Università Degli Studi Di Torino

Facoltà di Scienze Matematiche Fisiche e Naturali

Dipartimento di Fisica Teorica

Dottorato di Ricerca in Sistemi Complessi Applicati alla Biologia Post-Genomica

CICLO XX

PARALOGOUS ALIGNMENTS AS A TOOL TO INVESTIGATE GENOMIC INFORMATION

TESI PRESENTATA DA: Dott. Ivan Molineris	TUTOR: Prof. M. Caselle
COORDINATORE DEL CICLO: Prof. F. Bussolino	RELATORI ESTERNI: Prof. G. Valle Prof. C. Herrmann

Anni Accademici: 2005 - 2007
SETTORE SCIENTIFICO-DISCIPLINARE DI APPARTENENZA: FIS-02

probis viris

Abstract

In the last few years the amount of information about genomes, especially accurate complete sequences, has been exponentially increasing. Despite this abundance of information the interpretation in biological significant terms of the entire genomic sequence of an organism remains a challenge of the post-genomic scientific era.

In this study we propose paralogous alignments (i.e. the alignments of a genome with itself) as a tool to extract meaningful information from raw genomic sequences. We computed a complete database of such alignments for a few organisms and we developed a set of software tools to mine, collect, visualize and integrate these data with the present knowledge about genomic sequences.

As a first result we were able to identify previously unknown genes (chapter 2).

As a second step we adopted a more abstract perspective which was inspired by two considerations: on one hand many known languages are structured so that the used words are only few of the possible combinations of characters; on the other hand, a word is often present many times in the same text. Therefore we searched for sequences of nucleotides occurring many times in the genomes using paralogous alignments and managing them with graph theory concepts. Within the "genomic words" that we found there are many well known sequences, such as protein domains, but also previously uncharacterized sequences (chapter 3).

Alongside this principal project, which is centred on a genomic level, we investigated other ways to extract meaningful information from large sets of publicly available biological data. We developed a strategy able to handle gene expression profiles in order to infer interactions among genes at proteomic level starting from data at the transcriptomic level (appendix A).

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Nowadays the “systems biology” is one of the most fascinating and prospering research field. This term has a lot of definitions, but the underlying common perception of systems biology researchers is that the living realm of nature is characterized by complex behaviors which cannot be understood in a purely reductionist framework.

Each scale (from sub-cellular molecular processes to population dynamics and ecosystems) shows complex behaviors which emerge from the interaction of many subsystems that are in turn complex systems.

Each scale has a typical formalism in order to explain the observations and to make predictions. However, even when the scientific principles governing a specific scale are very well understood, the organization at the nearest higher scale is often impossible to predict starting from these principles. For instance the prediction of the structure and function of a protein is actually almost impossible starting from physical and biochemical principles.

This thesis aims to give some contributions to the study of the biomolecular scale and in particular we focused our attention to the processing of the information that lives in the genomes. This information gives rise to the molecules that are the fundamentals players in the complex team play which makes the cell a living being.

A commonly accepted theoretical framework able to completely describe this particular field of research in its wholeness has not been developped yet. Even if the Darwin’s theory of evolution provides the underlying principles, only in few cases this theory can be directly applied in a well formalized way in order to explain phenomena occurring at the biomolecular scale. Besides the Darwin’s theory another paradigm drives the biomolecular interpretations: the so called “central dogma of molecular biology”. It was proposed by Francis Crick in 1958, after the discovery of the structure of DNA molecule. According to this paradigm, the information flows in the cell become of prime interest.

The “central dogma of molecular biology” has the great merit to be clear and simple in its statement, but nowadays too many phenomena do not fit with the original formulation and a more general one is required.

1.1 The extended central dogma

1.1.1 The original formulation and the exceptions

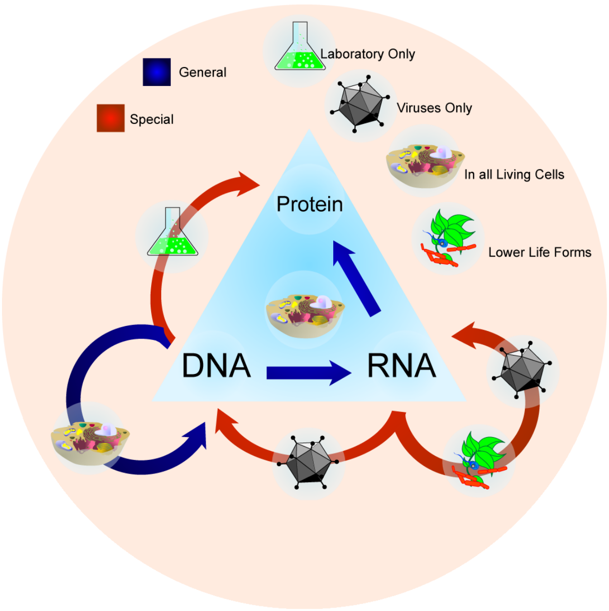


Figure 1.1: Central dogma: general and special kinds of information flows in molecular biology. Modified from [1].

The central dogma of molecular biology was first enunciated by Francis Crick in 1958 [2] and re-stated in a Nature paper published in 1970 [3]:

> *The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred back from protein to either protein or nucleic acid.*

In other words, "once information gets into protein, it can't flow back to nucleic acid."

Strictly speaking the principle is true, indeed there are not biomolecular mechanism, neither in living organisms nor in artificial ones, that can produce nucleic acid polymers with a sequence specified in a protein. Nevertheless the Crick paradigm is often intended in a more rich form asserting that:

- the genetic information is stored in DNA,
- each gene is a portion of DNA,
- each gene has a specific function,
- the function of a gene is expressed by a protein,
- the primary structure of a protein is encoded in the gene sequence,
- an RNA molecule carries the information from the DNA level to the protein (trough a transcription and translation process)

Nowadays this is a commonly accepted framework; the usual information flows in a cell (blue arrows in figure 1.1) are

- from DNA to DNA during replication,
- from DNA to RNA during transcription,
- from RNA to protein during translation.

Nevertheless some remarkable exceptions (red arrows in figure 1.1) have to be taken into account.

Retrotranscription

In 1970 the scientists Howard Temin and David Baltimore independently discovered an enzyme capable to make a double stranded DNA molecule from a single stranded RNA template. They denoted this enzyme reverse transcriptase because it acts in the opposite or reverse direction of transcription.

The retrotranscriptase is fundamental for the life-cycle of retroviruses such as the human immunodeficiency virus (HIV). The genome of this kind of viruses consists of RNA molecules; when the viruses entered in to a cell and uncoated, the genome is reverse transcribed into double stranded DNA which can be incorporated into the host cell and subsequently expressed.

Higher eukaryotes present transposable elements that duplicate themselves in the genome in a similar way (see section 1.5).

RNA-RNA replication

An example of RNA-RNA replication is provided by a class of viruses (including Coronavirus and SARS) which have their genome directly utilized as if it were mRNA, producing a single protein which is modified by host and viral proteins in order to form the various proteins needed for replication. One of these proteins is a RNA replicase, which copies the viral RNA to form a double-stranded replicative form, which in turn directs the formation of new virions [3].

Direct translation from DNA to protein

Direct translation from DNA to protein has been demonstrated in a test tube, using extracts from E. Coli that contained ribosomes, but the same result was impossible to obtain with intact cells [4].

1.1.2 Taking regulation and RNA-processing into account

The relatively simple scheme of central dogma depicted in figure 1.1 becomes much more complex if we take into account the regulatory information flow.

In the original enunciation (which is still valid) the central dogma refers to blueprint information that describes the primary molecular structure. In order to function, the cell needs blueprint and machinery able to build molecules, but it also needs a system able to decide when a certain molecular species has to be produced and the rate of its production. This system is often denoted by “gene regulation”.

Ultimately the information which drives the rate of production of various molecular species (mainly different proteins) originates from the environment in which the cell lives, or from changes in its internal state. Usually this information is picked up by specialized sensor proteins which initiate a complex information processing: the information returs to DNA in a way which is more or less opposite to that of standard molecular blueprint which flows form DNA to proteins.

The figure 1.2 is an attempt to depict in a simple and schematic way the principal processes, actors and flows in the information processing of the eucariotic cell, taking into account the regulation.

RNA molecules are involved in gene regulation

The transcription and the translation processes are fundamental steps in the canonical (molecular blueprint) information flow which goes from DNA to protein (blue arrows in figure 1.2). They are also the key step in which regulators proteins intervene in order to promote or to block the production of a certain molecular species.

As a consequence of the standard statement of the central dogma, it has been generally assumed that the RNA has only the role of passive messanger which carries blueprint information from DNA to protein[1], then only proteins may regulate the gene expression. In the early 2000’s a new class of regulatory RNA, sometimes denoted as smallRNA or sRNA (including miRNA and siRNA), became popular in the scientific community[2] and it

[1] The tRNA and rRNA class of RNA molecule are well known since the early 60’s and were considered as notable exception.

[2] In 1990, plant scientists at a biotechnology company were studying enzymes that formed anthocyanin, the pigment that makes petunias purple [5]. Testing whether chalcone synthase (CHS), was the rate-limiting enzyme in anthocyanin biosynthesis, they overexpressed chalcone synthase in petunias: "Unexpectedly the

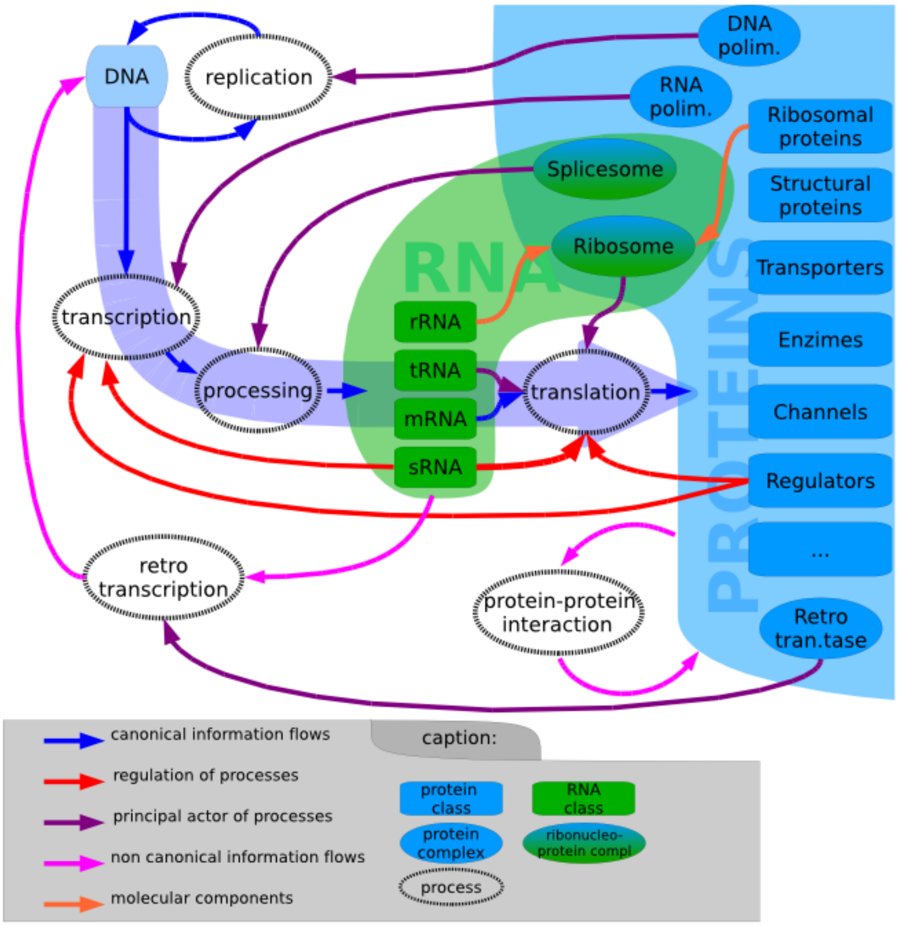


Figure 1.2: Extended central dogma.

is now clear that not only proteins intervene in the gene regulation process but also many kind of RNA molecules have an important role. The smalRNAs intervene not only at the transcription level but mainly at post-transcriptional and translation level.

RNA-processing

Now it is clear that a wide class of RNA molecules is fundamentally involved in gene regulation but also that the standard messanger RNA (mRNA) are not passive carriers.

The initial RNA copy of a gene encoded in the DNA (preRNA) undego a cut and paste processing (splicing) in which certain portions are removed (introns) and others (exons) are retained. The final product, called mature mesanger (mRNA) is ready to be potentially translated into a protein.

Each preRNA may undergo alternative splicing process and hence one single gene (that produce one single preRNA) can produce many different mRNA and as many different proteins.

The splicing mechanism is regulated in a complex and still unclear way. Remarkabley the principal molecular complex involved in this process (the spliceosome) is a ribonucleoprotein complex, i.e. its components are both proteins and RNA molecules. The ribosome has the same mixed composition and it is the crucial machinery involved in the translation process. Thus the RNA has several crucial roles: carrying and processing information, gene regulation, enzymatic activity in fundamental molecular machinery.

Protein-protein interaction

Alongside the transcription-processing-translation flow, regulated by sRNA and regulators proteins, there are other processes to take into account: the replication process (in which the DNA is duplicated in order to give the genome to offspring) and the retrotransposition processes (see section 1.1.1 and 1.5).

Moreover, considering the information processing of the cell as a whole, we cannot forget the protein-protein interactions: many responses to external or internal stimuli or to environmental changes, involve only a cascade or a network of interaction among proteins, while they do not involve the RNA or DNA level. The protein-protein interactions induced by a stimulus often produce variations in some proteins (cleavage, phosphorillation, etc) or in the topology of the interaction network itself; these changes allow new biological functions (sometimes very complex) which react to the initial stimulus.

1.2 Genes and genomes

The main topic of this thesis is a novel computational methodology devoted to gene prediction. The concept of gene remains central even when we assume a more abstract perspective: looking in the genomes for interesting patterns of symbols we often anchor our observation to genes. For these reasons a clear definition of a gene is necessary, and hence a brief historical evolution of the concept of gene is reported afterward.

The more expert scientists become in molecular genetics, the less easy it is to be sure about what a gene actually is [7].

In classical genetics, a gene was an abstract concept — a unit of inheritance that ferried a characteristic from parent to child. As bio-chemistry came into its own, those characteristics were associated with enzymes or proteins, one for each gene. And with the advent of molecular biology, genes became real, physical objects — sequences of DNA which when converted into strands of so-called messenger RNA could be used as the basis for building their associated protein piece by piece. The great coiled DNA molecules of the chromosomes were seen as long strings on which gene sequences sat like discrete beads.

This picture is still the working model for many scientists. But those at the forefront of genetic research see it as increasingly old-fashioned — a crude approximation that, at

introduced gene created a block in anthocyanin biosynthesis" [5] and 42% of the plants became white or had chimeric purple-white patterns [6]. This was the first experiment of RNA interferece and it opened the research field of RNA-mediated regulation.

best, hides fascinating new complexities and, at worst, blinds its users to useful new paths of enquiry [7].

Information seems to be parcelled out along chromosomes in a much more complex way than that originally supposed. Moreover, the RNA molecules are not just passive conduits through which the gene's message flows, but they are active regulators of cellular processes and messengers across generations [8].

1.2.1 Historical perspective on the concept of gene

1860s—1900s: Gene as a discrete unit of heredity

There are various definitions of the term "gene", although common initial descriptions include the ability to determine a particular characteristic of an organism and the heritability of this characteristic. In particular, the word gene was first used by Wilhelm Johannsen in 1909, based on the concept developed by Gregor Mendel in 1866 [9]. The word was a derivative of pangene, which was used by Hugo De Vries for entities involved in pangenesis, Darwin's hypothetical mechanism of heredity [10]. Johnannsen called a gene the "special conditions, foundations and determiners which are present [in the gametes] in unique, separate and thereby independent ways [by which] many characteristics of the organism are specified" [11]. The etymology of the term derives from the Greek genesis ("birth") or genos ("origin").

Mendel showed that when breeding plants, some traits such as height or flower color do not appear blended in their offspring — that is, these traits are passed on as distinct, discrete entities. His work also demonstrated that variations in traits were caused by variations in inheritable factors (or, in today's terminology, phenotype is caused by genotype). It was only after Mendel's work was repeated and rediscovered by Carl Correns, Erich von Tschermak-Seysenegg, and Hugo De Vries in 1900 that further work on the nature of the unit of inheritance truly began [12].

1910s: Gene as a distinct locus

In the next major development, the American genetist Thomas Hunt Morgan and his students were studying the segregation of mutations in Drosophila melanogaster. They were able to explain their data with a model that genes are arranged linearly, and their ability to cross-over is proportional to the distance that separated them [9]. The first genetic map was created in 1913 and Morgan and his students published "The Mechanism of Mendelian Inheritance" in 1915 [13]. To the early geneticists, a gene was an abstract entity whose existence was reflected in the way phenotypes were transmitted between generations. The methodology used by early geneticists involved mutations and recombination, so the gene was essentially a locus whose size was determined by mutations that inactivated (or activated) a trait of interest and by the size of the recombining regions [9]. The fact that genetic linkage corresponded to physical locations on chromosomes was shown later, in 1929, by Barbara McClintock, in her cytogenetic studies on maize [14] while the fact that genetic information reside on chromosomes was already observed by Theoder Boveri in 1907 [15].

1940s: Gene as a blueprint for a protein

Beadle and Tatum [16], who studied Neurospora metabolism, discovered that mutations in genes could cause defects in steps in metabolic pathways. This was stated as the "one gene, one enzyme" view, which later became "one gene, one polypeptide." In this viewpoint, the gene is being implicitly considered as the information behind the individual molecules in a biochemical pathway [9]. This view became progressively more explicit and mechanistic in later decades.

1950s: Gene as a physical molecule

The fact that heredity has a physical, molecular basis was demonstrated by the observation that X rays could cause mutations [17]. Griffith's [18] demonstration that something in virulent but dead Pneumococcus strains could be taken up by live nonvirulent Pneumococcus and transform them into virulent bacteria was further evidence in this direction. It was later shown that this substance could be destroyed by the enzyme DNase [19]. In 1955, Hershey and Chase established that the substance actually transmitted by bacteriophage to their progeny is DNA and not protein [20]. .

1960s: Gene as transcribed code

It was the solution of the three-dimensional structure of DNA by Watson and Crick in 1953 [21] that explained how DNA could function as the molecule of heredity. Base pairing explained how genetic information could be copied, and the existence of two strands explained how occasional errors in replication could lead to a mutation in one of the daughter copies of the DNA molecule.

From the 1960s on, molecular biology developed at a rapid pace. The RNA transcript of the protein-coding sequences was translated using the genetic code [22] into an amino acid sequence. Francis Crick [23] summarized the flow of information in gene expression as from nucleic acid to protein (the beginnings of the "Central Dogma", see section 1.1). However, there were some immediate exceptions to this: it was known that some genes code not for protein but for functional RNA molecules such as rRNA and tRNA. In addition, in RNA viruses the gene is made of RNA. The molecular view of the gene that developed through the 1960s can be summarized in general terms to be a code residing on nucleic acid that gives rise to a functional product [9].

1970s—1980s: Gene as open reading frame (ORF) sequence pattern

The development of cloning and sequencing techniques in the 1970s, combined with knowledge of the genetic code, revolutionized the field of molecular biology by providing a wealth of information on how genes are organized and expressed [9]. The first gene to be sequenced was from the bacteriophage MS2, which was also the first organism to be fully sequenced [24]. The parallel development of computational tools led to algorithms for the identification of genes based on their sequence characteristics (see section 1.3). In many cases, a DNA sequence could be used to infer structure and function for the gene and its products. This situation created a new concept of the "nominal gene", which is defined by its predicted sequence rather than as a genetic locus responsible for a phenotype [25]. The identification of most genes in sequenced genomes is based either on their similarity to other known genes, or the statistically significant signature of a protein-coding sequence [9].

1990s—2000s: Annotated genomic entity, enumerated in the databanks (current view, pre-ENCODE)

The current definition of a gene used by scientific organizations that annotate genomes still relies on the sequence view. Thus, a gene was defined by the Human Genome Nomenclature Organization as "a DNA segment that contributes to phenotype/function. In the absence of demonstrated function a gene may be characterized by sequence, transcription or homology" [26]. Recently, the Sequence Ontology Consortium called the gene a "locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions and/or other functional sequence regions" [27].

The sequencing of first the Haemophilus influenza genome and then the human genome [28, 29, 30] led to an explosion in the amount of sequence that definitions such as the above could be applied to. In fact, there was a huge popular interest in counting the number of genes in various organisms. This interest was crystallized originally by Gene Sweepstake's wager on the number of genes in the human genome, which received extensive media coverage. The landmark human genome sequencing surprised many with the small number (relative to simpler organisms) of protein-coding genes that sequence annotators could identify (21000 [31]),

It has been pointed out that these enumerations overemphasize traditional, protein-coding genes. In particular, when the number of genes present in the human genome was reported in 2003, it was acknowledged that too little was known about RNA-coding genes, such that the given number was that of protein-coding genes [9]. Moreover alternatively spliced transcripts, in the Gene Sweepstake's Wagner definition, all belong to the same gene, even if the proteins that are produced are different [32].

A current computational metaphor: Genes as "subroutines" in the genomic operating system

Given that counting genes in the genome is such a large-scale computational endeavor and that genes fundamentally deal with information processing, the lexicon of computer science naturally has been increasingly applied to describing them. In particular, people in the computational biology community have used the description of a formal language to describe the structure of genes in very much the same way that grammars are used to describe computer programs [33]. Moreover, one metaphor that is increasingly popular for describing genes is to think of them in terms of subroutines in a huge operating system. That is, insofar as the nucleotides of the genome are put together into a code that is executed through the process of transcription and translation, the genome can be thought of as an operating system for a living being. Genes are then individual subroutines in this overall system that are repetitively called in the process of transcription [33].

1.2.2 Recent development: FANTOM and ENCODE

Two recent international efforts devoted to functional annotations of genomes gave us a bumper crop of data that upset considerably the classical concept of gene.

The Functional ANnoTation of Mouse (FANTOM) consortium [34] was originally a Japanese national project for establishing a system for connecting genes with phenotypes and drug targets/effects by using the same platform in multiple biological systems [34]. It became an international consortium in 2001 with the aim to provide the ultimate characterization of the mouse transcriptome. It released the last data in 2005 [35].

The National Human Genome Research Institute (NHGRI) launched a public research consortium named ENCODE, the Encyclopedia Of DNA Elements, in September 2003, to carry out a project to identify all functional elements in the human genome sequence [36]. In the pilot phase (concluded in 2007), ENCODE researchers devised and tested high-throughput approaches for identifying functional elements in the genome. Those elements included genes that code for proteins; genes that do not code for proteins; regulatory elements that control the transcription of genes; and elements that maintain the structure of chromosomes and mediate the dynamics of their replication [37]. The pilot project focused on 44 targets, which together cover about 1 percent of the human genome sequence, or about 30 million DNA base pairs.

These projects represent a major milestone in the characterization of genomes, and the current findings show a striking picture of complex molecular activity [9]. Before the advent of FANTOM and ENCODE projects, there were a number of aspects of genes that where complex to explain and did not fit well with the common concept of gene, but much of this complexity was in some sense swept under the rug and did not really affect the fundamental definition of a gene [9]. Indeed much of the new observation are not really new: during the past decades almost each special characteristic shown by FANTOM or ENCODE where already depicted in some scientific paper. However these recent large international efforts demonstrated that these characteristics are not peculiar of few special genes but they are widespread in the genome. Therefore the scientific community realized that these observation must fit the proper concept of what a gene is.

Classical works on special features of genes

Concerning the definition of a gene it is hard to decide what is special and what is not, likewise it is not always clear who experimentally depicted a new phenomenon at first. Nevertheless it is remarkable that several complex gene characteristics which nowadays are fashionable, were discovered many years ago. Table 1.1 briefly reports some particular cases.

aspect involved	phenomenon	description	principal author	year	complex characteristics
Structural variation	Mobile elements	Genetic element appears in new locations over generations	McClintock	1948	A genetic element may be not constant in its location
Gene location and structure	Enhancers, silencers	Distant regulatory elements	Nash	1970	DNA sequences determining expression can be widely separated from one another in genome. Many-to-many relationship between genes and their enhancers.
Gene location and structure	Antisense transcript	A gene locus is transcribed in both strands	Spiegelman	1972	Multiple products from one genetic locus
Epigenetics and chromosome structure	Effect of chromatin structure	Chromatin structure, which does influence gene expression, only loosely associated with particular DNA sequences	Paul	1972	Gene expression depends on packing of DNA. DNA sequence is not enough to predict gene product.
Epigenetics and chromosome structure	Epigenetic modifications, imprinting	Inherited information may not be DNA-sequence based; a gene's expression depends on whether it is of paternal or maternal origin	Sager and Kitchin	1975	Phenotype is not determined strictly by genotype
Post-translational events	Protein splicing, viral polyproteins	Protein product self-cleaves and can generate multiple functional products	Villa-Komaroff	1975	Start and end sites of protein not determined by genetic code
Pseudogenes	Retrogenes	A retrogene is formed from reverse transcription of its parent gene's mRNA and by insertion of the DNA product into a genome	Jacq	1977	RNA-to-DNA flow of information
Post-transcriptional events	Alternative splicing of RNA	One transcript can generate multiple mRNAs, resulting in different protein products	Berget, Gelinas, Roberts	1977	Multiple products from one genetic locus; information in DNA not linearly related to that on protein
Gene location and structure	Genes with overlapping reading frames	A DNA region may code for two different protein products in different reading frames	Contreras	1977	No one-to-one correspondence between DNA and protein sequence

Table 1.1: Chronological report of the most important classical experimental work that depicted the complexity of the gene concept

aspect involved	phenomenon	description	principal author	year	complex characteristics
Structural variation	Copy-number variants	Copy number of genes/regulatory elements may differ between individuals	Schimke	1978	Genetic elements may differ in their number
Structural variation	Gene rearrangements	DNA rearrangement or splicing in somatic cells results in many alternative gene products	Early	1980	Gene structure is not hereditary, or structure may differ across individuals or cells/tissues
Post-translational events	Protein modification	Protein is modified to alter structure and function of the final product	Wold	1981	The information on the DNA is not encoded directly into protein sequence
Post-transcriptional events	RNA trans-splicing, homotypic trans-splicing	Distant DNA sequences can code for transcripts ligated in various combinations. Two identical transcripts of a gene can trans-splice to generate an mRNA where the same exon sequence is repeated.	Solnick	1985	A protein can result from the combined information encoded in multiple transcripts
Gene location and structure	Intronic genes	A gene exists within an intron of another	Henikoff	1986	Two genes in the same locus
Post-translational events	RNA editing	RNA is enzymatically modified	Eisen	1988	The information on the DNA is not encoded directly into RNA sequence
Post-transcriptional events	Alternatively spliced products with alternate reading frames	Alternative reading frames of the INK4a tumor suppressor gene encodes two unrelated proteins	Quelle	1995	Two alternative splicing products of a pre-mRNA produce protein products with no sequence in common
Post-translational events	Protein trans-splicing	Distinct proteins can be spliced together in the absence of a trans-spliced transcript	Handa	1996	Start and end sites of protein not determined by genetic code
Pseudogenes	Transcribed pseudogenes	A pseudogene is transcribed	Korneev	1999	Biochemical activity of supposedly dead elements

1.2.3 Some aspects of genomic complexity

ENCODE and FANTOM mainly analyzed the transcriptome, using different experimental techniques such as CAGE, tiling arrays, RACE and other [38]. A first finding from this techniques is that a vast amount of DNA, not annotated as known genes, is transcribed into RNA [9].

These novel transcribed regions are usually called TARs (i.e., transcriptionally active regions) or transcriptional fores . While the majority of the genome appears to be transcribed at the level of primary transcripts, only about half of the processed (spliced) transcription detected across all the cell lines and conditions mapped is currently annotated as genes [9]. Moreover a considerable fraction of TARs do not contain transcript with a well characterized ORF.

Thanks to these novel techniques the detectable transcribed genomic fraction is noticeably increased, but even more astonishing is the complexity of the transcription processes with respect to the previous expectation.

The following sections describe some of the most important observations about transcription complexity that FANTOM and ENCODE highlighted.

Gene overlapping

Some genes have been found to overlap one another, sharing the same DNA sequence in a different reading frame or on the opposite strand. The discontinuous structure of genes potentially allows one gene to be completely contained inside another one's intron, or one gene to overlap with another on the same strand without sharing any exons or regulatory elements [9]. Thousands of examples of ncNAT (non-coding Natural AnTisense) transcript have been discovered [39], many of them have a role in the expression regulation of the corresponding sense transcript.

Unannotated and alternative transcription start sites

There are a large number of unannotated transcription start sites (TSSs) [9]. Since the control of the rate of transcription is principally related to sequences located in the upstream region of the TSS, the fact that a gene could have multiple TSS complicates the interpretation of experiments devoted to expression profiling.

Moreover some of these new discovered TSSs use the promoter of an entirely different gene locus (i.e., there are transcript of different genes sharing the same transcription start site) [9].

Transcript encompassing multiple gene locus

Many known protein coding genes have alternative TSSs that are sometimes more than 100 kilobases upstream of the previously annotated TSS. Thus, some alternative isoforms are transcripts that span multiple gene loci [9].

Gene fusions

Many of the alternative isoforms encompassing multiple gene loci code for the same protein differing only in their 5' untranslated regions (UTRs). Therefore there are some case in which two consecutive genes are transcribed into a single RNA. The translation (after splicing) of such RNAs can lead to a new, fused protein, having parts from both original proteins [9].

Transplicing

There are also cases of ligation of two separate mRNA molecules. Clearly, the classical concept of the gene as “a locus” no longer applies for these gene products whose DNA sequences are widely separated across the genome.

Alternative splicing

The number of annotated alternative isoforms per locus has increased up to (on average) 5.4 transcripts per locus [9].

Dispersed regulation

The regulatory sites for a given gene are not necessarily directly upstream of it and they can, in fact, be located far away on the chromosome, closer to another gene. While the binding of many transcription factors appears to blanket the entire genome, it is not arranged according to simple random expectations and tends to be clumped into regulatory rich "forests" and poor "deserts"[40]. Moreover, it appears that some of regulatory elements may actually themselves be transcribed [9].

1.2.4 An updated definition of gene

According to Gerstein et al. [9] in this thesis I use the following definition of what is a gene: *the gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products.*

Some interesting consequences of this definition can be highlighted.

Definition based on genomic sequence

In the era of whole genome sequencing, the definition of what a genomic sequence is seems to be easier and clearer rather than the definition of what a gene is. Thus the definitions of gene rely on those of genome. It is a change of the historical perspective in which the knowledge about genes came before that about genomes.

Union of sequences

A single gene is a union of sequences and not a single one. The reasons for that is clear if we think about alternative splicing. Moreover now is clear that a single gene can be made starting from different primary transcript (transplicing), or by a very long single primary transcript and encompassing other gene locus.

Encoding functional products

A gene is not itself functional but contains the information to produce a functional product (there is no distintion on the biochemical nature of this product that can be either a protein or an RNA molecule). This definition of gene relies on the definition of function, but it is difficult to establish what is functional and what is not. The simple transcription is not taken as evidence of function.

Potentially overlapping

Different functional products of the same class (protein or RNA) that overlap in their usage of the primary DNA sequence are combined into the same gene. This overlap is done by projecting the sequence of the final product (either amino acid or RNA sequence) down onto the original genomic sequence from which it was derived [9]. For the protein conding genes the functional product is a protein, since potentially overlapping UTRs parts of two mRNAs are not considered in order to decide if they are two isoforms of the same gene or not.

The UTRs are considered regulative regions and regulation is simply too complex to be folded into the definition of a gene, and there is obviously a many-to-many (rather than one-to-one) relationship between regulatory regions and genes. Moreover in procariotic genomes the fact that genes in an operon share an operator and promoter region has traditionally not been considered to imply that their protein products are alternative products of a single gene [9].

Two functional products A and B may refer to the same gene even if they do not share any genomic sequences as blueprint: a third transcript C may overlap both A and B.

An obvious point that should still be stated is that, when looking at genomic products with common sequence segments, mere sequence identity is not enough; the products have to be encoded directly from the same genomic region. Thus, paralogous proteins may share sequence blocks, but DNA sequences coding for them reside in separate locations in the genome, and so they will not constitute one gene.

Coherent set of potentially overlapping products

Two distinct products are united in the same gene if they overlap; "coherent" means that this union is performed separately for final protein and RNA products.

1.3 Gene prediction

One of the main justifications for   achieve genome sequencing project is to identify new genes for which there is only partial or no previous information [41].

Essentially, two different types of methods are currently used to try to locate genes in a genomic sequence.

Extrinsic (homology): historically, the existence of a sufficient similarity with a biologically characterized sequence has been the main means in order to find genes. Similarity-based approaches have often been called extrinsic in opposition to others that try to capture some of the intrinsic properties of a gene.

Intrinsic (ab-initio): using statistic measures (like compositional bias, codon usage, the presence of the functional sites specific to a gene) these methods try to classify a DNA region into types, e.g. coding versus non-coding etc.

Each single type of information useful to determine the presence and the structure of a gene in a DNA sequence is retrieved by an algorithm, called sensor.

Complete gene prediction pipelines combine many sensors (therefore may type of information) to achieve more accurate prediction; in the Ensembl [42] pipeline, for instance the ab initio programs are called upon first. Then, to reduce the high incidence of false positives, the resulting gene predictions are 'fixed' by the incorporation of similarity information. Ensembl keeps only those exons that show sequence similarity to a gene or protein in the vertebrate databases, not necessarily the entire gene prediction and not necessarily from the same species.

1.3.1 Extrinsic (homology) sensors

Extrinsic sensors simply exploit a sufficient similarity between a genomic sequence region and a protein or DNA sequence present in a database in order to determine whether the region is transcribed and/or protein coding. The basic tools for detecting sufficient similarity between sequences are local alignment methods ranging from the optimal Smith–Waterman algorithm to fast heuristic approaches such as FASTA [43] and BLAST [44].

Often when one tries to map a given sequence (taken from a protein or from a RNA molecule) in the genome, beside a region that shares exact sequence similarity (excluding sequencing error, genotipic variability and sequence editing), it is possible to find other regions sharing statistically significant similarity; also this partial information could indicate the presence of a gene.

The obvious weakness of such extrinsic approaches is that nothing will be found if the database does not contain a sufficiently similar sequence. Furthermore, even when a good similarity is found, the limits of the regions of similarity, which should indicate exons, are not always very precise and do not enable an accurate identification of the structure of the complete gene or of the simpler single transcript.

Overall, similarities with three different types of sequences may provide information about exon/intron locations. The first and most widely used are protein sequences that can be found in databases such as SwissProt or PIR. It is estimated that almost 50% of the genes can be identified thanks to a sufficient similarity score with a homologous protein sequence. However, even when a good hit is obtained, a complete exact identification of the gene structure can still remain difficult because homologous proteins may not share all of their domains. Furthermore, UTRs cannot be delimited in this way.

The second type of sequences are transcripts, sequenced as cDNAs (a cDNA is a DNA copy of a mRNA) either in the classical way for targeted individual genes with high coverage sequencing of the complete clone or as expressed sequence tags (ESTs), which are one shot sequences from a whole cDNA library. ESTs and "classical" cDNAs are the most relevant information to establish the structure of a gene, especially if they come from the same source as the genome to be annotated. ESTs provide information that enable the identification of (partial) exons, either coding or non-coding, and give unbiased hints on alternative splicing. However, ESTs give only local and limited information on the gene structure as they only reflect a partial mRNA. Furthermore, the correct attribution of EST sequences to an individual member in a gene family is not a trivial task [45].

Finally, under the assumption that coding sequences are more conserved than non-coding ones, similarity with genomic DNA can also be a valuable source of information on exon/intron location. Two approaches are possible: intra-genomic comparisons can provide data for multigenic families, apparently representing a large percentage of the existing genes (e.g. 80% for Arabidopsis) [45]; inter-genomic (cross-species) comparisons can allow the identification of orthologous genes, even without any preliminary knowledge of them. Nevertheless, the similarity may not cover entire coding exons but be limited to the most conserved part of them. Alternatively, it may sometimes extend to introns and/or to the UTRs and promoter elements. This will be the case when genomes are evolutionarily close or when genome duplications are recent events. In both cases, exactly discriminating between coding and non-coding sequences is not an obvious task.

In all cases, an important strength of similarity-based approaches is that predictions rely on accumulated pre-existing biological data . They should thus produce biologically relevant predictions (even if only partial) [45].

1.3.2 Intrinsic (ab–initio) methods

Although an ab-initio gene predictor do not uses external information other that the raw (partial) genomic sequence, a set of well known gene is required, and the model inside the predictor summarize and generalize this previous knowledge. Therefore – although these methods are considered as "intrinsic" – the fact that the models are built from known sequences will inherently limit the applicability of the methods to sequences that, globally, behave in the same way as the learning set [41].

In the following sections I make a brief excursus on the ab-initio gene predictors and then I describe the gene model that is the core of the most used nowadays: GENSCAN.

Content sensors

Originally, intrinsic content sensors were defined for prokaryotic genomes. In such genomes, only two types of regions are usually considered: the regions that code for a protein and will be translated, and intergenic regions. Since coding regions will be translated, they are characterized by the fact that three successive bases in the correct frame define a codon which, using the genetic code rules, will be translated into a specific amino acid in the final protein.

In prokaryotic sequences, genes define long uninterrupted coding regions that must not contain stop codons. Therefore, the simplest approach for finding potential coding sequences is to look for sufficiently long open reading frames (ORFs), defined as sequences not containing stops, i.e. as sequences between a start and a stop codon. In eukaryotic sequences, however, the translated regions may be very short and the absence of stop codons becomes meaningless [46].

Several other measures have therefore been defined that try to more finely characterize the fact that a sequence is "coding" for a protein: nucleotide composition and especially (G+C) content (introns being more A/T-rich than exons, especially in plants) [45], codon composition, hexamer frequency, base occurrence periodicity, etc. Among the large variety of coding measures that have been tested, hexamer usage (i.e. usage of 6 nt long words) was shown to be the most discriminative variable between coding and non-coding sequences [46]. This characteristic has been widely exploited by a large number of algorithms through different methods.

Thus, hexamer frequency is one of the main variables used in SORFIND [47], Genview2 [48], the quadratic discriminant analysis approach of MZEF [49] and the neural network procedure of GeneParser [50]. This last program combines the use of hexamer frequency with local compositional complexity measures estimated on octanucleotide statistics. Such statistics are also efficiently used, among other variables, in the linear discriminant analysis of GeneFinder [51].

More generally, the kmer composition of coding sequences is the basis of the now ubiquitous so-called "three-periodic Markov model" introduced in the pioneering algorithm GeneMark [52]. Very briefly, a Markov model is a stochastic model which assumes that the probability of appearance of a given base (A, T, G or C) at a given position depends only on the k previous nucleotides (k is called the order of the Markov model). Such a model is defined by the conditional probabilities P(X|k previous nucleotides), where X = A, T, G or C. In order to build a Markov model, a learning set of sequences on which these probabilities will be estimated is required. Given a sequence and a Markov model, one can then very simply compute the probability that this sequence has been generated according to this model, i.e. the likelihood of the sequence, given the model [45].

The simplest Markov models are homogeneous zero order Markov models which assume that each base occurs independently with a given frequency. Such simple models are often used for non-coding regions, although it is now frequent to use higher order models to represent introns and intergenic regions as, for instance, in GeneMark, Genscan [53] and EuGène [54]. The more complex three-periodic Markov models have been introduced to characterize coding sequences. Coding regions are defined by three Markov models, one for each position inside a codon [45].

The larger the order of a Markov model, the finer it can characterize dependencies between adjacent nucleotides. However, a model of order k requires a very large number of coding sequences to be reliably estimated. Therefore, most existing gene prediction programs, such as GeneMark and Genscan, usually rely on a three-periodic Markov model of order five (thus exploiting hexamer composition) or less to characterize coding sequences. To cope with these limitations, interpolated Markov models (IMMs) have been introduced in the prokaryotic gene finder Glimmer [55]. For each conditional probability, an IMM combines statistics from several Markov models, from order zero to a given order k (typically k = 8), according to the information available. These IMMs are now also used in GlimmerM, a version dedicated to eukaryotes, and in EuGène. The new version of Glimmer introduces yet another sophistication of Markov models called interpolated context models, which can capture dependencies among 12 adjacent nucleotides [55].

Another type of refinement is often needed in eukaryotic genomes. It consists of estimating several gene models according to the G+C content of the genomic sequence. This is done by Genscan and GeneMark.hmm [56]. Indeed, it was shown that differences in gene structure and gene density along some genomes are closely related to their "isochore" organization [57, 58, 59].

In general, most currently existing programs use two types of content sensors: one for coding sequences and one for non-coding sequences, i.e. introns, UTRs and intergenic regions. A few software refine this by using a different model for the different types of non-coding regions (e.g. one model for introns, one for intergenic regions and an optional specific 3'- and 5'-UTR model in EuGène) [45].

Signal sensors

The basic and natural approach to finding a signal that may represent the presence of a functional site is to search for a match with a consensus sequence (with possible variations allowed), the consensus being determined from a multiple alignment of functionally related documented sequences.More sophisticated approachs are based on models of various type (HMMs for instance) but as for the previous "intrinsic" content sensors, the fact that the models are built from sets of known functional sequences inherently limits the sensors to canonical signals [41].

Simple match with a consensus sequence is used, for instance, for splice sites prediction in SPLICEVIEW and SplicePredictor [60] . A more flexible representation of signals is offered by the so-called positional weight matrices (PWMs[3]), which indicate the probability that a given base appears at each position of the signal. Equivalently, one can say that a PWM is defined by one classical zero order Markov model per position, which is called an inhomogeneous zero order Markov model. The PWM weights can also be optimized by a neural network method [61].

In order to capture possible dependencies between adjacent positions of a signal, one may use higher order Markov models. The so-called weight array model[4] (WAM) is essentially an inhomogeneous higher order Markov model. It was first proposed by Zhang and Marr [62] and later used by Salzberg [63].

Of course, higher-order WAM models capturing second-order (triplet) or third-order (tetranucleotide) dependencies in signal sequences could be used in principle, but typically there is insufficient data available to estimate the increased number of parameters in such models [53].

These methods assume a fixed length signal. Hidden Markov models (HMMs) further allow for insertions and deletions [64]. In order to capture the most significant dependencies between adjacent as well as non-adjacent positions, Burge [53] proposed another model for donor sites called the maximal dependence decomposition (MDD) method [5].

It was also shown that combining sequence-based metrics for splice sites (WAM) with

[3] In a PWM the frequency $p_j^{(i)}$ of each nucleotide j at each position i of a signal of length n is derived from a collection of aligned signal sequences and the product $P\{X\} = \prod_{i=1}^{n} p_{xi}^{(i)}$ is used to estimate the probability of generating a particular sequence, $X = x_1, x_2, \ldots, x_n$. [53]

[4] In a weight array model the probability of generating a particular sequence is:

$$P\{X\} = p_{x_1}^{(1)} \prod_{i=2}^{n} p_{x_{i-1},xi}^{(i-1,i)}$$

where $p_{j,k}^{(i-1,i)}$ is the conditional probability of generating nucleotide x_k at position i, given nucleotide x_j at position $i-1$ (which is estimated from the corresponding conditional frequency in the set of aligned signal sequences) [53].

[5] The goal of the MDD procedure is to generate, from an aligned set of signal sequences of moderate to large size (i.e. at least several hundred or more sequences), a model which captures the most significant dependencies between positions (allowing for non-adjacent as well as adjacent dependencies), essentially by replacing unconditional PWM probabilities by appropriate conditional probabilities provided that sufficient data is available to do so reliably. Given a data set D consisting of N aligned sequences of length k, the first step is to assign a consensus nucleotide or nucleotides at each position. For each pair of distinct positions $\{i, j\}$, a 2 by 4 contingency table was constructed for the indicator variable C_i (1 if the nucleotide at position i matches the consensus, 0 otherwise) versus the variable X_j identifying the nucleotide at position j, and the value of the χ^2 statistic for each such table was calculated.

If no significant dependencies are detected, then a simple PWM should be sufficient. If significant dependencies are detected, but they are exclusively or predominantly between adjacent positions, then a WAM model may be appropriate. If, however, there are strong dependencies between non-adjacent as well as adjacent positions, then we proceed as follows.

1. Calculate, for each position i, the sum $S_i = \sum_{j \neq i} \chi^2(C_i, X_j)$, which is a measure of the amount of dependence between the variable C_i and the nucleotides at the remaining positions of the site.
2. Choose the value i_1 such that S_{i_1} is maximal and partition D into two subsets: D_{i_1} all sequences which have the consensus nucleotide(s) at position i_1; and $D_{\overline{i_1}}$ all sequences which do not.

Now repeat steps (1) and (2) on each of the subsets, D_{i_1} and $D_{\overline{i_1}}$ and on subsets thereof, and so on, yielding a binary subdivision "tree" with (at most) $k-1$ levels. This process of subdivision is carried out successively on each branch of the tree until one of the following three conditions occurs:

1. the $(k-1)$th level of the tree is reached (so that no further subdivision is possible);
2. no significant dependencies between positions in a subset are detected (so that further subdivision is not indicated);
3. the number of sequences remaining in a subset becomes so small that reliable WMM frequencies could not be determined after further subdivision.

Finally, separate WMM models are derived for each subset of the tree, and these are combined to form a composite model.

secondary structure metrics could lead to valuable improvements in splice site prediction [65]. However, when using splice site prediction programs, one ends up with a list of potential splice sites, from which various gene structures may be built. The main purpose of such programs is not to find the gene structure but to try to find the correct exon boundaries [41]. They are thus very useful in addition to a content based sensor in order to refine an existing gene structure. These programs can also provide insights into possible alternative splicing .

1.3.3 A model of gene structure

Genscan is a widely used ab-initio gene prediction program based upon a model of gene structure encompassing many type of content and signal sensors. The information of the initial training set of well known genes is formulated as a probabilistic mathematical model, namely an explicit state duration Hidden Markov Model (HMM)[64].

The model is schematizzed in figure 1.3: each circle or diamond represents a functional unit (state) of a gene or genomic region [53]:

- N, intergenic region;
- P, promoter;
- F, 5' untranslated region (extending from the start of transcription up to the translation initiation signal);
- E_{sngl}, single-exon (intronless) gene (translation start → stop codon);
- E_{init}, initial exon (translation start → donor splice site);
- E_k $(0 \leq k \leq 2)$, phase k internal exon (acceptor splice site → donor splice site);
- E_{term}, terminal exon (acceptor splice site → stop codon);
- T, 3' untranslated region (extending from just after the stop codon to the polyadenylation signal);
- A, polyadenylation signal;
- I_k $(0 \leq k \leq 2)$, phase k intron.

For convenience, translation initiation/termination signals and splice sites are included as subcomponents of the associated exon state and intron states are considered to extend from just after a donor splice site to just before the branch point/acceptor splice site (i.e. the figure represents explicitly only content based sensors and theirs combinations). The upper half of the Figure corresponds to the states (designated with a superscript $+$) of a gene on the forward strand, while the lower half (designated with superscript -) corresponds to a gene on the opposite (complementary) strand. For example, proceeding in the 5' to 3' direction on the (arbitrarily chosen) forward strand, the components of an E_k^+ (forward-strand internal exon) state will be encountered in the order [53]:

1. acceptor site,
2. coding region,
3. donor site,

while the components of an E_k^- (reverse-strand internal exon) state will be encountered in the order:

1. inverted complement of donor site,
2. inverted complement of coding region,
3. inverted complement of acceptor site.

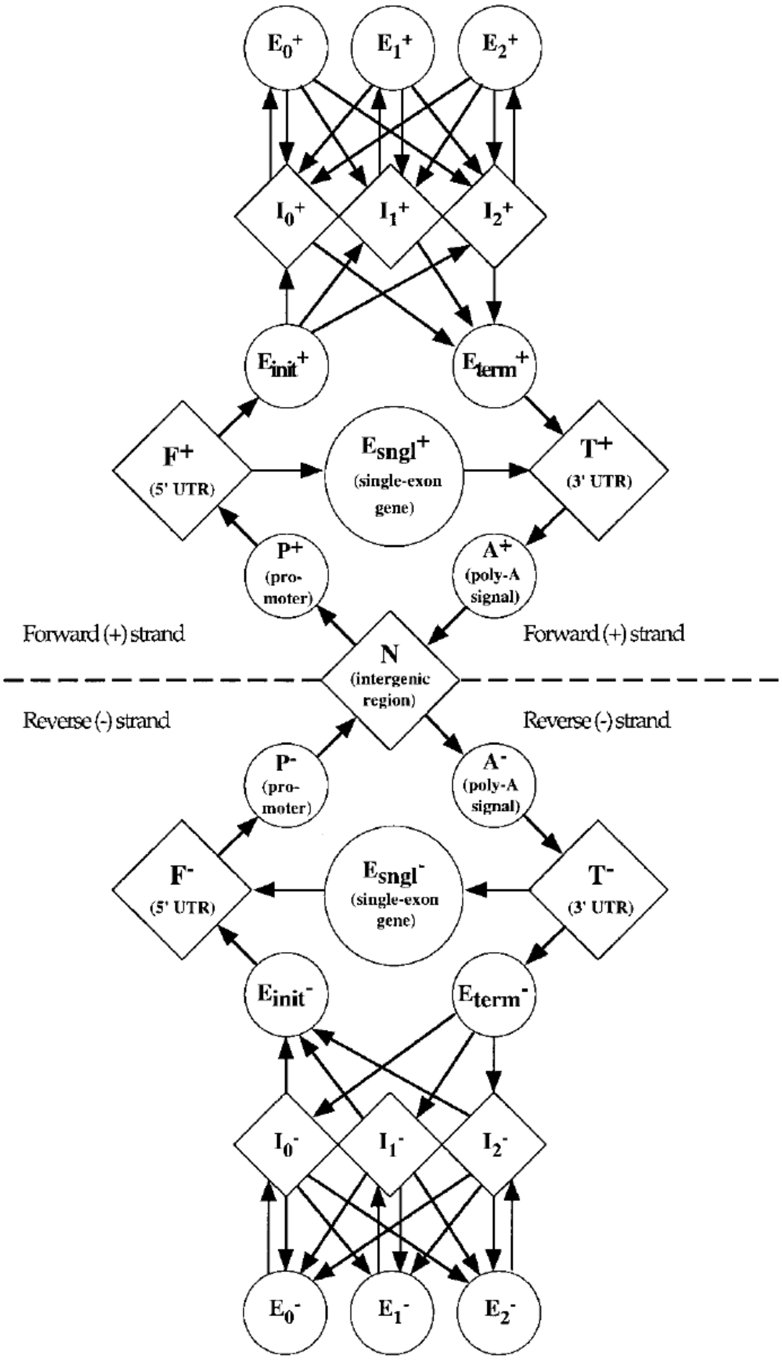


Figure 1.3: A representation of GENSCAN inner model. Modified from [53]

Only the intergenic state N is not divided according to strand.

The model is though of as generating a "parse" φ, consisting of an ordered set of states,

$$\vec{q} = \{q_1, q_2, \ldots, q_n\}$$

with an associated set of lengths (durations),

$$\vec{d} = \{d_1, d_2, \ldots, d_n\}$$

which, using probabilistic models of each of the state types, generates a DNA sequence S of length

$$L = \sum_{i=1}^{n} d_i$$

The generation of a parse corresponding to a (pre-defined) sequence length L is as follows

1. An initial state q_1 is chosen according to an initial distribution on the states, $\vec{\pi}$ where $\pi_i = P\{q_1 = Q^{(i)}\}$ where $Q^{(j)}(j = 1, 2, \ldots, 27)$ is an indexing of the state types (see figure 1.3).

2. A length (state duration), d_1, corresponding to the state q_1 is generated conditional on the value of $q_1 = Q^{(i)}$ from the length distribution $f_{Q^{(i)}}$.

3. A sequence segment s_1 of length d_1 is generated, conditional on d_1 and q_1, according to an appropriate sequence generating model for state type q_1.

4. The subsequent state q_2 is generated, conditional on the value of q_1, from the (first-order Markov) state transition matrix T, i.e.

$$T_{i,j} = P\{q_k + 1 = Q^{(j)} | q_k = Q^{(i)}\}$$

.

This process is repeated until the sum, $\sum_{i=1}^{n} d + i$, of the state durations first equals or exceeds the length L at which point the last state duration d_n is appropriately truncated, the final stretch of sequence is generated, and the process stops: the sequence generated is simply the concatenation of the sequence segments, $S = s_1 s_2 \ldots s_n$. Note that the sequence of states generated is not restricted to correspond to a single gene, but could represent a partial gene, several genes, or no genes at all. The model thus has four main components: a vector of initial probabilities $\vec{\pi}$, a matrix of state transition probabilities T, a set of length distributions f, and a set of sequence generating models P. Assuming for the moment that these four components have been specified, the model can be used for prediction in the following way [53].

Using the model for gene prediction

For a fixed sequence length L, consider the space $\Omega = \phi_L \times \mathcal{S}_L$ where ϕ_L is the set of (all possible) parses of length L and $\mathcal{S}_L$ is the set of (all possible) DNA sequences of length L. The model M can then be thought of as a probability measure on this space, i.e. a function which assigns a probability density to each parse/sequence pair. Thus, for a particular sequence $S \in \phi_L$, we can calculate the conditional probability of a particular parse $\varphi_i \in \phi_L$ (under the probability measure induced by M) using Bayes' Rule as:

$$P\{\varphi_i | S\} = \frac{P\{\varphi_i, S\}}{P\{S\}} = \frac{P\{\varphi_i, S\}}{\sum_{\varphi_j \in \phi_L} P\{\varphi_j, S\}}$$

The essential idea is that a precise probabilistic model of what a gene/genomic sequence looks like is specified in advance and then, given a sequence, one determines which of the vast number of possible gene structures (involving any valid combination of states/lengths) has highest likelihood given the sequence [53].

Given a sequence S of length L, the joint probability, $P\{\varphi_i, S\}$, of generating the parse φ_i and the sequence S is given by:

$$P\{\varphi_i, S\} = \pi_{q_1} f_{q_1}(d_1) P\{s_i|q_1, d_1\} \times \prod_{k=2}^{n} T_{q_{k-1},q_k}(d_k) P\{s_k|q_k, d_k\}$$

where the states of φ_i are $q_1, q_2, \ldots, q_n$ with associated state lengths $d_1, d_2, \ldots, d_n$, which break the sequence into segments $s_1, s_2, \ldots, s_n$. Here $P\{s_k|q_k, d_k\}$ is the probability of generating the sequence segment s_k under the appropriate sequence generating model for a type-q_k state of length d_k.

A recursive algorithm of the sort devised by Viterbi [66] may then be used to calculate φ_{opt}, the parse with maximal joint probability (under M), which gives the predicted gene or set of genes in the sequence. .

Assess the initial probability

Since genecan attempt to model a randomly chosen block of contiguous human genomic DNA the initial probability of each state should be chosen proportionally to its estimated frequency in bulk human genomic DNA. However, even this is not trivial since gene density and certain aspects of gene structure are known to vary quite dramatically in regions of differing C + G% content (so-called "isochores") of the human genome [67].

Assess the state transitions probabilities

The (biologically permissible) state transitions are shown as arrows in Figure 1.3. Certain transitions are obligatory (e.g. $P^+ \rightarrow F^+$) and hence are assigned probability one; all others are assigned (maximum likelihood) values equal to the observed state transition frequency in the learning set $\mathcal{L}$ for the appropriate C + G compositional group. Overall, transition frequencies varied to a lesser degree between groups than did initial probabilities. There was a trend for A + T-rich genes to have fewer introns, leading to slightly different estimates for the $I_j^+ \rightarrow E_{\text{term}}^+$ probabilities [53].

Assess the length distributions

For the length distribution genescan uses separate empirically derived length distribution functions for initial, internal, and terminal exons and for single-exon genes. Substantial differences in exon length distributions were not observed between the C + G compositional groups for the coding exons and the UTR while intron and intergenic lengths are modeled as geometric distributions with parameter q estimated for each C + G group separately.

Assess signals sensors

Polyadenylation signals are modeled as a 6 bp PWM (consensus: AATAAA). A 12 bp PWM model, beginning 6 bp prior to the initiation codon, is used for the translation initiation signal.

For the translation termination signal, one of the three stop codons is generated (according to its observed frequency in $\mathcal{L}$) and the next three nucleotides are generated according to a PWM. All these PWM are computed form $\mathcal{L}$.

Bases -20 to +3 relative to the intron/exon junction, encompassing the pyrimidine-rich region and the acceptor splice site itself, are modeled by a first-order WAM model.

In the donor splice there are highly significant dependencies between non-adjacent as well as adjacent positions, which are not adequately accounted for by WAM and which likely relate to details of donor splice site recognition by U1 snRNP and possibly other factors. The consensus region of the donor splice site comprises the last 3 bp of the exon (positions -3 to -1) and the first 6 bp of the succeeding intron (positions 1 through 6), with the almost invariant GT dinucleotide occurring at positions 1,2: consensus nucleotides are GUCCAUCCA. For this signal GENSCAN use a MDD model [53].

Assess conding exons content sensor

Coding portions of exons are modeled using an inhomogeneous 3-periodic fifth-order Markov model. In this approach, separate fifth-order Markov transition matrices are determined for hexamers ending at each of the three codon positions, denoted c_1, c_2, c_3, respectively; exons are modeled using the matrices c_1, c_2, c_3 in succession to generate each codon. These transition probabilities were derived from the training set $\mathcal{L}$. It as been shown [46] that frame-specific hexamer measures are generally the most accurate compositional discriminator of coding versus noncoding regions. A + T-rich genes are often not well predicted using such bulk hexamer-derived parameters. Accordingly, a separate set of fifth-order Markov transition matrices was derived for region of C + G composition below 43% [53].

In the model, the disruption of coding regions by introns in multi-exon genes is dealt with by keeping track of intron/exon phase, ensuring that a consistent reading frame is maintained throughout a gene.

1.4 Pseudogenes

Pseudogenes were originally defined as DNA sequences structurally similar to functional protein coding genes but containing important defects, which make them unable to produce functional proteins [68]. Such defects include, for example, the loss of the start codon, the presence of additional stop signals and the lack or abnormality of flanking regulatory regions. Nevertheless, this definition has to be revised in the light of data showing the possibility that pseudogenes can acquire novel functions during evolution [69] as well as the presence of many non-coding genes.

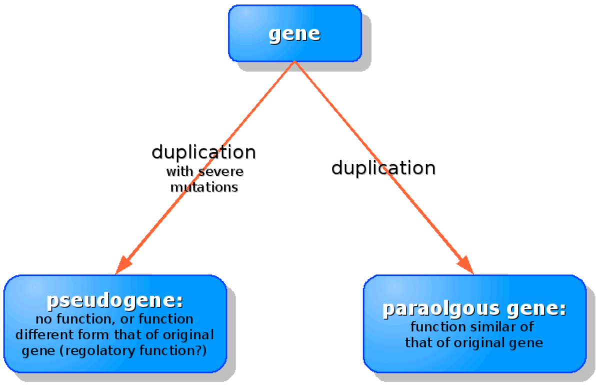


Figure 1.4: Pseudogenes and paralogous genes arise both from a duplication of an ancestor gene.

An updated definition is the following: a pseudogene is a genomic sequence that shares high homology with a gene but with important differences in its sequence or flanking regions. Important differences means that (if we think that the pseudogene came from a duplication of the gene) it underwent mutations that would make unfunctional the pseudogene if its supposed function was the same of the homologous gene. Therefore a genomic region, to be a pseudogene, must share sequence similarity with a gene and must have no function or a different function form the one carried by the homologous gene (i.e. must not be a paralogous gene).

If a pseudogenes has no function (as most or all pseudogenes, see section 1.4.4) it may continue to drift until it is either deleted or becomes unrecognizable as a genetic copy.

1.4.1 Non-processed pseudogenes

Non-processed pseudogenes are usually found on the same chromosome inside clusters of similar functional sequences; they may possess introns and flanking regulatory sequences like the functional gene. They usually originate from a gene duplication mechanism producing an extra copy of the gene which, being unnecessary, can accumulate mutations without damaging the organism, but they can also be generated by unequal crossing-over mechanisms. Premature stop codons, frameshift mutations, disablement of regulatory regions and alterations in splice sites are the most obvious characteristics of pseudogenes [69].

1.4.2 Processed pseudogenes

Processed pseudogenes are caused by a retrotransposition process in three stages. The first stage consists of an RNA synthesis starting from the DNA template. In the second stage, this primary transcript is deprived of introns, producing a mature messenger RNA (mRNA). At the last stage, the mRNA acts as a template in the reverse transcriptase process, producing a double-helix DNA sequence which is inserted in another chromosome [69].

The structural characteristics of processed pseudogenes are:

- lack of upstream regulatory region,
- absence of introns (some introns may be retained due to incomplete splicing),
- presence of 3’ end with a poly A tail,
- flanking direct repeats,

- copy possibly incomplete (not always the entire mRNA is copied).

Most of the known processed pseudogenes produced by a retrotransposition process lose their functionality as a consequence of defects in the mechanism generating them. In fact, reverse transcription is a process producing errors, and a lot of changes between the template RNA and the complementary DNA (cDNA) can be accumulated. The ENCODE processed and nonprocessed pseudogenes [33] share mean sequence identities of 67.6% (±14%) and 61.8% (±18%) with their parent proteins in alignment coverage of 82.4% (±26%) and 69.4% (±33%), respectively. In addition, 83.2% of processed and 79% of nonprocessed pseudogenes display disablements (defined as nonsense or frameshift mutations) in their putative ORFs, with average disablements of 6.2 per processed pseudogene and 2.4 per nonprocessed pseudogene. Overall, such disablements were located uniformly across the hypothetical coding regions of pseudogenes. The differences in sequence identity and disablements between processed and nonprocessed pseudogenes are significant ($P < 0.001$, Wilcoxon rank-sum test), appearing to suggest that the sequences giving rise to processed pseudogenes lose coding potential more quickly than those for nonprocessed pseudogenes [33].

Moreover, unless the processed gene is transcribed by RNA polymerase III, it should not contain the promoter (which usually lies in non-transcribed regions and) and so is quite likely to be inactive even though its coding region is intact.

Finally, a processed gene can be inserted in a genomic localization inappropriate for its expression which can also be a different chromosome compared with its functional counterpart. For this reason, a processed gene is 'dead on arrival' in most cases [69].

Due to the ubiquity of reverse transcription (see section 1.5), mammalian genomes are literally bombarded by copies of retrotranscribed sequences, and most of these copies become non-functional as soon as they integrate in the genome. Moreover, these sequences cannot be easily repaired through the gene conversion process [70], because they are mostly placed at long chromosomal distances from the parent functional gene.

As soon as a retropseudogene settles within a chromosome, it undergoes two different evolutionary processes [71]. The first process involves a rapid accumulation of point mutations which can hide the similarity between the pseudogene sequence and its functional homologue, which evolves much more slowly. The processed pseudogene nucleotide composition will tend to resemble more and more the surrounding non-functional region, enabling the pseudogene to blend with it. This process is called 'compositional assimilation' [69].

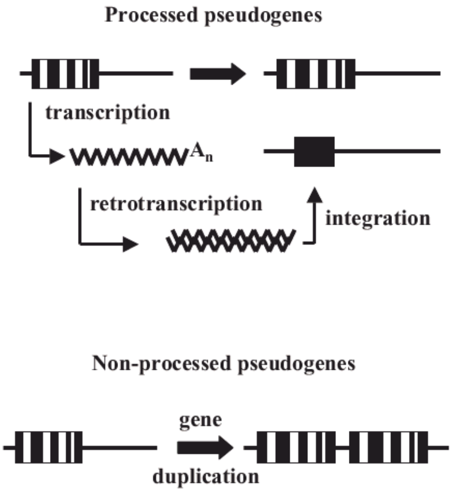


Figure 1.5: Pseudogenes origin. Genes are represented by black (exons) and white (introns) blocks. Single and double zigzag lines represent mRNA and cDNA produced by retrotranscription, respectively. Modified from [69].

The second evolutionary process involves the reduction of pseudogene size compared with the functional gene. This shrinkage is caused by an excess of deletions over insertions. It has been estimated that a processed pseudogene loses about one-half of its DNA in nearly 400 million years. This process is so slow that the human genome, for instance, still contains a large quantity of pseudogene DNA related to very distant ancestors [69]. Obviously, these ancient pseudogenes have often lost almost all their similarity with the functional genes. The shrinkage is too slow a process to counterbalance the increase in genome dimensions which results from the continuous retrotranspositions. So, the restriction in pseudogene number in the genome is probably due to other factors, such as natural selection [69].

Since processed pseudogenes are well characterized in term of sequence properties and structure it would be possible to define a processed pseudogene on the basis of these characteristics, even if the pseudogene as a paralogous function with respect to homologous gene.

As a matter of fact this is the operative definition used in computational pseudogene finding.

1.4.3 Mitochondrial pseudogenes

A few decades ago, the presence of sequences of many animal species having significant homology to mitochondrial DNA inside the nucleus was ascertained. These nuclear insertions of mitochondrial DNA are called pseudogenes because, unlike their homologous counterparts, they are not transcribed or translated into functional proteins, owing to the different mitochondrial genetic code [72].

The integration process of the mitochondrial fragments in the nucleus is probably very ancient; indeed, it is thought to have started soon after the settling of the first endosymbiont as an organelle. In fact, at least as far as the animal line is concerned, there has been a progressive thinning of the mitochondrial genome as a consequence of the transfer of genes coding for mitochondrial components to the nucleus. Thus, the unsuccessful transfers could have given rise to pseudogenes. There are essentially two mechanisms which could explain the integration of these mitochondrial fragments in the nucleus: direct DNA transfer [73] and RNA-mediated transfer [74]. Most of the experimental data available support the hypothesis that transfer is by DNA, although an origin of mitochondrial pseudogenes from RNA cannot be excluded [75].

From the available literature, it seems that mitochondrial pseudogenes are not equally distributed in all species – they are abundant in mammals and birds, but seem to be almost completely absent in fish [76], very few are found in Caenorhabditis elegans (two pseudogenes) and Drosophila melanogaster (three pseudogenes) compared with Homo sapiens (354 pseudogenes) [69].

Computational pseudogene detection

The prevalence of pseudogenes in mammalian genomes is problematic for gene annotation [77] and can introduce artifacts to molecular experiments targeted at functional genes. The correct identification of pseudogenes, therefore, is critical for obtaining a comprehensive and accurate catalog of structural and functional elements of the human genome [33].

Since many pseudogenes are expected to be non-functional and non-transcribed the pseudogenes identification depends almost exclusively on computational analysis.

Several computational algorithms have been described for annotating human pseudogenes. Although these methods often present similar estimates for the number of pseudogenes in the human genome, they can produce rather distinct pseudogene sets. For instance, the Gernstein group at the Yale University (as part of the ENCODE project, see section 1.2.2), examined five methods for detecting pseudogenes [33]. These methods, which have been developed independently, are:

1. the GIS-PET method, from the Genome Institute of Singapore [78];
2. the HAVANA method of manual pseudogene annotation, by the Human And Vertebrate Analysis aNd Annotation team (HAVANA) at the Wellcome Trust Sanger Institute as part of the GENCODE collaboration [79];
3. PseudoPipe [80], from the pseudogene research group at Yale University;
4. pseudoFinder, from the University of California Santa Cruz (UCSC);
5. retroFinder, also from UCSC but focused specifically on processed pseudogenes [81].

A simple union of these five sets yielded 252 (in the 1% portion of genome studied in ENCODE pilot project [37]) nonoverlapping pseudogenes, of which only 45 (17.9%) were identified by all methods (see Figure 1.6).

Almost all pseudogenes prediction methods detect pseudogenes by their sequence similarity to at least one entry in a collection of query sequences representing known human genes (referred to as the parent genes).

Each gene sequence is aligned back to the genome and all the aligned genomic regions outside the known gene position are retained. To decide that a region of this set host a

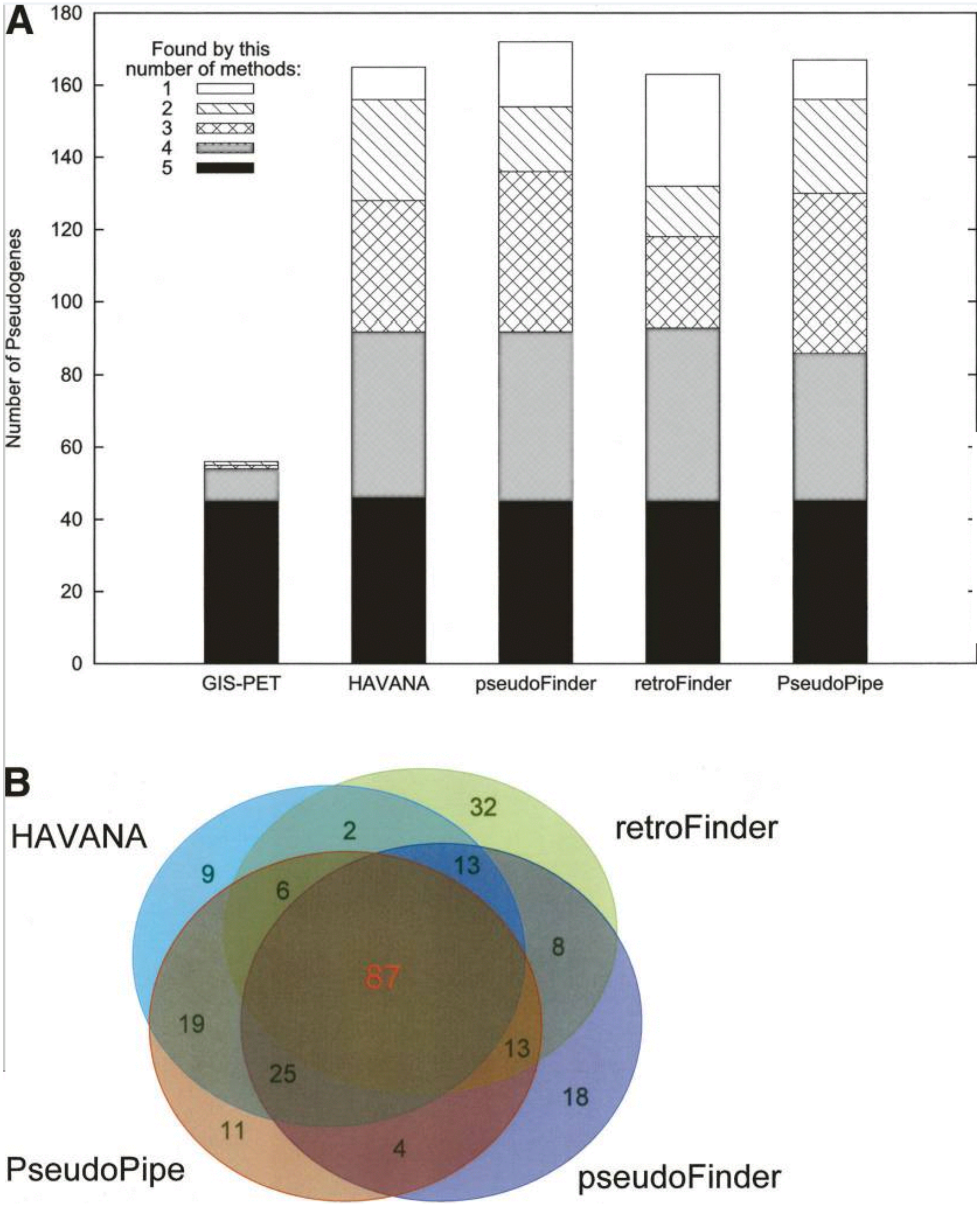


Figure 1.6: **(A)** pseudogenes annotated by a method were binned into groups based on the number of methods that recognized them as pseudogenes. In this scheme, method-specific pseudogenes were labelled as (found by) “1” method. **(B)** A four-way comparison of pseudogenes identified by HAVANA, PseudoPipe, retroFinder, and pseudoFinder. Note: one pseudogene could overlap more than one pseudogene from other method(s). Modified form [9].

processed pseudogene some typical characteristics must be present (severe mutations, lack of introns, poly A tail, flanking repeat, ...).

In our work (see chapter 2) we look for gene-pseudogenes couples in a given genome in a completely different way, that does not rely upon a set of known genes, thus allowing the finding of unknown pseudogenes and unknown genes.

1.4.4 Are there functional pseudogenes?

The common assumption is that pseudogenes are nonfunctional and thus evolve neutrally. As such, they are frequently considered as "genomic fossils" and are often used for calibrating parameters of various models in molecular evolution, such as estimates of neutral mutation rates [82]. However, a few pseudogenes have been indicated to have potential biological roles [83, 84, 85, 86, 87]. Whether these are anecdotal cases or pseudogenes do play cellular roles is still a matter of debate at this point [33].

Pseudogene transcription

Several studies have shown that a good fraction (5%) of the human pseudogenes were potentially transcribed [88, 89, 90, 91]. Within the ENCODE pseudogenes the $10\% \div 20\%$ are transcribed in at least one of 12 human tissues [33].

Some transcribed pseudogenes might possess their own promoters while in a few cases pseudogene transcription could have been initiated from the promoters of neighbouring genes or LINE elements [89, 92, 93].

Non-processed Pseudogenes

Certainly, recent non-processed pseudogenes can be transcriptionally active if the function of their promoters has not been lost entirely.

An example of this phenomenon is the human α-globin cluster of genes on chromosome 16 that has arisen by gene duplication and divergence. This cluster includes $\xi 2$, which is expressed in the embryonic yolk sac, and its non-processed pseudogene $\psi\xi 1$. The latter has a non-functional promoter but, in some individuals, its gene conversion [70] by $\xi 2$ has resulted in restoration of a functional promoter and the generation of $\xi 1$ from $\psi\xi 1$ [94]. Another reported case is that of bovine seminal ribonuclease, which has lain dormant for about 20 million years and which then appears to have been resurrected to form a functioning gene – probably via a gene conversion event.[95] This shows that, in some cases, 'resurrection' of duplicated pseudogenes, or of parts of them, can occur to result in an expressed protein [96] [69].

Nevertheless the evidences of non-processed pseudogenes function are few and unclear [97].

Processed pseudogenes

During the retrotransposition process processed pseudogenes lack their upstream regulatory region. Even if they do not possess all the transcriptional control regions present in the functional gene, they can use other transcriptional elements. For example, pseudogene transcription can be directed by a promoter apparently near a non-correlated sequence.21 [69] Indeed reverse transcription polymerase chain reaction experiments have shown, through transcript identification, that some pseudogenes can be transcribed and, in some cases, they can have a different role from that of the original gene. [69]

The Makorin example

Makorin1-p1 is the most famous putative functional pseudogene, even though the authors disagree upon the observations and the models about this pseudogene [98].

Hirotsune et al. [99] had been analysing mice in which copies of a Drosophila gene called Sex-lethal were randomly inserted in the mouse genome. In the course of their studies, they encountered one mouse line that died shortly after birth from multi-organ failure. As this occurred in only one mouse line out of many, the results could not be explained by

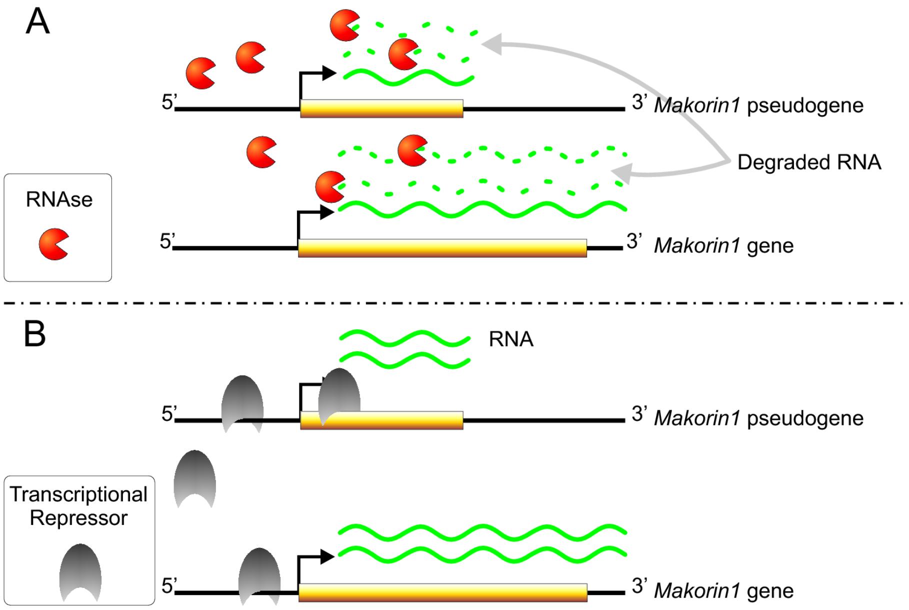


Figure 1.7: Plausible mechanism of gene-pseudogene interaction. **A**) A RNA-mediated mechanism: here, messenger RNA copies of the pseudogene and the gene compete for a destabilizing protein that binds a crucial 700-nucleotide region near the beginning of the mRNAs. This destabilizing protein might be an RNA-digesting enzyme (RNAse). **B**) A DNA-mediated mechanism: here, regulatory elements of the pseudogene and gene, located in the same region as above, compete for transcriptional repressors.

aberrant Sex-lethal expression. Instead, the authors attributed their finding to a disruption of the particular stretch of genomic information into which Sex-lethal had inserted in this case. The disrupted genomic region host Makorin1-p1 – a pseudogene copy of the functional Makorin1 gene[6].

Normally, Makorin1 mRNA is expressed throughout the organism. But Hirotsune et al. found that when Makorin1-p1 pseudogene was disrupted, the expression of Makorin1 was markedly reduced in embryos and during birth and weaning. This implies that the pseudogene is normally required for the high-level expression of Makorin1. Interestingly, of the two forms of Makorin1 mRNA, only the smaller 1.7-kilobase transcript was downregulated – the larger 2.9-kilobase copy was unaffected. The long and short forms are identical except in a region at the 3' UTR.

The authors found that the 700-nucleotide 5' region of Makorin1-p1 not only was required but was also sufficient for regulation in experiments in vitro. These experiments also suggested that the pseudogene acts sequence-specifically, affecting only those genes that show some sequence similarity to itself [99].

To explain this observations it as been proposed a model in which the first 700 nucleotides of the Makorin1 mRNA contain a recognition site for a destabilization factor. Because this 700-nucleotide domain is shared by the Makorin1-p1 mRNA, the expression of the pseudogene would provide a means of titrating out the destabilizing factor through direct competition. In this model, the longer Makorin1 mRNA is unaffected because its 3 untranslated region protects it from degradation (see figure 1.7 A) [100]. Another, more probable, mechanism could be involved (see figure 1.7 B). This is suggested by the fact that mRNA stability is usually controlled by elements in the 3'UTR regions – rather than at the 5' end, where the key 700-nucleotide region of Makorin1 is found – and by the fact that some authors claim that Makorin1-p1 is not expressed [98]. The alternative mechanism would involve the pseudogene DNA locus directly. Perhaps the 700-nucleotide region in the gene and pseudogene contains transcription factor binding sites that, on binding certain proteins, repress transcription. In this model the repressor proteins would be limited in availability,

[6] Makorin1 is an ancient gene that has been evolutionarily conserved from nematode worms to fruitflies and mammals, and encodes a putative RNA-binding protein. It is the prototype of a large family of Makorin genes and pseudogenes, and is located on mouse chromosome 6.

so that Makorin1-p1 would compete for repressor binding [100].

1.5 Transposons

Mobile elements are DNA sequences that have the ability to integrate into the genome at a new site within their cell of origin. The mechanism by which many of these elements move is well known, but for others, such as mammalian retrotransposons, there is still much to learn [101].

There are mobile, or transposable, elements in the genomes of all plants and animals. In mammals they and their recognizable remnants account for nearly half of the genome [102, 103], and in some plants they constitute up to 90% of the genome [104] [101].

Many authors discuss upon a possible mutually beneficial relationship between the mobile elements and the host organism. It is clear how mobile elements benefit from the hosting cell, because they use cellular apparatus to be transcribed, translated, to get energy and to be replicated through generations (together with the entire genome). But it is questionable if and how host organisms benefit from mobile elements. Probably the mobile elements are one of the major driving forces shaping the genome during evolution but there are not enough evidences to decide if the nowadays organisms live in spite of this force or if they take advantage from that.

name	family	n. of occurrences
Alu	SINE	1193407
L1	LINE	927393
MIR	SINE	590373
L2	LINE	409271
MaLR	LTR	334078
MER1	DNA	216179
ERV1	LTR	178385
ERVL	LTR	132638
MER2	DNA	83486
CR1	LINE	53926
Tip100	DNA	27006
AcHobo	DNA	18986
RTE	LINE	16406
Mariner	DNA	16259
DNA	DNA	13637
ERVK	LTR	10808
Tc2	DNA	7660
PiggyBac	DNA	2099
MuDR	DNA	1884
ERV	LTR	577
Merlin	DNA	54

Table 1.2: Transposons in the human genome

Usually mobile elements are divided in three different classes depending on the molecular mechanism of transposition:

- DNA transposons,
- autonomous retrotransposons,
- nonautonomous retrotransposons.

Retrotransposons need to be transcribed in a RNA molecule before being transposed while DNA transposons don't needed transcription. DNA transposons and autonomous retrotransposons encode the enzymes needed for the transposition while nonautonomous retrotransposons borrow the enzymes encoded by autonomous retrotransposons.

1.5.1 DNA transposons

DNA transposons are prevalent in bacteria (where they are called IS, or insertion sequences), but are also found in the genomes of many metazoa, including insects, worms, and humans. These elements are generally excised from one genomic site and integrated into another by a "cut and paste" mechanism.

Because sequence specificity of integration is limited to a small number of nucleotides – e.g., TA dinucleotides for Tc1 of Caenorhabditis elegans – insertions can occur at a large number of genomic sites. However, daughter insertions for most, but not all, DNA transposons occur in proximity to the parental insertion. This is called “local hopping” [101].

Active transposons encode a transposase enzyme between inverted-repeat termini. The transposase binds at or near the inverted repeats and to the target DNA. It then performs a DNA breakage reaction to remove the transposon from its “old” site and a joining reaction to insert the transposon into its “new” site. These reactions proceed with the hydrolysis of phosphodiester bonds between the transposon and flanking DNA to liberate 3'-OH residues that carry out the attack at the “new” site. Because the two strands of the “new” DNA are attacked at staggered sites, the inserted transposon is flanked by small gaps which, when filled in by host enzymes, leads to short duplications of sequence at the target sites. These are called target site duplications (TSDs), and their length is often characteristic for a particular transposon [101].

The reactions needed to move a piece of DNA use recombinase enzymes[7] encoded by the transposon itself.

Although these elements generally transpose to genomic sites less than 100 kb from their original site (e.g., the Drosophila P element), some are able to make distant “hops” (e.g., the fish Tc1/mariner element) [101].

1.5.2 LTR retrotransposons

Retrotransposons are transcribed into RNA, and then reverse transcribed and reintegrated into the genome, thereby duplicating the element. The major classes of retrotransposons either contain long terminal repeats at both ends (LTR retrotransposons) or lack LTRs and possess a polyadenylate sequence at their 3 termini (non-LTR retrotransposons).

LTR retrotransposons and retroviruses are quite similar in structure (see figure 1.8). They both contain gag and pol genes that encode a viral particle coat (GAG) and a reverse transcriptase (RT), ribonuclease H (RH), and integrase (IN) to provide enzymatic activities for making cDNA from RNA and inserting it into the genome. They differ in that retroviruses encode an envelope protein that facilitates their movement from one cell to another, whereas LTR retrotransposons either lack or contain a remnant of an env gene and can only reinsert into the genome from which they came. . For these similarities with retroviruses LTR retrotransposons are also called endogenous-retroviruses .

Many LTR retrotransposons target their insertions to relatively specific genomic sites. For example, Ty3 elements of Saccharomyces cerevisiae target specifically to a few nucleotides from RNA polymerase III (Pol III) transcription initiation sites [106]. Moreover, Pol III transcription factors, TFIIIB and TFIIIC, are essential for Ty3 integration. It is interesting to observe that in this way Ty3 elements maximize the probability to transpose themselves in a actively transcribed region. The integration of a transposon in a such regions probably disrupts the original gene that the region held; this behaviour suggests that the LTR (like viruses) are parasites and that the hosting cells can survive in spite of the presence of transposons.

In contrast to the Ty elements of S. cerevisiae, Tf elements of Schizosaccharomyces pombe cluster 100 to 400 nucleotides upstream of Pol II-transcribed genes. The retroviruses HIV (human immunodeficiency virus) and MLV (mouse leukemia virus) share many structural features with LTR retrotransposons. In general, HIV inserts into many sites throughout actively transcribed genes, whereas MLV integrates preferentially into the promoters of active genes. The preference of retroviruses for insertion sites in and around genes may explain the occurrence of leukemia-producing insertions into the promoter of the LMO-2 gene in 2 of 10 patients undergoing retroviral gene therapy for severe combined immunodeficiency [101].

There are also LTR transposons targetting specifically unfunctional zones: Ty5 targets the heterochromatin of telomeres and the silent mating loci. Ty5 requires a specific protein partner, Sir4, for tethering its cDNA to telomeric DNA, and the interaction sites of Ty5 (six amino acids in the integrase domain) with Sir4 (a region near the C terminus) have been characterized [107].

[7] There are two main classes of recombinase enzymes used by transposable elements. The first class is called conservative because the enzymes do not require high-energy cofactors, the total number of phosphodiester bonds remains unchanged, and no DNA degradation or resynthesis occurs. Examples of this recombinase type are the integrase protein of bacteriophage, Cre recombinase, and Flp recombinase. The second class is the transposases that catalyze a whole set of reactions necessary for DNA transposition. Examples are the transposases of Mu, P elements, and the Tc1/mariner family, and the integrases of long terminal repeat (LTR) retrotransposons and retroviruses. All of these enzymes share certain structural motifs such as a D,D35E sequence (aspartate, aspartate, 35 amino acid residues, then a glutamate) and a handlike three-dimensional structure [101, 105].

1.5.3 Non-LTR transposons

Non-LTR retrotransposons are typified by LINE-1 (long interspersed nucleotide elements-1, or L1) elements of mammals. Full-length non-LTR retrotransposons are 4 to 6 kb in length and usually have two open reading frames (ORFs), one encoding a nucleic acid binding protein, and the other encoding an endonuclease and a reverse transcriptase (see figure 1.8). Because these elements encode activities necessary for their retrotransposition, they are called autonomous even though they probably also require host proteins to complete retrotransposition [101]. Some non-LTR retrotransposons integrate at specific genomic sites. R1 and R2 of Drosophila melanogaster and Bombyx mori integrate at specific ribosomal RNA gene locations [108], whereas heT-A and TART elements help maintain the telomeres of Drosophila melanogaster chromosomes [109] and TRAS1 and SART1 integrate into telomeric repeats of B. mori [110]. In contrast, mammalian L1 elements apparently integrate at a very large number of sites in the genome because their endonuclease prefers to cleave DNA at a short consensus sequence (5-TTTT/A-3 , where / designates the cleavage site) [111, 112]. Our knowledge of most of the steps leading to retrotransposition of non-LTR retrotransposons is sketchy except for the reverse transcription process. In contrast to reverse transcription of LTR retrotransposons and retroviruses, this process takes place on nuclear genomic DNA through target primed reverse transcription, or TPRT [113, 114]. The great majority of mammalian L1 insertions are 5 truncated and much less than the full length of 6 kb. However, the mechanism of 5 truncation is still unclear. In about 30% of mammalian L1 insertions, but not in Drosophila R1 or R2 insertions, the 5 end of the insertion sequence is inverted. A likely explanation for this phenomenon is a variation on TPRT, called "twin priming" [115, 101].

Only a few members of the LINE1 family of highly repetitive retrotransposable sequences are capable of autonomous amplification. . Full-length LINE1 is bicistronic: the product of ORF1 is an RNA-binding protein (ORF1p), and ORF2 encodes a protein (ORF2p) with endonuclease and reverse transcriptase activities . The downstream location of ORF2 ensures that translational initiation of these catalytic activities is downregulated with respect to ORF1p expression. Both ORF1p and ORF2p are required for LINE1 retrotransposition, but surprisingly, ORF2p acts efficiently only on the active LINE element that encoded its expression (in cis). A retrotranspositionally successful LINE must sequester its limiting ORF2p, preventing the amplification of either defective LINE elements or entirely unrelated sequences, such as Alu (see section 1.5.4). [116]

1.5.4 Nonautonomous transposons

Other sequence elements, which do not encode their own reverse transcriptase, also transpose via RNA intermediates. These elements include the highly repetitive short interspersed elements (SINEs), of which there are nearly a million copies in mammalian genomes. SINEs arose by reverse transcription of small RNAs, including tRNAs and small cytoplasmic RNAs involved in protein transport. Since SINEs no longer encode functional RNA products, they represent pseudogenes that arose via RNA-mediated transposition. Since these elements do not include genes for reverse transcriptase or a nuclease, their transposition presumably involves the action of reverse transcriptases and nucleases that are encoded elsewhere in the genome —probably by class I or II retrotransposons, such as LINEs. This was shown at least for the Alu repeat that borrows the retrotransposition machinery of the L1 [117]

The most conspicuous human SINE is the Alu repeat family (so called because of early attempts at characterizing the sequence using the restriction nuclease AluI). Alu sequences containing $\simeq$ 300 base pairs are present at $\simeq$ 1 million sites in the human genome, accounting for about 10 percent of the total genomic DNA; similar sequences are abundant in other vertebrates [118]. In addition to full-length Alu sequences, many partial Alu-like sequences, clearly related to the Alu family but as short as 10 base pairs, have been found scattered between genes and within introns in human DNA. The Alu repeat contains an internal RNA polymerase III promoter sequence. Alu sequences are remarkably homologous to 7SL RNA, a small cellular RNA that is part of the signal-recognition particle. This cytoplasmic ribonucleoprotein particle aids in the secretion of newly formed polypeptides

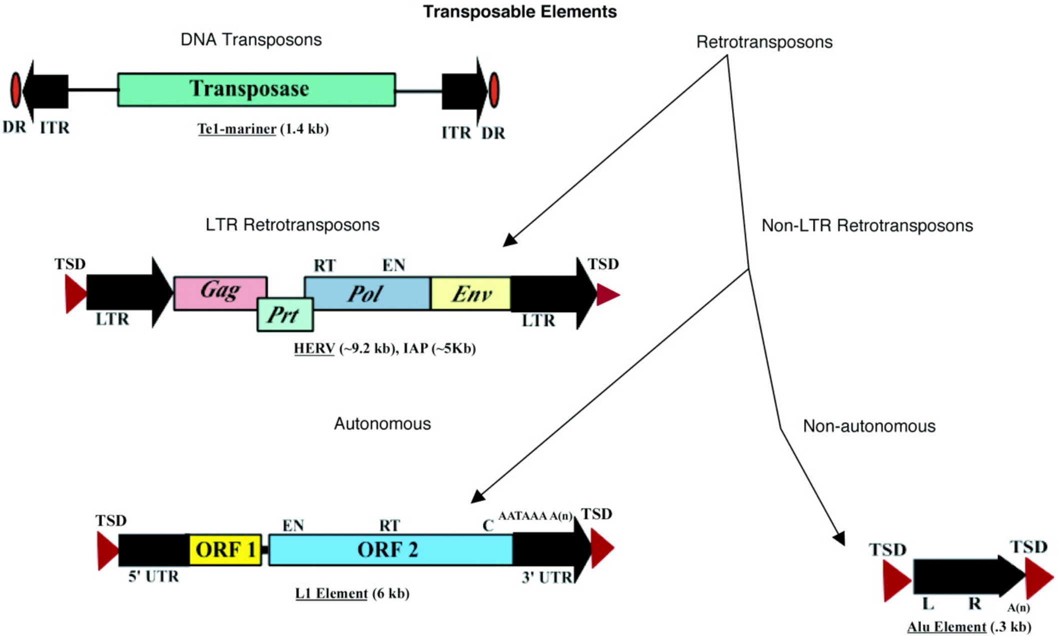


Figure 1.8: Classes of mobile elements. DNA transposons, e.g., Tc-1/mariner, have inverted terminal inverted repeats (ITRs) and a single open reading frame (ORF) that encodes a transposase. They are flanked by short direct repeats (DRs). Retrotransposons are divided into autonomous and nonautonomous classes depending on whether they have ORFs that encode proteins required for retrotransposition. Common autonomous retrotransposons are (i) LTRs or (ii) non-LTRs. Examples of LTR retrotransposons are human endogenous retroviruses (HERV) (shown) and various Ty elements of S. cerevisiae (not shown). These elements have terminal LTRs and slightly overlapping ORFs for their group-specific antigen (gag), protease (prt), polymerase (pol), and envelope (env) genes. They produce target site duplications (TSDs) upon insertion. Also shown are the reverse transcriptase (RT) and endonuclease (EN) domains. Other LTR retrotransposons that are responsible for most mobile-element insertions in mice are the intracisternal A-particles (IAPs), early transposons (Etns), and mammalian LTR-retrotransposons (MaLRs). These elements are not present in humans, and essentially all are defective, so the source of their RT in trans remains unknown. L1 is an example of a non-LTR retrotransposon. L1s consist of a 5'-untranslated region (5'UTR) containing an internal promoter, two ORFs, a 3'UTR, and a poly(A) signal followed by a poly(A) tail (An). L1s are usually flanked by 7- to 20-bp target site duplications (TSDs). The RT, EN, and a conserved cysteine-rich domain (C) are shown. An Alu element is an example of a nonautonomous retrotransposon. Alus contain two similar monomers, the left (L) and the right (R), and end in a poly(A) tail. Approximate full-length element sizes are given in parentheses. Modified from [101].

through the membranes of the endoplasmic reticulum. The 7SL sequence is highly conserved even in species as diverse as Drosophila, mouse, and man. The discovery of a small (≃ 100 nucleotide) E. coli RNA whose sequence is similar to eukaryotic 7SL RNA indicates that this molecule has existed since early in evolution. However, neither Drosophila nor single-celled organisms have any Alu-type intermediate repeats (at least in large numbers). These findings suggest that 7SL RNA genes existed before Alu sequences and that Alu sequences somehow arose fairly late in evolution from the 7SL sequences [118].

The Alu repeat is primate-specific but other mammals have similar types of sequence derived from the 7SL RNA gene such as the B1 family in mouse. Unlike the Alu repeat, another major human SINE family is not restricted to primates, with copies being found in marsupials and monotremes. In accordance with its distribution this family has been termed the MIR (mammalian-wide interspersed repeat) family [118].

Like all other mobile elements, Alu sequences usually are flanked by direct repeats. Although Alu sequences do not encode proteins and contain an A/T-rich region at one end, similar to L1 elements. Consequently, Alu sequences are thought to be retrotransposed by a mechanism similar to that proposed for L1 elements , possibly by the reverse transcriptase and other required proteins expressed from functional L1 elements [117].

Alu duplication mechanism and pseudogene formation

Both Alu and processed pseudogenes are duplicated by the LINE1 duplication machinery.

LINEs are extremely versatile genome modelers. First, they can transpose, via a high-efficiency process involving the LINE ORF products and the transcript encoding them (cis effect). Second, they can retrotranspose, with a reduced efficiency, transcribed DNAs (trans effect) and therefore mobilize transcribed sequences not necessarily associated with a LINE element. The trans effect results in the generation of retrotransposed copies disclosing features characteristic of the naturally found processed pseudogenes [119].

There is no RNA sequence specificity for the LINE-mediated trans effect. Indeed LINEs is also be responsible for the mobilization of the SINE [117] retrotransposons which are noncoding and therefore require complementation in trans for their retrotransposition. Concerning the higher relative efficiency of the cis versus trans LINE effects this is most probably due to the fact that there is a direct recognition of the LINE mRNA in the course of its translation as a target for retrotransposition [120].

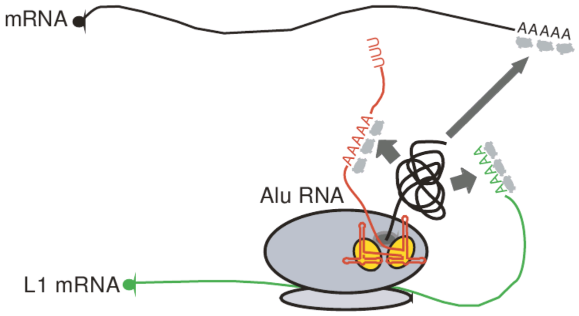


Figure 1.9: A model accounting for the LINE cis and trans effects. LINE proteins are particularly efficient for retrotransposition of the mRNA that encodes them, an effect (the cis effect) that can be partially explained by spatial proximity between the LINE nascent polypeptides (thick black line) and LINE RNA (ribosome in gray, LINE RNA in green). LINE proteins can retrotranspose cellular mRNA or non-coding LINE RNA (in black) as well, but with a much lower efficiency (1 mRNA retrotransposed every 3000 LINE retrotransposed) owing to greater distances and lower probabilities of interaction (trans effect). The highly structured Alu transcripts (in red) bind to the SRP9/14 proteins (shown in yellow), which in turn interact with the ribosome, positioning the Alu transcript close to the nascent LINE proteins (1 Alu retrotransposed every 300 LINE retrotransposed). Poly-A-binding proteins (light gray) interact with the Alu poly-A track (essential for Alu retrotransposition), as is the case for the poly-A of mRNAs, and target LINE ORF2p for a precise 3 end initiation of the target primed reverse transcription. Modified from [117].

Alu elements are ancestrally derived from the SRP RNA gene and Alu RNA binds SRP proteins. The translational role of SRP and its ribosomal location provide a plausible mechanism for co-compartmentalizing Alu RNA with nascent cis-acting ORF-2p [116]. This can explain the higher frequency of duplications of Alu versus other transcript. Furthermore, as expected, catalytic activities encoded by ORF2 are essential for efficient Alu retrotransposition but exogenous ORF1p is dispensable. If Alu does not require ORF1p, a large number of truncated LINE elements, lacking ORF1 and incapable of autonomous retrotransposition, might provide sufficient ORF2 activity for Alu amplification. Accordingly, the requirements of Alu for retrotransposition could be far more permissive than those of LINE1, partially explaining the relative success of Alu elements. [116]

Other reverse transcriptase expressing elements in eukaryotic genomes, like the LTR retrotransposons, were not likely to be responsible for processed pseudogene formation. In fact, attempts to generate such structures using retroviruses or retroviral-like elements failed to demonstrate canonical pseudogenes [121]. Therefore LINE-encoded retrotransposition machinery had specific enzymatic properties not shared by that of retroviruses (including MoMLV and HIV), allowing the reverse transcription of cellular mRNA with high efficiency and resulting in non-integrated cDNA copies [121]. Although the precise role of ORF1 is still unknown, it is absolutely required for pseudogene formation: it might be involved in the formation of 'particles'4,27,28 allowing the re-entry of the mRNA to be retrotransposed into the nucleus and integration of the reverse transcripts. A final hint for a role of LINEs in processed pseudogene formation arises from the systematic sequencing of the Saccharomyces cerevisiae genome, which provided evidence for the presence of numerous reverse transcriptase expressing elements (including functional telomerases and several active LTR-retrotransposons), but the correlated absence of LINEs, processed pseudogenes and Alu-like sequences [122].

1.5.5 Functional role of transposons

The initial evidence for the mobility of SINES came from analysis of DNA from a patient with neurofibromatosis, a genetic disorder marked by the occurrence of multiple neuronal tumors called neurofibromas due to mutation in the NF1 gene. Like the retinal tumors that occur in hereditary retinoblastoma, neurofibromas develop only when both NF1 alleles carry a mutation. In one individual with neurofibromatosis, one NF1 allele contains an inactivating Alu sequence; inactivating somatic mutations in the other NF1 allele in peripheral neurons lead to the development of neurofibromas. Several other inherited recessive mutations causing disease in humans also have been found to result from insertion of Alu sequences in exons, thereby disrupting protein-coding regions [118].

Alu sequences appear to have retrotransposed widely through the human genome and are tolerated, in both possible orientations, at sites where they do not disrupt gene function: flanking solitary genes and between duplicated genes, as well as within introns and the regions transcribed into the 5' and 3' untranslated regions of mRNAs. Alu sequences are thought to have no function, like other mobile elements, despite their widespread occurrence in mammalian genomes.

Although mobile DNA elements appear to have no direct function other than to maintain their own existence, their presence probably had a profound impact on the evolution of modern-day organisms. As mentioned earlier, many spontaneous mutations in Drosophila result from insertion of a mobile DNA element into or near a transcription unit, and mobile elements also have been found in mutant human genes. In addition, homologous recombination between mobile DNA elements dispersed throughout ancestral genomes may have been important in generating gene duplications and other DNA rearrangements during evolution . Cloning and sequencing of the β-globin gene cluster from various primate species have provided strong evidence that the human Gγ and Aγ genes arose from an unequal homologous crossover between two L1 sequences. Such duplications and DNA rearrangements contributed greatly to the evolution of new genes. Gene duplication probably preceded the evolution of a new member of a gene family, which subsequently acquired distinct, beneficial functions.

Mobile DNA most likely also influenced the evolution of genes that contain multiple copies of similar exons encoding similar protein domains (e.g., the fibronectin gene). Homologous recombination between mobile elements inserted into introns probably contributed to the duplication of introns within such genes. Some evidence suggests that during the evolution of higher eukaryotes, recombination between introns of distinct genes occurred, generating new genes made from novel combinations of preexisting exons. For example, tissue plasminogen activator, the Neu receptor, and epidermal growth factor all contain an

EGF domain . Evolution of the genes encoding these proteins may have involved recombinations between mobile DNA elements that resulted in the insertion of an EGF-encoding exon into an intron of the ancestral form of each of these genes. The term exon shuffling has been coined to refer to this type of evolutionary process. This phenomenon relies on the poor efficiency of the LINE polyadenylation sequence, which results in read-through LINE transcripts and the de facto transfer of the transcribed 3' genomic sequence to a new location upon LINE retrotransposition [123, 124, 125] [119].

Recombination between mobile elements also may have played a role in determining which specific genes are expressed in particular cell types and the amount of the encoded protein produced. Eukaryotic genes have transcription-control regions, called enhancers, that can operate over distances of tens of thousands of base pairs 1.2.3. Moreover the transcription of a gene can be controlled through the combined effects of several enhancers. Recombination between mobile elements inserted randomly near enhancers probably contributed to the evolution of the combinations of enhancers that control gene expression in modern organisms.

So, the early view of mobile DNA elements as completely selfish molecular parasites appears to be premature. Rather, they have probably indirectly made profound contributions to the evolution of higher organisms by serving as sites of recombination, leading to the evolution of novel genes and new controls on gene expression [118].

Chapter 2

Ab initio gene prediction through processed pseudogenes

2.1 Background

With the development of genome projects for many organisms, an increasing number of raw sequences needs to be annotated. In the case of unicellular organisms, gene identification by computational methods is quite straightforward, considering the limited amount of non-coding DNA. In contrast, in the case of metazoans, the annotation of all the different RNAs that may be produced by the genome still represents a daunting task, requiring the integration of predictive algorithms and experimental evidences.

Several sophisticated software algorithms have been devised to handle this problem [126]; these algorithms typically consist of one or more "sensors" designed to identify gene features. A single gene prediction program usually combines different sensors and complex pipelines combine together different tools. The ENSEMBL pipeline, for example, starts by using programs for signal and content terms (often called ab initio programs); then, to reduce the high incidence of false positives, the resulting gene predictions are "fixed" by the incorporation of similarity information [127]. Although gene-finding programs can correctly predict most exons of each gene, they are usually unable to cope with the complexity generated by the alternative use of transcription units, leading to the production of many mRNA variants.

Comparative genomics represents a very powerful strategy for the identification of exons and regulatory elements. The assumption behind this type of analysis is that phylogenetic conservation is related to functional relevance. Nevertheless, this approach is unable, by definition, to reveal species-specific genes and transcript variants. The generation of massive EST sequence data to be matched with the genomic sequence is probably the most direct and efficient method to tackle the annotation problem. However, even in the case of the human genome, for which more than 8 million EST sequences have been obtained, some annotated genes are still poorly represented in the corresponding databases (see for example ref. [128]). Moreover, splice variants unique to specific cell types, developmental stages and abnormal conditions may still escape detection. Obviously, these problems are much worse in most of the newly sequenced genomes, which lack such extensive transcriptome coverage.

Processed pseudogenes (PPGs) represent an alternative source of full length transcript information, contained in the raw genome sequence [129]. PPGs are copies of cellular RNAs, typically containing poly(A) and lacking introns, which have been reverse transcribed and inserted into the genome. Although PPGs cannot be expected to completely cover the transcriptome of most eukaryotes, they represent a rich sample of it [130], as they may derive from normal protein-coding mRNAs, alternatively spliced mRNAs [131], non-protein-coding RNAs [132] and antisense transcripts [123].

The methods so far devised to identify PPGs are based on the use of known mRNAs and protein sequences [131] or of gene predictions [133] as input data for suitable alignment programs. Therefore, they could be expected to perform poorly on genomes lacking extensive transcriptome annotation or on PPGs derived from non-canonical genes.

In this report, we describe a general method for the systematic identification of retrotransposition events, based on a completely different strategy. In particular, we search for generic paralogous alignments (i.e. alignments of a given genome with itself) and then select only those showing evidences of three or more splicing events. Using this approach we identified 987 human and 709 mouse genes. Most of the identified genes were already known or supported by EST tracks, but in a few cases they were completely new predictions, not supported by any type of evidence in the UCSC or ENSEMBL databases. We were able to experimentally validate some of these predictions.

2.2 Results and Discussion

2.2.1 Construction of the pseudogenes database.

We based our pipeline on the observation that a processed pseudogene can be recognized, in a set of pairwise paralogous alignments, as a cluster of nearby alignments (the exons of the retrotransposed gene) separated by unaligned sequences (the introns of the retrotransposed gene, see Fig. 2.1). Although one of these splicing events would be enough to identify a canditate gene, in order to decrease the false positive rate we required at least three splicing events. As mentioned above, the major reason of interest of our approach with respect to the existing ones [134, 135, 130, 136, 129, 131] is that it does not rely on known protein sequences, thus allowing to identify previously unknown genes.

We organized our pipeline as a set of consecutive steps which are discussed in detail in the methods section.

- **Construction of a database of paralogous alignments.**
- **Identification and refinement of the alignment clusters, which are our putative gene-pseudogene pairs.**

  This requires the careful identification and reconstruction of "corruption gaps" i.e. those portions of the pseudogene sequences which have been corrupted by random mutations.
- **Identification of the alignments due to DNA duplication events.**
- **Identification of the "gene side" with respect to the "pseudogene side" of putative pairs.**

  Both these steps are based on the identification of the gaps in the alignment due to introns. This allows to remove alignments not associated with the retrotranscription of processed mRNAs and at the same time it allows to distinguish in an unambiguous way the "gene side" of the alignments from the pseudogene one: this step is the cornerstone of our algorithm.
- **Refinement of the database.**

  Once the "gene side" has been identified we can filter our putative candidates using our knowledge of the expected mean size of introns and looking at possible "fake introns" due to repeat insertions.
- **Identification of the pseudogene families.**

  In some cases a single gene can give rise to many processed pseudogenes. They would appear as separate entries in our database, so it is mandatory to recognize these families and to associate them to the unique original gene.
- **Mapping of the candidate genes in the ENSEMBL and UCSC databases.**

  For each prediction we look at the ENSEMBL and UCSC databases in order to distinguish between already known genes and candidate new genes.

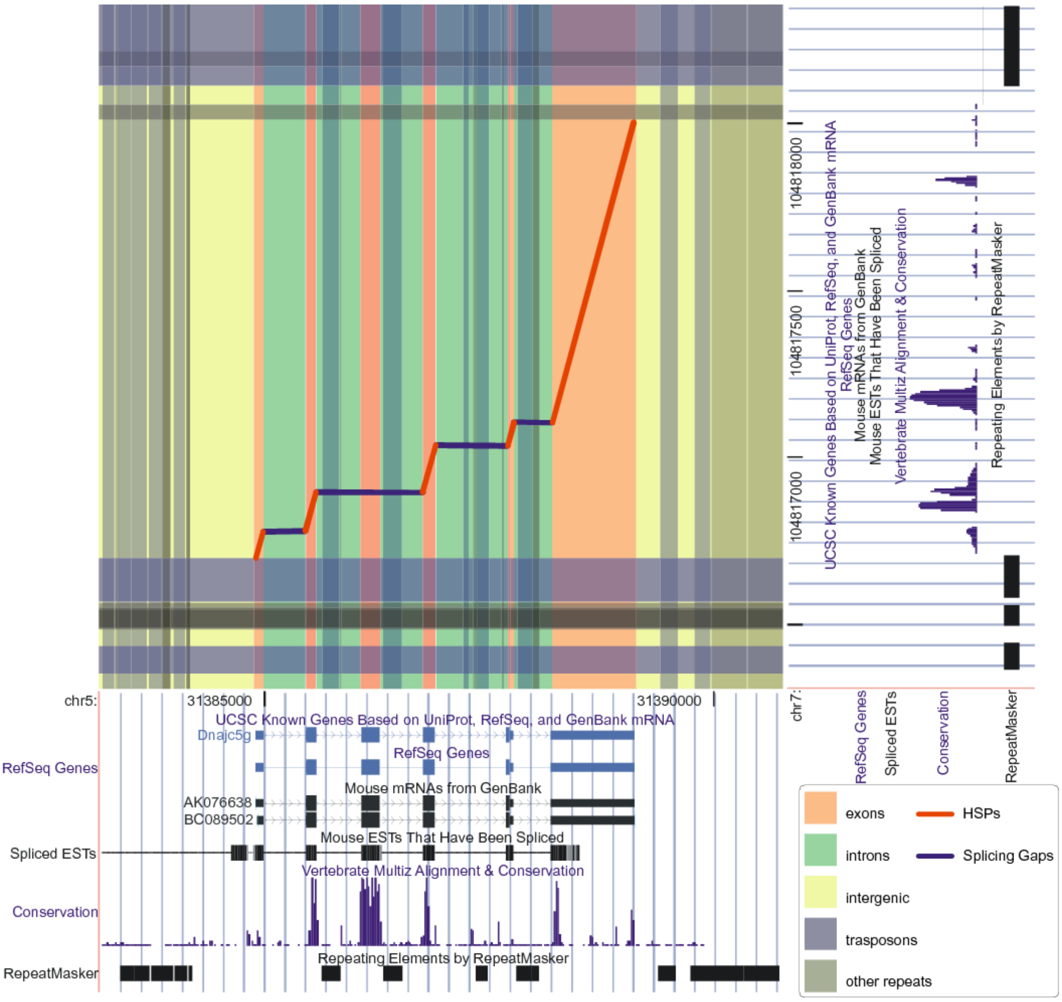


Figure 2.1: This figure shows a graphical representation of one entry of our dataset, corresponding to a gene-pseudogene pair in the mouse genome. The graph is similar to a dot-plot. On the horizontal axis we put the region where we identified the gene and its annotation from the UCSC genome browser; on the vertical axis the region corresponding to the pseudogene. Each alignment between the two regions is represented as a red segment in the central square, while blue segments are the splicing signatures recovered by our pipeline. Finally, the background is colored in horizontal and vertical stripes mirroring the sequence types (exons, introns, intergenic regions, transposons and other repeats). We can see tath each alignment (or HSP) corresponds quite well with an already known exon. The transcript that give rise to the processed pseudogene detected is a previously not annotated alternative transcrip in which the third exon was skipped. This report is produced by our visualization tool CGV (http://to444xl.to.infn.it/cgv).

Using this pipeline we found in the human genome 2288 gene-pseudogenes pairs, corresponding to 987 candidate genes. Out of these, 965 genes had at least one exon annotated in ENSEMBL and 943 had at least one overlapping UCSC known gene or RefSeq; in 15 cases we found neither (see table 2.1).

A similar analysis in mouse led to 29 candidates completely free of ENSEMBL, UCSC known genes or Refseq annotations. For the sake of clarity, in the following we shall only quote the results in the human case (the complete results for the mouse genome can be found in Table 2.1)

As a by-product of our analysis we identified several pseudogene families; we report in table 2.2 the largest ones.

	Human	**Mouse**
Total number of genes	987	709
Supported by UCSC known genes	928	649
Supported by RefSeq	922	655
Supported by ENSEMBL or VEGA genes	965	668
Supported by ENSEMBL or VEGA coding genes	948	661
Supported by UCSC, RefSeq, ENSEMBL, VEGA or EST	979	680
New predictions	8	29

Table 2.1: Summary of our analysis results

2.2.2 Comparison with other pseudogenes databases.

To obtain an independent validation of our approach we compared our results with processed pseudogenes reported in the human section of the Vertebrate Genome Annotation (VEGA) database, a central repository of high quality, manually curated annotations [137] and with the pseudogene.org pipeline dataset [131]. Overall, approximately 30% of our predictions were contained in the VEGA database. However, a more detailed analysis revealed that for chromosomes 1, 9, 10, 13, 20, X and Y the overlap was more than 90%, while in all the other cases it was far below 10%, being equal to zero in most cases, with the only exception of chromosome 18. This is clearly due to the fact that the present release of the VEGA database is still incomplete and, as far as pseudogenes are concerned, only a few chromosomes have been extensively annotated.

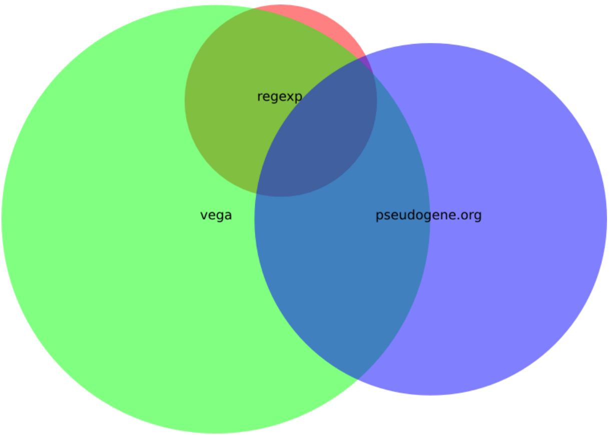


Figure 2.2: A Venn diagram showing the intersections among our dataset (red, 442 genes), ENSEMBL VEGA (green, 2194 genes) and pseudogene.org pipeline dataset (blue, 1484 genes) for the chromosomes for which the VEGA annotation has been already completed (1, 9, 10, 13, 20, X). The number of genes in common between datasets are: REGEXP – ENSEMBL VEGA: 417; REGEXP – pseudogene.org: 154; ENSEMBL VEGA – pseudogene.org: 607. 152 genes are shared by all three datasets.

We then compared our results with the pseudogene.org pipeline dateset [131], and also in this case we found a significant overall overlap, though not as good as in the case of VEGA. In particular, for the chomosomes in which the VEGA database seems to be complete our results show a much better overlap with the VEGA ones than with those of pseudogene.org (see fiugre. 2.2).

These results strongly validate the specificity of our procedure, indicating that it could be used to reliably predict pseudogenes in non annotated genomes.

regexp stable gene id	pseudogene number	retrotransposed genes
24030	88	SNORD102 small nucleolar RNA, C/D box 102 SNORA27 small nucleolar RNA, H/ACA box 27 RPL21 ribosomal protein L21
92497	62	KRT18 keratin 18
18623	52	RPSA ribosomal protein SA
133262	45	HNRPA1 heterogeneous nuclear ribonucleoprotein A1
84588	45	RPL7 ribosomal protein L7
178537	39	RPS3A ribosomal protein S3A
15729	38	GAPDH glyceraldehyde-3-phosphate dehydrogenase
17563	37	RPS2 ribosomal protein S2
200584	35	NPM1 nucleophosmin (nucleolar phosphoprotein B23, numatrin)
8698	32	SNORD36C small nucleolar RNA, C/D box 36C SNORD24 small nucleolar RNA, C/D box 24 SURF4 surfeit 4 SNORD36B small nucleolar RNA, C/D box 36B SNORD36A small nucleolar RNA, C/D box 36A
921125	32	SNORD21 small nucleolar RNA, C/D box 21 RPL5 ribosomal protein L5
200296	26	RPL12 ribosomal protein L12
12076	21	RPL7L1 ribosomal protein L7-like 1
87014	21	HMGB1 high-mobility group box 1
109645	20	KRT8 keratin 8
161788	20	RPS15A ribosomal protein S15a
18181	17	ACTG1 actin, gamma 1
237683	17	RPS4X ribosomal protein S4, X-linked
5449	17	RPL36A ribosomal protein L36a
41035	16	ACTB actin, beta
11000	15	RPL31 ribosomal protein L31
25634	15	RPL19 ribosomal protein L19
698758	15	RPL9 ribosomal protein L9
98433	15	RPL34 60S ribosomal protein L34
106898	14	SET SET translocation (myeloid leukemia-associated)
740881	13	ACTG2 actin, gamma 2, smooth muscle, enteric
27176	12	RPS18 ribosomal protein S18
296506	11	EEF1A1 eukaryotic translation elongation factor 1 alpha 1
51277	11	HMGB3 high-mobility group box 3
95436	11	LOC497661 putative NFkB activating protein

Table 2.2: Where more than one gene is reported as retrotransposed in the same family of pseudogenes the additionals genes are snoRNA. The only exception is represented by SURF4: its transcritps most likely didn't produce any processed pseudogene but it is reported because its exon at the 3' end overlaps with an exon of SNORD36A.

2.2.3 Processed pseudogenes as indicators of the transcriptional activity of the cell.

It was recently observed that the transcribed portion of mammalian genomes [138] is much larger than previously expected and that it is characterized by a high level of complexity, with an impressive amount of antisense and overlapping transcripts. Moreover, it was noticed that several of these new transcripts were processed, i.e. they underwent one or more splicing events. However it is not clear whether these new processed transcripts are abundant and stable products, or they should be rather considered as exceptional events (or possibly random "junk" events). Since a clear relationship exists between transcript stability and the frequency of retrotransposition [130] and since our approach for the identification of retrotransposition events is not biased for protein-coding transcripts, we reasoned that our database is ideally suited to address this issue. Interestingly we found that most of the entries in our database correspond to protein coding genes. A possible explanation for this strong preference is that it could be due to a special affinity of the retrotransposition machinery for particular functional classes of mRNAs. To address this point we studied the GO annotation of the genes in our database and weighted them with the size of the corresponding pseudogene families. Our results revealed a clear overrepresentation of sequences derived from ribosomal protein genes and, apparently, no other particular bias in the GO annotations (see table B.1) in good agreement with what already observed by Yao et al. [136]. These results indicate that the pseudogenes contained in our database (except for the ribosome related ones) represent a reasonably unbiased sample of the cell transcriptome. Thus the fact that we mostly find protein coding genes among our entries suggests that most of the non-coding transcripts produced by mammalian genomes are rather rare or less stable than protein coding ones.

2.2.4 Identification of new putative alternative splicing events.

In several cases we find instances of previously unknown alternatively spliced transcripts, in particular, for 75 human genes (out of the 965 which we could associate to ENSEMBL transcripts), we recovered additional exons. This is consistent with the idea already discussed in ref. [129] that PPGs can be effectively used to indentify alternative splicing events.

Out of these alternative exons 50% were supported by EST tracks, while the others are completely new. Similar results were obtained for the mouse genome (51 new alternative transcripts with 63% of the exons supported by ESTs).

2.2.5 Identification and validation of putative new genes.

One of the most interesting aspects of our method is that, in principle, it should be able to reveal the existence of functional genes independently from homology with previously identified cDNAs, even when they correspond to completely species-specific sequences. Among our predictions, 22 human sequences (2%) and 41 mouse sequences (6%) did not correspond to known genes in the ENSEMBL database. Nevertheless, for most of them we could find EST tracks covering the majority of the predicted exons. However, 8 human and 29 mouse new genes identified by our method did not correspond to available ESTs, and were not predicted by other gene finding programs. To discriminate whether these predictions correspond to false positive results or to actual new genes, we directly validated a sample of them. We reasoned that, if these sequences were produced by functional genes which are still active in the modern genomes, the corresponding mRNAs should be expressed at least in the germ line. To obtain direct support to this hypothesis, we designed for three of them in human and three in mouse specific PCR primers, perfectly matching the nucleotides of two different exons of the gene-side sequence, and performed RT-PCR on human and mouse testis cDNA, respectively. Remarkably, we recovered amplification products of the expected molecular weight (see figure 2.3) for two human and two mouse genes. Moreover the obtained products were further verified by direct sequencing (see B.4). Interestingly, the confermed transcripts appear to lack long open reading frames and may thus correspond to untraslated sequences.

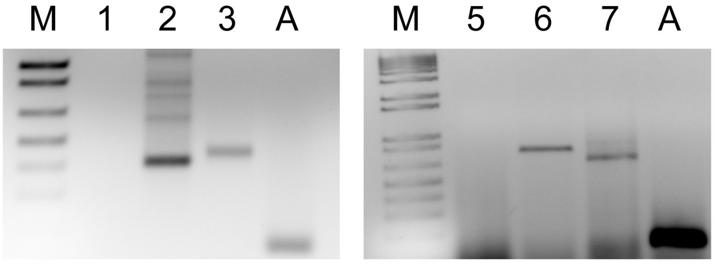


Figure 2.3: RT-PCR amplification of human (left panel) and mouse (right panel) testis cDNA, with primers specific for predictions 316640 (1), 338893 (2), 128365 (3), 564151 (5), 718689 (6) and 1033649 (7). The major band in the lanes shows an amplification product corresponding to the expected molecular weight. M: molecular weight standard; A: beta-actin positive control.

Detailed analysis of the genomic sequence revealed that one of the human candidates (ID 128365) contains an exonized DNA transposon belonging to the so called "mariner" family. It has been recently shown that this class of transposons had a burst of activity in the primate lineage [139]. Indeed a careful phylogenetic analysis showed that the mariner was transposed in the gene just before the separation of the galago (Otolemur garnettii), a prosimian primate, from the anthropoid lineage (i.e. between 60 and 80 My ago); it was then exonized and retrotransposed with the original gene. A copy of the original (i.e. without the mariner insertion) gene exists in all the other mammalians.

2.3 Conclusions

We have presented REGEXP, a new highly specific method for the annotation of retrotransposition events, which revealed new exons and genes even in the very well annotated human and mouse genomes. Since our pipeline does not require other information but the genomic sequence, we expect a similar efficiency even in the case of genomes lacking extensive transcriptome information.

2.4 Methods

The alignment database

We start from the full set of local alignments found by comparing the repeat masked sequence of the human genome (build 36) with itself; we compute these alignments with the Megablast software [140]. To avoid excessive memory occupation we split the chromosome sequences into smaller slices and compare them all. We choose to split the sequences when we find a repeat masked region longer than 1000 base pairs (usually a LINE); this way we don't need to postprocess the alignments to merge overlapping slices. We are confident that no alignment containing a masked region of 1000 bps or more can exist since its score would be under any reasonable statistical cutoff. The alignment database contains about 12 milions high scoring pairs (HSPs, pairs of regions sharing high sequence similarity) longer than 30 bps.

We label each HSP a using two aligned regions r_{a1} and r_{a2} which are identified by their starting and ending points in absolute chromosomal coordinates: $r_{a1} = (a_{11}, a_{12})$ and $r_{a2} = (a_{21}, a_{22})$. This induces a natural definition of distance between HSPs $a = (r_{a1}, r_{a2})$ and $b = (r_{b1}, r_{b2})$ as the length of the smallest segment joining two endpoints, i.e:

$$d(a,b) = \begin{cases} \infty & r_i \text{ and } r_j \text{ on different chrs} \\ \min\left(d(r_{a1}, r_{b1}), d(r_{a1}, r_{b2}), d(r_{a2}, r_{b1}), d(r_{a2}, r_{b2})\right) & \text{otherwise} \end{cases}$$

where $d(r_{ai}, r_{bj})$ is the euclidean distance between two points.

Location clusters

Since a processed pseudogene is the union of the exons of the original gene one would expect to find it in the alignment database looking for clusters of nearby HSPs. On one side of the alignment (the "pseudogene side") we expect multiple HSPs very close to each other (ideally, if no insertion occurred after the retrotransposition event they should be contiguous); on the other side (the "gene side") they will be near but separated by gaps corresponding to the introns that are missing from the pseudogene. Even if we allow for the presence of mutations in one or both the sequences, the scenario remains quite the same. Some of the original HSPs may now have a lower score, some may as well have disappeared; but the picture still consists of a number of HSPs clumped one next to the other. To extract these HSP clusters (which we shall denote in the following as "location clusters") from the alignment database we developed the following clustering procedure. Each HSP can be represented as a segment in the bidimensional plane spanned by the two sequences (in a way that closely resembles dot-plots); we cluster together two consecutive alignments / segments if the distance between the two segments is lower then a certain threshold (we chose 22Kbps because only 5% of known human introns are longer than that) along both directions. If at least three of these segments are concatenated together we consider the resulting group a location cluster.

As a result of this definition each location cluster can be considered as the bidimensional bounding box of a set of at least three nearby segments and any two location clusters are separated both horizontally and vertically by more than 22Kbps.

These location clusters are the starting point of our analysis. The remaining part of the computational pipeline is devoted to refine them and to filter out those that do not conform to certain requirements (described in the following sections). We consider each location cluster surviving the entire filtering process as a candidate gene-pseudogene pair.

Corruption gaps

In some cases processed pseudogenes may have accumulated so many mutations that only a small portion of the original duplicated region can be retrieved using a standard alignment algorithm. Typically this lack of homology with the original sequence shows up as a series of gaps in the alignment cluster: we call them "corruption gaps". Our goal is to separate these gaps from those due to the intron splicing.

To identify the corruption gaps we use as anchors the high scoring pairs (each HSP can have itself small gaps, as a consequence of standard alignment algorithms, but these usually do not create problems).

As mentioned above each alignment can be represented as a segment on the cartesian plane having as x and y axes the two genomic regions. Similarly to what happens in dot-plot graphs, these segments lie on lines with angular coefficent exactly ± 1 if there are no gaps in the HSP (the sign of the angular coefficient depends on the strand of the alignment). Given that we use a scoring system penalizing gaps, the angular coefficient of segments representing an HSP is always near ± 1.

We join two HSPs, represented by segments a and b, with a new segment c (that we define a "corruption gap") if the distance $d(a, b)$ is smaller than 3000 bps and if the angular coefficent of c is 45 ± 5 degrees.

We chose the values of these parameters considering some exemplar cases; the final results are only slightly influenced by such values.

We call a set of high scoring pairs joined by corruption gaps a "diagonal": its projections on the two axes define two regions that are a candidate exon or pseudoexon (homologous of an exon in a pseudogene).

Splicing gaps

The other class of gaps that we expect to find in the alignments clusters are those due to the splicing of introns in the processed pseudogenes. These are of great importance for our identification process since they allow us to distinguish the original gene from its retrotransposed copy.

Introns in the mRNA of a gene are expected to be spliced before the retrotransposition event, so we expect to see candidate pseudoexons that are close together while the corresponding candidate exons are separated by gaps that we call "splicing gaps".

A splicing gap is found by looking at the geometrical distribution of diagonals: if the segment joining two diagonals has a projection on one of the two axes that is less than σ bps in length while on the other axis the projection is larger than β bps, then we add this segment to the location cluster as a splicing gap.

We set the threshold β looking at the intron length distribution and choosing a value such that only the 5% of all introns are smaller than β bps; i.e. we expect to loose only a 5% of true introns because of this cutoff. In the human case the threshold turns out to be $\beta = 74$. The parameter σ accounts for the fuzziness of diagonals that may not be precise at the extremes; for this parameter we use a value of 15.

We can project a splicing gap on both axes of the cartesian plane: we consider the longest projection as a candidate intron.

Another reason for which the identification of the splicing gaps is of crucial importance is that it allows us to separete the "true" processed pseudogenes from alignments (and possibly unprocessed pseudogens) deriving from duplications of a portion of the genome. To this end we discard all location clusters without splicing gaps; to further reduce the number of false positives we actually require the presence of at least three splicing gaps in each location cluster to continue its processing along the pipeline (in fact only the 4% of the human genes contain only one intron).

In some cases it may happen that splicing gaps are found on both sides of a location cluster, for instance due to large repeat insertions on the pseudogene side. To avoid misclassification we eliminate these location clusters from our dataset (669 out of 22123 location clusters with splicing gaps).

For all the remaining location clusters we can unambiguously recognize which of the two axes holds a candidate exon (we call that side b) or a candidate pseudoexon (side s). The segments associated with the splicing gaps (which have projections only on the b side) denote our putative introns.

Trimming

Once we have identified the two sides (gene and pseudogene) of the location cluster we can perform a further refinement of our candidate. Indeed it often happens that the central alignment core, the signal of a retrotransposition event, is flanked by spurious alignments having no relation with the gene-pseudogene couple. We may eliminate them imposing the constraint that the pseudoexons on the pseudogene side should be "close enough" to each other.

To implement this constraint we evaluate the median μ of the gaps g_i between consecutive pseudoexons and the median s of their square variance defined as

$$s = \text{median}_i \left\{ (\mu - g_i)^2 \right\}$$

We then removed recursively alignments at the extremes of location clusters if the gap they open on the pseudogene side is larger than $\mu + 2\sqrt{s}$.

Analysis of the repeat content of candidate introns.

A possible source of misclassification in our analysis is the presence in a duplicated genomic region of one or more transposons inserted after the duplication. These inserted sequences could be erroneously interpreted as spliced introns by the pipeline described above thus leading to a wrong classification of the location cluster.

To avoid this problem we look at the transposon content of all the candidate introns and discard those whose sequence was composed for more than 90% by transposons. We then discard all the location clusters with less than two surviving introns.

Out of the initial 1588810 location clusters only 2288 survived all the steps of the above pipeline; they represent our database.

Description of our database

For each candidate gene-pseudogene pair we report in our database (see http://to444xl.to.infn.it/regexp2/ or tables B.2 and B.3) the genomic coordinates and further information which allow to better characterize the gene-pseudogene pair.

For both the gene and the pseudogene we report an annotation vector with seven entries correspondig to the seven possible annotations:

1. coding exon
2. non-coding exon
3. 5'UTR exon
4. 3'UTR exon
5. intron
6. upstream
7. intergenic

The category of non-coding exons includes all exons sequences that do not code for protein portions and are not marked as 5' or 3' UTRs (belonging to RNA genes for example). We get all these annotations from ENSEMBL with the only exception of upstreams: we define an upstream as the region ranging from 15 Kbps upstream of an annotated translation start site (TSS) to the TSS itself. Obviously a single nucleotide can belong to different categories: for example a 3' UTR can be within 15 Kbps from the TSS of another gene or an exon of a gene can fall inside an intron of another gene. In such cases we report, for each nucleotide, only the "stronger" category; we assume the strength of a category as indicated in the previous list (the coding exon being the strongest). We report in each entry of the annotation vector the number of nucleotides of the gene (or pseudogene) which belong to the corresponding category. For each intron of the gene we report also the fraction of nucleotides annotated as transposons.

For each gene we also report the fraction of its exons overlapping UCSC gene annotations tracks.

Retrieval of external datasets.

We obtained the lists of previously annotated genes from ENSEMBL [127] release 40 (August 2006), VEGA [137] release 40 (August 2006) and UCSC releases hg18 and mm8 (downloaded in September 2006). We obtained the lists of VEGA PPGs filtering the full VEGA gene dataset for the biotype "processed pseudogene". We also downloaded the full pseudogene set provided as the pseudogene.org pipeline output [131] (September 2006) and we later extracted all the processed pseudogenes linked to a valid ENSEMBL gene ID.

Identification of pseudogene families.

A relevant number of location clusters match with more than one other location cluster. This happens in two cases: either when a single gene produced many pseudogenes, or when a single processed pseudogene shares high sequence similarity with more than one gene belonging to the same family. In the first case we can define pseudogene families and associate them with a single original gene; in this way we classify 2288 total pseudogenes in 987 families. In the second case we report all the putative genes associated with the pseudogene and do not perform any further analysis. One or more of the candidate genes associated with a single PPG could be unprocessed pseudogenes; in principle one could distinguish them from the gene which originated the retrotransposition event looking in details at the alignments. Suppose that a single gene gives rise to both a processed and an unprocessed pseudogene: if the pseudogenes are free from selective pressure and therefore mutate randomly, the mutation events are independent and one could expect to find a better sequence homology between the PPG and the gene than between the PPG and the unprocessed pseudogene.

Experimental validation of the new candidate genes.

The amplification primers were designed on two consecutive exons on the gene side of our predictions. To ensure their specificity, all the sequences differed from the corresponding pseudogene sequences at least on their 3' end nucleotide. The sequence of primers was as follows:
Pred. 128365 (human) = TGATCAAATAAATATGACAAATG, TTTCACCCATTCTGGCACAATCT
Pred. 338893 (human) = AACGCCATAGGCCTGGGGCGGGT; CACAGCCCAGGGATCAGAAAAG
Pred. 316640 (human) = GACCCCAGTACTCATTTGCCAGG; GGAGCCACATCTATTCACCTATT
Pred. 718689 (mouse) = AGAAAGAGTATGATTTCATAATAGG; TTGATTAAAGTGGTATTTGGTGA
Pred. 1033649(mouse) = TTTTACAGGAGTGGAGTCCCTCA; TGTAGTCCATCTTCTAAGCCCAG
Pred. 564151 (mouse) = ACCAGCTGGTACTTAATGTGAAT; GGCTCACCAAGGTATTTCTGAAGA

Human testis cDNA was commercially obtained (Clontech). Mouse cDNA was obtained by reverse transcribing testis total RNA with MMLV reverse transcriptase (Promega), according to manufacturers specifications. In this case, negative controls lacking the reverse transcriptase were included. B-actin primers were used as positive control in both human and mouse samples.

Chapter 3

Genomic symbols

3.1 Introduction

Many known languages are structured so that the used words are only few of the possible combinations of characters; moreover a word, when used in a text, is often present many times. Looking at the genome like an usual language we may ask ourself which the "words" are. In order to find the equivalent of words in the genomic context, our first step was to search for nucleotide sequences occurring many times. Aiming at this result, we used paralogous alignments and we managed them with graph theory concepts. Within the found words, that we call "genomic symbols", there are several well known sequences, like protein domains, but also previously uncharacterized sequences.

A local sequence alignment algorithm is able to determine similar regions between two sequences. When a single sequence is aligned with itself, the best local alignment is obviously the one in which the two aligned subsequences both correspond to the entire sequence. However it is also possible to find suboptimal alignments, i.e. pairs of distinct regions (inside the same original sequence) that share sequence similarity (see figure 3.1). If the suboptimal alignment score [141, 142] cannot be statistically given by chance, a biological reason should be probably claimed in order to explain the presence of the similar (duplicated) subsequences producing the suboptimal alignment.

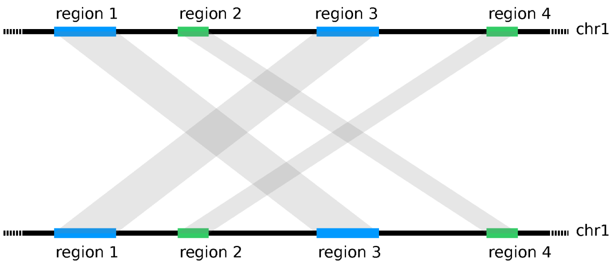


Figure 3.1: Local paralogous alignments. Inside the same sequence (chr1) there are four regions producing suboptimal alignments: the region 1 shares sequence similarity with region 3, the region 2 shares sequence similarity with the region 4. Are shown only the alignments which concern distinct regions.

These observations imply that it is possible to use paralogous alignments (i.e. the alignments of a genome with itself) in order to mine biologically significant subsequences: not every meaningful sequences in a genome are duplicated but duplicated sequences need a reason to be duplicated.

The so called "repeats" are a well known class of duplicated sequences, whose function is not well understood; a possibility is that repeats do not have any useful function for the organism hosting them. Good hypotheses about the mechanism of duplication for some subclasses of repeats are discussed in the literature (more details in section 1.5). The protein domains are other well known classes of duplicated genomic subsequences. Precise biological functions are often associated with protein domains. Both these two kinds of sequences (repeats and protein domains) have a biological reason to be duplicated, but they are duplicated according to two very different biological reasons. The repeats are sequences whose principal function is the duplication of themselves and they have a peculiar duplication mechanism. Instead there is not a mechanism that duplicates and inserts the sequence of a protein domain in several genomic locations. Most probably the proliferation of these domains is due to random events that duplicate entire genes or entire genomic regions. When such a duplication occurs a protein domain in the new genomic sequence may maintain a detectable sequence similarity with the ancestor, while the surrounding sequences drift neutrally; this happens when a duplicated gene that host the domain is still functional: therefore there is a selective pressure on the domain in order to maintain its sequence unchanged.

In this introduction we discussed the protein domains because they are the most famous duplicated functional sequences example. However our tool is not particularly suitable to find those protein domains whose function depends strictly from the 3D protein structure (and only indirectly from the nucleotide sequence). We are interested mainly in studing symbols with a regulatory role which occur in the non coding portion of genes or in intronic and intergenic regions.

3.2 Results and Discussion

At the end of our analysis we found 234 repeated sequences that we call genomic symbols (further details about symbols definition and retrival in section 3.4). Each symbol occurs many times in the human genome, with an average of about 100 occurrences for each symbol. Taking into account the characteristics of the genomic region in which a symbol preferentially occurs we could collect some information in order to establish if the given symbol is functional or not and which is its possible function.

annotation category	number of symbols	n. of symbols with GO assoc.
C	33	18
5	5	2
3	9	4
E	48	11
I	37	25
U	38	10
N	15	2

Table 3.1: For each annotation label (see section 3.4.3) the table reports the number of symbols that fall preferentially in this category (a single symbol may occur preferentially in more than one category) and the number of such symbols that are also associated with a Gene Ontology term.

We associated to each symbol occurrence one or more of the following labels:

- coding exon
- 5'UTR exon
- 3'UTR exon
- non-coding exon
- intron
- upstream (15 Kbps from the TSS)
- intergenic

(see section 3.4.3 for details about the label association with the occurrences). We observed that 173 symbols occur preferentially in a specific category, for example 9 of them occur preferentially in 3'UTR exons of genes and this is a strong evidence of some kind of function (see table 3.1 for the full report).

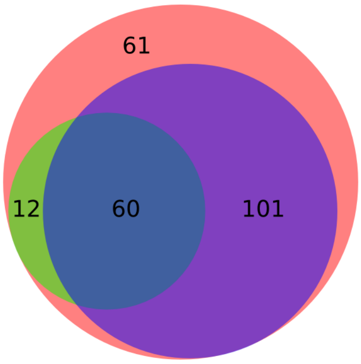


Figure 3.2: We found 234 symbols some of them (72, green set) have a Gene Ontology association and some (161, violet set) have other annotations anomalies (for example occours preferentially in UTR exons).

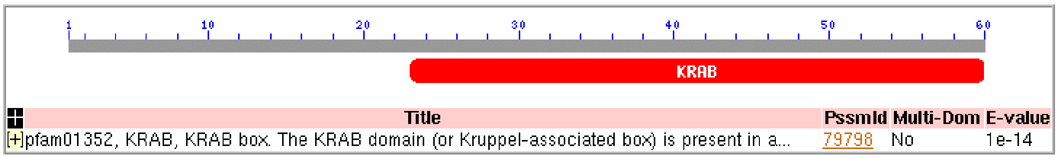
We associated to each symbol the set of genes in which the symbol occurs (see section 3.4.3 for details). Given a set of genes associated with a symbol we can look for overrepresented Gene Ontology [143] terms. To assess the statistical relevance of a symbol/term association we used the hypergeometrical model and we applied standard Bonferroni correction for multiple testing. In this way we can attribute a specific function to 72 symbols (see figure 3.2). The table 3.2 shows the GO terms associated to some symbols.

The KRAB domain example

There are biological meaningful repeated sequences that we do not expect to find; among all, those protein domains whose function depends strictly on the 3D structure. In this case there is a selective pressure to maintain the tertiary protein structure and not directly the primary structure, moreover we expect synonimous changes at the nucleotide level (for example in the third codon positions). Nevertheless we found 33 symbols that occour preferentially in the coding part of genes. An hypothesis may be that those protein coding symbols (besides some function depeding on the protein structure) have a further function that requires the conservation of the precise nucleotide sequence.

A typical example is given by the symbols $c1239$ (see section B.5 for the sequence) which occurs preferentially in the coding part of zinc finger proteins. Using the procedure described in the section 3.4.3 we associated this symbol with the KRAB domain (see figure 3.3).

The Krueppel-associated box (KRAB) is a domain of around 75 amino acids that is found in the N-terminal part of about one third of eukaryotic Krueppel-type C2H2[1] zinc finger proteins [145]. It is enriched in charged amino acids and can be divided into subregions A and B, which are predicted to fold into two amphipathic alpha-helices. The KRAB A and B boxes can be separated by variable spacer segments and many KRAB proteins contain only the A box [146] (further details in figure 3.5).



⊞ Title	PssmId	Multi-Dom	E-value
[+]pfam01352, KRAB, KRAB box. The KRAB domain (or Kruppel-associated box) is present in a...	79798	No	1e-14

Figure 3.3: The symbol c1239 corresponds to the KRAB protein domain.

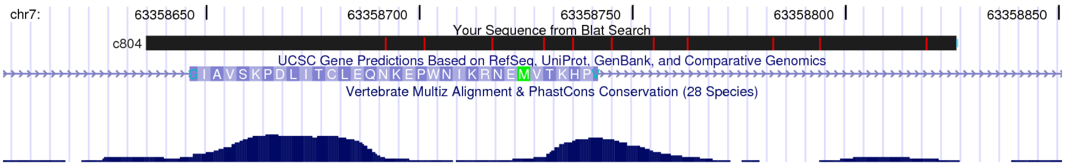

chr7:
63358650
63358700
63358750
63358800
63358850
Your Sequence from Blat Search
c804
UCSC Gene Predictions Based on RefSeq, UniProt, GenBank, and Comparative Genomics
G I A V S K P D L I T C L E Q N K E P W N I K R N E M V T K H P V
Vertebrate Multiz Alignment & PhastCons Conservation (28 Species)

Figure 3.4: The symbol c804 comprises an intronic portion.

Surprisingly we found another symbols (c804) associated with the KRAB domain, this symbol shows no nucleotide sequence similarity with the previous symbols c1239. If those symbols have a double function, it may be that the c804 proteic structural function is the same of the c1239 one, while the functions based on the nucleotide level are different.

The symbols c804 is longer than the KRAB domains and contains an intronic region (see figure 3.4).

There are other symbols (c1359, c690, c417, c307) associated with zinc finger proteins that occur both in the coding regions and in the introns of the genes. Among them only the symbol c417 corresponds to a known protein domain (DUF1220[2]).

[1] The C2H2 zinc finger is the classical zinc finger domain. The two conserved cysteines and histidines co-ordinate a zinc ion. The following pattern describes the zinc finger: $\sharp - X - C - X(1-5) - C - X3 - \sharp - X5 - \sharp - X2 - H - X(3-6) - [H/C]$ where X can be any amino acid, and numbers in brackets indicate the number of residues. The positions marked $\sharp$ are those that are important for the stable fold of the zinc finger. The final position can be either his or cys. The C2H2 zinc finger is composed of two short beta strands followed by an alpha helix. The amino terminal part of the helix binds the major groove in DNA binding zinc fingers. The accepted consensus binding sequence for Sp1 is usually defined by the asymmetric hexanucleotide core GGGCGG but this sequence does not include, among others, the GAG (=CTC) repeat

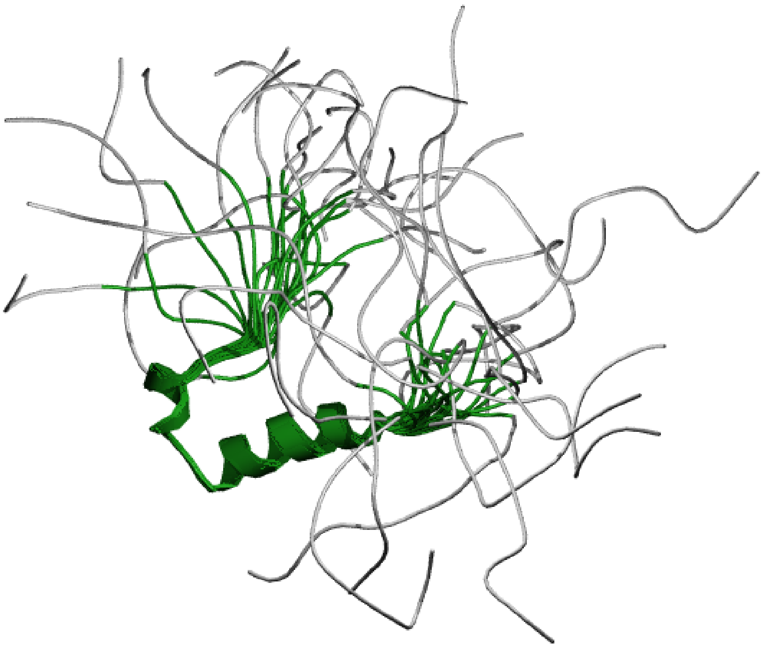


Figure 3.5: **Structure and function of the KRAB domain.** The functions currently known for members of the KRAB-containing protein family include transcriptional repression of RNA polymerase I, II, and III promoters, binding and splicing of RNA, and control of nucleolus function. The KRAB domain acts as a transcriptional repressor when tethered to the template DNA by a DNA-binding domain. A sequence of 45 amino acids in the KRAB A subdomain has been shown to be necessary and sufficient for transcriptional repression [146]. The B box does not repress by itself but does potentiate the repression exerted by the KRAB A subdomain. Gene silencing requires the binding of the KRAB domain to the RING-B box-coiled coil (RBCC) domain of the KAP-1/TIF1-beta corepressor [147, 148]. As KAP-1 binds to the heterochromatin proteins HP1, it has been proposed that the KRAB-ZFP-bound target gene could be silenced following recruitment to heterochromatin [146]. KRAB-ZFPs probably constitute the single largest class of transcription factors within the human genome. Although the function of KRAB-ZFPs is largely unknown, they appear to play important roles during development, cell differentiation, proliferation, apoptosis and neoplastic transformation [149].

number of symbols	Gene Ontology term association
31	regulation of transcription, DNA-dependent
6	positive regulation of translational initiation
5	gonadal mesoderm development
4	response to stimulus
3	organismal physiological process
2	ectoderm development
2	epidermis development
2	immune response
1	RNA processing
1	biopolymer metabolism
1	protein folding
1	G-protein coupled receptor protein signaling pathway
1	sensory perception of smell
1	fertilization (sensu Metazoa)

Table 3.2: The table reports the number of symbols associated to each relevant Gene Ontology term in the biological process ontology.

3.3 Conclusions

We found 234 genomic symbols, we have some statistical evidence of function for 173 of them. Genomes most probably contain many other nucleotide “words” that we are actually not able to find. Although we probably lost many interesting symbols, with our work we demonstrated that paralogous alignments are useful to extract previously unknown information form raw genomic sequences.

3.4 Methods

The pipeline used in this work is divided in four principal tasks.

Genomic alignments: in which a genome wide local paralogous alignment database is built and regions with high coverage (see later) are identified.

Symbol retrieval: in which the high coverage regions with similar sequences are clustered and the consensus sequence (the symbol) for each cluster is built.

Symbols remapping: in which we collected all the genomic occurrences of each symbol.

Symbols analysis: in which, for each symbol, various statistical observation are made and those symbols that seem to behave non randomly are identified.

In the following we describe in details each step of the pipeline.

3.4.1 Genomic alignments

Genome Sequence

We retrieved the complete, repeat masked, human genomic sequence from the ENSEMBL database (version 40) [42]. We used repeat masked sequences because we aren’t interested in retrieving already known repeats.

Genome Alignments

We performed a local alignment search in the entire genome using the Washington University implementation of BLAST [44, 150] with the PAM10 [151] scoring scheme. We retained each sobouptimal alignment longer than 30 base pairs. To chose this threshold we reasoned that the number of randomly occurring alignment is exponentially decreasing with the length (the longer is an alignment the less probability there is to find it by chance). Looking at the local alignments length distribution in the human genome, we have seen that this distribution is exponentially decreasing up to 30 base pairs. Then the distribution slope becomes weaker (approximately a power law): this means that the majority of the alignments that are shorter than 30 base pairs probably arise by chance. Instead the fraction of random alignments becomes negligible for alignments longer than this threshold. This is the first filter that we used to assess statistical significance of our results.

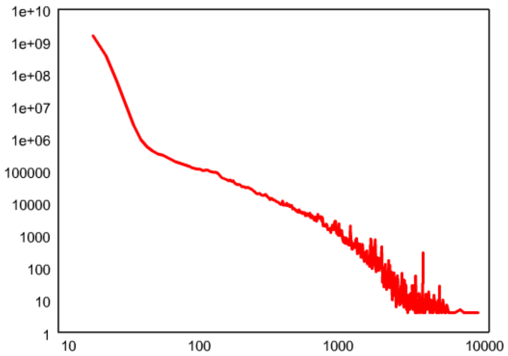


Figure 3.7: Alignment length distribution. The x-axis shows the alignment length, the y-axis shows the number of alignments with the given length.

that constitutes a high-affinity site for Sp1 binding to the wt1 promoter [144]

[2] The DUF1220 is a mammalian specific repeat of around 65 residues in length and is found in multiple copies in several human proteins. The function of this domain is unknown.

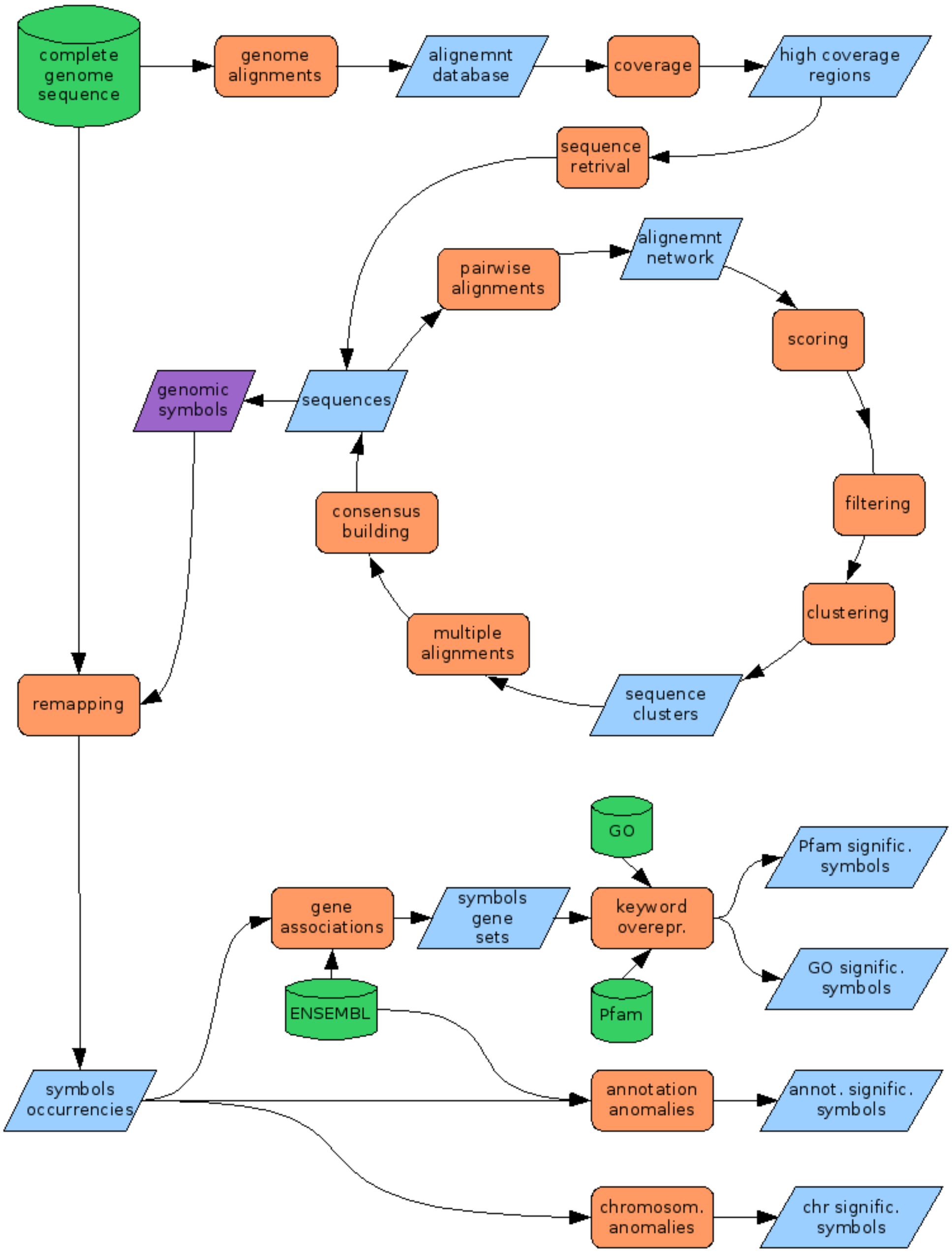


Figure 3.6: Genomic symbols pipeline.

Alignment database

We stored the computed alignments in an optimized data structure based on location clusters (see section 2.4) in order to speedup the following steps of the pipeline.

High coverage regions

An alignment in our framework may be seen as a couple of genomic regions (A, B) characterized by a detectable sequence similarity.

The coverage is a function associating each base in a genome to the number of alignments which overlap such base (see figure 3.8).

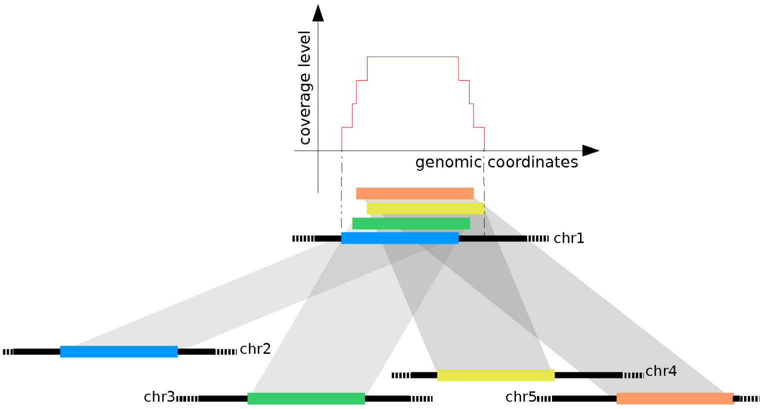


Figure 3.8: The alignment coverage level counts the number of alignments covering each base of a given region.

Suppose that a symbol s is present in three distinct places A, B and C in the genome; each occurrence of the symbol may have a slightly different sequence (for example due to mutations that do not influence the symbol functionality). Suppose also that the alignment algorithm could detect the sequence similarity between the three instances of s; in this case among the alignments present in the alignment database, there were 3 alignments: (A, B), (A, C), (B, C). We filtered out auto-alignments like (A, A) and we taken only one of the two symmetric alignments like (A, B) and (B, A).

If we have n distinct occurrences of a symbol s in the genome we find $\frac{n(n-1)}{2}$ alignments in the alignment database and each occurrence is involved in $n - 1$ alignments. So the coverage of each symbol occurrence is $n - 1$ (for each nucleotide of the occurrence).

To improve statistical significance of the observations we concentrated our attention on symbols with 50 occurrences or more (see figure 3.9).

In the simplified perspective exemplified before (symbol s with n occurrences), scanning the genome one could expect to see a coverage function in general equal to 0 and with instantaneous jumps to a value of $n-1$ at the edges of s occurrences. Actually this is not the case: the coverage landscape is complex and in particular the peaks are often “fuzzy“, i.e. there are not sharp jumps from 0 to a great value and viceversa but a smoother curve. Nevertheless we defined an initial set of putative symbol occurrences selecting each genomic region whose coverage is greater than 50 for the entire length of the region.

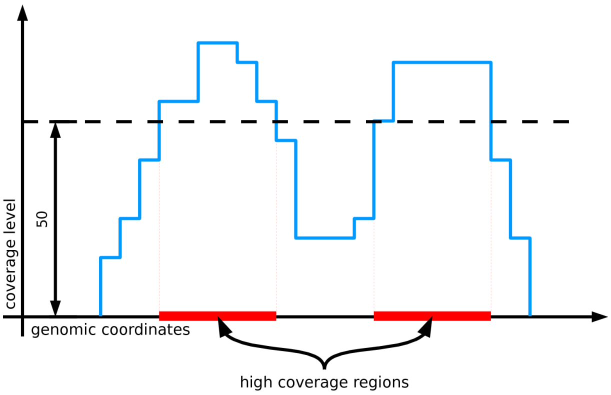


Figure 3.9: We require a minimum coverage level of 50.

Before proceeding to the next step of the pipeline some further considerations about the coverage cutoff are needed.

We are interested in two types of duplicated sequences: those that have a peculiar (maybe unknown) mechanism of duplication (like known repeats), and those that have a peculiar biological function (like protein domains). During the evolution large fractions of the genome have been duplicated, therefore it may happen that we find something that looks like a genomic symbol occurrence but it is actually only a non-functional region (nested inside a big duplicated genomic stretch) that didn't accumulate sufficient mutation to lose sequence similarity with the original region. The request of a minimum coverage level of 50 allows us to exclude this type of events: dozen of duplications of the same stretch should be required and all of them should be so recent that a random non-functional region should show sequence similarity with all the paralogous.

3.4.2 Symbol retrieval

The symbol retrieval is an iterative process: starting from a set of putative symbols they are refined by a circular pipeline. The output of this refinement process is then used as input for the next iteration of the pipeline. The same refinement process is made until the output of a cycle is the same of the previous one.

The initial input is the set of all the putative symbols occurrences (genomic regions with coverage greater than 50) i.e. we associate each occurrence with an initial putative distinct symbol.

Alignment network

The alignment network is a graph in which each sequence (i.e. symbol) is a vertex. Each possible pair of sequences is aligned: when a good sequence similarity is found we put a link between the two associated vertices.

In this context we use a score s_{ab} as a measure of sequence similarity between two sequences a and b. The score is given by the formula:

$$s_{ab} = \frac{\max_{\alpha \in \mathcal{A}}(l_\alpha)}{\max(l_a, l_b)}$$

where $\mathcal{A}$ is the set of all the local alignments between a and b, α is an arbitrary local alignment in $\mathcal{A}$, l_α is the length of the alignment, l_a and l_b are the lengths of the sequences. In other words s_{ab} is the length of the best local alignment between the two sequences divided by the length of the longest sequence.

We put a link between two vertices i and j if s_{ij} is greater than 0.9 .

Clustering

Given three symbol occurrences A, B and C if A is linked to B and B is linked to C not necessarily A is linked to C: the linkage propriety is not transitive. Nevertheless, since we use a stringent cutoff, the transitive property is "almost always" verified. If this property would be always valid, the network would be a set of distinct cliques. Instead the real network appears to be made of many connected components each one composed by different "communities". A community is a subset of nodes within which the node-node connections are dense, and the edges among communities are less dense. To separate each community we applied a clustering algorithm that maximize the modularity [152] (the modularity is the number of edges falling within groups minus the expected number in an equivalent network with edges placed at random).

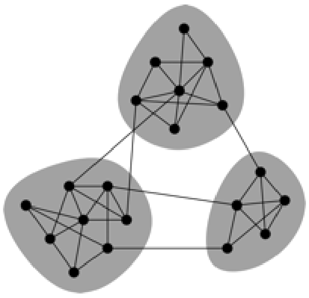


Figure 3.10: Small network with strong community structure: the network breaks into three distinct communities.

At each cycle of the pipeline each community is associated with a distinct putative symbol.

Multialignments and patterns

A community is a set of sequences. In order to associate to each community a single sequence, we performed a multiple alignment (using ClustalW with the standard scoring scheme [153]) of the sequences which belong to a community. Then we postprocessed the results in order to find stretch of highly conserved nucleotides that we call "patterns".

For each community it is possible to find zero, one or more than one patterns. The simplest situation is when there is only one pattern: in this case its sequence is the sequence of the symbol associated with the community. When it was not possible to find any pattern, we discarded the entire community: this way we lose the putative symbol whose occurrences share a low level of sequence similarity. When we found two or more distinct patterns it means that the community contains the occurrences of more than one symbol and we reported each pattern as a distinct symbol.

At the end of this process we have a set of putative symbols with their own sequences.

This output is given as input for a new cycle of the pipeline. This iterative process continue until there is no difference between the input symbol set and the output symbol set of a cycle.

3.4.3 Symbols analysis

Remapping

When a symbol set is defined, we remap it into the genome, looking for all the statistically significant alignments (found with the Washington University implementation of BLAST [44, 150]). We taken each of such alignment as a symbol occurrence.

Gene associations

We associated each symbol with a set of genes: a gene is associated with a symbol if a symbol occurrence falls inside the gene (in the exonic or intronic portion) or in the intergenic gene upstream or downstream region. We use gene annotations downloaded from the ENSEMBL database version 40 [42].

Gene Ontology analysis

Given a set of genes associated with a symbol we can look for overrepresented Gene Ontology [143] keywords.

To assess the statistical relevance of a symbol/keyword association we used the hypergeometrical model and we apply standard Bonferroni correction for multiple testing.

For each symbol σ and associated gene set $S(\sigma)$ we computed the prevalence of all Gene Ontology (GO) terms among the annotated genes in the set, and the probability that such prevalence would occur in a randomly chosen genes set of the same size. We always consider a gene annotated to a GO term if it is directly annotated to it or to any of its descendants in the GO graph. For a given GO term t let $M(t)$ be the total number of genes annotated to t in the genome, and $x(\sigma, t)$ the number of genes annotated to t in the set $S(\sigma)$. If N and $n(\sigma)$ denote the number of genes in the genome and in $S(\sigma)$ respectively, such probability is given by the right tail of the appropriate hypergeometric distribution [154]:

$$P(N, M(t), n(\sigma), x(\sigma, t)) = \sum_{h=x(\sigma,t)}^{\min(n(\sigma),M(t))} F(N, M(t), n(\sigma), h) \tag{3.1}$$

where

$$F(N, M, n, x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \tag{3.2}$$

In this way a P-value can be associated with each pair made of a symbol and a Gene Ontology term. A low P-value indicates that the symbol is correlated to the functional characterization described by the GO term.

The number B of indipendent statistical tests computed is $B = TS$ where T is the number of terms in Gene Ontology and S is the total number of symbols.

We apply a standard Bonferroni correction to P-values and we consider as significant only those P-value P so that $P < \frac{0.1}{B}$.

Annotation analysis

In order to identify symbols which show an uneven distribution in the genes (for instance those which always occur in the coding exons or in the UTRs) we reported for each symbol occurrence an annotation vector with seven entries corresponding to the seven possible annotation labels:

1. coding exon
2. non-coding exon
3. 5'UTR exon
4. 3'UTR exon
5. intron
6. upstream
7. intergenic

The category of non-coding exons includes all exons sequences that do not code for protein portions and are not marked as 5' or 3' UTRs (for example belonging to RNA genes); this category also comprises the pseudogenes. We get all these annotations from ENSEMBL version 40 [42] with the only exception of upstreams: we define an "upstream" as the region ranging from 15Kbps upstream of an annotated translation start site (TSS) to the TSS itself. Obviously a single nucleotide can belong to different categories, for example an exon of a gene can fall inside an intron of another gene (see section 1.2.3). In such cases we reported, for each nucleotide, only the "strongest" category; we assumed the strength of a category as indicated in the previous list (the coding exon being the strongest).

We associated each symbol occurrence with a 7 entries vector $\overrightarrow{v}$. We reported in each entry of $\overrightarrow{v}$ the fraction of nucleotides of the occurrence that belong to the corresponding annotation category of the previous list. We associated also a boolean vector $\overrightarrow{b}$ characterized by the following formula:

$$b_i = \begin{cases} 1 & \text{if } v_i \geq 0.15 \\ 0 & \text{otherwise} \end{cases}$$

The index i (with $1 \leq i \leq 7$) indicates an annotations category in the previous list.

For each symbol and each entries b_i we associated a P-value with a procedure similar to that described in the previous section (hypergeometrical model and Bonferroni correction).

A low P-value associated with a symbol and an b_i indicates that the symbol tends to occur preferentially in the genomic region indicated by i.

Protein domain analysis

In order to associate a protein domain with a symbol we searched the best alignment of the symbol in the database "Non-redundant protein sequences" (nr) using blastx at the NCBI [155].

Then we take the proteinic sequence that produces the best high scoring pair. Then, using this sequence as a query, we searched for a domain in Pfam [156] using the "Conserved Domain Database Search Service" at NCBI [157].

In this task all programs are used with standard parameters and the database versions are the latest published online at november 2007.

Chapter 4

Conclusions

The interpretation in biological significant terms of the genomic sequence of an organism is one of the most ambitious challenges of the post-genomic era. Even the more studied genomic structures, like genes, appear to be much more complex than expected and are still not completely understood.

In this thesis, after a short introduction on genomic sequences (chapter 1) and on our present understanding of the notion of gene (in particular on the implications of the results of the recent ENCODE and FANTOM projects), I discussed two main issues.

First I described a novel methodology to identify genes based on paralogous alignements (i.e. the alignments of a genome with itself) (chapter 2). This tool is not a general purpose gene predictor because it can only find genes with at least one processed pseudogene. Nevertheless our methodology provides novel information which can be integrated with those provided by other methodologies, in order to develop gene predictions pipelines. Our tool is completely unbiased because it does not require other information but the genomic sequence. Instead almost all other gene predictors use some kind of prior knowledge about genes, this infomation can be in the form of sequences (like ESTs) or models (like the Hidden Markov Model used in GENESCAN). We found about 1000 genes in the human genome and 8 of them where previously unknown. We tested 3 of them experimentally and we validated 2 of them as transcribed. Besides this results we could also identify 75 new alternative transcripts which correspond to known human genes but contain previously unknown exons.

The second research project discussed in this thesis deals with the identification of functional sequences (especially in intronic and intergenic regions) based again on paralogous alignements. We used these alignments to find duplicated sequences in the genome. We also elaborated a set of filters in order to select duplications with a statistically significant degree of conservation (see chapter 3). We call the resulting sequences “genomic symbols” because we think that they can be used in the genome like words in a text and that they can be combined in order to construct phrases or more complex structures. We found 234 of such sequences and we have some statistical evidence of function for 173 of them.

Even if probably we lose many interesting symbols, with our work we demonstrated that paralogous alignments are useful to extract information form raw genomic sequences.

Appendix A

A new computational approach to analyze human protein complexes and predict novel protein interactions

We propose a new approach to identify interacting proteins based on gene expression data. By using hypergeometric distribution and extensive Monte-Carlo simulations, we demonstrated that looking at synchronous expression peaks (in single time interval) is a high sensitivity approach to detect co-regulation among interacting proteins. Combining gene expression and GO similarity analyses resulted striking to extract novel interactors from microarray datasets. Applying this approach to PAK1, we validated α-tubulin and EEA1 as its novel interactors.

The cell is a complex system composed by a heterogeneous and highly dynamic set of proteins whose ability to interact and form complexes is critical for cellular activity regulation [158]. Therefore, the complete identification of the interactome is one of the major goals to be achieved. Different high-throughput experimental approaches have been developed to characterize the interactome of several organisms. Up to now, data have been mostly generated by studying simple organisms such as S. cerevisiae, C. elegans and D. melanogaster [159, 160]. For human cells, published experimental results are collected in databases like MINT and HPRD [161, 162] but the amount of information is still quite limited. Moreover data have been obtained from different cellular models and using different techniques, thus rendering it difficult to build a global network of interactions or to extrapolate information about the composition of multiprotein complexes. Computational approaches may help to address these crucial issues [163, 164, 165, 166, 167, 168, 169, 170]. The current idea is that proteins forming a supra-molecular complex are simultaneously transcribed and standard Pearson's analysis has been extensively applied on gene expression datasets to support this concept [165, 167, 168, 170, 171]. In general good results are obtained with this method when applied to protein interactions of stable protein complexes, but it is less efficient in other cases [165, 167]. A paradigmatic example is the application of Pearson's analysis on gene expression datasets of the yeast cell-cycle. A strong and significant correlation can be obtained for permanent protein complexes, but only weak correlations are seen for the transient ones [167]. A similar conclusion resulted from the analysis of some human gene profiles [165]. We present a new approach for the detection of putative protein interactions based on expression data. Besides the identification of permanent complexes, it is also capable (at least for well synchronized samples) to reliably identify interactions among proteins belonging to transient complexes. This approach is based on two observations. Firstly, protein-protein interactions are more easily identified if the interacting protein pair belongs to a multi-protein complex. This is a direct consequence of the fact that the features used to identify the interactions (i.e. correlations in expression data) displayed a much higher signal to noise ratio if multiple correlations were simultaneously looked for. Therefore, we focused on tracking interactions within protein complexes, even though our algorithm can in principle identify any type of protein-protein interaction. The second observation is that while Pearson's correlators are very effective for permanent complexes, which are assembled in most of the experimental time-points, they were found less suitable for transient complexes which are assembled only in few or unique time-points. To overcome this problem we propose a new method to extract putative human interacting proteins from microarray gene expression data looking at the presence of synchronous expression peaks in time course experiments of synchronized HeLa cells [172]. This is further supported by the recent observation in yeast that the timing of transcription during the cell-cycle is indicative of the timing of protein complex assembly [173]. This approach allowed us to address interactions characterized by low but not negligible statistical significance, which would instead be completely filtered out in the Pearson-based analysis. To further enhance the signal to noise ratio we combined this analytical procedure with a standard Gene Ontology [143] search. This filter turns out to be very effective, since it is based on input information completely independent from those exploited in the previous analysis step. To test the performances of our approach and compare it with the standard Pearson-based one, we established and tested a set of 32 permanent and transient complexes. The application of our method shows its effectiveness in detecting protein interactions in permanent and transient complexes. We also observed that, as expected, the proposed technique performs better as the synchronization of the dataset improves. To specifically test the applicability of our method in a precise biological context we applied it to explore novel putative interacting partners for the serine/threonine p21-activated kinase 1 (PAK1). PAK1 is a kinase downstream the Rho family of small GTPases, which participates in the formation of several dynamic and transient transductosomes [174]. We also provide experimental evidences confirming the interactions predicted by our algorithm between PAK1 and α-tubulin as well as PAK1 and early endosome antigen 1 (EEA1), a coiled coil dimer that is crucial for endosome fusion in vitro [175].

Appendix B

Supplementary data

B.1 Gene Ontology association of retrotransposed genes

Table B.1: Each GO terms with a significative p-value is reported, N is the number of genes associated with the given GO terms, M is the number of such genes that have also a more specific significative annotation.

GO key	N	M	P-value	ontology	description
GO:0051169	15	2	4.07e-03	biological process	nuclear transport
GO:0008104	62	60	4.89e-05	biological process	protein localization
GO:0051188	13	0	9.13e-03	biological process	cofactor biosynthesis
GO:0006260	19	0	4.93e-03	biological process	DNA replication
GO:0006177	2	0	8.30e-03	biological process	GMP biosynthesis
GO:0043285	34	25	9.77e-05	biological process	biopolymer catabolism
GO:0006445	21	9	1.66e-07	biological process	regulation of translation
GO:0006446	8	0	1.29e-03	biological process	regulation of translational initiation
GO:0006605	26	2	3.01e-05	biological process	protein targeting
GO:0030049	2	0	2.87e-03	biological process	muscle filament sliding
GO:0030163	25	20	1.44e-03	biological process	protein catabolism
GO:0006839	6	0	1.07e-03	biological process	mitochondrial transport
GO:0000398	12	0	6.74e-04	biological process	nuclear mRNA splicing, via spliceosome
GO:0016567	10	0	4.37e-03	biological process	protein ubiquitination
GO:0006259	54	20	9.96e-03	biological process	DNA metabolism
GO:0000050	3	0	7.00e-03	biological process	urea cycle
GO:0043037	34	31	1.09e-09	biological process	translation
GO:0006457	33	0	4.72e-06	biological process	protein folding
GO:0009060	11	9	4.46e-07	biological process	aerobic respiration
GO:0009064	7	0	8.32e-03	biological process	glutamine family amino acid metabolism
GO:0007067	23	6	5.61e-05	biological process	mitosis
GO:0044267	259	206	1.45e-17	biological process	cellular protein metabolism
GO:0044265	41	35	3.66e-06	biological process	cellular macromolecule catabolism
GO:0006888	10	0	3.94e-03	biological process	ER to Golgi vesicle-mediated transport
GO:0044260	261	259	3.22e-17	biological process	cellular macromolecule metabolism
GO:0006886	41	26	2.57e-06	biological process	intracellular protein transport
GO:0006880	2	0	2.87e-03	biological process	intracellular sequestering of iron ion
GO:0042254	17	15	1.52e-06	biological process	ribosome biogenesis and assembly
GO:0016043	156	147	3.44e-08	biological process	cell organization and biogenesis
GO:0006360	4	3	1.99e-03	biological process	transcription from RNA polymerase I promoter
GO:0042255	3	0	7.00e-03	biological process	ribosome assembly
GO:0007183	2	1	2.87e-03	biological process	SMAD protein heteromerization
GO:0007181	2	1	8.30e-03	biological process	transforming growth factor beta receptor complex assembly
GO:0007184	2	1	2.87e-03	biological process	SMAD protein nuclear translocation
GO:0019318	16	14	6.59e-03	biological process	hexose metabolism
GO:0006810	154	93	5.52e-03	biological process	transport
GO:0009889	25	23	1.41e-04	biological process	regulation of biosynthesis
GO:0009152	10	2	8.51e-03	biological process	purine ribonucleotide biosynthesis
GO:0009150	11	10	5.05e-03	biological process	purine ribonucleotide metabolism
GO:0009260	12	10	2.04e-03	biological process	ribonucleotide biosynthesis
GO:0008152	563	560	1.32e-22	biological process	metabolism
GO:0007059	9	6	1.60e-03	biological process	chromosome segregation
GO:0016070	67	61	1.32e-11	biological process	RNA metabolism
GO:0016071	39	33	4.52e-08	biological process	mRNA metabolism
GO:0050658	11	9	4.65e-04	biological process	RNA transport
GO:0015980	26	23	6.95e-07	biological process	energy derivation by oxidation of organic compounds
GO:0046907	77	64	1.36e-10	biological process	intracellular transport
GO:0006512	49	26	8.75e-06	biological process	ubiquitin cycle
GO:0006104	2	0	8.30e-03	biological process	succinyl-CoA metabolism

Continued on Next Page...

Table B.1 – Continued

GO key	N	M	P-value	ontology	description
GO:0006356	3	0	5.89e-04	biological process	regulation of transcription from RNA polymerase I promoter
GO:0006100	7	2	1.04e-04	biological process	tricarboxylic acid cycle intermediate metabolism
GO:0006417	23	21	1.57e-04	biological process	regulation of protein biosynthesis
GO:0006732	23	10	2.39e-05	biological process	coenzyme metabolism
GO:0030261	4	0	8.86e-03	biological process	chromosome condensation
GO:0007046	13	0	1.02e-04	biological process	ribosome biogenesis
GO:0050875	707	689	3.73e-29	biological process	cellular physiological process
GO:0009259	14	13	4.60e-04	biological process	ribonucleotide metabolism
GO:0007049	78	63	3.15e-07	biological process	cell cycle
GO:0000723	5	0	7.91e-03	biological process	telomere maintenance
GO:0006139	235	135	1.11e-04	biological process	nucleobase, nucleoside, nucleotide and nucleic acid metabolism
GO:0043161	6	0	2.12e-03	biological process	proteasomal ubiquitin-dependent protein catabolism
GO:0051246	28	23	3.07e-03	biological process	regulation of protein metabolism
GO:0008380	35	12	3.16e-09	biological process	RNA splicing
GO:0006996	87	54	3.48e-05	biological process	organelle organization and biogenesis
GO:0009987	743	707	4.18e-11	biological process	cellular process
GO:0006084	10	9	2.07e-06	biological process	acetyl-CoA metabolism
GO:0000051	4	0	8.86e-03	biological process	urea cycle intermediate metabolism
GO:0006337	3	0	2.72e-03	biological process	nucleosome disassembly
GO:0043170	402	397	1.07e-33	biological process	macromolecule metabolism
GO:0006007	14	11	6.81e-06	biological process	glucose catabolism
GO:0046365	15	14	1.00e-05	biological process	monosaccharide catabolism
GO:0006099	9	0	1.44e-06	biological process	tricarboxylic acid cycle
GO:0006092	21	19	2.31e-07	biological process	main pathways of carbohydrate metabolism
GO:0006096	11	0	1.12e-04	biological process	glycolysis
GO:0051301	23	0	8.42e-04	biological process	cell division
GO:0006511	20	7	5.76e-04	biological process	ubiquitin-dependent protein catabolism
GO:0000279	28	23	5.33e-05	biological process	M phase
GO:0000278	32	23	1.57e-05	biological process	mitotic cell cycle
GO:0007582	727	720	2.35e-13	biological process	physiological process
GO:0006913	18	2	5.35e-04	biological process	nucleocytoplasmic transport
GO:0009058	176	167	1.73e-29	biological process	biosynthesis
GO:0009059	134	127	1.80e-33	biological process	macromolecule biosynthesis
GO:0044262	33	25	1.48e-03	biological process	cellular carbohydrate metabolism
GO:0009056	68	64	1.43e-07	biological process	catabolism
GO:0006497	7	2	8.32e-03	biological process	protein amino acid lipidation
GO:0044238	540	501	1.32e-26	biological process	primary metabolism
GO:0000070	6	0	1.71e-03	biological process	mitotic sister chromatid segregation
GO:0051276	35	12	2.99e-03	biological process	chromosome organization and biogenesis
GO:0051028	9	0	1.39e-03	biological process	mRNA transport
GO:0000074	44	0	1.30e-03	biological process	regulation of progression through cell cycle
GO:0018348	2	0	8.30e-03	biological process	protein amino acid geranylgeranylation
GO:0044237	543	508	1.32e-24	biological process	cellular metabolism
GO:0015031	60	41	2.61e-06	biological process	protein transport
GO:0006415	3	0	4.56e-03	biological process	translational termination
GO:0006414	6	0	1.71e-03	biological process	translational elongation
GO:0006413	15	8	1.92e-06	biological process	translational initiation
GO:0006412	127	43	1.95e-34	biological process	protein biosynthesis
GO:0006396	56	39	1.16e-10	biological process	RNA processing
GO:0006397	33	12	6.07e-07	biological process	mRNA processing
GO:0043231	474	412	3.82e-24	cellular component	intracellular membrane-bound organelle
GO:0005784	2	0	8.30e-03	cellular component	translocon complex
GO:0015934	20	16	1.15e-12	cellular component	large ribosomal subunit
GO:0015935	29	26	1.16e-20	cellular component	small ribosomal subunit
GO:0030867	3	2	2.72e-03	cellular component	rough endoplasmic reticulum membrane
GO:0043232	183	166	6.05e-21	cellular component	intracellular non-membrane-bound organelle
GO:0005789	15	3	1.86e-03	cellular component	endoplasmic reticulum membrane
GO:0008043	2	0	2.87e-03	cellular component	ferritin complex
GO:0005832	4	0	1.13e-04	cellular component	chaperonin-containing T-complex
GO:0000228	12	9	5.11e-03	cellular component	nuclear chromosome
GO:0005737	409	317	1.26e-44	cellular component	cytoplasm
GO:0005819	9	0	6.77e-03	cellular component	spindle
GO:0005838	4	0	4.81e-04	cellular component	proteasome regulatory particle (sensu Eukaryota)
GO:0031090	65	48	6.70e-07	cellular component	organelle membrane
GO:0031981	54	52	3.92e-06	cellular component	nuclear lumen
GO:0005730	18	0	5.84e-04	cellular component	nucleolus
GO:0016282	31	26	6.56e-27	cellular component	eukaryotic 43S preinitiation complex
GO:0005634	299	100	1.61e-09	cellular component	nucleus
GO:0043233	70	67	3.15e-08	cellular component	organelle lumen
GO:0044444	315	232	2.08e-40	cellular component	cytoplasmic part
GO:0005739	105	45	2.04e-17	cellular component	mitochondrion
GO:0030529	111	99	1.50e-41	cellular component	ribonucleoprotein complex
GO:0005654	38	3	3.49e-04	cellular component	nucleoplasm
GO:0044432	17	15	1.23e-03	cellular component	endoplasmic reticulum part
GO:0044430	47	11	9.47e-03	cellular component	cytoskeletal part
GO:0005759	13	7	7.02e-04	cellular component	mitochondrial matrix
GO:0005694	40	36	8.76e-05	cellular component	chromosome
GO:0015630	30	9	3.77e-03	cellular component	microtubule cytoskeleton
GO:0031967	43	34	6.89e-06	cellular component	organelle envelope
GO:0031966	33	29	7.34e-07	cellular component	mitochondrial membrane

Continued on Next Page...

Table B.1 – Continued

GO key	N	M	P-value	ontology	description
GO:0012505	37	15	2.23e-03	cellular component	endomembrane system
GO:0000502	11	4	4.36e-05	cellular component	proteasome complex (sensu Eukaryota)
GO:0001674	2	0	8.30e-03	cellular component	female germ cell nucleus
GO:0005862	2	0	8.30e-03	cellular component	muscle thin filament tropomyosin
GO:0030530	4	0	6.93e-03	cellular component	heterogeneous nuclear ribonucleoprotein complex
GO:0043234	247	141	1.03e-35	cellular component	protein complex
GO:0005840	75	52	6.10e-37	cellular component	ribosome
GO:0005843	26	0	3.04e-25	cellular component	cytosolic small ribosomal subunit (sensu Eukaryota)
GO:0005842	16	0	1.29e-14	cellular component	cytosolic large ribosomal subunit (sensu Eukaryota)
GO:0005829	80	50	2.48e-22	cellular component	cytosol
GO:0044446	269	268	2.68e-32	cellular component	intracellular organelle part
GO:0005743	29	0	2.17e-06	cellular component	mitochondrial inner membrane
GO:0005740	34	33	1.89e-06	cellular component	mitochondrial envelope
GO:0044445	50	46	6.05e-32	cellular component	cytosolic part
GO:0044428	97	84	1.26e-11	cellular component	nuclear part
GO:0005761	7	0	9.41e-03	cellular component	mitochondrial ribosome
GO:0005666	3	0	7.00e-03	cellular component	DNA-directed RNA polymerase III complex
GO:0044427	33	9	1.01e-03	cellular component	chromosomal part
GO:0016281	4	0	8.30e-04	cellular component	eukaryotic translation initiation factor 4F complex
GO:0005681	22	0	1.66e-06	cellular component	spliceosome complex
GO:0044454	9	0	8.28e-03	cellular component	nuclear chromosome part
GO:0005200	11	0	9.96e-03	molecular function	structural constituent of cytoskeleton
GO:0008143	3	0	4.56e-03	molecular function	poly(A) binding
GO:0030508	3	0	2.72e-03	molecular function	thiol-disulfide exchange intermediate activity
GO:0048487	3	0	7.00e-03	molecular function	beta-tubulin binding
GO:0005488	656	561	5.27e-07	molecular function	binding
GO:0003676	251	137	6.94e-11	molecular function	nucleic acid binding
GO:0005485	2	0	8.30e-03	molecular function	v-SNARE activity
GO:0008168	14	0	6.10e-03	molecular function	methyltransferase activity
GO:0005198	97	84	4.52e-17	molecular function	structural molecule activity
GO:0005504	5	0	9.46e-03	molecular function	fatty acid binding
GO:0043566	10	9	3.18e-03	molecular function	structure-specific DNA binding
GO:0004659	4	3	5.30e-03	molecular function	prenyltransferase activity
GO:0016874	35	25	8.08e-05	molecular function	ligase activity
GO:0030911	2	0	2.87e-03	molecular function	TPR domain binding
GO:0051087	4	0	2.86e-03	molecular function	chaperone binding
GO:0051082	20	0	5.50e-06	molecular function	unfolded protein binding
GO:0000166	158	124	5.37e-07	molecular function	nucleotide binding
GO:0003735	73	0	6.66e-38	molecular function	structural constituent of ribosome
GO:0031202	6	0	8.26e-04	molecular function	RNA splicing factor activity, transesterification mechanism
GO:0004774	2	0	8.30e-03	molecular function	succinate-CoA ligase activity
GO:0017076	123	43	8.07e-04	molecular function	purine nucleotide binding
GO:0016491	55	37	7.50e-03	molecular function	oxidoreductase activity
GO:0003697	9	0	4.52e-04	molecular function	single-stranded DNA binding
GO:0005525	43	1	3.73e-07	molecular function	GTP binding
GO:0019144	2	0	2.87e-03	molecular function	ADP-sugar diphosphatase activity
GO:0015631	10	3	1.78e-03	molecular function	tubulin binding
GO:0016149	2	0	8.30e-03	molecular function	translation release factor activity, codon specific
GO:0008134	33	3	6.39e-03	molecular function	transcription factor binding
GO:0008135	24	22	1.99e-09	molecular function	translation factor activity, nucleic acid binding
GO:0003924	22	0	7.76e-05	molecular function	GTPase activity
GO:0017111	50	29	1.04e-04	molecular function	nucleoside-triphosphatase activity
GO:0004488	2	0	8.30e-03	molecular function	methylenetetrahydrofolate dehydrogenase (NADP+) activity
GO:0003824	317	162	1.49e-03	molecular function	catalytic activity
GO:0004004	7	0	1.04e-04	molecular function	ATP-dependent RNA helicase activity
GO:0045182	27	24	2.63e-10	molecular function	translation regulator activity
GO:0003865	2	0	8.30e-03	molecular function	3-oxo-5-alpha-steroid 4-dehydrogenase activity
GO:0047631	2	0	8.30e-03	molecular function	ADP-ribose diphosphatase activity
GO:0009055	20	0	3.20e-03	molecular function	electron carrier activity
GO:0043022	4	0	8.30e-04	molecular function	ribosome binding
GO:0016614	13	12	4.67e-03	molecular function	oxidoreductase activity, acting on CH-OH group of donors
GO:0016616	12	0	5.11e-03	molecular function	oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor
GO:0016879	22	0	1.16e-03	molecular function	ligase activity, forming carbon-nitrogen bonds
GO:0051920	3	0	2.72e-03	molecular function	peroxiredoxin activity
GO:0004477	2	0	8.30e-03	molecular function	methenyltetrahydrofolate cyclohydrolase activity
GO:0019843	11	0	7.30e-09	molecular function	rRNA binding
GO:0016462	55	52	1.17e-05	molecular function	pyrophosphatase activity
GO:0004662	2	0	2.87e-03	molecular function	CAAX-protein geranylgeranyltransferase activity
GO:0003723	115	21	1.74e-30	molecular function	RNA binding
GO:0005515	368	64	2.87e-05	molecular function	protein binding
GO:0004661	3	2	2.72e-03	molecular function	protein geranylgeranyltransferase activity
GO:0043014	2	0	8.30e-03	molecular function	alpha-tubulin binding
GO:0003746	5	0	6.55e-03	molecular function	translation elongation factor activity
GO:0051059	3	0	4.56e-03	molecular function	NF-kappaB binding
GO:0003743	15	0	1.03e-06	molecular function	translation initiation factor activity

B.2 Retrotrasposed genes (human)

Table B.2: Coordinates of found retrotransposed gene in the human genome (ENSEMBL version 40).

chr	start	stop	chr	start	stop	chr	start	stop	chr	start	stop
chr1	100468795	100487984	chr1	10382273	10402777	chr1	10995325	11006681			
chr1	111236517	111244055	chr1	111784403	111787540	chr1	111798354	111805472	chr1	112035483	112057251
chr1	112997743	113015757	chr1	114911837	114925710	chr1	115062156	115064757	chr1	11688940	11702712
chr1	117758862	117810569	chr1	144218974	144222800	chr1	1469112	1499792	chr1	148161834	148165811
chr1	148726985	148746370	chr1	148937545	148958637	chr1	149257565	149268016	chr1	149998762	150002222
chr1	150113159	150148570	chr1	151901134	151907658	chr1	152187598	152197639	chr1	152395466	152422287
chr1	152511723	152514975	chr1	153408509	153411993	chr1	153514174	153521707	chr1	153546455	153557080
chr1	153965551	153974937	chr1	154828211	154830537	chr1	158154524	158156952	chr1	158459089	158476425
chr1	159390399	159395267	chr1	159399331	159402116	chr1	159560026	159599607	chr1	160802435	160825227
chr1	165092371	165112096	chr1	166001444	166025131	chr1	167342563	167368582	chr1	173235784	173246030
chr1	174263333	174282084	chr1	174352386	174442834	chr1	17612157	17628282	chr1	176242584	176273724
chr1	177279472	177308295	chr1	180620053	180627548	chr1	181115072	181123500	chr1	183323307	183336812
chr1	19453861	19458058	chr1	201243523	201259136	chr1	209902738	209915586	chr1	210018882	210033139
chr1	220907980	220952436	chr1	222614820	222630210	chr1	22277507	22291472	chr1	224318652	224326106
chr1	229567401	229576468	chr1	231152988	231181118	chr1	23891685	23895498	chr1	24168185	24179499
chr1	24861737	24871338	chr1	27925188	27934742	chr1	28029910	28049640	chr1	28090638	28113267
chr1	28399657	28432123	chr1	31505276	31542197	chr1	31611091	31618499	chr1	31868059	31873711
chr1	32144609	32176482	chr1	32460733	32469760	chr1	32530324	32571813	chr1	33248111	33275064
chr1	36527850	36543076	chr1	3765171	3791849	chr1	37806469	37834023	chr1	37930911	37943755
chr1	39799094	39802002	chr1	40427312	40479179	chr1	40991411	41009857	chr1	41374088	41390926
chr1	42896680	42915012	chr1	42920685	42940600	chr1	45014292	45016996	chr1	45272570	45278798
chr1	45749310	45757314	chr1	45926491	45932691	chr1	47606726	47616126	chr1	52010348	52026187
chr1	52294459	52324431	chr1	52506782	52542185	chr1	52571026	52584892	chr1	53465195	53476778
chr1	54438782	54454594	chr1	54465188	54496413	chr1	54849867	54861770	chr1	6301144	6332488
chr1	6607827	6616208	chr1	67649548	67668666	chr1	68716079	68735407	chr1	71307547	71316824
chr1	76025786	76030537	chr1	84888674	84909579	chr1	92232302	92252566	chr1	93070193	93080069
chr1	9522115	9565175	chr1	9954663	9967452	chr10	100183125	100195161	chr10	102097842	102114570
chr10	102273490	102279608	chr10	103360412	103374560	chr10	103417633	103444695	chr10	104150412	104151260
chr10	104231744	104238885	chr10	105052563	105100820	chr10	105632444	105648748	chr10	105871914	105875510
chr10	1076850	1080113	chr10	112351745	112354380	chr10	118384392	118392749	chr10	120059252	120085923
chr10	120918680	120924093	chr10	12254769	12277893	chr10	124907233	124914845	chr10	126079429	126092995
chr10	126666891	126682053	chr10	13400074	13427035	chr10	14963503	14983253	chr10	27033614	27049383
chr10	27443455	27448291	chr10	33230398	33240529	chr10	42974187	43000743	chr10	5847199	5895311
chr10	59623353	59667582	chr10	59815183	59825728	chr10	6183273	6197669	chr10	69768338	69772956
chr10	70074453	70116448	chr10	70586768	70598372	chr10	71579964	71591689	chr10	71636026	71663154
chr10	73764276	73784673	chr10	74604451	74665211	chr10	7879025	7884369	chr10	79465115	79470465
chr10	82158284	82182729	chr10	88801110	88809010	chr10	90684961	90697125	chr10	96987324	97040708
chr10	97793309	97810610	chr10	99176008	99183187	chr10	99390772	99416316	chr11	100847140	100855016
chr11	106880235	106932933	chr11	109982591	110006690	chr11	111462807	111471710	chr11	113776638	113781731
chr11	116528367	116539824	chr11	118391759	118394262	chr11	122433411	122438054	chr11	124050342	124069723
chr11	124947565	124959781	chr11	125335614	125398642	chr11	13366120	13399963	chr11	13690050	13710431
chr11	14256047	14273979	chr11	14483001	14496102	chr11	16716801	16734014	chr11	17052516	17055369
chr11	18434846	18457093	chr11	30301265	30316350	chr11	32069118	32082855	chr11	33719875	33728760
chr11	3377001	3386785	chr11	36611408	36637375	chr11	47596842	47620594	chr11	50324789	50336409
chr11	57261656	57265018	chr11	58135016	58141310	chr11	59171837	59191013	chr11	59462466	59473066
chr11	59541399	59555141	chr11	60366146	60374996	chr11	61488614	61491682	chr11	62083650	62096791
chr11	62139347	62146032	chr11	63274041	63283916	chr11	63435304	63440605	chr11	63476455	63479162
chr11	63840770	63841586	chr11	65378884	65382224	chr11	65416271	65424435	chr11	65792655	65801537
chr11	65959887	65962815	chr11	67289421	67316547	chr11	69848154	69850567	chr11	71189207	71193336
chr11	73096113	73119517	chr11	73214398	73253289	chr11	74337954	74366427	chr11	74789384	74794381
chr11	74955005	74961444	chr11	75749648	75769593	chr11	77008299	77026517	chr11	7965654	7974292
chr11	82213058	82239197	chr11	82650150	82668949	chr11	8413639	8440036	chr11	8661324	8663967
chr11	8890574	8897631	chr11	89574295	89588048	chr11	95172074	95203927	chr11	9554608	9567888
chr12	100392384	100406442	chr12	100957777	100979942	chr12	10256766	10266966	chr12	103804245	103827661
chr12	10649879	10657435	chr12	108020694	108032575	chr12	108773132	108802658	chr12	109357091	109367740
chr12	111327374	111330845	chr12	115639938	115660053	chr12	117058291	117067770	chr12	119118895	119123019
chr12	122439930	122459850	chr12	122659179	122671432	chr12	14833255	14843932	chr12	15926719	15947670
chr12	21679537	21702011	chr12	23578126	23620042	chr12	25249450	25295056	chr12	26982628	27010844
chr12	2799496	2802339	chr12	40765990	40799236	chr12	4255430	4282565	chr12	4628539	4666633
chr12	47807837	47811445	chr12	48670223	48672399	chr12	51577242	51585109	chr12	51629172	51632961
chr12	51696521	51722258	chr12	52135934	52159819	chr12	52345222	52352700	chr12	52960777	52965543
chr12	54722475	54724762	chr12	54784704	54791360	chr12	54835104	54841623	chr12	54904913	54909904
chr12	54954837	54966076	chr12	55343403	55368295	chr12	55392482	55404574	chr12	55411251	55432338
chr12	55949433	55976596	chr12	56448689	56452158	chr12	61069668	61081322	chr12	63105862	63114630
chr12	6505555	6510937	chr12	6515918	6517795	chr12	67328744	67340070	chr12	6846959	6850250
chr12	6945862	6948986	chr12	74179698	74191664	chr12	74733252	74749041	chr12	75776626	75796930
chr12	7833298	7839862	chr12	7856380	7876679	chr12	7963095	7977767	chr12	8271596	8275775
chr12	8793724	8818264	chr12	892276	910751	chr12	8985304	8990225	chr12	97445811	97465987
chr13	113008835	113024026	chr13	114065512	114089379	chr13	18902609	18910290	chr13	24053803	24061484
chr13	24431871	24439593	chr13	26538293	26588730	chr13	26725896	26729315	chr13	29931994	29938097
chr13	40265346	40280733	chr13	40384650	40393874	chr13	44411383	44461606	chr13	44809021	44813286
chr13	47415192	47469161	chr13	49515648	49554167	chr13	51699157	51707414	chr13	51933042	51946029
chr13	52124976	52159988	chr13	97456413	97465888	chr13	97902881	97932578	chr14	101619655	101620995
chr14	19846254	19867373	chr14	20748559	20772233	chr14	22461490	22463114	chr14	22486658	22496120
chr14	22860652	22865208	chr14	23682425	23684198	chr14	30605105	30635284	chr14	31685218	31695324
chr14	34831414	34856424	chr14	38653272	38675869	chr14	49645103	49653005	chr14	50776738	50792507
chr14	57784220	57808429	chr14	59031498	59040532	chr14	67126044	67136735	chr14	67156473	67211081
chr14	72478209	72496093	chr14	74030221	74031815	chr14	74252407	74273170	chr14	88692287	88726548
chr14	89796208	89808713	chr14	92239914	92268916	chr14	92369340	92375900	chr14	96057093	96103177
chr14	99798422	99813565	chr15	22770579	22774822	chr15	23134732	23168054	chr15	32221080	32226785
chr15	36543613	36564167	chr15	38115536	38118654	chr15	39310822	39358382	chr15	40530631	40570598
chr15	40622076	40626984	chr15	41710109	41728258	chr15	41806589	41825679	chr15	42408188	42417807
chr15	46410925	46421858	chr15	48503893	48538632	chr15	50125319	50145746	chr15	53260813	53276427
chr15	55786207	55796494	chr15	57186842	57204535	chr15	61233109	61236739	chr15	62474649	62534520
chr15	64561145	64568668	chr15	64578712	64584238	chr15	67532211	67534937	chr15	70278428	70298507
chr15	71403253	71423198	chr15	72999672	73017490	chr15	73933782	73976456	chr15	76959907	76977126
chr15	78199724	78217757	chr15	80608216	80611596	chr15	81002559	81005939	chr15	98029246	98049055
chr15	99639238	99652939	chr16	11843058	11852828	chr16	15062857	15093971	chr16	18701782	18707942
chr16	18710519	18717658	chr16	1952056	1954630	chr16	19625350	19633797	chr16	21718470	21737582
chr16	21876251	21890938	chr16	23476771	23489679	chr16	23499874	23506135	chr16	23560289	23588683
chr16	28995479	29022985	chr16	48617810	48628490	chr16	5067909	5075045	chr16	52082793	52094717
chr16	55836642	55844916	chr16	57298540	57325669	chr16	66247870	66248957	chr16	66456503	66462724
chr16	671463	672744	chr16	67931415	67933266	chr16	68333286	68346444	chr16	72888278	72897023
chr16	79608437	79620384	chr16	79627051	79636280	chr16	79673432	79687455	chr16	82399131	82402773
chr16	84392309	84398109	chr16	85983518	85994474	chr16	87861543	87879349	chr16	8799201	8814456
chr16	88517310	88530000	chr17	10525113	10539874	chr17	11925398	11973329	chr17	1194621	1215104
chr17	15381031	15407557	chr17	18634946	18650343	chr17	20588983	20598117	chr17	20850804	20856821
chr17	24071850	24075500	chr17	2517037	2531991	chr17	25523684	25537584	chr17	26143587	26175787
chr17	27439295	27446065	chr17	28282604	28292773	chr17	31001663	31033935	chr17	33706643	33732362
chr17	34162543	34173965	chr17	34259894	34262886	chr17	34610990	34614508	chr17	36038533	36052337
chr17	37098666	37101403	chr17	37382805	37406115	chr17	37530541	37536134	chr17	37974018	37978257
chr17	38404286	38408490	chr17	39629823	39631010	chr17	39642253	39651235	chr17	39753565	39755446
chr17	42355496	42364564	chr17	43258999	43263900	chr17	43405883	43414118	chr17	44365566	44373426

Continued on Next Page...

Table B.2 – Continued

chr	start	stop	chr	start	stop	chr	start	stop	chr	start	stop
chr17	44836423	44845654	chr17	45507962	45509400	chr17	4799963	4801140	chr17	50393574	50401055
chr17	52370562	52393402	chr17	5276824	5283169	chr17	55386207	55395032	chr17	59259227	59262887
chr17	60176249	60185208	chr17	62767379	62793181	chr17	628645	642105	chr17	63463686	63473421
chr17	70675420	70690720	chr17	71354584	71362974	chr17	7153656	7156496	chr17	7156745	7159459
chr17	71592194	71611105	chr17	72065464	72071318	chr17	73676332	73680395	chr17	7418586	7421953
chr17	77091600	77093978	chr17	8288304	8304163	chr18	11883471	11898316	chr18	22120158	22161102
chr18	41939180	41958857	chr18	45268856	45271297	chr18	53367578	53391597	chr18	53418907	53424938
chr18	54737256	54772010	chr18	9092747	9124340	chr19	10362809	10375193	chr19	12768643	12773610
chr19	1389366	1391496	chr19	14069170	14078686	chr19	14380851	14385058	chr19	14535612	14536670
chr19	15326582	15345769	chr19	16048376	16074813	chr19	16522808	16544046	chr19	1771302	1776944
chr19	17833102	17835124	chr19	18503677	18513807	chr19	18545097	18547047	chr19	20897898	20928430
chr19	21116683	21159426	chr19	21266559	21304002	chr19	21444148	21449262	chr19	2194380	2199597
chr19	21949115	21985533	chr19	22155894	22190364	chr19	22261122	22288638	chr19	22733989	22758732
chr19	22806949	22832113	chr19	23091719	23122204	chr19	23195833	23224983	chr19	23337047	23370079
chr19	23729465	23733502	chr19	24061865	24103522	chr19	3045647	3074991	chr19	37767694	37770165
chr19	39377254	39410836	chr19	41325809	41329039	chr19	4314737	4317993	chr19	43557309	43566299
chr19	43801686	43819430	chr19	43900455	43910494	chr19	45017181	45021655	chr19	47480652	47490959
chr19	52325965	52365020	chr19	53810408	53812947	chr19	541121	568150	chr19	54160378	54161947
chr19	54684916	54687357	chr19	54691535	54694752	chr19	5641274	5642678	chr19	56517129	56525415
chr19	59396838	59403323	chr19	60589528	60591555	chr19	60604465	60610963	chr19	62354906	62370665
chr19	63590448	63597971	chr19	977635	990064	chr19	9807011	9821358	chr2	100549394	100559626
chr2	100985596	100989317	chr2	10180405	10184549	chr2	10499095	10502808	chr2	109660916	109728781
chr2	114390952	114431822	chr2	11503073	11511299	chr2	118295241	118305040	chr2	119841041	119846580
chr2	122201723	122210905	chr2	128321706	128332129	chr2	131821715	131832198	chr2	150134399	150151908
chr2	152367612	152379076	chr2	157040854	157078252	chr2	170168942	170202484	chr2	170363756	170389398
chr2	173936218	173941860	chr2	174647981	174655036	chr2	174796004	174821532	chr2	176902273	176914682
chr2	177788513	177793084	chr2	183515046	183554297	chr2	183701290	183734652	chr2	190817853	190834204
chr2	198059559	198071803	chr2	198073322	198076419	chr2	198089043	198125360	chr2	202779459	202811549
chr2	203453613	203484573	chr2	206732865	206735896	chr2	208809423	208828040	chr2	210575579	210589612
chr2	223481701	223507617	chr2	233129370	233142161	chr2	24047649	24076635	chr2	24144115	24152645
chr2	25310406	25321027	chr2	27758644	27765138	chr2	31946400	31970710	chr2	31996499	32021958
chr2	3570204	3583761	chr2	3601055	3606384	chr2	42720815	42740505	chr2	46661939	46695830
chr2	47240831	47243308	chr2	55313447	55316245	chr2	55317291	55334737	chr2	55645554	55679568
chr2	61259624	61268149	chr2	61953100	61964189	chr2	63675125	63687800	chr2	65168513	65185444
chr2	65308473	65351875	chr2	68122842	68143699	chr2	69476778	69512601	chr2	73308665	73313881
chr2	73325180	73332068	chr2	73810001	73815150	chr2	73981980	74000209	chr2	74216052	74228545
chr2	74286339	74295919	chr2	74635370	74638181	chr2	85676378	85678215	chr2	86224585	86276082
chr2	86559252	86573350	chr2	9465035	9471838	chr2	95116424	95151285	chr2	9641849	9649096
chr2	99169071	99180328	chr2	99304432	99319226	chr20	13688400	13704891	chr20	1372184	1395486
chr20	17542509	17550419	chr20	1843742	1853523	chr20	30871414	30901871	chr20	31899933	31905297
chr20	32140954	32163681	chr20	32331587	32354810	chr20	32667506	32728566	chr20	35794692	35841152
chr20	35901918	35933931	chr20	3683043	3696403	chr20	39177425	39186535	chr20	41519949	41523823
chr20	42947864	42970554	chr20	43876444	43878986	chr20	44412805	44417917	chr20	50133961	50148536
chr20	5043633	5048573	chr20	52257807	52269147	chr20	54378129	54391639	chr20	57041595	57051250
chr20	60395781	60396970	chr21	17841223	17855664	chr21	29170318	29179484	chr21	33744897	33763104
chr21	33872510	33876182	chr21	45188035	45222267	chr21	9936235	9943864	chr22	16454838	16491513
chr22	17543095	17545532	chr22	17817435	17846674	chr22	18486534	18494583	chr22	28206209	28215170
chr22	29302619	29315257	chr22	30129013	30160162	chr22	35193042	35207622	chr22	36533873	36542849
chr22	36601792	36614583	chr22	37393966	37399788	chr22	38038836	38044544	chr22	39552203	39582577
chr22	39679515	39699482	chr22	40250794	40254929	chr22	45018582	45022846	chr3	101540621	101554032
chr3	101911138	101950495	chr3	10266158	10296250	chr3	102775730	102795971	chr3	102882625	102888270
chr3	10317617	10332114	chr3	110527924	110535524	chr3	121028238	121078047	chr3	123561125	123584773
chr3	127130807	127136079	chr3	129253991	129261723	chr3	130749099	130752991	chr3	134775431	134790324
chr3	143105151	143127698	chr3	143877751	143899592	chr3	144238726	144256593	chr3	14462254	14501502
chr3	150192060	150228014	chr3	151768381	151782153	chr3	151803863	151830913	chr3	157132245	157138193
chr3	157743549	157755660	chr3	161556495	161585130	chr3	161700683	161732041	chr3	162441540	162452487
chr3	180552380	180586186	chr3	182185035	182188580	chr3	184143337	184166236	chr3	185029877	185085355
chr3	185442905	185446097	chr3	186844222	186893246	chr3	187985045	187990377	chr3	196724338	196751465
chr3	19967300	20001662	chr3	38514125	38523449	chr3	39405987	39413682	chr3	39424105	39428933
chr3	42800797	42821012	chr3	44949422	44975954	chr3	46543920	46561675	chr3	46595579	46598833
chr3	48456671	48460609	chr3	48869406	48911302	chr3	48974047	48996113	chr3	5187187	5195867
chr3	53894140	53897551	chr3	57532695	57558172	chr3	67628910	67662300	chr3	72924757	72976267
chr3	73179027	73198705	chr3	75530342	75554124	chr4	100020123	100042158	chr4	100179547	100201696
chr4	100212341	100225393	chr4	101021438	101027536	chr4	101039740	101063366	chr4	101088271	101090455
chr4	103936375	103968418	chr4	104218220	104236880	chr4	109150362	109175703	chr4	109761217	109765855
chr4	109791236	109808425	chr4	110854886	110870615	chr4	121200069	121207436	chr4	13071421	13094898
chr4	139287076	139300804	chr4	140198478	140224987	chr4	148758521	148775322	chr4	152240229	152245241
chr4	166220000	166243699	chr4	170887248	170911545	chr4	17225369	17236333	chr4	17423685	17429093
chr4	174528661	174535218	chr4	1785438	1804460	chr4	184807598	184817301	chr4	185789199	185796617
chr4	186405441	186425305	chr4	20315466	20338466	chr4	22043318	22084564	chr4	2440675	2484123
chr4	26035377	26043648	chr4	38358603	38375584	chr4	39132143	39136325	chr4	3985736	3995340
chr4	48582189	48601382	chr4	55920277	55934012	chr4	56997131	57020618	chr4	57032683	57062810
chr4	83563416	83568284	chr4	84035871	84040713	chr4	84230864	84248059	chr4	84598522	84612467
chr4	88575239	88591845	chr4	89235711	89298748	chr4	9314307	9341249	chr5	102483985	102511614
chr5	10303447	10318126	chr5	10671444	10705638	chr5	111092901	111120822	chr5	114580473	114605202
chr5	115195079	115205165	chr5	125908389	125913886	chr5	125964541	125989906	chr5	133335507	133354745
chr5	133520470	133540537	chr5	134061492	134087237	chr5	137871056	137881224	chr5	141333338	141348894
chr5	147754474	147799548	chr5	148705256	148711463	chr5	148855811	148884927	chr5	149803997	149807494
chr5	150050587	150058324	chr5	151131708	151160623	chr5	159781887	159788323	chr5	16506121	16516665
chr5	167917010	167939153	chr5	170747459	170770508	chr5	171250680	171270279	chr5	175705710	175707968
chr5	176663388	176666346	chr5	177509074	177513458	chr5	178054176	178057358	chr5	180596558	180603407
chr5	32391201	32455970	chr5	32624375	32637168	chr5	34951593	34961192	chr5	37733513	37758789
chr5	40868318	40871113	chr5	43157909	43209269	chr5	43571403	43591859	chr5	529135	541588
chr5	56545658	56581141	chr5	61678722	61694904	chr5	6686624	6722561	chr5	68549389	68561628
chr5	68713489	68746213	chr5	71653753	71690597	chr5	72830020	72837244	chr5	74098822	74109149
chr5	76362238	76395677	chr5	77747098	77753541	chr5	79315309	79320224	chr5	79957805	79981071
chr5	80636660	80644380	chr5	86725966	86736564	chr5	89805670	89845968	chr6	107126542	107184023
chr6	107184145	107220599	chr6	111302885	111322206	chr6	111386628	111395892	chr6	112498690	112509100
chr6	116528739	116545801	chr6	116648909	116673503	chr6	116999367	117021113	chr6	126361422	126401935
chr6	133177395	133180401	chr6	135398730	135417597	chr6	136594127	136605806	chr6	13899002	13915245
chr6	142510154	142561439	chr6	150099313	150108892	chr6	151377707	151399939	chr6	153357341	153365500
chr6	157639694	157664517	chr6	160119977	160129166	chr6	167263191	167290931	chr6	17649195	17665644
chr6	17708601	17716968	chr6	20588053	20601909	chr6	24517667	24532412	chr6	30796108	30800245
chr6	31240101	31246418	chr6	31606538	31612416	chr6	31806376	31812101	chr6	32255416	32256531
chr6	33348382	33352264	chr6	33648308	33655987	chr6	34312628	34318547	chr6	34497485	34501781
chr6	34838354	34849787	chr6	35544552	35546534	chr6	37043905	37054366	chr6	42282580	42290062
chr6	42955302	42965599	chr6	43553253	43579407	chr6	43651876	43665676	chr6	44324343	44329600
chr6	64344303	64349304	chr6	74172121	74182755	chr6	74283925	74286503	chr6	7806715	7826752
chr6	86443446	86444455	chr6	88441363	88448142	chr7	100007788	100009426	chr7	100651683	100654181
chr7	102502928	102527445	chr7	102782551	102795810	chr7	107899307	107930322	chr7	10939609	10946276
chr7	127819578	127828407	chr7	128382422	128407320	chr7	133777657	133787046	chr7	133900732	133913064
chr7	139799329	139825749	chr7	140961738	140999962	chr7	141065343	141077022	chr7	141085389	141096724
chr7	149614404	149623370	chr7	150794729	150819034	chr7	152128557	152151518	chr7	156822520	156871296
chr7	23316354	23325400	chr7	23348208	23357718	chr7	23513572	23538127	chr7	26212552	26219495
chr7	30628484	30639960	chr7	33100951	33105575	chr7	35869710	35911440	chr7	35952429	35980823
chr7	39846273	39874609	chr7	44207100	44217247	chr7	44578698	44580058	chr7	44802807	44807737
chr7	44840434	44849475	chr7	45734302	45771934	chr7	51207648	51228782	chr7	5533308	5536759
chr7	55829097	55854410	chr7	5598973	5612808	chr7	56136758	56141663	chr7	6030837	6065206
chr7	6380696	6409334	chr7	64001079	64028508	chr7	64162811	64171931	chr7	64856948	64866049
chr7	65963947	65980736	chr7	6915150	6937156	chr7	73226624	73249365	chr7	7642894	7646593
chr7	87345332	87375624	chr7	87673537	87677322	chr7	89722675	89741346	chr7	91579409	91601731

Continued on Next Page...

Table B.2 – Continued

chr	start	stop	chr	start	stop	chr	start	stop	chr	start	stop
chr7	91996045	92004334	chr7	92044352	92057632	chr7	94067914	94097082	chr7	97319519	97336407
chr7	98609131	98638797	chr7	98831546	98840514	chr7	98996096	99000262	chr8	101232103	101234806
chr8	101784320	101803488	chr8	101999980	102006444	chr8	10715112	10729688	chr8	11664574	11682258
chr8	12073629	12077777	chr8	12258021	12261339	chr8	12322912	12327022	chr8	125569933	125600305
chr8	125620631	125631399	chr8	130922966	130952924	chr8	145211141	145212600	chr8	145985961	145988100
chr8	17131108	17148724	chr8	20112373	20122138	chr8	25341466	25371892	chr8	26252338	26277302
chr8	30762679	30776815	chr8	38087417	38116376	chr8	39432016	39451320	chr8	41467516	41487650
chr8	42131146	42144010	chr8	43030644	43060080	chr8	47610125	47645293	chr8	55042916	55075164
chr8	55121492	55137438	chr8	57148167	57149616	chr8	59486487	59526604	chr8	66744671	66782790
chr8	68117874	68136844	chr8	74083699	74122538	chr8	74365428	74367586	chr8	74657559	74692242
chr8	74865403	74905257	chr8	76618605	76641598	chr8	7844062	7847387	chr8	82355325	82359568
chr8	87512750	87529563	chr8	99183758	99190158	chr9	101024384	101032716	chr9	109085664	109132985
chr9	112046123	112058595	chr9	115058212	115066590	chr9	122566272	122578296	chr9	127038336	127043290
chr9	129249977	129253505	chr9	130485929	130498485	chr9	131629473	131635614	chr9	132342077	132366461
chr9	135205597	135208102	chr9	138416200	138424722	chr9	139077828	139084802	chr9	19105775	19116308
chr9	19365782	19370267	chr9	19398917	19442486	chr9	33016476	33029897	chr9	33246799	33252828
chr9	35094221	35098737	chr9	35649074	35651177	chr9	35672022	35675138	chr9	35802963	35803777
chr9	37753631	37766402	chr9	4701158	4731061	chr9	70851135	70879606	chr9	720140	736101
chr9	72120182	72128878	chr9	74160604	74165524	chr9	80101880	80134827	chr9	85773644	85782993
chr9	88069285	88087274	chr9	96287521	96289971	chr9	98441673	98453597	chr9	99785462	99807256
chrX	100160941	100193718	chrX	100532630	100537443	chrX	103094029	103109802	chrX	106758501	106780844
chrX	107217573	107221424	chrX	107255959	107284259	chrX	11690189	11700736	chrX	116916462	116938361
chrX	118486436	118489303	chrX	119227722	119230652	chrX	119621918	119626240	chrX	129301546	129335006
chrX	135116230	135120623	chrX	135783286	135789251	chrX	13640301	13662623	chrX	149902417	149907894
chrX	152506580	152517780	chrX	152705449	152706553	chrX	153194126	153211195	chrX	153280872	153282447
chrX	153908283	153936574	chrX	153954844	153972310	chrX	15758106	15780488	chrX	19279356	19287064
chrX	21905099	21922876	chrX	23761410	23808992	chrX	30581568	30655588	chrX	40333221	40350774
chrX	48318508	48321613	chrX	48635678	48639086	chrX	48819151	48824439	chrX	54573340	54603888
chrX	55774526	55801482	chrX	56276342	56328267	chrX	69270509	69302854	chrX	69426548	69438976
chrX	70430808	70434032	chrX	70590744	70600464	chrX	71318333	71334120	chrX	71409179	71413673
chrX	77255911	77267580	chrX	77271921	77281787	chrX	99963171	99979999	chrY	18431080	18444810
chrY	18703743	18717478	chrY	19077364	19092148	chrY	19470799	19485581	chrY	24326976	24353681
chrY	24528313	24548828	chrY	26232378	26252883	chrY	26427466	26454179	chrY	2770205	2794955

B.3 Retrotrasposed genes (mouse)

Table B.3: Coordinates of found retrotransposed gene in the mouse genome (ENSEMBL version 40).

chr	start	stop	chr	start	stop	chr	start	stop	chr	start	stop
chr1	100468795	100487984	chr1	10382273	10402777	chr1	10995325	11006681			
chr1	111236517	111244055	chr1	111784403	111787540	chr1	111798354	111805472	chr1	112035483	112057251
chr1	112997743	113015757	chr1	114911837	114925710	chr1	115062156	115064757	chr1	11688940	11702712
chr1	117758862	117810569	chr1	144218974	144222800	chr1	1469112	1499792	chr1	148161834	148165811
chr1	148726985	148746370	chr1	148937545	148958637	chr1	149257565	149268016	chr1	149998762	150002222
chr1	150113159	150148570	chr1	151901134	151907658	chr1	152187598	152197639	chr1	152395466	152422287
chr1	152511723	152514975	chr1	153408509	153411993	chr1	153514174	153521707	chr1	153546455	153557080
chr1	153965551	153974937	chr1	154828211	154830537	chr1	158154524	158156952	chr1	158459089	158476425
chr1	159390399	159395267	chr1	159399331	159402116	chr1	159560026	159599607	chr1	160802435	160825227
chr1	165092371	165112096	chr1	166001444	166025131	chr1	167342563	167368582	chr1	173235784	173246030
chr1	174263333	174282084	chr1	174352386	174442834	chr1	17612157	17628282	chr1	176242584	176273724
chr1	177279472	177308295	chr1	180620053	180627548	chr1	181115072	181123500	chr1	183323307	183336812
chr1	19453861	19458058	chr1	201243523	201259136	chr1	209902738	209915586	chr1	210018882	210033139
chr1	220907980	220952436	chr1	222614820	222630210	chr1	22277507	22291472	chr1	224318652	224326106
chr1	229567401	229576468	chr1	231152988	231181118	chr1	23891685	23895498	chr1	24168185	24179499
chr1	24861737	24871338	chr1	27925188	27934742	chr1	28029910	28049640	chr1	28090638	28113267
chr1	28399657	28432123	chr1	31505276	31542197	chr1	31611091	31618499	chr1	31868059	31873711
chr1	32144609	32176482	chr1	32460733	32469760	chr1	32530324	32571813	chr1	33248111	33275064
chr1	36527850	36543076	chr1	3765171	3791849	chr1	37806469	37834023	chr1	37930911	37943755
chr1	39799094	39802002	chr1	40427312	40479179	chr1	40991411	41009857	chr1	41374088	41390926
chr1	42896680	42915012	chr1	42920685	42940600	chr1	45014292	45016996	chr1	45272570	45278798
chr1	45749310	45757314	chr1	45926491	45932691	chr1	47606726	47616126	chr1	52010348	52026187
chr1	52294459	52324431	chr1	52506782	52542185	chr1	52571026	52584892	chr1	53465195	53476778
chr1	54438782	54454594	chr1	54465188	54496413	chr1	54849867	54861770	chr1	6301144	6332488
chr1	6607827	6616208	chr1	67649548	67668666	chr1	68716079	68735407	chr1	71307547	71316824
chr1	76025786	76030537	chr1	84888674	84909579	chr1	92232302	92252566	chr1	93070193	93080069
chr1	9522115	9565175	chr1	9954663	9967452	chr10	100183125	100195161	chr10	102097842	102114570
chr10	102273490	102279608	chr10	103360412	103374560	chr10	103417633	103444695	chr10	104150412	104151260
chr10	104231744	104238885	chr10	105052563	105100820	chr10	105632444	105648748	chr10	105871914	105875510
chr10	1076850	1080113	chr10	112351745	112354380	chr10	118384392	118392749	chr10	120059252	120085923
chr10	120918680	120924093	chr10	12254769	12277893	chr10	124907233	124914845	chr10	126079429	126092995
chr10	126666891	126682053	chr10	13400074	13427035	chr10	14963503	14983253	chr10	27033614	27049383
chr10	27443455	27448291	chr10	33230398	33240529	chr10	42974187	43000743	chr10	5847199	5895311
chr10	59623353	59667582	chr10	59815183	59825728	chr10	6183273	6197669	chr10	69768338	69772956
chr10	70074453	70116448	chr10	70586768	70598372	chr10	71579964	71591689	chr10	71636026	71663154
chr10	73764276	73784673	chr10	74604451	74665211	chr10	7879025	7884369	chr10	79465115	79470465
chr10	82158284	82182729	chr10	88801110	88809010	chr10	90684961	90697125	chr10	96987324	97040708
chr10	97793309	97810610	chr10	99176008	99183187	chr10	99390772	99416316	chr11	100847140	100855016
chr11	106880235	106932933	chr11	109982591	110006690	chr11	111462807	111471710	chr11	113776638	113781731
chr11	116528367	116539824	chr11	118391759	118394262	chr11	122433411	122438054	chr11	124050342	124069723
chr11	124947565	124959781	chr11	125335614	125398642	chr11	13366120	13399963	chr11	13690050	13710431
chr11	14256047	14273979	chr11	14483001	14496102	chr11	16716801	16734014	chr11	17052516	17055369
chr11	18434846	18457093	chr11	30301265	30316350	chr11	32069118	32082855	chr11	33719875	33728760
chr11	3377001	3386785	chr11	36611408	36637375	chr11	47596842	47620594	chr11	50324789	50336409
chr11	57261656	57265018	chr11	58135016	58141310	chr11	59171837	59191013	chr11	59462466	59473066
chr11	59541399	59555141	chr11	60366146	60374996	chr11	61488614	61491682	chr11	62083650	62096791
chr11	62139347	62146032	chr11	63274041	63283916	chr11	63435304	63440605	chr11	63476455	63479162
chr11	63840770	63841586	chr11	65378884	65382224	chr11	65416271	65424435	chr11	65792655	65801537
chr11	65959887	65962815	chr11	67289421	67316547	chr11	69848154	69850567	chr11	71189207	71193336
chr11	73096113	73119517	chr11	73214398	73253289	chr11	74337954	74366427	chr11	74789384	74794381
chr11	74955005	74961444	chr11	75749648	75769593	chr11	77008299	77026517	chr11	7965654	7974292
chr11	82213058	82239197	chr11	82650150	82668949	chr11	8413639	8440036	chr11	8661324	8663967
chr11	8890574	8897631	chr11	89574295	89588048	chr11	95172074	95203927	chr11	9554608	9567888
chr12	100392384	100406442	chr12	100957777	100979942	chr12	10256766	10266966	chr12	103804245	103827661
chr12	10649879	10657435	chr12	108020694	108032575	chr12	108773132	108802658	chr12	109357091	109367740
chr12	111327374	111330845	chr12	115639938	115660053	chr12	117058291	117067770	chr12	119118895	119123019
chr12	122439930	122459850	chr12	122659179	122671432	chr12	14833255	14843932	chr12	15926719	15947670
chr12	21679537	21702011	chr12	23578126	23620042	chr12	25249450	25295056	chr12	26982628	27010844
chr12	2799496	2802339	chr12	40765990	40799236	chr12	4255430	4282565	chr12	4628539	4666633
chr12	47807837	47811445	chr12	48670223	48672399	chr12	51577242	51585109	chr12	51629172	51632961
chr12	51696521	51722258	chr12	52135934	52159819	chr12	52345222	52352700	chr12	52960777	52965543
chr12	54722475	54724762	chr12	54784704	54791360	chr12	54835104	54841623	chr12	54904913	54909904
chr12	54954837	54966076	chr12	55343403	55368295	chr12	55392482	55404574	chr12	55411251	55432338
chr12	55949433	55976596	chr12	56448689	56452158	chr12	61069668	61081322	chr12	63105862	63114630
chr12	6505555	6510937	chr12	6515918	6517795	chr12	67328744	67340070	chr12	6846959	6850250
chr12	6945862	6948986	chr12	74179698	74191664	chr12	74733252	74749041	chr12	75776626	75796930
chr12	7833298	7839862	chr12	7856380	7876679	chr12	7963095	7977767	chr12	8271596	8275775
chr12	8793724	8818264	chr12	892276	910751	chr12	8985304	8990225	chr12	97445811	97465987
chr13	113008835	113024026	chr13	114065512	114089379	chr13	18902609	18910290	chr13	24053803	24061484
chr13	24431871	24439593	chr13	26538293	26588730	chr13	26725896	26729315	chr13	29931994	29938097
chr13	40265346	40280733	chr13	40384650	40393874	chr13	44411383	44461606	chr13	44809021	44813286
chr13	47415192	47469161	chr13	49515648	49554167	chr13	51699157	51707414	chr13	51933042	51946029
chr13	52124976	52159988	chr13	97456413	97465888	chr13	97902881	97932578	chr14	101619655	101620995
chr14	19846254	19867373	chr14	20748559	20772233	chr14	22461490	22463114	chr14	22486658	22496120
chr14	22860652	22865208	chr14	23682425	23684198	chr14	30605105	30635284	chr14	31685218	31695324
chr14	34831414	34856424	chr14	38653272	38675869	chr14	49645103	49653005	chr14	50776738	50792507
chr14	57784220	57808429	chr14	59031498	59040532	chr14	67126044	67136735	chr14	67156473	67211081
chr14	72478209	72496093	chr14	74030221	74031815	chr14	74252407	74273170	chr14	88692287	88726548
chr14	89796208	89808713	chr14	92239914	92268916	chr14	92369340	92375900	chr14	96057093	96103177
chr14	99798422	99813565	chr15	22770579	22774822	chr15	23134732	23168054	chr15	32221080	32226785
chr15	36543613	36564167	chr15	38115536	38118654	chr15	39310822	39358382	chr15	40530631	40570598
chr15	40622076	40626984	chr15	41710109	41728258	chr15	41806589	41825679	chr15	42408188	42417807
chr15	46410925	46421858	chr15	48503893	48538632	chr15	50125319	50145746	chr15	53260813	53276427
chr15	55786207	55796494	chr15	57186842	57204535	chr15	61233109	61236739	chr15	62474649	62534520
chr15	64561145	64568668	chr15	64578712	64584238	chr15	67532211	67534937	chr15	70278428	70298507
chr15	71403253	71423198	chr15	72999672	73017490	chr15	73933782	73976456	chr15	76959907	76977126
chr15	78199724	78217757	chr15	80608216	80611596	chr15	81002559	81005939	chr15	98029246	98049055
chr15	99639238	99652939	chr16	11843058	11852828	chr16	15062857	15093971	chr16	18701782	18707942
chr16	18710519	18717658	chr16	1952056	1954630	chr16	19625350	19633797	chr16	21718470	21737582
chr16	21876251	21890938	chr16	23476771	23489679	chr16	23499874	23506135	chr16	23560289	23588683
chr16	28995479	29022985	chr16	48617810	48628490	chr16	5067909	5075045	chr16	52082793	52094717
chr16	55836642	55844916	chr16	57298540	57325669	chr16	66247870	66248957	chr16	66456503	66462724
chr16	671463	672744	chr16	67931415	67933266	chr16	68333286	68346444	chr16	72888278	72897023
chr16	79608437	79620384	chr16	79627051	79636280	chr16	79673432	79687455	chr16	82399131	82402773
chr16	84392309	84398109	chr16	85983518	85994474	chr16	87861543	87879349	chr16	8799201	8814456
chr16	88517310	88530000	chr17	10525113	10539874	chr17	11925398	11973329	chr17	1194621	1215104
chr17	15381031	15407557	chr17	18634946	18650343	chr17	20588983	20598117	chr17	20850804	20856821
chr17	24071850	24075500	chr17	2517037	2531991	chr17	25523684	25537584	chr17	26143587	26175787
chr17	27439295	27446065	chr17	28282604	28292773	chr17	31001663	31033935	chr17	33706643	33732362
chr17	34162543	34173965	chr17	34259894	34262886	chr17	34610990	34614508	chr17	36038533	36052337
chr17	37098666	37101403	chr17	37382805	37406115	chr17	37530541	37536134	chr17	37974018	37978257
chr17	38404286	38408490	chr17	39629823	39631010	chr17	39642253	39651235	chr17	39753565	39755446
chr17	42355496	42364564	chr17	43258999	43263900	chr17	43405883	43414118	chr17	44365566	44373426

Continued on Next Page...

Table B.3 – Continued

chr	start	stop	chr	start	stop	chr	start	stop	chr	start	stop
chr17	44836423	44845654	chr17	45507962	45509400	chr17	4799963	4801140	chr17	50393574	50401055
chr17	52370562	52393402	chr17	5276824	5283169	chr17	55386207	55395032	chr17	59259227	59262887
chr17	60176249	60185208	chr17	62767379	62793181	chr17	628645	642105	chr17	63463686	63473421
chr17	70675420	70690720	chr17	71354584	71362974	chr17	7153656	7156496	chr17	7156745	7159459
chr17	71592194	71611105	chr17	72065464	72071318	chr17	73676332	73680395	chr17	7418586	7421953
chr17	77091600	77093978	chr17	8288304	8304163	chr18	11883471	11898316	chr18	22120158	22161102
chr18	41939180	41958857	chr18	45268856	45271297	chr18	53367578	53391597	chr18	53418907	53424938
chr18	54737256	54772010	chr18	9092747	9124340	chr19	10362809	10375193	chr19	12768643	12773610
chr19	1389366	1391496	chr19	14069170	14078686	chr19	14380851	14385058	chr19	14535612	14536670
chr19	15326582	15345769	chr19	16048376	16074813	chr19	16522808	16544046	chr19	1771302	1776944
chr19	17833102	17835124	chr19	18503677	18513807	chr19	18545097	18547047	chr19	20897898	20928430
chr19	21116683	21159426	chr19	21266559	21304002	chr19	21444148	21449262	chr19	2194380	2199597
chr19	21949115	21985533	chr19	22155894	22190364	chr19	22261122	22288638	chr19	22733989	22758732
chr19	22806949	22832113	chr19	23091719	23122204	chr19	23195833	23224983	chr19	23337047	23370079
chr19	23729465	23733502	chr19	24061865	24103522	chr19	3045647	3074991	chr19	37767694	37770165
chr19	39377254	39410836	chr19	41325809	41329039	chr19	4314737	4317993	chr19	43557309	43566299
chr19	43801686	43819430	chr19	43900455	43910494	chr19	45017181	45021655	chr19	47480652	47490959
chr19	52325965	52365020	chr19	53810408	53812947	chr19	541121	568150	chr19	54160378	54161947
chr19	54684916	54687357	chr19	54691535	54694752	chr19	5641274	5642678	chr19	56517129	56525415
chr19	59396838	59403323	chr19	60589528	60591555	chr19	60604465	60610963	chr19	62354906	62370665
chr19	63590448	63597971	chr19	977635	990064	chr19	9807011	9821358	chr2	100549394	100559626
chr2	100985596	100989317	chr2	10180405	10184549	chr2	10499095	10502808	chr2	109660916	109728781
chr2	114390952	114431822	chr2	11503073	11511299	chr2	118295241	118305040	chr2	119841041	119846580
chr2	122201723	122210905	chr2	128321706	128332129	chr2	131821715	131832198	chr2	150134399	150151908
chr2	152367612	152379076	chr2	157040854	157078252	chr2	170168942	170202484	chr2	170363756	170389398
chr2	173936218	173941860	chr2	174647981	174655036	chr2	174796004	174821532	chr2	176902273	176914682
chr2	177788513	177793084	chr2	183515046	183554297	chr2	183701290	183734652	chr2	190817853	190834204
chr2	198059559	198071803	chr2	198073322	198076419	chr2	198089043	198125360	chr2	202779459	202811549
chr2	203453613	203484573	chr2	206732865	206735896	chr2	208809423	208828040	chr2	210575579	210589612
chr2	223481701	223507617	chr2	233129370	233142161	chr2	24047649	24076635	chr2	24144115	24152645
chr2	25310406	25321027	chr2	27758644	27765138	chr2	31946400	31970710	chr2	31996499	32021958
chr2	3570204	3583761	chr2	3601055	3606384	chr2	42720815	42740505	chr2	46661939	46695830
chr2	47240831	47243308	chr2	55313447	55316245	chr2	55317291	55334737	chr2	55645554	55679568
chr2	61259624	61268149	chr2	61953100	61964189	chr2	63675125	63687800	chr2	65168513	65185444
chr2	65308473	65351875	chr2	68122842	68143699	chr2	69476778	69512601	chr2	73308665	73313881
chr2	73325180	73332068	chr2	73810001	73815150	chr2	73981980	74000209	chr2	74216052	74228545
chr2	74286339	74295919	chr2	74635370	74638181	chr2	85676378	85678215	chr2	86224585	86276082
chr2	86559252	86573350	chr2	9465035	9471838	chr2	95116424	95151285	chr2	9641849	9649096
chr2	99169071	99180328	chr2	99304432	99319226	chr20	13688400	13704891	chr20	1372184	1395486
chr20	17542509	17550419	chr20	1843742	1853523	chr20	30871414	30901871	chr20	31899933	31905297
chr20	32140954	32163681	chr20	32331587	32354810	chr20	32667506	32728566	chr20	35794692	35841152
chr20	35901918	35933931	chr20	3683043	3696403	chr20	39177425	39186535	chr20	41519949	41523823
chr20	42947864	42970554	chr20	43876444	43878986	chr20	44412805	44417917	chr20	50133961	50148536
chr20	5043633	5048573	chr20	52257807	52269147	chr20	54378129	54391639	chr20	57041595	57051250
chr20	60395781	60396970	chr21	17841223	17855664	chr21	29170318	29179484	chr21	33744897	33763104
chr21	33872510	33876182	chr21	45188035	45222267	chr21	9936235	9943864	chr22	16454838	16491513
chr22	17543095	17545532	chr22	17817435	17846674	chr22	18486534	18494583	chr22	28206209	28215170
chr22	29302619	29315257	chr22	30129013	30160162	chr22	35193042	35207622	chr22	36533873	36542849
chr22	36601792	36614583	chr22	37393966	37399788	chr22	38038836	38044544	chr22	39552203	39582577
chr22	39679515	39699482	chr22	40250794	40254929	chr22	45018582	45022846	chr3	101540621	101554032
chr3	101911138	101950495	chr3	10266158	10296250	chr3	102775730	102795971	chr3	102882625	102888270
chr3	10317617	10332114	chr3	110527924	110535524	chr3	121028238	121078047	chr3	123561125	123584773
chr3	127130807	127136079	chr3	129253991	129261723	chr3	130749099	130752991	chr3	134775431	134790324
chr3	143105151	143127698	chr3	143877751	143899592	chr3	144238726	144256593	chr3	14462254	14501502
chr3	150192060	150228014	chr3	151768381	151782153	chr3	151803863	151830913	chr3	157132245	157138193
chr3	157743549	157755660	chr3	161556495	161585130	chr3	161700683	161732041	chr3	162441540	162452487
chr3	180552380	180586186	chr3	182185035	182188580	chr3	184143337	184166236	chr3	185029877	185085355
chr3	185442905	185446097	chr3	186844222	186893246	chr3	187985045	187990377	chr3	196724338	196751465
chr3	19967300	20001662	chr3	38514125	38523449	chr3	39405987	39413682	chr3	39424105	39428933
chr3	42800797	42821012	chr3	44949422	44975954	chr3	46543920	46561675	chr3	46595579	46598833
chr3	48456671	48460609	chr3	48869406	48911302	chr3	48974047	48996113	chr3	5187187	5195867
chr3	53894140	53897551	chr3	57532695	57558172	chr3	67628910	67662300	chr3	72924757	72976267
chr3	73179027	73198705	chr3	75530342	75554124	chr4	100020123	100042158	chr4	100179547	100201696
chr4	100212341	100225393	chr4	101021438	101027536	chr4	101039740	101063366	chr4	101088271	101090455
chr4	103936375	103968418	chr4	104218220	104236880	chr4	109150362	109175703	chr4	109761217	109765855
chr4	109791236	109808425	chr4	110854886	110870615	chr4	121200069	121207436	chr4	13071421	13094898
chr4	139287076	139300804	chr4	140198478	140224987	chr4	148758521	148775322	chr4	152240229	152245241
chr4	166220000	166243699	chr4	170887248	170911545	chr4	17225369	17236333	chr4	17423685	17429093
chr4	174528661	174535218	chr4	1785438	1804460	chr4	184807598	184817301	chr4	185789199	185796617
chr4	186405441	186425305	chr4	20315466	20338466	chr4	22043318	22084564	chr4	2440675	2484123
chr4	26035377	26043648	chr4	38358603	38375584	chr4	39132143	39136325	chr4	3985736	3995340
chr4	48582189	48601382	chr4	55920277	55934012	chr4	56997131	57020618	chr4	57032683	57062810
chr4	83563416	83568284	chr4	84035871	84040713	chr4	84230864	84248059	chr4	84598522	84612467
chr4	88575239	88591845	chr4	89235711	89298748	chr4	9314307	9341249	chr5	102483985	102511614
chr5	10303447	10318126	chr5	10671444	10705638	chr5	111092901	111120822	chr5	114580473	114605202
chr5	115195079	115205165	chr5	125908389	125913886	chr5	125964541	125989906	chr5	133335507	133354745
chr5	133520470	133540537	chr5	134061492	134087237	chr5	137871056	137881224	chr5	141333338	141348894
chr5	147754474	147799548	chr5	148705256	148711463	chr5	148855811	148884927	chr5	149803997	149807494
chr5	150050587	150058324	chr5	151131708	151160623	chr5	159781887	159788323	chr5	16506121	16516665
chr5	167917010	167939153	chr5	170747459	170770508	chr5	171250680	171270279	chr5	175705710	175707968
chr5	176663388	176666346	chr5	177509074	177513458	chr5	178054176	178057358	chr5	180596558	180603407
chr5	32391201	32455970	chr5	32624375	32637168	chr5	34951593	34961192	chr5	37733513	37758789
chr5	40868318	40871113	chr5	43157909	43209269	chr5	43571403	43591859	chr5	529135	541588
chr5	56545658	56581141	chr5	61678722	61694904	chr5	6686624	6722561	chr5	68549389	68561628
chr5	68713489	68746213	chr5	71653753	71690597	chr5	72830020	72837244	chr5	74098822	74109149
chr5	76362238	76395677	chr5	77747098	77753541	chr5	79315309	79320224	chr5	79957805	79981071
chr5	80636660	80644380	chr5	86725966	86736564	chr5	89805670	89845968	chr6	107126542	107184023
chr6	107184145	107220599	chr6	111302885	111322206	chr6	111386628	111395892	chr6	112498690	112509100
chr6	116528739	116545801	chr6	116648909	116673503	chr6	116999367	117021113	chr6	126361422	126401935
chr6	133177395	133180401	chr6	135398730	135417597	chr6	136594127	136605806	chr6	13899002	13915245
chr6	142510154	142561439	chr6	150099313	150108892	chr6	151377707	151399939	chr6	153357341	153365500
chr6	157639694	157664517	chr6	160119977	160129166	chr6	167263191	167290931	chr6	17649195	17665644
chr6	17708601	17716968	chr6	20588053	20601909	chr6	24517667	24532412	chr6	30796108	30800245
chr6	31240101	31246418	chr6	31606538	31612416	chr6	31806376	31812101	chr6	32255416	32256531
chr6	33348382	33352264	chr6	33648308	33655987	chr6	34312628	34318547	chr6	34497485	34501781
chr6	34838354	34849787	chr6	35544552	35546534	chr6	37043905	37054366	chr6	42282580	42290062
chr6	42955302	42965599	chr6	43553253	43579407	chr6	43651876	43665676	chr6	44324343	44329600
chr6	64344303	64349304	chr6	74172121	74182755	chr6	74283925	74286503	chr6	7806715	7826752
chr6	86443446	86444455	chr6	88441363	88448142	chr7	100007788	100009426	chr7	100651683	100654181
chr7	102502928	102527445	chr7	102782551	102795810	chr7	107899307	107930322	chr7	10939609	10946276
chr7	127819578	127828407	chr7	128382422	128407320	chr7	133777657	133787046	chr7	133900732	133913064
chr7	139799329	139825749	chr7	140961738	140999962	chr7	141065343	141077022	chr7	141085389	141096724
chr7	149614404	149623370	chr7	150794729	150819034	chr7	152128557	152151518	chr7	156822520	156871296
chr7	23316354	23325400	chr7	23348208	23357718	chr7	23513572	23538127	chr7	26212552	26219495
chr7	30628484	30639960	chr7	33100951	33105575	chr7	35869710	35911440	chr7	35952429	35980823
chr7	39846273	39874609	chr7	44207100	44217247	chr7	44578698	44580058	chr7	44802807	44807737
chr7	44840434	44849475	chr7	45734302	45771934	chr7	51207648	51228782	chr7	5533308	5536759
chr7	55829097	55854410	chr7	5598973	5612808	chr7	56136758	56141663	chr7	6030837	6065206
chr7	6380696	6409334	chr7	64001079	64028508	chr7	64162811	64171931	chr7	64856948	64866049
chr7	65963947	65980736	chr7	6915150	6937156	chr7	73226624	73249365	chr7	7642894	7646593
chr7	87345332	87375624	chr7	87673537	87677322	chr7	89722675	89741346	chr7	91579409	91601731

Continued on Next Page...

Table B.3 – Continued

chr	start	stop	chr	start	stop	chr	start	stop	chr	start	stop
chr7	91996045	92004334	chr7	92044352	92057632	chr7	94067914	94097082	chr7	97319519	97336407
chr7	98609131	98638797	chr7	98831546	98840514	chr7	98996096	99000262	chr8	101232103	101234806
chr8	101784320	101803488	chr8	101999980	102006444	chr8	10715112	10729688	chr8	11664574	11682258
chr8	12073629	12077777	chr8	12258021	12261339	chr8	12322912	12327022	chr8	125569933	125600305
chr8	125620631	125631399	chr8	130922966	130952924	chr8	145211141	145212600	chr8	145985961	145988100
chr8	17131108	17148724	chr8	20112373	20122138	chr8	25341466	25371892	chr8	26252338	26277302
chr8	30762679	30776815	chr8	38087417	38116376	chr8	39432016	39451320	chr8	41467516	41487650
chr8	42131146	42144010	chr8	43030644	43060080	chr8	47610125	47645293	chr8	55042916	55075164
chr8	55121492	55137438	chr8	57148167	57149616	chr8	59486487	59526604	chr8	66744671	66782790
chr8	68117874	68136844	chr8	74083699	74122538	chr8	74365428	74367586	chr8	74657559	74692242
chr8	74865403	74905257	chr8	76618605	76641598	chr8	7844062	7847387	chr8	82355325	82359568
chr8	87512750	87529563	chr8	99183758	99190158	chr9	101024384	101032716	chr9	109085664	109132985
chr9	112046123	112058595	chr9	115058212	115066590	chr9	122566272	122578296	chr9	127038336	127043290
chr9	129249977	129253505	chr9	130485929	130498485	chr9	131629473	131635614	chr9	132342077	132366461
chr9	135205597	135208102	chr9	138416200	138424722	chr9	139077828	139084802	chr9	19105775	19116308
chr9	19365782	19370267	chr9	19398917	19442486	chr9	33016476	33029897	chr9	33246799	33252828
chr9	35094221	35098737	chr9	35649074	35651177	chr9	35672022	35675138	chr9	35802963	35803777
chr9	37753631	37766402	chr9	4701158	4731061	chr9	70851135	70879606	chr9	720140	736101
chr9	72120182	72128878	chr9	74160604	74165524	chr9	80101880	80134827	chr9	85773644	85782993
chr9	88069285	88087274	chr9	96287521	96289971	chr9	98441673	98453597	chr9	99785462	99807256
chrX	100160941	100193718	chrX	100532630	100537443	chrX	103094029	103109802	chrX	106758501	106780844
chrX	107217573	107221424	chrX	107255959	107284259	chrX	11690189	11700736	chrX	116916462	116938361
chrX	118486436	118489303	chrX	119227722	119230652	chrX	119621918	119626240	chrX	129301546	129335006
chrX	135116230	135120623	chrX	135783286	135789251	chrX	13640301	13662623	chrX	149902417	149907894
chrX	152506580	152517780	chrX	152705449	152706553	chrX	153194126	153211195	chrX	153280872	153282447
chrX	153908283	153936574	chrX	153954844	153972310	chrX	15758106	15780488	chrX	19279356	19287064
chrX	21905099	21922876	chrX	23761410	23808992	chrX	30581568	30655588	chrX	40333221	40350774
chrX	48318508	48321613	chrX	48635678	48639086	chrX	48819151	48824439	chrX	54573340	54603888
chrX	55774526	55801482	chrX	56276342	56328267	chrX	69270509	69302854	chrX	69426548	69438976
chrX	70430808	70434032	chrX	70590744	70600464	chrX	71318333	71334120	chrX	71409179	71413673
chrX	77255911	77267580	chrX	77271921	77281787	chrX	99963171	99979999	chrY	18431080	18444810
chrY	18703743	18717478	chrY	19077364	19092148	chrY	19470799	19485581	chrY	24326976	24353681
chrY	24528313	24548828	chrY	26232378	26252883	chrY	26427466	26454179	chrY	2770205	2794955

B.4 Validated genes sequences

Predicted genes sequences.

```
>718689 (M.musculus)
TCCAGCCTCAACATGAGTTGGGAATCCTCAAGGACAAGGGATACCTGGTAGGTGAATAAGAAAACGAATAATTTAGATAA
AATGCTGGCAGACTGTTCTTGCAGTGAAATCTCCTAAACTCAGGAGTCAGAGTGCTTTACAGACTTGACTCAGCCCCAGT
CTGTTATTGAAATCATACTAGGAAGTTAGTTACAAAGTTTATGGTGGGGAGGGGCAGTAAATACTTCTCCATTTTCCTGT
GCCAATATTTTCTTGTGATTCAGAGTAGTATAGATTTAGAACAGAAAAGAAAGTAACTGTAAAATCTCAAACATAAAAGT
TTTGTAGCAAGCACAGGTTTGTTTGTTTGTTTATCTTCCAGGCACCAGTCTCTGGTGAGGTTGAGGGCAGCAGCATTGTG
CTTACAAAGTCAATATCATTCAGACATGTGAGAAAGAGTATGATTTCATAATAGGGCAAAAATGAGGTCATACTGCTAAC
CCTGCACTCT-GATCCTATATGTCTGTCTATAAAAACATGCTCATGTAATCTGTATGGAGAGTTCGTGGGGAGCTGCCAT
TATTGTTACTGAACTCAATGAACCCTTGCTGACTTTCTGGAAAATATCTGTATGTTATTCCATCCTCAGTTTTGTAATTA
AAATGCCCTATTTTCATTATGCAGAATTGAGACTAAGACTGGGGGGAACAGCTTCAGATTTACTCATCTCTACCCATGGC
TCTAGGTTACTTTCATTTTCACCTCAAGAGAGTGATTGTAACATAC-ACATTTCTCTTGAGTTCATGAATAGCTTTCACT
TTAATATTATAGGTAACATCTAGTTCTTGGATATCAAGCAGACTCCTATTAGGTGGTGCCAGATCACCAAATACCACTTT
AATCAAGGAAGAAGGAAAGAGAA-TAAATAGAACAGAAGAATAATAAGTTTTAAGAATAGTTGCAGAAATGACATCATGA
GGTGCTTATGCCAAAGAAGCAGACCTTGGGGCCTCTCAGTACTCATACTAGGGCCCACAAGAACCCATAGCCT

>1033649 (M.musculus)
GGTCCTTTTACAGGAGTGGAGTCCCTCACTCCAAGAGAGGTTAGCAACTGCTCAGACATTCAGGCCATTTTCCTCTGCTA
GCCTGTGCTTCTTTGAGTAGGTCCAATTCAGTGCTACTAGGTAT-TTCTAGGGAACGCTAACAAATCTGCATACTGCTAT
GTTCCAGAGGATGATAAAAGGTGCTGCAGTTGGCATGACTGCTGACAAGAGACACTCAGAGGTTCACTTCAAATTCAATA
TTCAGATCGTCTTCATGGTCTTTTCTTAGTCTCCCCTTAGAGACAGAACTGAAATTCACCATGGACCACATCTTTTACAT
GAAG-AGAGTTCATCCCAAAATTTATCATCATTCATAAGCTTTCCAACCAGCTCTGAGAATTAATTTCCCAGAAAACAGA
ATATCCTTGGATGGGAATAACATCTGTGGGCTCTCATAAGCCAATCCTAGTGGATGGTATTGCTCTGCCAGTAATTGGAA
CAATGAGTCTCTGCTGGTCCGTGATGAGCAGATTCCAGGCACCTTCGATCAAAATGTAATGAGATGATCTTGACCCGTAA
AATGTTAAGTAAGTATTCAAAATGGTTCTTGGACACCCATTTGCACAAAGGAG-TTCAGTCTCATGATACTGAACAGACT
GTTCTAGCTGATGAGGCTGGGCTTAGAAGATGGACTACAGCCAAGAACCCTGGCCACCATAACAACAAGTCACACCTTTA
GGAAACACTTGAAAAGCACTTATGCACAGTAGAAAAGTACAAGGATTAAAGTCTCCTCAGGCATTCCAGCATAGAGAGAG
TCTGACAGACTCAGGTCAGAGGTTACCTCCAAATAATAGAAGAGGAAGAAGGTCCTGCTTGACAGAAGAAGCCTGTGGCT
GAGGCCATTGCAACAAGGACTATGATTGTGAAACTTGTTTCTCCGTGGATACCGGTTTCAGAAATCTTGTGCCCAGGTAC
CCACAGTCCTCCTCACATACCTCATTTTATAGAACTTTGCAAGTCCTGACATTCTGTCATGTGCTTTGTTATCCCACATA
GGGTAACAAAGATTTTTACCCATTTAAGAGTCTGGAAATAGGCTGGACCCACTGTTGACATTAAAGGTGCATTTCT

>338893 (H.sapiens)
CATTCATTTATTTCACATTTATTCTCATTGCACCAGGTGAGGAGAGGAAGGAGTCATTCACTAACACACACAGATTGTGC
TGTTTTTCAGTCTTTCTGATGAATCAGGATCCAGATTCTAAATGTCTTCAAGACCTGGATCAGTCAGTAGAGATGGCCCA
CTGTGTCAGGGGGCCTGGGGCTGCCGGAGGCAAAGCAGGATACATATATGGAACATGACCATTCATGTTCCAGGGCTCCC
ATCCGGGTACCTGAGGATTTTCCACATAGATCACAGGTGTTGGGGCCCACATCGGCTCCTGAAAGACTATGGGAACGCCA
TAGGCCTGGGGCGGGTGCTCGGGAGGTGCGAGGTTAACCACATCTGCACAGGGAGGACCTGAGA-CCTGAGAAGACGCTG
ATCTTCCATCTGTACCCCAGAATCCTGTGCAGGGCCTCACCATAAGGACATGGCAAGGGTTGGCACCTTGGCCTCTGGTT
GGCCTCCTGAATAGTGAAGTATAAGTCCTGCAAGCTTATTAGCATCTGGAGACATTCCTTCCAGCTCTTGTTGATACCCC
TCTGCTTGAGACGCTGAGCAATTGCTTTTGATACTATGTGATACTTCTTCTTCACCCTGTACACCTCACGTTCAAGAAAT
TCCCATTCTTGCAGGAAACTCCGGATTTCCTGGTCACTCCAAGGTTTAACTGACTGGACTG-TGAGGGCTTTTCTGATCC
CTGGGCTGTGTTTTCCTGCTCCATTTTCTGGATGTTTATGGTAGTTTCAGTGGGAAGTAT-TCCATTTTTACTCCTGAGT
TTTCTCCCGGGAGTCTTCTCACTGTGCCTGCAGGGTCTGGTCAGTTCCTGAGTTGGCGGAACACTGGCACTTACTCTCCT
CTAGTGGAACCTAGGAGAGTCAGGAGGAACCCGAGTGTGGAAATGTGCTTGCTTCTGTCGGCTTCT

>128365 (H.sapiens)
ATTATATAAATCTTCAACTTCTTGATCAAATAAATATGACAAATGATGTTCTCTAAGAAAAACACCCTTCAATTTTATTC
CT-CTTCCCTGCATATTTTGAGTAATTATCTTCCAAGACCCATGTATCTTTTCTCAACATCTCTGAGAGTACAATTCCT-
CTTATTTCCTTCATTGTGGCAAAGTGTTTCAGAAAAGGGTTCCTTGAATTAAAAGTCGGCGTATCCTATTTGACTCCTGC
TCCTCCGGTATCACATACCTACAGCCAACCATGCCAAGAGCTTCCCCATTATCTCCGCATCGGAGAGC-CTTTCCTCCCG
ATATCCTCCAGTTTCAGAGACCGCACCCGGAGACCCATTGGCAGGTTCCTGGATTCGCCTCAATTTTGGTCCTGCCTCTC
TGCTTCGCATTTTCAGGCTTGGCCTCACAAGAAGGACGATGGCGCCAGATTGTGCCAGAATGGGTGAAAACAGAAGGAAA
ATAAACCGGTTGCAGCAAAACCCACTATTCC
```

All the predicted gene sequences are avaliable on the web-site http://to444xl.to.infn.it/regexp2/

Sequence obtained by PCR amplification products direct sequencing.

```
>718689 (M.musculus)
CNNNCNNGTATAGGAGTCTGCTTGATATCCAGAACTAGATGTTACCTATAATATTAAAGTGAAAGCTATTCATGAACTCA
AGAGAAATGTAACTATAGAGGAAACTCTACAAATTTGTAAAATTGCACACAATATAAAATTGGATATGTTGATAGTTACA
TCATCTTGGGGAAATTTGGTATGTTACAATCACTCTCTTGAGGTGAAAATGAAAGTAACCTAGAGCCATGGGTAGAGATG
AGTAAATCTGAAGCTGTTCCCCCCAGTCTTAGTCTCAATTCTGCATAATGAAAATAGGGCATTTTAATTACAAAACTGAG
GATGGAATAACATACAGATATTTTCCAGAAAGTCAGCAAGGGTTCATTGAGTTCAGTAACAATAATGGCAGCTCCCCACG
AACTCTCCATACAGATTACATGAGCATGTTTTTATAGACAGACATATAGGATCCCTGGCTGTTTTACATGGACTGGAGAT
GTTGTGGGATTAGGGTATAGGACTCAGTACTAGAAAGGGATATTGGTAAAAGAATGTACCGTTTACTATATTTTTAGGAT
CACCTTGGCACACATTGCCATGGGACTGTATTCTCAAAGAAAAGAGAATTAAGTTTCTACAAGAAAAATTGAAATGTAGA
AAATTAGGACAGCCAGGAGAATTAAATACATGGTGTGCAGGTCAGACCCAGGAAGAATTGCAGAGTGCAGGGTTAGCAGT
ATGACCTCATTTTTGCCCTATTATGAACN

>1033649 (M.musculus)
CNCCCTTAGCATNAAANGTTCTGTTAGTATCATGAGACTGTTATCCATGTACAGAAGGCAAACAAAGAGACTCGCATATA
GATCCTGTAGCAACTCCACCAAAAGTTTTCCCCTGGAGACTCAAAAGCTCGAGGCCTGGGACACAAATATGAACTCATGC
AGGCTTGAAGATGATGGCCAAAAGATTCTTCCTCCATAGCTTCATGTAAAAGATGTGGTCCATGGTGAATTTCAGTTCTG
TCTCTAAGGGGAGACTAAGAAAAGACCATGAAGACGATCTGAATATTGAATTTGAAGTGAACCTCTGAGTGTCTCTTGTC
AGCAGTCATGCCAACTGCAGCACCTTTTATCATCCTCTGGTTTTCATCACAGCAAAGGTTGTTGGTTATGTACAGCAGAT
GAAGATCCCACTGGAAAGTCAAGTTTTTCCCAGAACATCAGTGGTCTTGCTCAGACTCTCTGATAAAATCCAGAGCTTGA
CCTTCATTCCTTTGACACTTGGTCCAACCTCCTCACTGAGTTCAGGATACTGGAACAGGCATGACCTAGTAGCACTGAAT
TGGACCTACTCAAAGAAGCACAGGCTAGCAGAGGAAAATGGCCTGAATGTCTGAGCAGTTGCTAACCTCTCTTGGAGTGA
GGGACTCCACTCCTGTAAAAA

>338893 (H.sapiens)
CCNNNNNNNTCAGTTAACCTTGGAGTGACCAGGAATCCGGAGTTTCCTGCAAGAATGGGAATTTCTTGAACGTGAGGTGT
ACAGGGTGAAGAAGAAGTATCACATAGTATCAAAAGCAATTGCTCAGCGTCTCAAGCAGAGGGGTATCAACAAGAGCTGG
AAGGAATGTCTCCAGATGCTAATAAGCTTGCAGGACTTATACTTCACTATTCAGGAGGCCAACCAGAGGCCAAGGTGCCA
ACCCTTGCCATGTCCTTATGGTGAGGCCCTGCACAGGATTCTGGGGTACAGATGGAAGATCAGCGTCTTCTCAGGTCCTC
CCTGTGCAGATGTGGTTAACCTCGCACCTCCCGAGCACCCGCCCCAGGCCTATGGCGTAAAA

>128365 (H.sapiens)
GNNANNTNTGAAANACCCTTCATTTTATTCCTTCCCTGCATATTTTGAGTAATTATCTTCCAAGACCCATGTATCTTTTC
TCAACATCTCTGAGAGTACAATTCCTTATTTCCTTCATTGTGGCAAAGTGTTTCAGAAAAGGGTTCCTTGAATTAAAAGT
CGGCGTATCCTATTTGACTCCTGCTCCTCCGGTATCACATACCTACAGCCAACCATGCCAAGAGCTTCCCCATTATCTCC
GCATCGGAGAGCCTTTCCTCCCGATATCCTCCAGTTTCAGAGACCGCACCCGGAGACCCATTGGCAGGTTCCTGGATTCG
CCTCAATTTTGGTCCTGCCTCTCTGCTTCGCATTTTCAGGCTTGGCCTCACAAGAAGGACGATGGCGCCAGATTGTGCCA
GAAGGGGTGAAAAA
```

B.5 Sequence of symbols discussed in the text

```
>c1239
CCAGAGAGACCAGGTTTCTGTAGTTCTCTAACATCGGATGCTAATACAAATTCTGCTGTGCAGTGCCCAGGCATTGCCAC
CCCTCTGGAGAGAATTCTATGGCTACATCCCTGAATGTCAACAGTTCCATTTCTAGGCTTGCAGCAGGTTCTGGTGTCTT
AGCTATGGATCTCCCAATACCTGCTGGTC

>c804
AATAAAACAGGTATTGCTGTCTCTAAGCCAGACTTGATCACCTGTCTGGAGCAAAAAAAAGAGCCCTGGAATATAAAGAG
ACATGAGATGGTAGCCAAACCCCCAGGTAGGTGAGAGTGAATGAAGCAGATGACACAGATGAGAGGTACAAAAGTCAAAG
AGGAAGCCAGTCCTTAAAATGTGGTTTGGGAAGCTGTGCTCCAATGGAAATAGTTTCTG

>c307
TCTATCTTGAACGAACATCACATTAAATGTGTTTGCAAAATTACCTGTCCCAGATAGTTGTCCATCCTTTATTTCTGTGG
CCATATTCGAAACAGAATCTTCCTCGTCACTTGTAGCCTGAATGGAATTTGAAACAAAACAATCAATAAATAAAGTAGGT
TTCATAGACTATACAGTTAATAGTTCAAAATATAAATGAGACTTTAATTACCTTCAAAGCTGGTTGTTTATGAGAAGACA
CTGAAAAGCAAAAGGGATACATAATCACTCATACGTAAATATGATAAAGTTATCCATACATTCATACAGTGTTAGCATCA
AACTCTATCCTCCTGCCTGTAATAGTGTAGGCTTTGATGGCTTCTACTTTGTGTCTGGAGACAAGAACATGACAGAAATA
CACT

>c417
AAACTCTTCCTTACGTTAGCCATGAAATCTAGCTGGGGCTGTGTGGTTTCTGATTCCCCCTGGCTTATTCTTTACTTTTT
CCCACTTTTCCAGGCTCAGCAGGGAGCTGCTGGATGAGAAAGGGCCTGAAGTCTTGCAGGACTCACTGGATAGATGTTAT
TCAACTCCTTCAGGTTGTCTTGAACTGACTGACTCATGCCAGCCCTACAGAAGTGCCTTTTACGTATTGGAGCAACAGCG
TGTTGGCTTGGCTGTTGACATGGATGGTGAGTACCTTTCTATGAAGGTGATAAGGATCCACTGAGTCTTCTGGTTAGGGT
CATATTCCTACTGCAAGTGGCCCTTACTGAGCTGAGAGATGTCATTGCCACAGGGAGGACCTATAGGCACATGTAGGTTG
AATGAAACTCTAGTTCCACTTGGAAGCCCAGACAAGGGATGGGTCAGTGAGCAAGGCTCTCTTCCTAGTCTCAGGCCATG
CCTGTGGCGCCCTAATCCTACTCTCATGACGTTGGACCTGGGCAGATGTGACAAATTCACACAACTCTGATTTTGTCTCA
ATTTTGTAGATCTTGTAGATTTCATCCTTCACTCTAATTTCAGCGTCTAAAATCCTCGCTACCATGAACAATCTGAGTAT
TTGATGAGACAGGGCTGAATAGTGCAGTTTTTCTCCTAGCAACCATTTGGGGGCATTTGCTTTAAATCGATTGGAAAAAT
ATGGCATAACCATTTGCACAAACTTGGGACAAATGATATTGGGATAACGATCTACCAGAATAGGGAATTTTACCCACAGT
TTCTGGGACAAAAACCAAGGAATCTCTATGGTGATCAGCCTTCAGGCCTCCTGAAGACTATCTCTCACAGTGTCCTATTC
TCATGCTGAGGAGCCTGAAGTCCCTGTGTGAGGATTAGACAGTGGATTGTTATGTGTGTAGGAGAACCAGCTTAATATGT
CTGTCCATGTCTGAACTTATTGCAGAAATTGAAAAGTACCAAGAAGTGGAAGAAGACCAAGACCCATCATGCCCCAGGTA
ACTTTGAGCAATTATGGATGCTTAATTCTGTGTTGACACCTGGAGATGCCAGGTCCAGGGAAAACAAGAGTGTGTTCAAT
TTCATGTTTTCAACGAAGGTTGAATTACTCCTACTGACATTGCTGTTGGTTTTCATTGCAGTAGATGTTTAGGTTTCCAT
TTCTTCCTCCCCTTATCATTTACTAACTTACTATAGGTTGACCATACCTCAAAGGCTGTATGGCAACTGCATGGAATCTT
GAGCAAGTTTATGGAAAATTATTGAGCCCACTCTTTTCATGATCACTGTTCGCTGTGTGTCCCGAGGGCACTAACTCAGA
GTGTCCTTTGACCCCTTCATCAGTGTGTCACCCGGCCAATTCGCTGAGCTCAC

>c1359
TACCTATATCCTCTAGAGGAATGTTCATCCCAACTAGAATGACCATAATCACGGTATGCATAGCCTCTAGATGGTGGAGC
ATAATCCCTAGTTTCTCGGGAACTTGGATGATTTCTGTGTGCATAAGTTTAAGCAA

>c690
ACCTTTCTTTTCAGGCATTTCCTGCTTATCCAAGTTCACCATTTCAGGTCACCACTGGATATCAGTTGCCTGTATATAAT
TATCAGGTAATGTAAGAAGGAGTAAAATTATTTGCTTTCAGGTATTATTGAGGCCTTTAACTTGTTTATACAAATTTCCG
GAATAGTTGGTCATTTTAAACTAGTGAAGTGTACCTAAAATTTAAGGAAACACTTAGAATTAGTGTAGAATGAAGACCTC
TGTCTTATTGAGAAGTAATGAAGTCGAATTTTGACAGGAATATACTTGGGAATAACTTTCCTGTAGAACAGATTTCTGAG
ATTTGGTGTCCCATTCTTCATTTCTGGATGTAGTTTTCATCTTTACTGTCAAATAACTGAATGAAACATCCAAACTGACT
TTCATGAATTTTCTTAGGGAGATAGAGT
```

All the symbols sequences are avaliable on the web-site
http://www.to.infn.it/∼molineri/tesi/additional-files/symbols.grouped.fa

Acknowledgements

I would like to thanks...

Michele Caselle for all the ideas and friendly tutoring.

Gabriele Sales which worked whit me side by side.

Chiara Peyron for the great help in the bioinfomatic and statistical stuff.

Ferdinando Di Cunto, Sara Zanivan, Federico Bianchi, Ilaria Cascone and Serena Marchiò for the biological experiments and to give my the opportunity to appreciate the biological point of view.

All the students and all the professors of the Ph.D. in "Complex Systems Applied to Post-Genomic Biology" for the useful discussions.

Michele Caselle, Federico Bussolino, Istituto Nazionale di Fisica Nucleare (INFN), Università di Torino, Istituto per la Ricerca e Cura del Cancro (IRCC) and Ministero per l'Istruzone l'Università e la Ricerca (MIUR) for the financial support.

Bibliography

[1] Mike Jones. Flow of information in biological systems., 2007. [Wikipedia; accessed 11-September-2007].

[2] F. Crick. On protein synthesis. *Symposia of the Society for Experimental Biology*, 12:138–163, xx 1958.

[3] F. Crick. Central dogma of molecular biology. *Nature*, 227:561–563, Aug 1970.

[4] B. McCarthy and J. Holland. Denatured dna as a direct template for in vitro protein synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 54:880–886, Sep 1965.

[5] C. Napoli, C. Lemieux, and R. Jorgensen. Introduction of a chimeric chalcone synthase gene into petunia results in reversible co-suppression of homologous genes in trans. *THE PLANT CELL*, 2:279–289, Apr 1990. 10.1105/tpc.2.4.279.

[6] Gerald Weissmann. Americans sweep nobel prizes: Thinking inside and outside the box. *The FASEB Journal*, 21:1–4, Jan 2007. 10.1096/fj.07-0101.

[7] Elizabeth Pennisi. Genomics. dna study forces rethink of what it means to be a gene. *Science (New York, N.Y.)*, 316:1556–7, Jun 2007. 10.1126/science.316.5831.1556.

[8] Alyson Ashe and Emma Whitelaw. Another role for rna: a messenger across generations. *Trends in genetics : TIG*, 23:8–10, Jan 2007. 10.1016/j.tig.2006.11.008.

[9] Mark Gerstein, Can Bruce, Joel Rozowsky, Deyou Zheng, Jiang Du, Jan Korbel, Olof Emanuelsson, Zhengdong Zhang, Sherman Weissman, and Michael Snyder. What is a gene, post-encode?: History and updated definition. *Genome research*, 17:669–81, Jun 2007. 10.1101/gr.6339607.

[10] J. Heimans. Hugo de vries and the gene concept. *American Naturalist*, 96:93, xx 1962. 10.1086/282210.

[11] Nils Roll-Hansen. The crucial experiment of wilhelm johannsen. *Biology and Philosophy*, 4:303–29, xx 1989.

[12] Hans-Jorg Rheinberger. When did carl correns read gregor mendel's paper?: A research note. *Isis*, 86:612, xx 1995. 10.1086/357321.

[13] Thomas hunt morgan - wikipedia.

[14] Barbara McClintock. A cytological and genetical study of triploid maize. *Genetics*, 14:180–222, Mar 1929.

[15] K. Manchester. Theodor boveri and the origin of malignant tumours. *Trends in cell biology*, 5:384–7, Oct 1995.

[16] G. Beadle and E. Tatum. Genetic control of biochemical reactions in neurospora. *Proceedings of the National Academy of Sciences*, 27:499–506, Nov 1941. 10.1073/pnas.27.11.499.

[17] H. J. Muller. Artificial Transmutation of the Gene. *Science*, 66:84–87, July 1927.

[18] Griffith's experiment - wikipedia.

[19] Oswald Avery, Colin MacLeod, and Maclyn McCarty. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type iii. *The Journal of Experimental Medicine*, 79:137–58, Feb 1944. 10.1084/jem.79.2.137.

[20] A. Hershey. An upper limit to the protein content of the germinal substance of bacteriophage t2. *Virology*, 1:108–27, May 1955.

[21] J. Watson and F. Crick. Genetical implications of the structure of deoxyribonucleic acid. *Nature*, 171:964–7, May 1953.

[22] M. Nirenberg, P. Leder, M. Bernfield, R. Brimacombe, J. Trupin, F. Rottman, and C. O'Neal. Rna codewords and protein synthesis, vii. on the general nature of the rna code. *Proceedings of the National Academy of Sciences*, 53:1161–8, May 1965. 10.1073/pnas.53.5.1161.

[23] Francis Crick. Ideas on protein synthesis. *Symp. Soc. Exp. Biol.*, XII:138–63, xx 1956.

[24] W. Fiers, R. Contreras, F. Duerinck, G. Haegeman, D. Iserentant, J. Merregaert, W. Jou, F. Molemans, A. Raeymaekers, A. van den Berghe, G. Volckaert, and M. Ysebaert. Complete nucleotide sequence of bacteriophage ms2 rna: primary and secondary structure of the replicase gene. *Nature*, 260:500–7, Apr 1976.

[25] Paul Griffiths and Karola Stotz. Genes in the postgenomic era. *Theoretical medicine and bioethics*, 27:499–521, xx 2006. 10.1007/s11017-006-9020-y.

[26] Hester Wain, Elspeth Bruford, Ruth Lovering, Michael Lush, Mathew Wright, and Sue Povey. Guidelines for human gene nomenclature. *Genomics*, 79:464–70, Apr 2002. 10.1006/geno.2002.6748.

[27] Helen Pearson. Genetics: what is a gene? *Nature*, 441:398–401, May 2006. 10.1038/441398a.

[28] R. Fleischmann, M. Adams, O. White, R. Clayton, E. Kirkness, A. Kerlavage, C. Bult, J. Tomb, B. Dougherty, J. Merrick, and and. Whole-genome random sequencing and assembly of haemophilus influenzae rd. *Science*, 269:496–512, Jul 1995. 10.1126/science.7542800.

[29] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, and P. Predki. Initial sequencing and analysis of the human genome. *Nature.*, 409:860–921, Feb 2001. 10.1038/35057062.

[30] J. Venter, Mark Adams, Eugene Myers, Peter Li, Richard Mural, Granger Sutton, Hamilton Smith, Mark Yandell, Cheryl Evans, Robert Holt, Jeannine Gocayne, Peter Amanatides, Richard Ballew, Daniel Huson, Jennifer Wortman, Qing Zhang, Chinnappa Kodira, Xiangqun Zheng, Lin Chen, Marian Skupski, Gangadharan Subramanian, Paul Thomas, Jinghui Zhang, George Gabor, Catherine Nelson, Samuel Broder, Andrew Clark, Joe Nadeau, Victor McKusick, Norton Zinder, Arnold Levine, Richard Roberts, Mel Simon, Carolyn Slayman, Michael Hunkapiller, Randall Bolanos, Arthur Delcher, Ian Dew, Daniel Fasulo, Michael Flanigan, Liliana Florea, Aaron Halpern, Sridhar Hannenhalli, Saul Kravitz, Samuel Levy, Clark Mobarry, Knut Reinert, Karin Remington, Jane Abu-Threideh, Ellen Beasley, Kendra Biddick, Vivien Bonazzi, Rhonda Brandon, Michele Cargill, Ishwar Chandramouliswaran, Rosane Charlab, Kabir Chaturvedi, Zuoming Deng, Valentina di Francesco, Patrick Dunn, Karen Eilbeck, Carlos Evangelista, Andrei Gabrielian, Weiniu Gan, Wangmao Ge, Fangcheng Gong, Zhiping Gu, Ping Guan, Thomas Heiman, Maureen Higgins, Rui-Ru Ji, Zhaoxi Ke, Karen Ketchum, Zhongwu Lai, Yiding Lei, Zhenya Li, Jiayin Li, Yong Liang, Xiaoying Lin, Fu Lu, Gennady Merkulov, Natalia Milshina, Helen Moore, Ashwinikumar Naik, and Vaibhav Narayan. The sequence of the human genome. *Science*, 291:1304–51, Feb 2001. 10.1126/science.1058040.

[31] Nicholas Wade. Gene sweepstakes ends, but winner may well be wrong - new york times. *New York Times*, Jun 2003.

[32] Gene sweep 2000-2003.

[33] Deyou Zheng, Adam Frankish, Robert Baertsch, Philipp Kapranov, Alexandre Reymond, Siew Choo, Yontao Lu, France Denoeud, Stylianos Antonarakis, Michael Snyder, Yijun Ruan, Chia-Lin Wei, Thomas Gingeras, Roderic GuigÃ?, Jennifer Harrow, and Mark Gerstein. Pseudogenes in the encode regions: consensus annotation, analysis of transcription, and evolution. *Genome research*, 17:839–51, Jun 2007. 10.1101/gr.5586307.

[34] Yoshihide Hayashizaki and Piero Carninci. Genome network and fantom3: Assessing the complexity of the transcriptome. *PLoS Genetics*, 2:63, Apr 2006. 10.1371/journal.pgen.0020063.

[35] Fantom3::databases.

[36] Encode - wikipedia.

[37] Encode pilote project news release.

[38] Piero Carninci. Tagging mammalian transcription complexity. *Trends in genetics : TIG*, 22:501–10, Sep 2006. 10.1016/j.tig.2006.07.003.

[39] J. Timmons and L. Good. Does everything now make (anti)sense? *Biochemical Society transactions*, 34:1148–50, Dec 2006. 10.1042/BST0341148.

[40] Zhengdong Zhang, Alberto Paccanaro, Yutao Fu, Sherman Weissman, Zhiping Weng, Joseph Chang, Michael Snyder, and Mark Gerstein. Statistical analysis of the genomic distribution and correlation of regulatory elements in the encode regions. *Genome Research*, 17:787–97, Jun 2007. 10.1101/gr.5573107.

[41] Jun Wang, Shengting Li, Yong Zhang, Hongkun Zheng, Zhao Xu, Jia Ye, Jun Yu, and Gane Wong. Vertebrate gene predictions and the problem of large genes. *Nature Reviews Genetics*, 4:741–9, Sep 2003. 10.1038/nrg1160.

[42] T. Hubbard, B. Aken, K. Beal, B. Ballester, M. Caccamo, Y. Chen, L. Clarke, G. Coates, F. Cunningham, T. Cutts, T. Down, S. Dyer, S. Fitzgerald, J. Fernandez-Banet, S. Graf, S. Haider, M. Hammond, J. Herrero, R. Holland, K. Howe, N. Johnson, A. Kahari, D. Keefe, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, C. Melsopp, K. Megy, P. Meidl, B. Ouverdin, A. Parker, A. Prlic, S. Rice, D. Rios, M. Schuster, I. Sealy, J. Severin, G. Slater, D. Smedley, G. Spudich, S. Trevanion, A. Vilella, J. Vogel,

S. White, M. Wood, T. Cox, V. Curwen, R. Durbin, X. Fernandez-Suarez, P. Flicek, A. Kasprzyk, G. Proctor, S. Searle, J. Smith, A. Ureta-Vidal, and E. Birney. Ensembl 2007. *Nucleic acids research*, 35:610–7, Jan 2007. 10.1093/nar/gkl996.

[43] William Pearson and David Lipman. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences*, 85:2444–8, Apr 1988. 10.1073/pnas.85.8.2444.

[44] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman. Basic local alignment search tool. *Journal of molecular biology.*, 215:403–10, Oct 1990. 10.1006/jmbi.1990.9999.

[45] Catherine MathÃ?, Marie-France Sagot, Thomas Schiex, and Pierre RouzÃ? Current methods of gene prediction, their strengths and weaknesses. *Nucleic acids research*, 30:4103–17, Oct 2002.

[46] J. Fickett. Orfs and genes: how strong a connection? *Journal of computational biology : a journal of computational molecular cell biology*, 2:117–23, xx 1995.

[47] G B Hutchinson and M R Hayden. The prediction of exons through an analysis of spliceable open reading frames. *Nucleic Acids Res*, 20(13):3453–62, July 1992.

[48] T A Ronneberg, S J Freeland, and L F Landweber. Genview and gencode : a pair of programs to test theories of genetic code evolution. *Bioinformatics*, 17(3):280–1, March 2001.

[49] Chun-Ting Zhang and Ren Zhang. Evaluation of gene-finding algorithms by a content-balancing accuracy index. *J Biomol Struct Dyn*, 19(6):1045–52, June 2002.

[50] E E Snyder and G D Stormo. Identification of protein coding regions in genomic dna. *J Mol Biol*, 248(1):1–18, April 1995.

[51] Victor V. Solovyev and Asaf A. Salamov. The gene-finder computer tools for analysis of human and model organisms genome sequences. In Terry Gaasterland, Peter D. Karp, Kevin Karplus, Christos A. Ouzounis, Chris Sander, and Alfonso Valencia, editors, *ISMB*, pages 294–302. AAAI, 1997.

[52] M Borodovsky and A Peresetsky. Deriving non-homogeneous dna markov chain models by cluster analysis algorithm minimizing multiple alignment entropy. *Comput Chem*, 18(3):259–67, September 1994.

[53] C. Burge and S. Karlin. Prediction of complete gene structures in human genomic dna. *Journal of molecular biology*, 268:78–94, Apr 1997. 10.1006/jmbi.1997.0951.

[54] Olivier Gascuel and Marie-France Sagot, editors. *Computational Biology, First International Conference on Biology, Informatics, and Mathematics, JOBIM 2000, Montpellier, France, May 3-5, 2000, Selected Papers*, volume 2066 of *Lecture Notes in Computer Science*. Springer, 2001.

[55] Arthur L. Delcher, Kirsten A. Bratke, Edwin C. Powers, and Steven Salzberg. Identifying bacterial genes and endosymbiont dna with glimmer. *Bioinformatics*, 23(6):673–679, 2007.

[56] A V Lukashin and M Borodovsky. Genemark.hmm: new solutions for gene finding. *Nucleic Acids Res*, 26(4):1107–15, February 1998.

[57] G Bernardi. The isochore organization of the human genome. *Annu Rev Genet*, 23:637–61, 1989.

[58] G Matassi, L M Montero, J Salinas, and G Bernardi. The isochore organization and the compositional distribution of homologous coding sequences in the nuclear genome of plants. *Nucleic Acids Res*, 17(13):5273–90, July 1989.

[59] L Duret, D Mouchiroud, and C Gautier. Statistical analysis of vertebrate sequences reveals that long genes are scarce in gc-rich isochores. *J Mol Evol*, 40(3):308–17, March 1995.

[60] Alexander Churbanov, Igor B Rogozin, Jitender S Deogun, and Hesham Ali. Method of predicting splice sites based on signal interactions. *Biol Direct*, 1:10, 2006.

[61] S. Brunak, J. Engelbrecht, and S. Knudsen. Prediction of human mrna donor and acceptor sites from the dna sequence. *Journal of molecular biology*, 220:49–65, Jul 1991.

[62] M. Zhang and T. Marr. A weight array method for splicing signal analysis. *Computer applications in the biosciences : CABIOS*, 9:499–509, Oct 1993.

[63] S. Salzberg. A method for identifying splice sites and translational start sites in eukaryotic mrna. *Computer applications in the biosciences : CABIOS*, 13:365–76, Aug 1997.

[64] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77:257, xx 1989. 10.1109/5.18626.

[65] Donald J. Patterson, Ken Yasuhara, and Walter L. Ruzzo. Pre-mrna secondary structure prediction aids splice site prediction. In *Pacific Symposium on Biocomputing*, pages 223–234, 2002.

[66] Viterbi algorithm - wikipedia.

[67] José Oliver, Pedro Carpena, Michael Hackenberg, and Pedro Bernaola-Galván. Isofinder: computational prediction of isochores in genome sequences. *Nucleic acids research*, 32:287–92, Jul 2004. 10.1093/nar/gkh399.

[68] N. Proudfoot. Pseudogenes. *Nature*, 286:840–1, Aug 1980.

[69] Ilenia D'Errico, Gemma Gadaleta, and Cecilia Saccone. Pseudogenes in metazoa: origin and features. *Briefings in functional genomics & proteomics*, 3:157–67, Aug 2004.

[70] Gene conversion - wikipedia.

[71] D. Graur, Y. Shuali, and W. Li. Deletions in processed pseudogenes accumulate faster in rodents than in humans. *Journal of molecular evolution*, 28:279–285, Apr 1989.

[72] C. Saccone. The evolution of mitochondrial dna. *Current opinion in genetics & development*, 4:875–881, Dec 1994.

[73] J. Lopez, N. Yuhki, R. Masuda, W. Modi, and S. O'Brien. Numt, a recent transfer and tandem amplification of mitochondrial dna to the nuclear genome of the domestic cat. *Journal of molecular evolution*, 39:174–190, Aug 1994.

[74] J. Nugent and J. Palmer. Rna-mediated transfer of the gene coxii from the mitochondrion to the nucleus during flowering plant evolution. *Cell*, 66:473–481, Aug 1991.

[75] Markus Woischnik and Carlos Moraes. Pattern of organization of human mitochondrial pseudogenes in the nuclear genome. *Genome research*, 12:885–893, Jun 2002. 10.1101/gr.227202. Article published online before print in May 2002.

[76] D. Bensasson, D. Zhang, D. Hartl, and G. Hewitt. Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends Ecol Evol*, 16:314–321, Jun 2001.

[77] Marijke van Baren and Michael Brent. Iterative gene prediction and pseudogene removal improves genome annotation. *Genome research*, 16:678–85, May 2006. 10.1101/gr.4766206.

[78] Patrick Ng, Chia-Lin Wei, Wing-Kin Sung, Kuo Chiu, Leonard Lipovich, Chin Ang, Sanjay Gupta, Atif Shahab, Azmi Ridwan, Chee Wong, Edison Liu, and Yijun Ruan. Gene identification signature (gis) analysis for transcriptome characterization and genome annotation. *Nature methods*, 2:105–11, Feb 2005. 10.1038/nmeth733.

[79] Jennifer Harrow, France Denoeud, Adam Frankish, Alexandre Reymond, Chao-Kung Chen, Jacqueline Chrast, Julien Lagarde, James Gilbert, Roy Storey, David Swarbreck, Colette Rossier, Catherine Ucla, Tim Hubbard, Stylianos Antonarakis, and Roderic Guigo. Gencode: producing a reference annotation for encode. *Genome biology*, 7 Suppl 1:4–1, xx 2006. 10.1186/gb-2006-7-s1-s4.

[80] Deyou Zheng and Mark Gerstein. A computational approach for identifying pseudogenes in the encode regions. *Genome biology*, 7 Suppl 1:13–1, xx 2006. 10.1186/gb-2006-7-s1-s13.

[81] W. Kent, Robert Baertsch, Angie Hinrichs, Webb Miller, and David Haussler. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 100:11484–9, Sep 2003. 10.1073/pnas.1932072100.

[82] W. Li, T. Gojobori, and M. Nei. Pseudogenes as a paradigm of neutral evolution. *Nature*, 292:237–9, Jul 1981.

[83] Evgeniy S. Balakirev and Francisco J. Ayala. Pseudogenes: are they "junk" or functional dna? *Annual review of genetics*, 37:123–151, xx 2003. 10.1146/annurev.genet.37.040103.103949.

[84] A. Mighell, N. Smith, P. Robinson, and A. Markham. Vertebrate pseudogenes. *FEBS letters*, 468:109–114, Feb 2000.

[85] Sergei Korneev, Ji-Ho Park, and Michael O'Shea. Neuronal expression of neural nitric oxide synthase (nnos) protein is suppressed by an antisense rna transcribed from an nos pseudogene. *Journal of Neuroscience*, 19:7711–7720, Sep 1999.

[86] T. Ota and M. Nei. Evolution of immunoglobulin vh pseudogenes in chickens. *Molecular biology and evolution*, 12:94–102, Jan 1995.

[87] Örjan Svensson, Lars Arvestad, and Jens Lagergren. Genome-wide survey for biologically functional pseudogenes. *PLoS Computational Biology*, 2:46–46, May 2006. 10.1371/journal.pcbi.0020046.

[88] Yoshihisa Yano, Rintaro Saito, Noriyuki Yoshida, Atsushi Yoshiki, Anthony Wynshaw-Boris, Masaru Tomita, and Shinji Hirotsune. A new role for expressed pseudogenes as ncrna: regulation of mrna stability of its homologous coding gene. *Journal of molecular medicine (Berlin, Germany)*, 82:414–422, Jul 2004. 10.1007/s00109-004-0550-3.

[89] Paul Harrison, Deyou Zheng, Zhaolei Zhang, Nicholas Carriero, and Mark Gerstein. Transcribed processed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein-coding ability. *Nucleic acids research*, 33:2374–2383, xx 2005. 10.1093/nar/gki531.

[90] Deyou Zheng, Zhaolei Zhang, Paul Harrison, John Karro, Nick Carriero, and Mark Gerstein. Integrated pseudogene annotation for human chromosome 22: evidence for transcription. *Journal of molecular biology*, 349:27–45, May 2005. 10.1016/j.jmb.2005.02.072.

[91] Martin Frith, Laurens Wilming, Alistair Forrest, Hideya Kawaji, Sin Tan, Claes Wahlestedt, Vladimir Bajic, Chikatoshi Kai, Jun Kawai, Piero Carninci, Yoshihide Hayashizaki, Timothy Bailey, and Lukasz Huminiecki. Pseudo-messenger rna: phantoms of the transcriptome. *PLoS genetics*, 2:e23, Apr 2006. 10.1371/journal.pgen.0020023.

[92] Julie Bradley, Andrew Baltus, Helen Skaletsky, Morgan Royce-Tolland, Ken Dewar, and David Page. An x-to-autosome retrogene is required for spermatogenesis in mice. *Nature genetics*, 36:872–876, Aug 2004. 10.1038/ng1390.

[93] Nicolas Vinckenbosch, Isabelle Dupanloup, and Henrik Kaessmann. Evolutionary fate of retroposed gene copies in the human genome. *Proceedings of the National Academy of Sciences of the United States of America.*, 103:3220–5, Feb 2006.

[94] A. Hill, R. Nicholls, S. Thein, and D. Higgs. Recombination within the human embryonic xi-globin locus: a common xi-xi chromosome produced by gene conversion of the psi xi gene. *Cell*, 42:809–19, Oct 1985.

[95] Nathalie Trabesinger-Ruefa, Thomas Jermanna, Todd Zankela, Barbara Durrantc, Gerhard Frankb, and Steven Benner. Pseudogenes in ribonuclease evolution: a source of new biomacromolecular function? *FEBS Letters*, 382:319, xx 1996. 10.1016/0014-5793(96)00191-3.

[96] Paul Harrison and Mark Gerstein. Studying genomes through the aeons: protein families, pseudogenes and proteome evolution. *Journal of molecular biology*, 318:1155–74, May 2002.

[97] Slim Sassi, Edward Braun, and Steven Benner. The evolution of seminal ribonuclease: pseudogene reactivation or multiple gene inactivation events? *Molecular biology and evolution*, 24:1012–24, Apr 2007. 10.1093/molbev/msm020.

[98] Todd Gray, Alison Wilson, Patrick Fortin, and Robert Nicholls. The putatively functional mkrn1-p1 pseudogene is neither expressed nor imprinted, nor does it regulate its source gene in trans. *Proceedings of the National Academy of Sciences*, 103:12039–12044, Aug 2006. 10.1073/pnas.0602216103.

[99] Shinji Hirotsune, Noriyuki Yoshida, Amy Chen, Lisa Garrett, Fumihiro Sugiyama, Satoru Takahashi, Ken-Ichi Yagami, Anthony Wynshaw-Boris, and Atsushi Yoshiki. An expressed pseudogene regulates the messenger-rna stability of its homologous coding gene. *Nature*, 423:91–96, May 2003. 10.1038/nature01535.

[100] Jeannie Lee. Complicity of gene and pseudogene. *Nature*, 423:26–28, May 2003. 10.1038/423026a.

[101] Jr Haig Kazazian. Mobile elements: drivers of genome evolution. *Science*, 303:1626–1632, Mar 2004. 10.1126/science.1089670.

[102] E. Lander, L. Linton, B. Birren, C. Nusbaum, M. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. Fitzhugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. Levine, P. McEwan, K. McKernan, J. Meldrim, J. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. Waterston, R. Wilson, L. Hillier, J. McPherson, M. Marra, E. Mardis, L. Fulton, A. Chinwalla, K. Pepin, W. Gish, S. Chissoe, M. Wendl, K. Delehaunty, T. Miner, A. Delehaunty, J. Kramer, L. Cook, R. Fulton, D. Johnson, P. Minx, S. Clifton, T. Hawkins, E. Branscomb, and P. Initial sequencing and analysis of the human genome. *Nature.*, 409:860–921, Feb 2001. 10.1038/35057062.

[103] Robert Waterston, Kerstin Lindblad-Toh, Ewan Birney, Jane Rogers, Josep Abril, Pankaj Agarwal, Richa Agarwala, Rachel Ainscough, Marina Alexandersson, Peter An, Stylianos Antonarakis, John Attwood, Robert Baertsch, Jonathon Bailey, Karen Barlow, Stephan Beck, Eric Berry, Bruce Birren, Toby Bloom, Peer Bork, Marc Botcherby, Nicolas Bray, Michael Brent, Daniel Brown, Stephen Brown, Carol Bult, John Burton, Jonathan Butler, Robert Campbell, Piero Carninci, Simon Cawley, Francesca Chiaromonte, Asif Chinwalla, Deanna Church, Michele Clamp, Christopher Clee, Francis Collins, Lisa Cook, Richard Copley, Alan Coulson, Olivier Couronne, James Cuff, Val Curwen, Tim Cutts, Mark Daly, Robert David, Joy Davies, Kimberly Delehaunty, Justin Deri, Emmanouil Dermitzakis, Colin Dewey, Nicholas Dickens, Mark Diekhans, Sheila Dodge, Inna Dubchak, Diane Dunn, Sean Eddy, Laura Elnitski, Richard Emes, Pallavi Eswara, Eduardo Eyras, Adam Felsenfeld, Ginger Fewell, Paul Flicek, Karen Foley, Wayne Frankel, Lucinda Fulton, Robert Fulton, Terrence Furey, Diane Gage, Richard Gibbs, Gustavo Glusman, Sante Gnerre, Nick Goldman, Leo Goodstadt, Darren Grafham, Tina Graves, Eric Green, Simon Gregory, Roderic Guigó, Mark Guyer, Ross Hardison, David Haussler, and Yoshihide Hayashizaki. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420:520–562, Dec 2002. 10.1038/nature01262.

[104] Phillip Sanmiguel, Alexander Tikhonov, Young-Kwan Jin, Natasha Motchoulskaia, Dmitrii Zakharov, Admasu Melake-Berhan, Patricia Springer, Keith Edwards, Michael Lee, Zoya Avramova, and Jeffrey Bennetzen. Nested retrotransposons in the intergenic regions of the maize genome. *Science*, 274:765–768, Nov 1996. 10.1126/science.274.5288.765.

[105] M. Curcio and Keith Derbyshire. The outs and ins of transposition: from mu to kangaroo. *Nature reviews. Molecular cell biology*, 4:865–877, Nov 2003.

[106] D. Chalker and S. Sandmeyer. Ty3 integrates within the region of rna polymerase iii transcription initiation. *Genes and Development*, 6:117–128, Jan 1992. 10.1101/gad.6.1.117.

[107] Yunxia Zhu, Junbiao Dai, Peter Fuerst, and Daniel Voytas. From the cover: Controlling integration specificity of a yeast retrotransposon. *Proceedings of the National Academy of Sciences*, 100:5891–5895, May 2003. 10.1073/pnas.1036705100.

[108] J. Jakubczak, Y. Xiong, and T. Eickbush. Type i (r1) and type ii (r2) ribosomal dna insertions of drosophila melanogaster are retrotransposable elements closely related to those of bombyx mori. *Journal of molecular biology*, 212:37–52, Mar 1990.

[109] Mary-Lou Pardue and P. G. DeBaryshe. Retrotransposons provide an evolutionarily robust non-telomerase mechanism to maintain telomeres. *Annual review of genetics*, 37:485–511, xx 2003. 10.1146/annurev.genet.38.072902.093115.

[110] H. Takahashi, S. Okazaki, and H. Fujiwara. A new family of site-specific retrotransposons, sart1, is inserted into telomeric repeats of the silkworm, bombyx mori. *Nucleic acids research*, 25:1578–1584, Apr 1997.

[111] Q. Feng, J. Moran, Jr H. Kazazian, and J. Boeke. Human l1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell*, 87:905–916, Nov 1996.

[112] Jerzy Jurka. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proceedings of the National Academy of Sciences*, 94:1872–1877, Mar 1997. 10.1073/pnas.94.5.1872.

[113] D. Luan, M. Korman, J. Jakubczak, and T. Eickbush. Reverse transcription of r2bm rna is primed by a nick at the chromosomal target site: a mechanism for non-ltr retrotransposition. *Cell*, 72:595–605, Feb 1993.

[114] Gregory Cost, Qinghua Feng, Alain Jacquier, and Jef Boeke. Human l1 element target-primed reverse transcription in vitro. *The EMBO journal*, 21:5899–5910, Nov 2002.

[115] Eric Ostertag and Jr. Kazazian. Twin priming: A proposed mechanism for the creation of inversions in l1 retrotransposition. *Genome Research*, 11:2059–2065, Dec 2001. 10.1101/gr.205701.

[116] Carl Schmid. Alu: a parasite's parasite? *Nature genetics*, 35:15–16, Sep 2003. 10.1038/ng0903-15.

[117] Marie Dewannieux, Cecile Esnault, and Thierry Heidmann. Line-mediated retrotransposition of marked alu sequences. *Nat Genet*, 35:41–48, Sep 2003. 10.1038/ng1223.

[118] Harvey Lodish, Arnold Berk, Lawrence S. Zipursky, Paul Matsudaira, David Baltimore, and James Darnell. *Molecular Cell Biology*.

[119] C. Esnault, J. Maestre, and T. Heidmann. Human line retrotransposons generate processed pseudogenes. *Nature genetics*, 24:363–367, Apr 2000. 10.1038/74184.

[120] J. Boeke. Lines and alus–the polya connection. *Nature genetics*, 16:6–7, May 1997. 10.1038/ng0597-6.

[121] R. Dornburg and H. Temin. cdna genes formed after infection with retroviral vector particles lack the hallmarks of natural processed pseudogenes. *Molecular and cellular biology*, 10:68–74, Jan 1990.

[122] H. Hohjoh and M. Singer. Cytoplasmic ribonucleoprotein complexes containing human line-1 protein and rna. *The EMBO journal*, 15:630–639, Feb 1996.

[123] Yosuke Ejima and Lichun Yang. Trans mobilization of genomic dna as a mechanism for retrotransposon-mediated exon shuffling. *Human molecular genetics*, 12:1321–1328, Jun 2003.

[124] T. Eickbush. Exon shuffling in retrospect. *Science (New York, N.Y.)*, 283:1465–1467, Mar 1999.

[125] J. Moran, R. Deberardinis, and Jr H. Kazazian. Exon shuffling by l1 retrotransposition. *Science (New York, N.Y.)*, 283:1530–1534, Mar 1999.

[126] Wentian Li. Computational gene recognition.

[127] Ensembl.

[128] Tadatsugu Taniguchi and Akinori Takaoka. The interferon-alpha/beta system in antiviral responses: a multimodal machinery of gene regulation by the irf family of transcription factors. *Current opinion in immunology*, 14:111–6, Feb 2002.

[129] Ronen Shemesh, Amit Novik, Sarit Edelheit, and Rotem Sorek. Genomic fossils as a snapshot of the human transcriptome. *Proceedings of the National Academy of Sciences of the United States of America*, 103:1364–9, Jan 2006.

[130] Adam Pavlicek, Andrew Gentles, Jan Paces, Vaclav Paces, and Jerzy Jurka. Retroposition of processed pseudogenes: the impact of rna stability and translational control. *Trends in genetics : TIG*, 22:69–73, Feb 2006.

[131] John E Karro, Yangpan Yan, Deyou Zheng, Zhaolei Zhang, Nicholas Carriero, Philip Cayting, Paul Harrrison, and Mark Gerstein. Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation. *Nucleic Acids Res*, 35(Database issue):D55–60, January 2007.

[132] J. Jurka, T. F. Smith, and D. Labuda. Small cytoplasmic ro rna pseudogene and an alu repeat in the human alpha-1 globin gene. *Nucleic acids research*, 16:766, Jan 1988.

[133] Marijke J van Baren and Michael R Brent. Iterative gene prediction and pseudogene removal improves genome annotation. *Genome Res*, 16(5):678–85, May 2006.

[134] David Torrents, Mikita Suyama, Evgeny Zdobnov, and Peer Bork. A genome-wide survey of human pseudogenes. *Genome Res*, 13(12):2559–67, December 2003.

[135] Mikita Suyama, Eoghan Harrington, Peer Bork, and David Torrents. Identification and analysis of genes and pseudogenes within duplicated regions in the human and mouse genomes. *PLoS Comput Biol*, 2(6):e76, June 2006.

[136] Alison Yao, Rosane Charlab, and Peter Li. Systematic identification of pseudogenes through whole genome expression evidence profiling. *Nucleic Acids Research*, 34:4477, xx 2006.

[137] J. L. Ashurst, C. K. Chen, J. G. R. Gilbert, K. Jekosch, S. Keenan, P. Meidl, S. M. Searle, J. Stalker, R. Storey, S. Trevanion, L. Wilming, and T. Hubbard. The vertebrate genome annotation (vega) database. *Nucleic acids research*, 33:459–65, Jan 2005.

[138] P Carninci, T Kasukawa, S Katayama, J Gough, M C Frith, N Maeda, R Oyama, T Ravasi, B Lenhard, C Wells, R Kodzius, K Shimokawa, V B Bajic, S E Brenner, S Batalov, A R R Forrest, M Zavolan, M J Davis, L G Wilming, V Aidinis, J E Allen, A Ambesi-Impiombato, R Apweiler, R N Aturaliya, T L Bailey, M Bansal, L Baxter, K W Beisel, T Bersano, H Bono, A M Chalk, K P Chiu, V Choudhary, A Christoffels, D R Clutterbuck, M L Crowe, E Dalla, B P Dalrymple, B de Bono, G Della Gatta, D di Bernardo, T Down, P Engstrom, M Fagiolini, G Faulkner, C F Fletcher, T Fukushima, M Furuno, S Futaki, M Gariboldi, P Georgii-Hemming, T R Gingeras, T Gojobori, R E Green, S Gustincich, M Harbers, Y Hayashi, T K Hensch, N Hirokawa, D Hill, L Huminiecki, M Iacono, K Ikeo, A Iwama, T Ishikawa, M Jakt, A Kanapin, M Katoh, Y Kawasawa, J Kelso, H Kitamura, H Kitano, G Kollias, S P T Krishnan, A Kruger, S K Kummerfeld, I V Kurochkin, L F Lareau, D Lazarevic, L Lipovich, J Liu, S Liuni, S McWilliam, M Madan Babu, M Madera, L Marchionni, H Matsuda, S Matsuzawa, H Miki, F Mignone, S Miyake, K Morris, S Mottagui-Tabar, N Mulder, N Nakano, H Nakauchi, P Ng, R Nilsson, S Nishiguchi, S Nishikawa, F Nori, O Ohara, Y Okazaki, V Orlando, K C Pang, W J Pavan, G Pavesi, G Pesole, N Petrovsky, S Piazza, J Reed, J F Reid, B Z Ring, M Ringwald, B Rost, Y Ruan, S L Salzberg, A Sandelin, C Schneider, C Schönbach, K Sekiguchi, C A M Semple, S Seno, L Sessa, Y Sheng, Y Shibata, H Shimada, K Shimada, D Silva, B Sinclair, S Sperling, E Stupka, K Sugiura, R Sultana, Y Takenaka, K Taki, K Tammoja, S L Tan, S Tang, M S Taylor, J Tegner, S A Teichmann, H R Ueda, E van Nimwegen, R Verardo, C L Wei, K Yagi, H Yamanishi, E Zabarovsky, S Zhu, A Zimmer, W Hide, C Bult, S M Grimmond, R D Teasdale, E T Liu, V Brusic, J Quackenbush, C Wahlestedt, J S Mattick, D A Hume, C Kai, D Sasaki, Y Tomaru, S Fukuda, M Kanamori-Katayama, M Suzuki, J Aoki, T Arakawa, J Iida, K Imamura, M Itoh, T Kato, H Kawaji, N Kawagashira, T Kawashima, M Kojima, S Kondo, H Konno, K Nakano, N Ninomiya, T Nishio, M Okada, C Plessy, K Shibata, T Shiraki, S Suzuki, M Tagami, K Waki, A Watahiki, Y Okamura-Oho, H Suzuki, J Kawai, and Y Hayashizaki. The transcriptional landscape of the mammalian genome. *Science*, 309(5740):1559–63, September 2005.

[139] John K Pace and Cedric Feschotte. The evolutionary history of human dna transposons: Evidence for intense activity in the primate lineage. *Genome Research*, 17:422–32, Apr 2007.

[140] Z. Zhang, S. Schwartz, L. Wagner, and W. Miller. A greedy algorithm for aligning dna sequences. *Journal of computational biology : a journal of computational molecular cell biology*, 7:203–14.

[141] S F Altschul, R Bundschuh, R Olsen, and T Hwa. The estimation of statistical parameters for local alignment score distributions. *Nucleic Acids Res*, 29(2):351–61, January 2001.

[142] M. S. Waterman and M. Vingron. Rapid and accurate estimates of statistical significance for sequence data base searches. *Proceedings of the National Academy of Sciences*, 91:4625–4628, May 1994. 10.1073/pnas.91.11.4625.

[143] M. Ashburner, C. Ball, J. Blake, D. Botstein, H. Butler, J. Cherry, A. Davis, K. Dolinski, S. Dwight, J. Eppig, M. Harris, D. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. Matese, J. Richardson, M. Ringwald, G. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature genetics*, 25:25–29, May 2000. 10.1038/75556.

[144] Pfam: Family: zf-c2h2 (pf00096).

[145] Raul Urrutia. Krab-containing zinc-finger repressor proteins. *Genome biology*, 4:231, xx 2003. 10.1186/gb-2003-4-10-231.

[146] Pfam: Family: Krab (pf01352).

[147] J. Friedman, W. Fredericks, D. Jensen, D. Speicher, X. Huang, E. Neilson, and F. Rd. Kap-1, a novel corepressor for the highly conserved krab repression domain. *Genes & development*, 10:2067–2078, Aug 1996.

[148] S. Kim, Y. Chen, E. O'Leary, R. Witzgall, M. Vidal, and J. Bonventre. A novel member of the ring finger family, krip-1, associates with the krab-a transcriptional repressor domain of zinc finger proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 93:15299–15304, Dec 1996.

[149] H. Thiesen, E. Bellefroid, O. Revelant, and J. Martial. Conserved krab protein domain identified upstream from the zinc finger region of kox 8. *Nucleic acids research*, 19:3996, Jul 1991.

[150] W. Gish. Wu blast 2.0, 1996-2004.

[151] S. Altschul. Amino acid substitution matrices from an information theoretic perspective. *Journal of molecular biology*, 219:555–565, Jun 1991.

[152] M Girvan and M E J Newman. Community structure in social and biological networks. *Proc Natl Acad Sci U S A*, 99(12):7821–6, June 2002.

[153] J. Thompson, D. Higgins, and T. Gibson. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research.*, 22:4673–4680, Nov 1994.

[154] Davide Corà, Ferdinando Di Cunto, Paolo Provero, Lorenzo Silengo, and Michele Caselle. Computational identification of transcription factor binding sites by functional analysis of sets of genes sharing overrepresented upstream motifs. *BMC bioinformatics*, 5:57, May 2004. 10.1186/1471-2105-5-57.

[155] National center for biotechnology information.

[156] Alex Bateman, Lachlan Coin, Richard Durbin, Robert Finn, Volker Hollich, Sam Griffiths-Jones, Ajay Khanna, Mhairi Marshall, Simon Moxon, Erik Sonnhammer, David Studholme, Corin Yeats, and Sean Eddy. The pfam protein families database. *Nucleic acids research.*, 32:138–141, Jan 2004. 10.1093/nar/gkh121.

[157] Aron Marchler-Bauer, John Anderson, Myra Derbyshire, Carol Deweese-Scott, Noreen Gonzales, Marc Gwadz, Luning Hao, Siqian He, David Hurwitz, John Jackson, Zhaoxi Ke, Dmitri Krylov, Christopher Lanczycki, Cynthia Liebert, Chunlei Liu, Fu Lu, Shennan Lu, Gabriele Marchler, Mikhail Mullokandov, James Song, Narmada Thanki, Roxanne Yamashita, Jodie Yin, Dachuan Zhang, and Stephen Bryant. Cdd: a conserved domain database for interactive domain family analysis. *Nucleic acids research*, 35:D237–40, Jan 2007. 10.1093/nar/gkl951.

[158] B Alberts. The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell*, 1998.

[159] L Giot, JS Bader, C Brouwer, A Chaudhuri, B Kuang, Y Li, YL Hao, CE Ooi, B Godwin, and E Vitols. A protein interaction map of drosophila melanogaster. *Science*, 2003.

[160] S Li, CM Armstrong, N Bertin, H Ge, S Milstein, M Boxem, PO Vidalain, JD Han, A Chesneau, and T Hao. A map of the interactome network of the metazoan c. *elegans. Science*, 2004.

[161] S Peri, JD Navarro, TZ Kristiansen, R Amanchy, V Surendranath, B Muthusamy, TK Gandhi, KN Chandrika, N Deshpande, and S Suresh. Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res*, 2004.

[162] A Zanzoni, Palazzi L Montecchi, M Quondam, G Ausiello, Citterich M Helmer, and G Cesareni. Mint: a molecular interaction database. *FEBS Lett*, 2002.

[163] N Lin, B Wu, R Jansen, M Gerstein, and H Zhao. Information assessment on predicting protein-protein interactions. *BMC Bioinformatics*, 2004.

[164] Mering C von, R Krause, B Snel, M Cornell, SG Oliver, S Fields, and P Bork. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 2002.

[165] A Hahn, J Rahnenfuhrer, P Talwar, and T Lengauer. Confirmation of human protein interaction data by human expression data. *BMC Bioinformatics*, 2005.

[166] R Jansen, H Yu, D Greenbaum, Y Kluger, NJ Krogan, S Chung, A Emili, M Snyder, JF Greenblatt, and M Gerstein. A bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 2003.

[167] R Jansen, D Greenbaum, and M Gerstein. Relating whole-genome expression data with protein-protein interactions. *Genome Res*, 2002.

[168] DR Rhodes, SA Tomlins, S Varambally, V Mahavisno, T Barrette, Sundaram S Kalyana, D Ghosh, A Pandey, and AM Chinnaiyan. Probabilistic model of the human protein-protein interaction network. *Nat Biotechnol*, 2005.

[169] E Sprinzak, Y Altuvia, and H Margalit. Characterization and prediction of protein- protein interactions within and between complexes. *Proc Natl Acad Sci U S A*, 2006.

[170] M Pellegrino, P Provero, L Silengo, and Cunto F Di. Cloe: identification of putative functional relationships among genes by comparison of expression profiles between two species. *BMC Bioinformatics*, 2004.

[171] JD Han, N Bertin, T Hao, DS Goldberg, GF Berriz, LV Zhang, D Dupuy, AJ Walhout, ME Cusick, and FP Roth. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, 2004.

[172] ML Whitfield, G Sherlock, AJ Saldanha, JI Murray, CA Ball, KE Alexander, JC Matese, CM Perou, MM Hurt, and PO Brown. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol Biol Cell*, 2002.

[173] Lichtenberg U de, LJ Jensen, S Brunak, and P Bork. Dynamic complex formation during the yeast cell cycle. *Science*, 2005.

[174] GM Bokoch. Biology of the p21-activated kinases. *Annu Rev Biochem*, 2003.

[175] IG Mills, AT Jones, and MJ Clague. Involvement of the endosomal autoantigen eea1 in homotypic fusion of early endosomes. *Curr Biol*, 1998.