OXFORD

Briefings in Functional Genomics, 15(5), 2016, 374–384

doi: 10.1093/bfgp/elv054
Advance Access Publication Date: 9 December 2015
Review paper

Exome sequencing explained: a practical guide to its clinical application

Eleanor G. Seaby, Reuben J. Pengelly, and Sarah Ennis

Corresponding author: Sarah Ennis, Human Genetics & Genomic Medicine, Faculty of Medicine, University of Southampton, Duthie Building (MP 808), Tremona Road, Southampton SO16 6YD, UK. Tel.: +44(0)2381 208614; E-mail: s.ennis@soton.ac.uk

Abstract

Next-generation sequencing has catapulted healthcare into a revolutionary genomics era. One such technology, whole-exome sequencing, which targets the protein-coding regions of the genome, has proven success in identifying new causal mutations for diseases of previously unknown etiology. With a successful diagnostic rate approaching 25% for rare disease in recent studies, its clinical utility is becoming increasingly popular. However, the interpretation of whole-exome sequencing data requires expertise in genomic informatics and clinical medicine to ensure the accurate and safe reporting of findings back to the bedside. This is challenged by vast amounts of sequencing data harbouring approximately 25 000 variants per sequenced individual. Computational strategies and fastidious filtering frameworks are thus required to extricate candidate variants in a sea of common polymorphisms. Once prioritized, identified variants require intensive scrutiny at a biological level, and require judicious assessment alongside the clinical phenotype. In the final step, all evidence is collated and documented alongside pathogenicity guidelines to produce an exome report that returns to the clinic. This review provides a practical guide for clinicians and genomic informaticians on the clinical application of whole-exome sequencing. We address sequencing capture and methodology, quality control parameters at different stages of sequencing analysis and propose an exome data filtering strategy that includes primary filtering (for the removal of probable benign variants) and secondary filtering for the prioritization of remaining candidates.

Key words: whole-exome sequencing; next-generation sequencing; clinical genomics

Introduction

Since the completion of the Human Genome Project in 2003 [1], an innovative genomic era of next-generation sequencing (NGS) technologies has begun to revolutionize medical practice [2]. NGS is thriving where conventional genetic tests (i.e. candidate gene sequencing, array-based comparative genomic hybridization and karyotyping) have failed to elucidate a cause for Mendelian diseases [3]. With the power to detect novel variants from only a small number of individuals (including a singleton), NGS is proving invaluable for modern geneticists, boasting a putative diagnosis rate that ranges from 21 to 25% for rare disease of unknown etiology [4, 5]. Its application is not limited to rare diseases, and includes cancers, complex diseases and RNA sequencing; however, discussion of these is beyond the scope of this review.

NGS uses massively parallel nucleic acid sequencing technologies capable of providing a cost-effective approach to large-scale resequencing of human samples for both medical and population genetics [6, 7], so much so that the annotation and interpretation of sequencing data is now the rate limiting step [6]. Since the transition from Sanger sequencing to NGS, sequencing costs have declined exponentially [8], increasing the application of NGS in research domains and clinical diagnostics. The human genome comprises 3 billion base pairs;

Eleanor G. Seaby is a final year medical student who has just completed an intercalated MMedSc in Clinical Genomics under the supervision of professor Sarah Ennis.
Reuben J. Pengelly is a research fellow in the Genetic Epidemiology and Bioinformatics Research Group at the University of Southampton and is involved in method development and analysing NGS data sets encompassing diverse clinical phenotypes.
Sarah Ennis is professor of Genomics. She leads the Genomic Informatics Group at the University of Southampton and specializes in method development and application of high-throughput genomics to elucidate biological mechanisms of human disease using next-generation sequencing data.

yet, only 1–2% code for protein (the exome). Most known disease-causing variants alter the protein-coding sequence, and relatively little is known about the function of non-coding DNA, although more is being elucidated by the encyclopedia of DNA elements (ENCODE) project [9, 10]. Because it is estimated that ∼85% of disease-causing mutations reside in the exome [6, 11, 12], a cheaper alternative to whole-genome sequencing, whole-exome sequencing (WES) has become increasing popular owing to its compromise between cost, genome coverage, diagnostic yield and interpretability [7, 13]. Therefore, this article aims to review and outline the fundamental basis of WES applied to clinical medicine, providing a guide for both genomic informaticians and clinicians.

WES Methods

DNA sources and extraction

The first step of WES involves the acquisition of high-quality genomic DNA (gDNA) from biological samples, most commonly extracted from peripheral blood leukocytes. Common extraction methods include the traditional 'salting out' technique and spin column-based methods. Noteworthy, gDNA can also be extracted from saliva, which provides a non-invasive alternative to venesection, but at the expense of quantity and quality, particularly pertaining to risk of DNA contamination from oral microflora and food remnants [14]. Formalin-fixed paraffin-embedded (FFPE) samples are another viable source, i.e. in archival histopathology specimens and also in cancers (assessed alongside germ line samples from the same individual to differentiate somatic from germ line variants) [15]. However, FFPE tissue yields far poorer DNA quality, with discordant reports concerning the viable standard of sequencing output derived from FFPE-extracted DNA. Some studies report comparable analytical output [16–18], while Genomics England report substandard sequencing quality in 50% of FFPE samples used in the 100 000 Genomes Project [19].

Exome library preparation

The preparation of an exome enrichment library follows DNA extraction. Agilent, Illumina and NimbleGen are three commonly used exome capture kits, and are compared and summarized in Table 1. Product selection should be influenced by platform-specific strengths and weaknesses. Despite differences, all capture technologies obey the same three basic principles: (1) DNA fragmentation, (2) adaptor ligation and (3) target enrichment.

1. gDNA is sheared into random fragments either mechanically by ultrasonication methods, or biologically by enzymatic digestion [24].
2. Fragment ends are subsequently blunted and ligated with adaptors. Agilent's SureSelectQXT and the Illumina Nextera platforms used a shearing-free transposase-based library preparation.
3. Targeted enrichment of exonic regions follows library preparation, and methods vary between capture platforms.

Agilent's SureSelect Human All Exon, Nimblegen's SeqCap EZ Exome Library and Illumina's TruSeq use a hybridization method with complementary baits and magnetic bead pull-down, while Agilent's Haloplex platform uses an amplicon-based capture [25]. Non-targeted sequences are washed away,

Table 1. Comparison of exome capture kits

	Agilent HaloPlex Exome	Agilent SureSelect All Exon V5.0	NimbleGen SeqCap EZ Human Exome V3.0	Illumina Nextera Rapid Capture Expanded Exome	Illumina TruSeq Exome Enrichment
Target size	37 MB	50 MB	64 MB	62 MB	64 MB
Probe size (bases)	161 ± 75	120	55–105	95	95
Number of probes	2.49 million	789 000	2.1 million	347 517	340 427
Number of targeted exons	557 999	335 765	∼300 000	201 121	201 121
Reads on target (%)	∼80	∼80	>70	∼60	>65
(%) target bases covered at ≥10x	>90	>90	>95	>97	>97
Recommended DNA input	200 nanograms	3 micrograms	3 micrograms	50 nanograms	500 nanograms
Fragmentation method	Transposomes	Ultrasonication	Ultrasonication	Transposomes	Ultrasonication
Strengths	Smallest amounts of DNA required	Fewer duplicated reads	More uniform coverage	Best coverage of UTRs and micro RNAs	Best coverage of UTRs and micro RNAs
		Better sensitivity for indels	Most reads on target	Fast	
		High alignment rate			
Weaknesses	Variable length target regions	Most affected by high GC content	Alignment rate less than Agilent	High proportion of off target reads	High proportion of off-target reads
		Variable length target regions		Coverage bias in GC-rich regions	Variable length target regions
		Fewer reads on target compared with NimbleGen			Poorest for reads on target

Note. A tabulated comparison of Agilent's SureSelect and Haloplex capture kits, Illumina's Nextera and TruSeq kits and NimbleGen's SeqCap kit.
UTR = untranslated region [20–23].

hybridized fragments are eluted and the resultant enriched library is amplified [20].

Exome sequencing

Following exon enrichment, the resultant captured library is subject to high-throughput, massively parallel sequencing to produce millions of short reads. The exome-sequencing methodological workflow is visualized in Figure 1. Current sequencing platforms include Life Technologies SOLiD, Roche's 454 Genome Sequencer, Pacific Bioscience's RS, Life Technologies Ion Proton and the current market leader, Illumina's HiSeq range of sequencers, which use a sequencing by synthesis approach [35, 38]. Sequencing of both the forward and reverse strands allows for creation of paired-end reads. These provide longer-range information than single reads, yielding greater alignment accuracy when computationally mapped to the human genome reference sequence [6, 20]. This target-mapping process allows for the identification of coding nucleotide and splice site changes in the patient's DNA that vary from the reference sequence (variants).

Clinical application of WES

The application of WES has proven successful in the discovery of novel disease genes and pathogenic mutations across a wide range of disciplines, resulting in new diagnoses with considerable prognostic impact (Table 2). In 2009, Choi *et al.* reported the first diagnosis resolved by WES in a patient misdiagnosed with Bartter syndrome; WES revealed a novel homozygous mutation in *SLC26A3*, a gene in which previous mutations were causal for congenital chloride-losing diarrhoea (CLD). Re-evaluation of the clinical phenotype confirmed the diagnosis of CLD [43]. Worthey *et al.* published the first clinical case using exome sequencing to diagnose and cure a rare form of inflammatory bowel disease [39].

With the growing published successes of WES, there has been increased demand for novel informatics and analytical strategies to compute vast sequencing data into high-quality calls with sufficient sensitivity and specificity for clinical application. However, WES data are limited by current WES methodology; competing chemistries differ in their capture efficiency and probe design, with 5–15% of targeted regions suboptimally covered for sufficient variant detection [34, 47], and some regions are not amenable to the mapping of short reads [48]. In designing the ideal exome capture platform, the three following conditions would need to be met: (1) 100% coverage of all the coding regions at sufficiently high read depth to sensitively detect all variants, (2) all regions enriched by capture probes would correctly map to target and (3) the allelic biases of capture would be minimized to capture all indels and copy number variants (CNVs) with 100% sensitivity. Despite continual development, current capture platforms cannot meet such criteria. Many of these issues can be resolved using whole-genome sequencing (particularly for indels and CNV detection) by virtue of its continuous coverage; however, cost continues to be a barrier to the uptake of this in many settings [49].

Capture kits are vulnerable to off-target enrichment, particularly when enrichment probes share sequence similarity with non-coding sequences [20, 50]. Furthermore, there is difficulty in uniquely mapping to regions with high sequence identity, e.g. gene families or repeated domains, and in calling genotypes at the end of short reads. Additionally, sequencing platforms have systematic errors, which should be considered during data processing [51]. Nonetheless, sequencing data are usually of sufficient quality to undergo next-step data processing (Figure 2), subject to meeting quality control standards assessed continually throughout data analysis. Quality control procedures must (at the least) control for poor genotype call quality, sample mislabelling [53], exogenous DNA contamination and alignment errors.

Variant filtering

Annotated WES data typically identify ~25 000 coding variants, requiring high-throughput *in silico* methods to prioritize candidate variants amongst a sea of background noise. The differentiation of common single nucleotide variants (SNVs) that represent benign inter-individual variation from disease-causing variants is analogous to finding a needle in a pile of needles [54]. The difficulty lies when one of 25 000 variants is sufficient to cause a devastating disease, as is the case for many monogenic, Mendelian-inherited diseases. This challenge is further complicated when the mutant allele is completely novel, and there is no prior or established literature regarding pathogenicity of the variant. The most commonly called variants are synonymous SNVs, followed by non-synonymous SNVs and splice site variants; however, the less common frameshift and stopgain/stoploss variants are more likely to have deleterious effects at the protein level, and this provides a good starting place for variant prioritization [55]. There is an obvious demand to apply a filtering framework to reduce the vast number of variants to a manageable list of candidates. Strategies for extricating disease-causing alleles depend on multiple factors such as: phenotype segregation within families (where available), presumed mode of inheritance, extent of locus heterogeneity and computational predictive tools based on evolutionary conservation and impact of protein change [6]. Filtering strategies are discussed below in further detail and divided into primary and secondary filtering (Figure 3).

Primary filtering

The main objective of primary filtering is the exclusion of benign variants. Although relatively crude, primary filtering should be accepted as a semi-rigid strategy and not preclude the revisiting of disregarded variants.

Quality control

Quality control comprises a mandatory component of variant annotation and analysis. Variant calling can be prone to error, and where evident, low-quality variants should be immediately excluded. Sufficient read depth (>20) is crucial to the sensitivity and specificity of variant calling, particularly for heterozygous calls and in assessing allelic balance. Sequencing depth should be set much higher where somatic variation is to be evaluated because of the lower proportion of chromosomes that will harbour the variant [15].

False-positive (type I) errors can be minimized by paying attention to poor alignment around processed pseudogenes, alignment artefacts and variants occurring in homopolymer tracts. Strand bias, whereby called genotypes disagree between the forward and reverse strands, usually reflects an error and should have a low threshold for exclusion [56]. Certain genes from large gene families with increased homology between members, e.g. olfactory receptors, harbour a higher rate of mutability than others. These 'polymorphic' genes are enriched for type I errors with regards to true pathogenic potential, and are reflected by either true gene mutability or alignment artefacts

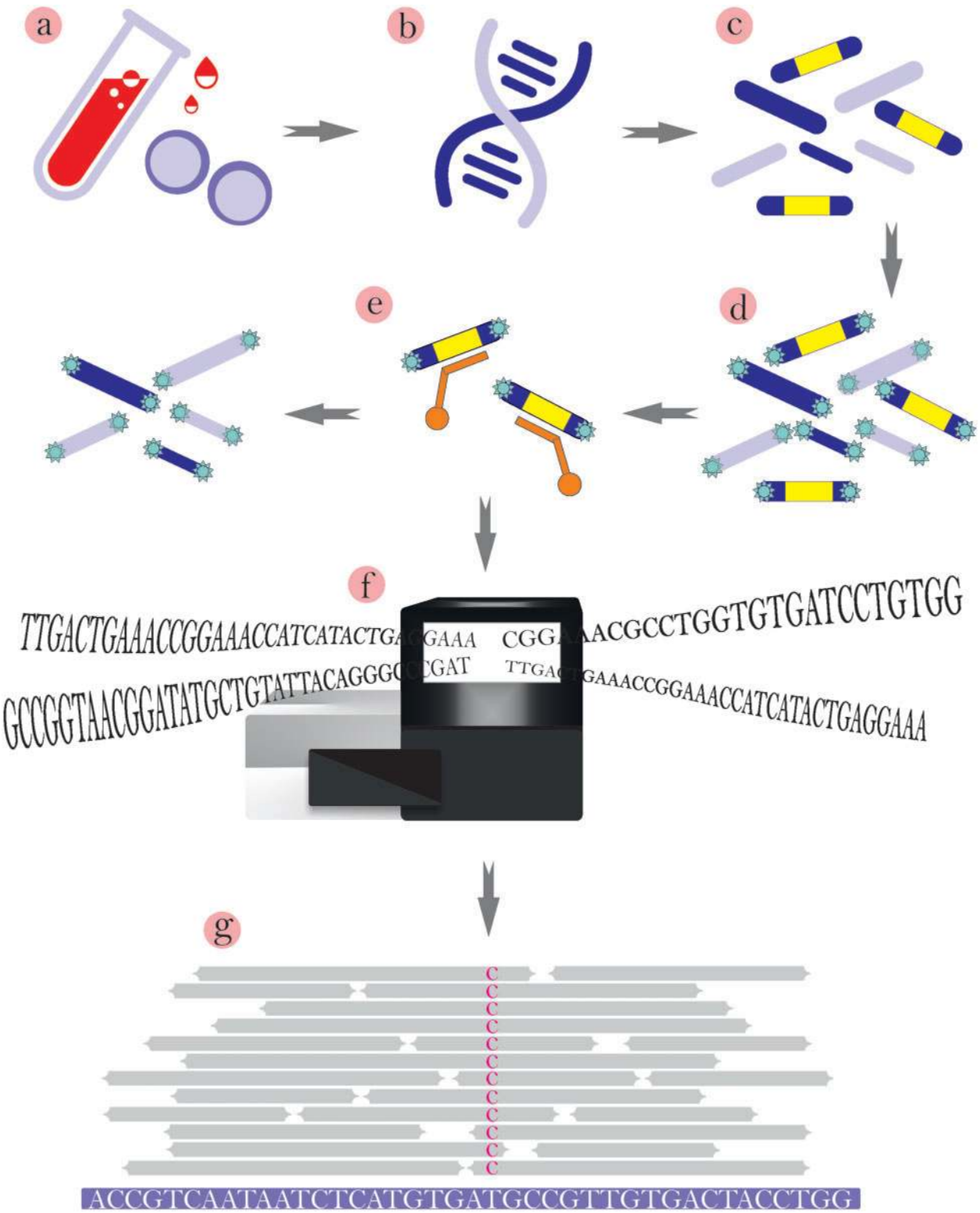

Figure 1. Workflow example of WES. (**A**) A blood sample is collected containing peripheral lymphocytes. (**B**) gDNA is extracted from white blood cells using extraction kits or the salting-out method, and the quality and quantity is assessed. (**C**) DNA is fragmented either by sonication or enzymatic methods, which vary between library preparations. (**D**) Fragment ends are repaired by removing overhanging nucleotides, and the ends are ligated to adaptors (stars). Rectangular regions within fragments represent exons present in DNA fragments. (**E**) Aqueous-phase hybridization capture enriches exonic sequences (rectangular regions) by ligation of fragments to biotinylated baits (probes) as used by most enrichment platforms. Hybridized fragments are recovered by biotin-streptavidin-based magnetic bead pulldown. Uncaptured regions are washed away. The enriched library of exonic fragments are eluted and amplified. (**F**) The resultant exome library is sequenced using massively parallel sequencing technologies, producing millions of sequenced reads. (**G**) Raw data are aligned to the human genome reference sequence and downstream *in silico* tools analyse output data. (A colour version of this figure is available online at: http://bfg.oxfordjournals.org)

Table 2. Small subset of causal genes identified by WES

Causal gene	Disease	OMIM ref	Reference
XIAP	X-linked inhibitor of apoptosis deficiency	300079	Worthey *et al.* [39]
WDR62	Severe cerebral cortical malformations	613583	Bilgüvar *et al.* [40]
ANGPTL3	Familial combined hyperlipaemia	604774	Musunuru *et al.* [41]
TGM6	Spinocerebellar ataxia	613900	Wang *et al.* [42]
SLC26A3	Congenital chloride diarrhoea	126650	Choi *et al.* [43]
MLL2	Kabuki syndrome	147920	Ng *et al.* [44]
ATP1A3	Alternating hemiplegia of childhood	182350	Rosewich *et al.* [45]
PRRT2	Paroxysmal kinesigenic dyskinesia	614386	Chen *et al.* [46]

Note. A sample of genes and associated diseases discovered by WES, illustrating the breadth of research areas using the technology.
OMIM ref—reference entry of gene and associated diseases in the Online Mendelian Inheritance in Man (OMIM) database.

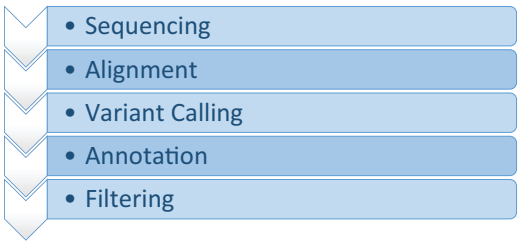

Figure 2. Data processing workflow of NGS data. Following (**1**) **sequencing**, the initial computational step involves the (**2**) **alignment** of sequencing data to the human genome reference sequence [2]. Following alignment and target mapping, *in silico* tools are used in (**3**) **variant calling**. SNV genotypes are called most reliably, and the two best used SNV discovery algorithms are SAMtools [26] and GATK [27], which use the principle of Bayesian detection. The challenge for these SNV detection algorithms is the differentiation of a true variant from a sequencing error, particularly given the high error per base rate of NGS [28]. Software are available for large indel calling, e.g. Pindel [29] and Softsearch [30], but these have suboptimal sensitivity and specificity [31, 32]. Small indels can be detected in SAMtools [26]. Calling CNVs poses a great challenge in WES analysis because of a non-uniform depth of coverage across regions of the exome [33]. (**4**) **Annotation** of called variants provides the essential information required for downstream analysis and interpretation. A commonly used tool is ANNOVAR [34]. The final step involves a (**5**) **filtering** process for the identification of causal genes.

[51]. Although tempting to disregard these variants, evidence has shown that laboratories underdiagnose because they dismiss pathogenic variants found in particularly polymorphic genes [57, 58].

Candidate gene analysis

Gene-specific filtering involves targeting variants in candidate genes associated with the clinical phenotype, and is somewhat analogous to a targeted gene panel. Of course, this raises the question of why an exome may be performed in preference, or even before a gene panel? For one, the cost of exome sequencing is often equivalent to a single gene panel and yields data from all known genes (but at lower read depth). Secondly, WES is capable of greater data throughput than that of a gene panel, and if no variants of interest are identified initially, the gene list can be revised with the option of expanding data interrogation across the entire exome. Thirdly, having access to raw data enables interrogation of call quality and alignment, allowing for more informed variant scrutiny. Furthermore, genes selected for candidate gene analysis are typically obtained from up-to-date curated databases in addition to the most current published literature; this minimizes the risk of missing a new disease-associated gene unavailable on a gene panel. The unbiased data capture of WES allows for *in silico* revision of candidate gene lists on the description of new candidate genes without cost duplication from ordering further gene panels. However, cost duplication should be weighed against data storage costs.

Exclusion of synonymous variants

Owing to the redundancy of the genetic code, synonymous variants are SNVs that do not cause an amino-acid change at that codon. Their removal forms an integral part of most downstream informatics pipelines and reduces variant lists by approximately 50% [6]. Although generally assumed benign and appropriately excluded, synonymous variants have been acknowledged to harbour pathogenic properties, particularly concerning changes in protein expression and splicing [59]. The difficulty lies in excluding a large repository of probably benign synonymous noise, at the expense of a small number of false negatives (type II errors). There is currently an unmet balance between minimizing type II errors and the lack of availability of affordable, high-throughput functional assays to assess the true effects of synonymous variants. The majority of downstream *in silico* prediction tools do not assess synonymous SNVs, and therefore our understanding of the functional consequences of these variants is limited by insufficient interpretation. Yet since 2014, there are three predictive algorithms capable of predicting the functional consequences of non-coding variants by using nucleotide-sequence conservation metrics: FATHMM-MKL [60], GWAVA [61] and CADD [62]. It is envisaged that as we accrue functional data on synonymous variants, analytical tools will continue to develop, providing improved predictive accuracy.

Filtering by minor allele frequency

Minor allele frequency (MAF) is the reported allelic frequency of a given variant in a given population. MAFs are available from publicly available repositories and can be used to differentiate 'rare' variants from polymorphisms using a cut-off value of <0.01. Although this somewhat arbitrary value can select for rare alleles, there are issues to consider. Certain databases, such as the single nucleotide polymorphism database (dbSNP), are contaminated with pathogenic variants (albeit in modest number). Some alleles that are inherited in an autosomal recessive manner may segregate at appreciable frequencies greater than a MAF of 1% [63], and some rare variants may segregate at higher frequencies in populations with an appreciable founder effect. Therefore, when using MAF to assess rare variants, it is prudent to consider the presumed mode of inheritance and ethnicity of the sequenced individuals, and the ethnic diversity

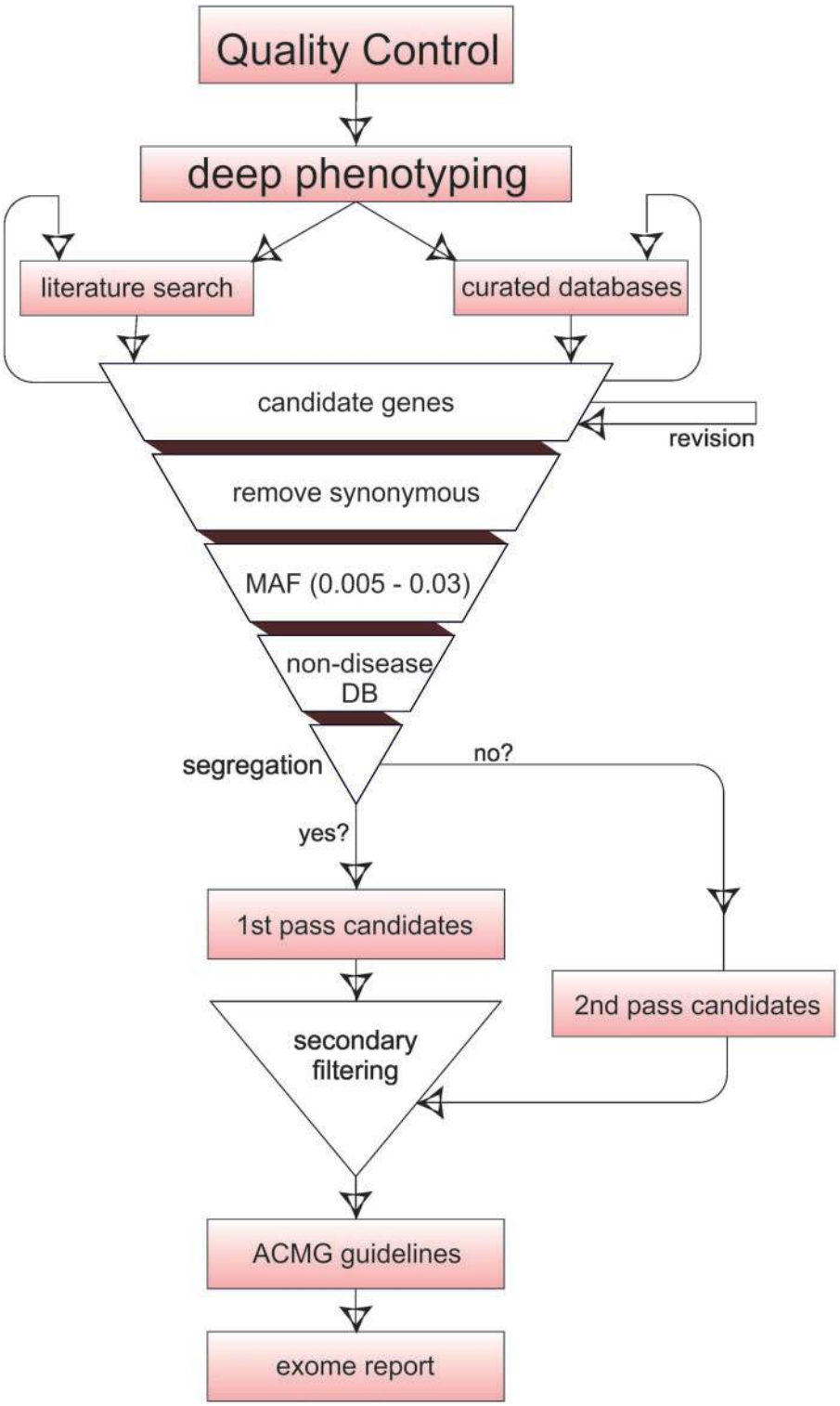

Figure 3. Suggested exome data filtering strategy. The suggested filtering strategy begins with quality control, and the removal of low-quality variants. Deep phenotyping and sharing of accurate clinical information best inform the selection of candidate gene lists from curated databases and published literature. If new information becomes available, the gene list can be revised, and the data iteratively analysed over time. Evidence of new disease–gene associations will be added to the literature and curated in disease databases, increasing the repository of candidate genes. Exome data are reduced by primary filtering, which often includes: the removal of synonymous variants, a MAF cut-off that can range between 0.005 and 0.03 depending on the expected mode of inheritance and variant exclusion from control group databases curated from in-house sequencing data. Variants that meet segregation criteria based on presumed mode of inheritance (first pass candidates) can undergo scrutiny using secondary analysis. Prioritized variants are cross-referenced with the ACMG guidelines [52] on pathogenicity and scored according to their criteria. If no suitable candidates are identified, segregation analysis can be expanded to include non-penetrant alleles (second pass candidates) and undergo secondary filtering. An exome report is compiled containing a list of prioritized variants. (A colour version of this figure is available online at: http://bfg.oxfordjournals.org)

captured within the database from which the MAF is obtained. For rare, dominant alleles, smaller MAFs tend to be used (0.01–0.05). For presumed recessive diseases, more conservative frequencies are used (i.e. >0.01–0.03), which provide a compromise between unwanted noise and minimizing type 2 errors from alleles segregating at >0.01 [64]. Where available, variants should be cross checked with in-house non-disease variant databases to minimize population-specific effects and remove errors resulting from technical artefacts specific to the in-house pipeline.

Filtering by segregation

Filtering by variant segregation in families can powerfully reduce the number of potential causal variants to a manageable list of candidates. However, it relies heavily on the optimal selection of individuals for sequencing informed by the apparent mode of inheritance. It is important that all available and relevant clinical information transitions from the clinic to the genomic informatics laboratory and are completed maintaining patient anonymity. A comprehensive family history, excellent pedigree documentation and deep phenotyping are essential. Pedigree information concerning ethnicity and consanguinity should always be provided where available. Excellent multidisciplinary communication ensures the most appropriate individuals are sequenced to maximize segregation filtering power. One way to achieve this is by minimizing the probability that alleles of sequenced individuals are shared by chance. Where possible, more distantly related affected individuals should be sequenced preferentially (i.e. first cousins). Of course, this relies on affected individuals being phenotypically identical, further justifying the need for accurate phenotyping with minimal bias. For presumed *de novo* inheritance, trio analysis (parents/child) is preferred where funding permits, as it can powerfully identify new causal variants in the offspring. Segregation analysis is not without caveats, particularly concerning late onset diseases, non-paternity, mosaicism and incomplete penetrance.

Incomplete penetrance

Segregation analysis relies on complete disease penetrance, and therefore pathogenic variants segregating in healthy individuals are filtered out. Where variants are known to display variable penetrance, a multidisciplinary team must judiciously agree on the likely variant status, particularly if the variant segregates in a pedigree with disparate phenotypes. In suspected incomplete penetrance, it is important not to falsely label non-specific symptoms in an unaffected individual as mild phenotypic characteristics in support of a diagnosis. For example, where a variable penetrant genotype can cause severe respiratory disease, an unaffected individual may report 'lots of coughs' when questioned about their respiratory history, and these may be entirely benign. Functional verification provides the only definitive conclusion; yet, this is costly and unfeasible, given the sheer volume of potential variants.

Compound heterozygosity

In homozygous recessive disease patterns, both alleles harbour the identical mutation at the same locus. Compound heterozygosity should be considered where heterozygous variants are called in the same gene and reside proximally to one another, or may functionally interact. Because these variants are not homozygotes in segregation analysis, it is prudent not to miss inheritance of proximal variants in *trans* (with one variant from each parent), even if the alleles alone are insufficient to cause disease. Trio analysis is advantageous in this regard, and can offer discovery of compound heterozygosity by observing segregation of alleles inherited from each parent. Unfortunately, NGS is limited with regards to determining phase of alleles from a single sample, therefore necessitating parental or long-range sequencing. For compound heterozygotes occurring in proximal regions where reads overlap, it is sometimes possible to scrutinize and predict *cis/trans* inheritance when the heterozygotes occur on different reads.

Secondary filtering

Distinct from primary filtering, which removes probable benign variants, second filtering uses strategies to prioritize remaining candidate variants by consideration of a conglomeration of factors. These include *in silico* prediction tools, re-evaluation of variants occurring in mutable genes, CNVs and multiallelic hits. The end result of secondary filtering is the cross-reference of best candidate variants with consensus pathogenicity guidance for reporting back to the clinic.

In silico prediction

Variant pathogenicity can be predicted using computational tools that consider the effects of a variant at the nucleotide, amino-acid and protein levels. Frameshift, nonsense and canonical splice site variants are considered most likely to disrupt gene function and are thus assigned greater pathogenic potential. This underpins much of the framework behind laboratory reporting (as discussed later). However, the more frequent missense variants require aggregation of a greater body of evidence in support of pathogenicity and should at least include: evolutionary conservation metrics; the biochemical consequence of an amino acid change; and for splicing variants, splicing prediction software (Table 3) [52, 75]. However, *in silico* tools have limitations, their sensitivity and specificity do not meet diagnostic standards and most annotation algorithms currently ignore gene-specific domains and multivariant interactions [76].

Copy number variants

CNVs comprise an integral part of genome analysis and have roles in both common and rare diseases, but WES has been traditionally poor at resolving them [77, 78]. CNVs in candidate genes warrant consideration, particularly when a patient is heterozygous for a recessive mutant allele.

Although considered poor at detecting CNVs because of the punctate nature of sequenced data, WES technologies are beginning to make headway; Agilent's OneSeq target enrichment assay promises huge improvements in the detection of CNVs, as well as copy neutral loss of heterozygosity and indels. Read depth software are also available to crudely assess changes in copy number; these software use read depth as a measure of the amount of DNA present at a given locus to predict copy number variations in the test sample compared with the given reference. Software examples include ExomeCNV [79] and ExomeDepth [80], but these are not highly sensitive nor specific and are prone to upstream errors occurring in the capture procedure [81]. Additionally, statistical *in silico* tools exist, which claim good sensitivity for CNVs, but input data are noisy because of non-uniform capture across non-contiguous sequence data from discrete exons [78].

Table 3. *In silico* prediction tools

Category	Algorithm	Source	Principle
Non-synonymous SNV prediction	*SIFT [65]	http://sift.jcvi.org	Evolutionary conservation
	*PolyPhen-2 [66]	http://genetics.bwh.harvard.edu/pph2	Evolutionary conservation and protein structure/function
	*MutationTaster [67]	http://www.mutationtaster.org	Evolutionary conservation and protein structure/function
	*Grantham [68]	Grantham *et al. Science*.	Biological consequence of amino-acid change
Synonymous SNV prediction	FATHMM-MKL	http://fathmm.biocompute.org.uk/fathmmMKL.htm	Sequence conservation within hidden Markov models
	GWAVA	https://www.sanger.ac.uk/sanger/StatGen_Gwava	Integration of various genomic and epigenomic annotations
	CADD	http://cadd.gs.washington.edu/	Multiple genomic annotations
Splicing prediction	*MaxEnt [69]	http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html	Maximum entropy model
	GeneSplicer [70]	https://ccb.jhu.edu/software/genesplicer/	Markov model
	Human Splicing Finder [71]	http://www.umd.be/HSF/	Position dependent logic
	MutPred Splice [72]	http://mutdb.org/mutpredsplice/about.htm	Machine learning prediction of exonic variants
Conservation prediction	*PhyloP [73]	http://compgen.bscb.cornell.edu/phast/	Conservation scoring
	*GERP++ [74]	http://mendel.stanford.edu/SidowLab/downloads/gerp/	Conservation scoring

Note. Predictive *in silico* algorithms available to aid in the interpretation of variants of unknown significance.

Reporting on prioritized variants

Application of the aforementioned filtering methodology will typically streamline ~25 000 variants to a manageable list of prioritized candidates (and occasionally a single candidate variant). To accurately assess candidate pathogenicity in a clinical context, an extensive literature search is required. Previous reporting of the variant (as either pathogenic or benign) is of huge importance, especially when backed up by functional studies. Where literature is sparse, assessment of protein function and the domains affected can help assign importance to variants, e.g. if a variant alters the protein sequence of a functional phosphorylation domain, this would be more salient than a variant altering an innocuous repeat region. Gene expression studies, animal models of gene knock outs or *in vitro* functional effects of stop gains and alternative splicing are all further sources that can help assess variant significance. Because the accurate interpretation of sequence variants is critical in influencing clinical outcomes, there has been a need for a standardized classification framework for variant pathogenicity, for example, as recommended by the American College of Genetics and Genomics (ACMG) in the USA and the Association for Clinical Genetic Science in the UK [52, 75, 82].

Assignment of pathogenicity

Application of American and UK guidelines is useful, but many prioritized variants fall into the category of 'variant of unknown significance'; a massive bottleneck effect is incurred by the sheer number of these variants that require a step-change in functional assessment to reliably predict clinical relevance. Most often, there are insufficient funds and resources to follow-up multiple variants; consequently, variant interpretation necessitates the convergence of relevant clinical disciplines and genomic informaticians to judiciously consider all available evidence and make the most informed decisions regarding diagnosis and/or treatment where applicable. Furthermore, bridging communication between genomic informatics and clinicians should not cease after the initial exome report. New clinical information can entirely revise analytical methodology and delineate a new list of candidate genes. One the biggest advantages of obtaining genomic data is the ability to return to it should new clinical information, disease–gene associations or improved analytical strategies become available.

Ethical considerations and incidental findings

WES raises ethical issues, most notably concerning consent, data sharing and return of information. There are concerns that informed consent is insufficient in educating patients about the scope of potential results identified, particularly in relation to the return of incidental findings.

When presented with data from an entire exome or genome, there is always the possibility of incidentally discovering pathogenic mutations unrelated to the presenting phenotype. This is a particularly topical and contentious issue, and there is much debate around whether 'actionable' incidental findings should be reported back to patients [82–84]. Current guidelines by the American College of Medical Genetics and Genomics recommend that constitutional mutations from a list of 56 genes should be reported back to the referring clinician, regardless of the initial indication for exome or genome sequencing [52]. These guidelines have been subject to heavy criticism within the literature, and alternative position statements exist from other organizations. The rationale behind reporting incidental findings ultimately concerns the sharing of medically valuable information to patients, providing them with the opportunity for medical interventions that carry substantial prognostic benefit. Although commendable, there are obvious flaws: some known pathogenic mutations will not be fully penetrant, variants of unknown significance in disease-associated genes may be entirely benign but cause undue concern and there are issues surrounding patient consent [85]. With an estimated frequency of 1–3% [86], incidental findings are not to be taken lightly,

particularly considering the financial burden they may have on the healthcare system. Furthermore, it is unknown how the reporting of such findings will materialize as NGS forms a ubiquitous component of modern-day diagnostics.

Limitations

WES has substantial diagnostic potential capable of uncovering causal mutations in rare monogenic diseases; but, it is not without limitations. Its limited target capture of only 1–2% of the genome completely disregards clinically relevant alleles occurring outside of these regions, missing deep intronic variants [87, 88]. Cost remains a substantial issue, although this may be offset by the unnecessary financial burden of 'reflex testing'. Other limitations include: the disregard for epigenetic modifications, variability in *in silico* sequence capture by different platforms, read depth and alignment errors, small CNVs and cryptic indels (poorly resolved and aligned) and the subjectivity of secondary filtering during data analysis—different laboratories will have their own methodology for variant prioritization and may use different thresholds for the inclusion or exclusion of variants [69].

In a wider context, WES is a disruptive technology that challenges the traditional practice of clinical genetics. The inevitable move away from traditional methods in favour of WES may threaten the jobs of cytogeneticists, technicians and other professionals trained in the pre-NGS era, unless appropriate retraining schemes are established. Nonetheless, training medical professionals in genomics will help to improve some of the poor communication between clinical and research disciplines that currently exist; interpretation of exome data requires analysis by genomic informaticians with limited clinical knowledge, and many clinicians are unfamiliar with this rapidly evolving technological discipline and require continued education. There is an obvious demand to train clinicians in genomic informatics to be able to close the gap between the two different disciplines and truly demonstrate personalized, translational medicine.

Future

There is no doubt that exome and genome sequencing will become increasingly prevalent. Concomitant with media interest, curiosity is rising amongst the lay population who are beginning to use commercially available personalized genomic services. But as it stands today, variants solely identified by NGS technologies do not meet clinical diagnostic standards, and this is often underappreciated. There is thus temptation to infer causation from unverified data, which is particularly problematic without genomic and clinical expertise. This necessitates the convergence of genomic informatics with multidisciplinary clinical medicine to nurture a new field of clinical genomics for the safe and accurate reporting of clinical variants. In England, this is being driven by the Department of Health, who are funding the 100 000 Genomes Project [89], which aims to sequence 100 000 genomes by 2017 by recruiting patients from the National Health Service (NHS) through Genomic Medicine Centres focusing particularly on rare disease and cancers. Similarly in the USA, a Precision Medicine Initiative has recently been announced [90]. These projects aim to combine genomic data with clinical medicine, advance medical research, develop new therapies and accelerate the genomic industry. This will ultimately catapult clinical medicine into the genomics era, requiring genomic literacy of many clinicians to ensure the best possible outcomes for patient-centred healthcare.

Key Points

- A practical guide to whole-exome sequencing that is relevant to both genomic informaticians and clinicians.
- A suggested filtering strategy for the extrication of causal variants.
- The value of next-generation sequencing technology in clinical medicine.
- Ethical issues and incidental findings.
- The challenges and limitations of whole exome sequencing.

Funding

Reuben J Pengelly is supported by the University of Southampton Doctoral Training Fund.

References

1. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* 2004;**431**(7011):931–45.
2. DePristo MA, Banks E, Poplin R, *et al*. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 2011;**43**(5):491–8.
3. Nielsen R, Paul JS, Albrechtsen A, *et al*. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet* 2011;**12**(6):443–51.
4. Yang Y, Muzny DM, Reid JG, *et al*. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N Engl J Med* 2013;**369**(16):1502–11.
5. Taylor JC, Martin HC, Lise S, *et al*. Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. *Nat Genet* 2015;**47**:717–26.
6. Bamshad MJ, Ng SB, Bigham AW, *et al*. Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* 2011;**12**(11):745–55.
7. Ng SB, Buckingham KJ, Lee C, *et al*. Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* 2010;**42**(1):30–5.
8. National Human Genome Research Institute. *Sequencing costs*. 2014. www.genome.gov/sequencingcosts/.
9. Ecker JR, Bickmore WA, Barroso I, *et al*. Genomics: ENCODE explained. *Nature* 2012;**489**(7414):52–5.
10. Consortium EP. The ENCODE (ENCyclopedia of DNA elements) project. *Science* 2004;**306**(5696):636–40.
11. Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genet* 2003;**33**:228–37.
12. Majewski J, Schwartzentruber J, Lalonde E, *et al*. What can exome sequencing do for you? *J Med Genet* 2011;**48**(9):580–9.
13. Rabbani B, Tekin M and Mahdieh N. The promise of whole-exome sequencing in medical genetics. *J Hum Genet* 2013;**59**(1):5–15.
14. Kidd JM, Sharpton TJ, Bobo D, *et al*. Exome capture from saliva produces high quality genomic and metagenomic data. *BMC Genomics* 2014;**15**(1):262.
15. Bao R, Huang L, Andrade J, *et al*. Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing. *Cancer Inform* 2014;**13**(Suppl 2):67–82.

16. Van Allen EM, Wagle N, Stojanov P, *et al*. Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. *Nat Med* 2014;**20**(6):682–8.
17. Hedegaard J, Thorsen K, Lund MK, *et al*. Next-generation sequencing of RNA and DNA isolated from paired fresh-frozen and formalin-fixed paraffin-embedded samples of human cancer and normal tissue. *Plos One* 2014;**9**(5):e98187.
18. Kerick M, Isau M, Timmermann B, *et al*. Targeted high throughput sequencing in clinical cancer settings: formaldehyde fixed-paraffin embedded (FFPE) tumor tissues, input amount and tumor heterogeneity. *BMC Med Genomics* 2011;**4**(1):68.
19. Genomics England. *100,000 Genomes Project Update*. 2015 26/05/2015. http://www.genomicsengland.co.uk/100000-genomes-project-update/.
20. Clark MJ, Chen R, Lam HY, *et al*. Performance comparison of exome DNA sequencing technologies. *Nat Biotechnol* 2011;**29**(10):908–14.
21. Chilamakuri CSR, Lorenz S, Madoui MA, *et al*. Performance comparison of four exome capture systems for deep sequencing. *BMC Genomics* 2014;**15**(1):449.
22. van der Werf IM, Kooy RF, Vandeweyer G. A robust protocol to increase NimbleGen SeqCap EZ multiplexing capacity to 96 samples. *PLoS One* 2015;**10**(4):e0123872.
23. Sulonen AM, Ellonen P, Almusa H, *et al*. Comparison of solution-based exome capture methods for next generation sequencing. *Genome Biol* 2011;**12**(9):R94.
24. van Dijk EL, Jaszczyszyn Y, Thermes C. Library preparation methods for next-generation sequencing: tone down the bias. *Exp Cell Res* 2014;**322**(1):12–20.
25. Samorodnitsky E, Jewell BM, Hagopian R, *et al*. Evaluation of hybridization capture versus amplicon-based methods for whole-exome sequencing. *Hum Mutat* 2015;**36**:903–14.
26. Li H, Handsaker B, Wysoker A, *et al*. The sequence alignment/map format and SAMtools. *Bioinformatics* 2009;**25**(16):2078–9.
27. McKenna A, Hanna M, Banks E, *et al*. The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;**20**(9):1297–303.
28. Stitziel NO, Kiezun A., Sunyaev S. Computational and statistical approaches to analyzing variants identified by exome sequencing. *Genome Biol* 2011;**12**(9):227.
29. Ye, K, Schulz MH, Long Q, *et al*. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* 2009;**25**(21):2865–71.
30. Hart SN, Sarangi V, Moore R, *et al*. SoftSearch: integration of multiple sequence features to identify breakpoints of structural variations. *PloS One* 2013;**8**(12):e83356.
31. O'Rawe J, Jiang T, Sun G, *et al*. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med* 2013;**5**(3):28.
32. Fang H, Wu Y, Narzisi G, *et al*. Reducing INDEL calling errors in whole genome and exome sequencing data. *Genome Med* 2014;**6**:89.
33. Belkadi A, Bolze A, Itan Y, *et al*. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proceedings of the National Academy of Sciences*, 2015;**112**(17):5473–8.
34. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010;**38**(16):e164.
35. Bentley DR, Balasubramanian S, Swerdlow HP, *et al*. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 2008;**456**(7218):53–9.
36. Fuller CW, Middendorf LR, Benner SA, *et al*. The challenges of sequencing by synthesis. *Nat Biotechnol* 2009;**27**(11):1013–23.
37. Lizardi PM. Next-generation sequencing-by-hybridization. *Nat Biotechnol* 2008;**26**(6):649–50.
38. Ju J, Kim DH, Bi L, *et al*. Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *Proc Nat Acad Sci USA* 2006;**103**(52):19635–40.
39. Worthey EA, Mayer AN, Syverson GD, *et al*. Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genet Med* 2010;**13**(3):255–62.
40. Bilgüvar K, Öztürk AK, Louvi A, *et al*. Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. *Nature* 2010;**467**(7312):207–10.
41. Musunuru K, Pirruccello JP, Do R, *et al*. Exome sequencing, ANGPTL3 mutations, and familial combined hypolipidemia. *N Engl J Med* 2010;**363**(23):2220–7.
42. Wang JL, Yang X, Xia K, *et al*. TGM6 identified as a novel causative gene of spinocerebellar ataxias using exome sequencing. *Brain* 2010;**133**(12):3510–18.
43. Choi M, Scholl UI, Ji W, *et al*. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Nat Acad Sci USA* 2009;**106**(45):19096–101.
44. Ng SB, Bigham AW, Buckingham KJ, *et al*. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat Genet* 2010;**42**(9):790–3.
45. Rosewich H, Thiele H, Ohlenbusch A, *et al*. Heterozygous de-novo mutations in ATP1A3 in patients with alternating hemiplegia of childhood: a whole-exome sequencing gene-identification study. *Lancet Neurol* 2012;**11**(9):764–73.
46. Chen WJ, Lin Y, Xiong ZQ, *et al*. Exome sequencing identifies truncating mutations in PRRT2 that cause paroxysmal kinesigenic dyskinesia. *Nat Genet* 2011;**43**(12):1252–5.
47. Biesecker LG. Exome sequencing makes medical genomics a reality. *Nat Genet* 2010;**42**(1):13.
48. Gilissen C, Hoischen A, Brunner HG, *et al*. Disease gene identification strategies for exome sequencing. *Eur J Hum Genet* 2012;**20**(5):490–7.
49. Belkadi A, Bolze A, Itan Y, *et al*. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc Nat Acad Sci USA* 2015;**112**(17):5473–8.
50. Meienberg J, Zerjavic K, Keller I, *et al*. New insights into the performance of human whole-exome capture platforms. *Nucleic Acids Res* 2015;**43**(11):e76.
51. Fuentes Fajardo KV, Adams D, Mason CE, *et al*. Detecting false-positive signals in exome sequencing. *Hum Mut* 2012;**33**(4):609–13.
52. 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):405–24.
53. Pengelly RJ, Gibson J, Andreoletti G, *et al*. A SNP profiling panel for sample tracking in whole-exome sequencing studies. *Genome Med* 2013;**5**(9):89.
54. Cooper GM, Shendure J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat Rev Genet* 2011;**12**(9):628–40.
55. Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012;**491**(7422):56–65.
56. Guo Y, Li J, Li CI, *et al*. The effect of strand bias in Illumina short-read sequencing data. *BMC Genomics* 2012;**13**(1):666.

57. MacArthur D, Manolio T, Dimmock D, *et al*. Guidelines for investigating causality of sequence variants in human disease. *Nature* 2014;**508**(7497):469–76.
58. Brownstein CA, Beggs AH, Homer N, *et al*. An international effort towards developing standards for best practices in analysis, interpretation and reporting of clinical genome sequencing results in the CLARITY challenge. *Genome Biol* 2014;**15**(3):R53.
59. Sauna ZE, Kimchi-Sarfaty C. Understanding the contribution of synonymous mutations to human disease. *Nat Rev Genet* 2011;**12**(10):683–91.
60. Shihab HA, Rogers MF, Gough J, *et al*. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* 2015;**31**(10):1536–43.
61. Ritchie GR, Dunham I, Zeggini E, *et al*. Functional annotation of noncoding sequence variants. *Nat Methods* 2014;**11**(3):294–6.
62. Kircher M, Witten DM, Jain P, *et al*. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 2014;**46**(3):310–15.
63. Hutton SM, Spritz RA. A comprehensive genetic study of autosomal recessive ocular albinism in Caucasian patients. *Invest Ophthalmol Vis Sci* 2008;**49**(3):868–72.
64. Foo JN, Liu JJ, Tan EK. Whole-genome and whole-exome sequencing in neurological diseases. *Nat Rev Neurol* 2012;**8**(9):508–17.
65. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res* 2003;**31**(13):3812–14.
66. Adzhubei IA, Schmidt S, Peshkin L, *et al*. A method and server for predicting damaging missense mutations. *Nat Methods* 2010;**7**(4):248–9.
67. Schwarz JM, Rödelsperger C, Schuelke M, *et al*. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* 2010;**7**(8):575–6.
68. Grantham R. Amino acid difference formula to help explain protein evolution. *Science* 1974;**185**(4154):862–4.
69. Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* 2004;**11**(2-3):377–94.
70. Pertea M, Lin X, Salzberg SL. GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Res* 2001;**29**(5):1185–90.
71. Desmet FO, Hamroun D, Lalande M, *et al*. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res* 2009;**37**(9):e67.
72. Mort M, Sterne-Weiler T, Li B, *et al*. MutPred Splice: machine learning-based prediction of exonic variants that disrupt splicing. *Genome Biol* 2014;**15**(1):R19.
73. Pollard KS, Hubisz MJ, Rosenbloom KR, *et al*. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* 2010;**20**(1):110–21.
74. Davydov EV, Goode DL, Sirota M, *et al*. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* 2010;**6**(12):e1001025.
75. Wallis Y, Payne S, McAnulty C, *et al*. *Practice guidelines for the Evaluation of Pathogenicity and the Reporting of Sequence Variants in Clinical Molecular Genetics*. Birmingham: Association for Clinical Genetic Science (ACGS), 2013, 18–16.
76. Crockett DK, Lyon E, Williams MS, *et al*. Utility of gene-specific algorithms for predicting pathogenicity of uncertain gene variants. *J Am Med Inform Assoc* 2012;**19**(2):207–11.
77. Fromer M, Moran JL, Chambert K, *et al*. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am J Hum Genet* 2012;**91**(4):597–607.
78. Magi A, Tattini L, Cifola I, *et al*. EXCAVATOR: detecting copy number variants from whole-exome sequencing data. *Genome Biol* 2013;**14**(10):R120.
79. Sathirapongsasuti JF, Lee H, Horst BA, *et al*. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics* 2011;**27**(19):2648–54.
80. Plagnol V, Curtis J, Epstein M, *et al*. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics* 2012;**28**(21):2747–54.
81. Ligt J, Boone PM, Pfundt R, *et al*. Detection of clinically relevant copy number variants with whole-exome sequencing. *Hum Mutat* 2013;**34**(10):1439–48.
82. Green RC, Berg JS, Grody WW, *et al*. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet Med* 2013;**15**(7):565–74.
83. Allyse M, Michie M. Not-so-incidental findings: the ACMG recommendations on the reporting of incidental findings in clinical whole genome and whole exome sequencing. *Trends Biotechnol* 2013;**31**(8):439–41.
84. McGuire AL, Joffe S, Koenig BA, *et al*. Ethics and genomic incidental findings. *Science* 2013;**340**(6136):1047–8.
85. Burke W, Antommaria AHM, Bennett R, *et al*. Recommendations for returning genomic incidental findings? We need to talk! *Genet Med* 2013;**15**(11):854–9.
86. Dorschner MO, Amendola LM, Turner EH, *et al*. Actionable, pathogenic incidental findings in 1,000 participants' exomes. *Am J Hum Genet* 2013;**93**(4):631–40.
87. Dhir A, Buratti E. Alternative splicing: role of pseudoexons in human disease and potential therapeutic strategies. *FEBS J* 2010;**277**(4):841–55.
88. King K, Flinter FA, Nihalani V, *et al*. Unusual deep intronic mutations in the COL4A5 gene cause X linked Alport syndrome. *Hum Genet* 2002;**111**(6):548–54.
89. Department of Health. *100K Genomes Project*. 2015. http://www.genomicsengland.co.uk/the-100000-genomes-project/.
90. Collins FS, Varmus H. A new initiative on precision medicine. *N Engl J Med* 2015;**372**(9):793–5.