APPLICATIONS OF NEXT-GENERATION SEQUENCING

Towards precision medicine

Euan A. Ashley

Abstract | There is great potential for genome sequencing to enhance patient care through improved diagnostic sensitivity and more precise therapeutic targeting. To maximize this potential, genomics strategies that have been developed for genetic discovery — including DNA-sequencing technologies and analysis algorithms — need to be adapted to fit clinical needs. This will require the optimization of alignment algorithms, attention to quality-coverage metrics, tailored solutions for paralogous or low-complexity areas of the genome, and the adoption of consensus standards for variant calling and interpretation. Global sharing of this more accurate genotypic and phenotypic data will accelerate the determination of causality for novel genes or variants. Thus, a deeper understanding of disease will be realized that will allow its targeting with much greater therapeutic precision.

The sequencing of the human genome led many to speculate on the near-term potential for clinical medicine[1]. Understanding the genetic basis of disease was naturally expected to lead to better targeted therapies. Indeed, the steep decline in the cost of sequencing, pursuant to the invention of 'next-generation' technologies, facilitated the discovery of many more causative genes[2,3] and, more recently, application to individual patients, including several widely reported examples of genome-driven medical decision making[4–6]. Pilot studies explored the use of genomic information more broadly in patient care[7–9] and the US National Human Genome Research Institute (NHGRI) laid out a 20-year plan for translating insights from genomics to medicine[10,11]. Additionally, direct-to-consumer companies put genotypes in the hands of interested participants[12]. However, the brightest spotlight was provided in 2015 by President Obama in his State of the Union address where he laid out a vision for a national Precision Medicine Initiative in the United States[13,14].

The term 'precision medicine' (BOX 1) was first given prominence by a publication from the US National Research Council that sought to inspire a new taxonomy for disease classification via a knowledge network[15]. In the appendix of that publication, the authors clarify that its coining, as opposed to the more commonly used term 'personalized medicine', was intended to convey the principle that although therapeutics were rarely developed for single individuals, increasingly, subgroups of patients could be defined, often by genomics, and targeted in more specific ways. Worldwide internet searches for the term increased dramatically after the State of the Union address and have remained at similar levels to that of 'personalized medicine' ever since (FIG. 1a).

The timing does seem right for a new approach: genomic data are more readily available, we have a greater understanding of population-scale genetic variation[16,17], and approaches to data integration with electronic medical records will lead to much improved characterization of phenotypes[18]. However, for precision medicine to succeed it also needs to be more accurate. The current algorithms for genome analysis were developed for population or cohort variant discovery where the consequences of reduced accuracy are a lost opportunity for discovery. By contrast, an inaccurate clinical genetic test could lead to very serious consequences for individuals and families with genetic disease. In this Review, I describe promising applications of precision medicine as it currently exists then move on to discuss the challenges our community needs to face, in the areas of sequencing technology, algorithm development and data sharing, to bring genomics up to clinical grade.

Promising applications of precision medicine

Cystic fibrosis. In the State of the Union address, President Obama specifically gave as an example the drug ivacaftor, which was developed for patients with cystic fibrosis. Cystic fibrosis is an autosomal recessive disease that affects approximately 70,000 people worldwide and that is caused by variants in the cystic fibrosis transmembrane conductance regulator (*CFTR*) gene. The protein product of this gene is an epithelial ion channel located on the cell surface where it regulates cellular chloride transit. Mutations of *CFTR* cause abnormal regulation of salt and water, which particularly affects the function of the lungs, pancreas and sweat glands. Recurrent pulmonary disease and resistant infection represent the major therapeutic challenges of cystic fibrosis, and traditional therapies have focused entirely

Center for Inherited Cardiovascular Disease, Falk Cardiovascular Research Building, Stanford Medicine, 870 Quarry Road, Stanford, California 94305, USA.
euan@stanford.edu

doi:10.1038/nrg.2016.86
Published online 16 Aug 2016

Box 1 | **Personalized medicine, precision medicine and precision health**

Semantically, precision and accuracy are distinct concepts. Precision reflects the extent to which repeated measurements are similar, whereas accuracy reflects the extent to which a given measurement reflects the truth. A common analogy is a target where precise but inaccurate shots cluster together away from the centre, whereas accurate but imprecise shots scatter widely around the centre. Although the US National Research Council explicitly includes the concepts of precision and accuracy in its definition of precision medicine[15], and we can paint both concepts in a genomic context, neither quite captures the essence of precision medicine as currently defined. The current definition — understanding disease at a deeper level in order to develop more targeted therapy — clearly requires genomic tools that are both accurate (the genome is represented faithfully) and precise (repeating the same test multiple times leads to the same result). Notably, the US National Research Council distinguished precision medicine from personalized medicine, which it defined as the situation in which therapeutics are synthesized for specific individuals[15]. However, most people probably believe personalized medicine instead to mean some degree of personalization that would incorporate, for example, pharmacogenomics-based tailoring of therapy, as well as the fruits of precision medicine approaches. Finally, precision medicine is increasingly recognized as synonymous with a technology-driven and participant-centred approach. A final extension includes the concept of precision health: using similar approaches for disease prevention and health promotion.

on the secondary consequences of the disease. Genetic understanding of cystic fibrosis has facilitated its categorization into molecular subgroups (FIG. 1b). In some subgroups, the channel reaches the cell surface but there is insufficient ensemble channel activity, but in other subgroups, trafficking leaves the channel in the cell cytoplasm. The oral agent ivacaftor was designed to increase the opening time of activated CFTR channels at the cell surface. Thus, for patients with mutant channels that do not reach the cell surface, ivacaftor would have minimal effect, whereas in patients with channels that are adequately transported, effect sizes for the improvement of pulmonary function could be dramatic. This was the case for the 5% of patients with the G551D mutation who were initially targeted[19,20]. A newer approach, which was recently approved by the US Food and Drug Administration (FDA), includes the use of ivacaftor and a second agent, lumacaftor, that improves the intracellular processing and delivery of the mutant channel[21]. This is particularly important for the 85% of patients with cystic fibrosis who have the most common genotype, F508del. For these patients, the mutant channel protein is misfolded, which leads to intracellular degradation. However, if there is proteasomal escape, the protein reaches the cell surface but with a gating abnormality that is similar to G551D. Thus, a combination approach may be optimal for these patients[21,22]. In this case, detailed understanding of the genetics of cystic fibrosis allows much more precise targeting of specific agents to individuals with specific functional defects.

Precision oncology. Another major area of promise for precision medicine is oncology. Traditional approaches to the classification of solid tumours focused on the tissue of origin. However, since the early success of the ABL1 kinase inhibitor imatinib for chronic myeloid leukaemia (which is driven by a BCR–ABL1 fusion protein), oncology has moved towards molecular classification. Crucial to this recognition of cancer as a genetic disease was the discovery of the central role of somatic mutation of genes that are involved in DNA repair, cell division and apoptosis. Genomic characterization has in fact been standard of care for some time for lung adenocarcinoma: testing for specific epidermal growth factor receptor (EGFR) mutations and anaplastic lymphoma receptor tyrosine kinase (ALK) rearrangements allows the personalization of therapy with targeted kinase inhibitors, such as gefitinib for EGFR and crizotinib for ALK[23,24]. Similarly, BRAF inhibition in BRAF-mutant melanoma[25] was a much heralded early application of precision targeting, but like many attempts to target 'driver' mutations with specific agents, the overall duration of response was disappointingly short owing to the acquisition of secondary resistance through additional somatic events.

A newer approach with the potential for longer term effects is the harnessing of the immune system[26] (FIG. 1c). Tumours present antigens in the form of oncogenic viruses, fetal developmental proteins or neoantigens that are formed by somatic mosaicism[27]. Initial attempts to harness T cell responses to such antigens through vaccination were disappointing but led to a greater appreciation of the importance of the antigen-presenting cell and co-stimulation with, for example, CD28. This led to the identification, not only of the critical steps required for T cell activation, but also of the autoinhibitory pathways mediated by the checkpoint receptors cytotoxic T lymphocyte-associated protein 4 (CTLA4), programmed cell death 1 (PD1; also known as PDCD1) and others. Antibodies to these proteins were rapidly developed and clinical trials of 'immune checkpoint therapy' found broad success in various tumours with, in some cases, prolonged effects[28]. It is speculated that the sustained effects of these therapies are mediated by memory T cells.

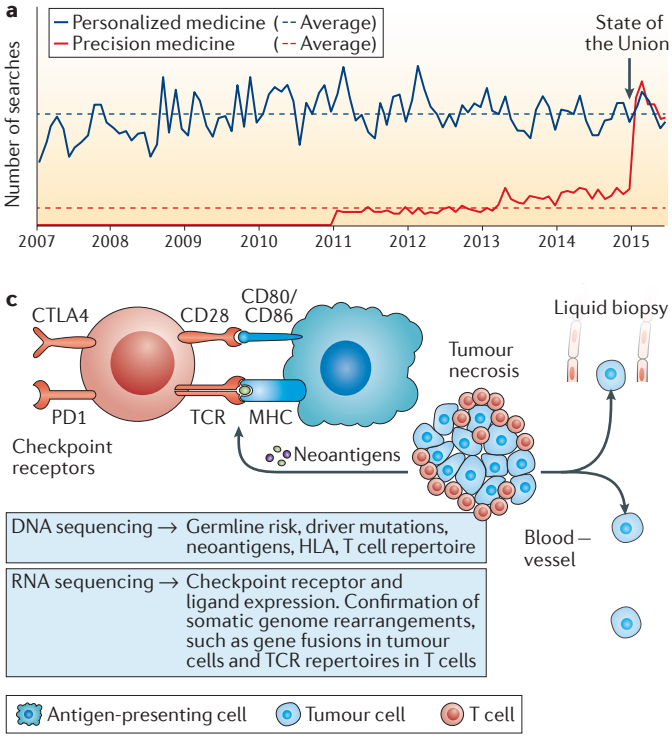
Most recently, trials combining genomic targeting with checkpoint therapy have begun[26]. In fact, genomic approaches, which have been greatly facilitated by resources such as The Cancer Genome Atlas (TCGA)[29], can also enable checkpoint targeting in other ways: RNA sequencing can confirm the expression of the checkpoint ligand in the tumour and the checkpoint receptor in the T cell. In addition, newer computational approaches to detecting neoantigens are beginning to show success[30]. Indeed, a seminal example of personalized tumour therapy is to combine a neoantigen-led vaccination strategy with the detection of circulating tumour cells and cell-free DNA from tumour cells in plasma[31,32].

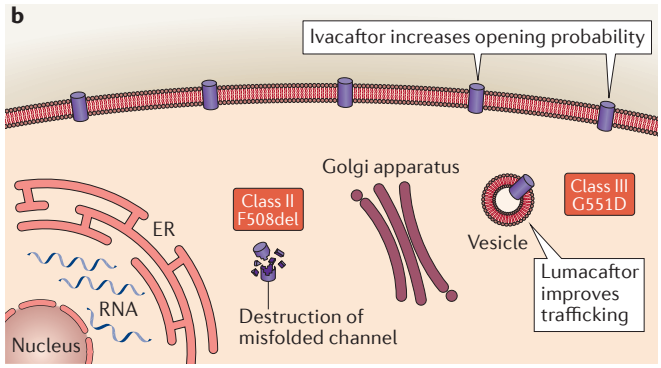
Pharmacogenomics. Pharmacogenomics was perhaps the earliest application of personalized medicine. Trials of genotyping *VKORC1* (which is involved in the biochemical activation of the blood clotting factor vitamin K) and *CYP2C9* (a member of the cytochrome p450 drug-metabolizing enzyme family) to optimize warfarin dosing led to some success, including approaches to automated dose estimation[33]. Indeed, the FDA embraced the possibility of such testing with black box warnings that encouraged the use of genetic testing where possible. However, some debate regarding cost-effectiveness[34,35] and the lack of readily available genomic information on large numbers of patients, left these potentially valuable

Checkpoint receptors
Mediate important immune autoinhibitory pathways, including programmed cell death 1 (PD1) and cytotoxic T lymphocyte-associated protein 4 (CTLA4).

Pharmacogenomics
The study and application of the effect of genetic variation on the response to pharmaceuticals.

Black box warnings
Named for the black border surrounding the text of the warning on the package insert or label of a drug. They detail the safety concerns that are of a more serious nature than those described elsewhere on the package or label. The border is used when a serious adverse event can be caused by the medication or can be prevented by appropriate use of the medication.





Subcategory	Functional effect	Example variant
I	No functional protein	G542x
II	Trafficking defect	F508del
III	Defective regulation	G551D
IV	Decreased conductance	R117H
V	Reduced synthesis	3120+1G>A
VI	Reduced stability	Q1412x

Figure 1 | **The emergence of precision medicine. a** | Worldwide searches using the Google search engine for the terms 'personalized medicine' and 'precision medicine' from January 2007 to June 2015. **b** | Precision medicine in cystic fibrosis. Subclasses of cystic fibrosis are defined according to the functional effects of specific genetic variants on the cystic fibrosis transmembrane conductance regulator (CFTR) channel. Six subclasses are defined. The first drug approved for a subclass of cystic fibrosis was ivacaftor, which increases the opening probability of channels on the cell surface. It was initially approved for patients with Class III cystic fibrosis (G551D patients), for which the trafficking of CFTR to the cell surface is intact but the major defect is in regulation. The most common variant, F508del, results in the destruction of a misfolded channel in the cytoplasm (Class II). For this variant, a combination of lumacaftor (to enhance intracellular processing and trafficking) and ivacaftor may be optimal. **c** | Precision oncology. Tumours are attacked by T cells. Tumour products, including neoantigens, are presented by major histocompatibility complex (MHC) molecules and bound by T cell receptors (TCRs). Antigen-presenting cells and co-stimulators such as CD28 lead to T cell activation and proliferation. Autoinhibition is mediated by checkpoint receptors such as cytotoxic T lymphocyte-associated protein 4 (CTLA4) and programmed cell death 1 (PD1). Genomic approaches to precision medicine include DNA sequencing of the germline for risk and typing of HLA genes (which encode MHC proteins), of the T cells to quantify T cell receptor (TCR) repertoires, and of the tumour for driver mutations and neoantigen prediction. RNA sequencing confirms the expression of checkpoint receptors and ligands in the tumour and the adjacent infiltrating T cells, and also confirms gene fusion events in the tumour and TCR repertoires in T cells. ER, endoplasmic reticulum.

tools in the hands of only a small number of clinics while pharmaceutical companies worked to develop drugs with alternative pharmacokinetics that did not require companion diagnostics. A similar situation emerged for clopidogrel, which is an anti-platelet agent used to prevent coronary artery stent thrombosis. A common loss-of-function polymorphism in CYP2C19 (*2), which is present in up to 35% of individuals of European and African ancestry and 60% of individuals of Asian ancestry, is associated with the reduced conversion of the pro-drug to the active metabolite[7]. Large studies showed adverse outcomes in poor metabolizers following coronary stent placement procedures[36,37] but other studies in different contexts did not show major effects on outcomes[38]. This was a confusing message for the cardiovascular community[39,40], and despite the development of point-of-care diagnostic monitoring[41] and a recommendation in the form of a black box warning from the FDA, the presence of platelet activation assays and newer agents that are not metabolized by this pathway[42] led to limited use.

However, the promise of pharmacogenomics remains very great as it could apply to every individual taking any medication. Indeed, there have been some estimates that 98% of people carry a high-risk pharmacogenomic diplotype[43]. Catalysed by carefully curated knowledge bases such as the Pharmacogenomics Knowledgebase (PharmGKB)[44,45], professional guidelines already detail many potential uses[46]. However, for pharmacogenomics to succeed broadly, genotype information that is relevant to drug metabolism needs to be available at the time of prescribing, which usually means a priori genotyping. Several major centres have deployed systems to enable this genotyping[47].

Companion diagnostics
Diagnostic tests that help to direct the appropriateness of a specific drug therapy.

In summary, applications of genomics to genetic diseases such as cystic fibrosis and cancer, as well as for pharmacogenomics sit within a broader landscape of promise for the application of genomics to medicine (TABLE 1). Applications that are further from routine medical application, such as microbiome sequencing and predictive analytics for common variants in complex disease, as well as some targeted approaches already in clinical practice, including non-invasive prenatal testing, are not discussed in this Review.

The US Precision Medicine Initiative

One of the central features of the US Precision Medicine Initiative is the establishment of a 1-million-person cohort of individuals willing to contribute their partnership and data for scientific discovery[13,14]. This was not, in fact, the first time either President Obama or the Director of the US National Institutes of Health, Francis Collins, had suggested such an idea. As Senator for Illinois, USA, Barack Obama introduced the Genomics and Personalized Medicine Act of 2006 (REF. 48) that included planning for a national biobanking initiative. Meanwhile, Francis Collins had called for a large-scale prospective cohort study of genes and environment as early as 2004 (REF. 49) to mirror those of the United Kingdom[50], Iceland, Denmark, Canada, Germany and others. In particular, Collins pointed out the advantages of studying the natural history of disease. One feature that seems particularly prominent in the planning of the US Precision Medicine Initiative

Table 1 | **Examples of precision medicine**

Condition	Gene	Action
Mendelian disease		
Cystic fibrosis	*CFTR*	Specific therapies such as ivacaftor and a combination of lumacaftor and ivacaftor
Long QT syndrome	*KCNQ1*, *KCNH2* and *SCN5A*	Specific therapy for patients with *SCN5A* mutations
Duchenne muscular dystrophy	*DMD*	Ongoing phase III clinical trials of exon-skipping therapies
Malignant hyperthermia susceptibility	*RYR1*	Avoid volatile anaesthetic agents; avoid extremes of heat
Familial hypercholesterolaemia (FH)	*PCSK9*, *APOB* and *LDLR*	• Heterozygous FH (HeFH): eligible for PCSK9 inhibitor drugs • Homozygous FH (HoFH): eligible for PCSK9 inhibitor drugs in addition to lomitapide and mipomersen
Dopa-responsive dystonia	*SPR*	Therapy with dopamine precursor L-dopa and the serotonin precursor 5-hydroxytryptophan
Thoracic aortic aneurysm	*SMAD3*, *ACTA2*, *TGFBR1*, *TGFBR2* and *FBN1*	Customization of surgical thresholds based on patient genotype
Left ventricular hypertrophy	*MYH7*, *MYBPC3*, *GLA* and *TTR*	Sarcomeric cardiomyopathy, Fabry disease and transthyretin cardiac amyloid disease have specific therapies
Precision oncology		
Lung adenocarcinoma	*EGFR* and *ALK*	Targeted kinase inhibitors, such as gefitinib and crizotinib
Breast cancer	*HER2*	HER2 (also known as ERBB2)-targeted treatment, such as trastuzumab and pertuzumab
Gastrointestinal stromal tumour	*KIT*	Targeted KIT kinase activity inhibitors, such as imatinib
Melanoma	*BRAF*	BRAF inhibitors, such as vemurafenib and dabrafenib
Pharmacogenomics		
Warfarin sensitivity	*CYP2C9* and *VKORC1*	Adjust dosage of warfarin or consider alternative anticoagulant
Clopidogrel sensitivity, post-stent procedure	*CYP2C19*	Consider alternative antiplatelet therapy (for example, prasugrel or ticagrelor)
Thiopurine sensitivity	*TPMT*	Reduce thiopurine dosage or consider alternative agent
Codeine sensitivity	*CYP2D6*	Avoid use of codeine; consider alternatives such as morphine and non-opioid analgesics
Simvastatin sensitivity	*SLCO1B1*	Reduce dose of simvastatin or consider an alternative statin; consider routine creatine kinase surveillance

ACTA2, actin, alpha 2, smooth muscle, aorta; *ALK*, anaplastic lymphoma receptor tyrosine kinase; *APOB*, apolipoprotein B; *CFTR*, cystic fibrosis transmembrane conductance regulator; CYP2, cytochrome P450 family 2; *EGFR*, epidermal growth factor receptor; *FBN1*, fibrillin 1; *GLA*, galactosidase alpha; KCN, potassium voltage-gated channel; *LDLR*, low-density lipoprotein receptor; *MYBPC3*, myosin-binding protein C, cardiac; *MYH7*, myosin heavy chain 7; *SCN5A*, sodium voltage-gated channel alpha subunit 5; *PCSK9*, proprotein convertase subtilisin/kexin type 9; *RYR1*, ryanodine receptor 1; *SLCO1B1*, solute carrier organic anion transporter family member 1B1; *SMAD3*, SMAD family member 3; *SPR*, sepiapterin reductase; *TGFBR*, transforming growth factor beta receptor; *TPMT*, thiopurine S-methyltransferase; *TTR*, transthyretin; *VKORC1*, vitamin K epoxide reductase complex subunit 1.

was the idea of including participants as partners and connecting participants and researchers via mobile technology devices. Such devices could be used for more sophisticated phenotyping or to monitor large populations at risk for disease[51].

The convergence of discovery and clinical genetics

Human discovery genetics and clinical genetics began together with family pedigrees and descriptions of inheritance in the absence of knowledge of the molecular cause. The advent of increasingly dense genome markers facilitated the first examples of forward genetics: positional cloning by linkage analysis of pedigrees followed by the discovery of causative genes and variants in those linkage regions. Fuelled by the HapMap project[52], the characterization of common variation at a genome-wide scale became possible when hundreds of thousands of markers could be simultaneously analysed on microarray platforms. When next-generation sequencing first became tractable, it was applied as low-coverage sequencing in large populations for single nucleotide variant (SNV) discovery (for example, the 1000 Genomes Project)[53]. These approaches were successful in the discovery of robustly replicable associations between traits and SNVs of small effect in mostly non-coding regions of the genome[54].

Meanwhile, in clinical medicine, diagnostic testing has historically focused on karyotyping to detect chromosomal abnormalities or fluorescence *in situ* hybridization for large-scale rearrangements. The association of genes to diseases and the facilitation of knowledge through curation in databases such as Online Mendelian Inheritance in Man (OMIM) led to an era of Sanger sequencing of the coding regions of small numbers of genes. If a rare or disrupting variant was not found in a control group of typically 100 Caucasian blood donors, it was deemed important. Meanwhile, crossing over from the discovery world, the microarray was the first high-throughput technology to truly affect medical genetics, offering the detection of deletions and duplications at increased resolution[55] and leading to the possibility of a genome-wide test that could be used for undiagnosed disease where no single candidate gene was identified[56,57]. In a similar manner, laboratories have extended gene panels using next-generation sequencing approaches to include many more genes, even including some for which gene–disease causality is less well established.

History, then, draws an interesting contrast between a clinical genetic testing community that was focused on large-scale disruptions to the open reading frame of genes, and an emerging population genetics community, who first defined themselves through genome-wide common SNV associations with complex disease. The excitement of the present era of precision medicine is driven by their convergence. This convergence was exemplified by a NHGRI-sponsored workshop that brought together clinical geneticists, population geneticists, genetic epidemiologists and statistical geneticists to agree on a framework for the determination of causality for sequence variants in human disease[58].

Making genomics more precise and accurate

Implicit in the term precision is an approach to genomics that includes accuracy. Although the formal definitions of precision and accuracy are distinct (BOX 1), the use of the term by the US National Research Council panel was intended to convey both meanings[15]. Semantics aside, there is clearly nothing more important to precision medicine than accurately representing the genomes of individual patients or their tumours[59] (FIG. 2). Key challenges to the attainment of accuracy in genomic medicine are described below along with their medical relevance and possible solutions.

Achieving accuracy: anatomy of the genome. The human genome has historically been defined by the reference sequence. The product of the publicly funded human genome project, the human reference was derived from the DNA of more than 50 individuals from whom clones representing single haplotypes were sequenced by a shotgun approach and then patched together in one haploid sequence[60]. Although the largest contribution probably came from one African American individual, this was an ethnically diverse group and so the reference genome switches from one ethnic haplotype to another at multiple places.

The newest human reference assembly GRCh38 was the result of many years of meticulous work from the Genome Reference Consortium. It adds 178 regions with 261 alternative loci[60] and 150 genes that were not previously represented. The genome itself (GRCh38.p5)[61] is 3.23 billion bases with (GRCh38.2) 51,087 genes and pseudogenes (of which 20,576 are protein-coding genes, although some algorithms estimate this may be as low as 19,000 (REF. 62)). The genes vary enormously in size from 8 base pairs (a transfer RNA) to 2,473,559 base pairs (the *CNTNAP2* gene encoding the CASPR2 protein). The genes may have as few as 1 exon (for example, a gene encoding a G-protein-coupled receptor) or as many as 363 exons (titin). In the original assembly, there was 198 Mb of heterochromatin gaps and 28 Mb of euchromatin gaps[63]. The GC richness of the genome, important as a challenge to DNA capture and sequencing chemistry, varies dramatically: first exons generally have a higher GC content than the overall average of around 40%[63]. The functional importance of GC-rich regions is driven by CpG motifs. These are thought to be particularly sensitive to mutation and are clustered in islands near the 5′ end of genes.

A major challenge to the accurate representation of the genome takes the form of repeated sequence, which represents more than 50% of the genome[64–67]. Common types of repeats include segmental duplications, simple repeats, short tandem repeats (recently shown to have an important role in gene expression[68]), transposon-derived repeats and processed pseudogenes.

Genome anatomy: challenges to clinical diagnosis. Much of this genomic complexity is only challenging because of the prevailing technology used to assess it: short-read sequencing. With extensive paralogy, originating in gene families, segmental duplication or pseudogenes, the genomic location of many short reads

Linkage analysis
An approach to establish the probability that a given genomic region is associated with a phenotype, usually in an extended pedigree.

HapMap project
An international consortium aimed at characterizing the haplotype diversity of the human genome.

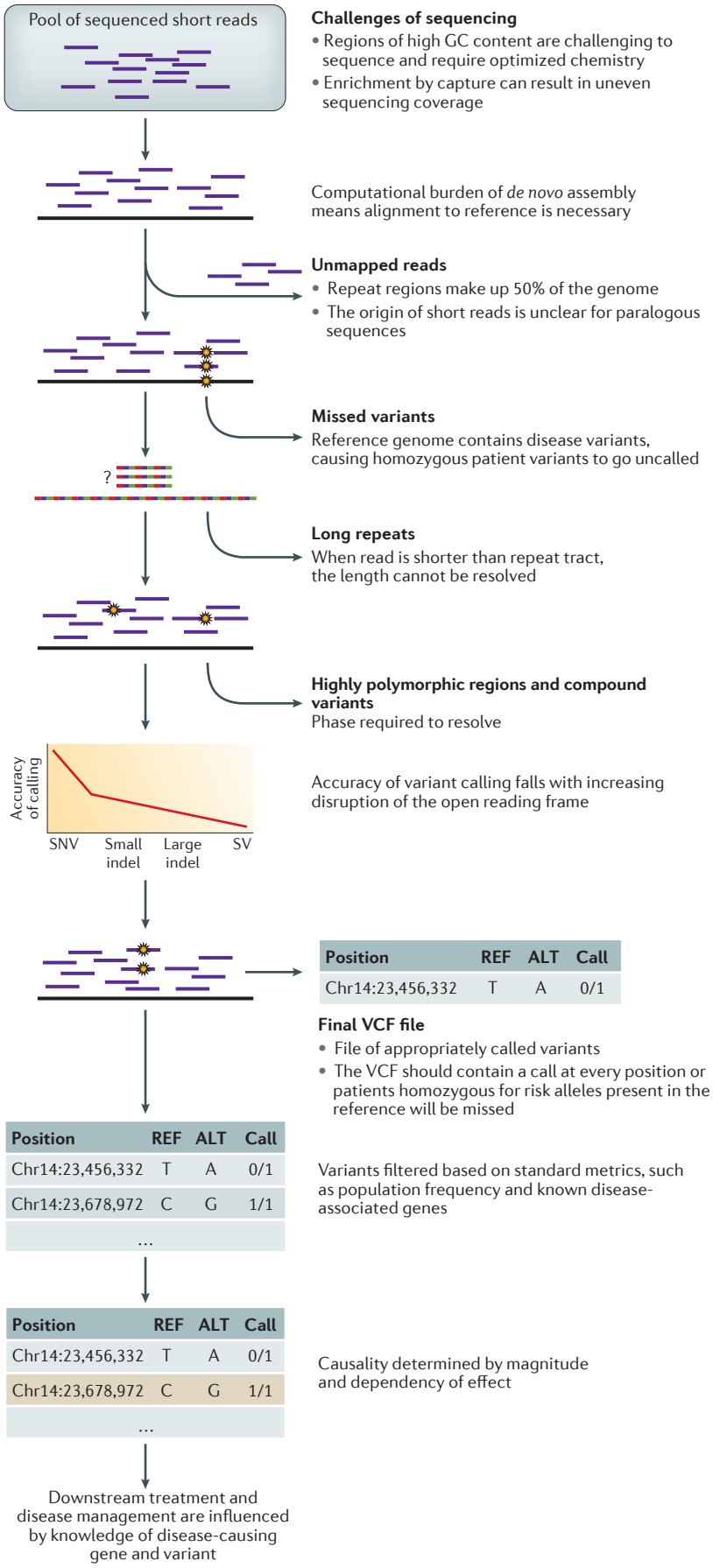
Shotgun
In shotgun sequencing, longer DNA fragments are broken into smaller fragments for sequencing using chain termination (Sanger) chemistry.

Pseudogenes
Copies of a gene that are no longer functional in the same way as the original gene, usually because of deactivating mutations, such as premature stop codons. Pseudogenes can be either processed (derived from retrotransposition of a mature transcript) or non-processed (derived from a DNA duplication event that includes a modification leading to a loss of transcription or translation).

Segmental duplications
Typically pericentromeric or subtelomeric duplications, concentrated in the Y chromosome, generally tens to hundreds of kilobases in length.

Short tandem repeats
Microsatellite DNA motifs consisting of 2–6 bp repeated elements of median length 25 bp and accounting for 1% of the genome. They predispose to DNA polymerase slippage events and high mutation rates. Recent work suggests an important role in gene expression.

Transposon-derived repeats
Repeats derived from transposons, which are DNA elements that can change their positions within the genome.



cannot be determined with confidence[67]. With simple repeats, the challenge is different. If the overall length of the repeat region is shorter than the read length, it is possible to resolve length by local re-assembly. However, if the repeat tract is longer than the read length, the length of the repeat region is very challenging to discern. Yet, important genetic diseases are encoded by simple repeats that expand because of the instability of the resulting secondary structures during replication. Indeed, most repeat tracts are pathogenic in a range greater than the typical size of a short read (100–250 base pairs)[69,70] (FIG. 3a). For example, in Huntington disease, the risk begins at 40 trinucleotide CAG repeats (120 base pairs) in *HTT* and increases from there.

Highly polymorphic regions also cause major challenges for short-read sequencing. The prototypical region is the major histocompatibility complex (MHC) that encodes human leukocyte antigens (HLAs). This is a 3.6 Mb segment on chromosome 6p21 that contains more than 100 genes of which six are the basis of the most commonly reported immune typing. The HLA region is fundamental for our definition of self and is associated with more than 100 diseases and many drug reactions, including some that are potentially fatal — for example, carbamazepine-associated toxic epidermal

Figure 2 | **Origins of reduced accuracy in clinical genomics from short sequencing reads.** Accuracy can be optimized at multiple steps in the route from DNA to variant calling and reporting. Regions of high GC content require tailored approaches both for capture and for sequencing. Enrichment by capture leads to uneven coverage. Alignment to the haploid reference sequence is required for short reads because of the computational burden of *de novo* assembly. Paralogous sequence is common throughout the genome, and the origin of a short read cannot be determined in 5% of cases. For diseases such as Huntington disease that are caused by repeat tracts where the most severe disease is associated with tracts longer than the short reads, length cannot be resolved. Similarly, highly polymorphic regions such as the major histocompatibility complex (MHC), which is used for HLA typing in transplantation and for risk quantification for multiple immune diseases, cannot be resolved. Along with compound variants such as multiple nucleotide variants, these cannot be adequately resolved without phasing. The accuracy of calling decreases with increasing disruption of the open reading frame. However, variants that are more disruptive of the open reading frame, such as structural variants (SVs), are generally more likely to cause disease. As the human reference sequence is made up of DNA from multiple individuals — and contains risk variants that reduce the accuracy of alignment and result in missed calls for homozygous risk alleles such as factor V Leiden — a call at every position should be included in the variant call file. Finally, causality is a complex construct with final effects determined by magnitude and dependency of the variant effect. Causal variants can often lead to changes in clinical management: in some cases, precision therapy for the patient and in other cases changes in screening are recommended for the family. ALT, alternative allele; indel, small insertion or deletion; REF, reference allele; SNV, single nucleotide variant; VCF, variant call format.

Paralogy
A paralogue is a gene related to another by duplication. In this Review, the words paralogy and paralogous are used as umbrella terms for areas of the human genome that are identical to each other. Note that paralogues can be formally distinguished from homologues (genes related to one another by descent from a common ancestor) and orthologues (genes related to one another by speciation).

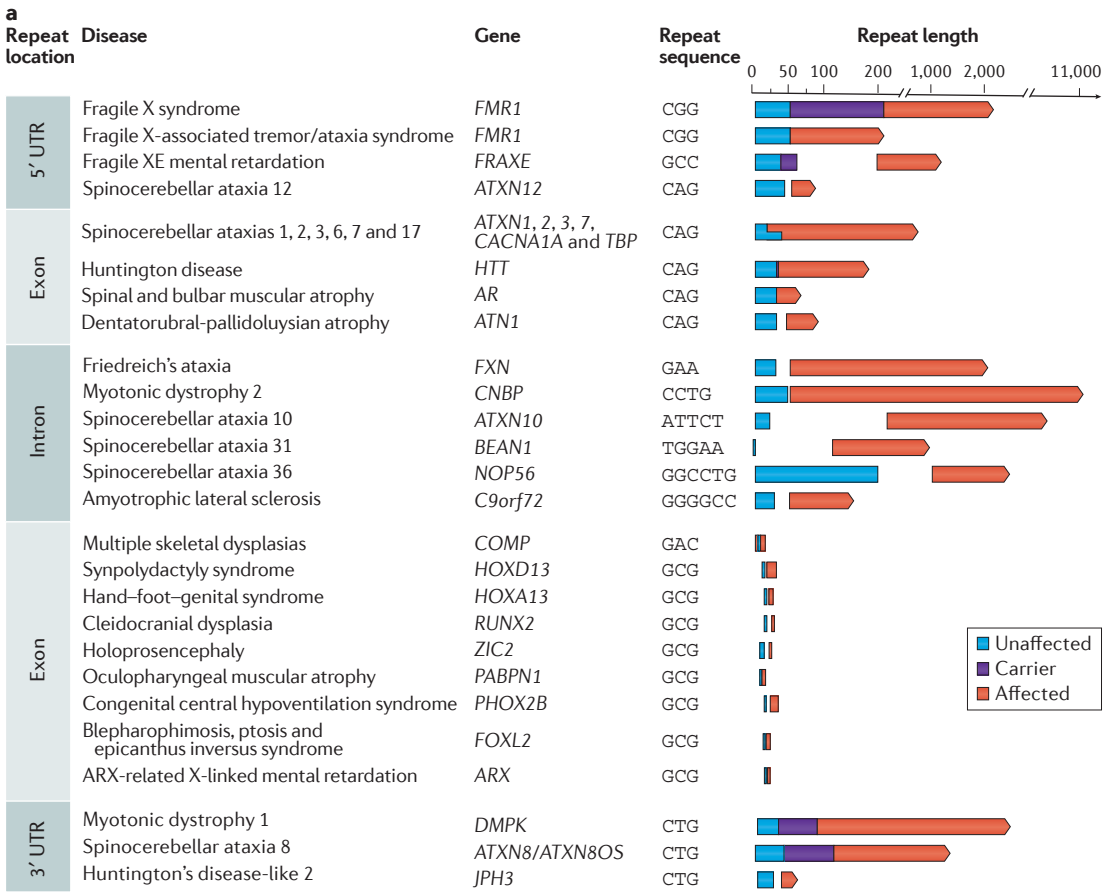


Figure 3 | **Repeats, compound variants and nomenclature as challenges to accuracy in clinical genome sequencing.** **a** | Diseases that are caused by unstable repeats. Repeats are thought to expand due to instability of resultant DNA secondary structures during replication. As shown, most repeat tracts are pathogenic in a range beginning at the typical size of a short read (100 base pairs), meaning that the most severely affected patients will not be diagnosed. **b** | Multiple nucleotide variants. If two variants are found at consecutive positions, standard approaches to genotype calling without phase do not resolve the appropriate amino acid sequences. A particularly important example would be where variants create a new start or stop codon only when in *cis*. **c** | Left and right justification of small insertions or deletions (indels). Confusion arises when different schema are used for locating indels. In the 1000 Genomes Project samples, the F508del variant is left justified with respect to the genome. In clinical practice, variants are right justified with respect to the transcript. As the transcript can be derived from either strand, even unifying to left or right justification would not be enough to solve the problem, which requires manual curation. Part **a** is adapted with permission from REF. 70, Annual Reviews.

necrolysis[71,72] and abacavir-induced liver injury[73]. However, the MHC is challenging to resolve using only short-read approaches because of the lack of a comprehensive catalogue of haplotypes and the intrinsic lack of phase information — that is, knowledge of the parental chromosome of origin — in short reads[74].

This lack of phase information is challenging in other clinically relevant situations, for example, the demonstration of compound heterozygosity, where two variants are found in the same gene. Knowing whether a mixture of doubly mutated and wild-type protein is expressed versus whether two singly mutated proteins are expressed is a critical distinction[75]. This is important in pharmacogenomics for which current standard practice is that combinations of variants mapped from association study evidence are assumed to be in *trans*. A related example of compound heterozygosity that can be resolved with a simple algorithm is the multiple nucleotide variant (MNV) where two variants appear at consecutive positions (FIG. 3b) — understanding the consequences for protein coding from each gene copy requires a variant-calling algorithm that distinguishes phase. MNVs seem to be particularly frequent in cancer[76].

Long-read sequencing approaches, which involve either barcoding fragments of longer molecules for short-read sequencing and subsequent *in silico* reassembly[77,78] or direct sequencing of the longer molecule[79], can theoretically provide answers to many of these currently unsolved challenges[80]. Long-read sequencing facilitates *de novo* assembly that automatically provides phase information[81–83]. It improves the likelihood that any given structural variant will be sequenced with a localizing non-duplicated region. Tracts of simple repeats that are even thousands of base pairs long can theoretically be captured. Length sizes for the currently available long-read technologies now have their median in the 5–10 kb range, with a long tail reaching to tens of thousands of base pairs or more[84,85], while short-read reconstruction approaches can have median haplotype blocks as large as several megabases[78]. In addition, such sequencing provides a more complete picture of the genome. Recently, interstitial euchromatic gaps with the human reference genome (GRCh37) were closed by a long-read sequencing method[79]. These gaps were identified as predominantly long runs of short tandem repeats embedded within GC-rich regions. In a second approach, combining two long-read technologies improved the length of assembly scaffold and structural variant detection. Such haploid[79] or diploid[86] approaches demonstrate previously unrecognized genomic complexity, particularly in structural variation. Unfortunately, however, long-read approaches remain between one and two orders of magnitude more expensive than short-read approaches[80] and also require larger amounts of DNA, partly to overcome a high error rate, delaying their widespread adoption for human genome sequencing.

Quality scores and compression. As more and more individuals are sequenced as part of clinical medicine, there will be an increasing need for long-term storage and retrieval of their data. Indeed, some have estimated that the data size of genomics will surpass that of online video and particle physics[87], making this a major challenge for precision medicine. Some methods encode differences from a reference sequence, while others focus on quality scores[88]. Each base that is called by a sequencing machine has an associated quality score. Traditionally, these values have been reported in a format known as 'Phred', which was originally derived from chromatogram traces of early sequencers[89,90]. The number is expressed as the negative log 10 of the probability of error: $q = -10\log_{10}(p)$.

The most common cut off is Q20, which corresponds to a 1-in-100 chance that a base call is incorrect. Notably, this score is calibrated by each sequencing vendor according to internal protocols. The scores, however, represent a large amount of data in an alignment file. And this provides an opportunity for compression. In fact, several approaches to the compression of genomic data, both lossy compression and lossless compression, have been proposed[91–93]. Some have even experimented with mapping these scores to a single byte (8 bits). Although compression of the reads themselves provides modest gains, compression of the quality scores offers much greater potential.

Alignment and assembly. The output from sequencing is a large text file of short or long reads along with their quality scores. Deriving a complete picture of a single human's genome from these 'raw' reads requires assembly and comparison with a reference genome. Typically, short reads are aligned to a reference genome using an algorithm that searches for the best match. First principles might suggest advantages to *de novo* assembly (assembling the genome by overlapping the reads without the aid of a reference sequence) using methods such as the De Bruijn or string graph[83,94]. However, *de novo* assembly, particularly of short reads, is computationally intense and impractical for clinical genome sequencing[79]. Currently, the vast majority of human exome and genome sequences are aligned to a reference sequence. The reference sequence itself has been the focus of some concern because it was derived from a pool of individuals and, as such, contains risk variants. In addition, it does not accurately represent longer range haplotypes owing to the switching between reference individuals in some regions[95]. Mapping quality will also be poorer in regions of variation[95].

Several algorithmic approaches to optimal alignment exist. One approach takes advantage of dynamic programming to yield an exact match for pairwise local or global alignment. It involves the generation of a similarity matrix of two sequences where a score or penalty is awarded for match or mismatch followed by a traceback step that identifies the highest scoring matrix cell. This was originally proposed by Needleman and Wunsch[96] and was later adapted for local alignment by Waterman and Smith[97]. The approach is computationally expensive but maximizes the sensitivity and specificity of downstream variant calling, especially with respect to gapped alignment[98]. Several methods for speeding up these algorithms have been recommended; for example, using graphics-processing units[99].

De novo assembly
Arranging DNA sequence reads in the most likely order of origination without alignment to a reference sequence.

Structural variant
A region of DNA usually greater than 500 bases variant from a defined reference.

Lossy compression
A class of data encoding that reduces data size for storage, handling and transmission at the expense of loss of content.

Lossless compression
A class of data encoding where the original can be perfectly restored from the compressed file.

Despite these advantages, a compression heuristic became the most commonly applied approach to alignment for human genomes[100,101]. This approach is based on a variant of the suffix array[102], which is an approach to the representation of sub-strings in a format that is efficient for searching and compression. Similar to many compression approaches, the Burrows–Wheeler transform (BWT) aims to group similar letters, sorting them lexicographically then storing the letter and the number of times it is repeated before changing. Importantly, what this approach offers over a simple sort is the possibility of inversion. That is, the original sequence can be recreated from the compressed output. For both compression and alignment use cases, reversing the transform is crucial. Notably, although much more rapid in a cohort discovery context, the BWT is less optimal in a single-patient ('$N=1$') clinical context. Although certain contexts demand speed[103,104], in most cases accuracy is primary for clinical genomics and an exact match global alignment generally performs better[105].

Even with an exact match algorithm, a major challenge for short-read sequencing arises when a read maps to more than one place. The read could be placed at the best aligning position, or, if it aligns equally well to more than one position, it could be placed at a randomly chosen position, at every position or not placed at all. Remarkably, there is no consensus regarding which of these placements is best, and different algorithms adopt different approaches with some allowing this placement to be specified in the command line. Clearly, the longer the read the less likely this issue of placement will be a problem, but for 100 bp reads, fully 5% of the genome will originate non-unique reads[67]. Given that a typical whole genome sequenced to 30× coverage generates approximately 1.3 billion reads, this represents 65 million reads that have no possibility of being accurately located[106,107]. In practice, it is typically closer to 10% of reads in a whole-genome sequence alignment that remain unplaced, meaning that a further 65 million reads are lost that will probably be enriched for paralogous areas under variable evolutionary constraint (for example, gene families or pseudogenes) or places where the genome being tested differs from the reference genome in ways more complex than single nucleotide variation. Unaligned reads may also represent non-human DNA, in which case, new approaches to the diagnosis of infectious diseases can take advantage by mapping these to databases of viruses and bacterial organisms.

Variant calling. After assembly or alignment comes variant calling. The most common approach is to compare the most likely genotype at each position to that of a standardized reference sequence. This is usually the most current version of the human genome reference but in tumour sequencing might be the patient's germline sequence. Notably, the human reference sequence is haploid. Thus, a homozygous disease-risk variant in a clinical genome sequence will not be called if it also occurs in the haploid reference sequence[95]. In the case of the factor V Leiden variant found in the reference, for example, a person with an up to 80-fold risk of thromboembolic disease would be undetected by the analysis of any standard variant call format (VCF) file[108]. Some solutions to this issue take the form of ethnicity-specific, major allele reference sequences[95] and family-based diploid reference approaches[56]. The move towards graph-based assembly approaches, in which the sequence and population variation are contained within a single structure, is underway[74]. Another solution involves calling all known risk-associated positions[8,105] or calling every position into a genomic variant call file, including both reference and variant calls: gVCF (FIG. 2). This has the advantage of distinguishing between a 'no call' and a 'homozygous reference call', which is unable to be distinguished using the standard approach. The challenge with calling every position is the loss of the advantageous drop in file size from raw data to variant call file (from ~five orders of magnitude drop to only ~two orders of magnitude drop).

Different classes of variation have widely varying call accuracy and reproducibility[109], which is something made more challenging by the lack, until recently, of a fully characterized single human's diploid genome. In its place, the NA12878 genome available in cell lines from Coriell has been adopted, led by a consortium from the US National Institute of Standards and Technology that is called Genome In A Bottle[67,110]. The consortium made a consensus call set freely available that was derived from 14 data sets from five sequencing technologies, including seven read mappers and three variant callers. The initial work demonstrated a lack of concordance across different technologies but a clear theme emerged, which was also reflected in work with a more clinical focus[8], that the accuracy of calling varied widely across different variant classes.

Single nucleotide variation. Overall, single nucleotide variation is called with high sensitivity and specificity for approximately 77% of the genome, approaching 99% concordance[8] with genotyping microarray-based approaches in those regions[110]. This is nevertheless encouraging not only because important Mendelian disease is encoded by this class of variation but also because of the ever-expanding genome-wide association study evidence of single nucleotide variation that is confidently associated with complex human disease. Notably, common single nucleotide variation associations remain overall less relevant from a clinical perspective, as there is currently only very limited evidence of clinical utility in predictive scores derived from common variation[111].

Insertions and deletions. In contrast to single nucleotide variation, calling of small insertions or deletions (indels) is less accurate. In one study, the concordance between two platforms for indel calling was only 57% across the genome and 33% for inherited disease risk genes[8]. This is particularly concerning for clinical genomics, as variation that disrupts the reading frame or that affects the structure of the protein in a major way is likely to be more clinically important. A further challenge

Compression heuristic
An approach to compression that is not designed to be optimal but is rather designed to be practical.

Variant call format
(VCF). A file format standard for the cataloguing of genetic variation in one or many genomes.

Major allele
The most common allele in a given population.

Mendelian disease
A genetic disease that follows traditionally recognized patterns of simple inheritance, for example, autosomal dominant.

to the appropriate identification of indels is the lack of standardization of nomenclature (FIG. 3c). The customary approach to naming genetic variants in the clinical domain is known as HGVS (from the Human Genome Variation Society) and relates the variant position relative to the gene rather than to the chromosome as is more common in discovery genetics[112]. Parsers now exist to map such variants to the more commonly used chromosome location[113,114] but this does not resolve all the issues. Although with single nucleotide variation there can be challenges in appropriate localization given its dependency on alignment and transcript diversity, with insertions or deletions the challenge is greater. Specifically, the locating coordinate could be left or right justified. This is not a theoretical problem, but rather one with very clear clinical implications. For example, the F508del variation in CFTR (discussed above) is the most common variant that is causative of cystic fibrosis (FIG. 1b) but it is represented in two different ways. HGVS convention requires right shifting or justification of ambiguous indel variants for reporting relative to the transcript (the most 3′ position possible should be assigned). However, when calling variants on the genome from aligned sequences, the convention for genomic reporting of ambiguous indel variants in VCF is to left shift or justify relative to the published reference sequence, which represents one, arbitrarily chosen, strand. Because transcripts can be notated in either direction (that is, on either strand), unifying the justification to the left or right would still lead to discordance approximately 50% of the time (FIG. 3c). Careful manual curation is currently the only approach that can resolve these issues. Notably, this error was recently reconciled by ClinVar (but not by dbSNP).

At their most fundamental, algorithms for calling indels remain inferior to those for calling SNVs. Dindel was widely adopted, including into the Genome Analysis ToolKit (GATK) framework[115], and local *de novo* assembly approaches as well as use of 'known' indel positions improved this further. However, sensitivity still drops very rapidly to below 50% in simulated genomes as the size of the indel increases, even above three base pairs. Newer approaches[116–119] show substantial improvements by including prior knowledge of existing indels and by the use of local *de novo* re-assembly. However, a great deal of work needs to be done before this class of variation can confidently be called for clinical purposes. Although false-positive indel calls may be resolved by validation with alternative approaches, false-negative calls remain a considerable concern for precision medicine because, if a convincingly causal disease-associated variant remains undetected, this represents a missed opportunity for diagnosis and intervention.

Structural variation. In discovery projects, structural variation has been detected through microarrays and sequencing, but algorithms to detect structural variation from short-read sequencing are fundamentally limited by the length of the short read. Indeed, the extent to which the discovery of structural variation has been missed has been illustrated by recent long-read sequencing approaches that have revealed novel variation that was previously undetected by short-read technologies[120]. In many ways, this is not surprising given that the aim of the Human Genome Project was to produce one sequence and the fact that long-read data on even one human genome have only recently become available[79]. However, the detection of structural variation remains a high priority for precision medicine because it is a especially important class of variation, particularly for neurodevelopmental disorders[121]. Current clinically deployed microarray approaches are limited by the distance between markers, by the lack of adequate control populations, and by the sensitivity of the detection technology (fluorescence). Improved algorithms for calling structural variants from genome data are a major prerequisite for the advancement of precision medicine, particularly as sequencing brings some intrinsic benefits, such as the ability to detect copy-number-neutral structural variants including balanced translocations. Approaches to maximize the diagnostic yield for structural variation from sequencing data have existed for some time[122] but have only recently been improved and rigorously tested[123,124].

Sequencing gene panels, exomes or genomes

Medical diagnostics to take advantage of next-generation sequencing can have various strategies that differ in the proportion of the genome that they interrogate (FIG. 4). These are discussed below and include: the capture of the coding regions of a limited panel of genes (often between ten and 100 genes); the capture of the coding regions of almost all genes (the exome); or whole-genome sequencing (sequencing all of the genome that is accessible to short-read sequencing).

Capture-based interrogation of gene panels and exomes. The enrichment of selected areas of the genome by hybridization to known sequences is known as capture. Capture was initially developed for the research market and, in the case of the exome, was designed to balance genome-wide coverage with commercial viability[2,3] (FIG. 4a,b). Coverage metrics for these products were typically quoted as a mean or median (for example, 100-fold coverage) but this average greatly belied the vast differences in coverage in different regions (FIG. 4b). Indeed, certain exons in medically important genes (for example, potassium voltage-gated channel subfamily H member 2 (*KCNH2*)) were effectively missing. In addition, capture oligonucleotides were designed to bait sequences that exactly matched the human reference assembly, so they captured the regions of the genome that we care most about less efficiently: those parts that are variant.

Although in a research context these issues reduced the power to detect variation in certain areas, they did not impede the overall goal of finding some important new variations and so the total incremental benefit over microarray-based methods for discovery in the coding regions overshadowed any major concerns.

By contrast, use in the clinical world is for a single patient with a potentially devastating medical problem. In this case, missing any region of a gene could have

Parsers
An algorithm with a specific application in translating one terminology to another.

ClinVar
A curated database of clinically relevant human genetic variation along with the evidence for its disease causality.

dbSNP
A minimally curated database of single nucleotide human genetic variation.

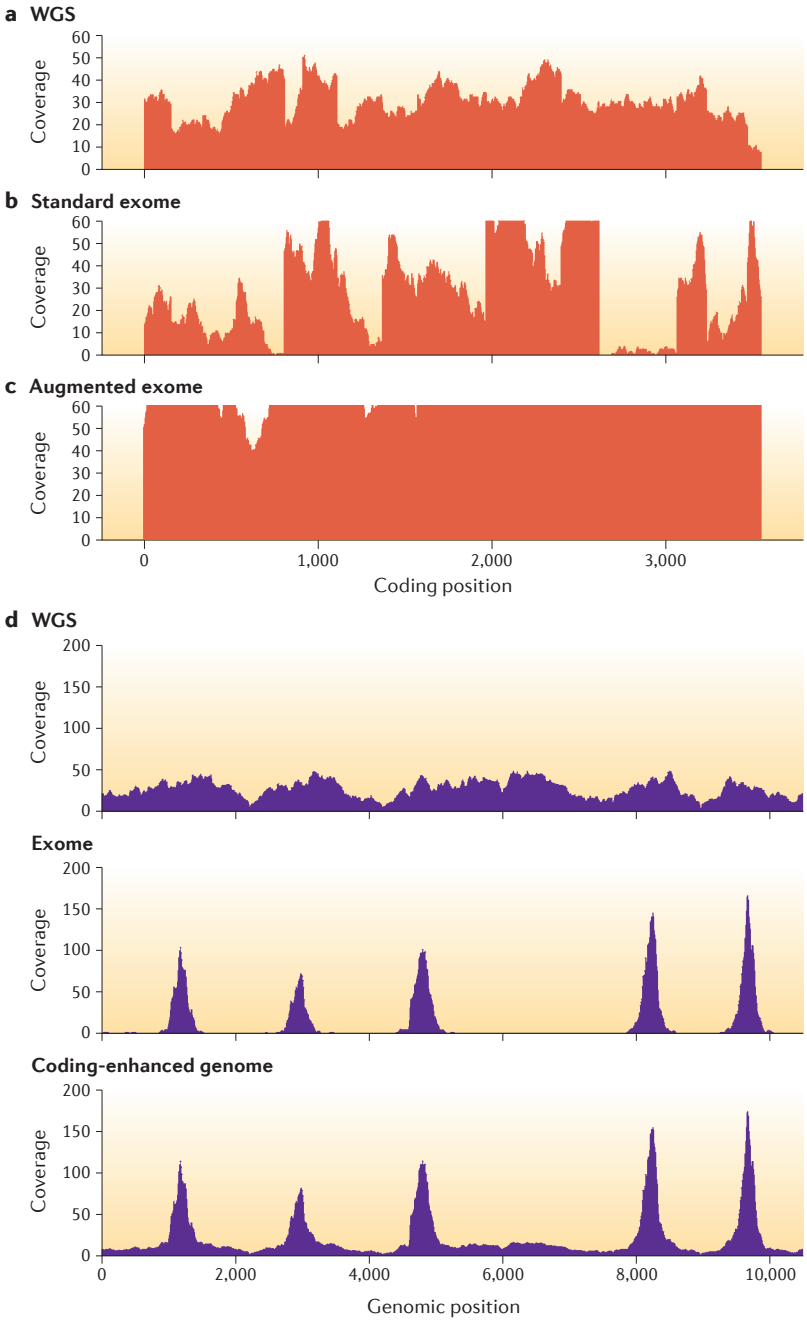


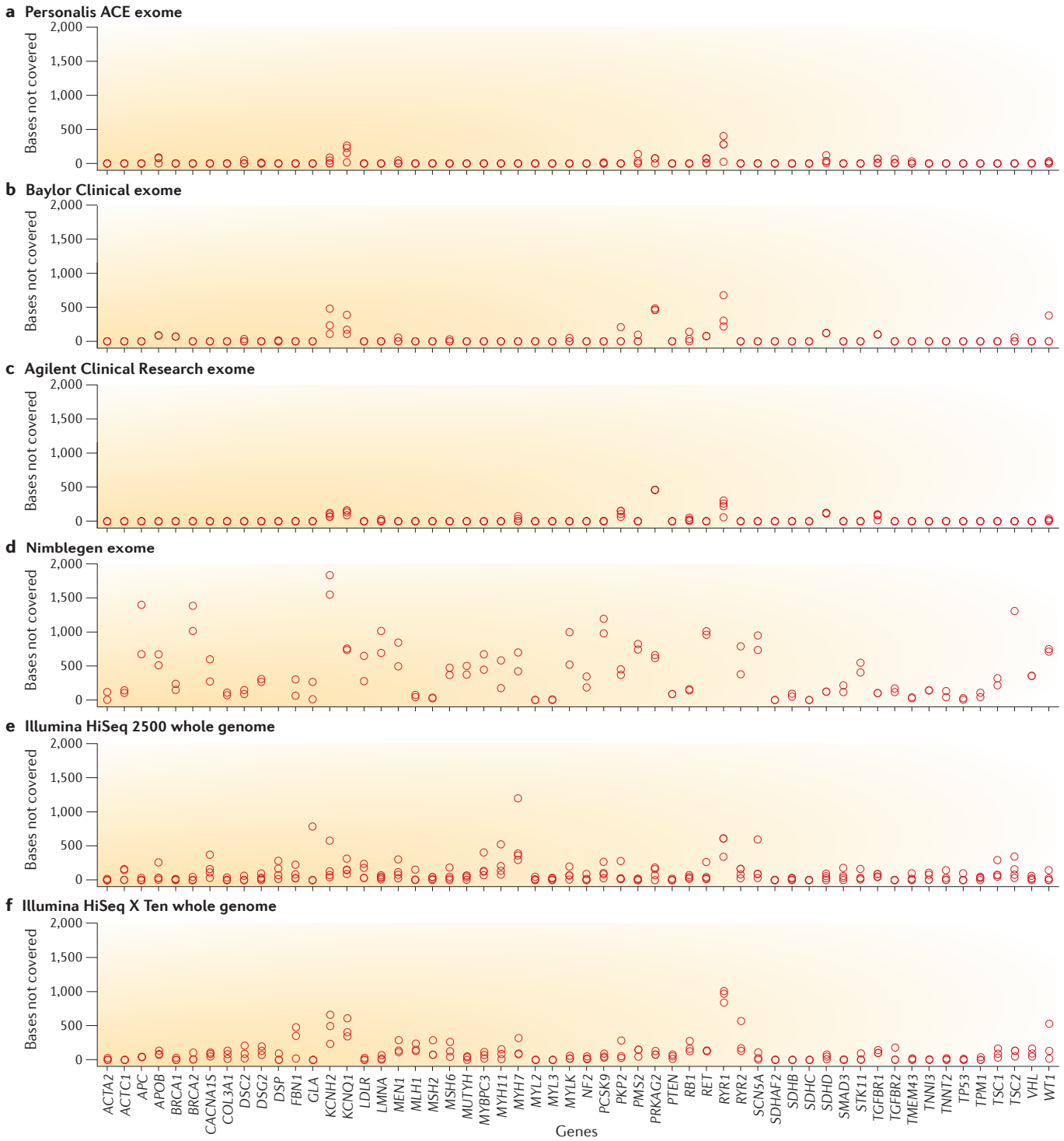
Figure 4 | **Exomes, genomes and augmentation. a** | Whole-genome sequencing (WGS) provides even coverage of the coding region of potassium voltage-gated channel subfamily H member 2 (*KCNH2*; introns removed for clarity), but at a typical clinical deliverable of median 30× coverage there are many positions that remain inadequately covered to make a confident call. **b** | Traditional exome capture results in highly variable coverage, with many sections of important genes not represented at sufficient quality coverage to make a confident call. In this case most of one exon is missing. **c** | Augmented exome capture targets medically relevant genes and fills in the gaps. **d** | Zoomed out view of multiple exons and introns. Whereas exome capture results in higher coverage of known disease-causing genes for less cost in sequencing and data storage, WGS at 30× results in lower coverage of the coding region of key genes, but in better coverage of the non-targeted genes and non-coding regions. A rational combination augments the coding regions of key genes but provides some genome-wide coverage to balance the strongest features of both approaches. Note that even exome capture at clinically deliverable depth for germline testing is not considered to be high enough coverage to adequately characterize tumour heterogeneity and to reliably detect somatic mosaicism[140]. A targeted panel is preferred for this purpose.

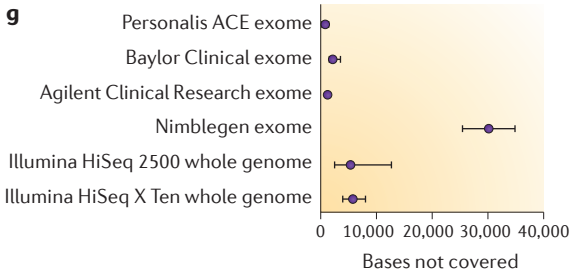
serious consequences were it to contain the causative variant. Alternatively, it could result in false reassurance. In either case, the consequences have meaning even beyond the individual to all family members who are potentially at risk of inheriting the disease. Metrics such as '90% coverage at 10× or more' that were common for exome research products are not appropriate for clinical diagnostics. This created a challenge for clinical laboratories for which the existing standard was that a clinical report would not be signed out unless every coding base pair (as well as the two bases on either side to account for splice dinucleotides) was called.

Groups have responded to these challenges by augmenting coverage in certain regions, both coding and non-coding[125,126] (FIGS 4c,5), through the addition of extra probes in these regions (known as augmented exome sequencing). Some laboratories also use targeted PCR to fill in gaps[127]. However, increasing capture in certain areas only goes so far in improving the ability to make a call at every position. GC-rich regions — for example, the first exon of most genes — cannot be optimized simply with extra coverage. These regions require library preparation and sequencing conditions that are tailored to their high GC chemistry[125].

Whole-genome sequencing. Sequencing the whole genome seems at first to be an answer to these problems. As all genomic DNA is included, concerns relating to capture are not relevant and coverage is clearly more evenly distributed (FIG. 4d). In addition, regulatory areas of the genome are included. Given that most variants that were associated with disease from genome-wide association studies (GWAS) were in non-coding regions, and given that ENCODE (The Encyclopedia of DNA Elements) suggested that large portions of the non-coding genome might be in some way important, this could be valuable data for clinical genomics. For discovery, this remains true. However, for clinical application, GWAS hits with low magnitude of effect remain of limited, though increasing, value as very few associations between Mendelian disease and regulatory variation have been described (and these regions can be added to a capture kit). The current major benefit of whole-genome sequencing for clinical medicine is likely to be in the identification of structural variation, but the algorithms have not so far been accurate enough for short reads to allow this at clinical grade. Overall, replacing exome sequencing with whole-genome sequencing at 30× would lead to the sacrifice of confident callability of the coding genome to provide coverage of the non-coding genome. This 50-fold increase in sequencing has an unclear value for clinical application, as well as for research groups looking to maximize study size for dollar sequencing spend[128].

Comparison of approaches. In comparing different genome diagnostics, a standard metric is helpful. Advances in diagnostics or therapeutics in medicine are judged by the standard of 'non-inferiority'. Here, non-inferiority to Sanger sequencing requires that every coding base pair +/– two bases should receive a





h

Sequencing approach	Vendor (sample size)	Mean sequencing yield (GB)	Mean bases missing
Exome	Personalis ACE Exome (*N*=4)	12.78	805
	Baylor Clinical Exome (*N*=3)	11.18	2,188
	Agilent Clinical Research Exome (*N*=4)	13.46	1,288
	Nimblegen Exome (*N*=2)	3.87	30,353
Whole genome	HiSeq 2500 (*N*=4)	124.07	5,377
	HiSeq X Ten (*N*=3)	114.00	5,800

Splice dinucleotides
The almost invariant canonical dinucleotides that are crucial for splicing (GT: donor; AG: acceptor).

Univariate
Depending on only one variable.

Multivariate
Depending on multiple variables.

confident call. In addition, as Sanger sequencing usually requires PCR of the specific exon, non-inferiority should include only uniquely mapped reads. Thus, the idea of a quality-coverage-mappability metric for comparing different research and commercial products has been gaining traction (FIG. 5). This metric quantifies the number of base pairs per gene of interest that are not covered by 20 or more uniquely mapped Q20 bases. An absolute base pair count is preferable to a percentage because any base can theoretically harbour a disease-causing variant and genes widely vary in their size. For example, if 10% of the gene titin is not callable this would represent >10,000 base pairs of potentially disease-causing variation that are missed (titin is associated with dilated cardiomyopathy). We recently made available[129] a tool to generate this metric for a given set of genes based on raw data output from various providers[8,67,129] (FIG. 5). An important finding from the application of this metric is that a clinical diagnostic that is based on whole-genome sequence at the standard coverage meets this standard far less often for known disease genes than does augmented exome sequencing (the reverse would be expected for non-disease genes, as these genes are not currently augmented by any vendor). Application of this metric may provide independent verification of new sequencing approaches. For example, data from the Illumina X Ten sequencing machine meets this standard less often than data from the prior HiSeq 2500 (REF. 130). Although this reduced calling confidence could potentially be overcome by increasing whole-genome coverage, this is at the cost of having to generate a large amount more genome-wide sequencing data and, notably, increasing the cost per genome beyond US$1,000.

Consistent with independent community verification is the emerging field of community-led regulatory science. Indeed, an important aspect of the US Precision Medicine Initiative is funding for new regulatory approaches. The first manifestation of this is precisionFDA — a website and development environment that will provide tools to allow easier comparison of products from different sequencing vendors and informatics companies[131,132]. The tool used to generate FIGURE 5 is one of the launch tools of precisionFDA[129].

A genome diagnostic combining multiple technologies. For clinical application, there is some value to whole-genome coverage, although it is more important to cover every coding base pair, especially of genes already known to be important for disease. This observation suggests the concept of a 'coding-enhanced' genome (FIG. 4d). In this concept, coding regions are covered at a high depth through specialized capture but there is some coverage of the whole genome to allow structural variant discovery from the same assay. Until such a time as long-read approaches are cost-effective for genome-wide coverage, then targeted capture of long molecules for complex areas of the genome maximizes the cost–discovery balance. Such a combination approach has the advantage of maximizing the opportunity to diagnose disease through excellent coverage of the coding medical genome, maximizing the accuracy of structural variant calling, repeat calling and variant characterization in complex areas of the genome, and at the same time rationalizing sequencing and data storage costs. As with all clinical genetic tools in an environment of rapidly expanding knowledge, the captured regions will probably need to be updated every few months to account for newly discovered genes and variants.

◀ Figure 5 | **Quality-coverage metrics for the American College of Medical Genetics 56 most actionable genes.** Plots are the number of bases per gene not callable with 20 uniquely-mapped Q20 bases. This standard is equivalent to 'non-inferiority' with Sanger sequencing. **a** | Personalis ACE exome. **b** | Baylor Clinical exome. **c** | Agilent Clinical Research exome. **d** | Nimblegen exome. **e** | Illumina HiSeq 2500 whole genome (2014). **f** | Illumina X Ten whole genome. **g,h** | Total number of bases not covered for the 56 genes. Adapted from REF. 129.

Causality and disease categorization

Accurate and precise genomic approaches will greatly facilitate the central tenet of precision medicine: more sophisticated definitions of disease[133]. The concept of causality is fundamental and recurrent in clinical genetics, as science has provided an abundance of association evidence. Indeed, discovery genetics has identified robust statistical associations between diseases and genetic variants but for a variant to be useful as a diagnostic test or therapeutic target, it is crucial to demonstrate a causal link. Achieving confidence in the determination of causality between a gene or variant and a disease is a complex task that requires various types of supportive data[58]. Clinical genetics has historically embraced a univariate paradigm in its approach to causality. Even the professional guidelines for variant classification[134] force variants into categories on a linear (but not proportional) scale between 'pathogenic' and 'benign'. However, it is clear that the clinical expressivity of a particular variant will depend on the magnitude and dependency of its effect. In this case, dependency incorporates genetic background and other factors such as age and environmental exposure in determining whether the clinical variant is expressed as disease. If the magnitude of the effect of a given variant is large and its dependency small (for example, a chromosomal abnormality) then the disease will generally always be evident if the particular variant is present. If the magnitude of the effect of a given variant is small and its dependency is large (for example, a common variant for a complex disease) then the effect of the variant may never be discernible in isolation. In between these two extremes is a highly variable relationship between variant and disease that is better conceptualized as a multivariate model with a large number of inputs. For Mendelian disease, one or more variants will be highly weighted, with other inputs having a substantially lower weighting (perhaps ten 'modifying variants'). For complex disease, recent data have suggested that there will probably be hundreds of variants with small weightings[135,136]. Significant weighting would also be given to environmental modifiers that may interact with genetic effects.

Therefore, a major challenge is the convenient storage and retrieval of the causal evidence for each variant. Until recently, data on clinically relevant variants were to be

found in the literature and in the proprietary databases of commercial testing companies. Sharing occurred but not in any structured or efficient way. The initiation of the ClinVar database and its population by the ClinGen project[137], as well as efforts such as Decipher in the UK, have led to more global sharing of rigorously curated evidence. However, the challenges of implementing even standardized guidelines for the interpretation of this evidence are considerable[138]. Nevertheless, the goal of accumulating and sharing clinical evidence is worth pursuing because for many the 'second case' — another patient who presents in a similar way with a variant in the same gene — provides the highest level of evidence possible for causality. In fact, the newest work from ClinGen indicates that gene-specific or disease-specific overlays should be added to guidelines to maximize concordance between interpreters in a domain-specific way[139].

Conclusions

The past decade has witnessed a rapid acceleration in our understanding of the genetic basis of many diseases. With this greater understanding comes the possibility of re-defining disease at higher resolution and, along with this, targeting with more precise therapy. However, for precision medicine to succeed, genomics must also be more accurate. Whereas in cohort discovery projects, if a base is not covered, or an algorithm is insensitive, all that is missed is an opportunity for discovery. In clinical medicine, failing to make a diagnosis, or making a diagnosis in error, could have devastating consequences for individuals and families. In discussing this extraordinary opportunity for precision medicine to fulfil the promise of the human genome project, I have described surmountable challenges in advancing the accuracy of clinical genomics. Reducing reliance on reference sequences, making phasing routine, improving calling of indels and structural variants, characterizing complex areas of the genome through long-read sequencing and maximizing the cost effectiveness of genomic coverage will all be crucial. Advancing regulatory processes in parallel will be a necessary step to ensure high standards and patient safety. Creating large cohorts of individuals committed to partnering in discovery will maximize the benefit and speed its global dispersion. Finally, educating the next generation of physicians and laboratory directors will be crucial to the generation of the workforce that is required to sustain the initial promise.

Fuelled by technological advancement, fundamental discovery of genetic elements related to health and disease has been the engine of human genetics for decades. Building on this foundation, precision medicine will use the knowledge gained to redefine disease, to realize new therapies and to provide hope for generations of patients to come.

1. Collins, F. S. Implications of the Human Genome Project for medical science. *JAMA* **285**, 540 (2001).
2. Ng, S. B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272–276 (2009).
3. Choi, M. *et al.* Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc. Natl Acad. Sci. USA* **106**, 19096–19101 (2009).
References 2 and 3 were among the earliest studies to show that exome sequencing could be used to diagnose a genetic condition.
4. Ashley, E. A. *et al.* Clinical assessment incorporating a personal genome. *Lancet* **375**, 1525–1535 (2010).
This paper presented a framework for clinical whole-genome interpretation and described the earliest example of whole-genome-based personalized medicine.
5. Worthey, E. A. *et al.* Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genet. Med.* **13**, 255–262 (2011).
This paper describes a diagnosis made by exome sequencing that led to a dramatic therapeutic response in a young boy.
6. Bainbridge, M. N. *et al.* Whole-genome sequencing for optimized patient management. *Sci. Transl. Med.* **3**, 87re3 (2011).
7. Johnson, J. A *et al.* Clopidogrel: a case for indication-specific pharmacogenetics. *Clin. Pharmacol. Ther.* **91**, 774–776 (2012).
8. Dewey, F. E. *et al.* Clinical interpretation and implications of whole-genome sequencing. *JAMA* **311**, 1035–1045 (2014).
9. Vassy, J. L. *et al.* The MedSeq Project: a randomized trial of integrating whole genome sequencing into clinical medicine. *Trials* **15**, 85 (2014).
10. Manolio, T. A. & Green, E. D. Leading the way to genomic medicine. *Am. J. Med. Genet. C. Semin. Med. Genet.* **166C**, 1–7 (2014).
11. Green, E. D., Guyer, M. S. & Human, N. Charting a course for genomic medicine from base pairs to bedside. *Nature* **470**, 204–213 (2011).
12. Bryc, K., Durand, E. Y., Macpherson, J. M., Reich, D. & Mountain, J. L. The genetic ancestry of African Americans, Latinos, and European Americans across the United States. *Am. J. Hum. Genet.* **96**, 37–53 (2015).
13. Collins, F. S. & Varmus, H. A. New initiative on precision medicine. *N. Engl. J. Med.* **372**, 793–795 (2015).
14. Ashley, E. A. The Precision Medicine Initiative: a new national effort. *JAMA* **313**, 2119–2120 (2015).
15. National Research Council (US) Committee on a Framework for Developing a New Taxonomy of Disease, 2011. *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease* (National Academies Press, 2011).
16. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* http://dx.doi.org/10.1038/nature19057 (in the press) (2016).
This paper describes the Exome Aggregation Consortium.
17. Homburger, J. R. *et al.* Multidimensional structure-function relationships in human β-cardiac myosin from population-scale genetic variation. *Proc. Natl Acad. Sci. USA* **113**, 6701–6706 (2016).
18. Waggott, D. *et al.* The next generation precision medical record — a framework for integrating genomes and wearable sensors with medical records. Preprint at *bioRxiv* http://dx.doi.org/10.1101/039651 (2016).
19. Ramsey, B. W. *et al.* A CFTR potentiator in patients with cystic fibrosis and the *G551D* mutation. *N. Engl. J. Med.* **365**, 1663–1672 (2011).
A paper describing precision therapy for cystic fibrosis.
20. Brodlie, M., Haq, I. J., Roberts, K. & Elborn, J. S. Targeted therapies to improve CFTR function in cystic fibrosis. *Genome Med.* **7**, 101 (2015).
21. Rehman, A., Baloch, N. U.-A. & Janahi, I. A. Lumacaftor-Ivacaftor in patients with cystic fibrosis homozygous for Phe508del *CFTR*. *N. Engl. J. Med.* **373**, 1783 (2015).
22. Brewington, J. J., McPhail, G. L. & Clancy, J. P. Lumacaftor alone and combined with ivacaftor: preclinical and clinical trial experience of F508del CFTR correction. *Expert Rev. Respir. Med.* **10**, 5–17 (2016).
23. Lindeman, N. I. *et al.* Molecular testing guideline for selection of lung cancer patients for EGFR and ALK tyrosine kinase inhibitors. *J. Thorac. Oncol.* **8**, 823–859 (2013).
24. Blumenthal, G. Next-generation sequencing in oncology in the era of precision medicine. *JAMA Oncol.* **2**, 13–*14* (2015).
25. Sosman, J. A. *et al.* Survival in BRAF V600–mutant advanced melanoma treated with vemurafenib. *N. Engl. J. Med.* **366**, 707–714 (2012).
26. Sharma, P. & Allison, J. P. Immune checkpoint targeting in cancer therapy: toward combination strategies with curative potential. *Cell* **161**, 205–214 (2015).
27. Linnemann, C. *et al.* High-throughput epitope discovery reveals frequent recognition of neo-antigens by CD4$^+$ T cells in human melanoma. *Nat. Med.* **21**, 81–85 (2015).
28. Schadendorf, D. *et al.* Pooled analysis of long-term survival data from phase II and phase III trials of ipilimumab in unresectable or metastatic melanoma. *J. Clin. Oncol.* **33**, 1889–1894 (2015).
29. Lawrence, M. S. *et al.* Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517**, 576–582 (2015).
30. Hackl, H., Charoentong, P., Finotello, F. & Trajanoski, Z. Computational genomics tools for dissecting tumour–immune cell interactions. *Nat. Rev. Genet.* **17**, 441–458 (2016).
31. Gubin, M. M. *et al.* Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens. *Nature* **515**, 577–581 (2014).
An important paper describing checkpoint blockade.
32. Newman, A. M. *et al.* An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nat. Med.* **20**, 548–554 (2014).
An early paper describing the concept of 'liquid biopsy'.
33. Klein, T. E. *et al.* Estimation of the warfarin dose with clinical and pharmacogenetic data. *N. Engl. J. Med.* **360**, 753–764 (2009).
A key paper describing warfarin pharmacogenomics.
34. Eckman, M. H., Rosand, J., Greenberg, S. M. & Gage, B. F. Cost-effectiveness of using pharmacogenetic information in warfarin dosing for patients with nonvalvular atrial fibrillation. *Ann. Intern. Med.* **150**, 73–83 (2009).

35. Epstein, R. S. *et al.* Warfarin genotyping reduces hospitalization rates results from the MM-WES (Medco-Mayo Warfarin Effectiveness study). *J. Am. Coll. Cardiol.* **55**, 2804–2812 (2010).
36. Mega, J. L. *et al.* Reduced-function *CYP2C19* genotype and risk of adverse clinical outcomes among patients treated with clopidogrel predominantly for PCI: a meta-analysis. *JAMA* **304**, 1821–1830 (2010).
37. Mega, J. L. *et al.* Genetic variants in *ABCB1* and *CYP2C19* and cardiovascular outcomes after treatment with clopidogrel and prasugrel in the TRITON–TIMI 38 trial: a pharmacogenetic analysis. *Lancet* **376**, 1312–1319 (2010).
38. Paré, G. Effects of *CYP2C19* genotype on outcomes of clopidogrel treatment. *N. Engl. J. Med.* **363**, 1704–1714 (2010).
39. Nissen, S. Pharmacogenomics and clopidogrel: irrational exuberance? *J. Am. Med. Assoc.* **306**, 2011–2012 (2012).
40. Johnson, J. A. *et al.* Clopidogrel: a case for indication-specific pharmacogenetics. *Clin. Pharmacol. Ther.* **91**, 774–776 (2012).
41. Roberts, J. D. *et al.* Point-of-care genetic testing for personalisation of antiplatelet treatment (RAPID GENE): a prospective, randomised, proof-of-concept trial. *Lancet* **379**, 1705–1711 (2012).
42. Wiviott, S. D. *et al.* Prasugrel versus clopidogrel in patients with acute coronary syndromes. *N. Engl. J. Med.* **357**, 2001–2015 (2007).
43. Dunnenberger, H. M. *et al.* Preemptive clinical pharmacogenetics implementation: current programs in five US medical centers. *Annu. Rev. Pharmacol. Toxicol.* **55**, 89–106 (2015).
44. Altman, R. B. PharmGKB: a logical home for knowledge relating genotype to drug response phenotype. *Nat. Genet.* **39**, 426 (2007).
45. Klein, T. E. *et al.* Integrating genotype and phenotype information: an overview of the PharmGKB project: an overview of the PharmGKB project. *Pharmacogenomics J.* **1**, 167–170 (2001).
46. Caudle, K. E. *et al.* Incorporation of pharmacogenomics into routine clinical practice: the Clinical Pharmacogenetics Implementation Consortium (CPIC) guideline development process. *Curr. Drug Metab.* **15**, 209–217 (2014).
47. Herr, T. M. *et al.* Practical considerations in genomic decision support: The eMERGE experience. *J. Pathol. Inform.* **6**, 50 (2015).
48. Obama, B. S.3822 — 109th Congress: Genomics and Personalized Medicine Act of 2006. *Congress.gov* https://www.congress.gov/bill/109th-congress/senate-bill/3822 (2006).
49. Collins, F. S. The case for a US prospective cohort study of genes and environment. *Nature* **429**, 475–477 (2004).
50. Collins, R. What makes UK Biobank special? *Lancet* **379**, 1173–1174 (2012).
51. Stanford Medicine MyHeart Counts iPhone Application. https://med.stanford.edu/myheartcounts.html (2016)
52. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
53. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
54. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
55. Henderson, L. B. *et al.* The impact of chromosomal microarray on clinical management: a retrospective analysis. *Genet. Med.* **16**, 1–8 (2014).
56. Gahl, W. A. *et al.* The National Institutes of Health Undiagnosed Diseases Program: insights into rare diseases. *Genet. Med.* **14**, 51–59 (2012).
57. Gahl, W. A., Wise, A. L. & Ashley, E. A. The Undiagnosed Diseases Network of the National Institutes of Health: a national extension. *JAMA* **314**, 1797–1798 (2015).
58. MacArthur, D. G. *et al.* Guidelines for investigating causality of sequence variants in human disease. *Nature* **508**, 469–476 (2013).
Detailed guidance on how to assess causality of variants for rare disease.
59. Biesecker, L. G. & Spinner, N. B. A genomic view of mosaicism and human disease. *Nat. Rev. Genet.* **14**, 307–320 (2013).
60. Church, D. M. *et al.* Extending reference assembly models. *Genome Biol.* **16**, 13 (2015).
61. Genome Reference Consortium. Human Genome Assembly Data. *GRC* http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/data (2015).
62. Ezkurdia, I. *et al.* Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Hum. Mol. Genet.* **23**, 5866–5878 (2014).
63. Platzer, M. The human genome and its upcoming dynamics. *Genome Dyn.* **2**, 1–16 (2006).
64. Pruitt, K. D., Tatusova, T., Brown, G. R. & Maglott, D. R. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.* **40**, D130–D135 (2012).
65. López-Flores, I. & Garrido-Ramos, M. A. The repetitive DNA content of eukaryotic genomes. *Genome Dyn.* **7**, 1–28 (2012).
66. National Center for Biotechnology Information. NCBI *Homo sapiens* annotation release 107. *NCBI* http://www.ncbi.nlm.nih.gov/genome/annotation_euk/Homo_sapiens/107 (2015).
67. Goldfeder, R. *et al.* Medical implications of technical accuracy in clinical genome sequencing. *Genome Med.* **8**, 1–12 (2016).
68. Gymrek, M. *et al.* Abundant contribution of short tandem repeats to gene expression variation in humans. *Nat. Genet.* **48**, 22–29 (2016).
69. Budworth, H. & McMurray, C. T. A brief history of triplet repeat diseases. *Methods Mol. Biol.* **1010**, 3–17 (2013).
70. Iyer, R. R., Pluciennik, A., Napierala, M. & Wells, R. D. DNA triplet repeat expansion and mismatch repair. *Annu. Rev. Biochem.* **84**, 199–226 (2015).
71. Rufini, S. *et al.* Stevens–Johnson syndrome and toxic epidermal necrolysis: an update on pharmacogenetics studies in drug-induced severe skin reaction. *Pharmacogenomics* **16**, 1989–2002 (2015).
72. Chung, W.-H. *et al.* Medical genetics: a marker for Stevens–Johnson syndrome. *Nature* **428**, 486 (2004).
73. Mallal, S. *et al.* Association between presence of *HLA-B*5701*, *HLA-DR7*, and *HLA-DQ3* and hypersensitivity to HIV-1 reverse-transcriptase inhibitor abacavir. *Lancet* **359**, 727–732 (2002).
74. Dilthey, A., Cox, C., Iqbal, Z., Nelson, M. R. & McVean, G. Improved genome inference in the MHC using a population reference graph. *Nat. Genet.* **47**, 682–688 (2015).
75. Tewhey, R., Bansal, V., Torkamani, A., Topol, E. J. & Schork, N. J. The importance of phase information for human genomics. *Nat. Rev. Genet.* **12**, 215–223 (2011).
76. Rosenfeld, J. A., Malhotra, A. K. & Lencz, T. Novel multi-nucleotide polymorphisms in the human genome characterized by whole genome and exome sequencing. *Nucleic Acids Res.* **38**, 6102–6111 (2010).
77. Tilgner, H., Grubert, F., Sharon, D. & Snyder, M. P. Defining a personal, allele-specific, and single-molecule long-read transcriptome. *Proc. Natl Acad. Sci. USA* **111**, 9869–9874 (2014).
78. Zheng, G. X. *et al.* Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat. Biotechnol.* **34**, 303–311 (2016).
79. Chaisson, M. J. P. *et al.* Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608–611 (2014).
80. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).
81. Chaisson, M. J. P., Wilson, R. K. & Eichler, E. E. Genetic variation and the *de novo* assembly of human genomes. *Nat. Rev. Genet.* **16**, 627–640 (2015).
82. Chakraborty, M., Baldwin-Brown, J. G., Long, A. D. & Emerson, J. J. Contiguous and accurate *de novo* assembly of metazoan genomes with modest long read coverage. Preprint at *bioRxiv* http://dx.doi.org/10.1101/029306 (2015).
83. Huang, Y.-T. & Liao, C.-F. Integration of string and de Bruijn graphs for genome assembly. *Bioinformatics* **32**, 1301–1307 (2016).
84. Korlach, J. Returning to more finished genomes. *Genom. Data* **2**, 46–48 (2014).
85. Goodwin, S. *et al.* Oxford Nanopore sequencing, hybrid error correction, and *de novo* assembly of a eukaryotic genome. *Genome Res.* **25**, 1750–1756 (2015).
86. Pendleton, M. *et al.* Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods* **12**, 780–786 (2015).
87. Stephens, Z. D. *et al.* Big Data: astronomical or genomical? *PLOS Biol.* **13**, e1002195 (2015).
88. Hsi-Yang Fritz, M., Leinonen, R., Cochran, G. & Birney, E. Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Res.* **21**, 734–740 (2011).
89. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**, 186–194 (1998).
90. Ewing, B., Hillier, L., Wendl, M. C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8**, 175–185 (1998).
91. Malysa, G. *et al.* QVZ: lossy compression of quality values. *Bioinformatics* **31**, 3122–3129 (2015).
92. Ochoa, I., Hernaez, M., Goldfeder, R., Weissman, T. & Ashley, E. Effect of lossy compression of quality scores on variant calling. *Brief. Bioinform.* http://dx.doi.org/10.1093/bib/bbw011 (2016).
93. Yu, Y. W., Yorukoglu, D., Peng, J. & Berger, B. Quality score compression improves genotyping accuracy. *Nat. Biotechnol.* **33**, 240–243 (2015).
94. Iqbal, Z., Caccamo, M., Turner, I., Flicek, P. & McVean, G. *De novo* assembly and genotyping of variants using colored de Bruijn graphs. *Nat. Genet.* **44**, 226–232 (2012).
95. Dewey, F. E. *et al.* Phased whole-genome genetic risk in a family quartet using a major allele reference sequence. *PLoS Genet.* **7**, e1002280 (2011).
96. Needleman, S. B. & Wunsch, C. D. A general method applicable to search for similarities in amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–*453* (1970).
97. Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197 (1981).
98. Priest, J. R. *et al. De novo* and rare variants at multiple loci support the oligogenic origins of atrioventricular septal heart defects. *PLoS Genet.* **12**, e1005963 (2016).
99. Korpar, M. & Šikic, M. SW#–GPU-enabled exact alignments on genome scale. *Bioinformatics* **29**, 2494–2495 (2013).
100. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
101. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
102. Manber, U. & Myers, G. Suffix arrays: a new method for on-line string searches. *SIAM J. Comput.* **22**, 935–948 (1993).
103. Saunders, C. J. *et al.* Rapid whole-genome sequencing for genetic disease diagnosis in neonatal intensive care units. *Sci. Transl. Med.* **4**, 154ra135 (2012).
104. Priest, J. R. *et al.* Molecular diagnosis of long QT syndrome at 10 days of life by rapid whole genome sequencing. *Heart Rhythm* **11**, 1707–1713 (2014).
105. Dewey, F. E. E. *et al.* Sequence to medical phenotypes (STMP): a clinical research tool for interpretation of next generation sequencing data. *PLoS Genet.* **11**, e1005496 (2015).
106. Kalyana-Sundaram, S. *et al.* Expressed pseudogenes in the transcriptional landscape of human cancers. *Cell* **149**, 1622–1634 (2012).
107. Poliseno, L. *et al.* A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* **465**, 1033–1038 (2010).
108. Bauer, K. A. The thrombophilias: well-defined risk factors with uncertain therapeutic implications. *Ann. Intern. Med.* **135**, 367–373 (2001).
109. Lam, H. Y. K. *et al.* Performance comparison of whole-genome sequencing platforms. *Nat. Biotechnol.* **30**, 78–82 (2012).
110. Zook, J. M. *et al.* Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.* **32**, 246–251 (2014).
The US National Institute for Standards and Technology paper providing a consensus resource for the community for one genome.
111. Krier, J., Barfield, R., Green, R. C. & Kraft, P. Reclassification of genetic-based risk predictions as GWAS data accumulate. *Genome Med.* **8**, 20 (2016).
112. Human Genome Variation Society. Human Genome Variation Society. *HGVS* http://www.hgvs.org/dblist/glsdb.html (updated 30 May 2016).
113. Vis, J. K., Vermaat, M., Taschner, P. E. M., Kok, J. N. & Laros, J. F. J. An efficient algorithm for the extraction of HGVS variant descriptions from sequences. *Bioinformatics* **31**, 3751–3757 (2015).

114. Hart, R. K. *et al.* A Python package for parsing, validating, mapping and formatting sequence variants using HGVS nomenclature. *Bioinformatics* **31**, 268–270 (2015).
115. Albers, C. a. *et al.* Dindel: accurate indel calls from short-read data. *Genome Res.* **21**, 961–973 (2011).
116. Narzisi, G. *et al.* Accurate *de novo* and transmitted indel detection in exome-capture data using microassembly. *Nat. Methods* **11**, 1–7 (2014).
117. Rimmer, A. *et al.* Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **46**, 1–9 (2014).
118. Ye, K. *et al.* Systematic discovery of complex insertions and deletions in human cancers. *Nat. Med.* **22**, 1–10 (2015).
119. Yang, R., Nelson, A. C., Henzler, C., Thyagarajan, B. & Silverstein, K. A. T. ScanIndel: a hybrid framework for indel detection via gapped alignment, split reads and *de novo* assembly. *Genome Med.* **7**, 127 (2015).
120. Huddleston, J. *et al.* Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Res.* **24**, 688–696 (2014).
121. Gilissen, C. *et al.* Genome sequencing identifies major causes of severe intellectual disability. *Nature* **511**, 344–347 (2014).
122. Xie, C. & Tammi, M. T. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics* **10**, 80 (2009).
123. Retterer, K. *et al.* Clinical application of whole-exome sequencing across clinical indications. *Genet. Med.* **18**, 696–704 (2016).
124. Retterer, K. *et al.* Assessing copy number from exome sequencing and exome array CGH based on CNV spectrum in a large clinical cohort. *Genet. Med.* **17**, 623–629 (2015).
125. Patwardhan, A. *et al.* Achieving high-sensitivity for clinical applications using augmented exome sequencing. *Genome Med.* **7**, 71 (2015).
126. Santani, A. *et al.* Medical Exome: Towards achieving complete coverage of disease related genes. (Abstract #371) *The 64th Annual Meeting of The American Society of Human Genetics, San Diego, California* http://www.ashg.org/2014meeting/pdf/2014_ASHG_Meeting_Platform_Abstracts.pdf (18–22 Oct 2014).
127. Mandelker, D. *et al.* Navigating highly homologous genes in a molecular diagnostic setting: a resource for clinical next-generation sequencing. *Genet. Med.* http://dx.doi.org/10.1038/gim.2016.58 (2016).
128. McRae, J. F. *et al.* Prevalence, phenotype and architecture of developmental disorders caused by de novo mutation. Preprint at *bioRxiv* http://dx.doi.org/10.1101/049056 (2016).
129. Goldfeder, R. & Ashley, E. A precision metric for clinical genomic sequencing. Preprint at *bioRxiv* http://dx.doi.org/10.1101/051490 (2016).
130. Li, H. On HiSeq X10 Base Quality. http://lh3.github.io/2014/11/03/on-hiseq-x10-base-quality (2014).
131. Altman, R. B., Khuri, N., Salit, M. & Giacomini, K. M. Unmet needs: Research helps regulators do their jobs. *Sci. Transl. Med.* **7**, 315ps22 (2015).
132. Kass-Hout, T. & Litwack, D. Advancing precision medicine by enabling a collaborative informatics community. *FDA Voice* http://blogs.fda.gov/fdavoice/index.php/2015/08/advancing-precision-medicine-by-enabling-a-collaborative-informatics-community (2015).
133. Pearl, J. *Causality* (Cambridge Univ. Press, 2009).
134. Richards, C. S. *et al.* ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007. *Genet. Med.* **10**, 294–300 (2008).
135. Kutalik, Z., Whittaker, J., Waterworth, D., Beckmann, J. S. & Bergmann, S. Novel method to estimate the phenotypic variation explained by genome-wide association studies reveals large fraction of the missing heritability. *Genet. Epidemiol.* **35**, 341–349 (2011).
136. Wood, A. R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46**, 1173–1186 (2014).
137. Rehm, H. L. *et al.* ClinGen — The Clinical Genome Resource. *N. Engl. J. Med.* **372**, 2235–2242 (2015). **A description of the Clinical Genome Resource (ClinGen).**
138. Amendola, L. M. *et al.* Performance of ACMG-AMP variant-interpretation guidelines among nine laboratories in the clinical sequencing exploratory research consortium. *Am. J. Hum. Genet.* **98**, 1067–1076 (2016).
139. Caleshu, C. & Ashley, E. Taming the genome. *Genome Med.* **8**, 70 (2016).
140. Fisher, K. E. *et al.* Clinical validation and implementation of a targeted next-generation sequencing assay to detect somatic variants in non-small cell lung, melanoma, and gastrointestinal malignancies. *J. Mol. Diagn.* **18**, 299–315 (2016).

Acknowledgements
The author extends his grateful thanks to R. Goldfeder, A. Dainis, M. Grove, D. Church, M.J. Clark, S. Garcia, G. Chandratillake and C. Caleshu for helpful discussion and suggestions on the manuscript.

Competing interests statement
The author declares competing interests: see Web version for details.

FURTHER INFORMATION
Centers for Mendelian Genomics: http://mendelian.org
Clinical Genome Resource (ClinGen): https://www.clinicalgenome.org
Clinical Sequencing Exploratory Research: https://cser-consortium.org
Online Mendelian Inheritance in Man (OMIM): http://www.omim.org
precisionFDA: https://precision.fda.gov
Precision Medicine Initiative: https://www.whitehouse.gov/precision-medicine
The Cancer Genome Atlas: http://cancergenome.nih.gov
The Encyclopedia of DNA Elements (ENCODE): https://www.encodeproject.org
The Pharmacogenomics Knowledgebase (PharmGKB): https://www.pharmgkb.org
The Undiagnosed Diseases Network: http://undiagnosed.hms.harvard.edu
UK Biobank: http://www.ukbiobank.ac.uk

ALL LINKS ARE ACTIVE IN THE ONLINE PDF