Genome Annotation

Josep F Abril, University of Barcelona (UB), Barcelona, Spain and Institute of Biomedicine of the University of Barcelona (IBUB), Barcelona, Spain
Sergi Castellano, Great Ormond Street Institute of Child Health (ICH), University College London (UCL), London, United Kingdom and UCL Genomics, University College London (UCL), London, United Kingdom

> "The genome sequence of an organism is an information resource unlike any that biologists have previously had access to. But the value of the genome is only as good as its annotation. It is the annotation that bridges the gap from the sequence to the biology of the organism." Lincoln Stein, 2001. Nat. Rev. Genet. 2(7), 493–503.

Introduction

The precise ordering of nucleotides and amino acids confer specific functional properties to biological sequences. Functional and structural features can then be understood as regions contained within those sequences, defined as segments delimited by an interval of positions, initial and terminal, on the coordinates axis provided by the sequence itself. In a wide sense, sequence annotation is defined by the procedures leading to determine the location and properties of these sequence features. Genomes are the organisms' blueprints, encoding for all the functional elements that lead to a fully developed life form. Therefore, genomes are complex and contain a mixture of features with different roles, which are interspersed with non-functional sequences under no evolutionary constraint. Genome annotation is the process of identifying any functional element along the DNA sequence of a genome, yet at initial stages often focuses on genes. The annotation gives meaning to the genome by providing the location and function of genes (protein coding or otherwise) and regulatory regions, which underlie the biology of living organisms. Completing the catalog of functional regions of the genome can facilitate further downstream analyses, as for instance the assessment of the effect of mutations caused by single nucleotide variants in individuals or populations. As it has been remarked since the start of the genomics era, raw genomes are thus of limited use to scientists and their annotation has become central to genome research.

Genome annotation itself has evolved over the years and it is possible to distinguish three, possibly four, major stages based on the techniques and data used. Each succeeding, yet overlapping, stage has refined previous annotations of the genome, characterized some of its previously unannotated parts, and provided novel insights into genome biology. The development in the mid 1990s of computational techniques to predict protein-coding genes, which span thousands of nucleotides in eukaryotes, marks the start of the efforts to provide genome-wide annotations. This first stage centered on the annotation of individual reference genomes (one per species), using a combination of the statistical properties of the sequences of the protein-coding genes themselves (to predict genes *ab initio*) and the similarity of mRNA and protein sequences to their genome of origin (to align known transcripts and proteins). Both approaches though rely on the existence of a set of already characterized genes, *a priori* knowledge based on experimental evidences, which will be used either as training set to build the models describing the genes species-specific signals and sequence content biases for subsequent computational predictions, or to be compared against other anonymous sequences to map their location based on similarity, thus annotating the genes in them (Fig. 1).

When the reference genomes of more species became available in the early and mid 2000s, they ushered in a comparative and second stage of genome annotation, where mRNA, protein and genome sequences from one species could be homologously aligned to annotate the genome of another. A third and regulatory stage of genome annotation was initiated in the mid and late 2000s, after genome-wide methods to profile DNA binding, conformation and alteration (*e.g.*, in their copy number) were introduced. Such genome-wide and multi-omics regulatory annotations have become standard in the last few years and short (in the tens or hundreds of nucleotides) regulatory elements are commonly annotated. Finally, genome annotation definitely moved away from one or few genomes per species in the late 2000s and early 2010s. This fourth and population stage of genome annotation, which uses hundreds if not thousands of genomes from the same species to annotate individual nucleotides, is where we find ourselves in the late 2010s. We discuss in some detail these different stages of genome annotation and how their increased resolution and inclusiveness of multi-omics approaches leads to more precise genome biology.

The *ab Initio* and Similarity Years

Ab initio gene prediction (also known as *de novo*) refers to the identification of protein-coding genes using the signals that define and characterize their gene structures along the genome. This is done without the aid of their encoded mRNAs and proteins. When the sequences of these mRNAs and proteins are sequenced they can be aligned to the genome to locate the genes that encode them. Thus, allowing the prediction of protein-coding genes by similarity. We discuss both approaches and the increased accuracy of similarity approaches over the *ab initio* ones.

doi:10.1016/B978-0-12-809633-8.20226-4

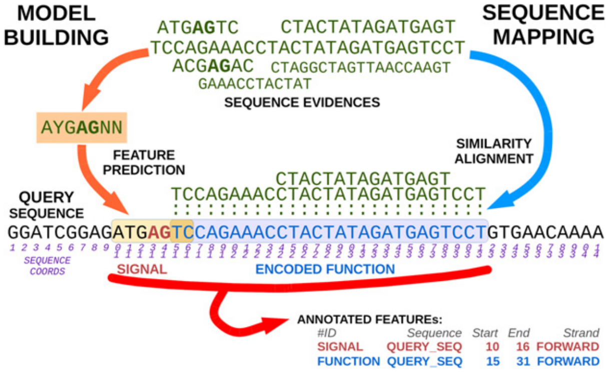

Fig. 1 In order to annotate an anonymous sequence (query) we have two main approaches. The first approach, depicted on the left, it is to build models that, based on known sequences, summarize some property or bias characteristic of the functional elements being annotated, and then apply an algorithm to score these elements along the sequence, taking the optimal prediction as the one to annotate. The second approach uses the alignment against the query of the set of known sequences, retrieved either from the same (similarity search) or other species (homology search). Again, the optimal alignment is taken as the one to annotate the functional element. The coordinates – relative to the original sequence – define the location for each of the functional elements found in the anonymous genome, also known as the annotation set.

Ab Initio Gene Prediction

The structure of prokaryotic and eukaryotic genes is such that start (AUG) and stop (UAA/UAG/UGA) codons signal the initiation and termination steps of protein translation. These codons, together with some nucleotide context around them, define the coding DNA sequence (CDS) of the exon or exons in a gene. In eukaryote genomes, in particular, codons in exons along the CDS are not necessarily adjacent but often interrupted by introns of hundreds or thousands of nucleotides in length (Sharp, 2005). These breaks are defined, in most cases, by the canonical donor [AG|**GT**RAGT, where R=A or G] and acceptor [YYTTYYYYYYNC**AG**|G, where Y=C or T and N=A, C, T or G] splice sites in mammalian genomes (Burset *et al.*, 2000; Abril *et al.*, 2005). Albeit GT-AG and the surrounding context defines a highly conserved motif, splicing is a complex biological process that tolerates small nuclear RNA variants in the spliceosome (Kyriakopoulou *et al.*, 2006), the RNA-protein complex that mediates the splicing process, and even non-canonical splice sites, such as those defined by the donor and acceptor pair AT-AC (Burset *et al.*, 2000; Patel and Steitz, 2003). Together with transcription initiation and termination sites, these signals have allowed the computational definition of prokaryotic and eukaryotic genes.

In addition, the genetic code is degenerate and the same amino acid can be encoded by different codons. These codons are synonymous and the frequency of their occurrence is known as codon usage or bias. Codon usage varies between species and between genes in a genome due to mutational biases and differences in the strength of natural selection on translational optimization (Plotkin and Kudla, 2011). Predating the sequencing of genomes there were tools that used codon bias to annotate expressed sequence tags (ESTs), fragments of mRNAs, which were even capable to correct frameshifts from sequencing errors like ESTSCAN (Iseli *et al.*, 1999). Thus, codon usage biases within the signals that define prokaryotic and eukaryotic genes can be used to computationally identify the CDS of genes along a particular species genome. Some of the first *ab initio* programs to predict genes this way did so in the short genomes of prokaryotes. For example, the seminal GENEMARK (Lukashin and Borodovsky, 1998) and GLIMMER (Delcher *et al.*, 1999) programs made use of Markov chains – a stochastic model describing a sequence of nucleotide-emitting states in which the probability of the next state only depends on the previous one – to capture the usage and dependence between successive nucleotides and codon positions, while defining the start and end of genes with models of the nucleotides around these signals (their sequence context). Interestingly, Markov chains of order-five, in which the probability of the next nucleotide depends on the five previous ones, work best in gene prediction as they also capture dependencies between consecutive amino acids in proteins (Durbin *et al.*, 1998). These models are used to score potential genes along the genome. The final prediction is then obtained using algorithms that find the optimal gene structure, that is, the combination of predicted exons that maximizes the score of the predicted genes. GENEMARK, for example, uses the Viterbi algorithm (Durbin *et al.*, 1998). This is a dynamic programming algorithm (Eddy, 2004a) – an optimization technique based on breaking down a problem into smaller ones and using them recursively to find the optimal solution – which, in a Hidden Markov Model (HMM) (Eddy, 2004b) – a statistical model following a Markov Chain process with hidden states that either emit nucleotides for coding or noncoding sequences – finds the best scoring gene structures. More recent developments include Prodigal (Hyatt *et al.*, 2010), which extends dynamic programming with a special set of rules to cope with overlapping genes while looking for the optimal gene structure prediction along the genome.

Hyatt *et al.* also established a reference annotation set to evaluate the accuracy of available prokaryotic gene-finders; a difficult task due to the lack of an experimentally verified translation start in many genes that changes the average length of genes, as well as the huge variability in GC content across prokaryota. The accuracy of gene prediction in prokaryotic genomes is generally high, with more than 95% of true protein coding genes being detected in most species (Hyatt *et al.*, 2010). Though the accuracy predicting the start of the gene is lower, usually around 80%. As a result, the 5′ end (upstream) of prokaryotic genes is less likely to be correct. In addition, prokaryotic gene prediction may still slightly overpredict the number of genes in a genome, with the resulting false positive genes being often short. It is however possible that some of these genes are indeed real. In any case, prokaryotic gene prediction has become remarkably accurate in the last two decades.

The identification of eukaryotic genes is, compared to their prokaryotic counterparts, more demanding due to their length (often in the hundreds of thousands of bases) and their split structure. Still, similar approaches to those in prokaryotes can be used. Early eukaryotic gene predictors included GENEID (Guigo *et al.*, 1992) and GENSCAN (Burge and Karlin, 1997). Both of these programs use Markov Chains of order five to assess codon usage and score coding exons as a combination of their coding bias and their start, end and splice sites signals; either in a rule-based (the GENEID Gene Model; (Guigo *et al.*, 1992)) or a generalized HMM (GHMM) framework for GENSCAN (Burge and Karlin, 1997). To cope with the combinatorial explosion of putative exons and predict the optimal gene structure along the genome dynamic programming was used (Burge and Karlin, 1997; Guigo, 1998).

The accuracy of such predictions is lower than in prokaryotic genomes (**Table 1**). An influential work on the assessment of gene prediction accuracy in eukaryotes (Burset and Guigo, 1996) established the ground metrics at the nucleotide, exon and gene levels, and provided one of the first reference gene test sets on this field. To begin with, there is in gene prediction a compromise between sensitivity (*Sn* in **Table 1**), the proportion of true coding nucleotides correctly predicted as such, and specificity (*Sp* in **Table 1**), the proportion of predicted coding nucleotides actually coding, so increasing one often decreases the other. At the nucleotide level this is not much of an issue, as fragments of most genes are predicted; yet missing the right start and end of an exon reduces the sensitivity and specificity of gene prediction at the exon level (**Table 1**). The accuracy for the overall gene is even lower due the difficulty in determining the right combination of exons along the genome. On the other hand, short genes encoding small proteins (Su *et al.*, 2013) are usually missed (missing exons, ME, in **Table 1**) as most genome annotation protocols apply a minimum length threshold, *i.e.*, of 100 amino acids, to reduce the number of falsely predicted genes (wrong exons, WE, in **Table 1**). These measures show that classic *ab initio* gene prediction programs identified no more than 80% of the true coding nucleotides while predicting as coding a large fraction of noncoding ones (**Table 1**). Also, a substantial fraction of coding exons had their boundaries mispredicted. Furthermore, around 40% of the predicted exons are completely wrong, a rather unreasonable figure. Still, *ab initio* gene prediction provided at the time a first set of annotations that could be later refined using comparative and/or experimental approaches (see below). To cope with annotation changes and multiple transcripts in a gene further accuracy measures have been proposed such as the Annotation Edit Distance, which quantifies changes to a gene annotation, and the Splice Complexity, which quantifies the complexity of alternative splicing in a gene (Eilbeck *et al.*, 2009). The terms recall and precision were later adopted to measure gene prediction accuracy. Recall is equivalent to the sensitivity measure described above, whereas precision is equivalent to specificity. Importantly, specificity or precision in the gene prediction field avoid the estimation of true negatives from incomplete genomes. Nowadays gene predictions are also compared with respect to the area under the curve in a receiver operating characteristics plot (known as AUC-ROC plot) (Powers, 2011).

The first large-scale assessment on gene-finding accuracy was performed over 2.4 megabases of the *Drosophila melanogaster* genome, which contained the *alcohol dehydrogenase* gene (*Adh*), in a collaborative experiment known as the Genome Annotation Assessment Project (GASP) (Reese *et al.*, 2000). This region was first annotated by human curators, with the raw sequence and a subset of the reference fly annotations later provided to the participants performing computational gene predictions. Similar assessments have subsequently been done in the worm (nGASP) (Coghlan *et al.*, 2008) and human genomes (EGASP and RGASP, see **Figs. 2** and **3**) (ENCODE Consortium, 2012; Guigo *et al.*, 2006).

Human curators have been invaluable in producing reference annotations for the above assessment experiments and in integrating the experimental evidence that validates (or refutes) individual gene predictions. For example, *D. melanogaster* was the first eukaryotic genome assembled from shotgun sequences (where DNA is broken up randomly) (Adams *et al.*, 2000) and Celera, the company sequencing it, organized an annotation jamboree with bioinformatics experts, protein specialists, and fruit fly biologists to functionally annotate it (Pennisi, 2000). Similar community annotation projects have since then become a standard for other genomes, among those it is worth citing the VErtebrate Genome Annotation database (VEGA, http://vega.sanger.ac.uk) maintained by the human and vertebrate analysis and annotation (HAVANA) group at the Wellcome Trust Sanger Institute (Harrow *et al.*, 2012), which also has defined guidelines to integrate gene annotations (Madupu *et al.*, 2010; Loveland *et al.*, 2012). To assist curators in the visualization and editing of gene annotations different tools have been created. First, to produce static maps of annotations along a genome sequence with GFF2PS (Abril and Guigo, 2000), GFF2APLOT (Abril *et al.*, 2003), and more recently CIRCOS (Krzywinski *et al.*, 2009). Later, tools to manually review gene predictions and integrate experimental evidence, such as APOLLO (Lewis *et al.*, 2002), the Integrative Genomics

Table 1 Accuracy of gene prediction programs on human chromosome 22

Program	*Nucleotide*			*Exon*				
	Sn	Sp	CC	Sn	Sp	$\frac{Sn+Sp}{2}$	*ME*	*WE*
"*ab initio*" gene finding								
GENEID	0.73	0.67	0.70	0.65	0.55	0.60	0.21	0.33
GENSCAN	0.79	0.53	0.64	0.68	0.41	0.55	0.15	0.48
Comparative genomics approach								
SGP2	0.75	0.73	0.73	0.66	0.58	0.62	0.19	0.28
TWINSCAN	0.72	0.67	0.69	0.69	0.59	0.64	0.18	0.29

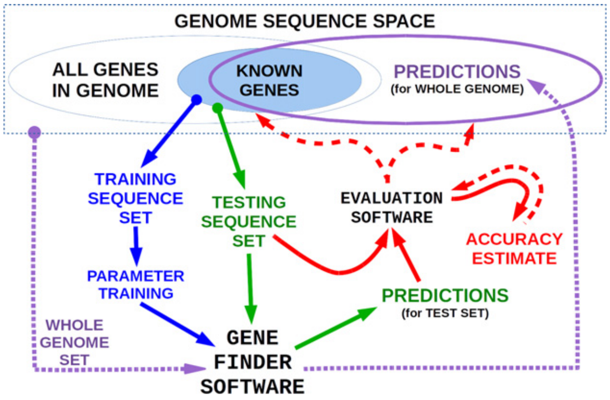

Fig. 2 The assessment of the accuracy of gene predictions (or other functional features) is usually done in three steps. First, a set of well characterized genes is split into the training and the testing sequence sets. Second, the training set is used to estimate the species-specific parameters, such as the frequencies of the different codons in coding exons. Finally, the test set is used to predict genes with the computational approach under assessment. The sequence coordinates of the predicted genes are then compared against those of the known genes. Accuracy metrics, calculated at the nucleotide, exon and gene levels, are used to assess accuracy of the method and the reliability of the predictions. This assessment can be used later on either to estimate overall genome predictions reliability or to iterate through the train/test sets again in order to improve the software parameters (dashed lines).

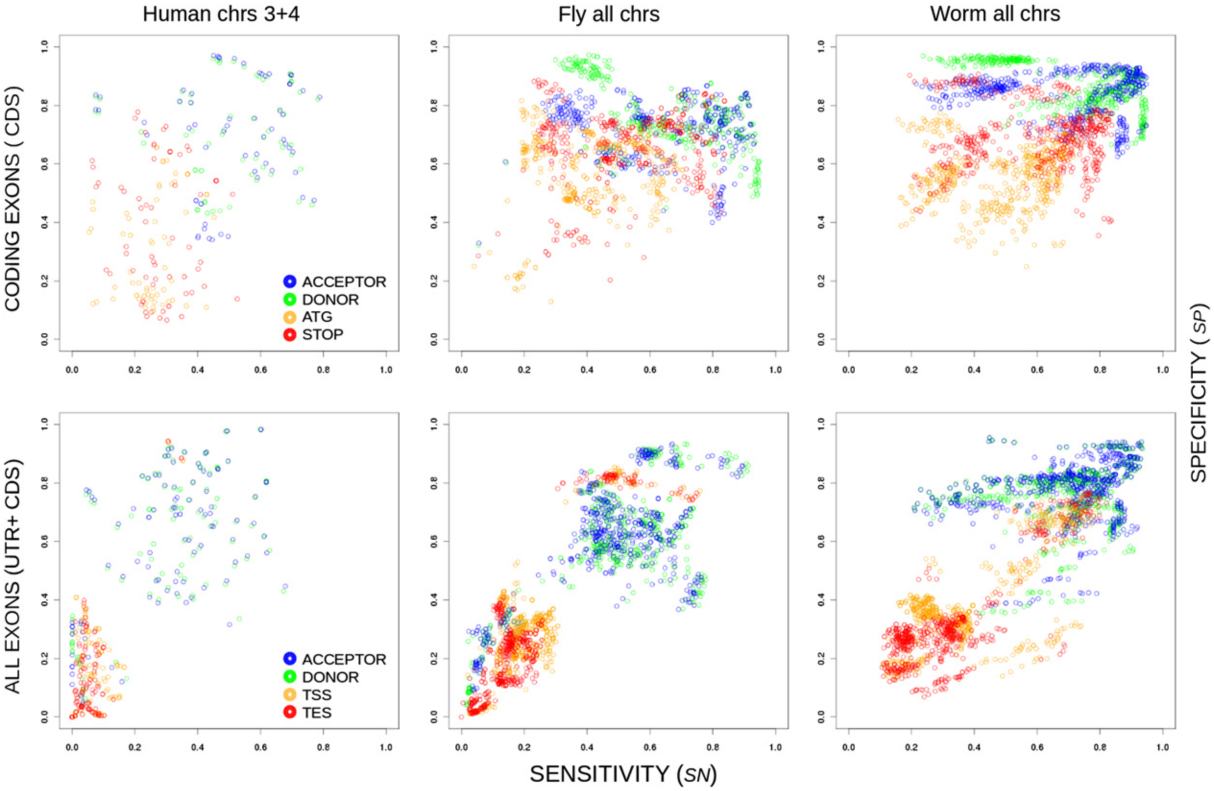

Fig. 3 Sensitivity versus specificity scatterplots for the prediction of a number of defining signals in coding exons (CDS), on top row, and full-length exons, on bottom row, for the submitted predictions to RGASP. The accuracy metrics were computed over the genome sequences from three species: the human chromosomes three and four, the *Drosophila melanogaster* (fly) genome and the *Caenorhabditis elegans* (worm) genome. Each dot corresponds to a set of predictions made by one of each RGASP participant over one of the species sequences. Note the lower accuracy of predictions in the larger human genome as well as the difficulty to predict the untranslated part of exons due to missing the correct transcription start (TSS) and end (TES) sites in the three species.

Viewer (IGV) (Thorvaldsdottir *et al.*, 2013) and the GALAXY server (Giardine *et al.*, 2005) grew in importance. Over time, genome browsers providing access to databases storing the analyses from genome annotation pipelines have become the standard. These include GBROWSE (Stein *et al.*, 2002), the UCSC Genome Browser (Kent *et al.*, 2002) and ENSEMBL (Hubbard *et al.*, 2002; Aken *et al.*, 2017), which are widely used today (**Fig. 4**).

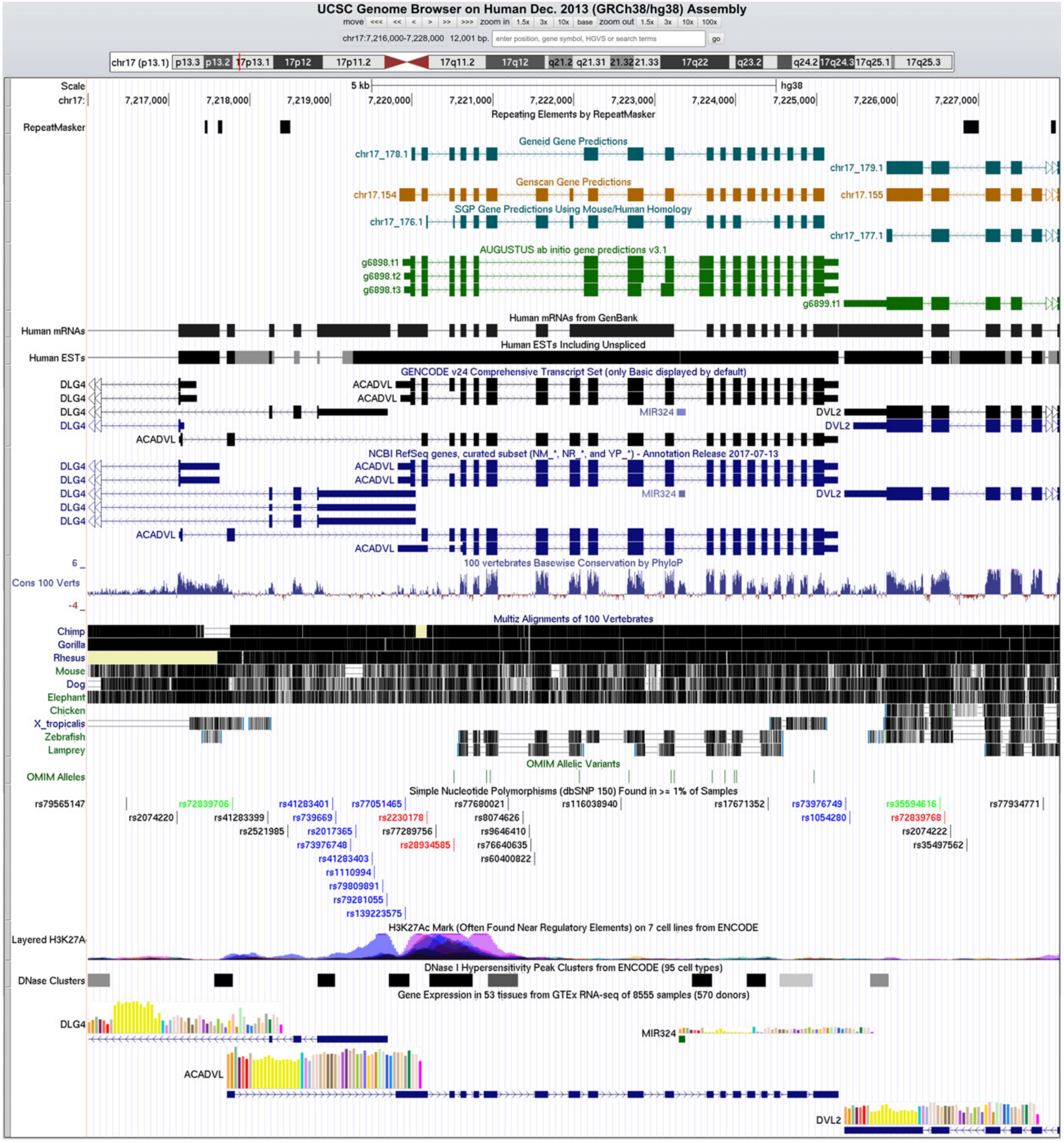
UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
chr17:7,216,000-7,228,000 12,001 bp.
enter position, gene symbol, HGVS or search terms
go
chr17 (p13.1) p13.3 p13.2 17p13.1 17p12 17p11.2 17q11.2 17q12 q21.2 q21.31 21.32 21.33 17q22 q23.2 q24.2 17q24.3 17q25.1 17q25.3
Scale
5 kb
hg38
chr17:
7,217,000
7,218,000
7,219,000
7,220,000
7,221,000
7,222,000
7,223,000
7,224,000
7,225,000
7,226,000
7,227,000
Repeating Elements by RepeatMasker
RepeatMasker
Geneid Gene Predictions
chr17_178.1
chr17_179.1
Genscan Gene Predictions
chr17.154
chr17.155
SGP Gene Predictions Using Mouse/Human Homology
chr17_176.1
chr17_177.1
AUGUSTUS ab initio gene predictions v3.1
g6898.t1
g6898.t2
g6898.t3
g6899.t1
Human mRNAs from GenBank
Human mRNAs
Human ESTs Including Unspliced
Human ESTs
GENCODE v24 Comprehensive Transcript Set (only Basic displayed by default)
DLG4
ACADVL
MIR324
DVL2
NCBI RefSeq genes, curated subset (NM_*, NR_*, and YP_*) - Annotation Release 2017-07-13
6
Cons 100 Verts
-4
100 vertebrates Basewise Conservation by PhyloP
Multiz Alignments of 100 Vertebrates
Chimp
Gorilla
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Lamprey
OMIM Allelic Variants
OMIM Alleles
Simple Nucleotide Polymorphisms (dbSNP 150) Found in >= 1% of Samples
rs79565147
rs2074220
rs72839706
rs41283399
rs2521985
rs41283401
rs739669
rs2017365
rs73976748
rs41283403
rs1110994
rs79809891
rs79281055
rs139223575
rs77051465
rs2230178
rs77289756
rs28934585
rs77680021
rs8074626
rs9646410
rs76640635
rs60400822
rs116038940
rs17671352
rs73976749
rs1054280
rs35594616
rs72839768
rs2074222
rs35497562
rs77934771
H3K27Ac Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27A
DNase I Hypersensitivity Peak Clusters from ENCODE (95 cell types)
DNase Clusters
Gene Expression in 53 tissues from GTEx RNA-seq of 8555 samples (570 donors)
DLG4
MIR324
ACADVL
DVL2

Fig. 4 UCSC Genome Browser view of the human chromosome 17, focused on the sequence annotation around the *ACADVL* gene. Topmost annotation tracks show several gene/transcript structures obtained by different methods; mid tracks summarize the conservation score and the human sequence aligned to different vertebrate genomes; bottom tracks include sequence variants (OMIM and common variants), histone marks from the ENCODE project and transcript gene expression. The gene annotation tracks on top displays predictions from GENEID, GENSCAN, SGP2, and AUGUSTUS, which can be compared with the current curated annotations from two reference sets, GENCODE and NCBI-RefSeq. From the aforementioned gene-finding programs, only AUGUSTUS predicts the three alternative spliced transcripts as it integrates experimental evidence, although it fails to predict some conserved exons detected by the other programs. Note that downstream *DVL2* gene was also detected by the gene-finders, but the upstream *DLG4* gene was missed.

Using mRNAs and Proteins

The sequencing of fragments (ESTs) or full-length mRNAs and proteins from the species whose genome is being annotated has been the main approach to increase the accuracy of *ab initio* gene annotation. Instead of relying on the statistical properties of the CDS and the signals that define them, the direct alignment of transcribed or translated sequences from a genome provides experimental evidence to a gene structure. The alignment of such sequences to a naked genome has been computationally addressed yet again as a dynamic programming algorithm that takes into account the splice sites and exon/intron structure of eukaryotic genes. In particular, ESTs or full-length transcripts can be properly aligned to the genome using variations of the Smith-Waterman (Smith and Waterman, 1981; Gotoh, 1982) and Needleman-Wunsch (Needleman and Wunsch, 1970) algorithms, the classic local and global sequence alignment algorithms, respectively. This is the approach used by the pioneering EST_GENOME program (Mott, 1997). For the sake of speed, other approaches use the heuristic BLAST algorithm (Altschul *et al.*, 1990) to produce the starting alignments of coding exons in the genome, which are later recursively chained together to obtain the best gene structure. SIM4 (Florea *et al.*, 1998) and SPIDEY (Wheelan *et al.*, 2001) were examples of the latter. A more recent tool to align transcripts to a genomic sequence is EXONERATE (Slater and Birney, 2005), which provides (slower) exhaustive or (faster) heuristic approaches to align transcripts using different dynamic programming algorithms. On the other hand, spliced-alignment algorithms with proteins require the conceptual translation of the genomic sequence and finding the optimal alignment of a multi-exon structure to a related protein. PROCRUSTES (Gelfand *et al.*, 1996) and GENEWISE (Birney and Durbin, 1997, 2000) are well-known examples, with the HMM-based GENEWISE at the core of the ENSEMBL gene annotation system (Curwen *et al.*, 2004; Aken *et al.*, 2016). The alignment of proteins to predict genes in large scale genomes can become extremely time- and space-demanding. This investment, however, is usually at the benefit of prediction accuracy. In this regard, the faster EXONERATE can also align proteins to DNA sequences and is also heavily used in the ENSEMBL pipeline.

In the last few years, the widespread use of next-generation sequencing technologies for genomes and transcriptomes (RNA sequencing, RNA-seq) and high-throughput mass spectrometry approaches for proteins have put the approaches above at the forefront of genome annotation. For example, the alignment of transcripts from RNA-seq experiments has allowed the comprehensive annotation of alternative splice forms of protein-coding genes. Humans, for example, produce on average four different protein-coding sequences per gene (ENCODE Consortium, 2012). Pseudogenes, which derived from functional genes through retrotransposition or duplication but have lost the original functions of their parental genes, are sometimes (about 10%) also transcribed (ENCODE Consortium, 2012). The high degree of conservation and similarity to functional genes of recent pseudogenes are significant enough to confound conventional gene prediction approaches (Zheng *et al.*, 2007). Other non-standard types of genes that have benefited from experimental evidence are fused genes, which combine part or all of the exons from transcripts of two collinear genes. In this way, they may contribute to the increase the protein repertoire in eukaryotes (Parra *et al.*, 2006). Similarly, trans-splicing, where the starting exons from one transcript are merged with exons from another transcript, far downstream or even located in another chromosome sequence, can be annotated using sequenced transcripts. Remarkably, up to 70% of the worm *Caenorhabditis elegans* mRNAs begin with the spliced leader sequence (SL, 22bp), which is not associated with the gene (Hastings, 2005).

Annotation of Non-Standard Protein Coding Genes

The identification of eukaryotic selenoprotein genes has challenged most computational gene prediction methods. The reason for this is the alternative use of the UGA codon, typically a termination signal, to code for selenocysteine (Chambers *et al.*, 1986; Zinoni *et al.*, 1986). The recoding of UGA confounds the computational identification of selenoprotein genes, whose exons containing a selenocysteine codon are truncated or skipped in their prediction. Selenocysteine, the 21st amino acid in the genetic code, is incorporated into selenoprotein with the help of an mRNA element, the selenocysteine insertion sequence (SECIS), located in the 3′ untranslated region of selenoprotein genes (48,49). This RNA structure directs the incorporation of selenocysteine through the interaction with several trans-acting factors. Selenocysteine mediates the biological functions of selenium, an essential micronutrient whose deficiency is associated to infertility, preterm birth and serious children's disorders of the bone and heart (Rayman, 2000).

Efforts to identify selenoproteins were first directed to the computational prediction of the SECIS element in Expressed Sequence Tags (ESTs, sequenced mRNA fragments). Known SECIS elements were used to derive deterministic models of their secondary structures (**Fig. 5**) which, in turn, were used to scan and identify two novel selenoprotein genes (Kryukov *et al.*, 1999; Lescure *et al.*, 1999). Efforts to identify additional selenoprotein genes in eukaryotic genomes soon followed. The inclusion of UGA as a selenocysteine codon in the prediction of optimal gene structures along the genome, however, led to the prediction of many erroneous selenoprotein genes (false positives). The low specificity of selenoprotein gene prediction was due to the difficulty of using the typical codon usage characteristic of proteins to extend them beyond an in-frame termination codon (Castellano *et al.*, 2001). A solution to this problem was to use the prediction of SECIS elements along the genome to limit the number of exons with a TGA in-frame that are incorporated into the dynamic programming recursion used to predict optimal genes. Several new selenoproteins in different species were identified this way (Castellano *et al.*, 2001; Kryukov *et al.*, 2003; Castellano *et al.*, 2005). Another example of stop codon recoding is pyrrolysine, which is encoded by UAG but may not require a specific stem loop structure in the 3′ untranslated region for translation (Zhang *et al.*, 2005).

The Comparative and Homology Years

The Zoo Blot is an experimental technique which uses Southern blot analysis to test the ability of a nucleic acid probe from one species to hybridize with DNA of a range of species. Zoo Blots were already used in the 1980s to characterize the protein coding

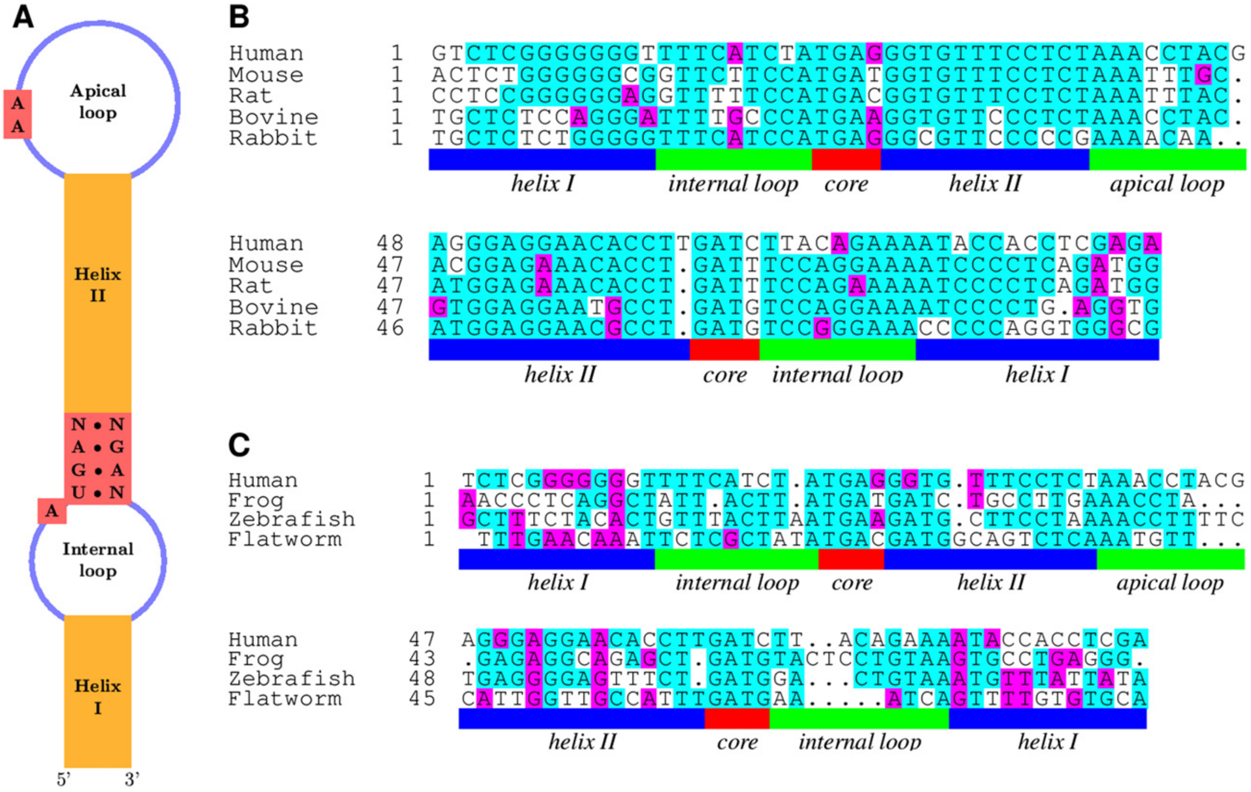
A
Apical loop
A
A
Helix II
N • N
A • G
G • A
U • N
A
Internal loop
Helix I
5'
3'
B
Human 1 GTCTCGGGGGGGTTTTCATCTATGAGGGTGTTTCCTCTAAACCTACG
Mouse 1 ACTCTGGGGGGCGGTTCTTCCATGATGGTGTTTCCTCTAAATTTGC.
Rat 1 CCTCCGGGGGGAGGTTTTTCCATGACGGTGTTTCCTCTAAATTTAC.
Bovine 1 TGCTCTCCAGGGATTTTGCCCATGAAGGTGTTCCCTCTAAACCTAC.
Rabbit 1 TGCTCTCTGGGGGTTTCATCCATGAGGGCGTTCCCCCGAAAACAA..
helix I
internal loop
core
helix II
apical loop
Human 48 AGGGAGGAACACCTTGATCTTACAGAAAATACCACCTCGAGA
Mouse 47 ACGGAGAAACACCT.GATTTCCAGGAAAATCCCCTCAGATGG
Rat 47 ATGGAGAAACACCT.GATTTCCAGAAAAATCCCCTCAGATGG
Bovine 47 GTGGAGGAATGCCT.GATGTCCAGGAAAATCCCCTG.AGGTG
Rabbit 46 ATGGAGGAACGCCT.GATGTCCGGGAAACCCCCAGGTGGGCG
helix II
core
internal loop
helix I
C
Human 1 TCTCGGGGGGTTTTCATCT.ATGAGGGTG.TTTCCTCTAAACCTACG
Frog 1 AACCCTCAGGCTATT.ACTT.ATGATGATC.TGCCTTGAAACCTA...
Zebrafish 1 GCTTTCTACACTGTTTACTTAATGAAGATG.CTTCCTAAAACCTTTTC
Flatworm 1 TTTGAACAAATTCTCGCTATATGACGATGGCAGTCTCAAATGTT...
helix I
internal loop
core
helix II
apical loop
Human 47 AGGGAGGAACACCTTGATCTT..ACAGAAAATACCACCTCGA
Frog 43 .GAGAGGCAGAGCT.GATGTACTCCTGTAAGTGCCTGAGGG.
Zebrafish 48 TGAGGGGAGTTTCT.GATGGA...CTGTAAATGTTTATTATA
Flatworm 45 CATTGGTTGCCATTTGATGAA.....ATCAGTTTTGTGTGCA
helix II
core
internal loop
helix I

Fig. 5 (a) Schematic SECIS element (form 1) divided into structural units. Mostly invariant nucleotides are indicated. (b) Alignment of human and other mammalian glutathione peroxidase 1 SECIS sequences. Note the strong conservation among orthologous sequences. (c) Alignment of human and other non-mammalian glutathione peroxidase 1 SECIS sequences. Compared to the previous alignment, sequence conservation drops significantly.

regions of newly discovered genes (Monaco *et al.*, 1986; Katzav *et al.*, 1989). The rationale behind this approach was that protein coding regions will be conserved across species and would thus retain the capability to hybridize. Testing for hybridization of genomic probes to genomic libraries from many species will reveal the fragments likely to correspond to coding exons. As often happens, bioinformatics techniques based on sequence similarity replicate in the computer those experimental techniques based on the hybridization properties of nucleic acid molecules. In this regard, comparative gene prediction methods that rely on the fact that protein coding regions are generally more conserved during evolution that non-functional ones (**Fig. 6**), can be though as a sort of electronic zoo blots. While phylogenetic footprinting – the phenomenon of the conservation of functional regions across species – has been used to discover, identify and characterize functional regions other than protein coding genes (Margulies *et al.*, 2003).

Essentially, there are two main classes of comparative approaches for gene identification: the comparison of the DNA query sequence with a target protein or mRNA sequence, or a database of such sequences (the computational equivalent of a Northern zoo blot), and the comparison of two or more genomic sequences. In both approaches query and target sequences may be from the same (as discussed above) or different species. If the latter, the selection of the target organism is not a negligible step; the organism should be chosen, when possible, at the phylogenetic distance maximizing the correlation between sequence conservation and coding function. Thus, allowing the inference of homology.

Using mRNAs and Proteins

The backbone of similarity-aided or similarity-based gene structure determination is constituted by those methods that rely on a comparison of the query sequence with protein, or mRNA sequences. Although mostly known as a database search program, `BLASTX` (Altschul *et al.*, 1990; Gish and States, 1993) illustrates the rationale behind this approach. With `BLASTX` a genomic query is translated into a set of amino acid sequences in the six possible frames and compared against a database of known protein sequences. The assumption is that those segments in the genomic query similar to database proteins are likely to correspond to homologous coding exons. A similar assumption is behind the comparison of the genomic query against a database of mRNA sequences (such as ESTs), using `BLASTN (`Altschul *et al.*, 1990`)`, `FASTA` (Pearson and Lipman, 1988; Pearson, 2000), or similar programs. The National Center for Biotechnology Information GENBANK (Benson *et al.*, 2013) and UNIPROT (Bateman *et al.*, 2017), and their European and Asian counterparts, have been the main sources of nucleotide and protein sequences, respectively, for homology-based searches.

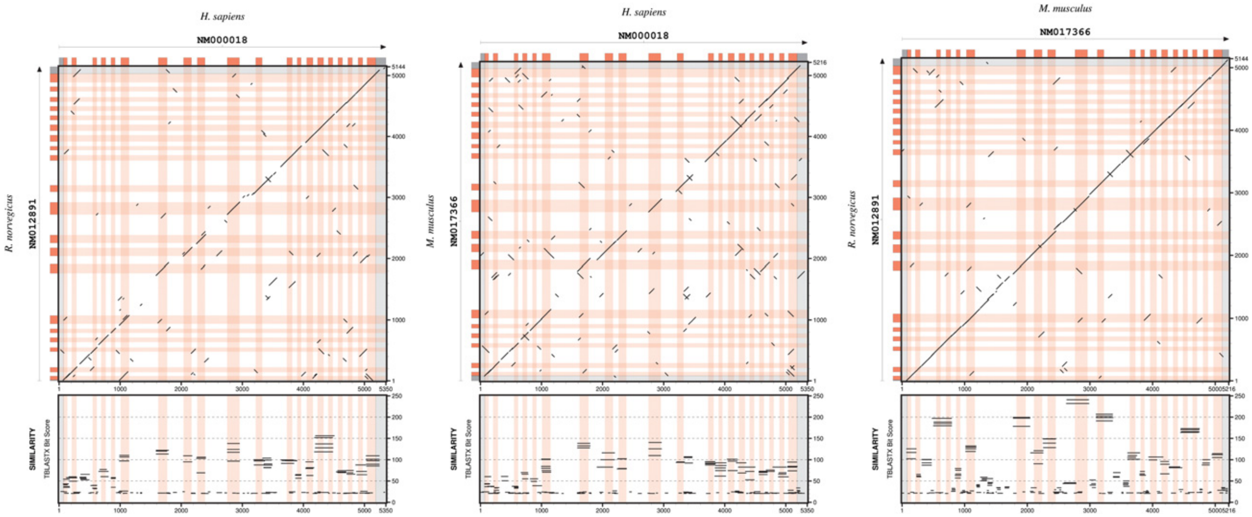

Fig. 6 Comparison of the orthologous genomic region of the acyl-Coenzyme A dehydrogenase very long chain (*ACADVL*) in human, mouse, and rat. The comparison was done at the amino acid level using the ungapped TBLASTX. Coding exons are shown as red boxes and non-coding ones (untranslated regions) are depicted in grey. Exons are projected to highlight their conservation (orange/grey ribbons). Note that introns are also conserved, albeit to a lesser extent than coding exons between human and any of the rodent genomes. The plot was obtained with GFF2APLOT.

Database search programs, however, are not dedicated gene prediction tools; they are not capable of automatically identifying start and stop codons or splice sites, the signals defining the exonic structure of the genes. Thus, after a database search and the identification of potential targets of homology, additional tools are required to define these structures. One approach is to use the top database match as target sequence and obtain a so-called spliced alignment between this and the genomic query. In such an alignment, large gaps – likely to correspond to introns – are only allowed at legal splice junctions. To calculate such spliced alignments between transcript and genomic sequences several programs we have already mentioned have been used: SIM4 (Florea *et al.*, 1998), EST_GENOME (Mott, 1997), SPIDEY (Wheelan *et al.*, 2001), and EXONERATE (Slater and Birney, 2005). Regarding the alignment of proteins of one species to another species genome, GENEWISE (Birney and Durbin, 1997, 2000) and EXONERATE (Slater and Birney, 2005) are used by ENSEMBL. In this direction, Meyer and Durbin (Meyer and Durbin, 2004) have developed the program PROJECTOR, which makes explicit use of the conservation of the exon-intron structure between two related genes. Because human and mouse orthologous genes show a remarkable conservation of their exonic structure, (Mouse Genome Sequencing Consortium, 2002), the PROJECTOR program outperforms GENEWISE predictions when human targets are used in mouse genomic sequences, and vice versa, in particular, when the conservation at the amino level is weak.

In an alternative approach, the results of a database search can be integrated *ad-hoc* into the framework of a typical *ab initio* gene prediction program. In essence, these methods promote candidate exons in the query sequence for which similar known coding sequences exist. Indeed, the score of the candidate exon – initially a function of the score of the splice (start, stop) sites and of the coding potential of the exon sequence – is increased as a function of the similarity between the candidate exon and the known coding sequences. In this way, candidate exons showing similarity to known coding sequences and thus likely orthologous, are promoted into the final gene prediction. In theory, this approach should produce predictions as accurate as pure *ab initio* programs when no similar target sequences exist, but more accurate ones (ideally, as accurate as those from splicing alignment tools) when such target sequences do exist. One example of this approach is the program GENOMESCAN (Yeh *et al.*, 2001), an extension of GENSCAN (Burge and Karlin, 1997) over which it reports increased accuracy. GRAILEXP (Xu *et al.*, 1997), CRASA (Chuang *et al.*, 2003) and AUGUSTUS (Stanke *et al.*, 2006a) are examples of methods using ESTs or mRNAs instead.

Using Whole-Genomes

With the increasing availability of the genome sequences for a wide range of eukaryotic organisms, whole genome sequence comparisons gained popularity as a mean of identifying protein coding genes. Under the assumption that regions conserved in genome sequence alignments will tend to correspond to coding exons from homologous genes, a number of programs were developed. The program EXOFISH (Crollius *et al.*, 2000) was one of the first such programs. Basically, it predicted human exons aided by the comparison of the human genome with sequences from *Tetraodon nigroviridis*, a puffer fish species (within Actinopterygii) that diverged about 400 Myr ago from the lineage later leading to humans. Latter developments followed notably different approaches.

In one such approach (Pedersen and Scharling, 2002; Blayo *et al.*, 2003), the problem was stated as a generalization of pairwise sequence alignment: given two genomic sequences coding for homologous genes, the goal was to obtain the predicted exonic structure in each sequence maximizing the score of the alignment of the resulting amino acid sequences. Both, (Blayo *et al.*, 2003) and Pedersen and Scharling (Pedersen and Scharling, 2002) solved this problem through a complex extension of the classical dynamic programming algorithm for sequence alignment.

In a different approach, the programs SLAM (Alexandersson *et al.*, 2003) and DOUBLESCAN (Meyer and Durbin, 2002) combined sequence alignment pair HMMs (Durbin *et al.*, 1998), with gene prediction generalized HMMs (Burge and Karlin, 1997) into the so-called generalized pair HMMs. In these, gene prediction is not the result of the sequence alignment, as in the programs above, but both gene prediction and sequence alignment are obtained simultaneously.

A third class of programs adopted a more heuristic approach, and clearly separated gene prediction from sequence alignment. The programs ROSETTA (Batzoglou *et al.*, 2000), SGP1 (from Syntenic Gene Prediction, (Wiehe *et al.*, 2001)), and CEM (from the Conserved Exon Method, (Bafna and Huson, 2000)) are representative of this approach. All these programs start by aligning two syntenic sequences – understood here as sequences from different species, but containing homologous genes in a similar order–, and then predict gene structures in which the exons are compatible with the alignment.

Although similarity-based gene prediction with homologous genomic sequences produced high quality results (Miller, 2001), an obvious shortcoming was the need for two homologous sequences. Also, genes without a homologue in the partner sequence will escape detection. This is particularly problematic if species are compared at genomic regions where synteny is not preserved. Given only a single query sequence, it is therefore desirable to automatically search for homologous sequences or syntenic chromosome stretches in other species that are suited to similarity-based approaches. The programs TWINSCAN (Korf *et al.*, 2001) and SGP2 (Parra *et al.*, 2003) attempted to address this limitation. The approach in these programs was reminiscent of that used in GENOMESCAN (Yeh *et al.*, 2001). Essentially, the sequence from the query genome is compared against a collection of sequences from the target or informant genome – which can be a single homologous sequence to the query sequence, a whole assembled genome, or a collection of shotgun reads –, and the results of the comparison are used to modify accordingly the scores of the exons produced by *ab initio* gene predictors. In TWINSCAN, the query and target genome sequences are compared at the nucleotide level with BLASTN and the results serve to modify the underlying probability of the potential exons predicted by GENSCAN. In SGP2, the genome sequences are compared using TBLASTX (Altschul *et al.*, 1990), and the results used to modify the scores of the potential exons predicted by GENEID.

TWINSCAN, SGP2, and SLAM were successfully applied to the annotation of the mouse genome (Mouse Genome Sequencing, C, 2002), and helped to identify previously unconfirmed genes (Guigo *et al.*, 2003). Importantly, a number of studies consistently indicate that comparative gene finders improve over their *ab initio* counterparts (Parra *et al.*, 2003; Wu *et al.*, 2004). Indeed, in **Table 1**, we show that, when evaluated in human chromosome 22, TWINSCAN and SGP2 outperformed GENSCAN and GENEID, respectively. The exhaustive scrutiny to which the sequence of human chromosome 22 (Dunham *et al.*, 1999) was subjected through the Vertebrate Genome Annotation database project at the Wellcome Sanger Institute offers an excellent platform to obtain estimates of the accuracy of current gene finders.

Annotating a genome has required a complex craftsmanship from its inception, given the many different layers of annotations that have to be considered and therefore put together. Common to the tools and approaches described above is the first task of masking the repetitive sequences, often taxa-specific, that populate the genome; this step facilitates and speeds up the downstream analyses just by replacing the segments where the repeats are found by N's, a common nucleotide wildcard – meaning A or C or G or T–. RepeatMasker (Smit *et al.*, 2013-2015), has been the classical tool for this purpose, yet it has been described as sensitive but computationally costly when applied to large eukaryotic genomes (Bedell *et al.*, 2000). After repeat-masking, genome sequences are scanned for protein-coding and non-coding genes using different prediction tools while, simultaneously, homology-based searches and whole-genome alignments to other species are performed. The result are sets of annotations with varying degrees of supporting evidence that need to be integrated into one set of annotated genes, also known as the gene-build of a genome. Additional genome annotation layers result from high-throughput experimental approaches, such as RNA-seq and the transcriptional evidence it produces (which may be used on the gene-build itself), or the typing and resequencing experiments that sample sequence variation across populations (human or otherwise), and so on. Genes from the gene-build, and the proteins they encode, require these other annotation layers to assign them a molecular function, a process known as functional annotation. These functions rely on controlled vocabularies to describe them, for instance the Gene Ontology (GO) (Ashburner *et al.*, 2000; The Gene Ontology Consortium, 2017).

As the complexity of the annotation process increased due to the need to integrate new layers of information, guidelines and/or automatic or semiautomatic workflows to annotate novel genomes or to re-annotate existing ones (every time the assemblies are improved) were developed (Mudge and Harrow, 2016; Elsik *et al.*, 2014). For example, the NCBI prokaryotic (Tatusova *et al.*, 2016) and eukaryotic (Thibaud-Nissen *et al.*, 2016) pipelines, or EnsEMBL (Potter *et al.*, 2004), which rely on in-house protocols for large-scale genome annotation. In addition, gene-finding tools were wrapped into specialized genome annotation workflows. For instance, the Fgenesh (Salamov and Solovyev, 2000), SNAP + Exonerate, and GeneMark + AUGUSTUS were integrated into the Fgenesh + + (Solovyev *et al.*, 2006), MAKER (Cantarel *et al.*, 2008), and BRAKER (Hoff *et al.*, 2016) annotation pipelines, respectively.

Comparative genomics can be also useful to assess the completeness of genome annotations. REFSEQ (Pruitt *et al.*, 2012) has been used as a reference annotation set for the annotation of novel gene-builds, but also to validate the completeness of genome assemblies. A set of 458 highly reliable core proteins, conserved across plants, fungi and mammals, was collected and integrated into a gene-finding validation pipeline known as CEGMA (Parra *et al.*, 2007). More recently, it has been superseded by a set of single-copy orthologs in the BUSCO pipeline (Simao *et al.*, 2015), which can be used to benchmark the completeness of both the annotation of genes and the genome assembly.

Annotation on Non-Standard Protein-Coding Genes

Comparative methods for gene prediction have been also useful to annotate selenoprotein genes. The fact that downstream of a recoded stop codon lies a real protein implies this region to have the typical pattern of sequence conservation among its protein homologues. At the least, to the extent of the conservation in the upstream protein sequence. Therefore, sequence conservation beyond a TGA codon in the genome strongly suggests selenocysteine coding function. It is then feasible to predict selenoprotein genes in two different species genomes, without information from their SECIS elements, and compare them to identify a pattern of symmetrical protein sequence conservation around a predicted selenocysteine codon. Application of this approach to the human and fugu genomes identified a new selenoprotein in this fish (Castellano *et al.*, 2004). More recently, selenoproteins have been annotated across eukaryotes, in a SECIS-independent manner, using alignment profiles derived from the known selenoprotein families (Mariotti and Guigo, 2010).

The RNA and Regulatory Years

The functional annotation of the regulatory and non-coding part of the genome has lagged behind the annotation of protein-coding genes. Systematic analysis of RNA genes and regulatory regions was made possible in the mid and late 2000s by a number of experimental approaches for the characterization of transcribed regions, transcription factor binding sites, DNA methylation sites and chromatin structure. The ENCODE project was an early adopter of these methods to systematically characterize the human genome, (2012ENCODE Project Consortium). Among these, tiling arrays and RNA-seq have been important to understand the transcriptional landscape of the human genome, which is pervasively transcribed (Djebali *et al.*, 2012; Bertone *et al.*, 2004). This was perhaps not unexpected if one acknowledged the high initiation promiscuity of *Pol II*, the eukaryotic RNA polymerase complex (Struhl, 2007). In any case, transcription from regions where protein-coding genes or known RNA genes – such as ribosomal and transfer RNAs, small nuclear RNAs involved in splicing, microRNAs, short interfering RNAs, and a few others – have so far not been annotated, produced thousands of long noncoding RNAs with no obvious function. These unknown RNAs (Derrien *et al.*, 2012), which standard gene prediction programs cannot detect using codon bias measures, constitute a vast array of potential genes that do not encode for proteins (Kowalczyk *et al.*, 2012), albeit some may do (Ji *et al.*, 2015). While a few of these new RNAs have been shown to have functions (Nesterova *et al.*, 2001; Sleutels *et al.*, 2002), there is little functional evidence for the majority of them. Their transcription is often extremely low and they tend to have limited sequence or RNA structure conservation (Rivas *et al.*, 2017), raising the possibility that they are transcriptional noise. Indeed, between and within species analyses indicate that most of these RNA genes are not under strong evolutionary constraint (purifying selection) (Wiberg *et al.*, 2015). Alternatively, it could be their transcription itself rather than their sequence that is under selection. The functional annotation of RNA genes thus remains open and additional work is needed before this controversy is solved.

The high-throughput sequencing of transcriptomes and other experimental approaches has required the development of fast aligners like `BWA` (Li and Durbin, 2009) and `Bowtie` (Langmead and Salzberg, 2012), which can align millions of short sequences to a reference genome sequence. Specialized tools can integrate those alignments into predicted gene structures, such as `AUGUSTUS` (Stanke *et al.*, 2006b), `mGENE` (Schweikert *et al.*, 2009), or the `Transomics` pipeline (Solovyev *et al.*, 2006). Other tools, like `SpliceGrapher` (Rogers *et al.*, 2012), have been developed to refine the alignments of mapped sequences to predict "splicing graphs", a representation of the splicing variants of a gene in which exons are shown as nodes that are connected by the intervening introns represented as edges. As discussed, obtaining simultaneously the genome and transcriptome of a species facilitates the task of gene annotation, by aligning the transcriptome sequences along the newly assembled genome, and provides an estimation of gene expression levels when translating the alignment coverage of read sequences into normalized counts (Mortazavi *et al.*, 2008). An extensive assessment of the sequence mappers (Engstrom *et al.*, 2013) and the gene-finders (Steijger *et al.*, 2013) was undertaken by the GENCODE consortium (Harrow *et al.*, 2012). Among the results obtained it is worth mentioning that 5′- and 3′-UTR exons were still difficult to predict (see **Fig. 3**) as well as the different exon combinations for splicing transcript isoforms, and that prediction accuracy correlated with the gene expression level, as highly expressed genes are often more abundant in the training sets. However, this is changing as single molecule sequencing methods, such as PacBio or ONT nanopore, produce long sequences recovering the full length splicing isoforms of a gene (Sharon *et al.*, 2013). Thus easing the task of protein-coding and non-coding mRNA genes annotation, even at single cell resolution (Byrne *et al.*, 2017).

Regulatory regions where proteins (*e.g.*, transcription factors) bind have been studied using chromatin immunoprecipitation followed by sequencing (ChIP-seq). Around 8% of the genome across different cell types is typically protein-bound with sequence-specific signals known to bind proteins being common in them. In addition, *DNase I* hypersensitivity has been used to identify those regions of the genome open to transcription. These regions can be bound by transcription factors and were found to be upstream of the predicted transcription start sites, coinciding with the regions defined by the ChIP-seq experiments. Methylation of cytosine, usually at CpG dinucleotides, is involved in the epigenetic regulation of gene expression, with promoter and gene methylation leading to repression and promotion of transcription, respectively. Levels of DNA methylation correlate with chromatin accessibility as defined by the *DNase I* hypersensitivity experiments. Interestingly, most variable regions in their methylation correspond to genes rather than their promoters.

One surprising result from the ENCODE project was the finding that the majority (about 80%) of the human genome participated in one or more of the experiments above (ENCODE Project Consortium, 2012), mostly in the RNA transcription experiments discussed above. This was initially interpreted as providing evidence for a functional role to most of the human DNA sequence but low specificity of the assays used is a simpler explanation. These assays often do not distinguish between meaningful function – some molecular function that impacts fitness in the individual – and one that does not. As a result, having the majority of the genome performing important roles contrasts with the small fraction of it – probably 5%–10% – that has been shown to be under evolutionary constraint (Asthana *et al.*, 2007; Davydov *et al.*, 2010). It follows that some fraction of non-coding DNA is functional in the typical meaning of this word but that the majority of it – mostly transposable elements that make up about half of the human genome (Eddy, 2012) – is not (albeit being biochemically active in different experiments). Indeed, one of the common tasks in gene prediction is masking about half of the human genome for repeats and transposons using the venerable REPEATMASKER program.

The Population Years

The functional annotation of a species genome provides a reference to which additional genomes from the same species can be compared. Thousands of human exomes (protein-coding fraction of the genome) or genomes have now been sequenced (1k Genomes Project Consortium, 2012, 2015) and, in doing so, many differences have been found between them and their reference – *e.g.*, between four and five million nucleotide differences in each human genome –. These differences, which include single nucleotide variants, insertions and deletions, copy number variants and others, are often novel and unannotated, which each ensuing genome having more of them. These rare variants in humans, which are found in specific populations and at low frequency, are to a large the result of the recent growth of human populations (Keinan and Clark, 2012). Indeed, humans have accumulated in the last 5000–10,000 years an enormous number of rare variants, particularly in non-African populations. Protein-coding variants at low frequency are often slightly deleterious as measured by approaches that infer the probability of amino acid changes impacting the structure or function of proteins (Fu *et al.*, 2013). Programs like POLYPHEN (Adzhubei *et al.*, 2010) and SIFT (Kumar *et al.*, 2009), using amino acid conservation (within and between species) and different amino acid properties, annotate their deleteriousness across the genome. In addition, a few hundred variants disrupt the translation of proteins in a typical human genome. Non-coding variants in regulatory regions have also been annotated with regard to their deleteriousness using, for example, their conservation among species with PHASTCONS (Siepel *et al.*, 2005). It follows that the functional annotation of single nucleotide variants, particularly those that change (amino acid substitutions) or disrupt proteins (e.g., changing splice sites or creating stop codons), are important to understand interindividual differences linked to diseases that are both rare and severe. Indeed, rare diseases, understood as those that occur in one out of 2000 individuals, have often a genetic basis and the annotation of the deleteriousness of the causal mutation(s) in single-gene disorders is the first step in their diagnosis and treatment.

Still, most of the genetic variation between individuals genomes of a species is shared among some (or most) of them. This common variation was described in a landmark study that provided not only a catalog of single nucleotide variants in humans but the genetic association among them (International HapMap Consortium, 2005). This association, known as linkage disequilibrium, reflects the coinheritance of sets of single nucleotide variants (haplotypes) and their functional annotation is needed to investigate, using genome-wide association studies, the hereditary factors involved in disease. This is important as each human genome has around 10,000 rare and common amino acid changes, a few hundred protein-disrupting ones and about 500,000 non-coding changes in regulatory regions (1k Genomes Project Consortium, 2015). Of these, about 2000 are linked through genome-wide studies with common disease whereas a few dozen are implicated with rare disease per genome. Today, thousands of mutations have been associated to common or rare disease in humans and their genome and functional annotation can be obtained from different databases, for example OMIM (http://omim.org), CLINVAR (Landrum *et al.*, 2016) and the NHGRI-EBI GWAS catalog (MacArthur *et al.*, 2017).

Conclusion

Genome annotation has come a long way since the first protein-coding genes were annotated using *ab initio* and later comparative approaches. Experimental evidence is now routinely integrated in the annotation of these genes and their multiple alternative splice forms. At the same time, promoters and smaller regulatory elements, as well as RNA genes (including controversial ones), are being annotated with a variety of experimental approaches genome-wide. Knowing the location of these coding and noncoding functional features has provided the framework to understand the role of sequence variants in health and disease, which remains largely undetermined. Still, since the early nineties, when the Human Genome Project was launched, our knowledge of genome biology has greatly improved. This is not only due to the ever increasing computational capabilities but also to novel discoveries in molecular biology from experimental methods with increased resolution and throughput, which has been harnessed in annotation consortiums like the ENCODE project. As a result, well annotated genomes like that of human, fly or mice, the result of years of intense human curation, as well as gold standards such REFSEQ or VEGA genes, provide reference annotations to other species.

This is important as the pace of sequencing new species genomes, transcriptomes, and environmental (metagenomics) samples is only accelerating given the ever decreasing cost of sequencing technologies. Thus, a shift towards more automated annotation pipelines is warranted to cope with the scale of the analyses in genome annotation. Automated annotation pipelines have room to improve though, with gene-builds still having a significant impact in the downstream analyses. For example, on the variants found in a genome and their predicted functional consequences (Frankish *et al.*, 2015). This is due to the fact that most gene-finding tools return a single and often different set of optimal gene structures. This also poses a problem for overlapping genes, including those on different strands, nested genes, genes located within introns of other genes, non-standard gene structures, and genes with translation recoding like selenoprotein genes, as they are often not included in the optimal prediction of genes. Furthermore, most gene prediction tools do not consider frame shifts, RNA editing, and the impact of sequencing errors in the identification of gene structures. Still, some tools are capable to infer more than one alternative splicing isoform provided that there is some sort of experimental evidence, for example, from RNA-seq experiments. However, problems may appear when transcripts contain start or stop codons split by introns. Together with non-cannonical splice-sites, these are some of the special cases that human curators solve in the final refinement steps of genome annotation when all sort of evidences are integrated. Many of aforementioned issues will be overcome by the use of single molecule sequencing technologies, which make possible to map the full length of individual protein-coding and non-coding transcript isoforms. In combination with single cell gene expression experiments, it will be finally feasible to produce gene-builds and their corresponding functional annotation at the single molecule and cell resolution, an astounding promise just two decades ago.

Acknowledgements

This study was partially funded (Sergi Castellano) by the NIHR HS&DR Programme (14/21/45) and supported by the NIHR GOSH BRC. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

See also: Algorithms for Strings and Sequences: Multiple Alignment. Algorithms for Strings and Sequences: Pairwise Alignment. Bioinformatics Approaches for Studying Alternative Splicing. Comparative and Evolutionary Genomics. Comparative Genomics Analysis. Data Mining: Classification and Prediction. Data Mining: Prediction Methods. Detecting and Annotating Rare Variants. Exome Sequencing Data Analysis. Functional Enrichment Analysis. Functional Genomics. Gene Mapping. Genome Annotation: Perspective From Bacterial Genomes. Genome Databases and Browsers. Genome Informatics. Genome-Wide Haplotype Association Study. Hidden Markov Models. Integrative Analysis of Multi-Omics Data. Linkage Disequilibrium. Metagenomic Analysis and its Applications. Natural Language Processing Approaches in Bioinformatics. Next Generation Sequencing Data Analysis. Ontology in Bioinformatics. Ontology in Bioinformatics. Ontology-Based Annotation Methods. Phylogenetic Footprinting. Prediction of Coding and Non-Coding RNA. Protein Functional Annotation. Quantitative Immunology by Data Analysis Using Mathematical Models. Sequence Analysis. Sequence Composition. Single Nucleotide Polymorphism Typing. Whole Genome Sequencing Analysis

References

Abril, J.F., Castelo, R., Guigo, R., 2005. Comparison of splice sites in mammals and chicken. Genome Res. 15 (1), 111–119.
Abril, J.F., Guigo, R., 2000. gff2ps: Visualizing genomic annotations. Bioinformatics 16 (8), 743–744.
Abril, J.F., Guigo, R., Wiehe, T., 2003. gff2aplot: Plotting sequence comparisons. Bioinformatics 19 (18), 2477–2479.
Adams, M.D., *et al*., 2000. The genome sequence of *Drosophila melanogaster*. Science 287 (5461), 2185–2195.
Adzhubei, I.A., *et al*., 2010. A method and server for predicting damaging missense mutations. Nat. Methods 7 (4), 248–249.
Aken, B.L., *et al*., 2016. The Ensembl gene annotation system. Database (Oxford) 2016.
Aken, B.L., *et al*., 2017. Ensembl 2017. Nucleic Acids Res. 45 (D1), D635–D642.
Alexandersson, M., Cawley, S., Pachter, L., 2003. SLAM: Cross-species gene finding and alignment with a generalized pair hidden Markov model. Genome Res. 13 (3), 496–502.
Altschul, S.F., *et al*., 1990. Basic local alignment search tool. J. Mol. Biol. 215 (3), 403–410.
Ashburner, M., *et al*., 2000. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25 (1), 25–29.
Asthana, S., *et al*., 2007. Widely distributed noncoding purifying selection in the human genome. Proc. Natl. Acad. Sci. USA 104 (30), 12410–12415.
Bafna, V., Huson, D.H., 2000. The conserved exon method for gene finding. Proc. Int. Conf. Intell. Syst. Mol. Biol. 8, 3–12.
Bateman, A., *et al*., 2017. UniProt: The universal protein knowledgebase. Nucleic Acids Res. 45 (D1), D158–D169.
Batzoglou, S., *et al*., 2000. Human and mouse gene structure: Comparative analysis and application to exon prediction. Genome Res. 10 (7), 950–958.
Bedell, J.A., Korf, I., Gish, W., 2000. MaskerAid: A performance enhancement to RepeatMasker. Bioinformatics 16 (11), 1040–1041.
Benson, D.A., *et al*., 2013. GenBank. Nucleic Acids Res. 41 (D1), D36–D42.
Bertone, P., *et al*., 2004. Global identification of human transcribed sequences with genome tiling arrays. Science 306 (5705), 2242–2246.
Birney, E., Durbin, R., 1997. Dynamite: A flexible code generating language for dynamic programming methods used in sequence comparison. In: Proceedings of Ismb-97 – Fifth International Conference on Intelligent Systems for Molecular Biology, pp. 56–64.
Birney, E., Durbin, R., 2000. Using GeneWise in the *Drosophila* annotation experiment. Genome Res. 10 (4), 547–548.
Blayo, P., Rouze, P., Sagot, M.F., 2003. Orphan gene finding – An exon assembly approach. Theor. Comput. Sci. 290 (3), 1407–1431.
Burge, C., Karlin, S., 1997. Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 268 (1), 78–94.
Burset, M., Guigo, R., 1996. Evaluation of gene structure prediction programs. Genomics 34 (3), 353–367.
Burset, M., Seledtsov, I.A., Solovyev, V.V., 2000. Analysis of canonical and non-canonical splice sites in mammalian genomes. Nucleic Acids Res. 28 (21), 4364–4375.
Byrne, A., *et al*., 2017. Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. Nat. Commun. 8, 16027.
Cantarel, B.L., *et al*., 2008. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 18 (1), 188–196.

Castellano, S., *et al*., 2001. *In silico* identification of novel selenoproteins in the *Drosophila melanogaster* genome. EMBO Rep. 2 (8), 697–702.

Castellano, S., *et al*., 2004. Reconsidering the evolution of eukaryotic selenoproteins: A novel nonmammalian family with scattered phylogenetic distribution. EMBO Rep. 5 (1), 71–77.

Castellano, S., *et al*., 2005. Diversity and functional plasticity of eukaryotic selenoproteins: Identification and characterization of the SelJ family. Proc. Natl. Acad. Sci. USA 102 (45), 16188–16193.

Chambers, I., *et al*., 1986. The structure of the mouse glutathione peroxidase gene: The selenocysteine in the active site is encoded by the 'termination' codon, TGA. EMBO J. 5 (6), 1221–1227.

Chuang, T.J., *et al*., 2003. A complexity reduction algorithm for analysis and annotation of large genomic sequences. Genome Res. 13 (2), 313–322.

Coghlan, A., *et al*., 2008. nGASP – The nematode genome annotation assessment project. BMC Bioinform. 9, 549.

E.PENCODE, Project Consortium, 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489 (7414), 57–74.

Crollius, H.R., *et al*., 2000. Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. Nat. Genet. 25 (2), 235–238.

Curwen, V., *et al*., 2004. The Ensembl automatic gene annotation system. Genome Res. 14 (5), 942–950.

Davydov, E.V., *et al*., 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP plus. PLOS Comput. Biol. 6 (12),

Delcher, A.L., *et al*., 1999. Improved microbial gene identification with GLIMMER. Nucleic Acids Res. 27 (23), 4636–4641.

Derrien, T., *et al*., 2012. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. Genome Res. 22 (9), 1775–1789.

Djebali, S., *et al*., 2012. Landscape of transcription in human cells. Nature 489 (7414), 101–108.

Dunham, I., *et al*., 1999. The DNA sequence of human chromosome 22. Nature 402 (6761), 489–495.

Durbin, R., *et al*., 1998. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge, United Kingdom: Cambridge University Press, (xi, 356 pages: Illustrations; 26 cm).

Eddy, S.R., 2004a. What is dynamic programming? Nat. Biotechnol. 22 (7), 909–910.

Eddy, S.R., 2004b. What is a hidden Markov model? Nat. Biotechnol. 22 (10), 1315–1316.

Eddy, S.R., 2012. The C-value paradox, junk DNA and ENCODE. Curr. Biol. 22 (21), R898–R899.

Eilbeck, K., *et al*., 2009. Quantitative measures for the management and comparison of annotated genomes. BMC Bioinform. 10, 67.

Elsik, C.G., *et al*., 2014. Finding the missing honey bee genes: Lessons learned from a genome upgrade. BMC Genomics 15, 86.

Engstrom, P.G., *et al*., 2013. Systematic evaluation of spliced alignment programs for RNA-seq data. Nat. Methods 10 (12), 1185–1191.

Florea, L., *et al*., 1998. A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res. 8 (9), 967–974.

Frankish, A., *et al*., 2015. Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction. BMC Genom. 16 (Suppl. 8), S2.

Fu, W.Q., *et al*., 2013. Analysis of 6515 exomes reveals the recent origin of most human protein-coding variants. Nature 493 (7431), 216–220.

Gelfand, M.S., Mironov, A.A., Pevzner, P.A., 1996. Gene recognition via spliced sequence alignment. Proc. Natl. Acad. Sci. USA 93 (17), 9061–9066.

1k Genomes Project Consortium, *et al*., 2012. An integrated map of genetic variation from 1092 human genomes. Nature 491 (7422), 56–65.

1k Genomes Project Consortium, *et al*., 2015. A global reference for human genetic variation. Nature 526 (7571), 68–74.

Giardine, B., *et al*., 2005. Galaxy: A platform for interactive large-scale genome analysis. Genome Res. 15 (10), 1451–1455.

Gish, W., States, D.J., 1993. Identification of protein coding regions by database similarity search. Nat. Genet. 3 (3), 266–272.

Gotoh, O., 1982. An improved algorithm for matching biological sequences. J. Mol. Biol. 162 (3), 705–708.

Guigo, R., *et al*., 1992. Prediction of gene structure. J. Mol. Biol. 226 (1), 141–157.

Guigo, R., 1998. Assembling genes from predicted exons in linear time with dynamic programming. J. Comput. Biol. 5 (4), 681–702.

Guigo, R., *et al*., 2003. Comparison of mouse and human genomes followed by experimental verification yields an estimated 1019 additional genes. Proc. Natl. Acad. Sci. USA 100 (3), 1140–1145.

Guigo, R., *et al*., 2006. EGASP: The human ENCODE genome annotation assessment project. Genome Biol. 7 (Suppl. 1), S2 1–31.

Harrow, J., *et al*., 2012. GENCODE: The reference human genome annotation for The ENCODE project. Genome Res. 22 (9), 1760–1774.

Hastings, K.E.M., 2005. SL trans-splicing: Easy come or easy go? Trends Genet. 21 (4), 240–247.

Hoff, K.J., *et al*., 2016. BRAKER1: Unsupervised RNA-seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics 32 (5), 767–769.

Hubbard, T., *et al*., 2002. The Ensembl genome database project. Nucleic Acids Res. 30 (1), 38–41.

Hyatt, D., *et al*., 2010. Prodigal: Prokaryotic gene recognition and translation initiation site identification. BMC Bioinform. 11, 119.

International HapMap Consortium, C., 2005. A haplotype map of the human genome. Nature 437 (7063), 1299–1320.

Iseli, C., Jongeneel, C.V., Bucher, P., 1999. ESTScan: A program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. Proc. Int. Conf. Intell. Syst. Mol. Biol. 138–148.

Ji, Z., *et al*., 2015. Many lncRNAs, 5′ UTRs, and pseudogenes are translated and some are likely to express functional proteins. eLife 4.

Katzav, S., Martin-Zanca, D., Barbacid, M., 1989. *vav*, a novel human oncogene derived from a locus ubiquitously expressed in hematopoietic cells. EMBO J. 8 (8), 2283–2290.

Keinan, A., Clark, A.G., 2012. Recent explosive human population growth has resulted in an excess of rare genetic variants. Science 336 (6082), 740–743.

Kent, W.J., *et al*., 2002. The human genome browser at UCSC. Genome Res. 12 (6), 996–1006.

Korf, I., *et al*., 2001. Integrating genomic homology into gene structure prediction. Bioinformatics 17 (Suppl. 1), S140–S148.

Kowalczyk, M.S., Higgs, D.R., Gingeras, T.R., 2012. Molecular biology: RNA discrimination. Nature 482 (7385), 310–311.

Kryukov, G.V., *et al*., 2003. Characterization of mammalian selenoproteomes. Science 300 (5624), 1439–1443.

Kryukov, G.V., Kryukov, V.M., Gladyshev, V.N., 1999. New mammalian selenocysteine-containing proteins identified with an algorithm that searches for selenocysteine insertion sequence elements. J. Biol. Chem. 274 (48), 33888–33897.

Kumar, P., Henikoff, S., Ng, P.C., 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc. 4 (7), 1073–1082.

Kyriakopoulou, C., *et al*., 2006. U1-like snRNAs lacking complementarity to canonical 5′ splice sites. RNA 12 (9), 1603–1611.

Krzywinski, M., *et al*., 2009. Circos: an Information Aesthetic for Comparative Genomics. Genome Res. 19, 1639–1645.

Landrum, M.J., *et al*., 2016. ClinVar: Public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 44 (D1), D862–D868.

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9 (4), 357–359.

Lescure, A., *et al*., 1999. Novel selenoproteins identified *in silico* and *in vivo* by using a conserved RNA structural motif. J. Biol. Chem. 274 (53), 38147–38154.

Lewis, S.E., *et al*., 2002. Apollo: A sequence annotation editor. Genome Biol. 3 (12), (RESEARCH0082).

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

Loveland, J.E., *et al*., 2012. Community gene annotation in practice. Database – J. Biol. Databases Curation.

Lukashin, A.V., Borodovsky, M., 1998. GeneMark.hmm: New solutions for gene finding. Nucleic Acids Res. 26 (4), 1107–1115.

MacArthur, J., *et al*., 2017. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 45 (D1), D896–D901.

Madupu, R., *et al*., 2010. Meeting report: A workshop on best practices in genome annotation. Database (Oxford) 2010, baq001.

Margulies, E.H., *et al*., 2003. Identification and characterization of multi-species conserved sequences. Genome Res. 13 (12), 2507–2518.

Mariotti, M., Guigo, R., 2010. Selenoprofiles: Profile-based scanning of eukaryotic genome sequences for selenoprotein genes. Bioinformatics 26 (21), 2656–2663.

Meyer, I.M., Durbin, R., 2002. Comparative *ab initio* prediction of gene structures using pair HMMs. Bioinformatics 18 (10), 1309–1318.

Meyer, I.M., Durbin, R., 2004. Gene structure conservation aids similarity based gene prediction. Nucleic Acids Res. 32 (2), 776–783.

Miller, W., 2001. Comparison of genomic DNA sequences: Solved and unsolved problems. Bioinformatics 17 (5), 391–397.

Monaco, A.P., *et al.*, 1986. Isolation of candidate cDNAs for portions of the Duchenne muscular dystrophy gene. Nature 323 (6089), 646–650.

Mortazavi, A., *et al.*, 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods 5 (7), 621–628.

Mott, R., 1997. EST_GENOME: A program to align spliced DNA sequences to unspliced genomic DNA. Comput. Appl. Biosci. 13 (4), 477–478.

Mouse Genome Sequencing Consortium, 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420 (6915), 520–562.

Mudge, J.M., Harrow, J., 2016. The state of play in higher eukaryote gene annotation. Nat. Rev. Genet. 17 (12), 758–772.

Needleman, S.B., Wunsch, C.D., 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48 (3), 443–453.

Nesterova, T.B., *et al.*, 2001. Characterization of the genomic Xist locus in rodents reveals conservation of overall gene structure and tandem repeats but rapid evolution of unique sequence. Genome Res. 11 (5), 833–849.

Parra, G., *et al.*, 2003. Comparative gene prediction in human and mouse. Genome Res. 13 (1), 108–117.

Parra, G., *et al.*, 2006. Tandem chimerism as a means to increase protein complexity in the human genome. Genome Res. 16 (1), 37–44.

Parra, G., Bradnam, K., Korf, I., 2007. CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23 (9), 1061–1067.

Patel, A.A., Steitz, J.A., 2003. Splicing double: Insights from the second spliceosome. Nat. Rev. Mol. Cell Biol. 4 (12), 960–970.

Pearson, W.R., 2000. Flexible sequence similarity searching with the FASTA3 program package. Methods Mol. Biol. 132, 185–219.

Pearson, W.R., Lipman, D.J., 1988. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA 85 (8), 2444–2448.

Pedersen, C.N.S., Scharling, T., 2002. Comparative methods for gene structure prediction in homologous sequences. Proc. Algorithms Bioinform. 2452, 220–234.

Pennisi, E., 2000. Ideas fly at gene-finding jamboree. Science 287 (5461), 2182. +.

Plotkin, J.B., Kudla, G., 2011. Synonymous but not the same: The causes and consequences of codon bias. Nat. Rev. Genet. 12 (1), 32–42.

Potter, S.C., *et al.*, 2004. The Ensembl analysis pipeline. Genome Res. 14 (5), 934–941.

Powers, D., 2011. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. Int. J. Mach. Learn. Technol. 2, 37–63.

Pruitt, K.D., *et al.*, 2012. NCBI Reference Sequences (RefSeq): Current status, new features and genome annotation policy. Nucleic Acids Res. 40 (Database issue), D130–D135.

Rayman, M.P., 2000. The importance of selenium to human health. Lancet 356 (9225), 233–241.

Reese, M.G., *et al.*, 2000. Genome annotation assessment in *Drosophila melanogaster*. Genome Res. 10 (4), 483–501.

Rivas, E., Clements, J., Eddy, S.R., 2017. A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. Nat. Methods 14 (1), 45–48.

Rogers, M.F., *et al.*, 2012. SpliceGrapher: Detecting patterns of alternative splicing from RNA-Seq data in the context of gene models and EST data. Genome Biol. 13 (1), R4.

Salamov, A.A., Solovyev, V.V., 2000. *Ab initio* gene finding in *Drosophila* genomic DNA. Genome Res. 10 (4), 516–522.

Schweikert, G., *et al.*, 2009. mGene: Accurate SVM-based gene finding with an application to nematode genomes. Genome Res. 19 (11), 2133–2143.

Sharon, D., *et al.*, 2013. A single-molecule long-read survey of the human transcriptome. Nat. Biotechnol. 31 (11), 1009–1014.

Sharp, P.A., 2005. The discovery of split genes and RNA splicing. Trends Biochem. Sci. 30 (6), 279–281.

Siepel, A., *et al.*, 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15 (8), 1034–1050.

Simao, F.A., *et al.*, 2015. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31 (19), 3210–3212.

Slater, G.S., Birney, E., 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinform. 6, 31.

Sleutels, F., Zwart, R., Barlow, D.P., 2002. The non-coding Air RNA is required for silencing autosomal imprinted genes. Nature 415 (6873), 810–813.

Smith, T.F., Waterman, M.S., 1981. Identification of common molecular subsequences. J. Mol. Biol. 147 (1), 195–197.

Smit, A.F.A., Hubley, R., Green, P., *RepeatMasker Open-4.0*, 2013-2015, http://www.repeatmasker.org.

Solovyev, V., *et al.*, 2006. Automatic annotation of eukaryotic genes, pseudogenes and promoters. Genome Biol. 7 (Suppl. 1), S10 1–12.

Stanke, M., *et al.*, 2006a. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinform. 7.

Stanke, M., Tzvetkova, A., Morgenstern, B., 2006b. AUGUSTUS at EGASP: Using EST, protein and genomic alignments for improved gene prediction in the human genome. Genome Biol. 7.

Steijger, T., *et al.*, 2013. Assessment of transcript reconstruction methods for RNA-seq. Nat. Methods 10 (12), 1177–1184.

Stein, L.D., *et al.*, 2002. The generic genome browser: A building block for a model organism system database. Genome Res. 12 (10), 1599–1610.

Struhl, K., 2007. Transcriptional noise and the fidelity of initiation by RNA polymerase II. Nat. Struct. Mol. Biol. 14 (2), 103–105.

Su, M., *et al.*, 2013. Small proteins: Untapped area of potential biological importance. Front. Genet. 4, 286.

Tatusova, T., *et al.*, 2016. NCBI prokaryotic genome annotation pipeline. Nucleic Acids Res. 44 (14), 6614–6624.

The Gene Ontology Consortium,, 2017. Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Res. 45 (D1), D331–D338.

Thibaud-Nissen, F., *et al.*, 2016. The NCBI eukaryotic genome annotation pipeline. J. Anim. Sci. 94, 184.

Thorvaldsdottir, H., Robinson, J.T., Mesirov, J.P., 2013. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. Brief. Bioinform. 14 (2), 178–192.

Wheelan, S.J., Church, D.M., Ostell, J.M., 2001. Spidey: A tool for mRNA-to-genomic alignments. Genome Res. 11 (11), 1952–1957.

Wiberg, R.A.W., *et al.*, 2015. Assessing recent selection and functionality at long noncoding RNA loci in the mouse genome. Genome Biol. Evol. 7 (8), 2432–2444.

Wiehe, T., *et al.*, 2001. SGP-1: Prediction and validation of homologous genes based on sequence alignments. Genome Res. 11 (9), 1574–1583.

Wu, J.Q., *et al.*, 2004. Identification of rat genes by TWINSCAN gene prediction, RT-PCR, and direct sequencing. Genome Res. 14 (4), 665–671.

Xu, Y., Mural, R.J., Uberbacher, E.C., 1997. Inferring gene structures in genomic sequences using pattern recognition and expressed sequence tags. In: Proceedings Ismb-97 – Fifth International Conference on Intelligent Systems for Molecular Biology, pp. 344–353.

Yeh, R.F., Lim, L.P., Burge, C.B., 2001. Computational inference of homologous gene structures in the human genome. Genome Res. 11 (5), 803–816.

Zhang, Y., *et al.*, 2005. Pyrrolysine and selenocysteine use dissimilar decoding strategies. J. Biol. Chem. 280 (21), 20740–20751.

Zheng, D., *et al.*, 2007. Pseudogenes in the ENCODE regions: Consensus annotation, analysis of transcription, and evolution. Genome Res. 17 (6), 839–851.

Zinoni, F., *et al.*, 1986. Nucleotide sequence and expression of the selenocysteine-containing polypeptide of formate dehydrogenase (formate-hydrogen-lyase-linked) from *Escherichia coli*. Proc. Natl. Acad. Sci. USA 83 (13), 4650–4654.

Biographical Sketch

Josep F. Abril, PhD Associate Professor at the Department of Genetics, Microbiology and Statistics of the Universitat de Barcelona. He earned his Bachelor's degree in Biology from the Universitat de Barcelona, and his PhD in Bioinformatics from the Universitat Pompeu Fabra in Barcelona. His research focuses on the computational analysis of sequences and their annotated features. He has worked on a variety of genomic and transcriptomic projects, but also on the functional characterization of proteins, and on different areas in computational biology. These include the development of protocols for the assembly and annotation of genomes, the use of comparative genomics methods, the analysis of gene expression, the modeling of splice sites and exonic structure of eukaryotic genes, the visualization of whole-genome annotations – this includes the human genome map published in *Science* in 2001 – the integration of expression and variation data into interaction networks, as well as the characterization of viral samples from metagenomic experiments. His organisms of interest are planarians, flies, and humans, not necessarily in that order, although he contributed to the annotation of other species too. Finally, he has been involved in the organization and analysis of three of the most relevant gene-prediction accuracy assessments in computational gene prediction, namely GASP in the *Drosophila* genome and EGASP and RGASP in the human one.

Sergi Castellano, PhD Associate Professor at the Genetics and Genomics Medicine Programme at University College London. He received his Bachelor's degree in Biology from the Universitat de Barcelona, and his PhD in Bioinformatics from the Universitat Pompeu Fabra, also in Barcelona. His group works on understanding the role of essential micronutrients, with particular emphasis on selenium, in the adaptation of human metabolism to the different environments encountered by archaic and modern humans. Much of his early work focused on the identification and characterization of functional elements in eukaryotic genomes, primarily selenium-containing genes, which are missannotated in standard databases. At the University of Hawaii, he developed the first database, `SELENODB`, with correct annotations for selenium-containing genes across animal genomes. Later, while at Cornell University and Howard Hughes Medical Institute, he used population genetics and molecular evolution approaches to study the evolution of the use of selenocysteine, the selenium-containing amino acid, compared to cysteine in proteins. More recently, at Max Planck in Germany, his group contributed to the population genetics analysis of Neandertals as it relates to their coding variation and their first encounters – as early as 100,000 years ago – with modern humans. His group is currently working on the role of micronutrient deficiencies in common and rare disease.