



**HAL**  
open science

## Phylogenomics and Genome Annotation

Anamaria Necsulea

► **To cite this version:**

Anamaria Necsulea. Phylogenomics and Genome Annotation. Scornavacca, Celine; Delsuc, Frédéric; Galtier, Nicolas. Phylogenetics in the Genomic Era, No commercial publisher | Authors open access book, pp.4.1:1–4.1:26, 2020. hal-02535669

**HAL Id: hal-02535669**

**<https://hal.science/hal-02535669v1>**

Submitted on 10 Apr 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

# Chapter 4.1 Phylogenomics and Genome Annotation

Anamaria Necsulea

Laboratoire de Biométrie et Biologie Évolutive  
UMR 5558, CNRS, Université de Lyon, Université Lyon 1  
Villeurbanne, France  
anamaria.necsulea@univ-lyon1.fr  
 <https://orcid.org/0000-0001-9861-7698>

---

## Abstract

Annotating a genome is a challenging endeavor, which aims to describe not only the protein-coding and non-coding gene catalogues, but also other functional elements involved in gene expression regulation, maintenance of genome integrity and genome transmission across generations. Recent technical developments have greatly improved the annotation process by providing large-scale assessments of transcription, translation, chromatin status and tri-dimensional conformation etc. . . Genome-wide maps of various biochemical activities can thus be readily obtained. However, biochemical activity is not synonymous with biological function and many active genomic elements may in fact be dispensable. Genome editing techniques allow for more direct tests of biological functions, but are still costly, time-consuming, and largely limited to phenotypes that can be observed in the laboratory. In this context, evolutionary approaches, which can identify genomic regions under purifying selection to preserve existing functions, or under positive selection following the acquisition of new biological roles, are an important asset for functional genome annotation. While evolutionary analyses cannot determine precise biological functions, they can be used to test for functionality at multiple levels, by assessing selective pressures on primary DNA or RNA sequences, on secondary RNA structures, transcription levels or patterns, transcription factor binding sites etc. . . Here, I review the proven and potential contributions of phylogenomic approaches to genome annotation, focusing on how these methods can be combined with insights from molecular biology and genetics to provide a comprehensive image of functional genomic landscapes.

**How to cite:** Anamaria Necsulea (2020). Phylogenomics and Genome Annotation. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter No. 4.1, pp. 4.1:1–4.1:26. No commercial publisher | Authors open access book. The book is freely available at <https://hal.inria.fr/PGE>.

## 1 Introduction

Understanding how complex biological functions are encoded in the DNA is a fundamental goal of genetics. An important step towards attaining this goal is the process of genome annotation, which aims to describe the localization, structure, biochemical activities and (ideally) biological roles of the functional elements present in a genome.

The scope of genome annotation has expanded in recent years. When the first complete DNA sequences of cellular organisms were obtained (Fleischmann et al., 1995), annotating a genome was largely synonymous with describing its catalogue of protein-coding genes. This endeavor is challenging in itself, as demonstrated by the fact that, almost twenty years after the initial publication of the human genome sequence (Lander et al., 2001), the number of protein-coding genes present in our genome has yet to reach a stable estimate (Pertea et al.,





© Anamaria Necsulea.

Licensed under Creative Commons License CC-BY-NC-ND 4.0.

*Phylogenetics in the genomic era*.

Editors: Celine Scornavacca, Frédéric Delsuc and Nicolas Galtier; chapter No. 4.1; pp. 4.1:1–4.1:26

 A book completely handled by researchers.

 No publisher has been paid.

## 4.1:2 Phylogenomics and Genome Annotation

2018b). For eukaryotes, annotating protein-coding genes is complicated by the presence of complex exon-intron structures and of multiple isoforms for each gene. Expectedly, alternative transcript annotations are even less stable than known gene repertoires, as thousands of new isoforms are added at each genome annotation release for human or mouse (Harrow et al., 2012). Thus, annotating the complete protein-coding gene repertoire is in itself an ambitious aim.

More recently, describing non-coding RNA genes has become an important part of the genome annotation process. Some categories of non-coding RNAs, such as ribosomal or transfer RNAs, which have essential roles in translating messenger RNAs (mRNAs) into proteins, have been extensively studied and are thus generally well annotated in most species (Abe et al., 2014). Other classes of non-coding RNAs are more elusive. These include both small RNAs (such as miRNAs, which regulate gene expression at the post-transcriptional and translational level (He and Hannon, 2004), or piRNAs, which are thought to protect the germline from transposable element invasion (Weick and Miska, 2014)) and large RNAs (such as long non-coding RNAs, which were proposed to act in a multitude of biological processes [Guttman et al. 2009]). In vertebrates, the number of annotated non-coding RNA genes has increased exponentially in the past few years, thanks to the development of sensitive transcriptome sequencing techniques (Wang et al., 2009). For example, the human genome may harbor as many as 60,000 long non-coding RNA (lncRNA) genes (Iyer et al., 2015; Pertea et al., 2018a), which vastly surpasses the number of known protein-coding genes.

Efforts to chart the functional components of a genome now go even beyond establishing a complete protein-coding and non-coding gene list. In addition to gene repertoires, comprehensive genome annotation projects aim to survey elements that are important for gene expression regulation, for the maintenance of genome integrity, genome transmission across generations, etc. . . Such integrative functional annotation projects are in progress for the human and mouse genomes (Carninci et al., 2005; ENCODE Project Consortium et al., 2007), as well as for other model organisms (modENCODE Consortium et al., 2010; Gerstein et al., 2010). These aspects of genome annotation were made possible by technological advances that enabled large-scale surveys of various biochemical activities, such as enhancer activity (Visel et al., 2009), transcription factor binding (Robertson et al., 2007) or initiation of DNA replication (Cadoret et al., 2008).

Regardless of the class of genomic element that is annotated, either genic or non-genic, biological function is far more difficult to assess than biochemical activity (see Chapter 4.2 [Robinson-Rechavi 2020]). Indeed, numerous genomic elements are biochemically active but functionally dispensable (Graur et al., 2015). For example, transcriptional activity is often observed for pseudogenes, long after the loss of biological functions (Nakamura et al., 2009). A direct test for functionality is to examine the phenotypes and fitness of individuals in which specific elements are inactivated through genetic manipulations. Until recently, genetic manipulation techniques could only be applied to a few targeted genomic loci at a time and were exclusively used with laboratory-grown model organisms or to cell cultures (Hérault et al., 1998; Hockemeyer et al., 2011; Barde et al., 2011). With the development of CRISPR/Cas-based gene editing techniques (Jinek et al., 2012), these approaches have become more broadly applicable, leading to functional surveys encompassing thousands of loci at a time (Shalem et al., 2014; Sanjana et al., 2016). However, at the moment these techniques are still costly, time-consuming and largely restricted to phenotypes that can be observed in the laboratory. In this context, evolutionary studies can bring important insights into the functionality of diverse genomic elements. While precise biological functions generally cannot be predicted through evolutionary analyses, they are useful tools for predicting genome

functionality, by revealing elements that have been under purifying selection to preserve existing biological roles, or under positive selection following the acquisition of new functions.

Here, I review the contributions of large-scale evolutionary genomic (or phylogenomic) approaches to genome annotation. Focusing on eukaryotes, I will present several aspects of genome annotation, such as delineating the protein-coding and non-coding gene repertoires, describing gene expression regulatory elements and identifying other functional genomic elements or structures. I will present the molecular biology and genetic techniques that are nowadays frequently employed to generate data for genome annotation, as well as the evolutionary approaches that can be used to bring insights into the functionality of various genetic elements. I will thus endeavor to show how these methods can be combined to provide a comprehensive image of functional genomic landscapes.

## 2 Annotating protein-coding and non-coding gene repertoires

Undoubtedly the most important step of the genome annotation process is to characterize gene repertoires. This is a complex procedure, which can be roughly divided into four steps: describing gene models, predicting broad functional categories of genes (e.g., protein-coding and non-coding genes), inferring gene functionality and annotating putative gene functions. Here, I will discuss how phylogenomic approaches can contribute to these four gene annotation steps.

### 2.1 Gene model description

Describing gene models in eukaryotes is a challenging task, which involves identifying transcribed regions, transcription start and end sites, exon-intron structures and alternative splicing variants. Gene model prediction can be performed either *ab initio*, using species-specific data and predictive methods, or through homology-based approaches, which use gene and protein information from closely-related species to predict genes in the species of interest. *Ab initio* methods are evidently required for organisms where genome sequences for closely-related species are lacking. Conversely, homology-based predictions are beneficial when data from closely-related species is abundant, and were notably used to annotate primate genomes (Chimpanzee Sequencing and Analysis Consortium, 2005; Rhesus Macaque Genome Sequencing and Analysis Consortium et al., 2007).

For *ab initio* gene model prediction, transcriptome sequencing has become an invaluable tool. In its many forms, transcriptome sequencing has long benefited genome annotation efforts, even before next-generation sequencing techniques became available. For example, analyses of expressed sequence tags (ESTs) helped compile the initial catalogue of human genes (Lander et al., 2001), and Cap Analysis of Gene Expression (CAGE) sequencing data were used to annotate mouse gene promoters (Carninci et al., 2005). More recently, massively parallel transcriptome sequencing methods (commonly termed RNA-seq), have become an indispensable aspect of the genome annotation process. Compared to previous transcriptomics assays, RNA-seq offers increased sequencing depth and thus higher transcript detection sensitivity, even for moderately expressed genes (Wang et al., 2009). To improve detection sensitivity even at low expression levels, RNA-seq can be used in combination with RT-PCR amplification (Howald et al., 2012) or with capture on tiling arrays (Clark et al., 2015; Bussotti et al., 2016), which considerably increases the sequencing depth for targeted transcripts or genomic regions. Several computational methods were developed to assemble transcript sequences from RNA-seq data, either using a genome sequence as a reference (Trapnell et al., 2010; Pertea et al., 2015) or entirely *de novo* (Grabherr et al., 2011). The

#### 4.1:4 Phylogenomics and Genome Annotation

application of transcriptome sequencing to genome annotation has revealed many forms of transcriptome complexity. These include the presence of numerous transcript variants for protein-coding genes, generated through canonical mechanisms such as alternative splicing, use of alternative transcription initiation or termination sites, read-through transcription or trans-splicing (ENCODE Project Consortium et al., 2007; Gerstein et al., 2007). These in-depth genome annotation studies also established that transcription is pervasive outside of protein-coding genes (ENCODE Project Consortium et al., 2007). In particular, in-depth transcriptome and chromatin accessibility surveys revealed that mammalian genomes contain tens of thousands of long non-coding RNAs (Guttman et al., 2009; Khalil et al., 2009; Iyer et al., 2015; Pertea et al., 2018b).

Homology-based gene model prediction approaches are of particular importance for non-model species, when other sources of data are insufficient. The quality of a genome annotation largely depends on the quality and quantity of transcriptomic and proteomic data available for that species (Mudge and Harrow, 2016). For widely-studied species such as human, mouse, fruitfly or nematode, extensive resources (including full-length or partial cDNA sequences, RNA-seq and proteomics data) have accumulated over time and are available as input for genome annotation (Mudge and Harrow, 2016). However, this is an exception rather than the rule, and for many species experimental data are scarce. In this case, homology-based annotation methods can be applied, with relative facility. The most frequently used gene model prediction software, including Augustus (Stanke et al., 2006), Gnomon (Suvorov et al., 2010), Exonerate (Slater and Birney, 2005) and GeneWise (Birney et al., 2004), can use as input protein and RNA sequences from closely related species. In the simplest implementations, the genome is scanned to identify local alignments between protein-sequences and nucleotide sequence translations. This is for example done in the Ensembl annotation pipeline (Zerbino et al., 2018), in which pairwise alignments between reference protein sequences and translated nucleotide sequences are generated and exploited to predict gene structures, with Exonerate (Slater and Birney, 2005) and GeneWise (Birney et al., 2004). The efficiency of this approach depends on the degree of sequence conservation between the proteins used as reference and the ones encoded in the target genome. To identify more divergent proteins, an extension of Augustus (Keller et al., 2011) uses multiple sequence alignments to construct protein conservation profiles and to identify blocks of ungapped, highly-conserved sequences. Predicted gene structures in the target genome are then compared with the resulting sequence conservation profiles, and are assigned higher confidence scores if they match the amino acid composition profiles of conserved alignment blocks.

Homology-based prediction methods can also be insightful for annotating non-coding RNA genes. For lncRNAs, which are generally weakly expressed, defining gene models with standard RNA-seq data is often not sufficient, as the low read coverage can result in gene model fragmentation (Howald et al., 2012). In these cases, for comparative analyses of lncRNAs across closely-related species, it can be beneficial to project annotations from one species to another, based on primary sequence similarity (Washietl et al., 2014; Necsulea et al., 2014). This method has obvious disadvantages, as it cannot correctly analyze homologous lncRNA loci that have diverged in terms of exon/intron structures, nor can it predict loci where transcription is species-specific (Hezroni et al., 2015). Homology-based annotation approaches, for protein-coding genes, non-coding RNAs or other types of functional genomic elements, all share these limitations, and it is important to complement these methods with species-specific “omics” data. Nevertheless, they provide a valuable starting point on which more comprehensive genome annotation resources can be built.

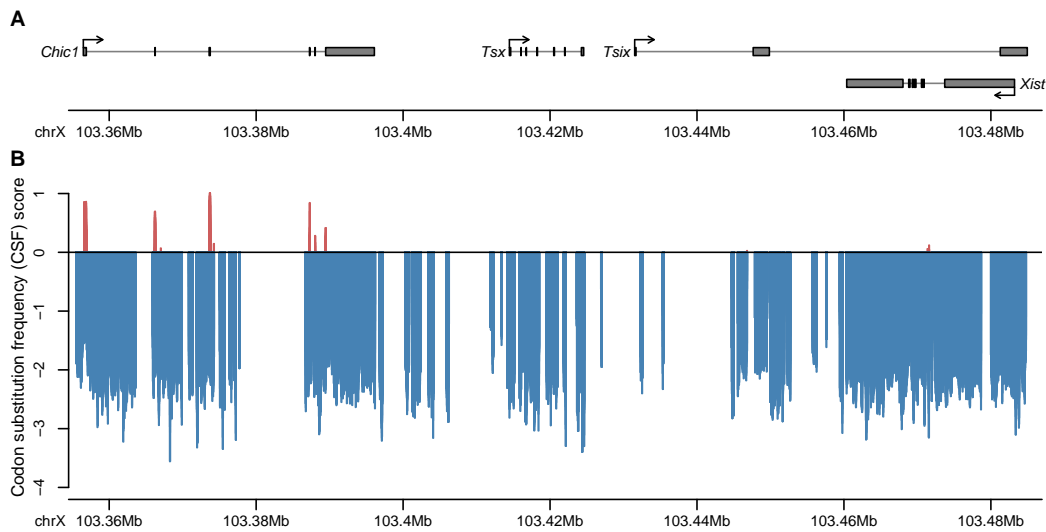
## 2.2 Gene classification

A second important step in the genome annotation process, after gene model description, is to provide a broad classification of the resulting loci into protein-coding and non-coding genes. This step is more difficult than it can seem at first sight, mainly because lncRNAs are structurally very similar to protein-coding mRNAs (Derrien et al., 2012).

To categorize genes as protein-coding or non-coding, direct proteome assays are an evident path. However, proteomics technologies, although in continuous progress (Richards et al., 2015), are still far from the throughput observed for RNA-seq. Large-scale investigations of the proteome based on mass spectrometry have only recently become available for humans (Kim et al., 2014; Wilhelm et al., 2014), and are still lacking for most other species. Recent studies were able to detect and quantify peptides for approximately 84% of annotated protein-coding genes, but generally lacked power to detect known alternative protein isoforms (Kim et al., 2014; Wilhelm et al., 2014). In the absence of high-throughput proteome sequencing, an alternative avenue towards large-scale investigations of the proteome (or at least of the translome) is provided by the development of ribosome profiling (Ingolia et al., 2009). This technique isolates and sequences RNA molecules that are bound by poly-ribosome complexes, which are thus likely actively translated (Ingolia et al., 2009). While more accessible than mass spectrometry, this technique is nevertheless considerably more complex than classical RNA-seq, and very little data has been generated so far. Thus, experimental data that could help distinguish between protein-coding and non-coding RNA genes are not readily available. Instead, computational methods, many of which are based on the patterns of sequence evolution, have been developed to determine the protein-coding potential of newly-annotated transcripts.

It is interesting to note that the first long non-coding RNA ever identified in mammals, namely the H19 lncRNA, was defined as such using an evolutionary approach (Brannan et al., 1990). Sequence analyses of the mouse transcript revealed the presence of several small open reading frames (ORFs). However, comparisons with the human homolog showed that none of these open reading frames were conserved during evolution, indicating that the locus did not encode a functional protein (Brannan et al., 1990). Indeed, the mere presence of ORFs is not a reliable indicator that an eukaryotic sequence is protein-coding, given that such stretches can appear by chance in long RNA molecules (Clamp et al., 2007). In contrast, their conservation during evolution, through negative selection that prevents the fixation of ORF-disrupting mutations, is a strong predictor of the presence of a constrained protein-coding sequence. The idea of exploiting the patterns of sequence evolution to predict the protein-coding potential of genomic sequences was later implemented into two computational methods that aimed to detect *bona fide* protein-coding genes in yeast and fruitfly genomes: the reading frame conservation (RFC) method (Kellis et al., 2004) and the codon substitution frequency (CSF) method (Lin et al., 2007). The RFC method assesses the presence of ORF-disrupting insertions and deletions in a multiple sequence alignment between the target species and other “informant” species (Kellis et al., 2004). The CSF method (Figure 1) analyzes the proportion of synonymous and non-synonymous single-nucleotide substitutions between the target and informant species, in all possible reading frames (Lin et al., 2007). Given that it relies on the presence of insertions and deletions, which are less frequent than point mutations, the RFC approach strongly depends on the degree of sequence conservation between the target and informant species (Lin et al., 2008). In contrast, the CSF method has high sensitivity and specificity values, although it may propose wrong classifications for protein-coding sequences that are subject to positive selection (Lin et al., 2008, 2011). This approach was used to distinguish protein-coding and non-coding regions in the first

## 4.1:6 Phylogenomics and Genome Annotation



**Figure 1** The codon substitution frequency (CSF) score exploits the pattern of nucleotide substitutions in a multiple species alignment to predict protein-coding regions. A) Genomic localization and exon-intron structure for mouse *Chic1*, *Tsx*, *Tsix* and *Xist* genes. The rectangles represent the exons and the arrows represent the direction of transcription. B) The codon substitution frequency (CSF) score variation in the same genomic region. Positive CSF scores, which indicate the presence of protein-coding regions under purifying selection to preserve protein sequences, mainly co-localize with *Chic1* annotated protein-coding exons. Another annotated protein-coding gene, *Tsx*, does not show any positive scores. Negative CSF scores are observed elsewhere, including on the exons of long non-coding RNAs *Xist* and *Tsix*. Whole-genome CSF data were taken from a previous publication (Necsulea et al., 2014); recently, whole-genome PhyloCSF data have become available (Mudge et al., 2019).

large-scale investigations of lncRNAs (Guttman et al., 2009; Khalil et al., 2009) and later in the first large-scale evolutionary analyses of lncRNA across vertebrates (Necsulea et al., 2014). Although its efficiency is higher for longer sequences, if sufficient “informants” are included in the analysis (including both distant and closely related species with respect to the species of interest), the CSF method can also detect short protein-coding regions. This approach can thus be applied to scan protein-coding regions in the whole genome, with a sliding window approach (Figure 1, Mudge et al., 2019).

Expectedly, gene classifications as protein-coding or non-coding obtained with biochemical or evolutionary approaches do not always agree. Notably, ribosome profiling studies revealed that numerous lncRNAs annotated with evolutionary approaches are in fact actively translated (Ingolia et al., 2014), and mass spectrometry assays were able to detect peptide sequences stemming from hundreds of lncRNAs (Kim et al., 2014). While some of this inconsistency could simply be attributed to imperfect sensitivity and specificity of the classification methods, the presence of ribosome footprints on lncRNA sequences is not itself evidence that these transcripts are translated into functional proteins. In fact, the ribosome occupancy profile is strikingly different between genuine protein-coding mRNAs and lncRNAs: while a sharp ribosome release at the stop codon and a strong reading frame preference is observed for the former, the profiles are much more uniform along lncRNA sequences, indicating that these transcripts are simply scanned by ribosomes, but likely do not generate functional proteins (Guttman et al., 2013).

Another intriguing cause of disagreement between the phylogenomic and biochemical

classification methods is the evolutionary history of the genes (see Chapter 4.2 [Robinson-Rechavi 2020]). Indeed, the RFC and CSF methods both rely on the pattern of sequence evolution, which can be assessed within varying evolutionary time frames, depending on the phylogenetic relatedness of the analyzed species. However, the functional category of the gene may itself evolve over time. For example, protein-coding genes may become pseudogenized, and potentially resurrected into functional lncRNAs, as is famously the case for *Xist* (Duret et al., 2006), as well as for other conserved lncRNAs (Hezroni et al., 2017). Conversely, lncRNAs may transform into protein-coding genes by acquiring functional ORFs (McLysaght and Hurst, 2016). This evolutionary plasticity highlights the importance of combining phylogenomic and biochemical approaches to determine the protein-coding potential of newly annotated transcripts, which may reveal insights into the evolutionary processes that lead to new gene origination (McLysaght and Hurst, 2016).

### 2.3 Gene functionality

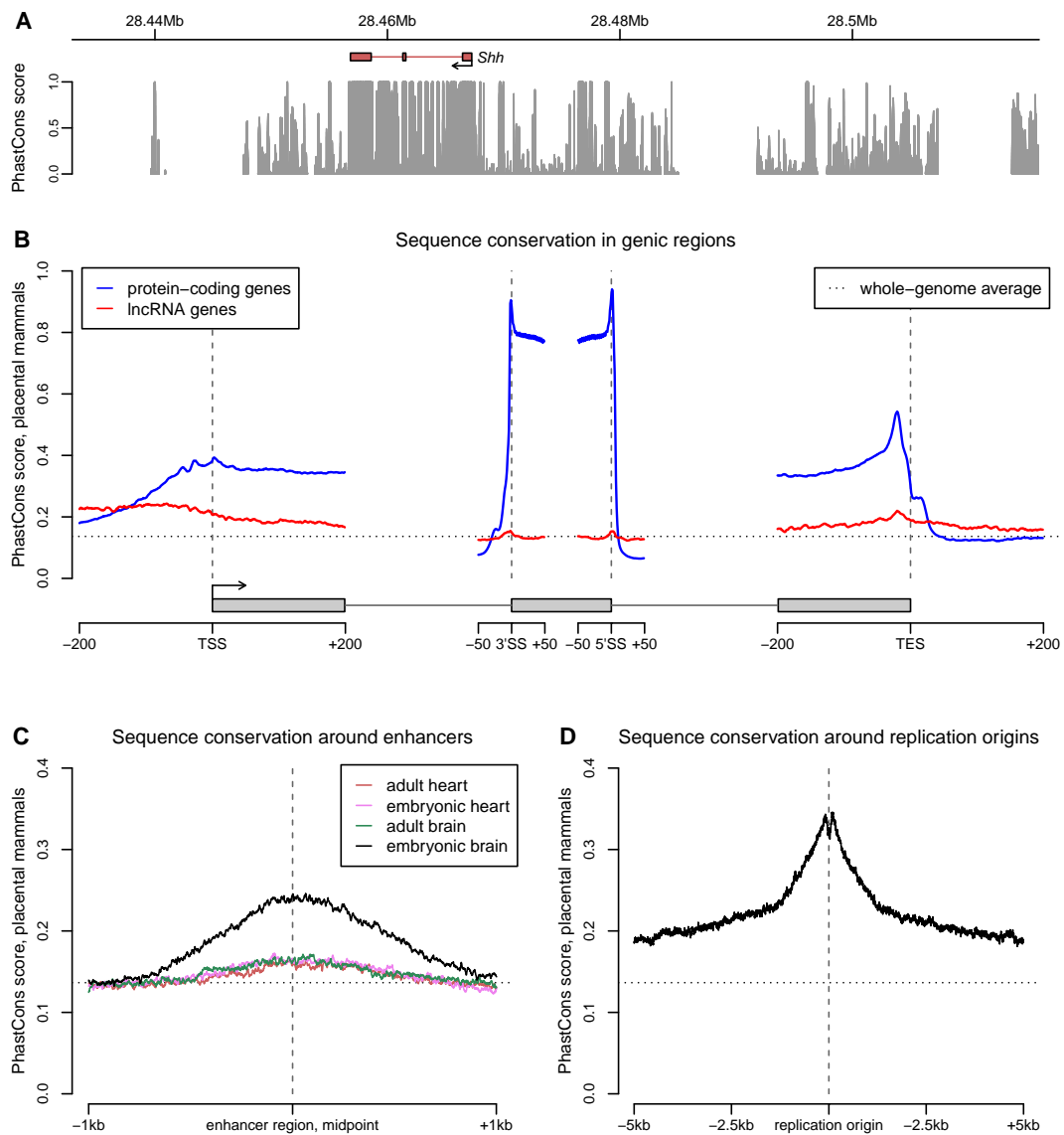
The staggering complexity of the human transcriptome (Pertea et al., 2018b; Iyer et al., 2015; Carninci et al., 2005) raises the question of its functionality. Many of the transcripts discovered with high-throughput transcriptome sequencing data, whether alternative isoforms of protein-coding genes, read-through transcripts that join neighboring genes and in particular long non-coding RNAs, may in fact be functionally dispensible, representing so-called “transcriptional noise” (Ponjavic et al., 2007). Experimental methods that directly address functionality typically rely on genetic manipulations that inactivate or over-express specific transcripts, followed by phenotypic evaluations. Although these methods have recently become more accessible, applicable to large numbers of loci (Shalem et al., 2014; Joung et al., 2017) and to a wider range of organisms (Mazo-Vargas et al., 2017), they are still costly, time-consuming and largely restricted to phenotypes that can be observed in the laboratory. In this context, phylogenomic approaches are extremely valuable, as they can provide solid predictions of biological functionality (Haerty and Ponting, 2014).

The ongoing search for lncRNA functionality is a good illustration of the usefulness of phylogenomic methods in this context. Indeed, in the absence of large-scale experimental data for this category of genes, the functionality of lncRNAs has often been investigated with evolutionary approaches. One such study compared the rates and patterns of sequence evolution between mammalian long non-coding RNAs and ancient transposable element insertions, which are likely neutrally-evolving (Ponjavic et al., 2007). This study revealed slightly, but significantly lower rates of evolution for lncRNAs than for ancient repeats, indicating the presence of purifying selection for at least a subset of lncRNAs (Ponjavic et al., 2007). These conclusions were confirmed by subsequent studies, which consistently showed that mammalian lncRNAs are more conserved than expected by chance, but that they display modest levels of primary sequence conservation compared to protein-coding genes (Guttman et al., 2009; Washietl et al., 2014; Necsulea et al., 2014; Marques and Ponting, 2009; Kutter et al., 2012; Haerty and Ponting, 2013; Wiberg et al., 2015). These studies assessed either long-term selective constraints, for example by analyzing PhastCons scores determined from whole-genome alignments of placental mammals or vertebrates (Figure 2), or short-term sequence evolution, contrasting single-nucleotide polymorphisms within populations and sequence divergence between closely related species (Haerty and Ponting, 2013; Wiberg et al., 2015).

In contrast, in fruitfly, lncRNAs are under strong purifying selection (Haerty and Ponting, 2013; Young et al., 2012). These observations are in agreement with the “transcriptional noise” hypothesis, and the differences between mammals and fruitfly likely reflect the reduced



## 4.1:8 Phylogenomics and Genome Annotation



**Figure 2** Sequence conservation patterns around functional elements in the mouse genome. A) Sequence conservation (PhastCons score (Siepel et al., 2005), computed on a whole-genome alignment of mouse and 59 other vertebrate species) variation around the *Shh* gene, in the mouse genome. The amount of sequence conservation reaches maximum values in *Shh* exons, but also in neighboring intergenic regions, potentially including regulatory elements. B) Average sequence conservation profile in protein-coding and lncRNA gene structures: transcription start sites, splice sites and transcription end sites. C) Average sequence conservation profiles around mouse transcriptional enhancers (Yue et al., 2014) from different tissues. D) Average sequence conservation profiles around mouse replication origins (Cayrou et al., 2015). B-D) The average sequence conservation profiles were based on the PhastCons score, computed on a whole-genome alignment of mouse and 39 other placental mammal species (Siepel et al., 2005). PhastCons scores were downloaded from the UCSC Genome Browser (Casper et al., 2018).

efficiency of natural selection in the former, due to low effective population sizes (Haerty and Ponting, 2013).

However, alternative hypotheses were proposed to explain the low levels of sequence constraint observed for mammalian lncRNAs without dismissing their potential functionality. A plausible hypothesis posits that lncRNA functions may be achieved by short sequence motifs, which may for example mediate their binding to genomic regions or protein sequences (Hezroni et al., 2015). This would explain why levels of evolutionary conservation, when computed on the entire length of lncRNAs, are only slightly above neutral expectations (Ponjavic et al., 2007). Interestingly, analyses of human lncRNAs revealed that almost all sequence constraint is indeed concentrated in very short sequence motifs, but that these small constrained regions are in fact splicing regulatory elements (Figure 2; Schüler et al., 2014; Haerty and Ponting, 2015). Purifying selection on sequences needed to achieve correct splicing of multi-exonic lncRNA loci could indeed be indicative of transcript functionality. However, a recent experimental investigation showed that splicing of lncRNA loci can influence the expression of neighboring genes (Engreitz et al., 2016). Thus, the presence of selection on lncRNA splicing motifs does not necessarily prove that lncRNA transcripts are themselves biologically functional.

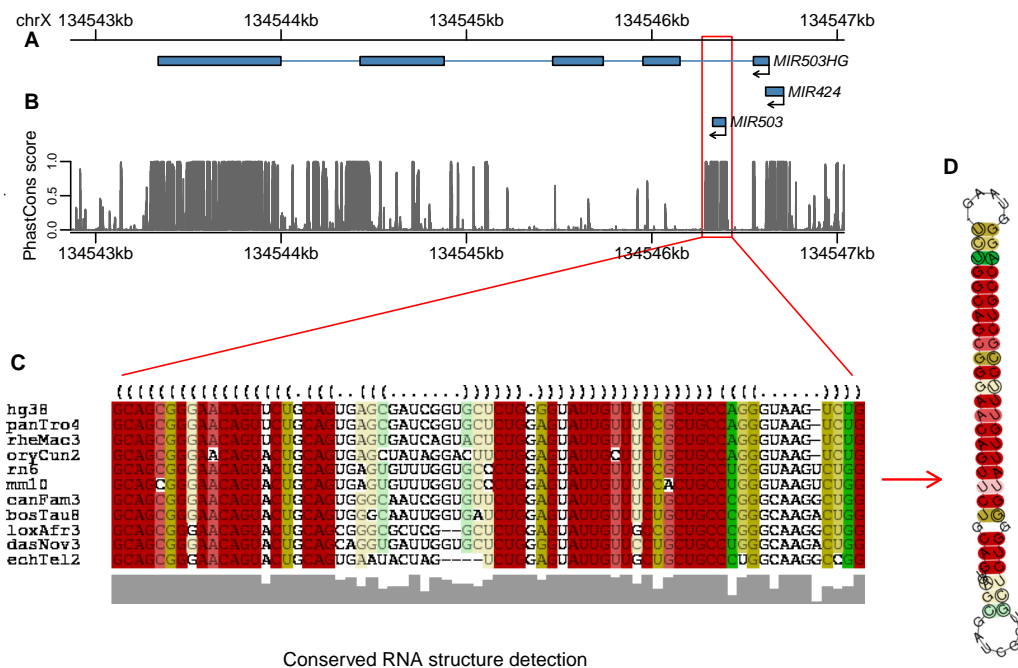
Another hypothesis that could explain the weak levels of lncRNA conservation is that selective pressures may act on secondary RNA structures, rather than on primary transcript sequences (Kapusta and Feschotte, 2014). This hypothesis can be directly tested, for example by contrasting the degree of RNA secondary structure conservation with the degree of primary sequence conservation, using RNA structures predicted with thermodynamic modeling and multiple sequence alignments (Washietl et al., 2005). Using this principle, genome-wide scans for conserved RNA secondary structures consistently confirmed selective pressures on miRNA, tRNA and rRNA structures (Figure 3), but revealed only limited such constraint within long non-coding RNA loci (Pedersen et al., 2006; Parker et al., 2011; Seemann et al., 2017).

Overall, there is increasing evidence that lncRNA functionality often does not reside in the RNA molecule encoded by the locus, but in the presence of additional regulatory elements that affect neighboring gene expression patterns (Latos et al., 2012; Engreitz et al., 2016; Amândio et al., 2016). Experimental studies of lncRNA functions must be carefully designed to address these strong confounding effects (Bassett et al., 2014). Likewise, phylogenomic studies of lncRNA functionality need to be adapted to account for additional targets of selective pressures (Haerty and Ponting, 2014).

## 2.4 Gene function

Even when gene models (*i.e.*, gene localization, exon-intron structure and alternative isoforms) can be predicted based on species-specific experimental data, gene functions are still overwhelmingly inferred based on homology. Indeed, experimental investigations of protein or RNA functions are lagging well behind the vast amounts of transcripts and proteins predicted from next-generation sequencing data. Functional annotations are thus commonly transferred across species based on homology relationships, with the underlying assumption that gene functions are generally conserved during evolution (see Chapter 4.2 [Robinson-Rechavi 2020]). As for homology-based gene model predictions, the efficacy and reliability of the transfer of functional annotations across species is dependent on the degree of sequence divergence between the reference sequences and the target genome to be annotated. Computational methods that can predict homologous gene families in the presence of high degrees of sequence divergence are thus of great interest (Vilella et al., 2009). Another important challenge is to correctly identify gene duplication events, and to predict the functional characteristics of the resulting gene copies. Indeed, gene duplication is believed to be an important driver of

#### 4.1:10 Phylogenomics and Genome Annotation



**Figure 3** Identification of conserved RNA structures using the pattern of sequence evolution (Seemann et al., 2017). A) Genomic position and exon-intron structure for lncRNA gene *MIR503-HG* and miRNA genes *MIR503* and *MIR424*, in the human genome. The rectangles represent the exons and the arrows represent the direction of transcription. B) Sequence conservation profile (PhastCons score [Siepel et al. 2005], computed on a whole-genome alignment of human and 99 other vertebrate genomes), on the same genomic region. PhastCons scores were provided by the UCSC Genome Browser (Casper et al., 2018). C) Sequence alignment and predicted consensus RNA structure in the *MIR503* region. D) Resulting conserved RNA structure for *MIR503*.

functional innovation, as the initially redundant gene copies can accumulate mutations that lead to sub-functionalization or to neo-functionalization (Conant and Wolfe, 2008). For both homologous and paralogous genes, the likelihood of functional conservation decreases with increasing divergence time (Studer and Robinson-Rechavi, 2009). The relationship between the extent of sequence (or structure) divergence and functional divergence cannot be readily defined, and it likely varies among functional categories of genes (Tian and Skolnick, 2003). Thus, cross-species projections of gene functions need to be interpreted with great caution.

Homology-based gene model annotation and functional assignment methods have been applied to both protein-coding and non-coding genes. However, these approaches are significantly more successful for the former than for the latter, as non-coding RNA sequences are generally much less conserved than protein sequences. Among non-coding RNA classes, lncRNAs in particular evolve very rapidly (Figure 2; Washietl et al., 2014; Necsula et al., 2014). This is well illustrated by the fact that lncRNA annotation efforts based on gene model projections across species could identify only approximately 2,000 lncRNAs conserved in placental mammals (Washietl et al., 2014; Necsula et al., 2014). These studies predicted conserved lncRNAs based on primary sequence conservation and required species-specific transcription evidence to confirm the activity of the lncRNA loci in other species (Washietl et al., 2014; Necsula et al., 2014). Here again, additional methodological developments are needed to exploit the specific patterns of lncRNA evolution, such as the presence of

short stretches of conserved regions within larger, overall divergent sequences (Hezroni et al., 2015). Transfer of functional annotations across species is particularly problematic for long non-coding RNAs, for which experimental data on biological functions are scarce even in model organisms. In this context, comparative transcriptomics analysis across species can provide crude functional assignments, for example by identifying evolutionarily conserved co-expression relationships between lncRNAs and protein-coding genes, which may indicate functional associations (Stuart et al., 2003; Necsulea et al., 2014).

### 3 Annotating non-genic functional elements with phylogenomic approaches

Eukaryotic genomes harbor numerous functional non-genic elements. These include non-coding sequences that regulate gene expression, such as transcriptional enhancers (Banerji et al., 1981) or silencers (Busturia et al., 1997), splicing regulatory elements (Lee and Rio, 2015), but also origins of DNA replication (Benbow et al., 1992), insulators that organize chromatin architecture in the nucleus (Van Bortle and Corces, 2012), recombination hotspots (Smith, 1994), etc. . . Some categories of non-coding functional elements can be now be identified with dedicated experimental assays, such as chromatin immunoprecipitation and sequencing (ChIP-seq) techniques that identify genomic sequences bound by specific proteins or by modified histones (Robertson et al., 2007; Visel et al., 2009), or nascent DNA strand sequencing to pinpoint origins of replication (Cadoret et al., 2008; Cayrou et al., 2015). However, by construction these techniques use the presence of biochemical activity to predict biological function, although the two concepts are far from being synonymous (Graur et al., 2013). Indeed, numerous biochemically active genomic elements are altogether dispensable from a biological point of view, either because most cellular mechanisms (including transcription, protein-DNA binding, etc. . .) are error-prone, or because of functional redundancy with other genomic elements (Graur et al., 2013). Additional data are thus needed to ascertain biological functionality, and phylogenomic approaches are again a valuable asset in this context.

Perhaps the most striking example of how phylogenomic approaches can be used to annotate functional non-coding elements is the discovery of ultra-conserved sequences (Duret et al., 1993; Bejerano et al., 2004). These elements were first identified through comparative analyses of nucleotide sequences across distant vertebrate species, which revealed the presence of regions with unexpectedly high degrees of conservation (more than 70% sequence similarity for species that diverged at least 300 million years ago, Duret et al., 1993). This pioneering study, which predates the genomic era, was later confirmed through genome-wide scans, which identified thousands of ultra-conserved elements outside of protein-coding genes in vertebrates and in other metazoan genomes (Bejerano et al., 2004; Siepel et al., 2005). Importantly, the low rate of sequence evolution in these regions is not due to overlap with mutational cold-spots. On the contrary, analyses of within-species polymorphism and between-species divergence rates showed that these elements are subject to intense purifying selective pressures (Katzman et al., 2007), which further underscores their functional relevance. *In vivo* experimental assays showed that a great proportion of ultraconserved elements have transcriptional enhancer capacity in the mouse embryo (Pennacchio et al., 2006), thus confirming the regulatory roles proposed upon their initial discovery (Duret et al., 1993). Some of these elements may also belong to non-coding RNA loci (Kern et al., 2015).

It is important to stress that phylogenomic approaches that focus on signatures of strong evolutionary conservation cannot discover all types of functional non-coding elements.

#### 4.1:12 Phylogenomics and Genome Annotation

For example, the extreme levels of sequence conservation observed for some embryonic transcriptional enhancers are not observed in all tissues and developmental stages: heart enhancers show much weaker levels of sequence conservation than brain enhancers (Blow et al., 2010), and enhancers active in adult brain are much less conserved than those active in embryonic brain (Figure 2). Other functional genomic elements, such as origins of replication, also display increased sequence conservation compared to the genomic background (Figure 2, Cadoret et al., 2008). However, much of the sequence conservation observed within experimentally predicted origins of DNA replication in the human genome stems from their overlap with transcriptional promoters (Cadoret et al., 2008).

In addition to overlooking genomic elements that are under weak purifying selection, which are difficult to distinguish from the neutrally evolving genomic background, phylogenomic scans may also bypass functional elements that evolve rapidly due to positive selection. Dedicated computational methods were developed to identify genomic regions that evolve faster than expected under a neutral regime (Pollard et al., 2010). However, an accelerated rate of sequence evolution, which is the main signal used to predict the footprints of adaptation in non-coding regions, is by no means synonymous with positive selection. Biased gene conversion, a non-adaptive mechanism that promotes the fixation of specific alleles in highly recombining regions, frequently leads to accelerated sequence evolution, thereby confounding positive selection scans (Duret and Galtier, 2009; Ratnakumar et al., 2010).

Phylogenomic approaches that aim to predict functional non-genic elements will likely further be improved by the increasing numbers of complete genome sequences, including population genomics datasets that enable investigations of DNA sequence variations within and between populations (1000 Genomes Project Consortium et al., 2015), in addition to between-species sequence divergence. Moreover, important efforts have been made to generate combined genome and transcriptome population datasets, such as Geuvadis (Lappalainen et al., 2013) or GTEx (GTEx Consortium, 2015). Joint analyses of genome and transcriptome variations within populations have already been used to predict putative regulatory variants, that is, polymorphisms that are statistically associated with expression level variations between individuals (Lappalainen et al., 2013; GTEx Consortium, 2015). Combined with between-species genome and transcriptome comparative analyses, these approaches could bring insights into the selective pressures that act on gene expression levels (Gilad et al., 2006; Romero et al., 2012), and thereby help annotate non-coding RNA transcripts whose expression patterns are constrained, rather than their RNA sequences (Latos et al., 2012).

### **4 Combining molecular biology, genetics and evolutionary biology to annotate functional genomic elements**

We have never been this close to truly uncovering the functional landscapes of the genomes. In the past decade, technological innovations have enabled us not only to investigate biochemical activities (such as transcription, translation or transcription factor binding) at a genome-wide level, but also to perform large-scale experimental assessments of biological functions through genetic manipulations (Jinek et al., 2012; Sanjana et al., 2016; Joung et al., 2017). The contributions of molecular biology and genetics methodologies to functional genome annotation are thus indisputable. However, even in this technology-dominated context, phylogenomic approaches are still an invaluable tool for the discovery and annotation of functional genomic elements.

Phylogenomic methods, such as genome-wide scans for regions under purifying or positive selection, can be used in combination with molecular biology assays and genetic manipulations

to obtain thorough functional characterizations for specific genomic elements. First of all, very often, genetic manipulation studies use the presence of evolutionary sequence conservation to prioritize elements for further experiments (Sauvageau et al., 2013). Moreover, evolutionary analyses can also provide information into the facet of a locus that is most likely the target of natural selection, and which should thus be perturbed through genetic manipulations to test for biological function. For example, for long non-coding RNAs the highest degrees of sequence conservation were observed on promoter regions and splicing regulatory elements (Figure 2, Guttman et al., 2009; Ponjavic et al., 2007; Schüler et al., 2014; Haerty and Ponting, 2015). Genetic manipulations later showed that the presence of transcription and splicing at multiple lncRNA loci affected neighboring gene expression, while the production of a specific RNA sequence was dispensable (Engreitz et al., 2016). Thus, the functional elements in lncRNA loci could be correctly predicted with an evolutionary approach.

While most phylogenomic studies can bring insights into the functionality of a given locus (that is, on its effect on the overall fitness of the organism), rather than on its specific biological functions, in some cases evolutionary studies can go even beyond and predict the mode of action or the phenotype in which a genomic element is involved. For example, genome-wide scans for evolutionarily conserved RNA secondary structures have uncovered thousands of genomic regions that are transcribed into structured non-coding RNAs, such as miRNAs, tRNAs or rRNAs (Pedersen et al., 2006; Parker et al., 2011; Seemann et al., 2017). Interestingly, while most phylogenomic scans for functional elements rely on the presence of evolutionary conservation, evolutionary losses of genes and other genomic elements can also bring insights into genomic functions. An elegant evolutionary approach aiming to discover genes and regulatory elements that are involved in specific phenotypes is the recently proposed “forward genomics” method, which analyzes phylogenies in which the same specific trait (e.g. the ability to synthesize vitamin C) was lost multiple times independently (Hiller et al., 2012). Genomic regions that were needed only to achieve the specific function under study are likely to accumulate substitutions in the lineages that have lost it, due to relaxation of purifying selection pressures. This approach can successfully predict genes and non-genic functional elements that are specifically associated with a given trait, if sufficient independent trait losses can be analyzed (Hiller et al., 2012). Although this methodology clearly has limitations, not least of which is the pervasive presence of pleiotropy in vertebrate genomes, it is an exciting use of phylogenomics for functional genome annotation, which bridges the gap between genome and phenotypes.

So far, phylogenomic methods have been successfully used to predict gene localization and structure, expression regulatory elements, conserved RNA secondary structures, as well as to distinguish between coding and non-coding transcribed regions. As molecular biology and genetic technologies continue to progress, bringing us closer to understanding genomic functions, the field of evolutionary genomics must also continue to develop and to propose new methods to assess selective pressures that act on newly discovered classes of functional elements. We can thus hope to make sense of the intricate functional architecture of our genomes, in the light of evolution (Haerty and Ponting, 2014).

## References

- 1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., and Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, 526(7571):68–74.

#### 4.1:14 REFERENCES

- Abe, T., Inokuchi, H., Yamada, Y., Muto, A., Iwasaki, Y., and Ikemura, T. (2014). tRNADB-CE: tRNA gene database well-timed in the era of big sequence data. *Front Genet*, 5:114.
- Amândio, A. R., Necsulea, A., Joye, E., Mascrez, B., and Duboule, D. (2016). Hotair is dispensible for mouse development. *PLoS Genet.*, 12(12):e1006232.
- Banerji, J., Rusconi, S., and Schaffner, W. (1981). Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, 27(2 Pt 1):299–308.
- Barde, I., Verp, S., Offner, S., and Trono, D. (2011). Lentiviral Vector Mediated Transgenesis. *Curr Protoc Mouse Biol*, 1(1):169–184.
- Bassett, A. R., Akhtar, A., Barlow, D. P., Bird, A. P., Brockdorff, N., Duboule, D., Ephrussi, A., Ferguson-Smith, A. C., Gingeras, T. R., Haerty, W., Higgs, D. R., Miska, E. A., and Ponting, C. P. (2014). Considerations when investigating lncRNA function in vivo. *Elife*, 3:e03058.
- Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S., and Haussler, D. (2004). Ultraconserved elements in the human genome. *Science*, 304(5675):1321–1325.
- Benbow, R. M., Zhao, J., and Larson, D. D. (1992). On the nature of origins of DNA replication in eukaryotes. *Bioessays*, 14(10):661–670.
- Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and Genomewise. *Genome Res*, 14(5):988–995.
- Blow, M. J., McCulley, D. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., Afzal, V., Bristow, J., Ren, B., Black, B. L., Rubin, E. M., Visel, A., and Pennacchio, L. A. (2010). ChIP-Seq identification of weakly conserved heart enhancers. *Nat. Genet.*, 42(9):806–810.
- Brannan, C. I., Dees, E. C., Ingram, R. S., and Tilghman, S. M. (1990). The product of the H19 gene may function as an RNA. *Mol. Cell. Biol.*, 10(1):28–36.
- Bussotti, G., Leonardi, T., Clark, M. B., Mercer, T. R., Crawford, J., Malquori, L., Notredame, C., Dinger, M. E., Mattick, J. S., and Enright, A. J. (2016). Improved definition of the mouse transcriptome via targeted RNA sequencing. *Genome Res.*, 26(5):705–716.
- Busturia, A., Wightman, C. D., and Sakonju, S. (1997). A silencer is required for maintenance of transcriptional repression throughout Drosophila development. *Development*, 124(21):4343–4350.
- Cadoret, J.-C., Meisch, F., Hassan-Zadeh, V., Luyten, I., Guillet, C., Duret, L., Quesneville, H., and Prioleau, M.-N. (2008). Genome-wide studies highlight indirect links between human replication origins and gene regulation. *Proc. Natl. Acad. Sci. U.S.A.*, 105(41):15837–15842.
- Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., Kodzius, R., Shimokawa, K., Bajic, V. B., Brenner, S. E., Batalov, S., Forrest, A. R. R., Zavolan, M., Davis, M. J., Wilming, L. G., Aidinis, V., Allen, J. E., Ambesi-Impiombato, A., Apweiler, R., Aturaliya, R. N., Bailey, T. L., Bansal, M., Baxter, L., Beisel, K. W., Bersano, T., Bono, H., Chalk, A. M., Chiu, K. P., Choudhary, V., Christoffels, A., Clutterbuck, D. R., Crowe, M. L., Dalla, E., Dalrymple, B. P., de Bono, B., Della Gatta, G., di Bernardo, D., Down, T., Engstrom, P., Fagiolini, M., Faulkner, G., Fletcher, C. F., Fukushima, T., Furuno, M., Futaki, S., Gariboldi, M., Georgii-Hemming, P., Gingeras, T. R., Gojobori, T., Green, R. E., Gustincich, S., Harbers, M., Hayashi, Y., Hensch, T. K., Hirokawa, N., Hill, D., Huminiecki, L., Iacono, M., Ikeo, K., Iwama, A., Ishikawa, T., Jakt, M., Kanapin, A., Katoh, M., Kawasawa, Y., Kelso, J., Kitamura, H., Kitano, H., Kollias, G., Krishnan, S. P. T., Kruger, A., Kummerfeld, S. K., Kurochkin, I. V., Lareau, L. F., Lazarevic, D., Lipovich, L., Liu, J., Liuni, S.,



- McWilliam, S., Madan Babu, M., Madera, M., Marchionni, L., Matsuda, H., Matsuzawa, S., Miki, H., Mignone, F., Miyake, S., Morris, K., Mottagui-Tabar, S., Mulder, N., Nakano, N., Nakauchi, H., Ng, P., Nilsson, R., Nishiguchi, S., Nishikawa, S., Nori, F., Ohara, O., Okazaki, Y., Orlando, V., Pang, K. C., Pavan, W. J., Pavesi, G., Pesole, G., Petrovsky, N., Piazza, S., Reed, J., Reid, J. F., Ring, B. Z., Ringwald, M., Rost, B., Ruan, Y., Salzberg, S. L., Sandelin, A., Schneider, C., Schönbach, C., Sekiguchi, K., Semple, C. a. M., Seno, S., Sessa, L., Sheng, Y., Shibata, Y., Shimada, H., Shimada, K., Silva, D., Sinclair, B., Sperling, S., Stupka, E., Sugiura, K., Sultana, R., Takenaka, Y., Taki, K., Tammoja, K., Tan, S. L., Tang, S., Taylor, M. S., Tegner, J., Teichmann, S. A., Ueda, H. R., van Nimwegen, E., Verardo, R., Wei, C. L., Yagi, K., Yamanishi, H., Zabarovskiy, E., Zhu, S., Zimmer, A., Hide, W., Bult, C., Grimmond, S. M., Teasdale, R. D., Liu, E. T., Brusica, V., Quackenbush, J., Wahlestedt, C., Mattick, J. S., Hume, D. A., Kai, C., Sasaki, D., Tomaru, Y., Fukuda, S., Kanamori-Katayama, M., Suzuki, M., Aoki, J., Arakawa, T., Iida, J., Imamura, K., Itoh, M., Kato, T., Kawaji, H., Kawagashira, N., Kawashima, T., Kojima, M., Kondo, S., Konno, H., Nakano, K., Ninomiya, N., Nishio, T., Okada, M., Plessy, C., Shibata, K., Shiraki, T., Suzuki, S., Tagami, M., Waki, K., Watahiki, A., Okamura-Oho, Y., Suzuki, H., Kawai, J., Hayashizaki, Y., FANTOM Consortium, and RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group) (2005). The transcriptional landscape of the mammalian genome. *Science*, 309(5740):1559–1563.
- Casper, J., Zweig, A. S., Villarreal, C., Tyner, C., Speir, M. L., Rosenbloom, K. R., Raney, B. J., Lee, C. M., Lee, B. T., Karolchik, D., Hinrichs, A. S., Haeussler, M., Guruvadoo, L., Navarro Gonzalez, J., Gibson, D., Fiddes, I. T., Eisenhart, C., Diekhans, M., Clawson, H., Barber, G. P., Armstrong, J., Haussler, D., Kuhn, R. M., and Kent, W. J. (2018). The UCSC Genome Browser database: 2018 update. *Nucleic Acids Res.*, 46(D1):D762–D769.
- Cayrou, C., Ballester, B., Peiffer, I., Fenouil, R., Coulombe, P., Andrau, J.-C., van Helden, J., and Méchali, M. (2015). The chromatin environment shapes DNA replication origin organization and defines origin classes. *Genome Res.*, 25(12):1873–1885.
- Chimpanzee Sequencing and Analysis Consortium (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437(7055):69–87.
- Clamp, M., Fry, B., Kamal, M., Xie, X., Cuff, J., Lin, M. F., Kellis, M., Lindblad-Toh, K., and Lander, E. S. (2007). Distinguishing protein-coding and noncoding genes in the human genome. *Proc. Natl. Acad. Sci. U.S.A.*, 104(49):19428–19433.
- Clark, M. B., Mercer, T. R., Bussotti, G., Leonardi, T., Haynes, K. R., Crawford, J., Brunck, M. E., Cao, K.-A. L., Thomas, G. P., Chen, W. Y., Taft, R. J., Nielsen, L. K., Enright, A. J., Mattick, J. S., and Dinger, M. E. (2015). Quantitative gene profiling of long noncoding RNAs with targeted RNA sequencing. *Nat Methods*, 12(4):339–342.
- Conant, G. C. and Wolfe, K. H. (2008). Turning a hobby into a job: How duplicated genes find new functions. *Nat. Rev. Genet.*, 9(12):938–950.
- Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D. G., Lagarde, J., Veeravalli, L., Ruan, X., Ruan, Y., Lassmann, T., Carninci, P., Brown, J. B., Lipovich, L., Gonzalez, J. M., Thomas, M., Davis, C. A., Shiekhhattar, R., Gingeras, T. R., Hubbard, T. J., Notredame, C., Harrow, J., and Guigó, R. (2012). The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res.*, 22(9):1775–1789.
- Duret, L., Chureau, C., Samain, S., Weissenbach, J., and Avner, P. (2006). The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. *Science*, 312(5780):1653–5.



## 4.1:16 REFERENCES

- Duret, L., Dorkeld, F., and Gautier, C. (1993). Strong conservation of non-coding sequences during vertebrates evolution: Potential involvement in post-transcriptional regulation of gene expression. *Nucleic Acids Res*, 21(10):2315–2322.
- Duret, L. and Galtier, N. (2009). Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet*, 10:285–311.
- ENCODE Project Consortium, Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigó, R., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., Dermitzakis, E. T., Thurman, R. E., Kuehn, M. S., Taylor, C. M., Neph, S., Koch, C. M., Asthana, S., Malhotra, A., Adzhubei, I., Greenbaum, J. A., Andrews, R. M., Flicek, P., Boyle, P. J., Cao, H., Carter, N. P., Clelland, G. K., Davis, S., Day, N., Dhami, P., Dillon, S. C., Dorschner, M. O., Fiegler, H., Giresi, P. G., Goldy, J., Hawrylycz, M., Haydock, A., Humbert, R., James, K. D., Johnson, B. E., Johnson, E. M., Frum, T. T., Rosenzweig, E. R., Karnani, N., Lee, K., Lefebvre, G. C., Navas, P. A., Neri, F., Parker, S. C. J., Sabo, P. J., Sandstrom, R., Shafer, A., Vetrie, D., Weaver, M., Wilcox, S., Yu, M., Collins, F. S., Dekker, J., Lieb, J. D., Tullius, T. D., Crawford, G. E., Sunyaev, S., Noble, W. S., Dunham, I., Denoeud, F., Reymond, A., Kapranov, P., Rozowsky, J., Zheng, D., Castelo, R., Frankish, A., Harrow, J., Ghosh, S., Sandelin, A., Hofacker, I. L., Baertsch, R., Keefe, D., Dike, S., Cheng, J., Hirsch, H. A., Sekinger, E. A., Lagarde, J., Abril, J. F., Shahab, A., Flamm, C., Fried, C., Hackermüller, J., Hertel, J., Lindemeyer, M., Missal, K., Tanzer, A., Washietl, S., Korb, J., Emanuelsson, O., Pedersen, J. S., Holroyd, N., Taylor, R., Swarbreck, D., Matthews, N., Dickson, M. C., Thomas, D. J., Weirauch, M. T., Gilbert, J., Drenkow, J., Bell, I., Zhao, X., Srinivasan, K. G., Sung, W.-K., Ooi, H. S., Chiu, K. P., Foissac, S., Alioto, T., Brent, M., Pachter, L., Tress, M. L., Valencia, A., Choo, S. W., Choo, C. Y., Ucla, C., Manzano, C., Wyss, C., Cheung, E., Clark, T. G., Brown, J. B., Ganesh, M., Patel, S., Tammana, H., Chrast, J., Henrichsen, C. N., Kai, C., Kawai, J., Nagalakshmi, U., Wu, J., Lian, Z., Lian, J., Newburger, P., Zhang, X., Bickel, P., Mattick, J. S., Carninci, P., Hayashizaki, Y., Weissman, S., Hubbard, T., Myers, R. M., Rogers, J., Stadler, P. F., Lowe, T. M., Wei, C.-L., Ruan, Y., Struhl, K., Gerstein, M., Antonarakis, S. E., Fu, Y., Green, E. D., Karaöz, U., Siepel, A., Taylor, J., Liefer, L. A., Wetterstrand, K. A., Good, P. J., Feingold, E. A., Guyer, M. S., Cooper, G. M., Asimenos, G., Dewey, C. N., Hou, M., Nikolaev, S., Montoya-Burgos, J. I., Löytynoja, A., Whelan, S., Pardi, F., Massingham, T., Huang, H., Zhang, N. R., Holmes, I., Mullikin, J. C., Ureta-Vidal, A., Paten, B., Sringhaus, M., Church, D., Rosenbloom, K., Kent, W. J., Stone, E. A., NISC Comparative Sequencing Program, Baylor College of Medicine Human Genome Sequencing Center, Washington University Genome Sequencing Center, Broad Institute, Children's Hospital Oakland Research Institute, Batzoglou, S., Goldman, N., Hardison, R. C., Haussler, D., Miller, W., Sidow, A., Trinklein, N. D., Zhang, Z. D., Barrera, L., Stuart, R., King, D. C., Ameer, A., Enroth, S., Bieda, M. C., Kim, J., Bhingre, A. A., Jiang, N., Liu, J., Yao, F., Vega, V. B., Lee, C. W. H., Ng, P., Shahab, A., Yang, A., Moqtaderi, Z., Zhu, Z., Xu, X., Squazzo, S., Oberley, M. J., Inman, D., Singer, M. A., Richmond, T. A., Munn, K. J., Rada-Iglesias, A., Wallerman, O., Komorowski, J., Fowler, J. C., Couttet, P., Bruce, A. W., Dovey, O. M., Ellis, P. D., Langford, C. F., Nix, D. A., Euskirchen, G., Hartman, S., Urban, A. E., Kraus, P., Van Calcar, S., Heintzman, N., Kim, T. H., Wang, K., Qu, C., Hon, G., Luna, R., Glass, C. K., Rosenfeld, M. G., Aldred, S. F., Cooper, S. J., Halees, A., Lin, J. M., Shulha, H. P., Zhang, X., Xu, M., Haidar, J. N. S., Yu, Y., Ruan, Y., Iyer, V. R., Green, R. D., Wadelius, C., Farnham, P. J., Ren, B., Harte, R. A., Hinrichs, A. S., Trumbower, H., Clawson, H., Hillman-Jackson, J., Zweig, A. S., Smith, K., Thakkapallayil, A., Barber, G., Kuhn, R. M., Karolchik, D., Armengol, L., Bird, C. P., de Bakker, P. I. W., Kern,

- A. D., Lopez-Bigas, N., Martin, J. D., Stranger, B. E., Woodroffe, A., Davydov, E., Dimas, A., Eyras, E., Hallgrímsson, I. B., Huppert, J., Zody, M. C., Abecasis, G. R., Estivill, X., Bouffard, G. G., Guan, X., Hansen, N. F., Idol, J. R., Maduro, V. V. B., Maskeri, B., McDowell, J. C., Park, M., Thomas, P. J., Young, A. C., Blakesley, R. W., Muzny, D. M., Sodergren, E., Wheeler, D. A., Worley, K. C., Jiang, H., Weinstock, G. M., Gibbs, R. A., Graves, T., Fulton, R., Mardis, E. R., Wilson, R. K., Clamp, M., Cuff, J., Gnerre, S., Jaffe, D. B., Chang, J. L., Lindblad-Toh, K., Lander, E. S., Koriabine, M., Nefedov, M., Osoegawa, K., Yoshinaga, Y., Zhu, B., and de Jong, P. J. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816.
- Engreitz, J. M., Haines, J. E., Perez, E. M., Munson, G., Chen, J., Kane, M., McDonel, P. E., Guttman, M., and Lander, E. S. (2016). Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature*, 539(7629):452–455.
- Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., and Merrick, J. M. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, 269(5223):496–512.
- Gerstein, M. B., Bruce, C., Rozowsky, J. S., Zheng, D., Du, J., Korbil, J. O., Emanuelsson, O., Zhang, Z. D., Weissman, S., and Snyder, M. (2007). What is a gene, post-ENCODE? History and updated definition. *Genome Res.*, 17(6):669–681.
- Gerstein, M. B., Lu, Z. J., Van Nostrand, E. L., Cheng, C., Arshinoff, B. I., Liu, T., Yip, K. Y., Robilotto, R., Rechtsteiner, A., Ikegami, K., Alves, P., Chateigner, A., Perry, M., Morris, M., Auerbach, R. K., Feng, X., Leng, J., Vielle, A., Niu, W., Rhrissorakkrai, K., Agarwal, A., Alexander, R. P., Barber, G., Brdlik, C. M., Brennan, J., Brouillet, J. J., Carr, A., Cheung, M.-S., Clawson, H., Contrino, S., Dannenberg, L. O., Dernburg, A. F., Desai, A., Dick, L., Dosé, A. C., Du, J., Egelhofer, T., Ercan, S., Euskirchen, G., Ewing, B., Feingold, E. A., Gassmann, R., Good, P. J., Green, P., Gullier, F., Gutwein, M., Guyer, M. S., Habegger, L., Han, T., Henikoff, J. G., Henz, S. R., Hinrichs, A., Holster, H., Hyman, T., Iniguez, A. L., Janette, J., Jensen, M., Kato, M., Kent, W. J., Kephart, E., Khivansara, V., Khurana, E., Kim, J. K., Kolasinska-Zwierz, P., Lai, E. C., Latorre, I., Leahey, A., Lewis, S., Lloyd, P., Lochovsky, L., Lowdon, R. F., Lubling, Y., Lyne, R., MacCoss, M., Mackowiak, S. D., Mangone, M., McKay, S., Mecnas, D., Merrihew, G., Miller, D. M., Muroyama, A., Murray, J. I., Ooi, S.-L., Pham, H., Phippen, T., Preston, E. A., Rajewsky, N., Rättsch, G., Rosenbaum, H., Rozowsky, J., Rutherford, K., Ruzanov, P., Sarov, M., Sasidharan, R., Sboner, A., Scheid, P., Segal, E., Shin, H., Shou, C., Slack, F. J., Slightam, C., Smith, R., Spencer, W. C., Stinson, E. O., Taing, S., Takasaki, T., Vafeados, D., Voronina, K., Wang, G., Washington, N. L., Whittle, C. M., Wu, B., Yan, K.-K., Zeller, G., Zha, Z., Zhong, M., Zhou, X., modENCODE Consortium, Ahringer, J., Strome, S., Gunsalus, K. C., Micklem, G., Liu, X. S., Reinke, V., Kim, S. K., Hillier, L. W., Henikoff, S., Piano, F., Snyder, M., Stein, L., Lieb, J. D., and Waterston, R. H. (2010). Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. *Science*, 330(6012):1775–1787.
- Gilad, Y., Oshlack, A., and Rifkin, S. A. (2006). Natural selection on gene expression. *Trends Genet.*, 22(8):456–461.
- Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N.,

- and Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.*, 29(7):644–652.
- Graur, D., Zheng, Y., and Azevedo, R. B. R. (2015). An evolutionary classification of genomic function. *Genome Biol Evol*, 7(3):642–645.
- Graur, D., Zheng, Y., Price, N., Azevedo, R. B. R., Zufall, R. A., and Elhaik, E. (2013). On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol Evol*, 5(3):578–590.
- GTEX Consortium (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660.
- Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., Feldser, D., Huarte, M., Zuk, O., Carey, B. W., Cassady, J. P., Cabili, M. N., Jaenisch, R., Mikkelsen, T. S., Jacks, T., Hacohen, N., Bernstein, B. E., Kellis, M., Regev, A., Rinn, J. L., and Lander, E. S. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, 458(7235):223–227.
- Guttman, M., Russell, P., Ingolia, N. T., Weissman, J. S., and Lander, E. S. (2013). Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell*, 154(1):240–251.
- Haerty, W. and Ponting, C. P. (2013). Mutations within lncRNAs are effectively selected against in fruitfly but not in human. *Genome Biol.*, 14(5):R49.
- Haerty, W. and Ponting, C. P. (2014). No gene in the genome makes sense except in the light of evolution. *Annu Rev Genomics Hum Genet*, 15:71–92.
- Haerty, W. and Ponting, C. P. (2015). Unexpected selection to retain high GC content and splicing enhancers within exons of multiexonic lncRNA loci. *RNA*, 21(3):333–346.
- Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., Steward, C., Harte, R., Lin, M., Howald, C., Tanzer, A., Derrien, T., Chrast, J., Walters, N., Balasubramanian, S., Pei, B., Tress, M., Rodriguez, J. M., Ezkurdia, I., van Baren, J., Brent, M., Haussler, D., Kellis, M., Valencia, A., Reymond, A., Gerstein, M., Guigó, R., and Hubbard, T. J. (2012). GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res.*, 22(9):1760–1774.
- He, L. and Hannon, G. J. (2004). MicroRNAs: Small RNAs with a big role in gene regulation. *Nat. Rev. Genet.*, 5(7):522–531.
- Hérault, Y., Rassoulzadegan, M., Cuzin, F., and Duboule, D. (1998). Engineering chromosomes in mice through targeted meiotic recombination (TAMERE). *Nat. Genet.*, 20(4):381–384.
- Hezroni, H., Ben-Tov Perry, R., Meir, Z., Housman, G., Lubelsky, Y., and Ulitsky, I. (2017). A subset of conserved mammalian long non-coding RNAs are fossils of ancestral protein-coding genes. *Genome Biology*, 18(1):1–15.
- Hezroni, H., Koppstein, D., Schwartz, M. G., Avrutin, A., Bartel, D. P., and Ulitsky, I. (2015). Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell Rep*, 11(7):1110–1122.
- Hiller, M., Schaar, B. T., Indjeian, V. B., Kingsley, D. M., Hagey, L. R., and Bejerano, G. (2012). A "forward genomics" approach links genotype to phenotype using independent phenotypic losses among related species. *Cell Rep*, 2(4):817–823.
- Hockemeyer, D., Wang, H., Kiani, S., Lai, C. S., Gao, Q., Cassady, J. P., Cost, G. J., Zhang, L., Santiago, Y., Miller, J. C., Zeitler, B., Cheron, J. M., Meng, X., Hinkley, S. J., Rebar,

- E. J., Gregory, P. D., Urnov, F. D., and Jaenisch, R. (2011). Genetic engineering of human pluripotent cells using TALE nucleases. *Nat. Biotechnol.*, 29(8):731–734.
- Howald, C., Tanzer, A., Chrast, J., Kokocinski, F., Derrien, T., Walters, N., Gonzalez, J. M., Frankish, A., Aken, B. L., Hourlier, T., Vogel, J.-H., White, S., Searle, S., Harrow, J., Hubbard, T. J., Guigó, R., and Reymond, A. (2012). Combining RT-PCR-seq and RNA-seq to catalog all genic elements encoded in the human genome. *Genome Res.*, 22(9):1698–1710.
- Ingolia, N. T., Brar, G. A., Stern-Ginossar, N., Harris, M. S., Talhouarne, G. J. S., Jackson, S. E., Wills, M. R., and Weissman, J. S. (2014). Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep*, 8(5):1365–1379.
- Ingolia, N. T., Ghaemmighami, S., Newman, J. R. S., and Weissman, J. S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, 324(5924):218–223.
- Iyer, M. K., Niknafs, Y. S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., Barrette, T. R., Prensner, J. R., Evans, J. R., Zhao, S., Poliakov, A., Cao, X., Dhanasekaran, S. M., Wu, Y.-M., Robinson, D. R., Beer, D. G., Feng, F. Y., Iyer, H. K., and Chinnaiyan, A. M. (2015). The landscape of long noncoding RNAs in the human transcriptome. *Nat Genet*, 47(3):199–208.
- Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., and Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, 337(6096):816–821.
- Joung, J., Engreitz, J. M., Konermann, S., Abudayyeh, O. O., Verdine, V. K., Aguet, F., Gootenberg, J. S., Sanjana, N. E., Wright, J. B., Fulco, C. P., Tseng, Y.-Y., Yoon, C. H., Boehm, J. S., Lander, E. S., and Zhang, F. (2017). Genome-scale activation screen identifies a lncRNA locus regulating a gene neighbourhood. *Nature*, 548(7667):343–346.
- Kapusta, A. and Feschotte, C. (2014). Volatile evolution of long noncoding RNA repertoires: Mechanisms and biological implications. *Trends Genet.*, 30(10):439–452.
- Katzman, S., Kern, A. D., Bejerano, G., Fewell, G., Fulton, L., Wilson, R. K., Salama, S. R., and Haussler, D. (2007). Human genome ultraconserved elements are ultraselected. *Science*, 317(5840):915.
- Keller, O., Kollmar, M., Stanke, M., and Waack, S. (2011). A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics*, 27(6):757–763.
- Kellis, M., Patterson, N., Birren, B., Berger, B., and Lander, E. S. (2004). Methods in comparative genomics: Genome correspondence, gene identification and regulatory motif discovery. *J. Comput. Biol.*, 11(2-3):319–355.
- Kern, A. D., Barbash, D. A., Chang Mell, J., Hupaló, D., and Jensen, A. (2015). Highly constrained intergenic *Drosophila* ultraconserved elements are candidate ncRNAs. *Genome Biol Evol*, 7(3):689–698.
- Khalil, A. M., Guttman, M., Huarte, M., Garber, M., Raj, A., Rivea Morales, D., Thomas, K., Presser, A., Bernstein, B. E., van Oudenaarden, A., Regev, A., Lander, E. S., and Rinn, J. L. (2009). Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl. Acad. Sci. U.S.A.*, 106(28):11667–11672.
- Kim, M.-S., Pinto, S. M., Getnet, D., Nirujogi, R. S., Manda, S. S., Chaerkady, R., Madugundu, A. K., Kelkar, D. S., Isserlin, R., Jain, S., Thomas, J. K., Muthusamy, B., Leal-Rojas, P., Kumar, P., Sahasrabudde, N. A., Balakrishnan, L., Advani, J., George, B., Renuse, S., Selvan, L. D. N., Patil, A. H., Nanjappa, V., Radhakrishnan, A., Prasad, S., Subbannayya, T., Raju, R., Kumar, M., Sreenivasamurthy, S. K., Marimuthu, A., Sathe, G. J., Chavan, S., Datta, K. K., Subbannayya, Y., Sahu, A., Yelamanchi, S. D., Jayaram,

- S., Rajagopalan, P., Sharma, J., Murthy, K. R., Syed, N., Goel, R., Khan, A. A., Ahmad, S., Dey, G., Mudgal, K., Chatterjee, A., Huang, T.-C., Zhong, J., Wu, X., Shaw, P. G., Freed, D., Zahari, M. S., Mukherjee, K. K., Shankar, S., Mahadevan, A., Lam, H., Mitchell, C. J., Shankar, S. K., Satishchandra, P., Schroeder, J. T., Sirdeshmukh, R., Maitra, A., Leach, S. D., Drake, C. G., Halushka, M. K., Prasad, T. S. K., Hruban, R. H., Kerr, C. L., Bader, G. D., Iacobuzio-Donahue, C. A., Gowda, H., and Pandey, A. (2014). A draft map of the human proteome. *Nature*, 509(7502):575–581.
- Kutter, C., Watt, S., Stefflova, K., Wilson, M. D., Goncalves, A., Ponting, C. P., Odom, D. T., and Marques, A. C. (2012). Rapid turnover of long noncoding RNAs and the evolution of gene expression. *PLoS Genet.*, 8(7):e1002841.
- Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, Y., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglu, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kasprzyk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de

- Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., Szustakowki, J., and International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.
- Lappalainen, T., Sammeth, M., Friedländer, M. R., 't Hoen, P. A. C., Monlong, J., Rivas, M. A., González-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P. G., Barann, M., Wieland, T., Greger, L., van Iterson, M., Almlöf, J., Ribeca, P., Pulyakhina, I., Esser, D., Giger, T., Tikhonov, A., Sultan, M., Bertier, G., MacArthur, D. G., Lek, M., Lizano, E., Buermans, H. P. J., Padioleau, I., Schwarzmayr, T., Karlberg, O., Ongen, H., Kilpinen, H., Beltran, S., Gut, M., Kahlem, K., Amstislavskiy, V., Stegle, O., Pirinen, M., Montgomery, S. B., Donnelly, P., McCarthy, M. I., Flicek, P., Strom, T. M., Geuvadis Consortium, Lehrach, H., Schreiber, S., Sudbrak, R., Carracedo, A., Antonarakis, S. E., Häsler, R., Syvänen, A.-C., van Ommen, G.-J., Brazma, A., Meitinger, T., Rosenstiel, P., Guigó, R., Gut, I. G., Estivill, X., and Dermitzakis, E. T. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511.
- Latos, P. A., Pauler, F. M., Koerner, M. V., Şenergin, H. B., Hudson, Q. J., Stocsits, R. R., Allhoff, W., Stricker, S. H., Klement, R. M., Warczok, K. E., Aumayr, K., Pasierbek, P., and Barlow, D. P. (2012). Airn transcriptional overlap, but not its lncRNA products, induces imprinted Igf2r silencing. *Science*, 338(6113):1469–1472.
- Lee, Y. and Rio, D. C. (2015). Mechanisms and Regulation of Alternative Pre-mRNA Splicing. *Annu. Rev. Biochem.*, 84:291–323.
- Lin, M. F., Carlson, J. W., Crosby, M. A., Matthews, B. B., Yu, C., Park, S., Wan, K. H., Schroeder, A. J., Gramates, L. S., St Pierre, S. E., Roark, M., Wiley, K. L., Kulathinal, R. J., Zhang, P., Myrick, K. V., Antone, J. V., Celniker, S. E., Gelbart, W. M., and Kellis, M. (2007). Revisiting the protein-coding gene catalog of *Drosophila melanogaster* using 12 fly genomes. *Genome Res.*, 17(12):1823–1836.
- Lin, M. F., Deoras, A. N., Rasmussen, M. D., and Kellis, M. (2008). Performance and scalability of discriminative metrics for comparative gene identification in 12 *Drosophila* genomes. *PLoS Comput. Biol.*, 4(4):e1000067.
- Lin, M. F., Jungreis, I., and Kellis, M. (2011). PhyloCSF: A comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*, 27(13):i275–282.
- Marques, A. C. and Ponting, C. P. (2009). Catalogues of mammalian long noncoding RNAs: Modest conservation and incompleteness. *Genome Biol.*, 10(11):R124.
- Mazo-Vargas, A., Concha, C., Livraghi, L., Massardo, D., Wallbank, R. W. R., Zhang, L., Papador, J. D., Martinez-Najera, D., Jiggins, C. D., Kronforst, M. R., Breuker, C. J., Reed, R. D., Patel, N. H., McMillan, W. O., and Martin, A. (2017). Macroevolutionary shifts of WntA function potentiate butterfly wing-pattern diversity. *Proc. Natl. Acad. Sci. U.S.A.*, 114(40):10701–10706.
- McLysaght, A. and Hurst, L. D. (2016). Open questions in the study of de novo genes: What, how and why. *Nat. Rev. Genet.*, 17(9):567–578.
- modENCODE Consortium, Roy, S., Ernst, J., Kharchenko, P. V., Kheradpour, P., Negre, N., Eaton, M. L., Landolin, J. M., Bristow, C. A., Ma, L., Lin, M. F., Washietl, S., Arshinoff, B. I., Ay, F., Meyer, P. E., Robine, N., Washington, N. L., Di Stefano, L., Berezikov, E., Brown, C. D., Candeias, R., Carlson, J. W., Carr, A., Jungreis, I., Marbach, D., Sealfon, R., Tolstorukov, M. Y., Will, S., Alekseyenko, A. A., Artieri, C., Booth, B. W., Brooks, A. N., Dai, Q., Davis, C. A., Duff, M. O., Feng, X., Gorchakov, A. A., Gu, T., Henikoff, J. G., Kapranov, P., Li, R., MacAlpine, H. K., Malone, J., Minoda, A., Nordman, J., Okamura, K., Perry, M., Powell, S. K., Riddle, N. C., Sakai, A., Samsonova, A., Sandler, J. E., Schwartz, Y. B., Sher, N., Spokony, R., Sturgill, D., van Baren, M., Wan, K. H.,

## 4.1:22 REFERENCES

- Yang, L., Yu, C., Feingold, E., Good, P., Guyer, M., Lowdon, R., Ahmad, K., Andrews, J., Berger, B., Brenner, S. E., Brent, M. R., Cherbas, L., Elgin, S. C. R., Gingeras, T. R., Grossman, R., Hoskins, R. A., Kaufman, T. C., Kent, W., Kuroda, M. I., Orr-Weaver, T., Perrimon, N., Pirrotta, V., Posakony, J. W., Ren, B., Russell, S., Cherbas, P., Graveley, B. R., Lewis, S., Micklem, G., Oliver, B., Park, P. J., Celniker, S. E., Henikoff, S., Karpen, G. H., Lai, E. C., MacAlpine, D. M., Stein, L. D., White, K. P., and Kellis, M. (2010). Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science*, 330(6012):1787–1797.
- Mudge, J. M. and Harrow, J. (2016). The state of play in higher eukaryote gene annotation. *Nat. Rev. Genet.*, 17(12):758–772.
- Mudge, J. M., Jungreis, I., Hunt, T., Gonzalez, J. M., Wright, J. C., Kay, M., Davidson, C., Fitzgerald, S., Seal, R., Tweedie, S., He, L., Waterhouse, R. M., Li, Y., Bruford, E., Choudhary, J. S., Frankish, A., and Kellis, M. (2019). Discovery of high-confidence human protein-coding genes and exons by whole-genome PhyloCSF helps elucidate 118 GWAS loci. *Genome Res.*, 29(12):2073–2087.
- Nakamura, K., Akama, T., Bang, P. D., Sekimura, S., Tanigawa, K., Wu, H., Kawashima, A., Hayashi, M., Suzuki, K., and Ishii, N. (2009). Detection of RNA expression from pseudogenes and non-coding genomic regions of *Mycobacterium leprae*. *Microb. Pathog.*, 47(3):183–187.
- Necsulea, A., Soumillon, M., Warnefors, M., Liechti, A., Daish, T., Grutzner, F., and Kaessmann, H. (2014). The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature*, 505(7485):635–640.
- Parker, B. J., Moltke, I., Roth, A., Washietl, S., Wen, J., Kellis, M., Breaker, R., and Pedersen, J. S. (2011). New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. *Genome Res*, 21(11):1929–43.
- Pedersen, J. S., Bejerano, G., Siepel, A., Rosenbloom, K., Lindblad-Toh, K., Lander, E. S., Kent, J., Miller, W., and Haussler, D. (2006). Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comp Biol*, 2(4):e33.
- Pennacchio, L. A., Ahituv, N., Moses, A. M., Prabhakar, S., Nobrega, M. A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K. D., Plajzer-Frick, I., Akiyama, J., De Val, S., Afzal, V., Black, B. L., Couronne, O., Eisen, M. B., Visel, A., and Rubin, E. M. (2006). In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, 444(7118):499–502.
- Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., and Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.*, 33(3):290–295.
- Pertea, M., Shumate, A., Pertea, G., Varabyou, A., Breitwieser, F. P., Chang, Y.-C., Madugundu, A. K., Pandey, A., and Salzberg, S. L. (2018a). CHES: A new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biol.*, 19(1):208.
- Pertea, M., Shumate, A., Pertea, G., Varabyou, A., Chang, Y.-C., Madugundu, A. K., Pandey, A., and Salzberg, S. (2018b). Thousands of large-scale RNA sequencing experiments yield a comprehensive new human gene list and reveal extensive transcriptional noise. *Biorxiv*.
- Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res*, 20(1):110–121.
- Ponjavic, J., Ponting, C. P., and Lunter, G. (2007). Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res*, 17(5):556–65.



- Ratnakumar, A., Mousset, S., Glémin, S., Berglund, J., Galtier, N., Duret, L., and Webster, M. T. (2010). Detecting positive selection within genomes: The problem of biased gene conversion. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, 365(1552):2571–2580.
- Rhesus Macaque Genome Sequencing and Analysis Consortium, Gibbs, R. A., Rogers, J., Katze, M. G., Bumgarner, R., Weinstock, G. M., Mardis, E. R., Remington, K. A., Strausberg, R. L., Venter, J. C., Wilson, R. K., Batzer, M. A., Bustamante, C. D., Eichler, E. E., Hahn, M. W., Hardison, R. C., Makova, K. D., Miller, W., Milosavljevic, A., Palermo, R. E., Siepel, A., Sikela, J. M., Attaway, T., Bell, S., Bernard, K. E., Buhay, C. J., Chandrabose, M. N., Dao, M., Davis, C., Delehaunty, K. D., Ding, Y., Dinh, H. H., Dugan-Rocha, S., Fulton, L. A., Gabisi, R. A., Garner, T. T., Godfrey, J., Hawes, A. C., Hernandez, J., Hines, S., Holder, M., Hume, J., Jhangiani, S. N., Joshi, V., Khan, Z. M., Kirkness, E. F., Cree, A., Fowler, R. G., Lee, S., Lewis, L. R., Li, Z., Liu, Y.-S., Moore, S. M., Muzny, D., Nazareth, L. V., Ngo, D. N., Okwuonu, G. O., Pai, G., Parker, D., Paul, H. A., Pfannkoch, C., Pohl, C. S., Rogers, Y.-H., Ruiz, S. J., Sabo, A., Santibanez, J., Schneider, B. W., Smith, S. M., Sodergren, E., Svatek, A. F., Utterback, T. R., Vattathil, S., Warren, W., White, C. S., Chinwalla, A. T., Feng, Y., Halpern, A. L., Hillier, L. W., Huang, X., Minx, P., Nelson, J. O., Pepin, K. H., Qin, X., Sutton, G. G., Venter, E., Walenz, B. P., Wallis, J. W., Worley, K. C., Yang, S.-P., Jones, S. M., Marra, M. A., Rocchi, M., Schein, J. E., Baertsch, R., Clarke, L., Csürös, M., Glasscock, J., Harris, R. A., Havlak, P., Jackson, A. R., Jiang, H., Liu, Y., Messina, D. N., Shen, Y., Song, H. X.-Z., Wylie, T., Zhang, L., Birney, E., Han, K., Konkel, M. K., Lee, J., Smit, A. F. A., Ullmer, B., Wang, H., Xing, J., Burhans, R., Cheng, Z., Karro, J. E., Ma, J., Raney, B., She, X., Cox, M. J., Demuth, J. P., Dumas, L. J., Han, S.-G., Hopkins, J., Karimpour-Fard, A., Kim, Y. H., Pollack, J. R., Vinar, T., Addo-Quaye, C., Degenhardt, J., Denby, A., Hubisz, M. J., Indap, A., Kosiol, C., Lahn, B. T., Lawson, H. A., Marklein, A., Nielsen, R., Vallender, E. J., Clark, A. G., Ferguson, B., Hernandez, R. D., Hirani, K., Kehrer-Sawatzki, H., Kolb, J., Patil, S., Pu, L.-L., Ren, Y., Smith, D. G., Wheeler, D. A., Schenck, I., Ball, E. V., Chen, R., Cooper, D. N., Giardine, B., Hsu, F., Kent, W. J., Lesk, A., Nelson, D. L., O'Brien, W. E., Prüfer, K., Stenson, P. D., Wallace, J. C., Ke, H., Liu, X.-M., Wang, P., Xiang, A. P., Yang, F., Barber, G. P., Haussler, D., Karolchik, D., Kern, A. D., Kuhn, R. M., Smith, K. E., and Zwing, A. S. (2007). Evolutionary and biomedical insights from the rhesus macaque genome. *Science*, 316(5822):222–234.
- Richards, A. L., Merrill, A. E., and Coon, J. J. (2015). Proteome sequencing goes deep. *Curr Opin Chem Biol*, 24:11–17.
- Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., Thiessen, N., Griffith, O. L., He, A., Marra, M., Snyder, M., and Jones, S. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods*, 4(8):651–657.
- Robinson-Rechavi, M. (2020). Molecular evolution and gene function. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 4.2, pages 4.2:1–4.2:20. No commercial publisher | Authors open access book.
- Romero, I. G., Ruvinsky, I., and Gilad, Y. (2012). Comparative studies of gene expression and the evolution of gene regulation. *Nat. Rev. Genet.*, 13(7):505–516.
- Sanjana, N. E., Wright, J., Zheng, K., Shalem, O., Fontanillas, P., Joung, J., Cheng, C., Regev, A., and Zhang, F. (2016). High-resolution interrogation of functional elements in the noncoding genome. *Science*, 353(6307):1545–1549.
- Sauvageau, M., Goff, L. A., Lodato, S., Bonev, B., Groff, A. F., Gerhardinger, C., Sanchez-



- Gomez, D. B., Hacisuleyman, E., Li, E., Spence, M., Liapis, S. C., Mallard, W., Morse, M., Swerdel, M. R., D'Ecclessis, M. F., Moore, J. C., Lai, V., Gong, G., Yancopoulos, G. D., Friendewey, D., Kellis, M., Hart, R. P., Valenzuela, D. M., Arlotta, P., and Rinn, J. L. (2013). Multiple knockout mouse models reveal lincRNAs are required for life and brain development. *Elife*, 2:e01749.
- Schüler, A., Ghanbarian, A. T., and Hurst, L. D. (2014). Purifying selection on splice-related motifs, not expression level nor RNA folding, explains nearly all constraint on human lincRNAs. *Mol. Biol. Evol.*, 31(12):3164–3183.
- Seemann, S. E., Mirza, A. H., Hansen, C., Bang-Berthelsen, C. H., Garde, C., Christensen-Dalsgaard, M., Torarinsson, E., Yao, Z., Workman, C. T., Pociot, F., Nielsen, H., Tommerup, N., Ruzzo, W. L., and Gorodkin, J. (2017). The identification and functional annotation of RNA structures conserved in vertebrates. *Genome Res.*, 27(8):1371–1383.
- Shalem, O., Sanjana, N. E., Hartenian, E., Shi, X., Scott, D. A., Mikkelsen, T., Heckl, D., Ebert, B. L., Root, D. E., Doench, J. G., and Zhang, F. (2014). Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science*, 343(6166):84–87.
- Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., Workman, G. M., Wilson, R. K., Gibbs, R. A., Kent, W. J., Miller, W., and Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, 15(8):1034–50.
- Slater, G. S. C. and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, 6:31.
- Smith, G. R. (1994). Hotspots of homologous recombination. *Experientia*, 50(3):234–241.
- Souvorov, A., Kapustin, Y., Kiryutin, B., Chetvernin, V., Tatusova, T., and Lipman, D. (2010). Gnomon – NCBI eukaryotic gene prediction tool. *NCBI*.
- Stanke, M., Tzvetkova, A., and Morgenstern, B. (2006). AUGUSTUS at EGASP: Using EST, protein and genomic alignments for improved gene prediction in the human genome. *Genome Biol.*, 7 Suppl 1:S11.1–8.
- Stuart, J. M., Segal, E., Koller, D., and Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643):249–255.
- Studer, R. A. and Robinson-Rechavi, M. (2009). How confident can we be that orthologs are similar, but paralogs differ? *Trends Genet.*, 25(5):210–216.
- Tian, W. and Skolnick, J. (2003). How well is enzyme function conserved as a function of pairwise sequence identity? *J. Mol. Biol.*, 333(4):863–882.
- Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, 28(5):511–515.
- Van Bortle, K. and Corces, V. G. (2012). Nuclear organization and genome function. *Annu. Rev. Cell Dev. Biol.*, 28:163–187.
- Vilella, A. J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. (2009). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.*, 19(2):327–335.
- Visel, A., Blow, M. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., Afzal, V., Ren, B., Rubin, E. M., and Pennacchio, L. A. (2009). ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, 457(7231):854–858.
- Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10(1):57–63.

- Washietl, S., Hofacker, I. L., and Stadler, P. F. (2005). Fast and reliable prediction of noncoding RNAs. *Proc. Natl. Acad. Sci. U.S.A.*, 102(7):2454–2459.
- Washietl, S., Kellis, M., and Garber, M. (2014). Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. *Genome Res*, 24:616–28.
- Weick, E.-M. and Miska, E. A. (2014). piRNAs: From biogenesis to function. *Development*, 141(18):3458–3471.
- Wiberg, R. A. W., Halligan, D. L., Ness, R. W., Necsulea, A., Kaessmann, H., and Keightley, P. D. (2015). Assessing Recent Selection and Functionality at Long Noncoding RNA Loci in the Mouse Genome. *Genome Biol Evol*, 7(8):2432–2444.
- Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A. M., Lieberenz, M., Savitski, M. M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H., Mathieson, T., Lemeer, S., Schnatbaum, K., Reimer, U., Wenschuh, H., Mollenhauer, M., Slotta-Huspenina, J., Boese, J.-H., Bantscheff, M., Gerstmaier, A., Faerber, F., and Kuster, B. (2014). Mass-spectrometry-based draft of the human proteome. *Nature*, 509(7502):582–587.
- Young, R. S., Marques, A. C., Tibbit, C., Haerty, W., Bassett, A. R., Liu, J.-L., and Ponting, C. P. (2012). Identification and properties of 1,119 candidate lincRNA loci in the *Drosophila melanogaster* genome. *Genome Biol Evol*, 4(4):427–442.
- Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., Sandstrom, R., Ma, Z., Davis, C., Pope, B. D., Shen, Y., Pervouchine, D. D., Djebali, S., Thurman, R. E., Kaul, R., Rynes, E., Kirilusha, A., Marinov, G. K., Williams, B. A., Trout, D., Amrhein, H., Fisher-Aylor, K., Antoshechkin, I., DeSalvo, G., See, L.-H., Fastuca, M., Drenkow, J., Zaleski, C., Dobin, A., Prieto, P., Lagarde, J., Bussotti, G., Tanzer, A., Denas, O., Li, K., Bender, M. A., Zhang, M., Byron, R., Groudine, M. T., McCleary, D., Pham, L., Ye, Z., Kuan, S., Edsall, L., Wu, Y.-C., Rasmussen, M. D., Bansal, M. S., Kellis, M., Keller, C. A., Morrissey, C. S., Mishra, T., Jain, D., Dogan, N., Harris, R. S., Cayting, P., Kawli, T., Boyle, A. P., Euskirchen, G., Kundaje, A., Lin, S., Lin, Y., Jansen, C., Malladi, V. S., Cline, M. S., Erickson, D. T., Kirkup, V. M., Learned, K., Sloan, C. A., Rosenbloom, K. R., Lacerda de Sousa, B., Beal, K., Pignatelli, M., Flicek, P., Lian, J., Kahveci, T., Lee, D., Kent, W. J., Ramalho Santos, M., Herrero, J., Notredame, C., Johnson, A., Vong, S., Lee, K., Bates, D., Neri, F., Diegel, M., Canfield, T., Sabo, P. J., Wilken, M. S., Reh, T. A., Giste, E., Shafer, A., Kutayavin, T., Haugen, E., Dunn, D., Reynolds, A. P., Neph, S., Humbert, R., Hansen, R. S., De Bruijn, M., Selleri, L., Rudensky, A., Josefowicz, S., Samstein, R., Eichler, E. E., Orkin, S. H., Levasseur, D., Papayannopoulou, T., Chang, K.-H., Skoultschi, A., Gosh, S., Disteche, C., Treuting, P., Wang, Y., Weiss, M. J., Blobel, G. A., Cao, X., Zhong, S., Wang, T., Good, P. J., Lowdon, R. F., Adams, L. B., Zhou, X.-Q., Pazin, M. J., Feingold, E. A., Wold, B., Taylor, J., Mortazavi, A., Weissman, S. M., Stamatoyannopoulos, J. A., Snyder, M. P., Guigo, R., Gingeras, T. R., Gilbert, D. M., Hardison, R. C., Beer, M. A., Ren, B., and Mouse ENCODE Consortium (2014). A comparative encyclopedia of DNA elements in the mouse genome. *Nature*, 515(7527):355–364.
- Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C. G., Gil, L., Gordon, L., Haggerty, L., Haskell, E., Hourlier, T., Izuogu, O. G., Janacek, S. H., Juettemann, T., To, J. K., Laird, M. R., Lavidas, I., Liu, Z., Loveland, J. E., Maurel, T., McLaren, W., Moore, B., Mudge, J., Murphy, D. N., Newman, V., Nuhn, M., Ogeh, D., Ong, C. K., Parker, A., Patricio, M., Riat, H. S., Schuilenburg, H., Sheppard, D., Sparrow, H., Taylor, K., Thormann, A., Vullo, A., Walts, B., Zadissa, A., Frankish, A., Hunt, S. E., Kostadima, M., Langridge, N., Martin, F. J., Muffato, M., Perry, E., Ruffier, M., Staines, D. M., Trevanion, S. J., Aken,

#### 4.1:26 REFERENCES

B. L., Cunningham, F., Yates, A., and Flicek, P. (2018). Ensembl 2018. *Nucleic Acids Res.*, 46(D1):D754–D761.