to the genome. Besides, it was impossible to distinguish between different transcripts expressed in a gene[4].

Over the last decade, whole-transcriptome sequencing using 'next-' or 'second-' generation sequencing (NGS) technologies, also know as RNA sequencing (RNA-seq), unleashed a set of revolutionary tools to reveal the complex landscape and dynamics of the transcriptome in different species with high sensitivity and accuracy (Marguerat & Bahler, 2010; Ozsolak & Milos, 2011; Z. Wang, Gerstein, & Snyder, 2009; B. T. Wilhelm & Landry, 2009). This resolved the main limitations of previous approaches. The first NGS platforms, including Illumina, SOLiD and 454, only produced very short reads (35-50 bp) (Metzker, 2010). Over time, read lengths have increased substantially and NGS is the current standard. However, except for small classes of RNAs, it is necessary to apply informatics tools to reconstruct a comprehensive transcriptome.

Most recently, third-generation sequencing uses advances in nanotechnology to process unique RNA molecules through a real time synthesis sequencing system like PacBio or NANOPORE. The latter moves the sample through nano-sensors that identify each residue in the DNA strand. However, such technologies still have high error rates and lower yields.

4 Despite those limitations, CAGE- and SAGE-like methods are yet employed to have a precise snapshot of the transcriptional start and end sites.

1.2.2. Assembling the transcriptome

Transcript reconstruction, consisting in the assembly of all expressed isoforms and genes in a transcriptome using RNA-seq datasets, usually has high computational requirements. Datasets are composed of dozens of millions of reads; transcripts might have different sequencing depth, varying by several orders of magnitude; they might overlap in different strands; and isoform variants from the same gene can share exons that are difficult to unambiguously resolve.

Two main strategies were developed to reconstruct the transcriptome (Figure 1.1), depending on whether the reference genome assembly is available. The combination of both strategies is also feasible and can lead to improved results (Bingxin, Zhenbing, & Tieliu, 2013).

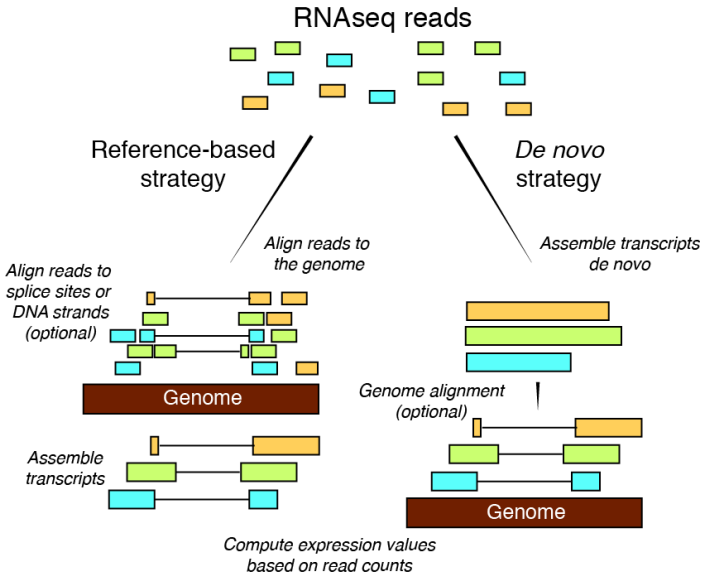

Figure 1.1. Overview of the two main methods for transcript assembly. Adapted from Haas & Zody, 2010.

-Reference-based strategy: This strategy involves mapping RNA-seq reads to an available reference genome using a splice-aware aligner, such as Tophat (C Trapnell, Pachter, & Salzberg, 2009), bwa (H. Li & Durbin, 2009), STAR (Dobin et al., 2012), or HISAT (D. Kim, Langmead, & Salzberg, 2015). Later, overlapping reads in each locus are merged into graph clusters that are traversed to represent all possible isoforms. Cufflinks (Cole Trapnell et al., 2010), Scripture (Guttman et al., 2010), and Stringtie (Pertea et al., 2015) are different softwares that resolve these clusters to build the final transcript assemblies.

-*De novo* strategy: This method does not require a reference genome and it is highly advisable when the quality of the reference genome is low since some regions might not be correctly assembled. It builds on the generation of k-mers from RNA-seq reads that are depicted as nodes in a De Bruijn graph[5]. Pairs of nodes are then connected when an overlap is found. This approach usually requires high computational power and a high sequencing depth, and it is more prone to sequencing errors and to the presence of chimeric molecules. Examples of *de novo* assemblers are Trinity (Grabherr et al., 2011) and Trans-ABySS (Robertson, 2010).

These two different strategies, especially when produced from shallow sequencing runs, may contain a significant fraction of partial transcript fragments. Hence, it is important to analyze the quality of the assembled transcriptome to discard these partial transcripts as well as alignment artifacts due to the presence of DNA contamination or multiple aligned

5 This directed graph is constructed by connecting pairs of k-mers with overlaps between the first k-1 nucleotides and the last k-1 nucleotides.

reads into repetitive regions (Weirick et al., 2015). Reference-based aligners use existing transcriptome assemblies to guide the reconstruction; even so, thousands of transcripts might not be previously annotated. Selecting a coverage or expression threshold or discarding short loci are usually the strategies to overcome such problems. Filtering out monoexonic transcripts can improve the specificity as well (Cabili et al., 2011), but would ignore the fraction of the transcriptome that is single-exon. Finally, software like TransRate analyze assembled transcriptomes to discard low-quality transcripts (Smith-Unna, Boursnell, Patro, Hibberd, & Kelly, 2015).

Moreover, the use of RNA-seq datasets with longer and/or paired-end reads lead to a better quality transcriptome, as does the use of strand-specific RNA-seq, which permits to identify the strand of each generated read. This information is used by the transcriptome assemblers to improve the assembly quality and the detection of overlapping transcripts in antisense orientation (Levin, 2010).

1.2.3. The pervasive transcription of the genome

After the sequencing of the human genome, the next big challenge was to locate and identify the functional elements including genes, transcripts, promoters, and other regulatory sequences. Large-scale studies were developed to identify full-length transcripts in a wide variety of tissues and cell types using the sequencing technologies present in the first decade of the 21st century, which were used to update the existing annotations in species like human and mouse (Babak, Blencowe, & Hughes, 2005; Carninci et al., 2005; Imanishi et al., 2004; Okazaki et al., 2002).

In a parallel effort, a public consortium named ENCODE (Encyclopedia of DNA elements) was launched in September of 2004 (T. E. P. Consortium, 2004). Initially, a set of regions representing approximately 1% of the genome was targeted for a pilot project and revealed the pervasive transcription of the genome, the relationship between transcriptional start sites (TSS) and specific regulatory sequences, and features of chromatin accessibility, structure and histone modification (T. E. P. Consortium et al., 2007).

Afterwards, the production scale-effort of the ENCODE project was launched (T. E. P. Consortium, 2011) and found that the 80% of the genome is involved in at least one biochemical RNA- and/or chromatin-associated event in at least one human cell type[6] (Dunham et al., 2012).

6 When using a conservative estimate, this percentage decreased to ~20%; in any case, both estimates are higher than the 8.2% of the genome that was detected as constrained in a different study (Rands, Meader, Ponting, & Lunter, 2014).

Moreover, ENCODE used high-throughput technologies to provide a genome-wide catalogue of human transcripts and to identify their subcellular localization, observing that 62.1-74.7% of the human genome is covered by transcripts, often overlapping and expressing many isoforms simultaneously (Djebali et al., 2012). In addition, other studies found that bidirectional transcription of promoters is widespread (Core, Waterfall, & Lis, 2008; Pickrell et al., 2010; Trinklein et al., 2004), being an inherent feature of most promoters (Wei, Pelechano, Järvelin, & Steinmetz, 2011).

A parallel study uncovered substantial conservation of the potential functional sequences in mouse and human, but also found a high divergence in sequences involved in transcriptional regulation or chromatin organization (Yue et al., 2014); and a high RNA expression diversity between humans and mice (S. Lin et al., 2014).

The observation of a pervasively and highly interleaved transcribed genome in such projects prompted the reconsideration of the definition of a gene: *"A transcript should be considered as the basic unit of inheritance whilst a gene would denote a different concept intended to include all the transcripts that contribute to a certain phenotypic trait"* (Djebali et al., 2012).

1.2.4. Deciphering the coding transcriptome

Until recently, distinguishing protein-coding RNAs and well-characterized non-coding RNAs such as transfer, ribosomal, nuclear, and nucleolar RNAs, was a straightforward procedure. Nevertheless, RNA-seq transcriptomics revealed the existence of a myriad of transcripts with low protein-coding potential and an unclear functional significance.

In order to experimentally check if a transcript has a protein product, high throughput tandem mass spectrometry (MS/MS) emerged as a method for the identification of peptides and proteins in a complex protein mixture (Choudhary, Blackstock, Creasy, & Cottrell, 2001; Keller, Nesvizhskii, Kolker, & Aebersold, 2002; Yates III, McCormack, Schieltz, Carmack, & Link, 1997). This approach generates thousands of peptide spectra with ion signatures that can be identified using a database search algorithm such as Sequest (Eng, McCormack, & Yates, 1994) or Mascot (Pappin, Creasy, & Cottrell, 1999). They assume that the peptide exists in the database and hence limit the analysis to known and predicted proteins.

The advent of RNA-seq permitted researchers to combine transcriptomics and mass-spectrometry data in an approach denominated 'proteogenomics' that expands protein sequence databases to identify novel peptides in transcripts with unknown coding potential (M.-S.-. S. Kim et al., 2014; Nesvizhskii, 2014; M. Wilhelm et al., 2014) and in alternative frames from protein-coding genes (Vanderperre et al., 2013). However, short proteins are usually difficult to identify by mass-spectrometry experiments (Slavoff et al., 2013) as well as rare and short-lived proteins that are rapidly degraded in purification procedures (Fälth et

al., 2006) so most short and recent proteins are expected to lack experimental evidence. Even so, some modifications in these approaches led to the detection of several human short proteins (Ma et al., 2014, 2016; Oyama et al., 2007; Schwaid et al., 2013; Slavoff et al., 2013) including MRI-2, a 69 aa protein isoform that stimulates end-joining DNA repair (Slavoff, Heo, Budnik, Hanakahi, & Saghatelian, 2014).

There are several methods to predict the protein-coding potential of a sequence. A coding statistic is a function that, for a DNA sequence, calculates a real number related to the likelihood that the sequence is coding for a protein (R. Guigo, 1999). Most of the methods that have been developed measure codon or dicodon usage bias, base compositional bias between codon positions, or periodicity in base occurrence (Fickett & Tung, 1992). Codon usage has been postulated to be the result of selection, mutation and genetic drift, although the relative contribution of each process is not clear (Akashi, 1997; Duret, 2002; Hershberg & Petrov, 2008; Ziheng Yang & Nielsen, 2008). These biases, which are associated with the GC content and the evolutionary age of genes (Prat, Fromer, Linial, & Linial, 2009), help distinguish between protein-coding genes and non-coding regions (Toll-Riera et al., 2009). Sequence-based statistics, combined with open reading frame (ORF) length and cross-species conservation, are employed in diverse computational tools such as Coding Potential Calculator (CPC) (Kong et al., 2007), Coding Potential Assessment Tool (CPAT) (L. Wang et al., 2013), or PhyloCSF (M. F. Lin et al., 2007). Such approaches are based on well-defined properties of conserved protein-coding genes and they filter out most of the spurious ORFs that randomly appear in long non-coding and intergenic regions, but

do not have a good sensitivity for detecting young or short proteins (Dinger, Pang, Mercer, & Mattick, 2008). So far, only a few functional small proteins (smORFs) with a length shorter than 100 aa have been functionally characterized. Examples include *Humanine* (HN), a 24 aa peptide that acts as a neuron death suppressor (Zapała et al., 2010), *myoregulin* (MLN), a 46 aa protein that regulates muscle performance (Anderson et al., 2015), and *hemotin/stannin* (SNN), an 88 aa conserved protein acting as a regulator of phagocytosis (Pueyo et al., 2016).

The number of protein-coding genes annotated in Ensembl has decreased over the years (Figure 1.2), probably due to the use of more conservative definition criteria. The current number of annotated protein-coding genes is around 20,000, but it may drop to ~19,000 genes when only considering long and conserved protein-coding genes that are more likely to be detected by proteomics experiments (Ezkurdia et al., 2014).

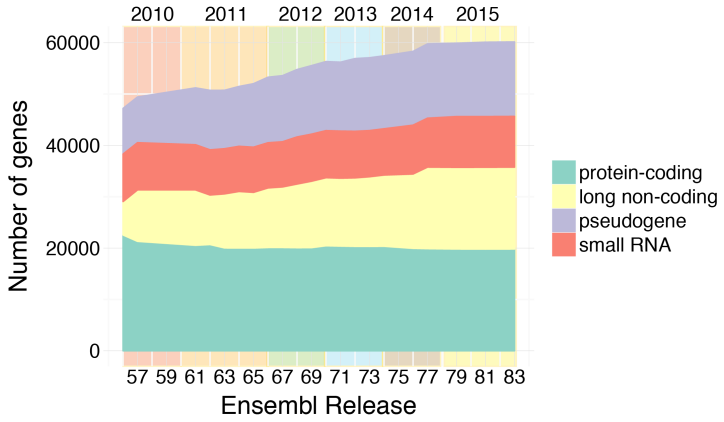

Figure 1.2. Number of annotated genes in different Ensembl releases (57-83) for different biotypes. Numbers were extracted from the primary assemblies in Ensembl (Flicek, Ahmed, et al., 2012) and GENCODE (Harrow et al., 2012) databases.

1.2.5. Ribosome profiling deciphers the translatome

Deep sequencing of ribosome-protected fragments, or ribosome profiling (Ribo-seq), has recently emerged as a technique that takes advantage of modern sequencing technologies to perform a global scale analysis of the regions that are translated in the transcriptome with the same precision and using similar pipelines to RNA-seq (Ingolia, Ghaemmaghami, Newman, & Weissman, 2009). Ribo-seq has been used to annotate translated sequences, to decipher the mechanisms of protein synthesis, and to study the translational control of gene expression (Ingolia, 2014).

This technique unveiled the presence of numerous small upstream ORFs (uORFs or leader peptides[7]) in protein-coding genes (Ingolia et al., 2009), which may act as regulators of translation or RNA levels (Johnstone, Bazzini, & Giraldez, 2016; Juntawong, Girke, Bazin, & Bailey-Serres, 2014). Small ORFs with similar features as uORFs were found in non-coding genes (Chew et al., 2013; Ingolia et al., 2014; Ingolia, Lareau, & Weissman, 2011; Juntawong et al., 2014) and, although some of them did not resemble coding ORFs (Guttman, Russell, Ingolia, Weissman, & Lander, 2013), a high fraction showed true hallmarks of translation in humans (Ingolia et al., 2014; Ji, Song, Regev, & Struhl, 2015; Ruiz-Orera, Messeguer, Subirana, & Alba, 2014), mammals (Bazzini et al., 2014) and yeast (Smith et al., 2014). Since reads are mapped with single nucleotide resolution, software like RiboTaper (Calviello et al., 2016) and RiboORF (Ji et al., 2015) can determine if reads span the correct frame and follow a 3-periodicity distribution as an indicator of true translation (Figure 1.3).

7 Leader peptides were originally discovered on the basis of their impact on the regulation of genes involved in the synthesis or transport of amino acids.

Hence, ribosome profiling has emerged as a method to detect ORFs that are not covered by proteomics techniques or coding sequence prediction software. Although several uORFs are regulatory and are expected to produce non-functional peptides, other small ORFs might translate functional peptides.

Moreover, this technique has shown that polycistronic translation is probably quite common, as it was previously hypothesized (Tautz, 2009). This occurs in the gene *pri* in fruit fly, which regulates tarsal development (Galindo, Pueyo, Fouix, Bishop, & Couso, 2007) and translates several small redundant ORFs (Kondo et al., 2007), the gene family *mlpt* in insects, that also regulates development (Savard, Marques-Souza, Aranda, & Tautz, 2006), and *meloe,* a gene that translates three antigenic peptides of different immunogenicity in melanoma cells (Charpentier et al., 2016; Godet et al., 2008).

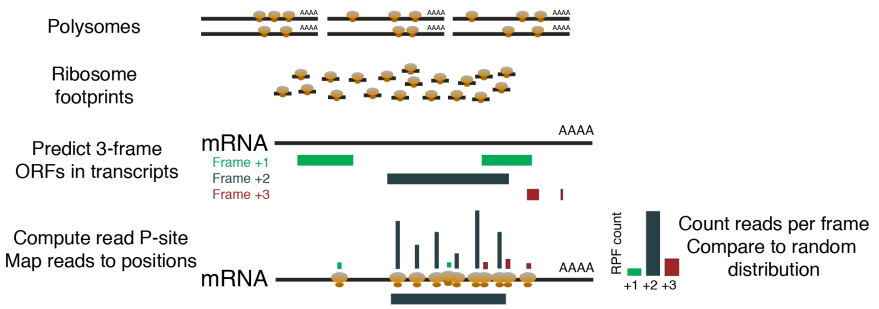

Figure 1.3. Representation of the ribosome profiling analysis. Ribosome-protected fragments are sequenced and mapped to the predicted ORFs in transcripts. Initiation and termination positions in the ORF are used as a training to decipher the exact read nucleotide that corresponds to the P-site of the ribosome. Bars represent the density of mapped reads to the three possible frames; in this example, the blue ORF is translated since it is sufficiently covered by ribosomes and the fraction of reads spanning the correct frame (+2) is higher than randomly expected (1/3 of reads in each frame).

1.2.6. Characterizing non-coding transcription

While the exact number of protein-coding genes is a subject of debate, it is clear that there are many transcripts that do not encode proteins. Small non-coding RNAs are a well-defined category of short transcripts (< 200 nucleotides) that include tRNAs, rRNAs, miRNAs, snRNAs and snoRNAs. There are also thousands of annotated long non-coding RNAs (lncRNAs)[8] longer than 200 nucleotides and expressed in different cell lines or tissues, many of them without a known function (Cabili et al., 2011; Iyer et al., 2015; Kazemian et al., 2015; Jun Liu et al., 2012; Necsulea et al., 2014; Okazaki et al., 2002; Pauli et al., 2012; Ponting, Oliver, & Reik, 2009; Ulitsky & Bartel, 2013).

Long intergenic non-coding RNAs (lincRNA) are the most abundant lncRNA class annotated in Ensembl and GENCODE. It includes all expressed non-overlapping transcripts without significant coding potential. A second category of non-coding RNAs is composed of the transcripts that overlap protein-coding genes, denominated Natural Antisense Transcripts (NATs).

LincRNAs, antisense RNAs and other secondary classes of non-coding RNAs[9] are defined based on their genomic location. Compared to well-defined protein-coding genes, lncRNAs are poorly expressed and often tissue-specific; some of them are localized in the nucleus but most of them

8 Unlike protein-coding and small RNAs, these genes are usually defined by negative descriptors and therefore have diverse properties and mechanisms of action.

9 Ensembl has different structural biotypes of non-coding RNAs as processed transcripts, sense intronic genes, sense overlapping genes, or TECs.

are polyadenylated and exported to the cytoplasm, and they tend to evolve very rapidly (Derrien et al., 2012; Heesch et al., 2014; Kutter et al., 2012; Necsulea et al., 2014; Ulitsky & Bartel, 2013).

Because lncRNAs are poorly conserved and most of them have no known functions, it has been hypothesized that they are not functional and are just spurious by-products of the noisy transcriptional machinery (Struhl, 2007). It has been estimated that the cost of transcription in multicellular organisms is probably too low for selection to counteract (Lynch & Marinov, 2015). Thus, it may be more costly to maintain mechanisms of control than simply tolerating some level of non-functional genomic transcription. On the other hand, transcription itself can produce changes in the chromatin- or transcription-factor-binding landscape resulting in the activation or repression of other genes in *cis* or *trans*. In the second case, those RNAs might be under pressure to conserve the structure or the promoter but not the sequence. Indeed, promoters are often more conserved than exon sequences in lncRNAs (Cabili et al., 2011; Derrien et al., 2012; Guttman et al., 2009; Necsulea et al., 2014).

Besides, some studies reported a significant number of lncRNAs showing purifying selection in their genomic loci, including splice sites and secondary structure motifs (Guttman et al., 2009; Haerty & Ponting, 2015; Nitsche, Rose, Fasold, Reiche, & Stadler, 2015; Pegueroles et al., 2016; Ponjavic, Ponting, & Lunter, 2007).

Necsulea et al. reconstructed homologous families of lncRNAs and detected that 3% originated more than 300 Myr ago and showed similar

conservation patterns as protein-coding genes (Necsulea et al., 2014). Sequence conservation may be explained by the role of some lncRNAs in post-transcriptional regulation by antisense base complementarity, by competing for miRNAs, circRNAs or mRNAs, or by hosting small RNAs in its sequence, or by acting as molecular scaffolds of proteins (Guttman & Rinn, 2012; Kung, Colognori, & Lee, 2013). For instance, TERC is a component of the telomerase complex and it has a conserved structure that serves as a template for telomere replication (Theimer CA, Blois CA, 2005). It is also possible that some of the lncRNAs analyzed in these studies are miss-annotated protein-coding genes. It has been shown that a significant fraction of annotated lncRNAs show ribosome protection patterns that are consistent with translation. In general, these transcripts contain ORFs which are shorter than 100 aa and often disregarded as possible protein-coding sequences.

There are some cases of nuclear RNAs involved in different well-defined non-coding functions (Figure 1.4). X-inactive specific transcript (*Xist*), *Kcnq1* overlapping transcript 1 (*Kcnq1ot1*) and *Airn* (antisense Igf2r (insulin-like growth factor 2 receptor) induce the formation of repressive chromatin in *cis* and they are involved in dosage compensation and genomic imprinting (J. T. Lee & Bartolomei, 2013). *Jpx* is a transcript that competes with CTCF, a transcriptional repressor that inhibits *Xist* promoter (S. Sun et al., 2013; Tian, Sun, & Lee, 2010). *HOXA* is a gene from the *HOX* gene family[10] that produces a myriad of non-coding RNAs (Rinn et al., 2007), some of which are functional antisense loci. First,

10 HOX genes are an evolutionary conserved family of transcription factors that regulate embryo development and cell specification in several adult differentiation processes. In mammals, this family is composed of 39 HOX genes.

distal transcript antisense RNA (*HOTTIP*) functions through the *cis*-recruitment of the MLL1 complex, which drives the formation of the activating histone H3K4me3 (K. C. Wang et al., 2011). Second, *HOXA* transcript antisense RNA (*HOTAIR*) is a *trans*-acting regulator of the *HOXD* genes by forming a scaffold that recruits two different repressor complexes (Rinn et al., 2007; Tsai et al., 2010). The RNAs *lncRNA-ES1* and *lncRNA-ES2* regulate embryonic stem cell pluripotency by associating with the transcription factor sex-determining region Y-box 2 (SOX2) and PRC2 (Ng, Johnson, & Stanton, 2012).

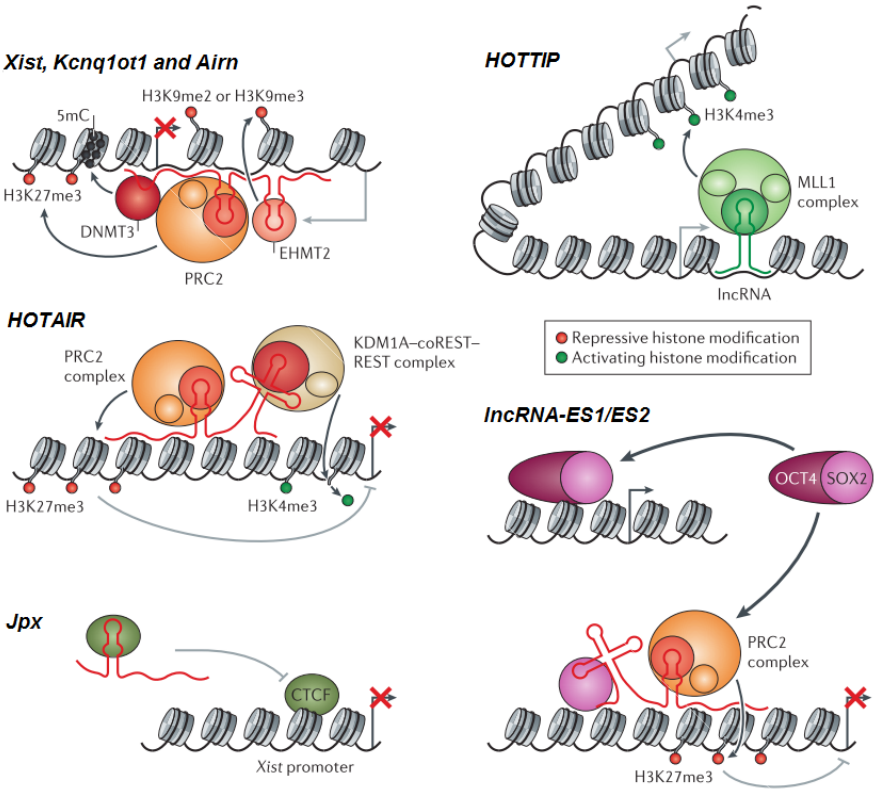

Figure 1.4. Examples of genes with nuclear non-coding functions. Modified from Fatica & Bozzoni, 2014.

Besides, there are different models by which lncRNAs can affect the stability or translation of other transcripts in the cytoplasm by base-pairing of complementary regions between the non-coding RNA and the RNA target (Figure 1.5).

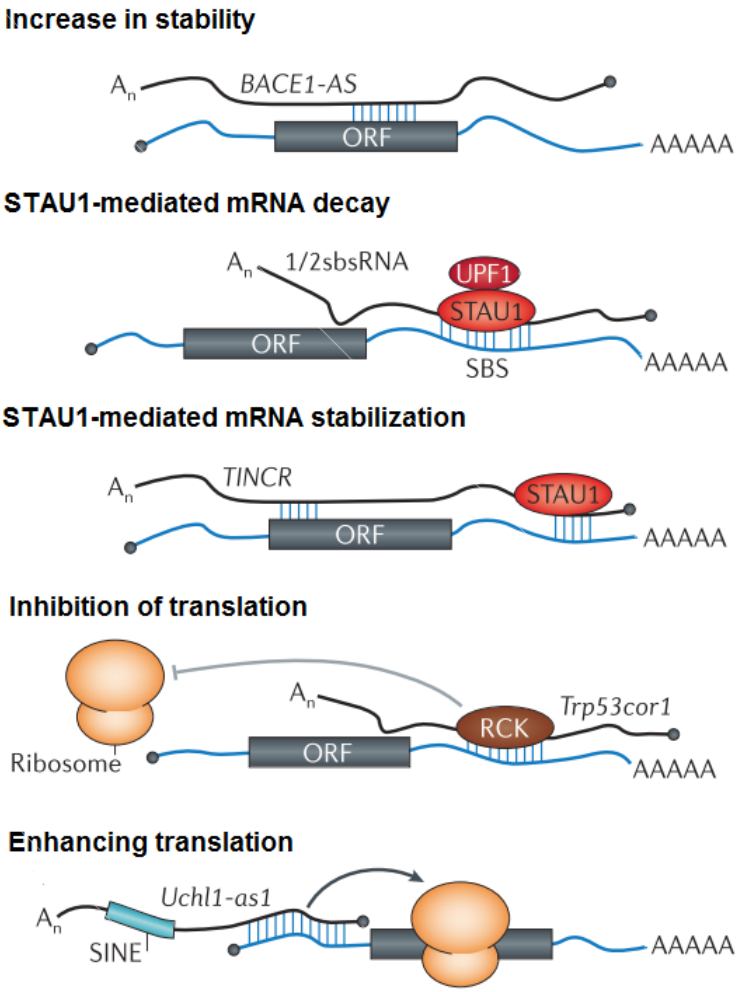

Figure 1.5. Models of cytoplasmatic RNA non-coding functions through base-pairing of complementary functions. Modified from Fatica & Bozzoni, 2014.

BACE1 interacts with an antisense transcript (*BACE1-AS*) in specific regions of the sense gene stabilizing it and increasing protein expression (Faghihi et al., 2008). Staufen double-stranded RNA-binding protein 1 (STAU1)-mediated mRNA decay (*1/2sbsRNA*) or stabilization (*TINCR*) is induced when base pairing is formed in an Alu or SINE element in the 3′ UTR of a mRNA as well as in other required motifs (Gong & Maquat, 2011; Kretz et al., 2013; Jiashi Wang, Gong, & Maquat, 2013). RNA pairing can also control translation. A repressive effect on translation was shown for the targets of tumour protein p53 pathway corepressor 1 (*Trp53cor1*) RNA (Yoon et al., 2012). Conversely, *Uchl1-as1* is a member of an antisense class of transcripts known as SINEUPs, whose activity requires an embedded inverted SINEB2 sequence to increase translation and for the overlapping region to target the protein-coding gene (Carrieri et al., 2012; Zucchelli et al., 2015), although other examples with a different architecture are functional as well (Tran et al., 2016).

Regardless of the observation that some lncRNAs can translate proteins, the list of biological events where lncRNAs play major roles as regulatory molecules is quickly growing. The described biological processes include cell-cycle regulation, apoptosis, lineage differentiation, and organogenesis (Grote & Herrmann, 2015; Pauli et al., 2012; Pickard & Williams, 2015; Ponting et al., 2009; Rinn & Chang, 2012). Dysregulation of lncRNAs is linked to several human diseases and cancers (Bhan & Mandal, 2014; Du et al., 2013; Gibb, Brown, & Lam, 2011; Mitra, Mitra, & Triche, 2012; Zhi et al., 2014) and hence they are potential biomarkers and therapeutic targets (Saus et al., 2016; Tsai, Spitale, & Chang, 2011).

> "All of today's DNA, strung through all the cells of the earth,
> is simply an extension and elaboration of [the] first molecule."
> **Lewis Thomas**

1.3. The origin of new genes

One of the long-standing questions in evolutionary biology has been how new genes emerge in genomes and how they evolve over time. A well-standing process is gene duplication followed by sequence divergence (Ohno, 1970). For a long period of time, gene duplication was the only accepted mechanism for new gene formation. In 1977, François Jacob stated: *"The probability that a functional protein would appear de novo by random association of amino acids is practically zero"* (Jacob, 1977). Later, other mechanisms that involved the reuse of DNA or RNA were also considered (reviewed in Long, Betran, Thornton, & Wang, 2003). More recently, several studies have described how completely new gene sequences can arise *de novo* (Figure 1.6) (McLysaght & Hurst, 2016; reviewed in Tautz & Domazet-Lošo, 2011). While all these mechanisms were initially discovered in protein-coding genes, they also explain how non-coding genes can arise in genomes (Kaessmann, 2010).

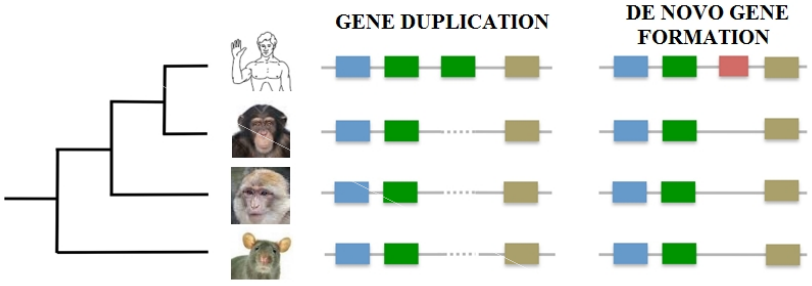

Figure 1.6. Comparison of gene duplication and *de novo* gene formation. In the latter mechanism, one new gene starts to be expressed from a DNA region in a single species. Thus, the gene has no paralogs or orthologs in other species.

1.3.1. Gene duplication

Gene duplication is the mechanism by which one DNA region[11] containing a gene is duplicated by a process involving DNA recombination and gives rise to a new gene that is called a paralog. It is the major force for new gene origination (Magadum, Banerjee, Murugan, Gangapur, & Ravikesavan, 2013) and there are numerous examples of gene families expanded through gene duplication; for instance the human and mouse olfactory receptors (Gilad, Man, & Glusman, 2005; Waterston et al., 2002).

John Haldane suggested that duplicated genes could be a source of novelty, given the redundancy of having two or more copies of a gene in a genome (JBS, 1932). Susumu Ohno proposed that, unless having two copies is advantageous for the organism, one of the genes might acquire a completely new function (neofunctionalization) (Ohno, 1970). Years later, a different outcome was proposed, in which both genes might evolve to retain different subfunctions of the original one (subfunctionalization) (Force et al., 1999; Stoltzfus, 1999).

11 This is a very generic definition, since a duplication can be genomic, segmental (> 1 kb DNA) or genic. In the latter case, the whole gene or just a partial region can be duplicated.

1.3.2. Other sources of new genes

Other mechanisms for new gene formation involving the reuse of DNA/RNA elements have been proposed in more recent years. These may occur separately or jointly.

-Retroposition: RNA-based duplication. A DNA copy is created through reverse transcription of a RNA transcript and inserted into the genome. These copies are usually pseudogenized since they lack regulatory sequences and introns (Maestre, Tchénio, Dhellin, & Heidmann, 1995) and are classified as processed pseudogenes in databases. Nevertheless, some retrocopies might adopt nearby promoters and show new expression patterns and functions (Kaessmann, 2010; Kaessmann, Vinckenbosch, & Long, 2009; Long et al., 2003). Examples of retroposition-based genes have been found in human, most of them expressed in testis and involved in spermatogenesis (Lahn & Page, 1999; Marques, Dupanloup, Vinckenbosch, Reymond, & Kaessmann, 2005).

-Exaptation of Transposable Elements: Use of DNA/RNA mobile elements such as SINEs, LINEs, or LTRs, as parts of genes. Several primate-specific protein-coding (Toll-Riera et al., 2009) and especially non-coding genes (Kapusta et al., 2013) have a significant fraction of the sequence covered by transposable elements. They may drive novel functions (Nekrutenko & Li, 2001) or rewire new sequence elements in conserved non-coding RNAs (Hezroni et al., 2015).

-Horizontal gene transfer: Process involving any occurrence of heritable material passing between organisms, asynchronous with their

reproduction (Heinemann & Bungard, 2006). It is more frequent in prokaryota (Boucher et al., 2003). Moreover, there are cases of lateral transfer described in associations of parasites and endosymbionts (Conaco et al., 2016; Hotopp et al., 2007).

-Gene fusion or fission: Gene fusion refers to the formation of a hybrid gene formed from two previously separate genes. *Kua-UEV*, for instance, is a single human gene that is expressed as two separate loci in insects (Thomson et al., 2000). *Tre2* is a chimeric gene which resulted from the fusion of a highly conserved gene and a recent segmental duplication (Paulding, Ruvolo, & Haber, 2003). Gene fission refers to the split of one gene into two different genes, as it occured with the *MITF* gene that is expressed as two different genes in some fish species (Altschmied et al., 2002). Both mechanisms are major contributors to the evolution of multi-domain proteins in bacteria (Pasek, Risler, & Brézellec, 2006).

-Overprinting: In this process, an alternative open reading frame acquires the capacity to be translated. In contrast to gene duplication, the new protein has a completely different sequence and therefore a novel function (Ohno, 1984). Initially, newly translated frames will be random byproducts of a gene, which may eventually acquire a function. Overprinted genes have been found in viruses (Carter et al., 2013; Pavesi, Magiorkinis, & Karlin, 2013), bacteria (Delaye, Deluna, Lazcano, & Becerra, 2008; Fellner et al., 2015), and eukaryotes (Chung, Wadhawan, Pond, & Nekrutenko, 2007; R Neme & Tautz, 2013). They constitute a strong argument for the viability of genes with completely new sequences, and thus for *de novo* gene emergence (Neme & Tautz, 2013).

1.3.3. The continuous emergence of new genes

The full sequencing of the chromosome 3 in yeast was completed in 1992 (Oliver et al., 1992) and four years later Bernard Dujon discussed 'the mystery of the orphans', following the observation that almost half of the protein-coding ORFs had no clear homologs in other organisms (Dujon, 1996). These genes were called ORFans or orfans, which implies that they should have parent genes that were somehow missing. Fischer and Eisenberg confirmed that the ORFans were a real phenomenon in prokaryotes (Fischer & Eisenberg, 1999). A different study confirmed that such genes evolve fast (Schmid & Tautz, 1997). However, it was thought that the parents of these genes would eventually be found with the sequencing of new genomes.

Orphan genes were continuously discovered in newly sequenced genomes and a new definition was coined: “taxonomically restricted genes” (TRGs) (Domazet-Loso & Tautz, 2003; Khalturin, Hemmrich, Fraune, Augustin, & Bosch, 2009; Toll-Riera et al., 2009). With this term, it was recognized that genes had been born at different time points across the evolutionary history of life. These genes might be linked to the emergence of lineage-specific adaptations. The idea of new genes arising relatively fast and slowing down as they became functional led to the development of a procedure called “phylostratigraphy” to study the phylogenetic age of the genes by comparing the presence of homologs in different species (Domazet-Loso, Brajkovic, & Tautz, 2007).

The classification of genes in different conservation levels revealed an inverse relationship between gene age and evolutionary rate (Albà &

Castresana, 2005). While this pattern could in principle be due to a lack of sensitivity of BLAST for fast evolving proteins (Elhaik, Sabath, & Graur, 2006), sequence evolution simulations along a phylogenetic tree revealed that the percentage of error for proteins is relatively small (4.7%) even when performing searches from mammals to fungi or plants (Albà & Castresana, 2007). Besides, this issue practically disappears at short evolutionary distances. For instance, in human and chimpanzee comparisons the percentage of nucleotide differences in neutrally evolving regions is only about 7%, so it should be possible to detect all homologous genes if they existed. In addition, other methodologies have been developed to improve the detection of homologous genes. For example, the alignment of domain arrangements can find protein homologs missed by other comparison methods (Terrapon, Weiner, Grath, Moore, & Bornberg-Bauer, 2014).

TRGs were initially proposed to evolve by a model of gene duplication followed by a fast evolving phase, which would explain the loss of similarity to parental genes and the absence of homologous sequence matches (Domazet-Loso & Tautz, 2003). However, species- and lineage-specific genes continued to be detected when many more genomes became available. This is better explained by *de novo* gene formation than by gene duplication (Neme & Tautz, 2013). The analysis of genomic syntenic regions across closely related species provided additional evidence for the birth of genes from previously non-genic regions (Begun, Lindfors, Kern, & Jones, 2007; J. Cai, Zhao, Jiang, & Wang, 2008; Heinen, Staubach, Häming, & Tautz, 2009; Knowles & Mclysaght, 2009; Toll-Riera et al., 2009).

1.3.4. *De novo* gene origination

The first *de novo* genes were described in Drosophila using genomic comparisions and gene expression analyses (Begun et al., 2007; Begun, Lindfors, Thompson, & Holloway, 2006; S.-T. Chen, Cheng, Barbash, & Yang, 2007; Levine, Jones, Kern, Lindfors, & Begun, 2006). In 2009, Toll-Riera et al. concluded that at least 5.5% of primate-specific genes had emerged out of non-coding DNA (Toll-Riera et al., 2009). Over the next few years, *de novo* genes were found in many other eukaryotic species or lineages with complete proteome or transcriptome information (Table 1.1). Moreover, *de novo* genes have been found in bacteria as well. A recent study has reported the presence of 72 *de novo* genes that were previously unannotated in *E. Coli*. These genes are also translated according to ribosome profiling data and 7 of them have additional evidence by mass-spectrometry (Neuhaus et al., 2016).

The number of *de novo* genes with well-characterized functions is still quite limited, but it is likely to increase in the future. *MDF1*, a *de novo* gene discovered in yeast, binds to two different proteins to suppress mating and promote vegetative growth, thus conferring a selective advantage in this species (D. Li et al., 2010; D. Li, Yan, Lu, Jiang, & Wang, 2014). *BSC4* is another *de novo* gene from yeast involved in DNA repair and that has evidence of peptides from mass spectrometry (J. Cai et al., 2008). *QQS*, a *de novo* gene found in Arabidopsis, is a regulator of starch biosynthesis and it modulates the allocation of carbon and nitrogen (L. Li et al., 2009; L. Li & Wurtele, 2015).

	Number	Gene dataset	Expression analysis	Translation analysis	References
D. melanogaster	5	Annotations	ESTs		Levine et al. 2006
D. yakuba / D. erecta	7 + 3	ESTs	ESTs		Begun et al. 2006
D. yakuba	11	ESTs	ESTs		Begun et al. 2007
D. melanogaster	14	ESTs + annotations	ESTs		Zhou et al. 2008
Drosophila	104	Annotations			Zhang et al. 2010
D. melanogaster	16	Annotations			Chen et al. 2010
D. melanogaster	248	RNA-seq transcriptome	RNAseq		Zhao et al. 2014
A. Thaliana	534	Annotations			Donoghue et al. 2011
Brassicaceae	25	Annotations			Donoghue et al. 2011
Hominoid	15	Annotations			Toll-Riera et al. 2009
H. Sapiens	3	Annotations		Proteomics	Knowles et al. 2009
H. Sapiens	60	Annotations			Wu et al. 2011
Hominoid	24	Annotations	RNAseq		Xie et al. 2012
Hominoid	56 (64)	Annotations	RNAseq	Proteomics	Chen et al. 2015
Hominoid	21	Annotations + RNA-seq transcriptome	RNAseq	Proteomics + RiboSeq	Ruiz-Orera et al. 2015
Hominoid	35	Annotations	RNAseq	Proteomics	Guerzoni et al. 2016
Murine	75	Annotations			Murphy et al. 2012
S. cerevisiae	1139	ORF prediction	RNAseq	RiboSeq	Carvunis et al. 2012
P. vivax	13	Annotations			Yang et al. 2011
Insects	230 – 13,181	Annotations	RNAseq + ESTs		Wissler et al. 2013

Table 1.1. Global studies identifying recently originated *de novo* protein-coding genes in different eukaryotic lineages: description of the initial gene datasets, and expression and translation analyses performed on these genes.

PBOV1 is a primate-specific *de novo* gene overexpressed in prostate and breast cancer which increases cell proliferation through promoting G1/S transition (Pan et al., 2016; Samusik, Krukovskaya, Meln, Shilov, & Kozlov, 2013). *C20orf203* is a primate-specific gene under purifying selection and it is expressed in brain and overexpressed in Alzheimer's disease (C.-Y. Li et al., 2010). *NCYM* is a gene conserved in human and chimpanzee that acts as an oncopromoting factor in human cancer stabilizing its antisense pair and thus probably acting as a bi-functional

RNA (Suenaga et al., 2014; Vadie et al., n.d.). Finally, *TDRG1* is a primate-specific gene which is overexpressed in testis that promotes the development and migration of seminoma cells via the regulation of the PI3K/Akt/mTOR signaling pathway (Jiang et al., 2011; Y. Wang et al., 2016). Additional *de novo* translated genes with unknown function were found in other studies: *DNAH10OS, CLLU1*[12], and *C22orf45* are human-specific genes with evidence of translation based on the presence of different peptides (Knowles & Mclysaght, 2009). *RDT1* is a yeast gene with a small 28 aa ORF that is translated according to ribosome profiling signatures (Wilson & Masel, 2011). Finally, several *de novo* genes with no evidence of translation have been identified as well. In 2009, it was demonstrated that a mouse gene, *Poldi*, arose *de novo* within the past 2.5-3.5 million from an intergenic region. The gene knockout resulted in reduced sperm motility and reduced testis weight (Heinen et al., 2009). Other examples are *Hydra* in fruit fly (S.-T. Chen et al., 2007), *OsDR10* in rice (Xiao et al., 2009), and *ESRG* in human (Jichang Wang et al., 2014).

De novo genes differ significantly from non-coding sequences in that they show interspecific and intraspecific purifying selection signatures (Carvunis et al., 2012; C.-Y. Li et al., 2010; Palmieri, Kosiol, & Schlötterer, 2014), indicating that a significant fraction of them is probably functional. Nonetheless, because *de novo* proteins are usually short and not conserved across species, many of them are likely to be missing from the current gene catalogs, as it occurs with short proteins in general. For example, the majority of *de novo* protein-coding genes reported in human in different studies are not present in the latest Ensembl

12 CLLU1 is a gene highly expressed in chronic lymphocytic leukemia.

version. Additionally, these genes are generally unstable across different dataset versions. Only 14 of them are annotated as protein-coding genes in Ensembl v.83, and 39 have been re-annotated as non-coding genes (Figure 1.7).

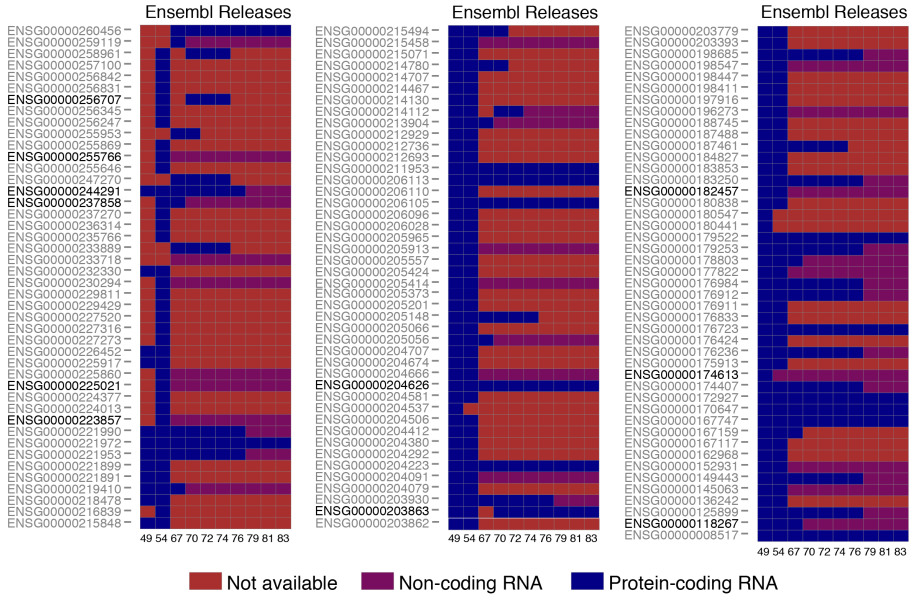

Figure 1.7. Evolution of Ensembl biotypes for the 133 human-/hominoid-protein-coding *de novo* genes reported in different studies (J.-Y. Chen et al., 2015; Guerzoni & McLysaght, 2016; Jiang et al., 2011; Knowles & Mclysaght, 2009; C.-Y. Li et al., 2010; Suenaga et al., 2014; Toll-Riera et al., 2009; D.-D. Wu, Irwin, & Zhang, 2011; Xie et al., 2012). Ensembl releases 49,67, 74, and 83 were available at 2008, 2012, 2013, and 2015, respectively. Not available: Cases where the ID is not found in that Ensembl version, and no other similar genes are found in the same region. Cases in bold were assigned to a distinct ID in certain Ensembl releases.

"Heritability pertains to the entirety of the genome, not to a single gene."
Steve Pinker

1.4. The life cycle of the transcriptome

Large-scale genomics, transcriptomics and epigenomics analyses have unveiled the existence of a dynamic transcriptome that continuously evolves and changes across mammalian lineages, organs, developmental stages, chromosomes, and sexes. This implies that events of gene birth and death are frequent across the genome.

1.4.1. Transcription explores the genomic space

The high number of detected expressed transcripts in every species exposes a high and variable fraction of the genomic space to transcription. This is due to the pervasive transcription of the genome, and the high turnover of poorly conserved transcripts (Kutter et al., 2012). For instance, 81% of the lncRNA families found in human are primate-specific (Necsulea et al., 2014). Strikingly, when analyzing transcriptomes from different mouse taxa spanning a phylogenetic distance of 10 Mya, no transcript-free regions are observed indicating that nearly the entire genome can be transcribed into poly-adenylated RNA when viewed from an evolutionary perspective (Neme & Tautz, 2016).

Transcriptional dynamics are highly asymmetric between different tissues. Transcriptional changes are slow in some tissues, such as the nervous system and high in others, such as testis (Brawand et al., 2011). Testis tissue is subject to strong selective pressures associated with sperm competition, sexual conflict, reproductive isolation, germline pathogens,

and mutations that cause segregation distortions in the male germline (Nielsen et al., 2005). A large number of protein-coding genes and non-coding genes (Heinen et al., 2009; Kaessmann, 2010; Levine et al., 2006; Paulding et al., 2003; She et al., 2004) are expressed in testis. Different testis cell types at different time points exhibit very different transcriptomes, which partially explains the high transcriptional diversity found in that tissue. However, the high transcriptional levels may be better explained by the overall permissive chromatin states resulting from numerous rounds of chromatin remodeling (Soumillon et al., 2013). This may allow gene expression from simple or cryptic promoters (Kleene, 2005), which is also consistent with the high expression of retrogenes observed in testis (Betrán, Thornton, & Long, 2002; Marques et al., 2005).

1.4.2. The making of a new gene

A genome that is pervasively transcribed and whose RNA products change over time implies the continuous *de novo* birth of numerous expressed loci with no apparent functionality. The life cycle of a gene is defined by different scenarios that a gene endures over evolutionary time (Figure 1.8) (Neme & Tautz, 2014).

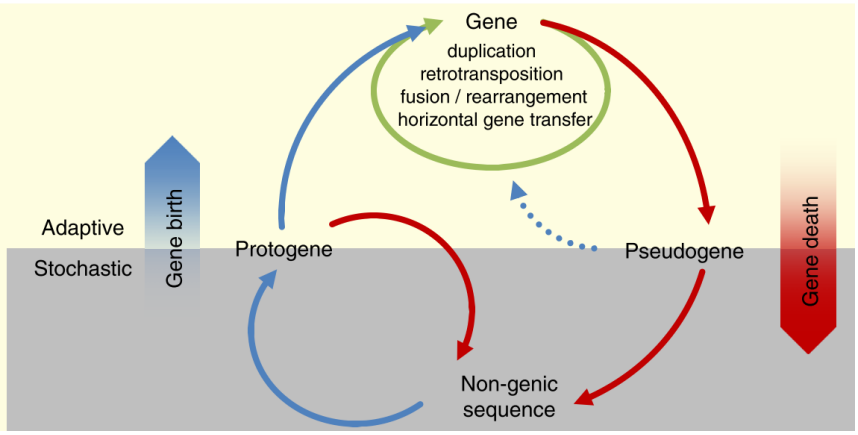

Figure 1.8. The life cycle of genes. Blue arrows represent transitions which lead, either partially or completely, to newly originated genes. Red arrows represent the loss of features that result in the degradation of the genic potential of a sequence. Green arrows represent the mechanisms which increase the gene repertoire from existing genes. Adapted from Rafik Neme & Tautz, 2014.

Since it is generally considered that a gene is a functional unit, most of these recently expressed loci are merely 'protogenes', genes subject to no or weak selective pressure that arise from previously non-expressed DNA sequences (Carvunis et al., 2012). Some of these sequences might become functional genes over time, and subsequently be expanded through gene duplication or other similar mechanisms. Eventually, some newly emerged genes might become non-functional again and pseudogenized (Demuth,

Bie, Stajich, Cristianini, & Hahn, 2006). A small fraction of the pseudogenes can become functional again (Bekpen et al., 2009), something which is more likely when they still conserve the capacity of being transcribed and/or translated (Shidhi et al., 2014). The relative constant number of genes over time means that most protogenes do not ever become functional genes. Accordingly, it has been observed than the probability of loss-of-function mutations is higher for *de novo* genes than for older ones (Palmieri et al., 2014). Selection will eliminate deleterious protogenes before they become established (Masel, 2006; Wilson & Masel, 2011). Therefore, the pool of protogenes will be enriched for those that have a higher chance of becoming a gene.

Given the high expression levels in testis and the gene life cycle model, it is not surprising that many young genes have been identified in this tissue (She et al., 2004; Xie et al., 2012). The 'out of testis' hypothesis proposes that a high fraction of genes arise in testis because of the aforementioned characteristics of this tissue. Over time, some new genes may evolve more efficient promoters, more diverse expression patterns, and gain functions in other tissues (Kaessmann, 2010; Light, Basile, & Elofsson, 2014).

The origination of a new gene not only involves the expression of a new locus but the acquisition of a translatable ORF (Figure 1.9). The RNA-first model describes that a non-coding loci would eventually acquire an ORF through DNA mutations (Xie et al., 2012). The ORFs may become longer and more structured over time (Bornberg-Bauer, Schmitz, & Heberlein, 2015), as supported by the presence of *de novo* genes with truncated ORFs in other species (J. Cai et al., 2008; Zhou et al., 2008).

1.4. THE LYFE CICLE OF THE TRANSCRIPTOME

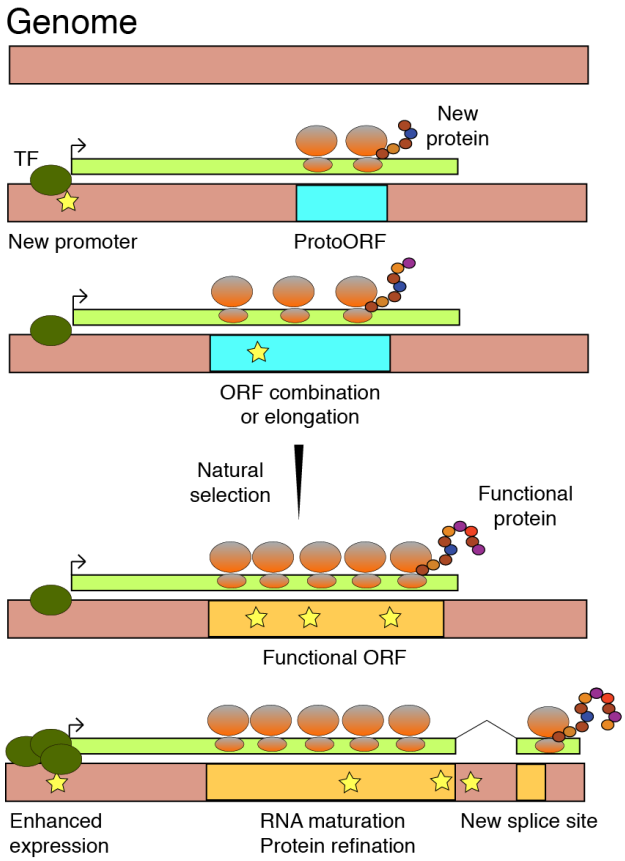

Figure 1.9. Model of *de novo* gene origination in a pervasively transcribed genome. New promoters continuously arise in the genome, which results in the production of novel non-functional transcripts. Some ORFs can become translated and eventually be elongated and combined with other ORFs. If the protein is useful it will continue to evolve under purifying selection. Subsequently, the new gene will tend to acquire a stronger promoter, a refined protein subject to stronger constraints, and a more complex exonic structure.

On the other hand, as the genome is full of random non-expressed ORFs, an alternative ORF-first model is also plausible. For example, ~60% of 800 bp intergenic sequences in *Drosophila* harbor ORFs of at least 150 bp (Zhao, Saelao, Jones, & Begun, 2014). Some of these ORFs might acquire regulatory elements that allow them to be transcribed and translated. Zhao et al. identified pre-existing ORFs in loci only expressed in a subset of *Drosophila* individuals, observing that the expression polymorphism was linked to *cis*-sequence variation (Zhao et al., 2014).

Finally, a gene acquiring a new translated ORF may keep or develop a non-coding function indistinctly. There are some examples of bi-functional RNAs in the literature, such as the human *Steroids Receptor Activator* (SRA), a RNA that co-activates steroid hormone receptors at the transcript level (Lanz et al., 1999) but which also encodes a protein that acts antagonistically to its non-coding function (Chooniedass-Kothari et al., 2004). Furthermore, numerous pairs of coding RNAs are stabilized by overlapping regions in their UTRs, and hence they might be considered bi-functional as well (Su et al., 2012).

2 RESULTS

2.1. Long non-coding RNAs as a source of new peptides

Authors: Jorge Ruiz-Orera, Xavier Messeguer, Juan A Subirana, M.Mar Albà

Published in: eLife (2014) 3: 1-24. doi:10.7554/eLife.03523

Full text: http://elifesciences.org/content/3/e03523v1

Abstract: Deep transcriptome sequencing has revealed the existence of many transcripts that lack long or conserved open reading frames (ORFs) and which have been termed long non-coding RNAs (lncRNAs). The vast majority of lncRNAs are lineage-specific and do not yet have a known function. Here we test the hypothesis that they may act as a repository for the synthesis of new peptides. We find that a large fraction of the lncRNAs expressed in cells from six different species is associated with ribosomes. The patterns of ribosome protection are consistent with the translation of short peptides. lncRNAs show similar coding potential and sequence constraints than evolutionary young protein coding sequences, indicating that they play an important role in *de novo* protein evolution.

Ruiz-Orera J, Messeguer X, Subirana JA, Alba MM. Long non-coding RNAs as a source of new peptides. Elife. 2014 Sep 16;3:e03523. doi: 10.7554/eLife.03523.

2.2. Origins of *De novo* genes in Human and Chimpanzee

Authors: Jorge Ruiz-Orera, Jessica Hernandez-Rodriguez, Cristina Chiva, Eduard Sabidó, Ivanela Kondova, Ronald Bontrop, Tomàs Marqués-Bonet, M.Mar Albà

Published in: PLOS Genetics (2015) *11*(12), e1005721. doi:10.1371/journal.pgen.1005721

Full text: http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1005721

Abstract: The birth of new genes is an important motor of evolutionary innovation. Whereas many new genes arise by gene duplication, others originate at genomic regions that did not contain any genes or gene copies. Some of these newly expressed genes may acquire coding or non-coding functions and be preserved by natural selection. However, it is yet unclear which is the prevalence and underlying mechanisms of *de novo* gene emergence. In order to obtain a comprehensive view of this process we have performed in-depth sequencing of the transcriptomes of four mammalian species - human, chimpanzee, macaque, and mouse - and subsequently compared the assembled transcripts and the corresponding syntenic genomic regions. This has resulted in the identification of over five thousand new multiexonic transcriptional events in human and/or chimpanzee that are not observed in the rest of species. Using comparative genomics we show that the expression of these transcripts is associated with the gain of regulatory motifs upstream of the transcription start site (TSS) and of U1 snRNP sites downstream of the TSS. In general, these transcripts show little evidence of purifying selection, suggesting that many of them are not functional. However, we find signatures of selection in a subset of *de novo* genes which have evidence of protein translation. Taken together, the data supports a model in which frequently-occurring new transcriptional events in the genome provide the raw material for the evolution of new proteins.

Ruiz-Orera J, Hernandez-Rodriguez J, Chiva C, Sabidó E, Kondova I, Bontrop R, Marqués-Bonet T, Albà MM. Origins of De Novo Genes in Human and Chimpanzee. PLoS Genet. 2015 Dec 31;11(12):e1005721. doi: 10.1371/journal.pgen.1005721.

2.3. Functional and non-functional classes of peptides produced by long non-coding RNAs

Authors: Jorge Ruiz-Orera, Pol Verdaguer-Grau, José Luis Villanueva-Cañas, Xavier Messeguer, M.Mar Albà

Submitted. Preprint published in: BioRxiv (2016).
doi: http://dx.doi.org/10.1101/064915

Full text: http://biorxiv.org/content/early/2016/07/21/064915

Abstract: Cells express thousands of transcripts that show weak coding potential. Known as long non-coding RNAs (lncRNAs), they typically contain short open reading frames (ORFs) having no homology with known proteins. Recent studies have reported that a significant proportion of lncRNAs are translated, challenging the view that they are essentially non-coding. These results are based on the selective sequencing of ribosome-protected fragments, or ribosome profiling. The present study used ribosome profiling data from eight mouse tissues and cell types, combined with ~330,000 synonymous and non-synonymous single nucleotide variants, to dissect the biological implications of lncRNA translation. Using the three-nucleotide read periodicity that characterizes actively translated regions, we found that about 23% of the transcribed lncRNAs was translated (1,365 out of 6,390). About one fourth of the translated sequences (350 lncRNAs) showed conservation in humans; this is likely to produce functional micropeptides, including the recently discovered myoregulin. For other lncRNAs, the ORF codon usage bias distinguishes between two classes. The first has significant coding scores and contains functional proteins which are not conserved in humans. The second large class, comprising >500 lncRNAs, produces proteins that show no significant purifying selection signatures. We showed that the neutral translation of these lncRNAs depends on the transcript expression level and the chance occurrence of ORFs with a favorable codon composition. This provides the first evidence to data that many lncRNAs produce non-functional proteins.

Ruiz-Orera J, Verdaguer-Grau P, Villanueva-Cañas JL, Messeguer X, Alba MM. Functional and non-functional clases of peptides produced by long non-coding RNAs. BioRxiv (2016).

3 DISCUSSION

3.1. Molecular processes involved in *de novo* gene origination

The advent of next-generation sequencing technologies led to the discovery of thousands of expressed loci that are not conserved across distant species. Most of these transcripts have no reported functionality and some authors have proposed that they are 'transcriptional noise' (Struhl, 2007; Jun Wang et al., 2004). This is consistent with the finding that pervasive transcription probably has a negligible cost for the cell (Lynch & Marinov, 2015). Despite this, the existence of highly dynamic transcriptomes can have important consequences for the evolution of new genes and functions. Expressed transcripts, even if they are not functional, can provide the raw material necessary for the birth of new genes. The discovery of thousands of non-characterized translated ORFs in different species by ribosome profiling has opened up new questions about the extent of transcript translation and its significance for protein-coding gene evolution.

In this thesis, we focused on the identification and characterization of the evolutionary plasticity of the transcriptome and translatome in different eukaryotes, including human and mouse. With the data we gathered, we propose an evolutionary model to explain how protein-coding genes can arise *de novo* in different species. This process can be broken into four steps: transcription, translation, selection, and maturation. Transcription

and translation are the source of new genes and proteins. Selection will act as a filter to eliminate deleterious new genes and favor useful ones. Finally, maturation is required to explain the differences between recently evolved functional genes and older ones. Alternatively, genes with non-coding functions can concurrently emerge in the genome. These genes will not be translated (or only secondarily) and will be selected for non-coding functions.

Transcription: The fraction of the genome that is transcribed is much larger than the fraction that corresponds to annotated genes. As a result, transcript assembly from RNA-seq typically identifies many transcripts that are not yet annotated in the databases (Iyer et al., 2015; Necsulea et al., 2014), and genes can also emerge from intronic regions (Kumar, 2009; Toll-Riera et al., 2009; Zhao et al., 2014). In any case, genomes are exposed to a fast transcriptional turnover over evolutionary time (Neme & Tautz, 2016). In line with this, we found thousands of non-annotated loci expressed in human and other mammals. This included both conserved and non-conserved genes and extended the current gene catalogs in those species (Ruiz-Orera et al., 2015). To ensure high transcript coverage, we used deep-sequencing transcriptomics for four mammalian species and analyzed data from several tissues.

To date, our study is the only one that identifies *de novo* transcription in mammals by using reconstructed assemblies instead of the annotations. This greatly increased the power of BLAST to detect expressed homologs in other species, since the annotations in many species are poor and based on predictions and projections from other species. Moreover, syntenic genomic alignments are available from the UCSC genome browser and

provide numerous alignment blocks of chromosome sequences for two different species, when possible. We used these syntenic alignments to directly check for the expression of the corresponding loci in related species. This analysis identified 2,714 *de novo* genes that are specific to human and/or chimpanzee, which constitutes 4% of the total set of genes. Only 8 of these genes were annotated as protein-coding in the databases, highlighting the importance of considering long non-coding RNAs and novel reconstructed genes to characterize the whole *de novo* transcriptome. The use of strand-specific RNA-seq detected a myriad of antisense transcripts that are not yet annotated since they would be missed by conventional RNA-seq technologies.

One of the fundamental questions regarding *de novo* gene emergence is what triggers the transcription of a new locus. It is known that animals and plants have evolved complex transcriptional regulation factors. However, the promoters of young and old genes show distinct features and interact with different transcription factors (de Mendoza et al., 2013). Conserved genes usually have GC-rich promoters, which have been hypothesized to result from the acquisition of numerous CpG islands across time (Almada et al., 2013). In contrast, *de novo* expressed loci are enriched in A/T-rich motifs (Carvunis et al., 2012; Necsulea et al., 2014; Toll-Riera et al., 2009). In mouse, a *de novo* gene has been proposed to arise after gaining new regulatory motif sequences (Heinen et al., 2009; Tautz & Domazet-Lošo, 2011).

In my thesis, I shed new light on this question by comparing data from closely related species that differ in the expression of the syntenic region.

We found an enrichment of transcription factor binding sites in *de novo* promoters when compared to syntenic regions not expressing the gene; in agreement with the hypothesis that the gain of regulatory motifs underlies *de novo* gene origination (Ruiz-Orera et al., 2015). We found an over-representation of two non-annotated motifs, which may correspond to binding sites for the polymerase II complex, as well as many sites recognized by RFX2, which is highly expressed in testis and has been involved in spermiogenesis (Kistler et al., 2015). This may explain why the expression of most *de novo* genes was restricted to testis. In addition, we found that *de novo* gene expression is stable in the human population as the majority of *de novo* genes were detected in all or the vast majority of individuals with testis sequencing data in GTEx (T. G. Consortium, 2013).

The birth of new motifs can be explained by the random accumulation of mutations in the genome. In addition, we found that in 13% of *de novo* genes, transposable elements (TEs) were an important source of regulatory sequences. TEs were previously found to be enriched in these types of sequences (Jordan, Rogozin, Glazko, & Koonin, 2003) and have been proposed to contribute to non-coding RNA origination (Faulkner et al., 2009; Kapusta et al., 2013).

Moreover, new genes are expected to emerged from bidirectional promoters of older genes (Gotea, Petrykowska, & Elnitski, 2013; X. Wu & Sharp, 2013). We found that 20% of *de novo* genes were associated with bidirectional promoters. The expression of these genes tends to be positively correlated with the older member of the pair and, as a result,

this class of *de novo* genes was expressed in a broader range of tissues than other *de novo* genes.

Translation: Translation is likely to be a highly pervasive process that produces both functional and non-functional products as judged by ribosome profiling analyses (Bazzini et al., 2014; Heesch et al., 2014; Ingolia et al., 2011; Ji et al., 2015; Juntawong et al., 2014; Ruiz-Orera et al., 2014). Since most of transcripts are polyadenylated, they accumulate in the cytoplasm (Carninci et al., 2005; Heesch et al., 2014) and therefore are exposed to the translation machinery.

In this thesis, we addressed what drives the translation of ORFs in codRNAs and lncRNAs. We performed the first meta-analysis of ribosome profiling data for several species and we observed that translation, while not so common as in codRNAs, is a pervasive process in lncRNAs regardless of the conservation of the ORF in other species. LncRNAs are often poorly expressed and we observed how this low abundance had a direct effect on the detection of ribosome association. Moreover, ORFs in lncRNAs were often small (< 100 aa) and had lower translational efficiency (TE) values than protein-coding genes, although they were still significantly higher than untranslated regions (UTRs). In protein-coding genes, one ORF harbored most of the ribosome profiling signal; however, in lncRNAs we observed that translation of multiple small ORFs was a relatively common event. Polycistronic transcripts have been reported in insects (Savard et al., 2006) and may extend to other species (Tautz, 2009).

The use of mouse ribosome profiling datasets with high read coverage resulted in the discovery of higher levels of translated peptides. We analyzed the ribosome protection patterns using the 3-nucleotide bias of ribosome profiling as indicator of true translation (Bazzini et al., 2014; Calviello et al., 2016; Ingolia et al., 2009; Ji et al., 2015). Previous studies used different models to distinguish true translation from noise. We identified ORFs with a significant frame bias by using randomization of the reads as a null model. This, combined with the use of 8 different datasets, led to the identification of thousands of translated ORFs. The translation patterns of many of these peptides were similar across tissues, indicating that their translation is relatively stable and reproducible. In humans, we only mapped ribosome profiling reads from brain and HeLa cells to the transcriptome. We detected translation patterns in many lncRNAs, including several human- or hominoid-specific *de novo* genes (Ruiz-Orera et al., 2015, 2014). We expect that, when ribosome profiling data from other human tissues become available, the number of translated *de novo* genes will increase dramatically.

The translation of numerous non-conserved peptides in mouse strengthens the idea that non-coding RNAs can be used as raw material for the birth of protein-coding genes (J. Cai et al., 2008; Carvunis et al., 2012; Levine et al., 2006; Wilson & Masel, 2011). In principle, non-sense mediated decay (NMD) could degrade many of these peptides. However, by analyzing the relative position of the STOP codon and splice junctions, we predict that only a minority of them will be targeted by this mechanism.

3.1. MOLECULAR PROCESSES IN DE NOVO GENE ORIGINATION

Selection: Translation is a highly pervasive process; this leads to the production of more peptides than was previously considered. The translation of peptides that are toxic to the cell will be the target of negative selection and so we will, in general, not observe them (Wilson & Masel, 2011). Our data, based on single nucleotide variants, suggests that a large number of new genes, and translated peptides, have no function and evolve neutrally. Many of these genes are therefore expected to degenerate over time. Consistent with this idea, studies in insects have found an excess of new genes in the terminal branches when compared to genes originated in more internal branches (Palmieri et al., 2014; Wissler et al., 2013) and we observed the same trend in primates (Ruiz-Orera et al., 2015).

The large number of recently expressed loci in a species provides a large reservoir for the appearance of new molecular functions. Some of the genes will turn out to be useful and be preserved by natural selection. In our study, we observed evidence of purifying selection in a subset of species or lineage-specific genes. We also measured the coding score of predicted ORFs using a statistic based on dicodon frequencies. Coding score captures mutational biases as well as the effect of natural selection for translation optimization. We observed that translated lncRNAs exhibit higher coding scores than non-translated lncRNAs and intron sequences, but lower than protein-coding and pseudogene sequences. More importantly, translated lncRNAs showed similar coding potential and sequence constraints than evolutionary young protein-coding genes defined in other studies (Ruiz-Orera et al., 2014). Resemblance between young protein-coding genes and translated non-coding genes in different

species probably indicates that translated lncRNAs form a heterogeneous group of genes ranging from non-functional new genes to well-established protein-coding genes.

We observed that a subset of recent mouse genes with high coding scores exhibited signatures of purifying selection, whereas the rest of sequences, despite signals of translation, evolved neutrally. To some extent, the coding score reflects the functionality of an ORF. In this study, we provided the first evidence that many lncRNAs produce non-functional proteins. These genes correspond to a previously defined class of genes from a study performed in yeast known as 'protogenes' (Carvunis et al., 2012). We also obtained evidence that a subset of species-specific genes showed evidence of purifying selection. These genes fit a more classical definition of *de novo* genes, which assumes that the genes are functional (McLysaght & Hurst, 2016).

It was previously hypothesized that the translation of a lncRNA may be linked to the relative amount of transcripts in the nucleus and the cytoplasm, but we observed translation of some lncRNAs with nuclear functions suggesting that the cytosolic fraction of any lncRNA may be translated regardless of the role or preferred location of the transcript. We determined that the neutral translation of these lncRNAs depends on the transcript expression level and the chance occurrence of ORFs with a favorable codon composition. *Poldi*, one *de novo* lncRNA described to have a non-coding function (Heinen et al., 2009), harbors one long ORF with a very low coding score. Accordingly, it is not translated in any of the samples analyzed in our study.

Maturation: A *de novo* gene that is useful for the cell and that is subjected to negative selection will mature into a more complex structure over time. Since older genes will have had more time to evolve complex coding sequences and promoters, it is not surprising that several studies have found a correlation between gene age and properties like number of exons, protein length, and gene expression level (Arendsee et al., 2014; Carvunis et al., 2012; Neme & Tautz, 2013). We found the expected patterns in *de novo* genes from human and chimpanzee, which were shorter and had more tissue-restricted expression than conserved genes (Ruiz-Orera et al., 2015).

It is likely that numerous *de novo* protogenes and genes are initially monoexonic since this is the simplest gene structure. For example, up to 5% of annotated human protein-coding genes are single-exon. Our study detected the expression of numerous non-conserved monoexonic loci, but the analysis of a RNA-seq sample without reverse transcriptase revealed the presence of numerous monoexonic artifacts that we could only discard by removing genes with one exon. It has been proposed that the formation of introns is a process that depends on the age of the gene. The number of U1snRNP sites (GGUAAG-like, 5'splice site) in the first kilobase after the transcriptional start site increases with the age of the gene while the number of PAS (poly-adenylation termination signals) sites decreases, which results in an increased rate of transcript elongation over time (Almada et al., 2013). Besides, the first intron is usually longer and more conserved than other downstream introns (Bradnam & Korf, 2008; Park et al., 2014) since it harbors most *cis* regulatory sequences (Chorev & Carmel, 2012). In this thesis, we found an enrichment of transcription

factor binding sites and U1snRNP motifs in the first 200bp of *de novo* genes in human and chimpanzee; this is consistent with the idea that the gain of regulatory motifs underlies *de novo* gene origination.

Other elements involved in transcript elongation are transposable elements (TEs) (Kapusta et al., 2013; Toll-Riera et al., 2009), and some of them harbor ORFs that can form fusion proteins with proximal exons (Denli et al., 2015). We found that *de novo* transcripts were enriched in such elements; about 20% of their total transcript length was covered by TEs, compared to only 8% in conserved genes.

Over time, a gene may acquire multiple isoforms, stronger promoters that might produce antisense transcripts, and more structured proteins that associate with other proteins and become more integrated into cellular networks (Abrusán, 2013; Bornberg-Bauer et al., 2015; Moore & Bornberg-Bauer, 2012).

Non-coding functions in RNA: The evolutionary model that was explained above intends to explain how a new protein-coding gene can arise in a genome. Nevertheless, several lncRNAs exert non-coding functions directly or by regulating the transcription and/or translation of other genes. These functional lncRNAs will be maintained by natural selection, although non-overlapping sequences might experience important changes in sequence and exon-intron structure. Thus, functional lncRNAs in the cytoplasm will be continuously exposed to the translation of new ORFs. A significant fraction of identified *de novo* genes were located in an antisense configuration and hence some of those cases might

have regulatory functions at transcriptional, translational, or post-transcriptional level. The PN/PS ratio will not be of any use to detect the non-coding functionality of such sequences. Other approaches, based on whole sequence values of diversity and divergence, will be more appropriate (Wiberg et al., 2015).

Some RNAs may be bi-functional displaying both coding and non-coding functions. For example, *Malat1* is a well-characterized RNA that is often retained in the nucleus, where it forms ribonucleoprotein complexes and it regulates the expression of numerous genes (Tripathi et al., 2010). However, some RNA molecules are exported to the cytoplasm and can be translated via a 3' triple helical structure (Marzluff, Wilusz, Jnbaptiste, & Lu, 2012). We found a small protein of 57 aa with evidence of proteomics and ribosome profiling that is translated from *Malat1*. Other functional lncRNAs, such as *Neat1*, *Jpx*, and *Cyrano*, also contained translated ORFs.

On the other hand, several protein-coding genes were revealed to have an intrinsic non-coding function (Karapetyan, Buiting, Kuiper, & Coolen, 2013; Kumari & Sampath, 2015). Such functionality is usually performed by the UTR regions, that resemble lncRNAs in base composition and structure. Finally, non-coding isoforms from protein-coding genes may have functional roles, as it occurs with the circular RNAs (Lasda, Parker, & Parker, 2014).

3.2. Improving the gene annotation

As previously explained, numerous long non-coding transcripts are discovered and annotated based on the analysis of high coverage RNA-seq data from different tissues and conditions. While the number of long non-coding RNAs is continuously raising, the number of protein-coding genes is slightly decreasing as some of them are re-annotated as non-coding or discarded as annotation artifacts.

The observation of a significant proportion of translated non-coding and non-annotated transcripts, though, calls into question how genes should be annotated. While most coding prediction software shows high precision in the detection of long and conserved protein-coding genes, it usually fails when trying to characterize the coding potential of short and non-conserved translated sequences. Other studies proposed the existence of hundreds of conserved functional smORFs (Bazzini et al., 2014; Hanada et al., 2013; Mackowiak et al., 2015). In our study, we observed how the analysis of the sequence conservation in other species, combined with the global analysis of purifying selection with polymorphism data, is an effective method to characterize translated groups of functional smORFs that are often non-annotated. The SNP analysis is limited to groups of genes, as individual coding sequences do not usually contain enough polymorphisms to perform statistical analysis (Gayà-Vidal & Albà, 2014).

Numerous non-conserved ORFs, most of them shorter than 100 aa, have been observed to be translated in our study in mouse. We developed a new computational tool, CIPHER, that predicts the coding potential of a sequence by computing an hexamer-based coding score. Unlike other

properties of coding sequences, the hexamer score is a robust metric that is independent of the length and that separates translated and non-translated transcripts regardless of the age of the gene. Widely-used predictors such as PhyloCSF or CPAT use multiple sequence conservation or ORF length metrics to predict coding sequences (M. F. Lin, Jungreis, & Kellis, 2011; L. Wang et al., 2013). This is not appropriated for the ORFs mentioned here, which are not phylogenetically conserved. Using the hexamer scoring metric we could identify many of the ORFs that were translated according to ribosome profiling data. However, there were also translated ORFs that had low scores and which would have not been detected by CIPHER. Hence, other unknown properties might induce the translation of ORFs.

Another current limitation is that predictions and annotations are usually limited to a single ORF per transcript that is often the longest one. However, many transcripts are polycistronic and numerous small ORFs are translated, including regulatory uORFs and non-canonical ORFs (Aspden et al., 2014; Ingolia et al., 2014; Ji et al., 2015; Juntawong et al., 2014), but those sequences usually escape the analysis of coding prediction tools because they are small and/or lack standard start and stop signals. Moreover, there are a few examples of genes translating multiple functional ORFs in separate or overlapping frames (Andrews & Rothnagel, 2014). Hence, future annotation efforts should aim at identifying all the multiple translated sequences. CIPHER can predict multiple translated ORFs per sequence, although the number of false positives will likely increase. Thus, the integration of ribosome profiling and ORF prediction data should lead to improved gene annotations.

So far, genes that are translated are only classified as protein-coding, but the data shown indicates that there are different types of translated sequences (ORFs): functional translated, regulatory translated, and non-functional translated. Depending on how we define a protein-coding gene, we may include thousands of sequences that are translated but not conserved, or instead focus on those that are conserved across species.

Once the number of sequences that are transcribed and translated in any organism has been defined, the question of how many genes are functional arises. While the term 'gene' is usually linked to functionality, we found several loci with no assigned functions that are neither conserved nor subjected to negative selection. Without ignoring the possibility of these loci having a regulatory non-coding function, the term 'protogene' define this set of non-functional sequences that might be retained or lost over time (Carvunis et al., 2012). In the future, annotations should distinguish between functional genes and non-functional protogenes, as they do now separate genes and pseudogenes. In this thesis, we observed how the coding score is a powerful metric to identify functional coding sequences that are not conserved in other distant species.

Moreover, the number of annotated lncRNAs continuously rises, but the annotations are based on expression, and the gene biotypes rely on structural properties, instead of functionality. Finally, the aforementioned possibility of the existence of numerous bi-functional genes, with coding and non-coding properties, adds more complexity to this problem.

4 CONCLUSIONS

1. We have identified thousands of non-annotated genes in human, chimpanzee, macaque, and mouse by analyzing strand-specific RNA-seq for different tissues, including brain and testis.

2. We have published the first catalog of multiexonic *de novo* genes in human and chimpanzee that includes protein-coding genes, long non-coding RNA genes and novel genes. These genes are short and tissue-restricted.

3. We have obtained evidence that genes which have only recently begun to be transcribed have gained regulatory motifs in promoters and exons.

4. We have developed CIPHER, a method using hexanucleotide frequencies to measure the coding potential of a sequence (coding score) independently of open reading frame length or phylogenetic conservation.

5. We have identified a significant fraction of long non-coding RNA transcripts associated with ribosomes in different eukaryote species. These transcripts exhibit lower abundance and translation efficiency than protein-coding genes. Their similarity to young

protein-coding genes indicates that they play an important role in *de novo* protein-coding gene evolution.

6. We have identified 350 long non-coding RNAs with evidence of translation in mouse that show sequence similarity to human transcripts. These genes are probably producing functional micropeptides.

7. We have found little evidence of purifying selection in recently evolved genes in general. However, some translated non-conserved genes show signatures of selection and are likely to be functional.

8. We have observed that translation is pervasive and some of the translated proteins show no evidence of purifying selection. Translation of neutral sequences depends on the coding score and the transcript expression level.

5 FUTURE RESEARCH

In this thesis I have described a series of analysis to quantify and characterize the regions of a genome that are expressed and/or translated, greatly extending the current gene annotation catalogs. Our study reveals the presence of a pervasive transcriptome with a significant fraction that continuously evolves *de novo* from not expressed DNA, and a pervasive translatome, with the translation of numerous short coding sequences into proteins.

Naturally, these findings raise new and interesting questions to be solved in the next years. As explained in the first chapter of the thesis, the definition of a gene has been changing over time, and it will probably continue to change for a long time. This will be triggered by the development of new technologies that will permit researchers to better assemble genomes and transcriptomes, as well as measuring the translatome with increasing accuracy and coverage.

Our work contributes to a better understanding of gene evolution and highlights the necessity to rethink what a gene is and how we define it. Investigating how new loci arise in genomes is a key step to understand the present architecture of genomes. Once the loci are defined, we can apply population and functional analyses to find out which fraction of the dynamic transcriptome is functional and how *de novo* genes are fixed in populations. Genes should then be tagged on the basis of functionality

rather than the presence or absence of expression. A parallel effort should be done to characterize the translatome in every cell type and condition in order to identify how many coding sequences produce functional short and long proteins. We should also investigate how cells cope with the high amount of translated proteins that is observed with ribosome profiling experiments and how much of this material is quickly degraded. Thus, what we consider a protein-coding gene might be subdivided into different functional categories, and this information should be combined with the non-coding functionality that was developed in this thesis.

6 ANNEX

Journal articles

Ruiz-Orera J, Messeguer X, Subirana JA., & Albà MM (2014). Long non-coding RNAs as a source of new peptides. ELife, 3, 1–24.

Ruiz-Orera J, Hernandez-Rodriguez J, Chiva C, Sabidó E, Kondova I, Bontrop R, Marqués-Bonet T, & Albà MM (2015). Origins of Genes in Human and Chimpanzee. PLoS Genet 11(12): e1005721.

Ruiz-Orera J, Verdaguer-Grau P, Villanueva-Cañas JL, Messeguer X, Albà MM (2016). Functional and non-functional classes of peptides produced by long non-coding RNAs (Submitted).

Oral communications

XXI Seminario de Genética de Poblaciones y Evolución. Sitges, 2016
Title: Functional and non-functional classes of peptides produced by long non-coding RNAs.

XII Symposium on Bioinformatics. Sevilla 2014
Title: *Do long non-coding RNAs make proteins?*

I Jornadas de Bioinformatica IEC. Barcelona 2013
Title: *Identification of recently evolved genes in human and chimpanzee using next generation sequencing technologies.*

Poster presentations

Annual meeting of the Society for Molecular Biology and Evolution. Vienna, 2015
Title: Origins of de novo genes in human and chimpanzee.

4th Meeting of the Spanish Society for Evolutionary Biology (SESBE). Barcelona, 2013
Title: *Identification of recently evolved genes in human and chimpanzee using next generation sequencing technologies.*

XIII Jornada de Biologia Evolutiva IEC. Barcelona, 2013
Title: *Identification of recently evolved genes in human and chimpanzee using next generation sequencing technologies.*

7 REFERENCES

Abrusán, G. (2013). Integration of new genes into cellular networks, and their structural maturation. *Genetics*, *195*(4), 1407–17. http://doi.org/10.1534/genetics.113.152256

Akashi, H. (1997). Codon bias evolution in Drosophila. Population genetics of mutation-selection drift. *Gene*, *205*(1–2), 269–278.

Albà, M. M., & Castresana, J. (2005). Inverse relationship between evolutionary rate and age of mammalian genes. *Molecular Biology and Evolution*, *22*(3), 598–606. http://doi.org/10.1093/molbev/msi045

Albà, M. M., & Castresana, J. (2007). On homology searches by protein Blast and the characterization of the age of genes. *BMC Evolutionary Biology*, *7*, 53. JOUR. http://doi.org/10.1186/1471-2148-7-53

Almada, A. E., Wu, X., Kriz, A. J., Burge, C. B., & Sharp, P. A. (2013). Promoter directionality is controlled by U1 snRNP and polyadenylation signals. *Nature*, *499*(7458), 360–3. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3720719&tool=pmcentrez&rendertype=abstract

Altschmied, J., Delfgaauw, J., Wilde, B., Duschl, J., Bouneau, L., Volff, J.-N., & Schartl, M. (2002). Subfunctionalization of duplicate mitf genes associated with differential degeneration of alternative exons in fish. *Genetics*, *161*(1), 259–267. JOUR. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1462118/

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.

Nucleic Acids Research, *25*(17), 3389–402. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=146917&tool=pmcentrez&rendertype=abstract

Anderson, D. M. M., Anderson, K. M. M., Chang, C.-L.-. L., Makarewich, C. A. A., Nelson, B. R. R., McAnally, J. R. R., … Olson, E. N. (2015). A Micropeptide Encoded by a Putative Long Noncoding RNA Regulates Muscle Performance. *Cell*, *160*(4), 595–606. article. http://doi.org/10.1016/j.cell.2015.01.009

Andrews, S. J., & Rothnagel, J. a. (2014). Emerging evidence for functional peptides encoded by short open reading frames. *Nature Reviews. Genetics*, *15*(3), 193–204. article. http://doi.org/10.1038/nrg3520

Arendsee, Z. W., Li, L., & Wurtele, E. S. (2014). Coming of age: orphan genes in plants. *Trends in Plant Science*, 1–11. http://doi.org/10.1016/j.tplants.2014.07.003

Artieri, C. G., & Fraser, H. B. (2014). Evolution at two levels of gene expression in yeast. *Genome Research*, *24*(3), 411–21. http://doi.org/10.1101/gr.165522.113

Aspden, J. L., Eyre-Walker, Y. C., Philips, R. J., Amin, U., Mumtaz, M. A. S., Brocard, M., & Couso, J.-P. (2014). Extensive translation of small ORFs revealed by Poly-Ribo-Seq. *eLife*, e03528. http://doi.org/10.7554/eLife.03528

Babak, T., Blencowe, B. J., & Hughes, T. R. (2005). A systematic search for new mammalian noncoding RNAs indicates little conserved intergenic transcription. *BMC Genomics*, *6*(1), 1–12. article. http://doi.org/10.1186/1471-2164-6-104

Bailey, T. L., Boden, M., Buske, F. a., Frith, M., Grant, C. E., Clementi, L., … Noble, W. S. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research*, *37*(SUPPL. 2), W202-8. http://doi.org/10.1093/nar/gkp335

7. REFERENCES

Basu, K., Graham, L. A., Campbell, R. L., & Davies, P. L. (2015). Flies expand the repertoire of protein structures that bind ice. *Proceedings of the National Academy of Sciences*, *112*(3), 737–742. article.

Bazzini, A. A., Johnstone, T. G., Christiano, R., Mackowiak, S. D., Obermayer, B., Fleming, E. S., … Giraldez, A. J. (2014). Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *The EMBO Journal*. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/24705786

Beadle, G. W., & Tatum, E. L. (1941). Genetic control of biochemical reactions in Neurospora. *Proc. Natl. Acad. Sci.*, *27*, 499–506.

Begun, D. J., Lindfors, H. A., Kern, A. D., & Jones, C. D. (2007). Evidence for de novo evolution of testis-expressed genes in the Drosophila yakuba/Drosophila erecta clade. *Genetics*, *176*(2), 1131–1137. Journal Article, Research Support, N.I.H., Extramural, Research Support, U.S. Gov't, Non-P.H.S. http://doi.org/10.1534/genetics.106.069245

Begun, D. J., Lindfors, H. A., Thompson, M. E., & Holloway, A. K. (2006). Recently evolved genes identified from Drosophila yakuba and D. erecta accessory gland expressed sequence tags. *Genetics*, *172*(3), 1675–81. http://doi.org/10.1534/genetics.105.050336

Bekpen, C., Marques-Bonet, T., Alkan, C., Antonacci, F., Leogrande, M. B., Ventura, M., … Eichler, E. E. (2009). Death and Resurrection of the Human IRGM Gene. *PLoS Genetics*, *5*(3), e1000403. http://doi.org/10.1371/journal.pgen.1000403

Bellora, N., Farré, D., & Mar Albà, M. (2007). PEAKS: identification of regulatory motifs by their position in DNA sequences. *Bioinformatics (Oxford, England)*, *23*(2), 243–4. http://doi.org/10.1093/bioinformatics/btl568

Berget, S. M., Moore, C., & Sharp, P. A. (1977). Spliced segments at the 5'terminus of adenovirus 2 late mRNA. *Proceedings of the National Academy of Sciences*, *74*(8), 3171–3175. article.

Betrán, E., Thornton, K., & Long, M. (2002). Retroposed new genes out of the X in Drosophila. *Genome Research*, *12*(12), 1854–1859. article.

Bhan, A., & Mandal, S. S. (2014). Long noncoding RNAs: emerging stars in gene regulation, epigenetics and human disease. *ChemMedChem*, *9*(9), 1932–1956. article.

Bingxin, L. U., Zhenbing, Z., & Tieliu, S. H. I. (2013). Comparative study of de novo assembly and genome-guided assembly strategies for transcriptome reconstruction based on RNA-Seq, *56*(2), 143–155. http://doi.org/10.1007/s11427-013-4442-z

Biswas, S., & Akey, J. M. (2006). Genomic insights into positive selection. *TRENDS in Genetics*, *22*(8), 437–446. article.

Black, D. L. (2003). Mechanisms of alternative pre-messenger RNA splicing. *Annual Review of Biochemistry*, *72*(1), 291–336. article.

Boguski, M. S., Tolstoshev, C. M., & Bassett, D. E. (1994). Gene discovery in dbEST. *Science*, *265*, 1993–1994. JOUR. Retrieved from http://dx.doi.org/10.1126/science.8091218

Bornberg-Bauer, E., Schmitz, J., & Heberlein, M. (2015). Emergence of de novo proteins from “dark genomic matter” by “grow slow and moult”. *Biochemical Society Transactions*, *43*(5), 867–873. Journal Article, Research Support, Non-U.S. Gov’t, Review. http://doi.org/10.1042/BST20150089

Bosch, T. C. G. (2014). Rethinking the role of immunity: lessons from Hydra. *Trends in Immunology*, *35*(10), 495–502. article.

Boucher, Y., Douady, C. J., Papke, R. T., Walsh, D. A., Boudreau, M. E. R., Nesbo, C. L., … Doolittle, W. F. (2003). Lateral gene transfer and the origins of prokaryotic groups. *Annual Review of Genetics*, *37*, 283–328. Journal Article, Review. http://doi.org/10.1146/annurev.genet.37.050503.084247

Bowler, P. J. (1989). Evolution: The history of an idea. *University Edition*.

7. REFERENCES

Bradnam, K. R., & Korf, I. (2008). Longer First Introns Are a General Property of Eukaryotic Gene Structure. *PLoS ONE*, *3*(8), e3093. http://doi.org/10.1371/journal.pone.0003093

Brar, G. a, Yassour, M., Friedman, N., Regev, A., Ingolia, N. T., & Weissman, J. S. (2012). High-resolution view of the yeast meiotic program revealed by ribosome profiling. *Science (New York, N.Y.)*, *335*(6068), 552–7. http://doi.org/10.1126/science.1215110

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., … Kaessmann, H. (2011). The evolution of gene expression levels in mammalian organs. *Nature*, *478*(7369), 343–8. http://doi.org/10.1038/nature10532

Brenner, S. (2000). Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature Biotechnol.*, *18*, 630–634. JOUR. Retrieved from http://dx.doi.org/10.1038/76469

Brockdorff, N., Ashworth, A., Kay, G. F., McCabe, V. M., Norris, D. P., Cooper, P. J., … Rastan, S. (1992). The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. *Cell*, *71*(3), 515–26.

Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., … Rinn, J. L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & Development*, *25*(18), 1915–27. article. http://doi.org/10.1101/gad.17446611

Cai, J. J., Borenstein, E., Chen, R., & Petrov, D. a. (2009). Similarly strong purifying selection acts on human disease genes of all evolutionary ages. *Genome Biology and Evolution*, *1*, 131–44. http://doi.org/10.1093/gbe/evp013

Cai, J. J., & Petrov, D. A. (2010). Relaxed purifying selection and possibly high rate of adaptation in primate lineage-specific genes. *Genome*

Biology and Evolution, *2*, 393–409. http://doi.org/10.1093/gbe/evq019

Cai, J., Zhao, R., Jiang, H., & Wang, W. (2008). De novo origination of a new protein-coding gene in Saccharomyces cerevisiae. *Genetics*, *179*(1), 487–96. http://doi.org/10.1534/genetics.107.084491

Calviello, L., Mukherjee, N., Wyler, E., Zauber, H., Hirsekorn, A., Selbach, M., … Ohler, U. (2016). Detecting actively translated open reading frames in ribosome profiling data. *Nat Meth*, *13*(2), 165–170. JOUR. Retrieved from http://dx.doi.org/10.1038/nmeth.3688

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., … Hayashizaki, Y. (2005). The transcriptional landscape of the mammalian genome. *Science (New York, N.Y.)*, *309*(5740), 1559–63. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16141072

Carrieri, C., Cimatti, L., Biagioli, M., Beugnet, A., Zucchelli, S., Fedele, S., … Gustincich, S. (2012). Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. *Nature*, *491*(7424), 454–7. http://doi.org/10.1038/nature11508

Carter, J. J., Daugherty, M. D., Qi, X., Bheda-Malge, A., Wipf, G. C., & Robinson, K. (2013). Identification of an overprinting gene in Merkel cell polyomavirus provides evolutionary insight into the birth of viral genes. *Proc Natl Acad Sci U S A*, *110*. article. http://doi.org/10.1073/pnas.1303526110

Carvunis, A.-R. R., Rolland, T., Wapinski, I., Calderwood, M. A., Yildirim, M. A., Simonis, N., … Vidal, M. (2012). Proto-genes and de novo gene birth. *Nature*, *487*(7407), 370–4. article. http://doi.org/10.1038/nature11184

Castaneda, J., Genzor, P., van der Heijden, G. W., Sarkeshik, A., Yates, J. R. 3rd, Ingolia, N. T., & Bortvin, A. (2014). Reduced pachytene piRNAs and translation underlie spermiogenic arrest in Maelstrom mutant mice. *The EMBO Journal*, *33*(18), 1999–2019. Journal

Article, Research Support, N.I.H., Extramural, Research Support, Non-U.S. Gov't. http://doi.org/10.15252/embj.201386855

Charpentier, M., Croyal, M., Carbonnelle, D., Fortun, A., Florenceau, L., Rabu, C., … Lang, F. (2016). IRES-dependent translation of the long non coding RNA meloe in melanoma cells produces the most immunogenic MELOE antigens.

Chen, J.-Y., Shen, Q. S., Zhou, W.-Z., Peng, J., He, B. Z., Li, Y., … Li, C.-Y. (2015). Emergence, Retention and Selection: A Trilogy of Origination for Functional De Novo Proteins from Ancestral LncRNAs in Primates. *PLoS Genetics*, *11*(7), e1005391. http://doi.org/10.1371/journal.pgen.1005391

Chen, S., Zhang, Y. E., & Long, M. (2010). New genes in Drosophila quickly become essential. *Science (New York, N.Y.)*, *330*(6011), 1682–5. http://doi.org/10.1126/science.1196380

Chen, S.-T., Cheng, H.-C., Barbash, D. A., & Yang, H.-P. (2007). Evolution of hydra, a recently evolved testis-expressed gene with nine alternative first exons in Drosophila melanogaster. *PLoS Genetics*, *3*(7), e107. Journal Article, Research Support, Non-U.S. Gov't. http://doi.org/10.1371/journal.pgen.0030107

Chew, G.-L.-. L., Pauli, A., Rinn, J. L., Regev, A., Schier, A. F., & Valen, E. (2013). Ribosome profiling reveals resemblance between long non-coding RNAs and 5{\textasciiacutex} leaders of coding RNAs. *Development.*, *140*(13), 2828–34. article. http://doi.org/10.1242/dev.098343

Cho, J., Yu, N.-K., Choi, J.-H., Sim, S.-E., Kang, S. J., Kwak, C., … Kaang, B.-K. (2015). Multiple repressive mechanisms in the hippocampus during memory formation. *Science (New York, N.Y.)*, *350*(6256), 82–87. Journal Article, Research Support, Non-U.S. Gov't. http://doi.org/10.1126/science.aac7368

Chooniedass-Kothari, S., Emberley, E., Hamedani, M. K., Troup, S., Wang, X., Czosnek, A., … Leygue, E. (2004). The steroid receptor

RNA activator is the first functional RNA encoding a protein. *Febs Letters*, *566*(1–3), 43–47. article.

Chorev, M., & Carmel, L. (2012). The function of introns. *Frontiers in Genetics*, *3*, 55. http://doi.org/10.3389/fgene.2012.00055

Choudhary, J. S., Blackstock, W. P., Creasy, D. M., & Cottrell, J. S. (2001). Interrogating the human genome using uninterpreted mass spectrometry data. *Proteomics*, *1*(5), 651–667. article. http://doi.org/10.1002/1615-9861(200104)1:5<651::AID-PROT651>3.0.CO;2-N

Chung, W.-Y., Wadhawan, S., Szklarczyk, R., Pond, S. K., & Nekrutenko, A. (2007). A first look at ARFome: dual-coding genes in mammalian genomes. *PLoS Comput Biol*, *3*(5), e91. article.

Clemson, C. M., Hutchinson, J. N., Sara, S. A., Ensminger, A. W., Fox, A. H., Chess, A., & Lawrence, J. B. (2009). An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. *Molecular Cell*, *33*(6), 717–726. Journal Article, Research Support, N.I.H., Extramural, Research Support, Non-U.S. Gov't. http://doi.org/10.1016/j.molcel.2009.01.026

Conaco, C., Tsoulfas, P., Sakarya, O., Dolan, A., Werren, J., & Kosik, K. S. (2016). Detection of Prokaryotic Genes in the <italic>Amphimedon queenslandica</italic> Genome. *PLoS ONE*, *11*(3), e0151092. JOUR. Retrieved from http://dx.doi.org/10.1371%2Fjournal.pone.0151092

Consortium, T. E. P. (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, *306*(5696), 636–640. JOUR. Retrieved from http://science.sciencemag.org/content/306/5696/636.abstract

Consortium, T. E. P. (2011). A User's Guide to the Encyclopedia of DNA Elements (ENCODE). *PLoS Biol*, *9*(4), e1001046. JOUR. Retrieved from http://dx.doi.org/10.1371%2Fjournal.pbio.1001046

7. REFERENCES

Consortium, T. E. P., Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigo, R., & Gingeras, T. R. (2007). Identification and analysis of functional elements in 1{%} of the human genome by the ENCODE pilot project. *Nature*, *447*(7146), 799–816. JOUR. http://doi.org/10.1038/nature05874

Consortium, T. F., others, Pmi, R., & Dgt, C. (2014). A promoter-level mammalian expression atlas. *Nature*, *507*(7493), 462–70. article. http://doi.org/10.1038/nature13182

Consortium, T. G. (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, *45*(6), 580–5. http://doi.org/10.1038/ng.2653

Consortium, T. U. (2014). Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Research*, *42*(1), D191-8. http://doi.org/10.1093/nar/gkt1140

Core, L. J., Waterfall, J. J., & Lis, J. T. (2008). Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. *Science*, *322*(5909), 1845–1848. JOUR. Retrieved from http://science.sciencemag.org/content/322/5909/1845.abstract

Crappé, J., Ndah, E., Koch, A., Steyaert, S., Gawron, D., De Keulenaer, S., … Menschaert, G. (2015). PROTEOFORMER: deep proteome coverage through ribosome profiling and MS integration. *Nucleic Acids Research*, *43*(5), e29. http://doi.org/10.1093/nar/gku1283

Crappé, J., Van Criekinge, W., Trooskens, G., Hayakawa, E., Luyten, W., Baggerman, G., … Menschaert, G. (2013). Combining in silico prediction and ribosome profiling in a genome-wide search for novel putatively coding sORFs. *BMC Genomics*, *14*, 648. http://doi.org/10.1186/1471-2164-14-648

Crick, F. H. . (1958). On protein synthesis. *Symp. Soc. Exp. Biol.*, *XII*, 138–163.

Darwin, C. (1859). On the Origin of Species. *London, Murray Edition.*

Darwin, C. (1868). *The variation of animals and plants under domestication* (Vol. 2). book, O. Judd.

de Mendoza, A., Sebe-Pedros, A., Sestak, M. S., Matejcic, M., Torruella, G., Domazet-Loso, T., & Ruiz-Trillo, I. (2013). Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages. *Proceedings of the National Academy of Sciences*, *110*(50), E4858–E4866. http://doi.org/10.1073/pnas.1311818110

Delaye, L., Deluna, A., Lazcano, A., & Becerra, A. (2008). The origin of a novel gene through overprinting in Escherichia coli. *BMC Evol Biol*, *8*. article. http://doi.org/10.1186/1471-2148-8-31

Demuth, J. P., Bie, T. De, Stajich, J. E., Cristianini, N., & Hahn, M. W. (2006). The Evolution of Mammalian Gene Families, (1). http://doi.org/10.1371/journal.pone.0000085

Deng, W., & Roberts, S. G. E. (2006). Core promoter elements recognized by transcription factor IIB. *Biochemical Society Transactions*, *34*(Pt 6), 1051–3. http://doi.org/10.1042/BST0341051

Denli, A. M., Narvaiza, I., Kerman, B. E., Pena, M., Benner, C., Marchetto, M. C. N., … Gage, F. H. (2015). Primate-Specific ORF0 Contributes to Retrotransposon-Mediated Diversity. *Cell*, *163*(3), 583–593. JOUR. http://doi.org/http://dx.doi.org/10.1016/j.cell.2015.09.025

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., … Guigó, R. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Research*, *22*(9), 1775–1789. article. http://doi.org/10.1101/gr.132159.111

Diaz-Munoz, M. D., Bell, S. E., Fairfax, K., Monzon-Casanova, E., Cunningham, A. F., Gonzalez-Porta, M., … Turner, M. (2015). The RNA-binding protein HuR is essential for the B cell antibody

response. *Nat Immunol*, *16*(4), 415–425. JOUR. Retrieved from http://dx.doi.org/10.1038/ni.3115

Dinger, M. E., Pang, K. C., Mercer, T. R., & Mattick, J. S. (2008). Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS Computational Biology*, *4*(11), e1000176. article. http://doi.org/10.1371/journal.pcbi.1000176

Djebali, S., Davis, C. a, Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., … Gingeras, T. R. (2012). Landscape of transcription in human cells. *Nature*, *489*(7414), 101–8. http://doi.org/10.1038/nature11233

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., … Gingeras, T. R. (2012). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* . JOUR. http://doi.org/10.1093/bioinformatics/bts635

Doherty, A., & McInerney, J. O. (2013). Translational selection frequently overcomes genetic drift in shaping synonymous codon usage patterns in vertebrates. *Molecular Biology and Evolution*, *30*(10), 2263–7. http://doi.org/10.1093/molbev/mst128

Domazet-Loso, T., Brajkovic, J., & Tautz, D. (2007). A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends in Genetics : TIG*, *23*(11), 533–539. Journal Article, Research Support, Non-U.S. Gov't. http://doi.org/10.1016/j.tig.2007.08.014

Domazet-Loso, T., & Tautz, D. (2003). An evolutionary analysis of orphan genes in Drosophila. *Genome Research*, *13*(10), 2213–2219. Journal Article, Research Support, Non-U.S. Gov't. http://doi.org/10.1101/gr.1311003

Donoghue, M. T., Keshavaiah, C., Swamidatta, S. H., & Spillane, C. (2011). Evolutionary origins of Brassicaceae specific genes in Arabidopsis thaliana. *BMC Evolutionary Biology*, *11*(1), 47. http://doi.org/10.1186/1471-2148-11-47

dos Reis, M., & Wernisch, L. (2009). Estimating translational selection in eukaryotic genomes. *Molecular Biology and Evolution*, *26*(2), 451–61. http://doi.org/10.1093/molbev/msn272

Du, Z., Fei, T., Verhaak, R. G. W., Su, Z., Zhang, Y., Brown, M., … Liu, X. S. (2013). Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. *Nature Structural & Molecular Biology*, *20*(7), 908–913. article.

Dujon, B. (1996). The yeast genome project: what did we learn? *Trends in Genetics : TIG*, *12*(7), 263–270. Journal Article, Research Support, Non-U.S. Gov't, Review.

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. a, Doyle, F., … Consortium, T. E. P. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. JOUR. http://doi.org/10.1038/nature11247

Dunn, J. G., Foo, C. K., Belletier, N. G., Gavis, E. R., Weissman, J. S., & Sonenberg, N. (2013). Ribosome profiling reveals pervasive and regulated stop codon readthrough in Drosophila melanogaster. *eLife*, *2*, e01179. article. http://doi.org/10.7554/eLife.01179

Duret, L. (2002). Evolution of synonymous codon usage in metazoans. *Current Opinion in Genetics & Development*, *12*(6), 640–649.

Ekman, D., Björklund, A. K., & Elofsson, A. (2007). Quantification of the elevated rate of domain rearrangements in metazoa. *Journal of Molecular Biology*, *372*(5), 1337–48. http://doi.org/10.1016/j.jmb.2007.06.022

Ekman, D., & Elofsson, A. (2010). Identifying and quantifying orphan protein sequences in fungi. *Journal of Molecular Biology*, *396*(2), 396–405. http://doi.org/10.1016/j.jmb.2009.11.053

Elhaik, E., Sabath, N., & Graur, D. (2006). The "inverse relationship between evolutionary rate and age of mammalian genes" is an artifact of increased genetic distance with rate of evolution and time

of divergence. *Molecular Biology and Evolution*, *23*(1), 1–3. http://doi.org/10.1093/molbev/msj006

Eng, J. K., McCormack, A. L., & Yates, J. R. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, *5*(11), 976–989. article.

Eyre-Walker, A., Woolfit, M., & Phelps, T. (2006). The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics*, *173*(2), 891–900. http://doi.org/10.1534/genetics.106.057570

Ezkurdia, I., Juan, D., Rodriguez, J. M., Frankish, A., Diekhans, M., Harrow, J., … Tress, M. L. (2014). Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Human Molecular Genetics*. http://doi.org/10.1093/hmg/ddu309

Faghihi, M. A., Modarresi, F., Khalil, A. M., Wood, D. E., Sahagan, B. G., Morgan, T. E., … Wahlestedt, C. (2008). Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. *Nature Medicine*, *14*(7), 723–30. http://doi.org/10.1038/nm1784

Fälth, M., Sköld, K., Norrman, M., Svensson, M., Fenyö, D., & Andren, P. E. (2006). SwePep, a database designed for endogenous peptides and mass spectrometry. *Molecular & Cellular Proteomics*, *5*(6), 998–1005. article.

Fatica, A., & Bozzoni, I. (2014). Long non-coding RNAs: new players in cell differentiation and development. *Nat Rev Genet*, *15*(1), 7–21. http://doi.org/10.1038/nrg3606

Faulkner, G. J., Kimura, Y., Daub, C. O., Wani, S., Plessy, C., Irvine, K. M., … Carninci, P. (2009). The regulated retrotransposon transcriptome of mammalian cells. *Nature Genetics*, *41*(5), 563–71. http://doi.org/10.1038/ng.368

Fellner, L., Simon, S., Scherling, C., Witting, M., Schober, S., Polte, C., … Neuhaus, K. (2015). Evidence for the recent origin of a bacterial protein-coding, overlapping orphan gene by evolutionary overprinting. *BMC Evolutionary Biology*, *15*(1), 1–14. article. http://doi.org/10.1186/s12862-015-0558-z

Fickett, J. W., & Tung, C. S. (1992). Assessment of protein coding measures. *Nucleic Acids Research*, *20*(24), 6441–6450.

Fiers, W., Contreras, R., De Wachter, R., Haegeman, G., Merregaert, J., Jou, W. M., & Vandenberghe, A. (1971). Recent progress in the sequence determination of bacteriophage MS2 RNA. *Biochimie*, *53*(4), 495–506. article.

Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., … others. (1976). Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature*, *260*(5551), 500–507. article.

Fischer, D., & Eisenberg, D. (1999). Finding families for genomic ORFans. *Bioinformatics (Oxford, England)*, *15*(9), 759–762. Journal Article.

Flicek, P., Ahmed, I., Amode, M. R., Barrell, D., Beal, K., Brent, S., … others. (2012). Ensembl 2013. *Nucleic Acids Research*, gks1236. article.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., … Searle, S. M. J. (2012). Ensembl 2012. *Nucleic Acids Research*, *40*(Database issue), D84-90. http://doi.org/10.1093/nar/gkr991

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L., & Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, *151*(4), 1531–1545. Journal Article, Research Support, U.S. Gov't, Non-P.H.S., Research Support, U.S. Gov't, P.H.S., Review.

7. REFERENCES

Frith, M. C., Forrest, A. R., Nourbakhsh, E., Pang, K. C., Kai, C., Kawai, J., … Grimmond, S. M. (2006). The abundance of short proteins in the mammalian proteome. *PLoS Genetics*, *2*(4), e52. article. http://doi.org/10.1371/journal.pgen.0020052

Galindo, M. I., Pueyo, J. I., Fouix, S., Bishop, S. A., & Couso, J. P. (2007). Peptides encoded by short ORFs control development and define a new eukaryotic gene family. *PLoS Biology*, *5*(5), e106. article. http://doi.org/10.1371/journal.pbio.0050106

Gayà-Vidal, M., & Albà, M. M. (2014). Uncovering adaptive evolution in the human lineage. *BMC Genomics*, *15*(1), 599. http://doi.org/10.1186/1471-2164-15-599

Gerhard, D. S. (2004). The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). *Genome Res.*, *14*, 2121–2127. JOUR. Retrieved from http://dx.doi.org/10.1101/gr.2596504

Gibb, E. A., Brown, C. J., & Lam, W. L. (2011). The functional role of long non-coding RNA in human carcinomas. *Molecular Cancer*, *10*(1), 1. article.

Gilad, Y., Man, O., & Glusman, G. (2005). A comparison of the human and chimpanzee olfactory receptor gene repertoires. *Genome Research*, *15*(2), 224–230. Comparative Study, Journal Article, Research Support, Non-U.S. Gov't. http://doi.org/10.1101/gr.2846405

Godet, Y., Moreau-Aubry, A., Guilloux, Y., Vignard, V., Khammari, A., Dreno, B., … Labarriere, N. (2008). MELOE-1 is a new antigen overexpressed in melanomas and involved in adoptive T cell transfer efficiency. *The Journal of Experimental Medicine*, *205*(11), 2673–2682. http://doi.org/10.1084/jem.20081356

Gong, C., & Maquat, L. E. (2011). lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu

elements. *Nature*, *470*(7333), 284–8. http://doi.org/10.1038/nature09701

Gonzalez, C., Sims, J. S., Hornstein, N., Mela, A., Garcia, F., Lei, L., … Sims, P. a. (2014). Ribosome profiling reveals a cell-type-specific translational landscape in brain tumors. *J Neurosci.*, *34*(33), 10924–36. article. http://doi.org/10.1523/JNEUROSCI.0084-14.2014

Gotea, V., Petrykowska, H. M., & Elnitski, L. (2013). Bidirectional Promoters as Important Drivers for the Emergence of Species-Specific Transcripts. *PLoS ONE*, *8*(2). http://doi.org/10.1371/journal.pone.0057323

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. a, Amit, I., … Regev, A. (2011). Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nature Biotech.*, *29*(7), 644–652. JOUR. http://doi.org/10.1038/nbt.1883

Grote, P., & Herrmann, B. G. (2015). Long noncoding RNAs in organogenesis: making the difference. *Trends in Genetics*, *31*(6), 329–335. JOUR. http://doi.org/http://dx.doi.org/10.1016/j.tig.2015.02.002

Guerzoni, D., & McLysaght, A. (2016). De novo genes arise at a slow but steady rate along the primate lineage and have been subject to incomplete lineage sorting. *Genome Biology and Evolution*, evw074. http://doi.org/10.1093/gbe/evw074

Guo, H., Ingolia, N. T., Weissman, J. S., & Bartel, D. P. (2010). Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*, *466*(7308), 835–40. http://doi.org/10.1038/nature09267

Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., Feldser, D., … Lander, E. S. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, *458*(7235), 223–7. Retrieved from

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2754849&tool=pmcentrez&rendertype=abstract

Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J., Adiconis, X., … Regev, A. (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotechnology*, *28*(5), 503–10. JOUR. http://doi.org/10.1038/nbt.1633

Guttman, M., & Rinn, J. L. (2012). Modular regulatory principles of large non-coding RNAs. *Nature*, *482*(7385), 339–46. article. http://doi.org/10.1038/nature10887

Guttman, M., Russell, P., Ingolia, N. T., Weissman, J. S., & Lander, E. S. (2013). Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell*, *154*(1), 240–51. http://doi.org/10.1016/j.cell.2013.06.009

Haas, B. J., & Zody, M. C. (2010). Advancing RNA-seq analysis. *Nature Biotech.*, *28*(5), 421–423. JOUR. http://doi.org/10.1038/nbt0510-421

Haerty, W., & Ponting, C. P. (2015). Unexpected selection to retain high GC content and splicing enhancers within exons of multiexonic lncRNA loci. *RNA*, *21*(3), 320–332. article.

Hanada, K., Higuchi-Takeuchi, M., Okamoto, M., Yoshizumi, T., Shimizu, M., Nakaminami, K., … Matsui, M. (2013). Small open reading frames associated with morphogenesis are hidden in plant genomes. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(6), 2395–400. http://doi.org/10.1073/pnas.1213958110

Harbers, M., & Carninci, P. (2005). Tag-based approaches for transcriptome research and genome annotation. *Nature Methods*, *2*, 495–502. JOUR. Retrieved from http://dx.doi.org/10.1038/nmeth768

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., … Hubbard, T. J. (2012). GENCODE: the reference

human genome annotation for The ENCODE Project. *Genome Research*, *22*(9), 1760–74. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3431492&tool=pmcentrez&rendertype=abstract

Hashimoto, Y., Niikura, T., Tajima, H., Yasukawa, T., Sudo, H., Ito, Y., … Nishimoto, I. (2001). A rescue factor abolishing neuronal cell death by a wide spectrum of familial Alzheimer's disease genes and Abeta. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(11), 6336–41. http://doi.org/10.1073/pnas.101133498

Heesch, S., Iterson, M., Jacobi, J., Boymans, S., Essers, P. B., Bruijn, E., … Simonis, M. (2014). Extensive localization of long noncoding RNAs to the cytosol and mono- and polyribosomal complexes. *Genome Biol.*, *15*(1), R6. article. http://doi.org/10.1186/gb-2014-15-1-r6

Heinemann, J. A., & Bungard, R. A. (2006). Horizontal Gene Transfer. In *Reviews in Cell Biology and Molecular Medicine*. CHAP, Wiley-VCH Verlag GmbH & Co. KGaA. http://doi.org/10.1002/3527600906.mcb.200400141

Heinen, T. J. a J., Staubach, F., Häming, D., & Tautz, D. (2009). Emergence of a new gene from an intergenic region. *Current Biology : CB*, *19*(18), 1527–31. http://doi.org/10.1016/j.cub.2009.07.049

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., … Glass, C. K. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell*, *38*(4), 576–89. http://doi.org/10.1016/j.molcel.2010.05.004

Hershberg, R., & Petrov, D. A. (2008). Selection on codon bias. *Annual Review of Genetics*, *42*, 287–299. http://doi.org/10.1146/annurev.genet.42.110807.091442

7. REFERENCES

Hezroni, H., Koppstein, D., Schwartz, M. G. G., Avrutin, A., Bartel, D. P. P., & Ulitsky, I. (2015). Principles of Long Noncoding RNA Evolution Derived from Direct Comparison of Transcriptomes in 17 Species. *Cell Reports*, *11*, 1–13. article. http://doi.org/10.1016/j.celrep.2015.04.023

Hotopp, J. C. D., Clark, M. E., Oliveira, D. C. S. G., Foster, J. M., Fischer, P., Torres, M. C. M., … Werren, J. H. (2007). Widespread Lateral Gene Transfer from Intracellular Bacteria to Multicellular Eukaryotes. *Science*, *317*(5845), 1753–1756. JOUR. Retrieved from http://science.sciencemag.org/content/317/5845/1753.abstract

Housman G, & Ulitsky I. (2016). Methods for distinguishing between protein-coding and long noncoding RNAs and the elusive purpose of translation of long noncoding RNAs. *Biochim Biophys Acta*, *1859*(1), 31–40.

Huang, Y., Ainsley, J. A., Reijmers, L. G., & Jackson, F. R. (2013). Translational profiling of clock cells reveals circadianly synchronized protein synthesis. *PLoS Biology*, *11*(11), e1001703. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3864454&tool=pmcentrez&rendertype=abstract

Imanishi, T., Itoh, T., Suzuki, Y. Y., O'Donovan, C., Fukuchi, S., Koyanagi, K. O., … Sugano, S. (2004). Integrative Annotation of 21,037 Human Genes Validated by Full-Length cDNA Clones. *PLoS Biol*, *2*(6), e162. JOUR. Retrieved from http://dx.doi.org/10.1371%2Fjournal.pbio.0020162

Ingolia, N. T. (2014). Ribosome profiling: new views of translation, from single codons to genome scale. *Nature Reviews Genetics*, *15*(3), 205–13. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/24468696

Ingolia, N. T., Brar, G. A., Stern-Ginossar, N., Harris, M. S., Talhouarne, G. J. S., Jackson, S. E., … Weissman, J. S. (2014). Ribosome

Profiling Reveals Pervasive Translation Outside of Annotated Protein-Coding Genes. *Cell Reports*, 1–15. http://doi.org/10.1016/j.celrep.2014.07.045

Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S., & Weissman, J. S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science (New York, N.Y.)*, *324*(5924), 218–23. article. http://doi.org/10.1126/science.1168978

Ingolia, N. T., Lareau, L. F., & Weissman, J. S. (2011). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*, *147*(4), 789–802. http://doi.org/10.1016/j.cell.2011.10.002

Iyer, M. K., Niknafs, Y. S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., … Chinnaiyan, A. M. (2015). The landscape of long noncoding RNAs in the human transcriptome. *Nature Genetics*, *47*. article. http://doi.org/10.1038/ng.3192

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

JBS, H. (1932). The causes of evolution. *New York: Harper and Bros.*

Jensen, J. D., Wong, A., & Aquadro, C. F. (2007). Approaches for identifying targets of positive selection. *TRENDS in Genetics*, *23*(11), 568–577. article.

Ji, Z., Song, R., Regev, A., & Struhl, K. (2015). Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *eLife*, *4*, e08890. JOUR. http://doi.org/10.7554/eLife.08890

Jiang, X., Li, D., Yang, J. J., Wen, J., Chen, H., Xiao, X., … Tang, Y. (2011). Characterization of a novel human testis-specific gene: testis developmental related gene 1 (TDRG1). *The Tohoku Journal of Experimental Medicine*, *225*(4), 311–318. Journal Article, Research Support, Non-U.S. Gov't.

7. REFERENCES

Johannsen, W. (1909). Elemente der exakten Erblichkeitslehre. *Gustav Fischer, Jena.*

Johnstone, T. G., Bazzini, A. A., & Giraldez, A. J. (2016). Upstream ORFs are prevalent translational repressors in vertebrates. *The EMBO Journal*. JOUR. Retrieved from http://emboj.embopress.org/content/early/2016/02/19/embj.201592759.abstract

Jordan, I. K., Rogozin, I. B., Glazko, G. V, & Koonin, E. V. (2003). Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends in Genetics*, *19*(2), 68–72. article.

Juntawong, P., Girke, T., Bazin, J., & Bailey-Serres, J. (2014). Translational dynamics revealed by genome-wide profiling of ribosome footprints in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(1), E203-12. http://doi.org/10.1073/pnas.1317811111

Kaessmann, H. (2010). Origins, evolution, and phenotypic impact of new genes. *Genome Research*, *20*(10), 1313–26. http://doi.org/10.1101/gr.101386.109

Kaessmann, H., Vinckenbosch, N., & Long, M. (2009). RNA-based gene duplication: mechanistic and evolutionary insights. *Nat Rev Genet*, *10*(1), 19–31. JOUR. Retrieved from http://dx.doi.org/10.1038/nrg2487

Kall, L., Canterbury, J. D., Weston, J., Noble, W. S., & MacCoss, M. J. (2007). Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Meth*, *4*(11), 923–925. JOUR. Retrieved from http://dx.doi.org/10.1038/nmeth1113

Kapranov, P., Cheng, J., Dike, S., Nix, D. A., Duttagupta, R., Willingham, A. T., … Gingeras, T. R. (2007). RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science (New York, N.Y.)*, *316*(5830), 1484–8. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17510325

Kapusta, A., Kronenberg, Z., Lynch, V. J., Zhuo, X., Ramsay, L. a., Bourque, G., … Feschotte, C. (2013). Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs. *PLoS Genetics*, *9*(4), e1003470. http://doi.org/10.1371/journal.pgen.1003470

Karapetyan, A. R., Buiting, C., Kuiper, R. A., & Coolen, M. W. (2013). Regulatory Roles for Long ncRNA and mRNA. *Cancers*, *5*(2), 462–490. Journal Article. http://doi.org/10.3390/cancers5020462

Kastenmayer, J. P., Ni, L., Chu, A., Kitchen, L. E., Au, W., Yang, H., … Basrai, M. A. (2006). Functional genomics of genes with small open reading frames (sORFs) in S . cerevisiae, 365–373. http://doi.org/10.1101/gr.4355406.7

Kazemian, M., Ren, M., Lin, J.-X., Liao, W., Spolski, R., & Leonard, W. J. (2015). Comprehensive assembly of novel transcripts from unmapped human RNA-Seq data and their association with cancer. *Molecular Systems Biology*, *11*(8), 826. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/26253570

Keller, A., Nesvizhskii, A. I., Kolker, E., & Aebersold, R. (2002). Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical Chemistry*, *74*(20), 5383–5392. article.

Khalturin, K., Hemmrich, G., Fraune, S., Augustin, R., & Bosch, T. C. G. (2009). More than just orphans: are taxonomically-restricted genes important in evolution? *Trends in Genetics : TIG*, *25*(9), 404–13. http://doi.org/10.1016/j.tig.2009.07.006

Khorkova, O., Myers, A. J., Hsiao, J., & Wahlestedt, C. (2014). Natural antisense transcripts. *Human Molecular Genetics*. http://doi.org/10.1093/hmg/ddu207

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nat Meth*, *12*(4), 357–360. JOUR. Retrieved from http://dx.doi.org/10.1038/nmeth.3317

7. REFERENCES

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, *14*(4), R36. http://doi.org/10.1186/gb-2013-14-4-r36

Kim, M.-S.-. S., Pinto, S. M., Getnet, D., Nirujogi, R. S., Manda, S. S., Chaerkady, R., … Pandey, A. (2014). A draft map of the human proteome. *Nature*, *509*(7502), 575–581. article. http://doi.org/10.1038/nature13302

Kistler, W. S., Baas, D., Lemeille, S., Paschaki, M., Seguin-Estevez, Q., Barras, E., … others. (2015). RFX2 is a major transcriptional regulator of spermiogenesis. *PLoS Genet*, *11*(7), e1005368. article.

Kleene, K. C. (2005). Sexual selection, genetic conflict, selfish genes, and the atypical patterns of gene expression in spermatogenic cells. *Developmental Biology*, *277*(1), 16–26. Journal Article, Research Support, U.S. Gov't, Non-P.H.S., Review. http://doi.org/10.1016/j.ydbio.2004.09.031

Knowles, D. G., & Mclysaght, A. (2009). Recent de novo origin of human protein-coding genes. *Genome Research*, *19*(10), 1752–9. http://doi.org/10.1101/gr.095026.109

Kodzius, R. (2006). CAGE: cap analysis of gene expression. *Nature Methods*, *3*, 211–222. JOUR. Retrieved from http://dx.doi.org/10.1038/nmeth0306-211

Kondo, T., Hashimoto, Y., Kato, K., Inagaki, S., Hayashi, S., & Kageyama, Y. (2007). Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. *Nature Cell Biology*, *9*(6), 660–5. article. http://doi.org/10.1038/ncb1595

Kong, L., Zhang, Y., Ye, Z.-Q., Liu, X.-Q., Zhao, S.-Q., Wei, L., & Gao, G. (2007). CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Research*, *35*(Web Server issue), W345–W349. JOUR. http://doi.org/10.1093/nar/gkm391

Kretz, M., Siprashvili, Z., Chu, C., Webster, D. E., Zehnder, A., Qu, K., … Khavari, P. A. (2013). Control of somatic tissue differentiation by the long non-coding RNA TINCR. *Nature*, *493*(7431), 231–5. http://doi.org/10.1038/nature11661

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., … Marra, M. A. (2009). Circos: an information aesthetic for comparative genomics. *Genome Research*, *19*(9), 1639–1645. article.

Ku, A. (2011). C ON D E T RI - A Content Dependent Read Trimmer for Illumina Data, *6*(10). http://doi.org/10.1371/Citation

Kumar, A. (2009). An overview of nested genes in eukaryotic genomes. *Eukaryotic Cell*, *8*(9), 1321–9. http://doi.org/10.1128/EC.00143-09

Kumari, P., & Sampath, K. (2015). cncRNAs: Bi-functional RNAs with protein coding and non-coding functions. *Seminars in Cell & Developmental Biology*, *47–48*, 40–51. JOUR. http://doi.org/10.1016/j.semcdb.2015.10.024

Kung, J. T. Y., Colognori, D., & Lee, J. T. (2013). Long noncoding RNAs: past, present, and future. *Genetics*, *193*(3), 651–669. article.

Kutter, C., Watt, S., Stefflova, K., Wilson, M. D., Goncalves, A., Ponting, C. P., … Marques, A. C. (2012). Rapid turnover of long noncoding RNAs and the evolution of gene expression. *PLoS Genetics*, *8*(7), e1002841. article. http://doi.org/10.1371/journal.pgen.1002841

Lahn, B. T., & Page, D. C. (1999). Retroposition of autosomal mRNA yielded testis-specific gene family on human Y chromosome. *Nat Genet*, *21*(4), 429–433. JOUR. Retrieved from http://dx.doi.org/10.1038/7771

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., … International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human

genome. *Nature*, *409*(6822), 860–921. http://doi.org/10.1038/35057062

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, *10*(3), R25. http://doi.org/10.1186/gb-2009-10-3-r25

Lanz, R. B., McKenna, N. J., Onate, S. A., Albrecht, U., Wong, J., Tsai, S. Y., … O'Malley, B. W. (1999). A steroid receptor coactivator, SRA, functions as an RNA and is present in an SRC-1 complex. *Cell*, *97*(1), 17–27. article.

Lasda, E., Parker, R., & Parker, R. O. Y. (2014). Circular RNAs : diversity of form and function Circular RNAs : diversity of form and function, 1829–1842. http://doi.org/10.1261/rna.047126.114.DIFFERENT

Lee, C., Yen, K., & Cohen, P. (2013). Humanin: a harbinger of mitochondrial-derived peptides? *Trends in Endocrinology and Metabolism: TEM*, *24*(5), 222–8. http://doi.org/10.1016/j.tem.2013.01.005

Lee, J. T., & Bartolomei, M. S. (2013). X-inactivation, imprinting, and long noncoding RNAs in health and disease. *Cell*, *152*(6), 1308–23. http://doi.org/10.1016/j.cell.2013.02.016

Lepoivre, C., Belhocine, M., Bergon, A., Griffon, A., Yammine, M., Vanhille, L., … Spicuglia, S. (2013). Divergent transcription is associated with promoters of transcriptional regulators. *BMC Genomics*, *14*, 914. http://doi.org/10.1186/1471-2164-14-914

Levin, J. Z. (2010). Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nature Methods*, *7*, 709–715. JOUR. Retrieved from http://dx.doi.org/10.1038/nmeth.1491

Levine, M. T., Jones, C. D., Kern, A. D., Lindfors, H. a, & Begun, D. J. (2006). Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased

expression. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(26), 9935–9. http://doi.org/10.1073/pnas.0509809103

Lewin, B. (2007). Genes IX. *Jones and Bartlett, Sudbury, Massachusetts*. article.

Li, C.-Y., Zhang, Y., Wang, Z., Zhang, Y., Cao, C., Zhang, P.-W., … Wei, L. (2010). A human-specific de novo protein-coding gene associated with human brain functions. *PLoS Computational Biology*, *6*(3), e1000734. http://doi.org/10.1371/journal.pcbi.1000734

Li, D., Dong, Y., Jiang, Y., Jiang, H., Cai, J., & Wang, W. (2010). A de novo originated gene depresses budding yeast mating pathway and is repressed by the protein encoded by its antisense strand. *Cell Research*, *20*(4), 408–20. http://doi.org/10.1038/cr.2010.31

Li, D., Yan, Z., Lu, L., Jiang, H., & Wang, W. (2014). Pleiotropy of the de novo-originated gene MDF1. *Scientific Reports*, *4*, 7280. http://doi.org/10.1038/srep07280

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows--Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. article.

Li, L., Foster, C. M., Gan, Q., Nettleton, D., James, M. G., Myers, A. M., & Wurtele, E. S. (2009). Identification of the novel protein QQS as a component of the starch metabolic network in Arabidopsis leaves. *The Plant Journal*, *58*(3), 485–498. article.

Li, L., & Wurtele, E. S. (2015). The QQS orphan gene of Arabidopsis modulates carbon and nitrogen allocation in soybean. *Plant Biotechnology Journal*, *13*(2), 177–187. Journal Article, Research Support, Non-U.S. Gov't, Research Support, U.S. Gov't, Non-P.H.S. http://doi.org/10.1111/pbi.12238

Light, S., Basile, W., & Elofsson, A. (2014). Orphans and new gene origination, a structural and evolutionary perspective. *Current*

Opinion in Structural Biology, *26C*, 73–83. http://doi.org/10.1016/j.sbi.2014.05.006

Lin, M. F., Carlson, J. W., Crosby, M. A., Matthews, B. B., Yu, C., Park, S., … others. (2007). Revisiting the protein-coding gene catalog of Drosophila melanogaster using 12 fly genomes. *Genome Research*, *17*(12), 1823–1836. article.

Lin, M. F., Jungreis, I., & Kellis, M. (2011). PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* , *27*(13), i275–i282. JOUR. http://doi.org/10.1093/bioinformatics/btr209

Lin, S., Lin, Y., Nery, J. R., Urich, M. a., Breschi, A., Davis, C. a., … Snyder, M. P. (2014). Comparison of the transcriptional landscapes between human and mouse tissues. *Proceedings of the National Academy of Sciences*, 201413624. http://doi.org/10.1073/pnas.1413624111

Liu, J., Hutchison, K., Perrone-Bizzozero, N., Morgan, M., Sui, J., & Calhoun, V. (2010). Identification of genetic and epigenetic marks involved in population structure. *PloS One*, *5*(10), e13209. http://doi.org/10.1371/journal.pone.0013209

Liu, J., Jung, C., Xu, J., Wang, H., Deng, S., Bernad, L., … Chua, N.-H. (2012). Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in Arabidopsis. *The Plant Cell*, *24*(11), 4333–45. http://doi.org/10.1105/tpc.112.102855

Liu, J., Zhang, Y., Lei, X., & Zhang, Z. (2008). Natural selection of protein structural and functional properties: a single nucleotide polymorphism perspective. *Genome Biology*, *9*(4), 1–17. article. http://doi.org/10.1186/gb-2008-9-4-r69

Long, M., Betran, E., Thornton, K., & Wang, W. (2003). The origin of new genes: glimpses from the young and old. *Nature Reviews. Genetics*, *4*(11), 865–875. Journal Article, Research Support, Non-

U.S. Gov't, Research Support, U.S. Gov't, Non-P.H.S., Research Support, U.S. Gov't, P.H.S., Review. http://doi.org/10.1038/nrg1204

Long, M., VanKuren, N. W., Chen, S., & Vibranovski, M. D. (2013). New gene evolution: little did we know. *Annual Review of Genetics*, *47*, 307–33. http://doi.org/10.1146/annurev-genet-111212-133301

Lynch, M., & Marinov, G. K. (2015). The bioenergetic costs of a gene. *Proceedings of the National Academy of Sciences*, *112*(51), 201514974. http://doi.org/10.1073/pnas.1514974112

Ma, J., Diedrich, J. K., Jungreis, I., Donaldson, C., Vaughan, J., Kellis, M., … Saghatelian, A. (2016). Improved Identification and Analysis of Small Open Reading Frame Encoded Polypeptides. *Analytical Chemistry*. JOUR. http://doi.org/10.1021/acs.analchem.6b00191

Ma, J., Ward, C. C., Jungreis, I., Slavo, S. A., Schwaid, A. G., Neveu, J., … Saghatelian, A. (2014). Discovery of Human sORF-Encoded Polypeptides (SEPs) in Cell Lines and Tissue. *Journal of Proteome Research*. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/24490786

Mackowiak, S. D., Zauber, H., Bielow, C., Thiel, D., Kutz, K., Calviello, L., … Obermayer, B. (2015). Extensive identification and analysis of conserved small ORFs in animals. *Genome Biology*, *16*(1), 1–21. article. http://doi.org/10.1186/s13059-015-0742-x

Maestre, J., Tchénio, T., Dhellin, O., & Heidmann, T. (1995). mRNA retroposition in human cells: processed pseudogene formation. *The EMBO Journal*, *14*(24), 6333–6338. JOUR. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC394758/

Magadum, S., Banerjee, U., Murugan, P., Gangapur, D., & Ravikesavan, R. (2013). Gene duplication as a major force in evolution. *Journal of Genetics*, *92*(1), 155–161. Journal Article, Review.

Magny, E. G., Pueyo, J. I., Pearl, F. M. G., Cespedes, M. A., Niven, J. E., Bishop, S. a, & Couso, J. P. (2013). Conserved regulation of cardiac

calcium uptake by peptides encoded in small open reading frames. *Science (New York, N.Y.)*, *341*(6150), 1116–20. http://doi.org/10.1126/science.1238802

Marguerat, S., & Bahler, J. (2010). RNA-seq: from technology to biology. *Cell. Mol. Life Sci.*, *67*, 569–579. JOUR. Retrieved from http://dx.doi.org/10.1007/s00018-009-0180-6

Marques, A. C., Dupanloup, I., Vinckenbosch, N., Reymond, A., & Kaessmann, H. (2005). Emergence of Young Human Genes after a Burst of Retroposition in Primates. *PLoS Biol*, *3*(11), e357. JOUR. Retrieved from http://dx.doi.org/10.1371%2Fjournal.pbio.0030357

Marzluff, W. F., Wilusz, J. E., Jnbaptiste, C. K., & Lu, L. Y. (2012). Novel 3' ends that support translation. *Genes & Development*, *26*(22), 2457–60. http://doi.org/10.1101/gad.207233.112

Masel, J. (2006). Cryptic Genetic Variation Is Enriched for Potential Adaptations. *Genetics*, *172*(3), 1985–1991. http://doi.org/10.1534/genetics.105.051649

Matys, V., Kel-Margoulis, O. V, Fricke, E., Liebich, I., Land, S., Barre-Dirrie, a, … Wingender, E. (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Research*, *34*(Database issue), D108-10. http://doi.org/10.1093/nar/gkj143

Mayr, E. (2002). What Evolution is. *London, Phoenix Edition*.

McLysaght, A., & Guerzoni, D. (2015). New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *370*(1678). JOUR. Retrieved from http://rstb.royalsocietypublishing.org/content/370/1678/20140332.abstract

McLysaght, A., & Hurst, L. D. (2016). Open questions in the study of de novo genes: what, how and why. *Nat Rev Genet*, *advance on*. JOUR. Retrieved from http://dx.doi.org/10.1038/nrg.2016.78

Mcmanus, C. J., May, G. E., Spealman, P., & Shteyman, A. (2014). Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast, 422–430. http://doi.org/10.1101/gr.164996.113.Freely

Mendel, J. G. (1866). Verhandlungen des naturforschenden Vereines in Brünn 4 Abhandlungen. *Versuche Über Pflanzenhybriden.*, 3–47.

Metzker, M. L. (2010). Sequencing technologies — the next generation. *Nature Rev. Genet.*, *11*, 31–46. JOUR. Retrieved from http://dx.doi.org/10.1038/nrg2626

Michel, M., & Michel, A. M. (2013). Visualising ribosome profiling and using it for reading frame detection and exploration of eukaryotic translation initiation . by PhD (Science).

Mitra, S. A., Mitra, A. P., & Triche, T. J. (2012). A central role for long non-coding RNA in cancer. *Genomic "dark Matter": Implications for Understanding Human Disease Mechanisms, Diagnostics, and Cures*, 70. article.

Moore, A. D., & Bornberg-Bauer, E. (2012). The dynamics and evolutionary potential of domain loss and emergence. *Molecular Biology and Evolution*, *29*(2), 787–96. http://doi.org/10.1093/molbev/msr250

Murphy, D. N., & McLysaght, A. (2012). De novo origin of protein-coding genes in murine rodents. *PloS One*, *7*(11), e48650. http://doi.org/10.1371/journal.pone.0048650

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., & Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science (New York, N.Y.)*, *320*(5881), 1344–9. Retrieved from

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2951732&tool=pmcentrez&rendertype=abstract

Nakamura, M., & Carninci, P. (2004). [Cap analysis gene expression: CAGE]. *Tanpakushitsu Kakusan Koso*, *49*, 2688–2693. JOUR.

Necsulea, A., Soumillon, M., Warnefors, M., Liechti, A., Daish, T., Zeller, U., … Kaessmann, H. (2014). The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature*. http://doi.org/10.1038/nature12943

Nei, M., & Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution*, *3*, 418–426.

Nekrutenko, A., & Li, W.-H. (2001). Transposable elements are found in a large number of human protein-coding genes. *TRENDS in Genetics*, *17*(11), 619–621. article.

Neme, R., & Tautz, D. (2013). Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. *BMC Genomics*, *14*(1), 117. article. http://doi.org/10.1186/1471-2164-14-117

Neme, R., & Tautz, D. (2014). Evolution: dynamics of de novo gene emergence. *Current Biology : CB*, *24*(6), R238-40. http://doi.org/10.1016/j.cub.2014.02.016

Neme, R., & Tautz, D. (2016). Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to de novo gene emergence. *eLife*, *5*, e09977. JOUR. http://doi.org/10.7554/eLife.09977

Nesvizhskii, A. I. (2014). Proteogenomics: concepts, applications and computational strategies. *Nature Methods*, *11*(11), 1114–1125. http://doi.org/10.1038/nmeth.3144

Neuhaus, K., Landstorfer, R., Fellner, L., Simon, S., Schafferhans, A., Goldberg, T., … Scherer, S. (2016). Translatomics combined with

transcriptomics and proteomics reveals novel functional, recently evolved orphan genes in Escherichia coli O157:H7 (EHEC). *BMC Genomics*, *17*(1), 133. http://doi.org/10.1186/s12864-016-2456-1

Ng, S.-Y., Johnson, R., & Stanton, L. W. (2012). Human long non-coding RNAs promote pluripotency and neuronal differentiation by association with chromatin modifiers and transcription factors. *The EMBO Journal*, *31*(3), 522–33. http://doi.org/10.1038/emboj.2011.459

Nielsen, R., Bustamante, C., Clark, A. G., Glanowski, S., Sackton, T. B., Hubisz, M. J., … others. (2005). A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol*, *3*(6), e170. article.

Nitsche, A., Rose, D., Fasold, M., Reiche, K., & Stadler, P. F. (2015). Comparison of splice sites reveals that long noncoding RNAs are evolutionarily well conserved, 1–12. http://doi.org/10.1261/rna.046342.114.studies

Ohno, S. (1970). *Evolution by gene duplication*. book, New York: Business Media. http://doi.org/10.1007/978-3-642-86659-3

Ohno, S. (1984). Birth of a unique enzyme from an alternative reading frame of the preexisted, internally repetitious coding sequence. *Proceedings of the National Academy of Sciences of the United States of America*, *81*(8), 2421–5.

Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., … Hayashizaki, Y. (2002). Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature*, *420*(6915), 563–73. http://doi.org/10.1038/nature01266

Okoniewski, M. J., & Miller, C. J. (2006). Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. *BMC Bioinformatics*, *7*, 276. JOUR. Retrieved from http://dx.doi.org/10.1186/1471-2105-7-276

7. REFERENCES

Oliver, S. G., van der Aart, Q. J., Agostoni-Carbone, M. L., Aigle, M., Alberghina, L., Alexandraki, D., … Benit, P. (1992). The complete DNA sequence of yeast chromosome III. *Nature*, *357*(6373), 38–46. Comparative Study, Journal Article, Research Support, Non-U.S. Gov't. http://doi.org/10.1038/357038a0

Ovcharenko, I., Loots, G. G., Nobrega, M. A., Hardison, R. C., Miller, W., & Stubbs, L. (2005). Evolution and functional classification of vertebrate gene deserts. *Genome Research*, *15*(1), 137–45. http://doi.org/10.1101/gr.3015505.

Oyama, M., Kozuka-Hata, H., Suzuki, Y., Semba, K., Yamamoto, T., & Sugano, S. (2007). Diversity of Translation Start Sites May Define Increased Complexity of the Human Short ORFeome. *Molecular & Cellular Proteomics*, *6*(6), 1000–1006. http://doi.org/10.1074/mcp.M600297-MCP200

Ozsolak, F., & Milos, P. M. (2011). RNA sequencing: advances, challenges and opportunities. *Nature Rev. Genet.*, *12*, 87–98. JOUR. Retrieved from http://dx.doi.org/10.1038/nrg2934

Palmieri, N., Kosiol, C., & Schlötterer, C. (2014). The life cycle of Drosophila orphan genes. *eLife*, *3*, e01311. http://doi.org/10.7554/eLife.01311

Pan, T., Wu, R., Liu, B., Wen, H., Tu, Z., Guo, J., … Shen, G. (2016). PBOV1 promotes prostate cancer proliferation by promoting G(1)/S transition. *OncoTargets and Therapy*, *9*, 787–795. JOUR. http://doi.org/10.2147/OTT.S92682

Pappin, D. J. C., Creasy, D. M., & Cottrell, J. S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data Proteomics and 2-DE.

Park, S., Hannenhalli, S., Choi, S., Chow, L., Gelinas, R., Broker, T., … Gardner, M. (2014). Conservation in first introns is positively associated with the number of exons within genes and the presence

of regulatory epigenetic signals. *BMC Genomics*, *15*(1), 526. http://doi.org/10.1186/1471-2164-15-526

Pasek, S., Risler, J.-L., & Brézellec, P. (2006). Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins. *Bioinformatics* , *22*(12), 1418–1423. JOUR. http://doi.org/10.1093/bioinformatics/btl135

Paulding, C. A., Ruvolo, M., & Haber, D. A. (2003). The Tre2 (USP6) oncogene is a hominoid-specific gene. *Proceedings of the National Academy of Sciences*, *100*(5), 2507–2511. article.

Pauli, A., Valen, E., Lin, M. F., Garber, M., Vastenhouw, N. L., Levin, J. Z., … Schier, A. F. (2012). Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res*, *3*, 577–591. article. http://doi.org/10.1101/gr.133009.111

Pauli, A., Valen, E., & Schier, A. F. (2014). Identifying (non-)coding RNAs and small peptides: Challenges and opportunities. *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology*, 1–10. http://doi.org/10.1002/bies.201400103

Pavesi, A., Magiorkinis, G., & Karlin, D. G. (2013). Viral proteins originated de novo by overprinting can be identified by codon usage: application to the ``gene nursery'' of Deltaretroviruses. *PLoS Comput Biol*, *9*. article. http://doi.org/10.1371/journal.pcbi.1003162

Pearson, H. (2006). Genetics: what is a gene? *Nature*, *441*(7092), 398–401. article.

Pegueroles, C., Gabaldón, T., Ulitsky, I., Bartel, D., Ponting, C., Oliver, P., … Wickham, H. (2016). Secondary structure impacts patterns of selection in human lncRNAs. *BMC Biology*, *14*(1), 60. http://doi.org/10.1186/s12915-016-0283-0

7. REFERENCES

Peiffer, J. A. (2008). A spatial dissection of the Arabidopsis floral transcriptome by MPSS. *BMC Plant Biol.*, *8*, 43. JOUR. Retrieved from http://dx.doi.org/10.1186/1471-2229-8-43

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotech*, *33*(3), 290–295. JOUR. Retrieved from http://dx.doi.org/10.1038/nbt.3122

Pesole, G. (2008). What is a gene? An updated operational definition. *Gene*, *417*(1–2), 1–4. JOUR. http://doi.org/http://dx.doi.org/10.1016/j.gene.2008.03.010

Pickard, M. R., & Williams, G. T. (2015). Molecular and cellular mechanisms of action of tumour suppressor GAS5 LncRNA. *Genes*, *6*(3), 484–499. article.

Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., … Pritchard, J. K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, *464*(7289), 768–772. JOUR. Retrieved from http://dx.doi.org/10.1038/nature08872

Ponjavic, J., Ponting, C. P., & Lunter, G. (2007). Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Research*, *17*(5), 556–65. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1855172&tool=pmcentrez&rendertype=abstract

Ponting, C. P., Oliver, P. L., & Reik, W. (2009). Evolution and functions of long noncoding RNAs. *Cell*, *136*(4), 629–41. http://doi.org/10.1016/j.cell.2009.02.006

Prat, Y., Fromer, M., Linial, N., & Linial, M. (2009). Codon usage is associated with the evolutionary age of genes in metazoan genomes. *BMC Evolutionary Biology*, *9*(i), 285. http://doi.org/10.1186/1471-2148-9-285

Pruitt, K. D., Brown, G. R., Hiatt, S. M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., … Ostell, J. M. (2014). RefSeq: an update on mammalian reference sequences. *Nucleic Acids Research*, *42*(1), D756-63. http://doi.org/10.1093/nar/gkt1114

Pruitt, K. D., Tatusova, T., Klimke, W., & Maglott, D. R. (2009). NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Research*, *37*(Database), D32–D36. http://doi.org/10.1093/nar/gkn721

Pueyo, J. I., Magny, E. G., Sampson, C. J., Amin, U., Evans, I. R., Bishop, S. A., & Couso, J. P. (2016). Hemotin, a Regulator of Phagocytosis Encoded by a Small ORF and Conserved across Metazoans. *PLoS Biol*, *14*(3), e1002395. JOUR. Retrieved from http://dx.doi.org/10.1371%2Fjournal.pbio.1002395

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, *26*(6), 841–2. article. http://doi.org/10.1093/bioinformatics/btq033

R. Guigo. (1999). DNA composition, codon usage and exon prediction. *Genetics Databases. Oxford (UK): Academic Press.*

R Development Core Team. (2016). R: a language and environment for statistical computing.

Raj, A., Wang, S. H., Shim, H., Harpak, A., Li, Y. I., Engelmann, B., … Pritchard, J. K. (2016). Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. *eLife*, *5*. http://doi.org/10.7554/eLife.13328

Rands, C. M., Meader, S., Ponting, C. P., & Lunter, G. (2014). 8.2% of the Human Genome Is Constrained: Variation in Rates of Turnover across Functional Element Classes in the Human Lineage. *PLoS Genet*, *10*(7), e1004525. JOUR. Retrieved from http://dx.doi.org/10.1371%2Fjournal.pgen.1004525

7. REFERENCES

Reinartz, J. (2002). Massively parallel signature sequencing (MPSS) as a tool for in-depth quantitative gene expression profiling in all organisms. *Brief. Funct. Genomic Proteomic*, *1*, 95–104. JOUR. Retrieved from http://dx.doi.org/10.1093/bfgp/1.1.95

Reinhardt, J. A., Wanjiru, B. M., Brant, A. T., Saelao, P., Begun, D. J., & Jones, C. D. (2013). De novo ORFs in Drosophila are important to organismal fitness and evolved rapidly from previously non-coding sequences. *PLoS Genetics*, *9*(10), e1003860. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3798262&tool=pmcentrez&rendertype=abstract

Rinn, J. L., & Chang, H. Y. (2012). Genome regulation by long noncoding RNAs. *Annual Review of Biochemistry*, *81*. article.

Rinn, J. L., Kertesz, M., Wang, J. K., Squazzo, S. L., Xu, X., Brugmann, S. A., … Chang, H. Y. (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell*, *129*(7), 1311–1323. Journal Article, Research Support, N.I.H., Extramural, Research Support, Non-U.S. Gov't, Research Support, U.S. Gov't, Non-P.H.S. http://doi.org/10.1016/j.cell.2007.05.022

Robertson, G. (2010). De novo assembly and analysis of RNA-seq data. *Nature Methods*, *7*, 909–912. JOUR. Retrieved from http://dx.doi.org/10.1038/nmeth.1517

Royce, T. E., Rozowsky, J. S., & Gerstein, M. B. (2007). Toward a universal microarray: prediction of gene expression through nearest-neighbor probe sequence identification. *Nucleic Acids Res.*, *35*, e99. JOUR. Retrieved from http://dx.doi.org/10.1093/nar/gkm549

Ruiz-Orera, J., Hernandez-Rodriguez, J., Chiva, C., Sabidó, E., Kondova, I., Bontrop, R., … Albà, M. M. (2015). Origins of De Novo Genes in Human and Chimpanzee. *PLOS Genetics*, *11*(12), e1005721. http://doi.org/10.1371/journal.pgen.1005721

Ruiz-Orera, J., Messeguer, X., Subirana, J. A., & Alba, M. M. (2014). Long non-coding RNAs as a source of new peptides. *eLife*, *3*, 1–24. article. http://doi.org/10.7554/eLife.03523

Saghatelian, A., & Couso, J. P. (2015). Discovery and characterization of smORF-encoded bioactive polypeptides. *Nature Chemical Biology*, *11*(12), 909–916. Journal Article. http://doi.org/10.1038/nchembio.1964

Samusik, N., Krukovskaya, L., Meln, I., Shilov, E., & Kozlov, A. P. (2013). PBOV1 is a human de novo gene with tumor-specific expression that is associated with a positive clinical outcome of cancer. *PloS One*, *8*(2), e56162. http://doi.org/10.1371/journal.pone.0056162

Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, J. C., … Smith, M. (1977). Nucleotide sequence of bacteriophage φX174 DNA. *Nature*, *265*(5596), 687–695. http://doi.org/10.1038/265687a0

Saus, E., Brunet-Vega, A., Iraola-Guzmán, S., Pegueroles, C., Gabaldón, T., & Pericay, C. (2016). Long Non-Coding RNAs As Potential Novel Prognostic Biomarkers in Colorectal Cancer. *Frontiers in Genetics*, *7*, 54. JOUR. http://doi.org/10.3389/fgene.2016.00054

Savard, J., Marques-Souza, H., Aranda, M., & Tautz, D. (2006). A segmentation gene in tribolium produces a polycistronic mRNA that codes for multiple conserved peptides. *Cell*, *126*(3), 559–69. article. http://doi.org/10.1016/j.cell.2006.05.053

Schlötterer, C. (2015). Genes from scratch – the evolutionary fate of de novo genes. *Trends in Genetics*, 1–5. http://doi.org/10.1016/j.tig.2015.02.007

Schmid, K. J., & Tautz, D. (1997). A screen for fast evolving genes from Drosophila. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(18), 9746–9750. Journal Article, Research Support, Non-U.S. Gov't.

7. REFERENCES

Schwaid, A. G., Shannon, D. A., Ma, J., Slavoff, S. A., Levin, J. Z., Weerapana, E., & Saghatelian, A. (2013). Chemoproteomic discovery of cysteine-containing human short open reading frames. *Journal of the American Chemical Society*, *135*(45), 16750–3. article. http://doi.org/10.1021/ja406606j

Schwartz, S., Kent, W. J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R. C., … Miller, W. (2003). Human-mouse alignments with BLASTZ. *Genome Research*, *13*(1), 103–7. http://doi.org/10.1101/gr.809403

Scofield, D. G., Hong, X., & Lynch, M. (2007). Position of the final intron in full-length transcripts: determined by NMD? *Molecular Biology and Evolution*, *24*(4), 896–9. http://doi.org/10.1093/molbev/msm010

Sebestyén, E., Singh, B., Miñana, B., Pagès, A., Mateo, F., Pujana, M. A., … Eyras, E. (2016). Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. *Genome Research*, *26*(6), 732–744. http://doi.org/10.1101/gr.199935.115

She, X., Horvath, J. E., Jiang, Z., Liu, G., Furey, T. S., Christ, L., … others. (2004). The structure and evolution of centromeric transition regions within the human genome. *Nature*, *430*(7002), 857–864. article.

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*, *29*(1), 308–11. article. http://doi.org/10.1093/nar/29.1.308

Shidhi, P. R., Suravajhala, P., Nayeema, A., Nair, A. S., Singh, S., & Dhar, P. K. (2014). Making novel proteins from pseudogenes. *Bioinformatics (Oxford, England)*, 1–7. http://doi.org/10.1093/bioinformatics/btu615

Shiraki, T. (2003). Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of

promoter usage. *Proc. Natl Acad. Sci. USA*, *100*, 15776–15781. JOUR. Retrieved from http://dx.doi.org/10.1073/pnas.2136655100

Siepel, A. (2009). Darwinian alchemy : Human genes from noncoding DNA, (607), 1693–1695. http://doi.org/10.1101/gr.098376.109.19

Slack, J. (2014). *Genes: A Very Short Introduction*. book, OUP Oxford.

Slavoff, S. A., Heo, J., Budnik, B. A., Hanakahi, L. A., & Saghatelian, A. (2014). A Human Short Open Reading Frame (sORF)-encoded Polypeptide That Stimulates DNA End Joining. *The Journal of Biological Chemistry*, *289*(16), 10950–7. http://doi.org/10.1074/jbc.C113.533968

Slavoff, S. A., Mitchell, A. J., Schwaid, A. G., Cabili, M. N., Ma, J., Levin, J. Z., … Saghatelian, A. (2013). Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nature Chemical Biology*, *9*(1), 59–64. http://doi.org/10.1038/nchembio.1120.Peptidomic

Smeds, L., & Künstner, A. (2011). ConDeTri--a content dependent read trimmer for Illumina data. *PloS One*, *6*(10), e26314. http://doi.org/10.1371/journal.pone.0026314

Smit, A., Hubley, & P, R. & G. (n.d.). RepeatMasker Open-4.0. *Available: Http://www.repeatmasker.org.*

Smith, J. E., Alvarez-Dominguez, J. R., Kline, N., Huynh, N. J., Geisler, S., Hu, W., … Baker, K. E. (2014). Translation of Small Open Reading Frames within Unannotated RNA Transcripts in Saccharomyces cerevisiae. *Cell Reports*, 1–9. http://doi.org/10.1016/j.celrep.2014.05.023

Smith-Unna, R. D., Boursnell, C., Patro, R., Hibberd, J. M., & Kelly, S. (2015). TransRate: reference free quality assessment of de-novo transcriptome assemblies. *bioRxiv*. JOUR. Retrieved from http://biorxiv.org/content/early/2015/06/27/021626.abstract

7. REFERENCES

Soumillon, M., Necsulea, A., Weier, M., Brawand, D., Zhang, X., Gu, H., … Kaessmann, H. (2013). Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. *Cell Reports*, *3*(6), 2179–90. http://doi.org/10.1016/j.celrep.2013.05.031

Stoltzfus, A. (1999). On the possibility of constructive neutral evolution. *Journal of Molecular Evolution*, *49.2*, 169–181.

Struhl, K. (2007). Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nature Structural & Molecular Biology*, *14*(2), 103–105. article.

Su, W.-Y., Li, J.-T., Cui, Y., Hong, J., Du, W., Wang, Y.-C., … Fang, J.-Y. (2012). Bidirectional regulation between WDR83 and its natural antisense transcript DHPS in gastric cancer. *Cell Research*, *22*(9), 1374–1389. Journal Article, Research Support, Non-U.S. Gov't. http://doi.org/10.1038/cr.2012.57

Suenaga, Y., Islam, S. M. R., Alagu, J., Kaneko, Y., Kato, M., Tanaka, Y., … Nakagawara, A. (2014). NCYM, a Cis-antisense gene of MYCN, encodes a de novo evolved protein that inhibits GSK3β resulting in the stabilization of MYCN in human neuroblastomas. *PLoS Genetics*, *10*(1), e1003996. http://doi.org/10.1371/journal.pgen.1003996

Sun, L., Luo, H., Bu, D., Zhao, G., Yu, K., Zhang, C., … Zhao, Y. (2013). Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Research*, *41*(17), e166. article. http://doi.org/10.1093/nar/gkt646

Sun, S., Del Rosario, B. C., Szanto, A., Ogawa, Y., Jeon, Y., & Lee, J. T. (2013). Jpx RNA activates Xist by evicting CTCF. *Cell*, *153*(7), 1537–51. http://doi.org/10.1016/j.cell.2013.05.028

Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T., Lehner, B., Bailey, T. L., … Adelson, D. L. (2014). Synonymous Mutations Frequently Act as Driver Mutations in Human Cancers. *Cell*, *156*(6), 1324–1335. JOUR. http://doi.org/http://dx.doi.org/10.1016/j.cell.2014.01.051

Tani, H., Torimura, M., & Akimitsu, N. (2013). The RNA degradation pathway regulates the function of GAS5 a non-coding RNA in mammalian cells. *PloS One*, *8*(1), e55684. http://doi.org/10.1371/journal.pone.0055684

Tarek, R., & Garrido, N. (2014). Evolutionary analyses of orphan genes in mouse lineages in the context of de novo gene birth Dissertation Rafik Tarek Neme Garrido.

Tautz, D. (2009). Polycistronic peptide coding genes in eukaryotes--how widespread are they? *Briefings in Functional Genomics & Proteomics*, *8*(1), 68–74. http://doi.org/10.1093/bfgp/eln054

Tautz, D. (2014). The discovery of de novo gene evolution. *Perspect Biol Med*, *57*(1), 149–61. article. http://doi.org/10.1353/pbm.2014.0006

Tautz, D., & Domazet-Lošo, T. (2011). The evolutionary origin of orphan genes. *Nature Reviews. Genetics*, *12*(10), 692–702. http://doi.org/10.1038/nrg3053

Team, R. D. C. (2009). *R: A language and environment for statistical computing*. book, Vienna, Austria: R Foundation for Statistical Computing.

Terrapon, N., Weiner, J., Grath, S., Moore, A. D., & Bornberg-Bauer, E. (2014). Rapid similarity search of proteins using alignments of domain arrangements. *Bioinformatics (Oxford, England)*, *30*(2), 274–281. Journal Article, Research Support, Non-U.S. Gov't. http://doi.org/10.1093/bioinformatics/btt379

Theimer CA, Blois CA, F. J. (2005). Structure of the human telomerase RNA pseudoknot reveals conserved tertiary interactions essential for function. *Molecular Cell*, *17 (5)*, 671–82.

Thomson, T. M., Lozano, J. J., Loukili, N., Carrió, R., Serras, F., Cormand, B., … Guigó, R. (2000). Fusion of the Human Gene for the Polyubiquitination Coeffector UEV1 with Kua, a Newly

Identified Gene. *Genome Research* , *10*(11), 1743–1756. JOUR. http://doi.org/10.1101/gr.GR-1405R

Tian, D., Sun, S., & Lee, J. T. (2010). The long noncoding RNA, Jpx, is a molecular switch for X chromosome inactivation. *Cell*, *143*(3), 390–403. article. http://doi.org/10.1016/j.cell.2010.09.049

Toll-Riera, M., Bosch, N., Bellora, N., Castelo, R., Armengol, L., Estivill, X., & Albà, M. M. (2009). Origin of primate orphan genes: a comparative genomics approach. *Molecular Biology and Evolution*, *26*(3), 603–12. http://doi.org/10.1093/molbev/msn281

Tran, N., Su, H., Khodadadi-jamayran, A., Lin, S., Zhang, L., Zhou, D., … Zhao, X. (2016). The AS-RBM 15 lncRNA enhances RBM 15 protein translation during megakaryocyte differentiation, 1–14.

Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-seq. *Bioinformatics*, *25*, 1105–1111. JOUR. Retrieved from http://dx.doi.org/10.1093/bioinformatics/btp120

Trapnell, C., Williams, B. a, Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., … Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, *28*(5), 511–5. JOUR. http://doi.org/10.1038/nbt.1621

Trinklein, N. D., Aldred, S. F., Hartman, S. J., Schroeder, D. I., Otillar, R. P., & Myers, R. M. (2004). An Abundance of Bidirectional Promoters in the Human Genome. *Genome Research* , *14*(1), 62–66. JOUR. http://doi.org/10.1101/gr.1982804

Tripathi, V., Ellis, J. D., Shen, Z., Song, D. Y., Pan, Q., Watt, A. T., … Prasanth, K. V. (2010). The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Molecular Cell*, *39*(6), 925–38. http://doi.org/10.1016/j.molcel.2010.08.011

Tsai, M.-C., Manor, O., Wan, Y., Mosammaparast, N., Wang, J. K., Lan, F., … Chang, H. Y. (2010). Long noncoding RNA as modular scaffold of histone modification complexes. *Science (New York, N.Y.)*, *329*(5992), 689–93. http://doi.org/10.1126/science.1192002

Tsai, M.-C., Spitale, R. C., & Chang, H. Y. (2011). Long intergenic noncoding RNAs: new links in cancer progression. *Cancer Research*, *71*(1), 3–7. article.

Uesaka, M., Nishimura, O., Go, Y., Nakashima, K., Agata, K., & Imamura, T. (2014). Bidirectional promoters are the major source of gene activation-associated non-coding RNAs in mammals. *BMC Genomics*, *15*, 35. http://doi.org/10.1186/1471-2164-15-35

Ulitsky, I., & Bartel, D. P. (2013). Review lincRNAs : Genomics , Evolution , and Mechanisms.

Ulitsky, I., Shkumatava, A., Jan, C. H., Sive, H., & Bartel, D. P. (2011). Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell*, *147*(7), 1537–50. http://doi.org/10.1016/j.cell.2011.11.055

Vadie, N., Saayman, S., Lenox, A., Ackley, A., Clemson, M., Burdach, J., … Morris, K. V. (n.d.). MYCNOS functions as an antisense RNA regulating MYCN. *RNA Biology*, *0*(ja), 0. article. http://doi.org/10.1080/15476286.2015.1063773

Vanderperre, B., Lucier, J.-F.-. F., Bissonnette, C., Motard, J., Tremblay, G., Vanderperre, S., … Roucou, X. (2013). Direct detection of alternative open reading frames translation products in human significantly expands the proteome. *PloS One*, *8*(8), e70698. article. http://doi.org/10.1371/journal.pone.0070698

Vasquez, J.-J., Hon, C.-C., Vanselow, J. T., Schlosser, A., & Siegel, T. N. (2014). Comparative ribosome profiling reveals extensive translational complexity in different Trypanosoma brucei life cycle stages. *Nucleic Acids Research*, *42*(6), 3623–37. http://doi.org/10.1093/nar/gkt1386

7. REFERENCES

Velculescu, V. E., Zhang, L., Vogelstein, B., & Kinzler, K. W. (1995). Serial analysis of gene expression. *Science*, *270*, 484–487. JOUR. Retrieved from http://dx.doi.org/10.1126/science.270.5235.484

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., … Zhu, X. (2001). The sequence of the human genome. *Science (New York, N.Y.)*, *291*(5507), 1304–51. http://doi.org/10.1126/science.1058040

Vogel, F. (1964). A preliminary estimate of the number of human genes. *Nature*, *201*, 847. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/14161239

Vries, H. De. (1889). Intracellular Pangenesis. *Open Court Pub. Co.*

Wain, H. M., Bruford, E. A., Lovering, R. C., Lush, M. J., Wright, M. W., & Povey, S. (2002). Guidelines for human gene nomenclature. *Genomics*, *79*(4), 464–470. article.

Wallace, A. R. (1858). On the Tendency of Varieties to Depart Indefinitely From the Original Type.

Wang, J., Gong, C., & Maquat, L. E. (2013). Control of myogenesis by rodent SINE-containing lncRNAs. *Genes & Development*, *27*(7), 793–804. http://doi.org/10.1101/gad.212639.112

Wang, J., Xie, G., Singh, M., Ghanbarian, A. T., Rasko, T., Szvetnik, A., … Izsvak, Z. (2014). Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. *Nature*, *516*(7531), 405–409. JOUR. Retrieved from http://dx.doi.org/10.1038/nature13804

Wang, J., Zhang, J., Zheng, H., Li, J., Liu, D., Li, H., … Wong, G. K.-S. (2004). Mouse transcriptome: Neutral evolution of /`non-coding/' complementary DNAs. *Nature*, *431*(7010). JOUR. Retrieved from http://dx.doi.org/10.1038/nature03016

Wang, K. C., Yang, Y. W., Liu, B., Sanyal, A., Corces-Zimmerman, R., Chen, Y., … Chang, H. Y. (2011). A long noncoding RNA maintains

active chromatin to coordinate homeotic gene expression. *Nature*, *472*(7341), 120–4. http://doi.org/10.1038/nature09819

Wang, L., Park, H. J., Dasari, S., Wang, S., Kocher, J.-P., & Li, W. (2013). CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Research*, *41*(6), e74–e74. JOUR. http://doi.org/10.1093/nar/gkt006

Wang, Y., Gan, Y., Tan, Z., Zhou, J., Kitazawa, R., Jiang, X., … Yang, J. (2016). TDRG1 functions in testicular seminoma are dependent on the PI3K/Akt/mTOR signaling pathway. *OncoTargets and Therapy*, *9*, 409. article.

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*(1), 57–63. JOUR. Retrieved from http://dx.doi.org/10.1038/nrg2484

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., … Carninci, P. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, *420*. article. http://doi.org/10.1038/nature01262

Watson, J. D., & Crick, F. H. . (1953). A Structure for Deoxyribose Nucleic Acid. *Nature*, *171*, 737–738.

Wei, W., Pelechano, V., Järvelin, A. I., & Steinmetz, L. M. (2011). Functional consequences of bidirectional promoters. *Trends in Genetics*, *27*(7), 267–276. JOUR. http://doi.org/http://dx.doi.org/10.1016/j.tig.2011.04.002

Weirick, T., Militello, G., Müller, R., John, D., Dimmeler, S., & Uchida, S. (2015). The identification and characterization of novel transcripts from RNA-seq data. *Briefings in Bioinformatics*, (May), 1–8. http://doi.org/10.1093/bib/bbv067

Wiberg, R. A. W., Halligan, D. L., Ness, R. W., Necsulea, A., Kaessmann, H., & Keightley, P. D. (2015). Assessing Recent Selection and Functionality at Long Noncoding RNA Loci in the Mouse Genome.

Genome Biology and Evolution , *7*(8), 2432–2444. JOUR. http://doi.org/10.1093/gbe/evv155

Wilhelm, B. T., & Landry, J. R. (2009). RNA-seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods*, *48*, 249–257. JOUR. Retrieved from http://dx.doi.org/10.1016/j.ymeth.2009.03.016

Wilhelm, M., Schlegl, J., Hahne, H., Moghaddas Gholami, A., Lieberenz, M., Savitski, M. M., … Savitski, M. M. (2014). Mass-spectrometry-based draft of the human proteome. *Nature*, *509*(7502), 582–7. article. http://doi.org/10.1038/nature13319

Wilson, B. A., & Masel, J. (2011). Putatively noncoding transcripts show extensive association with ribosomes. *Genome Biology and Evolution*, *3*, 1245–52. http://doi.org/10.1093/gbe/evr099

Wissler, L., Gadau, J., Simola, D. F., Helmkampf, M., & Bornberg-Bauer, E. (2013). Mechanisms and dynamics of orphan gene emergence in insect genomes. *Genome Biology and Evolution*, *5*(2), 439–55. http://doi.org/10.1093/gbe/evt009

Wu, D.-D., Irwin, D. M., & Zhang, Y.-P. (2011). De novo origin of human protein-coding genes. *PLoS Genetics*, *7*(11), e1002379. http://doi.org/10.1371/journal.pgen.1002379

Wu, X., & Sharp, P. a. (2013). Divergent transcription: a driving force for new gene origination? *Cell*, *155*(5), 990–6. http://doi.org/10.1016/j.cell.2013.10.048

Xiao, W., Liu, H., Li, Y., Li, X., Xu, C., Long, M., & Wang, S. (2009). A Rice Gene of <italic>De Novo</italic> Origin Negatively Regulates Pathogen-Induced Defense Response. *PLoS ONE*, *4*(2), e4603. JOUR. Retrieved from http://dx.doi.org/10.1371%2Fjournal.pone.0004603

Xie, C., Zhang, Y. E., Chen, J.-Y., Liu, C.-J., Zhou, W.-Z., Li, Y., … Li, C.-Y. (2012). Hominoid-Specific De Novo Protein-Coding Genes

Originating from Long Non-Coding RNAs. *PLoS Genetics*, *8*(9), e1002942. http://doi.org/10.1371/journal.pgen.1002942

Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., … Shmueli, O. (2005). Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics*, *21*(5), 650–9. article. http://doi.org/10.1093/bioinformatics/bti042

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, *24*(8), 1586–1591. article.

Yang, Z., & Huang, J. (2011). De novo origin of new genes with introns in Plasmodium vivax. *FEBS Letters*, *585*(4), 641–644. JOUR. http://doi.org/10.1016/j.febslet.2011.01.017

Yang, Z., & Nielsen, R. (2008). Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Molecular Biology and Evolution*, *25*(3), 568–579. http://doi.org/10.1093/molbev/msm284

Yates III, J. R., McCormack, A. L., Schieltz, D., Carmack, E., & Link, A. (1997). Direct analysis of protein mixtures by tandem mass spectrometry. *Journal of Protein Chemistry*, *16*(5), 495–497. article.

Yoon, J.-H., Abdelmohsen, K., Srikantan, S., Yang, X., Martindale, J. L., De, S., … Gorospe, M. (2012). LincRNA-p21 suppresses target mRNA translation. *Molecular Cell*, *47*(4), 648–55. http://doi.org/10.1016/j.molcel.2012.06.027

Yu, C., Dang, Y., Zhou, Z., Wu, C., Zhao, F., Sachs, M., & Liu, Y. (2015). Codon Usage Influences the Local Rate of Translation Elongation to Regulate Co-translational Protein Folding. *Molecular Cell*, *59*(5), 744–754.

Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., … Ren, B. (2014). A comparative encyclopedia of DNA elements in the mouse genome. *Nature*, *515*(7527), 355–364. Comparative Study, Journal

Article, Research Support, N.I.H., Extramural, Research Support, Non-U.S. Gov't, Research Support, U.S. Gov't, Non-P.H.S. http://doi.org/10.1038/nature13992

Zapała, B., Kaczyński, Ł., Kieć-Wilk, B., Staszel, T., Knapp, A., Thoresen, G. H., … Dembińska-Kieć, A. (2010). Humanins, the neuroprotective and cytoprotective peptides with antiapoptotic and anti-inflammatory properties. *Pharmacological Reports : PR*, *62*(5), 767–77. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/21098860

Zhang, Y. E., Vibranovski, M. D., Krinsky, B. H., & Long, M. (2010). Age-dependent chromosomal distribution of male-biased genes in Drosophila. *Genome Research* , *20*(11), 1526–1533. JOUR. http://doi.org/10.1101/gr.107334.110

Zhao, L. L., Saelao, P., Jones, C. D., & Begun, D. J. (2014). Origin and Spread of de Novo Genes in Drosophila melanogaster Populations. *Science*, *769*(6172), 769–72. article. http://doi.org/10.1126/science.1248286

Zhi, H., Ning, S., Li, X., Li, Y., Wu, W., & Li, X. (2014). A novel reannotation strategy for dissecting DNA methylation patterns of human long intergenic non-coding RNAs in cancers. *Nucleic Acids Research*, gku575. article.

Zhou, Q., Zhang, G., Zhang, Y., Xu, S., Zhao, R., Zhan, Z., … Wang, W. (2008). On the origin of new genes in Drosophila. *Genome Research*, *18*(9), 1446–55. http://doi.org/10.1101/gr.076588.108

Zucchelli, S., Cotella, D., Takahashi, H., Carrieri, C., Cimatti, L., Fasolo, F., … Gustincich, S. (2015). SINEUPs: A new class of natural and synthetic antisense long non-coding RNAs that activate translation. *RNA Biology*, *12*(8), 771–779. JOUR. http://doi.org/10.1080/15476286.2015.1060395

La impressió d'aquesta tesi ha estat possible gràcies a l'ajut per a la finalització de tesis doctorals de la Fundació IMIM.