A bioinformatics toolkit: *in silico* tools and online resources for investigating genetic variation

Simon J. Webster, PhD,[1] Maryam A. Aldossary, MSc,[1] and Daniel J. Hampshire, PhD[1,2]

[1]Department of Infection, Immunity and Cardiovascular Disease, University of Sheffield, Sheffield, UK and [2]Department of Biomedical Sciences, University of Hull, Hull, UK

Co-authors:

Dr. Simon J. Webster, Department of Infection, Immunity and Cardiovascular Disease, Faculty of Medicine, Dentistry and Health, University of Sheffield, Beech Hill Road, Sheffield S10 2RX, UK

Tel.: +44 (0)114 215 9589; e-mail: simon.j.webster@sheffield.ac.uk

Mrs. Maryam A. Aldossary, Department of Infection, Immunity and Cardiovascular Disease, Faculty of Medicine, Dentistry and Health, University of Sheffield, Beech Hill Road, Sheffield S10 2RX, UK

Tel.: +44 (0)114 215 9590; e-mail: maaldossary2@sheffield.ac.uk

Corresponding author:

Dr. Daniel J. Hampshire, Department of Biomedical Sciences, Faculty of Health Sciences, University of Hull, Cottingham Road, Hull HU6 7RX, UK

Tel.: +44 (0)1482 46 2316; e-mail: d.hampshire@hull.ac.uk

Running title: A bioinformatics toolkit

Abstract: 73

Main text: 3,471

Figures: 3

Tables: 4

References: 104

This is an Accepted Manuscript of an article published by Thieme in Seminars in Thrombosis and Hemostasis on 5 Aug 2019, available online at https://www.thieme-connect.de/products/ejournals/abstract/10.1055/s-0039-1692978

Abstract

With the advent of large-scale next generation sequencing initiatives, there is an increasing importance to interpret and understand the potential phenotypic influence of identified genetic variation and its significance in the human genome. Bioinformatics analyses can provide useful information to assist with variant interpretation. This review provides an overview of tools / resources currently available, and how they can help predict the impact of genetic variation at the DNA, RNA and protein level.

Key words

Bioinformatics, genetic variation, *in silico* tools, online resources, sequence variation

Introduction

Clinical, diagnostic and research groups working in the field of hemostasis and thrombosis generate considerable data concerning genetic variation. Traditionally, this information has derived from targeted analysis of genes linked to a specific disease phenotype (e.g. investigating von Willebrand factor (*VWF*) in patients diagnosed with von Willebrand disease).[1] Additional data also derives from genome-wide association studies (GWAS) aimed at identifying genetic loci that may influence plasma protein levels[2,3] or that are associated with a specific phenotype, e.g. coronary artery disease.[4,5] The advent of next generation sequencing (NGS) has increased the amount of genetic information obtained from targeted analysis[6-8] and is also generating a wealth of information on genetic variation throughout the human genome.[9,10]

Although this information on genetic variation represents an invaluable resource, it is essential to properly interpret and understand the relevance of identified genetic variants within the human genome in order to determine whether they have a potential functional effect. Current guidelines from the American College of Medical Genetics and Genomics highlight that many lines of evidence are required to effectively classify genetic variants and assign pathogenicity,[11] one of which is information obtained from bioinformatics analyses. This review aims to provide an overview of the many free *in silico* tools and resources currently available online that can help clinicians / scientists predict the potential impact of genetic variants at the DNA, RNA and protein level, and therefore assist with variant classification.

Online resources for DNA level investigations

Descriptions for the majority of reported genetic variants would usually be at the DNA level using either genomic coordinates (e.g. chr12:g.6044368T>C) or a specific location within a genetic locus (e.g. *VWF*:c.2365A>G). Usually, the first stage in evaluating genetic variants is to investigate the literature and databases for existing knowledge.

Genome browsers and variant databases

Genome browsers (Ensembl, the National Center for Biotechnology Information (NCBI) Genome Data Viewer (GDV) and University of California Santa Cruz (UCSC) Genomics Institute (Table 1)) can be useful initial resources as they bring together extensive information on the human genome and other species. This includes information on known genetic variants, genotype-phenotype correlations, sequence conservation, transcription factor binding sites (TFBS) and expressed gene transcripts. These browsers also allow investigation of genetic variation at the precise nucleotide location or within the wider genomic context. While it may be difficult to identify relevant information on these browsers, especially for first-time users, useful tutorials on how to utilize Ensembl, GDV and UCSC are available online (Table 1).

Several online variant databases detail the population frequency of genetic variants (Table 1). These resources provide indications of variant pathogenicity because common variants in the general population are generally less likely to be disease causing. However, variant frequencies can differ between ethnicities and frequency data may derive from disease-specific populations, which may influence data interpretation. The Exome Aggregation Consortium (ExAC) database and the Genome Aggregation Database (gnomAD) act as repositories of exonic and/or genomic sequencing data aligned to the human GRCh37/hg19 genome assembly.[10] Data derive from a variety of large-scale sequencing projects (e.g. 1000 Genomes) and various disease-specific population studies (e.g. the Framingham Heart Study), and includes populations from varying ethnicities. Currently, there is data from 60,706 unrelated individuals in ExAC (including data on copy number variation [CNV]) and from 138,632 unrelated individuals (15,496 screened via whole genome sequencing) in gnomAD.

NCBI also has databases of annotated genetic variant information, including population frequencies where available, which link to the various genome browsers. Information about simple genetic variation, including single nucleotide variants (SNV) and small insertion / deletion (indel) variants, catalogued in the database of single nucleotide polymorphisms (dbSNP; Table 1), are given rs# identifiers. Large CNV (>50 bp in length), catalogued in the database of human genomic structural variation (dbVar; Table 1), are given nsv# identifiers. Similar to dbVar, the Database of

Genomic Variants (DGV; Table 1) also provides annotated information on large CNV >50 bp in length.

Another NCBI database, ClinVar (Table 1), links genetic variants with reported phenotypic information to provide an assessment of their clinical significance.[12] Data included are derived from clinical testing, research or extraction from the literature. Of particular use, each entry has a confidence score, which reflects the accuracy of the variant information and the evidence supporting clinical significance.

Similar to ClinVar, locus-specific databases (LSDBs) such as those for *VWF*[13] and coagulation factor IX (*F9*)[14] are highly useful clinical and scientific resources. LSDBs available through the Leiden Open Variation Database (LOVD) installation provide searchable lists of genetic variants and relevant phenotypic information where available (Table 1). However, many genes associated with hemostatic / thrombotic disorders currently have limited data available due to a lack of a dedicated curator(s) to help maintain and populate the relevant LOVD installation. A notable exception is the recent establishment of the European Association for Haemophilia and Allied Disorders Coagulation Factor Variant Databases (EAHAD-CFDB; Table 1). This initiative is a combined set of LSDBs (currently incorporating *F7*, *F8*, *F9* and *VWF*) using LOVD installations to provide genotype-phenotype correlations while also establishing enhanced databases for each factor focusing on nucleotide / amino acid sequence conservation and protein structure.[14,15]

Mutalyzer

Mutalyzer (Table 1) is an online suite of tools that at a basic level are designed to help ensure that genetic variants are described correctly according to current Human Genome Variation Society guidelines,[16,17] maintaining consistency in the reporting of variant descriptions. However, the tools also convert NCBI dbSNP identifiers (e.g. rs1063856) or genomic coordinates (e.g. chr12:g.6044368T>C; NC_000012.12:g.6044368T>C) to coding DNA nomenclature, which can be useful when working with variants identified via GWAS or NGS strategies.

The conversion of genomic coordinates in Mutalyzer also provides an indication as to whether a variant could affect various expressed gene transcripts. Genes (e.g. *F7*, *GP6* and *FLI1*) can have several transcripts that may vary in length, number of

exons and/or exon/intron boundaries. This can be particularly relevant when investigating genetic variants because a coding variant in one transcript may be non-coding in another transcript (Figure 1) and alternate transcripts can have different patterns of tissue expression.

Tools for assessing the potential impact of DNA variation

Computational alignments of nucleotide or amino acid sequences can provide an indication as to whether specific regions have functional importance because these regions are likely to demonstrate high evolutionary conservation. Several online tools are available that can produce multiple sequence alignments (Table 1) and both the GDV and UCSC browsers can create alignments of up to 100 vertebrate species. GDV, UCSC and the Exome Variant Server (Table 1) also provide measurement scores of evolutionary conservation utilizing either phylogenetic analysis with space/time models conservation (phastCons), phylogenetic *P*-values (phyloP), genomic evolutionary rate profiling (GERP) and/or GERP++ predictions. phastCons provides probability scores from 0 to 1 that each nucleotide belongs to a conserved element based on multiple alignments and the flanking nucleotide sequence, where a score closer to 1 indicates greater conservation.[18] phyloP assigns positive scores for conserved regions and negative scores for regions predicted to be evolving at a fast rate.[19] Both GERP and GERP++ provide maximum likelihood evolutionary rate estimation scores from -12.3 to 6.17, with positive scores representing conserved regions.[20,21]

Highly conserved regions may indicate the presence of important nucleotide motifs regulating transcription such as TFBS. Genetic variants occurring in these locations can influence gene expression (e.g. the well-characterized hemophilia B Leyden variants in *F9*[22] and c.-1522_-1510del variant in *VWF*).[23] Several online tools (ConTra v3, GenomeTraFac, GPMiner; Table 1) will screen inputted nucleotide sequence and/or specified genomic regions and predict potential regulatory features. In addition, the Ensembl browser provides data on regulatory regions derived from the Blueprint, ENCODE and Roadmap Epigenomics projects and indicates the activity level of regulatory features in specific cells / tissues.[24] Likewise, the UCSC browser also provides data derived from the ENCODE project[25] along with

information from Open Regulatory Annotation[26] and information on CpG islands (which can indicate potential transcription start sites[27]).

Online resources are also beginning to evaluate gene intolerance to provide additional evidence of variant pathogenicity. For example, a gene that has a comparatively high frequency of variants predicted to result in loss-of-function (LoF, e.g. nonsense or splicing mutations) is less likely to have disease-causing variants (i.e. LoF tolerant). ExAC provides a probability of being LoF intolerant (pLI) value for each gene,[10] dividing them into LoF intolerant (pLI ≥ 0.9) or LoF tolerant (pLI ≤ 0.1) categories. Similarly, the residual variation intolerance score (RVIS; Table 1) uses data derived from both ExAC and gnomAD to rank genes based on whether they have more or less common functional genetic variation relative to the genome-wide expectation.[28] A negative RVIS score and low percentile highlights a gene with fewer common functional mutations than expected (LoF intolerant) while a positive score and high percentile highlights a LoF tolerant gene.

Online resources for RNA level investigations

Analysis of genetic variation at a RNA level primarily concerns those tools applicable to predicting their effect on RNA splicing. However, genetic variants can influence RNA in other ways, so additional tools / resources can also be of use.

RNA splicing prediction tools

Genetic variants that occur within consensus motifs for 5' splice acceptors, 3' splice donors or intronic branch points can interfere with the interaction of the spliceosome complex, influencing the splicing of intronic sequence from the mature RNA causing full / partial exon skipping[29-32] or intron retention.[30] In addition, deep intronic variants can activate cryptic splice acceptors or donors causing intron retention[33] or the formation of a pseudo-exon.[34,35]

There are several *in silico* tools available to help predict the effect of variants on RNA splicing (Table 2), usually based on the comparison of inputted wild-type and variant DNA sequence via specific algorithms. Although several tools utilize their own custom prediction algorithms (i.e. GeneSplicer, Human Splicing Finder (HSF)

and SplicePort),[36-38] the majority use either maximum entropy modelling (MEM) or neural network algorithms.[39-42] Each algorithm will interpret inputted DNA sequence differently; therefore, it is important to obtain a consensus from several RNA splicing tools in order to generate the most accurate predictions.[43] However, even consensus predictions do not always signify a genuine effect on RNA splicing as has recently been observed for a c.5998+182A>G variant in *F8*.[33]

The influence of genetic variants on RNA may be commonly overlooked, except when variants occur within introns or exon/intron boundaries. Analyzing variants in coding regions using *in silico* RNA splicing prediction tools should however be standard practice. There are several examples where synonymous variants[31,44] and even coding variants predicted to influence the protein (e.g. resulting in a missense change[45]) disrupt splicing. Furthermore, in addition to the spliceosome interaction, serine-arginine repeat proteins and heterogeneous nuclear ribonucleoproteins act to promote and inhibit RNA splicing respectively.[46] These proteins interact with the RNA via exonic / intronic splice enhancer (ESE / ISE) and exonic / intronic splice silencer (ESS / ISS) motifs. Genetic variants creating or disrupting these motifs can influence splicing[47,48] and investigations including assessment of these motifs have begun in the field of hemostasis / thrombosis.[49,50]

Currently, there are few *in silico* tools available to investigate whether variants create or disrupt enhancer / silencer motifs (Table 2). Both ESEfinder and RESCUE-ESE are limited because as their names suggest they focus only on ESE motif predictions. However, both SFmap and HSF provide enhancer and silencer motif predictions. As with regular RNA splicing *in silico* tools, consensus predictions from several tools are likely to be the most accurate, but given the limited tools available this is difficult to achieve when investigating enhancer / silencer motifs.

As an initial tool to investigate the effect of genetic variants on RNA splicing, HSF is probably the most appropriate as it incorporates predictions for all motifs currently known to be involved in RNA splicing, including predictions from other sources (i.e. MEM algorithms, ESEfinder and RESCUE-ESE).[38] HSF also allows for multiple input options and provides its own consensus prediction, but the use of additional tools is still likely to improve overall accuracy.

Additional RNA prediction tools

Not all genetic variants will influence RNA splicing, but may still have an impact at the RNA level. Micro RNAs (miRNAs) play a role in regulating gene expression and studies have highlighted interactions with coagulation factors.[51,52] A useful (regularly updated) online resource for investigating whether genetic variants influence reported / potential miRNA binding targets or generate a potential miRNA binding target is miRBase (Table 2).

Genetic variants (e.g. c.2365A>G and c.2385T>C in *VWF*[53]) can also influence the secondary structure of transcribed mRNA, thereby impacting on the overall RNA stability, which in turn can influence RNA production.[54] Rtools provides a useful suite of prediction programs designed to compare inputted wild-type and variant DNA sequence and to highlight any differences in RNA secondary structure (Table 2).

The abundance of tRNA molecules available for a given amino acid codon sequence can affect the rate at which mature mRNA is translated into protein via a process called codon usage bias, and this in turn can be influenced by genetic variation. For example, a synonymous c.459G>A variant in *F9* reduces factor IX translation rate partly via an effect on codon usage (as the non-reference valine codon is less abundant; GTG = 28.1 vs. GTA = 7.1 codons present per 1000 codons).[55] Several online tools provide information on codon usage frequency or calculation of codon usage frequency either in the human genome or for a specific gene, e.g. Graphical Codon Usage Analyser (GCUA) and the Codon Usage Database (CUD; Table 2).

Online resources for protein level investigations

Analysis at the protein level utilizes those tools applicable to predicting the effect of non-synonymous amino acid variation and those resources that provide further information on the structure and function of proteins found to harbor potentially pathogenic variants.

Amino acid prediction tools

Non-synonymous amino acid substitutions can have profound effects on protein structure and function leading to disease. It is therefore useful to predict the impact of these changes on a protein in order to differentiate disease causing variants from

variants that have neutral effect.[56] Several studies have demonstrated that variants affecting protein function are more frequently found at positions conserved throughout evolution.[57] In addition, variants that affect protein stability are crucial for molecular function and are also more likely to be deleterious.[58,59] Based on these assumptions, multiple prediction tools have been developed that use sequence and/or structural information to predict the pathogenicity of a given variant (Table 3).

Two commonly used prediction algorithms include, sorting intolerant from tolerant (SIFT)[60] and polymorphism phenotyping v2 (PolyPhen-2).[61] SIFT is a popular prediction tool that utilizes sequence homology and the physical properties of amino acids to determine a variant's impact.[60] SIFT constructs a multiple sequence alignment (MSA) and then considers the composition of amino acids appearing at the site of the substitution. A SIFT score is then calculated, ranging from 0 to 1, reflecting the probability of the new amino acid being observed (tolerated) at that site. Scores ranging from 0 to 0.05 are considered to impact protein function. Other prediction tools incorporate the SIFT algorithm into their analysis pipelines, notably Mutation Predictor (MutPred) and nonsynonymous single nucleotide polymorphism analyzer (nsSNPAnalyzer; Table 3).

PolyPhen-2 uses both sequence and structural information to predict the effect of a given variant.[61] This is achieved by constructing a MSA, performing functional annotation of SNV, extracting protein sequence and structural information and building a conservation profile. Based on these properties PolyPhen-2 then estimates the probability that the missense mutation is 'probably damaging', 'possibly damaging' or 'benign'.[62]

It is important to remember that each tool and the algorithm it employs will provide varying levels of prediction accuracy. When compared to known deleterious variants, impact predictions of most tools were found to be accurate in ~60-80% of cases.[63] A recent study assessing the use of *in silico* tools to predict the pathogenicity of known deleterious variants in antithrombin found that performance varied depending on the localization of the substitution within the secondary structure, with those in α-helices often misclassified as benign.[64] In addition, variants known to disrupt posttranslational modifications were also misclassified.[64] As with RNA splicing predictions, it is therefore useful to utilize several prediction tools to achieve an accurate consensus (e.g. hemostasis / thrombosis studies investigating variants in

αIIbβ3, ADAMTS13, FVIII and VWF used MutationTaster, PolyPhen-2, PROVEAN and SIFT).[7,65-68]

Tools for assessing protein stability

The effect of an amino acid substitution on the protein stability and function is an important consideration when trying to determine pathogenicity. Using a protein databank (PDB) file and a specified variant, tools such as Site Directed Mutator (SDM; Table 3) can calculate a stability difference score between the wild-type and the variant protein.[69] Where the tertiary structure of a protein of interest is unknown and no PDB structure file exists, machine learning programs such as MUpro (Table 3) predict protein stability changes using primary sequence data alone.[70] However, while these tools may be useful in a research context, providing an extra line of evidence, they do not make any predictions about whether a substitution is damaging or deleterious.

Other useful protein tools and resources

There are several resources that can be utilized in the analysis of proteins (e.g. to identify protein domains / motifs or to investigate protein-protein interactions). The Swiss Institute for Bioinformatics ExPASy resource (Table 3) contains a comprehensive list of protein analysis tools along with useful summary descriptions.[71] PDB provides 3D protein models that when imported into specialized molecular graphics programs such as Jmol and PyMOL (Table 3) allows visualization of a variant at the molecular level (e.g. to elucidate the impact of a novel deletion in αIIbβ3[72]) or simulate molecular interactions (e.g. to interpret variation in the DNA-binding domain of FLI1[73]). This may be particularly useful to resolve instances of variant misclassification by amino acid prediction tools. For many proteins of interest, the 3D structure is currently unknown, so no PDB structure entry exists. In these instances computational homology based modelling servers such as SWISS-MODEL provide a useful alternative.[74]

Variants causing amino acid substitutions in the signal peptide (SP) region of a protein may cause disruption or loss of function due to defective localization of the protein and/or defective SP cleavage. For secretory proteins therefore, it is important

to consider the effect of substitutions that occur in the SP region. The SignalP 4.1 server (Table 3) predicts the presence and location of SP cleavage sites in an amino acid sequence and can predict the effect of substitutions or deletions/duplications on SP cleavage.[75]

Additional tools / resources for variant analysis

While most *in silico* tools focus on specific predictions at the DNA, RNA or protein level, Combined Annotation Dependent Depletion (CADD) and the Ensembl browser Variant Effect Predictor (VEP; Table 4) both use multiple lines of evidence to provide an assessment of variant pathogenicity. CADD integrates multiple and diverse annotations to produce a single measure of deleteriousness (C-score) for a particular SNV; a C-score of ≥10 indicates a variant is in the top 10% of the most deleterious, a score of ≥20 in the top 1%, etc.[76] VEP provides a detailed annotation for variant effects on transcripts, proteins and regulatory regions, but is also a flexible and customizable software suite, allowing the addition of tools such as CADD into the analysis pipeline.[77] However, while CADD and VEP can complement other predictions to provide further consensus, they are not stand-alone tools.

Studies involving whole genome sequencing, whole exome sequencing (WES) or transcriptome profiling (using RNA sequencing or expression array approaches), generate large gene / protein lists. It is desirable to be able to make sense of these lists and extract the biological information they contain. The Database for Annotation, Visualization and Integrated Discovery (DAVID; Table 4) is a high-throughput and integrated data mining tool able to map genes / proteins to a biological annotation and then highlight statistically over-represented or enriched annotations.[78] This can enable clustering of gene / protein lists to a range of criteria including diseases, functional categories, gene ontology terms, pathways, protein domains, protein interactions and tissue expression.

Finally, the use of protein abundance information from different tissue types may be helpful when prioritizing candidate genes, e.g. to identify proteins present in the platelet proteome following WES of patients with inherited platelet function disorders. The Protein Abundance Database (PaxDb; Table 4) is a useful meta-resource of

protein abundance data for model organisms, tissues and cell-lines that enables a quick check of a protein of interest, aiding gene prioritization.[79]

Concluding remarks

In silico tools and online resources serve as useful sources of information for clinicians / scientists investigating genetic variation. However, this information is only a prediction and not a definitive answer; it will provide evidence to link a variant to disease pathogenicity or help confirm / direct further investigations, e.g. *in vitro* and *in vivo* studies. For the most accurate and informative analyses of a variant(s) users should consider its effect at the DNA, RNA and protein level (Figure 2) utilizing all the tools / resources highlighted in this review as a bioinformatics toolkit (Figure 3).

Acknowledgements

The British Heart Foundation (PG/15/61/31634) supports S.J.W. M.A.A would like to acknowledge the support of Imam Abdulrahman bin Faisal University, Dammam, Saudi Arabia.

References

1. Hampshire DJ, Abuzenadah AM, Cartwright A, et al. Identification and characterisation of mutations associated with von Willebrand disease in a Turkish patient cohort. *Thromb Haemost*. 2013;110(2):264-274.

2. Smith NL, Chen M-H, Dehghan A, et al. Novel associations of multiple genetic loci with plasma levels of factor VII, factor VIII, and von Willebrand factor: The CHARGE (Cohorts for Heart and Aging Research in Genome Epidemiology) Consortium. *Circulation*. 2010;121(12):1382-1392.

3. de Vries PS, Chasman DI, Sabater-Lleal M, et al. A meta-analysis of 120 246 individuals identifies 18 new loci for fibrinogen concentration. *Hum Mol Genet*. 2016;25(2):358-370.

4. Nelson CP, Goel A, Butterworth AS, et al. Association analyses based on false discovery rate implicate new loci for coronary artery disease. *Nat Genet*. 2017;49(9):1385-1391.

5. Verweij N, Eppinga RN, Hagemeijer Y, van der Harst P. Identification of 15 novel risk loci for coronary artery disease and genetic risk of recurrent events, atrial fibrillation and heart failure. *Sci Rep*. 2017;7(1):2761.

6. Leo VC, Morgan NV, Bem D, et al. Use of next-generation sequencing and candidate gene analysis to identify underlying defects in patients with inherited platelet function disorders. *J Thromb Haemost*. 2015;13(4):643-650.

7. Borràs N, Batlle J, Pérez-Rodríguez A, et al. Molecular and clinical profile of von Willebrand disease in Spain (PCM-EVW-ES): comprehensive genetic analysis by next-generation sequencing of 480 patients. *Haematologica*. 2017;102(12):2005-2014.

8. Johnsen JM, Fletcher SN, Huston H, et al. Novel approach to genetic analysis and results in 3000 hemophilia patients enrolled in the My Life, Our Future initiative. *Blood Adv*. 2017;1(13):824-834.

9. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74.

10. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016;536(7616):285-291.

11. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*. 2015;17(5):405-424.

12. Landrum MJ, Lee JM, Benson M, et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res*. 2018;46(Database issue):D1062-D1067.

13. Hampshire DJ, Goodeve AC. The International Society on Thrombosis and Haematosis von Willebrand disease database: an update. *Semin Thromb Hemost*. 2011;37(5):470-479.

14. Rallapalli PM, Kemball-Cook G, Tuddenham EG, Gomez K, Perkins SJ. An interactive mutation database for human coagulation factor IX provides novel insights into the phenotypes and genetics of hemophilia B. *J Thromb Haemost*. 2013;11(7):1329-1340.

15. Hampshire DJ, Cairo A, Dolan G, et al. EAHAD-DB: a combined coagulation factor variant databases resource for the clinical and scientific communities. *J Thromb Haemost*. 2015;13(Suppl. 2):abst. PO676-WED.

16. Wildeman M, van Ophuizen E, den Dunnen JT, Taschner PEM. Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. *Hum Mutat*. 2008;29(1):6-13.

17. den Dunnen JT, Dalgleish R, Maglott DR, et al. HGVS recommendations for the description of sequence variants: 2016 update. *Hum Mutat*. 2016;37(6):564-569.

18. Siepel A, Bejerano G, Pedersen JS, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*. 2005;15(8):1034-1050.

19. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res*. 2010;20(1):110-121.

20. Cooper GM, Stone EA, Asimenos G, et al. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res*. 2005;15(7):901-913.

21. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol*. 2010;6(12):e1001025.

22. Funnell APW, Crossley M. Hemophilia B Leyden and once mysterious *cis*-regulatory mutations. *Trends Genet*. 2014;30(1):18-23.

23. Othman M, Chirinian Y, Brown C, et al. Functional characterization of a 13-bp deletion (c.-1522_-1510del13) in the promoter of the von Willebrand factor gene in type 1 von Willebrand disease. *Blood*. 2010;116(18):3645-3652.

24. Zerbino DR, Achuthan P, Akanni W, et al. Ensembl 2018. *Nucleic Acids Res*. 2018;46(Database issue):D754-D761.

25. Rosenbloom KR, Sloan CA, Malladi VS, et al. ENCODE data in the UCSC Genome Browser: year 5 update. *Nucleic Acids Res*. 2012;41(Database issue):D56-D63.

26. Griffith OL, Montgomery SB, Bernier B, et al. ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic Acids Res*. 2008;36(Database issue):D107-D113.

27. Gardiner-Garden M, Frommer M. CpG islands in vertebrate genomes. *J Mol Biol*. 1987;196(2):261-282.

28. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet*. 2013;9(8):e1003709.

29. Corrales I, Ramírez L, Altisent C, Parra R, Vidal F. The study of the effect of splicing mutations in von Willebrand factor using RNA isolated from patients' platelets and leukocytes. *J Thromb Haemost*. 2011;9(4):679-688.

30. Martorell L, Corrales I, Ramirez L, et al. Molecular characterization of ten *F8* splicing mutations in RNA isolated from patient's leucocytes: assessment of *in silico* prediction tools accuracy. *Haemophilia*. 2015;21(2):249-257.

31. Nuzzo F, Bulato C, Nielsen BI, et al. Characterization of an apparently synonymous *F5* mutation causing aberrant splicing and factor V deficiency. *Haemophilia*. 2015;21(2):241-248.

32. Hawke L, Bowman ML, Poon M-C, Scully M-F, Rivard G-E, James PD. Characterization of aberrant splicing of von Willebrand factor in von Willebrand disease: an underrecognized mechanism. *Blood*. 2016;128(4):584-593.

33. Bach JE, Müller CR, Rost S. Mini-gene assays confirm the splicing effect of deep intronic variants in the factor VIII gene. *Thromb Haemost*. 2016;115(1):222-224.

34. Castaman G, Giacomelli SH, Mancuso ME, et al. Deep intronic variations may cause mild hemophilia A. *J Thromb Haemost*. 2011;9(8):1541-1548.

35. Castoldi E, Duckers C, Radu C, et al. Homozygous *F5* deep-intronic splicing mutation resulting in severe factor V deficiency and undetectable thrombin generation in platelet-rich plasma. *J Thromb Haemost*. 2011;9(5):959-968.

36. Pertea M, Lin X, Salzberg SL. GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Res*. 2001;29(5):1185-1190.

37. Dogan RI, Getoor L, Wilbur WJ, Mount SM. SplicePort--an interactive splice-site analysis tool. *Nucleic Acids Res*. 2007;35(Web Server issue):W285-W291.

38. Desmet F-O, Hamroun D, Lalande M, Collod-Béroud G, Claustres M, Béroud C. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res*. 2009;37(9):e67.

39. Brunak S, Engelbrecht J, Knudsen S. Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J Mol Biol*. 1991;220(1):49-65.

40. Reese MG, Eeckman FH, Kulp D, Haussler D. Improved splice site detection in Genie. *J Comput Biol*. 1997;4(3):311-323.

41. Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol*. 2004;11(2-3):377-394.

42. Wang M, Marín A. Characterization and prediction of alternative splice sites. *Gene*. 2006;366(2):219-227.

43. Houdayer C, Dehainault C, Mattler C, et al. Evaluation of in silico splice tools for decision-making in molecular diagnosis. *Hum Mutat*. 2008;29(7):975-982.

44. Daidone V, Gallinaro L, Grazia Cattini M, et al. An apparently silent nucleotide substitution (c.7056C>T) in the von Willebrand factor gene is responsible for type 1 von Willebrand disease. *Haematologica*. 2011;96(6):881-887.

45. James PD, O'Brien LA, Hegadorn CA, et al. A novel type 2A von Willebrand factor mutation located at the last nucleotide of exon 26 (3538G>A) causes skipping of 2 nonadjacent exons. *Blood*. 2004;104(9):2739-2745.

46. Lee Y, Rio DC. Mechanisms and regulation of alternative pre-mRNA splicing. *Annu Rev Biochem*. 2015;84(1):291-323.

47. Otsuka H, Sasai H, Nakama M, et al. Exon 10 skipping in *ACAT1* caused by a novel c.949G>A mutation located at an exonic splice enhancer site. *Mol Med Rep*. 2016;14(5):4906-4910.

48. Palhais B, Dembic M, Sabaratnam R, et al. The prevalent deep intronic c.639+919G>A *GLA* mutation causes pseudoexon activation and Fabry disease by

abolishing the binding of hnRNP A1 and hnRNP A2/B1 to a splicing silencer. *Mol Genet Metab*. 2016;119(3):258-269.

49. Balestra D, Barbon E, Scalet D, et al. Regulation of a strong *F9* cryptic 5′ss by intrinsic elements and by combination of tailored U1snRNAs with antisense oligonucleotides. *Hum Mol Genet*. 2015;24(17):4809-4816.

50. Mufti AH, Alyami NH, Peake IR, Goodeve AC, Hampshire DJ. Silent von Willebrand factor variant c.4146G>T (p.Leu1382=) causes type 1 von Willebrand disease via disruption of an exonic splice enhancer motif. *Res Pract Thromb Haemost*. 2017;1(Suppl. 1):abst. OC 22.25.

51. Fort A, Borel C, Migliavacca E, Antonarakis SE, Fish RJ, Neerman-Arbez M. Regulation of fibrinogen production by microRNAs. *Blood*. 2010;116(14):2608-2615.

52. Vossen CY, van Hylckama Vlieg A, Teruel-Montoya R, et al. Identification of coagulation gene 3'UTR variants that are potentially regulated by microRNAs. *Br J Haematol*. 2017;177(5):782-790.

53. Mufti AH, Ogiwara K, Swystun LL, et al. The common *VWF* single nucleotide variants c.2365A>G and c.2385T>C modify VWF biosynthesis and clearance. *Blood Adv*. 2018;2(13):1585-1594.

54. Manning KS, Cooper TA. The roles of RNA processing in translating genotype to phenotype. *Nat Rev Mol Cell Biol*. 2017;18(2):102-114.

55. Simhadri VL, Hamasaki-Katagiri N, Lin BC, et al. Single synonymous mutation in factor IX alters protein properties and underlies haemophilia B. *J Med Genet*. 2017;54(5):338-345.

56. Ng PC, Henikoff S. Predicting the effects of amino acid substitutions on protein function. *Annu Rev Genomics Hum Genet*. 2006;7(1):61-80.

57. Miller MP, Kumar S. Understanding human disease mutations through the use of interspecific genetic variation. *Hum Mol Genet*. 2001;10(21):2319-2328.

58. Sunyaev S, Ramensky V, Bork P. Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends Genet*. 2000;16(5):198-200.

59. Wang Z, Moult J. SNPs, protein structure, and disease. *Hum Mutat*. 2001;17(4):263-270.

60. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res*. 2003;31(13):3812-3814.

61. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nat Methods*. 2010;7(4):248-249.

62. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet*. 2013;76(1):7.20.21-27.20.41.

63. Thusberg J, Olatubosun A, Vihinen M. Performance of mutation pathogenicity prediction methods on missense variants. *Hum Mutat*. 2011;32(4):358-368.

64. Luxembourg B, D`Souza M, Körber S, Seifried E. Prediction of the pathogenicity of antithrombin sequence variations by in silico methods. *Thromb Res*. 2015;135(2):404-409.

65. Nurden AT, Pillois X, Fiore M, et al. Expanding the mutation spectrum affecting αIIbβ3 integrin in Glanzmann Thrombasthenia: screening of the *ITGA2B* and *ITGB3* genes in a large international cohort. *Hum Mutat*. 2015;36(5):548-561.

66. Sengupta M, Sarkar D, Ganguly K, Sengupta D, Bhaskar S, Ray K. *In silico* analyses of missense mutations in coagulation factor VIII: identification of severity determinants of haemophilia A. *Haemophilia*. 2015;21(5):662-669.

67. Stoll M, Rühle F, Witten A, et al. Rare variants in the ADAMTS13 von Willebrand factor-binding domain contribute to pediatric stroke. *Circ Cardiovasc Genet*. 2016;9(4):357-367.

68. Pillois X, Peters P, Segers K, Nurden AT. In silico analysis of structural modifications in and around the integrin αIIb genu caused by *ITGA2B* variants in human platelets with emphasis on Glanzmann thrombasthenia. *Mol Genet Genomic Med*. 2018;6(2):249-260.

69. Pandurangan AP, Ochoa-Montaño B, Ascher DB, Blundell TL. SDM: a server for predicting effects of mutations on protein stability. *Nucleic Acids Res*. 2017;45(Web Server issue):W229-W235.

70. Cheng J, Randall A, Baldi P. Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins*. 2006;62(4):1125-1132.

71. Artimo P, Jonnalagedda M, Arnold K, et al. ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Res*. 2012;40(Web Server issue):W597-W603.

72. Nurden P, Bordet J-C, Pillois X, Nurden AT. An intracytoplasmic β3 Leu718 deletion in a patient with a novel platelet phenotype. *Blood Adv*. 2017;1(8):494-499.

73. Saultier P, Vidal L, Canault M, et al. Macrothrombocytopenia and dense granule deficiency associated with FLI1 variants: ultrastructural and pathogenic features. *Haematologica*. 2017;102(6):1006-1016.

74. Waterhouse A, Bertoni M, Bienert S, et al. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res* 2018;46(Web Server issue):W296-W303.

75. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods*. 2011;8(10):785-786.

76. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*. 2014;46(3):310-315.

77. McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. *Genome Biol*. 2016;17(1):122.

78. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. 2009;4(1):44-57.

79. Wang M, Herrmann CJ, Simonovic M, Szklarczyk D, von Mering C. Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics*. 2015;15(18):3163-3168.

80. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 2018;46(Database issue):D8-D13.

81. Casper J, Zweig AS, Villarreal C, et al. The UCSC Genome Browser database: 2018 update. *Nucleic Acids Res*. 2018;46(Database issue):D762-D769.

82. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 2016;44(Database issue):D7-D19.

83. MacDonald JR, Ziman R, Yuen RKC, Feuk L, Scherer SW. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res*. 2014;42(Database issue):D986-D992.

84. Sievers F, Wilm A, Dineen D, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol*. 2011;7(1):539.

85. Corpet F. Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res*. 1988;16(22):10881-10890.

86. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004;32(5):1792-1797.

87. Kreft Ł, Soete A, Hulpiau P, Botzki A, Saeys Y, De Bleser P. ConTra v3: a tool to identify transcription factor binding sites across species, update 2017. *Nucleic Acids Res*. 2017;45(Web Server issue):W490-W494.

88. Jegga AG, Chen J, Gowrisankar S, et al. GenomeTrafac: a whole genome resource for the detection of transcription factor binding site clusters associated with conventional and microRNA encoding genes conserved between mouse and human gene orthologs. *Nucleic Acids Res*. 2007;35(Database issue):D116-D121.

89. Lee T-Y, Chang W-C, Hsu JB-K, Chang T-H, Shien D-M. GPMiner: an integrated system for mining combinatorial *cis*-regulatory elements in mammalian gene group. *BMC Genomics*. 2012;13(Suppl. 1):S3.

90. Stothard P. The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *Biotechniques*. 2000;28(6):1102-1104.

91. Cartegni L, Wang J, Zhu Z, Zhang MQ, Krainer AR. ESEfinder: a web resource to identify exonic splicing enhancers. *Nucleic Acids Res*. 2003;31(13):3568-3571.

92. Fairbrother WG, Yeh R-F, Sharp PA, Burge CB. Predictive identification of exonic splicing enhancers in human genes. *Science*. 2002;297(5583):1007-1013.

93. Paz I, Akerman M, Dror I, Kosti I, Mandel-Gutfreund Y. SFmap: a web server for motif analysis and prediction of splicing factor binding sites. *Nucleic Acids Res*. 2010;38(Web Server issue):W281-W285.

94. Nakamura Y, Gojobori T, Ikemura T. Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Res*. 2000;28(1):292.

95. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. miRBase: tools for microRNA genomics. *Nucleic Acids Res*. 2008;36(Database issue):D154-D158.

96. Kozomara A, Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res*. 2011;39(Database issue):D152-D157.

97. Hamada M, Ono Y, Kiryu H, et al. Rtools: a web server for various secondary structural analyses on single RNA sequences. *Nucleic Acids Res*. 2016;44(Web Server issue):W302-W307.

98. Tavtigian SV, Deffenbaugh AM, Yin L, et al. Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. *J Med Genet*. 2006;43(4):295-305.

99. Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods*. 2014;11(4):361-362.

100. Li B, Krishnan VG, Mort ME, et al. Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics*. 2009;25(21):2744-2750.

101. Niroula A, Urolagin S, Vihinen M. PON-P2: prediction method for fast and reliable identification of harmful variants. *PLoS One*. 2015;10(2):e0117380.

102. Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics*. 2015;31(16):2745-2747.

103. Sim N-L, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res*. 2012;40(Web Server issue):W452-W457.

104. Berman H, Henrick K, Nakamura H, Markley JL. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res*. 2007;35(Database issue):D301-D303.

Tables

Table 1. Tools / resources for investigating genetic variation at the DNA level

Tool / resource	Web address
Genome browsers with multiple functionality	
Ensembl[24]	http://www.ensembl.org/index.html
(online user guide: https://www.ensembl.org/info/website/tutorials/index.html)	
GDV[80]	https://www.ncbi.nlm.nih.gov/genome/gdv/
(online user guide: https://www.ncbi.nlm.nih.gov/genome/gdv/browser/help/)	
UCSC[81]	https://genome.ucsc.edu/cgi-bin/hgGateway
(online user guide: https://genome.ucsc.edu/training/)	
Annotated genetic variation and population frequency databases	
1000 Genomes[9]	http://www.internationalgenome.org/
dbSNP[82]	https://www.ncbi.nlm.nih.gov/snp
dbVAR[82]	https://www.ncbi.nlm.nih.gov/dbvar
DGV[83]	http://dgv.tcag.ca/dgv/app/home
ExAC[10]	http://exac.broadinstitute.org/
Exome Variant Server	http://evs.gs.washington.edu/EVS/
gnomAD[10]	http://gnomad.broadinstitute.org/
Databases providing genotype-phenotype correlations	
ClinVar[12]	https://www.ncbi.nlm.nih.gov/clinvar/
EAHAD-CFDB[15]	http://www.eahad-db.org/
LOVD	https://databases.lovd.nl/shared/genes
Sequence alignment tools[a]	
Clustal Omega[84]	https://www.ebi.ac.uk/Tools/msa/clustalo/
MultAlin[85]	http://multalin.toulouse.inra.fr/multalin/
MUSCLE[86]	https://www.ebi.ac.uk/Tools/msa/muscle/

Regulatory motif prediction tools	
ConTra v3[87]	http://bioit2.irc.ugent.be/contra/v3/#/step/1
GenomeTraFaC[88]	https://genometrafac.cchmc.org/genome-trafac/index.jsp
GPMiner[89]	http://gpminer.mbc.nctu.edu.tw/index.php
Other useful tools / resources	
Mutalyzer[16]	https://mutalyzer.nl/
RVIS[28]	http://genic-intolerance.org/
Sequence Manipulation Suite[a,90]	http://www.bioinformatics.org/sms2/

[a]Tools that analyze both nucleotide and amino acid sequences.

Table 2. Tools / resources for investigating genetic variation at the RNA level

Tool / resource	Web address
RNA splicing prediction tools	
ASSP[42]	http://wangcomputing.com/assp/index.html
BDGP[40]	http://www.fruitfly.org/seq_tools/splice.html
ESEfinder[91]	http://krainer01.cshl.edu/cgi-bin/tools/ESE3/esefinder.cgi?process=home
GeneSplicer[36]	http://www.cs.jhu.edu/~genomics/GeneSplicer/gene_spl.html
HSF[38]	http://www.umd.be/HSF3/
MaxEntScan[41]	http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq_acc.html
	http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html
NetGene2[39]	http://www.cbs.dtu.dk/services/NetGene2/
RESCUE-ESE[92]	http://genes.mit.edu/burgelab/rescue-ese/
SFmap[93]	http://sfmap.technion.ac.il/
SplicePort[37]	http://spliceport.cbcb.umd.edu/
Other useful tools / resources	
CUD[94]	https://www.kazusa.or.jp/codon/
GCUA	http://gcua.schoedl.de/index.html
miRBase[95,96]	http://www.mirbase.org/
Rtools[97]	http://rtools.cbrc.jp/

Table 3. Tools / resources for investigating genetic variation at the protein level

Tool / resource	Web address
Amino acid prediction tools	
Align-GVGD[98]	http://agvgd.hci.utah.edu/
MutationTaster[99]	http://www.mutationtaster.org/
MutPred[100]	http://mutpred1.mutdb.org/
nsSNPAnalyzer	http://snpanalyzer.uthsc.edu/
PolyPhen-2[61]	http://genetics.bwh.harvard.edu/pph2/
PON-P2[101]	http://structure.bmc.lu.se/PON-P2/
PROVEAN[102]	http://provean.jcvi.org/index.php
SIFT[103]	http://sift.bii.a-star.edu.sg/
Tools for assessing protein stability	
MUpro[70]	http://mupro.proteomics.ics.uci.edu/
SDM[69]	http://marid.bioc.cam.ac.uk/sdm2
Other useful tools / resources	
ExPASy[71]	https://www.expasy.org/proteomics
Jmol	http://jmol.sourceforge.net/
PyMOL	https://pymol.org/2/
PDB[104]	https://www.wwpdb.org/
SignalP 4.1[75]	http://www.cbs.dtu.dk/services/SignalP/
SWISS-MODEL[74]	https://swissmodel.expasy.org

Table 4. Other useful tools / resources

Tool / resource	Web address
CADD[76]	https://cadd.gs.washington.edu/
DAVID[78]	https://david.ncifcrf.gov/home.jsp
PaxDb[79]	https://pax-db.org/
VEP[77]	http://www.ensembl.org/Tools/VEP

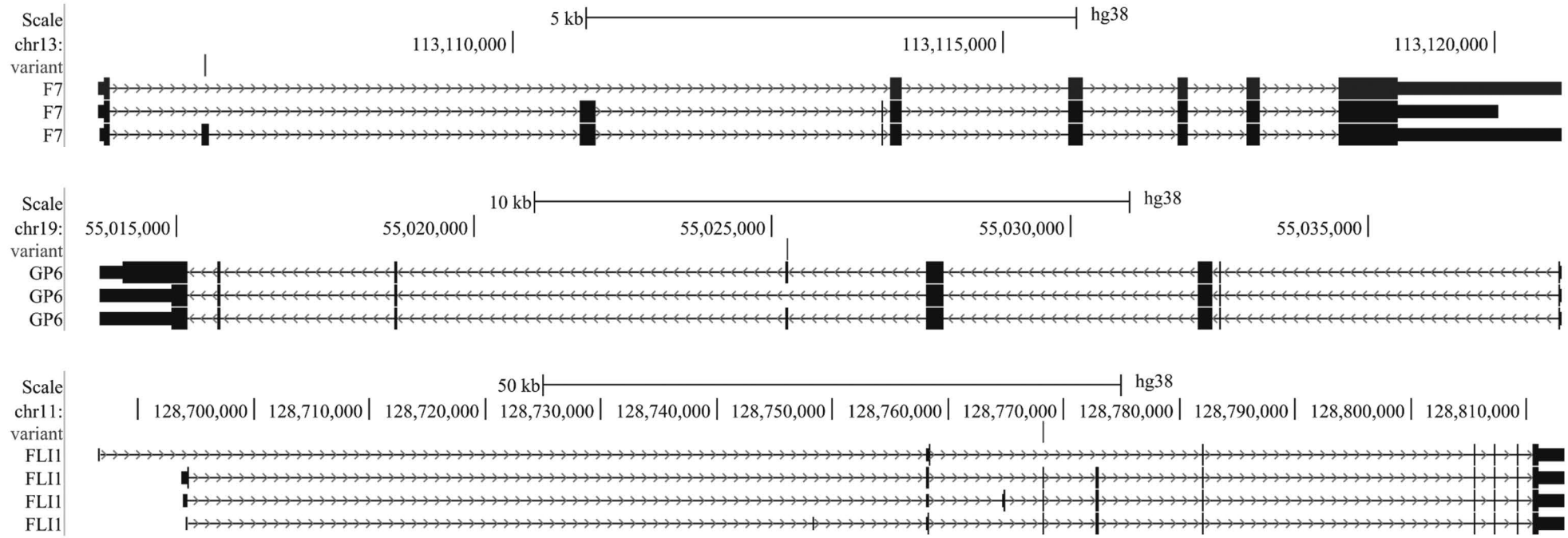
Figure legends

Figure 1. The location of genetic variants can vary depending on the gene transcript. A) Expressed gene transcripts for *F7*, *GP6* and *FLI1*. Shaded boxes and vertical lines represent exonic sequence. B) Examples of genetic variants in *F7*, *GP6* and *FLI1* are reported using genomic coordinates, and their corresponding location in each gene transcript.

Figure 2. Example bioinformatics analyses for three sequence variants in *VWF* (A: c.55G>A, p.(Gly19Arg); B: c.2365A>G, p.(Thr789Ala); C: c.3614G>A, p.(Arg1205His)). Different analyses at the DNA (green), RNA (yellow) and protein (blue) level can each provide useful information concerning a sequence variant, and analyses at one level (e.g. DNA) may prompt additional analyses at the other two levels (e.g. RNA and protein). Tools such as CADD and VEP provide information relating to all three levels of analysis. [a]Consensus utilizing MutationTaster, PolyPhen-2, PROVEAN and SIFT. [b]Consensus utilizing ASSP, BDGP, HSF and NetGene2. [c]Consensus utilizing HSF, RESCUE-ESE and SFmap. ESS, exonic splice silencer; SE, splice enhancer; SS, splice silencer; WT, wild-type.

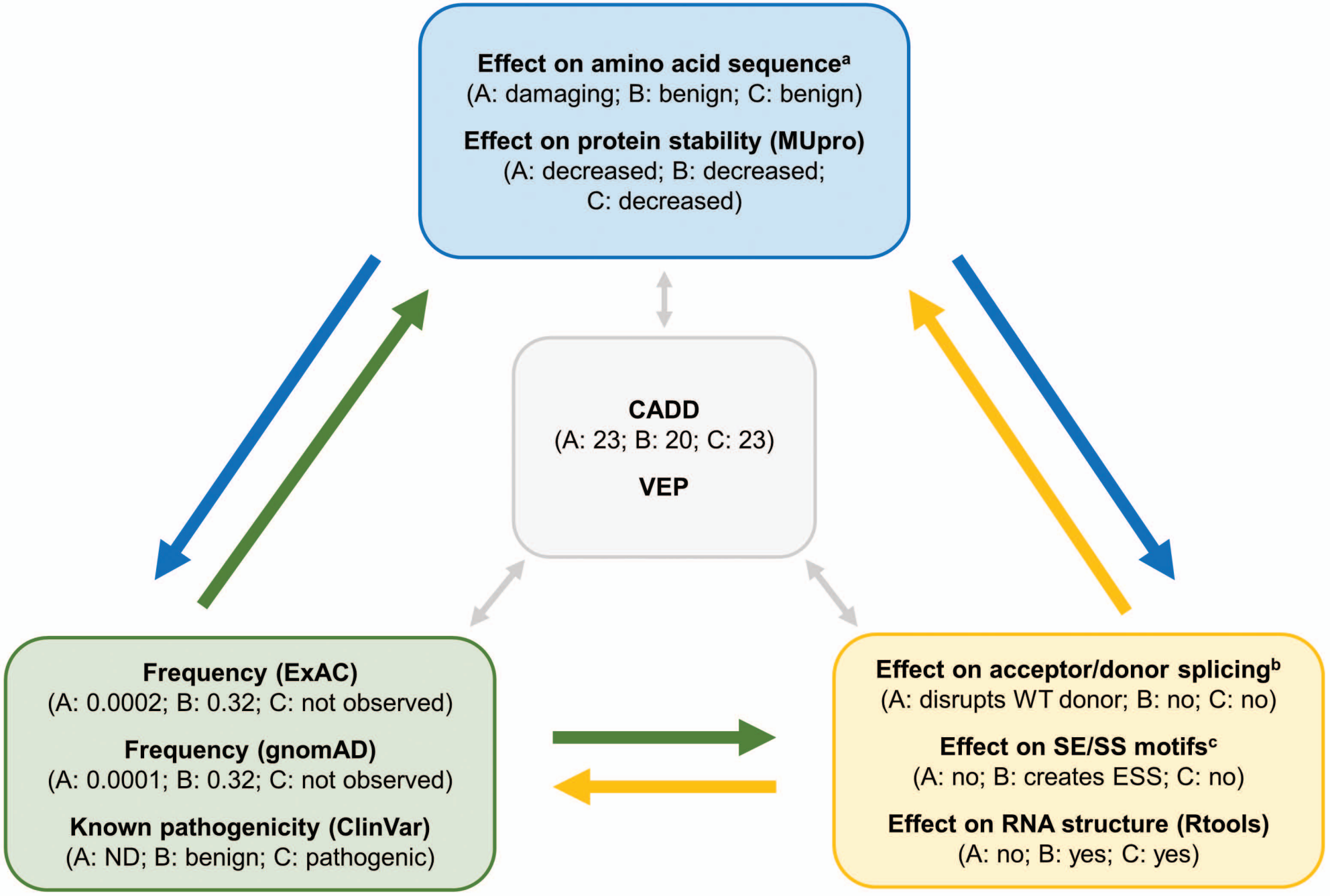
Figure 3. A bioinformatics toolkit quick reference guide. Suggested analyses at the DNA (green), RNA (yellow) and protein (blue) level are highlighted, along with the tools / resources that could be used.

A



B

Gene	Genetic variant	Variant nomenclature for each transcript
F7	chr13:g.113106866G>T	NM_000131:c.86G>T; NM_001267554:c.64+961G>T; NM_019616:c.64+961G>T
GP6	chr19:g.55025244G>T	NM_001083899:c.638C>A; NM_001256017:c.610+2334C>A; NM_016363:c.638C>A
FLI1	chr11:g.128768187G>T	NM_001167681:c.201G>T; NM_001271010:c.102G>T; NM_001271012:c.10+9861G>T; NM_002017:c.300G>T


Effect on amino acid sequence[a]
(A: damaging; B: benign; C: benign)
Effect on protein stability (MUpro)
(A: decreased; B: decreased;
C: decreased)
CADD
(A: 23; B: 20; C: 23)
VEP
Frequency (ExAC)
(A: 0.0002; B: 0.32; C: not observed)
Frequency (gnomAD)
(A: 0.0001; B: 0.32; C: not observed)
Known pathogenicity (ClinVar)
(A: ND; B: benign; C: pathogenic)
Effect on acceptor/donor splicing[b]
(A: disrupts WT donor; B: no; C: no)
Effect on SE/SS motifs[c]
(A: no; B: creates ESS; C: no)
Effect on RNA structure (Rtools)
(A: no; B: yes; C: yes)

Evaluation of variants using *in silico* tools and online resources

Prior to assessing pathogenicity using DNA, RNA and protein level tools, initial filtering of large-scale variant data sets may be based on:

- Frequency e.g. ExAC, gnomAD
- Protein abundance database e.g. PaxDb
- Gene intolerance to loss-of-function (LoF) e.g. ExAC pLI score
- Variant pathogenicity using multiple lines of evidence e.g. CADD, VEP
- Functional enrichment e.g. DAVID

DNA

- Mine the literature and databases for existing knowledge and variant frequencies e.g. Ensembl, GDV, LSDBs, NCBI, UCSC
- Evaluate if the variant has an effect on an alternative transcripts e.g. Mutalyzer
- Assess the evolutionary conservation of the affected nucleotide e.g. GDV, UCSC
- Predict if the variant is located within a regulatory feature e.g. ConTra v3, Ensembl, GenomeTraFaC, GPMiner, UCSC

RNA

- Predict the effect of the variant on RNA splicing e.g. ESEfinder, GeneSplicer, HSF, SplicePort, RESCUE-ESE, SFmap
- Predict if the variant creates/disrupts miRNA binding site e.g. miRBase
- Evaluate if the variant has an effect on RNA secondary structure e.g. Rtools
- Evaluate the codon usage frequency e.g. CUD, GCUA

Protein

- Predict the amino acid evolutionary conservation, sequence and structural attributes e.g. CADD, PolyPhen-2, SIFT, VEP
- Assess the variant effect on protein stability e.g. MUpro, SDM
- Visualize variants and predict changes in protein interactions using 3D models e.g. Jmol, PyMOL
- Evaluate any disruption to the localization of the protein e.g. SignalP 4.1