Uncovering Missing Heritability in Rare Diseases

Tatiana Maroilley, Maja Tarailo-Graovac

genes

MDPI

Review

Uncovering Missing Heritability in Rare Diseases

Tatiana Maroilley [1, 2] and Maja Tarailo-Graovac [1, 2, *]

[1] Departments of Biochemistry, Molecular Biology and Medical Genetics, Cumming School of Medicine, University of Calgary, Calgary, AB T2N 4N1, Canada; tatiana.maroilley@ucalgary.ca

[2] Alberta Children's Hospital Research Institute, University of Calgary, Calgary, AB T2N 4N1, Canada

* Correspondence: maja.tarailograovac@ucalgary.ca

Received: 1 March 2019; Accepted: 1 April 2019; Published: 4 April 2019

Abstract: The problem of 'missing heritability' affects both common and rare diseases hindering: discovery, diagnosis, and patient care. The 'missing heritability' concept has been mainly associated with common and complex diseases where promising modern technological advances, like genome-wide association studies (GWAS), were unable to uncover the complete genetic mechanism of the disease/trait. Although rare diseases (RDs) have low prevalence individually, collectively they are common. Furthermore, multi-level genetic and phenotypic complexity when combined with the individual rarity of these conditions poses an important challenge in the quest to identify causative genetic changes in RD patients. In recent years, high throughput sequencing has accelerated discovery and diagnosis in RDs. However, despite the several-fold increase (from ~10% using traditional to ~40% using genome-wide genetic testing) in finding genetic causes of these diseases in RD patients, as is the case in common diseases—the majority of RDs are also facing the 'missing heritability' problem. This review outlines the key role of high throughput sequencing in uncovering genetics behind RDs, with a particular focus on genome sequencing. We review current advances and challenges of sequencing technologies, bioinformatics approaches, and resources.

Keywords: missing heritability; rare disease; genome sequencing; long/short read sequencing; bioinformatics; variant detection; variant annotation; variation databases

1. Introduction

Heritability is a measure that estimates the proportion of a phenotypic trait variability that is genetic in origin (i.e., could not be explained by the environment or random chance). The 'missing heritability' problem term was first coined by Brendan Maher in 2008 [1], mainly to describe unmet expectations from the human genome project combined with promising modern technological advances, such as genome-wide association studies (GWAS), to uncover genetic components of common traits and diseases [1]. Although the problem of 'missing heritability' has been mostly (read exclusively) associated with common and complex diseases in the medical research field [1,2], rare diseases also face 'missing heritability' problem despite the state-of-the-field technological advances [3].

Rare diseases (RDs) are mostly genetic diseases that are defined as life-threatening or chronically debilitating disorders affecting a small number of people (fewer than 5 per 10,000) [4]. Some 7000 RDs have been reported to date (see ORPHANET [5] and OMIM for Online Mendelian Inheritance in Man [6] databases) and new syndromes continue to be described, making the RDs quite common overall. An estimated 350 million people in the world suffer from a rare disease and approximately 50% of those are children. In Canada, this represents approximately 1 in 12 people according to the Canadian Organization for Rare Diseases (CORD).

Traditionally, clinical genetic tests for diagnosing RD patients have involved high resolution molecular single-gene tests (e.g., Sanger sequencing), low resolution genome-wide cytogenetic tests (e.g., G-banded karyotype) or microarrays have achieved a diagnostic success rate of ~10% [3]. While

the GWAS had uncovered new associations in common diseases, this approach was not adaptable to RDs, due to genetic and phenotypic heterogeneity combined with the rarity of individual conditions, and the unavailability of large cohorts. It is only the crucial technological advances in high throughput sequencing (HTS) and the bioinformatics field that have enabled unprecedented opportunity to accelerate diagnosis and discovery in RDs [3,7–9]. However, even after almost a decade of HTS applications in RD patients, the majority of RD patients remain without genetic answers [3].

Here, we focus on the concept of the 'missing heritability' problem in the rare disease research field. We review the HTS approaches used so far, and highlight the potential of genome sequencing to uncover 'missing heritability' in RDs, with particular attention to types of sequencing technologies, bioinformatics approaches used, and available resources on 'normal' variation within populations. We conclude with future perspectives.

2. Complexity of Rare Diseases

2.1. Heterogeneity

PHENOTYPIC HETEROGENEITY refers to strikingly different phenotypes associated with different variants of the same gene. For example, variants in *TRPV4* have been reported in more than 10 different dominant disorders, from various forms of skeletal disorders (e.g., Brachyolmia type 3, Parastremmatic dwarfism), to neuromuscular disorders (e.g., Hereditary motor and sensory neuropathy, type IIc, various forms of Spinal muscular atrophy) [6,10]. Similarly, variants in *FLNA* have been reported in various X-linked dominant (XLD) and recessive disorders (XLR), such as Periventricular Heterotopia 1, various malformation syndromes (e.g., XLD Otopalatodigital syndrome, XLR Frontometaphyseal dysplasia) and others [6]. Recently, we [11] and others [12] have associated heterozygous variants in the *ATP1A1* to human diseases, either an inherited dominant Charcot-Marie-Tooth type 2 disease [12] or a more severe condition due to de novo variants with major features of renal hypomagnesemia, refractory seizures, and intellectual disability [11]. Another example of an emerging rare disease with phenotypic heterogeneity is Glutaminase deficiency. While a homozygous copy number variant (duplication) in *GLS* was associated with autosomal recessive spastic ataxia and optic atrophy in two brothers from a consanguineous family [13], homozygote loss of functional variants (e.g., nonsense and frameshift) were associated with severe neonatal Epileptic encephalopathy and death before 40 days [14]. Thus, with the discoveries of new genes related to human diseases (like *ATP1A1* and *GLS*), it is clear that phenotypic heterogeneity continues to play an important role, and must be considered when interpreting the data.

GENETIC HETEROGENEITY, on the other hand, is defined as variations in distinct genes (two or more) that produce the same or similar phenotypes, either biochemical or clinical. Beyond the phenotypic heterogeneity, the genetic heterogeneity of RDs poses substantial diagnostic challenge. The degree of heterogeneity varies between different diseases. For example, thus far cystic fibrosis had only been associated with variants in *CFTR* [6], while tuberous sclerosis had only been associated with *TSC1* and *TSC2* [15]. These are good examples of currently no known (cystic fibrosis) or low (tuberous sclerosis) genetic heterogeneity. On the other hand, retinitis pigmentosa is an inherited degenerative disease resulting in severe retinal dystrophy and visual impairment mainly with onset in infancy or adolescence. It is usually diagnosed by a clinical exam and electrophysiological recordings, but a genetic diagnosis requires a multi-gene approach since more than 60 different genes had been associated with monogenic retinal disorders [16]. While retinitis pigmentosa may be considered to be an example of moderate heterogeneity, intellectual disability with more than 800 different gene associations [17] exemplifies substantial heterogeneity in human genetic diseases. Thus, considering phenotypic/genotypic heterogeneities in RDs is crucial for a successful approach to diagnosis.

2.2. Mutation Spectrum

ClinVar [18], a freely accessible repository of human variation, summarizes reports of variants related to human phenotypes with an evaluation of pathogenicity (likely/benign, uncertain significance, likely/pathogenic) and the potential source of supporting evidence. As of December 2018, more than 412,000 variants were available in ClinVar. Importantly, of those 13% (n = 52,424) were variants other than single nucleotide variants (SNVs) (Figure 1).

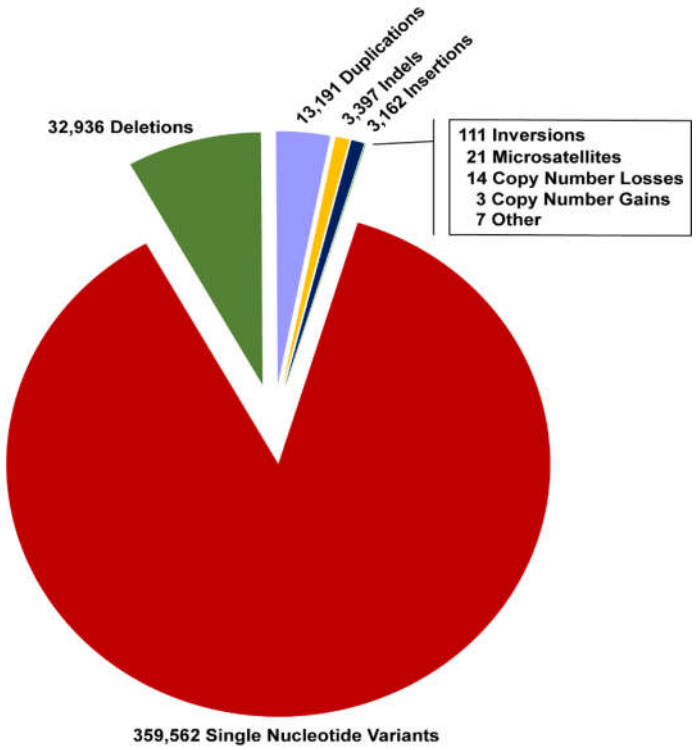

Figure 1. ClinVar variome. Representation of ClinVar variant types (as of December 2018). About 13% were structural variants. The annotation of variants is according to sequence ontology [19].

Most of the well-described monogenic diseases display a spectrum of gene-inactivation mechanisms [15,18,20,21]. For example, in patients with a clinical diagnosis of tuberous sclerosis, a spectrum of heterozygous variants affecting *TSC1* and *TSC2* had been described [15,20]. The variants range from SNVs resulting in missense, nonsense, splice-site changes, to structural variants (SVs), such as large deletions and duplications [20]. Interestingly, somatic, rather than germline variants, (in *TSC1* and *TSC2*) were identified in patients resistant to conventional diagnostic approaches [15,20]. Furthermore, in recent years HTS technologies revealed another type of SV termed chromothripsis, a type of chromosomal rearrangement with massive and complex clustered SVs that leave the affected genomic region changed beyond recognition [22]. Although chromothripsis had been predominantly associated with somatic genome instability (e.g., cancer), it had also been reported in individuals with severe congenital abnormalities [23] as well as in the striking case of spontaneous recovery in a patient with WHIM syndrome [24]. Given the variety of genetic mechanisms in gene inactivation, a holistic approach to assessment of individual genomes, including large insertions (such as mobile element insertions (MEI)), deletions, duplications, as well as translocations, inversions, repeat expansions, and other complex changes (Figure 2) would be a desired approach to discovery of functional variants in patients with rare disease.

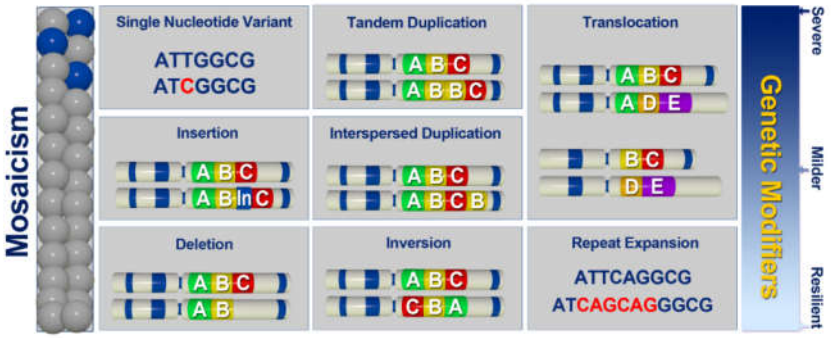

Figure 2. Uncovering missing heritability. A spectrum of variants, beyond the SNVs (single nucleotide variants), contributes to human genetic conditions as either germline or somatic variations. In addition, different types of variants, such as large insertions (including mobile element insertions (MEI)), deletions, duplications, as well as translocations, inversions, repeat expansions and other complex changes may be the source of genetic modifiers with the capacity to alleviate or exacerbate the effect of the primary pathogenic variant, and thus contribute to phenotypic variability (severe-mild-none).

2.3. Phenotypic Variability

MULTILOCUS GENETIC INHERITANCE contributes to phenotypic variability and subsequent diagnostic difficulty in patients with RDs. With the advent of HTS, it had been recognized that phenotypic variability or atypical presentation of a disease may be due to two or more genetic conditions with overlapping (blended) or discrete (composite) manifestations [25–27]. Newly discovered genetic conditions may also co-occur with another genetic condition(s) [28,29] (e.g., *NPL* and *GJB2* composite effects in a patient with sialuria, exercise intolerance/muscle wasting, cardiac symptoms, and deafness) [28]. Thus, considering multiple diagnoses in a patient is important in presumed monogenic disorders, especially the ones with atypical 'ultra' rare phenotypes [30] and/or substantial phenotypic variability [31] before a conclusion on expanded clinical presentation of a monogenic disease is made.

Beyond composite and/or blended effects of two or more genetic conditions, an increasing number of RDs is being reported where mutations in two or more genes need to co-occur for the disease to manifest. DIDA [32], a database on digenic diseases compiles information on 44 different digenic diseases and 213 of their corresponding digenic combinations [33]. For example, in ciliopathies, digenic compound heterozygous inheritance is repeatedly reported (e.g., Joubert syndrome; one heterozygous variant in *CEP41* and another in *KIF7*) [34]. Importantly, recent findings suggest that oligogenic inheritance may explain missing heritability problem in multiple genetic diseases classically considered to be monogenic, such as Long QT [35] syndrome, Holoprosencephaly [36] and others [33,35].

GENETIC MODIFIERS are important contributors to phenotypic variability (Figure 2). As modulators, these variants may alleviate or exacerbate the effect of the primary pathogenic variant leading to variable penetrance and expressivity of RDs and poor genotype-phenotype correlations even among the siblings. The extent of variation of any individual genome, combined with a known/expected property of genetic modifiers (variants of modest effects, not necessarily rare, also likely to affect non-coding regions) makes it difficult to identify these in small patient cohorts, typical for RDs. However, large-scale sequencing projects that combine phenotypic information are proving to be invaluable resources for assessing penetrance and expressivity in RDs [37,38], and thus the potential effect of genetic modifiers [39,40].

2.4. Unknowns

UNKNOWN GENE-DISEASE ASSOCIATIONS contribute to missing heritability in RDs. OMIM (Online Mendelian Inheritance in Man) database [6], daily updated, makes the inventory of the described and

published disease-related phenotypes with the causing genes and variants. To date, OMIM contains information on more than 15,000 genes and more than 8000 human disease phenotypes with a suspected Mendelian basis [6]. However, for more than 3000 phenotypes there is no known molecular basis of the disease. Given the rate at which new gene-disease associations are established [41], it is expected that the next decade will establish the majority of the currently unknown gene-disease associations, and thus facilitate better diagnostic success in patients with RDs.

UNKNOWN GENETIC MECHANISMS continue to be an important possible cause of missing heritability in RDs. For example, non-coding genome (~98% of the human genome) remains largely unexplored, yet emerging studies reinforce the importance of considering these variants in RD patients [42]. Similarly, recently described promoter epimutation [43] or allelic imbalance due to untranslated (UTR) variations [44] are some examples of not routinely screened genetic mechanisms that may cause unexplained RDs.

3. High Throughput Sequencing—Untangling Complexity

3.1. Exome Sequencing

Over the last decade, HTS has had a substantial impact on RDs by improving the likelihood of reaching a diagnosis. In particular, exome sequencing has emerged as an endorsed approach, mainly due to its cost-effectiveness and practicality.

GENE PANEL SEQUENCING refers to a type of HTS approach where a subset of known disease regions or known disease genes is targeted for sequencing. Gene panels can be of various sizes, from only two genes to thousands of genes, with the most comprehensive panels targeting all exons of the genes currently known to be associated with monogenic disease (e.g., Illumina's TruSight One ~4800 genes or TruSight One Expanded ~6700 genes). Panels offer the advantage of limiting the search for pathogenic variants to known disease gene set [45,46]; thus, circumventing the need for time-consuming interpretation of potentially unrelated variants and/or incidental findings (IFs). However, gene panels may result in missed or incomplete diagnoses, due to limited ability to address: (1) heterogeneity, (2) variability due to multiple diagnoses where one or more conditions may not be included on the panel, (3) novel genetic diseases and/or (4) genetic mechanisms of the disease due to limited capacity of the panel to detect a spectrum of gene-inactivation mechanisms.

WHOLE EXOME SEQUENCING (WES), on the other hand, simultaneously targets an entire set of protein-coding genes and allows a more comprehensive approach to uncovering missing heritability in RDs. An effective compromise between cost-effectiveness (e.g., targeting exome, a small part of the genome, <2%) and inclusion (e.g., most of the coding gene regions), WES had enabled unprecedented discoveries. These include, but are not limited to, discoveries of novel gene-disease and genotype-phenotype associations [6], unexpected role of somatic mosaicism in undiagnosed cohorts [15,47,48], as well as novel discoveries of causes of phenotypic variability (e.g., multiple genetic diagnoses in a single patient [25–27]). Moreover, WES effectively improved the diagnostic success rate well beyond the ~10% diagnostic rate of high resolution molecular single-gene tests (e.g., Sanger sequencing), low resolution genome-wide cytogenetic tests (e.g., G-banded karyotype) or microarrays [3]. While the diagnostic rate of WES varies widely depending on disease type, patient selection and type of the WES test (e.g., singleton-WES analyzing only the proband vs. trio-WES including the proband and two unaffected relatives, in most cases parents) [3], the overall diagnostic rate of trio-WES for RDs is estimated to be between 30% and 50% [3,49,50]. While WES had played a pivotal role in addressing multiple levels of complexity associated with deciphering RDs, it is still a test limited to a very small portion of a genome and exome-capture technologies [51]. This limitation of WES may explain persistent missing heritability in RDs, including the RDs with well-established clinical diagnosis [15,52].

3.2. Genome Sequencing

Unlike targeted sequencing approaches, whole genome sequencing (WGS) enables untargeted view of the entire human genome, and thus is the most comprehensive test with the potential to

identify every genetic variation that plays a role in human disease, causing either primary or secondary clinical features, or modifying the primary disease-causing variant (Figure 2). However, since sequencing human genomes became affordable, there have been mixed reports on the benefits of genome sequencing as opposed to exome sequencing in RDs. Some report marginal benefit [3,53,54], while others report a substantial benefit [55,56]. Nonetheless, all of these studies demonstrate that WGS facilitates discoveries not possible using exome sequencing (Table 1). For example, we recently reported on a family with a biochemical diagnosis of Dihydropyrimidine Dehydrogenase Deficiency (DPDD) in three members of one family [52]. Thus far, the only known genetic cause of DPDD is the alteration of *DPYD* resulting in autosomal recessive inheritance. While one member of the family received a genetic diagnosis (compound heterozygote for two *DPYD* variants), two family members with a confirmed biochemical DPDD remained only with partial genetic diagnosis despite clinical genetic tests including WES. Indeed, only one heterozygous *DPYD* variant was identified in these individuals, while the second variant expected for this recessive condition was missing [52]. It was only by WGS (Illumina short-read) that we were able to resolve the 'missing heritability' problem in this family, which was due to a complex SV, an imperfect >100Kb inversion with breakpoints in introns 8 and 12 and 4 bp deletion in *DPYD* [52]. Recently, a role of short repeat expansions in 'missing heritability' was demonstrated by identifying a cause of Benign Adult Familial Myoclonic Epilepsy (BAFME) [57]. Using single-molecule, real-time sequencing of BAC clones and Nanopore sequencing of genomic DNA, Ishiura et al. (2018) identified the same abnormal expansions of TTTCA and TTTTA repeats in introns of several different genes (*SAMD12*, *TNRC6A* and *RAPGEF2*), suggesting that it is the repeat expansion that is the cause of pathogenesis in BAFME rather than one of these genes specifically [57]. These and other examples (Table 1) clearly show the potential of WGS to uncover missing heritability, in particular variants other than SNVs, as well as variants located in a region not captured by WES, such as deep intronic variants (Table 1). In fact, Brendan Maher, who broached the concept of missing heritability over a decade ago, had already suggested that perhaps it makes sense to stop relying on SNV-gnostic technologies (e.g., GWAS in common disease and exome sequencing in RDs), and start looking for other types of variation as structural variants (SVs) via genome sequencing [1]. Although it is clear that WGS surpasses exome sequencing in its ability to uncover more (Table 1), the question remains whether it is possible to enhance the discovery and diagnostic potential of WGS beyond the currently reported rates [3,55].

SHORT-READ SEQUENCING is a type of HTS also known as second- or next-generation sequencing that could be further sub-divided into two categories: (1) sequencing by ligation (e.g., Complete Genomics and SOLiD platforms) and (2) sequencing by synthesis (proposed by Illumina, Qiagen, 454 pyrosequencing and IonTorrent platforms). These sequencing approaches allow high-throughput analyses with low error rate (Illumina accuracy rate >99.5%) and affordable per base costs. However, the short reads (typically 100 to 400 bp in length [8]) are challenging for accurate mapping (e.g., resolution of pseudogenes) and the detection of SVs [58].

LONG-READ SEQUENCING is a type of HTS known as third-generation sequencing that also could be sub-divided into two main categories: (1) single-molecule real-time sequencing approaches (SMRT, e.g., Pacific BioSciences, PacBio [59] and MinION, PromethION from Oxford Nanopore Technologies [60–62] and (2) synthetic long-read approaches proposed by Illumina and 10X Genomics.

Table 1. Examples of diagnoses facilitated by Whole Genome Sequencing (WGS).

Authors	Year	Gene	Disease	Type of Variation	Type of WGS	Ref.
Kloosterman et al.	2011	Multiple	Severe congenital abnormalities	De novo SV (chromothripsis)	SOLiD	[23]
Gilissen et al.	2014	*SHANK3*	Phelan-McDermid syndrome	De novo 66 kb deletion	Complete Genomics	[53]
Gilissen et al.	2014	*VPS13B*	Cohen syndrome	1.7 kb and 122 kb deletions	Complete Genomics	[53]
Gilissen et al.	2014	*MECP2*	Rett syndrome	De novo 0.6 kb deletion	Complete Genomics	[53]
Gilissen et al.	2014	*IQSEC2*	Intellectual disability	De novo 62 kb interspersed duplication	Complete Genomics	[53]
Gilissen et al.	2014	*SMC1A*	Cornelia de Lange syndrome	De novo 2.1 kb deletion	Complete Genomics	[53]
Gilissen et al.	2014	Multiple	16p11.2 deletion syndrome	De novo 611 kb deletion	Complete Genomics	[53]
Gilissen et al.	2014	*STAG1*	Intellectual Disability	De novo 382 kb deletion	Complete Genomics	[53]
van Kuilenburg et al.	2017	*DPYD*	DPDD	Large intragenic inversion	Illumina	[52]
Chiu et al.	2017	Multiple	Pulmonary alveolar proteinosis	425 kb deletion	Illumina	[63]
Borràs et al.	2017	*PKD1*	Polycystic kidney disease	Various, 18/19 probands	PacBio	[64]
Cretu Stancu et al.	2017	Multiple	Severe congenital abnormalities	De novo SV (chromothripsis)	ONT [1] + Illumina	[65]
Alfares et al.	2018	*PHOX2B*	Central hypoventilation syndrome	GCN (25) repeat expansion [+25]	Illumina	[54]
Alfares et al.	2018	*TPM3*	Nemaline myopathy 1	Large deletion	Illumina	[54]
Alfares et al.	2018	*TSC2*	Tuberous sclerosis type 2	De novo deep intronic SNV	Illumina	[54]
Lionel et al. [2]	2018	*GPR143*	Ocular albinism	Deep intronic variant	Illumina	[55]
Lionel et al. [2]	2018	*OTC*	Ornithine transcarbamylase deficiency	Deep intronic variant	Illumina	[55]
Ostrander et al.	2018	Multiple	Global developmental delay	Balanced inverted translocation	Illumina	[56]
Ostrander et al.	2018	*CDKL5*	Global developmental delay	De novo 63 kb tandem duplication	Illumina	[56]
Tavares et al.	2018	*BBS1*	Bardet-Biedl syndrome	Retrotransposon insertion	Illumina	[66]
Cowley et al.	2018	*SYNGAP1*	Epileptic encephalopathy	De novo 13 bp duplication	Illumina	[67]
Miao et al.	2018	*G6PC*	Glycogen storage disease type Ia	7.1 kb deletion	ONT [1]	[68]
Merker et al.	2018	*PRKAR1A*	Carney complex	De novo 2184 bp deletion	PacBio	[69]
Sanchis-Juan et al.	2018	*ARID1B*	Coffin-Siris syndrome	De novo complex SV dupINVinvDEL	Illumina	[70]
Sanchis-Juan et al.	2018	*HNRNPU*	Seizures; Intellectual disability	De novo complex SV delINVdup	Illumina	[70]
Sanchis-Juan et al.	2018	*CEP78*	Cone-rod dystrophy; Hearing loss	complex homozygous SV delINVdel	Illumina	[70]
Sanchis-Juan et al.	2018	*CDKL5*	Birth asphyxia; Fetal distress	De novo complex SV dupINVdup	Illumina + ONT [1]	[70]
Ishiura et al.	2018	*SAMD12*	BAFME [3]	TTTCA and TTTTA repeat expansions	PacBio +ONT	[57]
Ishiura et al.	2018	*TNRC6A*	BAFME [3]	TTTCA and TTTTA repeat expansions	PacBio + ONT	[57]
Ishiura et al.	2018	*RAPGEF2*	BAFME [3]	TTTCA and TTTTA repeat expansions	PacBio + ONT	[57]
Mizuguchi et al.	2019	*SAMD12*	BAFME [3]	4.6 kb intronic repeat insertion	PacBio	[71]

[1] Oxford Nanopore Tech. [2] Lionel et al., reported 18 diagnoses by WGS; however, the majority was missed by exome panels since panels did not include the corresponding gene. The two deep intronic variants included in this table would not have been detected by exome sequencing approaches. [3] Benign adult familial myoclonic epilepsy.

The Nanopore sequencers are able to produce on average 7–8 kb long reads and PacBio 10–15 kb long reads which may facilitate better detection of SVs as a result of more accurate alignments and better likelihood for detection of repetitive regions and tandem repeats [72]. However, there are many limitations associated with long-read sequencing technology, such as (1) significantly lower throughput; (2) higher per-sample sequencing cost (e.g., human WGS at 30Xcoverage is ~30-fold more expensive using PacBio than Illumina); (3) high error rates of >10% [8,73]; and (4) less resources of the available bioinformatics tools.

HOLISTIC/COMPREHENSIVE APPROACHES: Despite the advantages and disadvantages of both the short- and long-read sequencing technologies, both of these were successfully utilized to uncover a spectrum of SVs not easily/detectable by other approaches (Table 1). For example, short-read sequencing WGS successfully detected variants, such as deletions, duplications, inversions, repeat expansions, translocations, mobile element insertions, as well as complex structural variants (e.g., duplication-inversion-inversion-deletion or chromothripsis) (Table 1). Similarly, long-read sequencing had been successfully applied to detect SVs (Table 1). A combined approach may also be a possibility, as demonstrated by several studies where combining Nanopore and Illumina technologies (Table 1) helped resolve complex SVs [65,70] or synthetic long-read technology may be considered (10X Genomics/Illumina). This technology re-builds long reads *in silico* using barcodes in existing short-reads, and thus could potentially bypass issues related to the cost, error rates, and throughput of true long-read sequencers [73]. Nonetheless, we believe that in order to maximize holistic potential of WGS, besides the detection of a variation spectrum (Figures 1 and 2, Table 1), good coverage is desired in order to reliably call variants in both homozygous and heterozygous states, as well as somatic mosaicism, an emerging cause of missing heritability [15,47,48].

Currently, short-read sequencing technology has been very well positioned to lead the way in comprehensive genomics (Table 1), and the emerging computational approaches may effectively address the limitations of short-reads [8] (Table 2). For example, the recently developed ExpansionHunter uses PCR-free WGS short-read data to identify long repeat expansions, addressing the problem of identifying repetitive variation that is longer than the sequencing read itself [74]. Considering that just some 20 tandem repeat diseases have been described to date [75], and the fact that the repeatome (all repetitive or repeat-derived DNA sequences in a genome) represents a substantial source of variation in humans [75–77], is suggestive that with tools like ExpansionHunter [74] and GangSTR [78], we are likely to uncover many more causes of missing heritability (both germline [57] and somatic, Figure 2). Beyond the repeatome, other SVs represent a substantial potential for individual variation [79] (estimated to be up to 10-fold larger than that of SNVs) [80], and mobile elements (~45% of the human genome [81]) also play an important role (Table 1) [82]. Many tools had been developed for a specific type of SVs and continue to be tested and evaluated (Table 2). Genome sequencing has already been shown to be at least as sensitive as microarrays in discovery of CNVs, both germline, de novo and somatic [83], using Canvas [84,85] (Table 2), and data mining/machine learning algorithms are being developed to assess performance and to merge calls from various SV-calling algorithms [86,87].

Table 2. Examples of bioinformatics tools that facilitate comprehensive genome analyses.

Authors	Year	Tool	Method	Input [1]	Variants Detected	Reference
Abyzov et al.	2011	CNVnator	Read Depth	PE[2] Short read WGS	Copy Number Variants	[88]
Rausch et al.	2012	DELLY	Paired-ends, Read depth, Split-reads	Short read WGS	Structural Variants	[89]
Calabrese et al.	2014	MToolBox	Read re-alignment	WGS or WES	Mitochondrial Variants	[90]
Layer et al.	2014	LUMPY	Paired-ends, Read depth, Split-reads	PE short read WGS	Structural Variants	[91]
Roller et al.	2016	Canvas	Read Depth	WGS or WES	Copy Number Variations	[84,85]
Chen et al.	2016	Manta	Pair Read, Split Read	PE short read WGS	Indels, Structural Variants	[92]
Dolzhenko et al.	2017	Expansion-Hunter	Sequence-graph	PE short read WGS	Large Expansion of Short Tandem Repeats	[74]
Ebler et al.	2017	DIGTYPER	Breakpoint-Spanning, Split Alignments	PE short read WGS	Inversions, Tandem Duplications	[93]
Liang et al.	2017	Seeksv	Split Read, Discordant Paired-End, Read Depth, 2 Ends Unmapped	SE/PE [2] short read WGS	Structural Variants + Virus Integration	[94]
Mousavi et al.	2018	GangSTR	Enclosing, Fully Repetitive, Spanning and Off-target Fully Repetitive Read Pairs	PE short read WGS	Tandem Repeat expansions	[78]
Kim et al.	2018	Strelka2	Mixture-model	PE short read WGS	Single Nucleotide Variants, Indels	[95]
Ye et al.	2018	Pindel	Split-reads	PE short read WGS	Indels, Structural Variants (small and medium-size)	[96,97]
Wala et al.	2018	SvABA	Local assembly	PE short read WGS	Indels, Structural Variants (20–300 bp)	[98]
Becker et al.	2018	SVE/ FusorSV	8 SV callers combination + Data mining	PE short read WGS	Deletions + Duplications + Inversions [3]	[86]
Antaki et al.	2018	SV2	Supervised support vector machine classifiers	PE short read WGS	Deletions + Duplications	[87]

[1] All tools take BAM files as input. MToolBoxaccepts FASTQ files. Strelka2, SV2, SvABA, ExpansionHunter, Manta also accept CRAM files, SV2 requires SVs to genotype, SNV VCF files and PED files. SVE/FusorSV accepts FASTQ, BAM and VCF files. SvABA also accepts SAM files. [2] PE = Paired-Ended; SE = Single-Ended
[3] Other SVs could be explored if they are present in the training dataset.

3.3. Genome and Phenome Resources

REFERENCE GENOME: A crucial step of HTS bioinformatics pipelines is the read mapping with the following scenarios: (1) alignment along a reference genome; (2) alignment along a personalized genome; (3) de novo alignment or (4) alignment-free process. The most widely used approach is the alignment along a reference genome. A human reference genome is an assembly of sequenced DNA from a number of people, which is stored in a database in its digital form. It provides a haploid mosaic of different DNA from each donor, and thus not any single person in particular. For example, the Genome Reference Consortium human genome, build 37 (GRCh37/hg19) released in February 2009, is derived from 13 anonymous volunteers from Buffalo, New York [99], and the new build GRCh38/hg38 (release in December 2013) contains the same DNA but with more than 100 gaps that were present in hg19 now closed in hg38, some using Nanopore sequencing [100]. One disadvantage of the widely used read mapping via the reference genome approach is the assumption that the 13 volunteer genomes are representative of the genetic background of various populations subjected to genome/exome sequencing, which is unlikely to be the case. First, it has been shown that the human reference genome contains only an allele of O blood type of the ABO blood groups [101] and misses segments of DNA present in other populations [102], and additionally, it harbors some 20,000 ultra-low frequency alleles [103]. Thus, alternative approaches, such as ethnically concordant synthetic human reference sequence [104] or genome graphs (a mathematical graph of variation missing from the reference) [105] may play an important role in improving unique read mapping and variant calling for disease-associated variants [104,105], and thus further help to address the problem of missing heritability.

VARIOME RESOURCES: Another crucial component of the rare disease HTS bioinformatics pipelines is the assessment of the frequency of the variants identified in the patient by comparison against 'untargeted populations' or 'normal variation' databases. This step in variant interpretation can reduce the number of candidate variants several fold by deprioritizing the ones seen more frequently than expected in these databases, and thus focus analysis on the ultra/rare variants that are more likely to play a role. dbSNP [106], and databases such as Exome Aggregation Consortium (ExAC, 60,706 individual exomes) [37], DiscovEHR (50,726 individual exomes [38]), Genome Aggregation Database (gnomAD, 125,748 exomes and 15,708 genomes) [37] and TOPMed project BRAVO dataset (62,784 genomes) [107], aggregate exome/genome data on thousands of unrelated individuals not affected by severe pediatric genetic conditions, and thus represent invaluable resources. Even so, despite their large number of exomes/genomes, these databases are not representative of the global human population and variations, making interpretation difficult, especially in underrepresented populations (Figure 3). First, all of these resources use the GRCh37/hg19 and/or GRCh38/hg38 as the reference genome when calling the variants. Second, all of these resources predominantly contain the information on European 'normal' variation (e.g., 60% and 55% of ExAC and gnomAD data sets, respectively) (Figure 3), while other genomes are substantially under-represented (e.g., 67 Japanese individuals in gnomAD) or not at all (no information on Indigenous people) (Figure 3). This problem has been recognized and multiple efforts have been initiated to bridge these gaps, such as Iranome project [108], the Ashkenazi Jewish [109] reference panel, the Genome Russia project [110,111], as well as the Silent Genomes project (Canadian Indigenous people) [112]. Beyond these challenges with reference population data, another problem with the current population databases is that these aggregate predominantly SNVs. Thus, to effectively use WGS to uncover missing heritability, we will need both equitable representation of populations as well as robust methods to identify, compile and compare SVs across different populations.

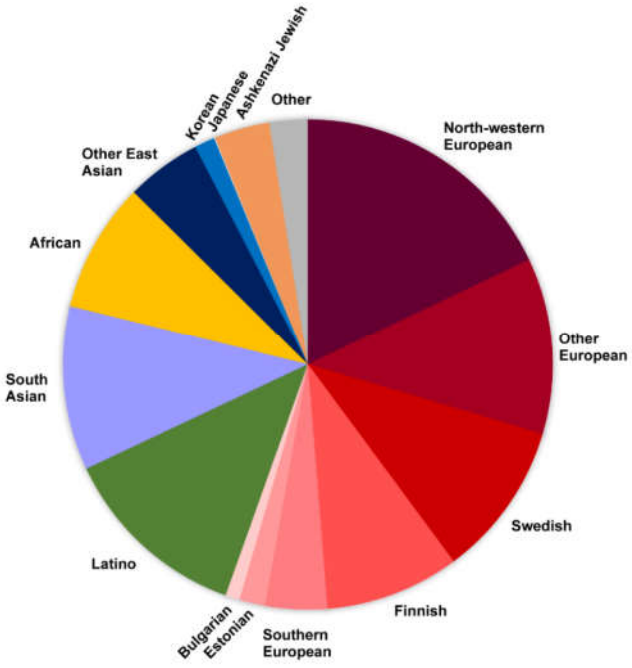

Figure 3. Populations represented in the gnomAD database. An example of various population exomes/genomes aggregated in the most comprehensive database, gnomAD (European populations are depicted in a spectrum of red colors).

Beyond the 'normal variation' resources, databases on variants already implicated in human disease are very important as well. These include already mentioned freely accessible database ClinVar [18], as well as Leiden Open Variation Database (LOVD) [20], Human Gene Mutation Database (HGMD) [113] and ClinGen resources [114]. Additional more specialized databases compile information on structural variants, such as a dbVar [115], a database housing over 3 million submitted

structural variants (SSV) from 120 human studies or an HmtVar [116], a dataset of over 40,000 human mitochondrial variants.

PHENOME RESOURCES: Accurate and detailed phenotyping is essential for correct and timely gene/variant-disease associations. Beyond the resources on human genetic variations, the resources on human phenomes, such as OMIM [6] and ORPHANET [5] compile the information on human rare phenotypes, as well as information on corresponding genes in cases where the associations had been made. The Human Phenotype Ontology (HPO) database contains HPO terms, a standardized vocabulary used to describe/communicate phenotypic abnormalities associated with disorders [117]. The HPO vocabulary not only helps link genes to diseases but also helps in standardizing health records around the world and thus connecting patients with the same disease [118]. In terms of matchmaking tools, there are a number of resources that facilitated the matching of patients with similar rare phenotypes who may have the same candidate gene identified from exome/genome sequencing studies. These include GeneMatcher [119], PhenomeCentral [120], as well as Matchmaker Exchange [121]. Since thousands of genes remain to be associated with rare disease, these matchmaking tools are effectively helping the missing heritability problem (e.g., by providing additional evidence; more than one patient with the same novel genotype-phenotype association). Similarly, international efforts, like the International Rare Diseases Research Consortium (IRDiRC) [49], Canadian Organization for Rare Diseases (CORD), UK10k project [122], the National Institute of Health (NIH) initiatives, Undiagnosed Diseases Program [123] and others are determined to work together in order to resolve the missing heritability in RDs and to understand the genetic origin of disease [124].

4. Uncovering Missing Heritability—"No Longer Just Looking under the Lamppost"

In his William Allan Award address, Dr. Francis Collins used an "under the lamppost" search metaphor to illustrate his view of the difficulty associated with searching for genetic answers in the small regions of the genome only [124]. It relates to the story of a man losing his car keys in the street at night. He was only looking under the lamppost justifying that this is where he is likely to find his keys since this is where the light is. It is clear that in RDs, we are exhausting the "lamppost", and thus it is time to search beyond for causes of "missing heritability". With affordable sequencing of genomes, we are undeniably *en route* to find more variations (Table 1), to be inclusive of underrepresented populations (Figure 3), and well positioned to comb the genome base-by-base for answers. The search beyond the obvious truly opens windows to the wonders of genomics, and while it untangles some complexity, it informs us of another complexity of human genetic conditions that we did not even consider (e.g., complex mosaicisms [47], chromothripsis [23,24]).

In this review, we discuss the 'missing heritability' paradigm through the rare disease lens. Heritability (H^2) $H^2 = \frac{\text{Var G}}{\text{Var P}} = \frac{\text{Var G}}{\text{Var G + Var E}}$ is a measure that estimates the proportion (0 to 1) of a phenotypic trait or phenotypic variance (Var P) that is genetic (Var G) in origin (i.e., it could not be explained by the environment (Var E) or random chance). We argue that missing heritability affects RDs in a fashion similar to common and complex diseases. Furthermore, we believe that given the fact that the majority of rare disease phenotypes are mostly due to genetics (Var G), RDs are the best phenotypic traits where causes of missing heritability, applicable also to common disease, can be effectively explored.

Funding: The authors were supported by Genome Canada (275SIL)/Genome BC/CIHR (GP1-155868) LSARP Genomics and Precision Health Silent Genomes Project and Alberta Children's Hospital Research Institute Foundation.

Acknowledgments: We thank Drs. Arbour, Lehman, Mwenifumbo and Stasiuk for thoughtful comments on the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Maher, B. Personal genomes: The case of the missing heritability. *Nature* **2008**, *456*, 18–21.
2. Turkheimer, E. Still missing. *Res. Hum. Dev.* **2011**, *8*, 227–241.
3. Wright, C.F.; FitzPatrick, D.R.; Firth, H.V. Paediatric genomics: Diagnosing rare disease in children. *Nat. Rev. Genet.* **2018**, *19*, 253–268.
4. Montserrat Moliner, A.; Waligóra, J. The European Union policy in the field of rare diseases. *Public Health Genomics* **2013**, *16*, 268–277.
5. Orphanet. Available online: https://www.orpha.net/consor/cgi-bin/index.php (accessed on Jan 6, 2019).
6. OMIM - Online Mendelian Inheritance in Man Available online: https://www.omim.org/ (accessed on Jan 6, 2019).
7. Chakravorty, S.; Hegde, M. Gene and variant annotation for Mendelian disorders in the era of advanced sequencing technologies. *Annu. Rev. Genomics Hum. Genet.* **2017**, *18*, 229–256.
8. Caspar, S.M.; Dubacher, N.; Kopps, A.M.; Meienberg, J.; Henggeler, C.; Matyas, G. Clinical sequencing: From raw data to diagnosis with lifetime value. *Clin. Genet.* **2018**, *93*, 508–519.
9. Prokop, J.W.; May, T.; Strong, K.; Bilinovich, S.M.; Bupp, C.; Rajasekaran, S.; Worthey, E.A.; Lazar, J. Genome sequencing in the clinic: The past, present, and future of genomic medicine. *Physiol. Genom.* **2018**, *50*, 563–579.
10. Schindler, A.; Sumner, C.; Hoover-Fong, J.E. *TRPV4*-Associated Disorders. In *GeneReviews®*; Adam, M.P., Ardinger, H.H., Pagon, R.A., Wallace, S.E., Bean, L.J., Stephens, K., Amemiya, A., Eds.; University of Washington, Seattle: Seattle, WA, USA, 1993.
11. Schlingmann, K.P.; Bandulik, S.; Mammen, C.; Tarailo-Graovac, M.; Holm, R.; Baumann, M.; König, J.; Lee, J.J.Y.; Drögemöller, B.; Imminger, K.; et al. Germline de novo mutations in *ATP1A1* cause renal hypomagnesemia, refractory seizures, and intellectual disability. *Am. J. Hum. Genet.* **2018**, *103*, 808–816.
12. Lassuthova, P.; Rebelo, A.P.; Ravenscroft, G.; Lamont, P.J.; Davis, M.R.; Manganelli, F.; Feely, S.M.; Bacon, C.; Brožková, D.Š.; Haberlova, J.; et al. Mutations in *ATP1A1* cause dominant Charcot-Marie-Tooth type 2. *Am. J. Hum. Genet.* **2018**, *102*, 505–514.
13. Lynch, D.S.; Chelban, V.; Vandrovcova, J.; Pittman, A.; Wood, N.W.; Houlden, H. *GLS* loss of function causes autosomal recessive spastic ataxia and optic atrophy. *Ann. Clin. Transl. Neurol.* **2018**, *5*, 216–221.
14. Rumping, L.; Büttner, B.; Maier, O.; Rehmann, H.; Lequin, M.; Schlump, J.-U.; Schmitt, B.; Schiebergen-Bronkhorst, B.; Prinsen, H.C.M.T.; Losa, M.; et al. Identification of a loss-of-function mutation in the context of glutaminase deficiency and neonatal epileptic encephalopathy. *JAMA Neurol.* **2018**.
15. Peron, A.; Au, K.S.; Northrup, H. Genetics, genomics, and genotype-phenotype correlations of TSC: Insights for clinical practice. *Am. J. Med. Genet. C Semin. Med. Genet.* **2018**, *178*, 281–290.
16. Bravo-Gil, N.; González-Del Pozo, M.; Martín-Sánchez, M.; Méndez-Vidal, C.; Rodríguez-de la Rúa, E.; Borrego, S.; Antiñolo, G. Unravelling the genetic basis of simplex Retinitis Pigmentosa cases. *Sci. Rep.* **2017**, *7*, 41937.
17. Chiurazzi, P.; Pirozzi, F. Advances in understanding - genetic basis of intellectual disability. *F1000Research* **2016**, *5*, doi:10.12688/f1000research.7134.1.
18. Landrum, M.J.; Lee, J.M.; Benson, M.; Brown, G.; Chao, C.; Chitipiralla, S.; Gu, B.; Hart, J.; Hoffman, D.; Hoover, J.; et al. ClinVar: Public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **2016**, *44*, D862-868.
19. Sequence Ontology Available online: http://www.sequenceontology.org/ (accessed on Jan 7, 2019).

20. Fokkema, I.F.A.C.; Taschner, P.E.M.; Schaafsma, G.C.P.; Celli, J.; Laros, J.F.J.; den Dunnen, J.T. LOVD v.2.0: The next generation in gene variant databases. *Hum. Mutat.* **2011**, *32*, 557–563.
21. Ulirsch, J.C.; Verboon, J.M.; Kazerounian, S.; Guo, M.H.; Yuan, D.; Ludwig, L.S.; Handsaker, R.E.; Abdulhay, N.J.; Fiorini, C.; Genovese, G.; et al. The genetic landscape of diamond-blackfan anemia. *Am. J. Hum. Genet.* **2018**, *103*, 930–947.
22. Piazza, A.; Heyer, W.-D. Homologous recombination and the formation of complex genomic rearrangements. *Trends Cell Biol.* **2019**, *29*, 135–149.
23. Kloosterman, W.P.; Guryev, V.; van Roosmalen, M.; Duran, K.J.; de Bruijn, E.; Bakker, S.C.M.; Letteboer, T.; van Nesselrooij, B.; Hochstenbach, R.; Poot, M.; et al. Chromothripsis as a mechanism driving complex de novo structural rearrangements in the germline. *Hum. Mol. Genet.* **2011**, *20*, 1916–1924.
24. McDermott, D.H.; Gao, J.-L.; Liu, Q.; Siwicki, M.; Martens, C.; Jacobs, P.; Velez, D.; Yim, E.; Bryke, C.R.; Hsu, N.; et al. Chromothriptic cure of WHIM syndrome. *Cell* **2015**, *160*, 686–699.
25. Tarailo-Graovac, M.; Shyr, C.; Ross, C.J.; Horvath, G.A.; Salvarinova, R.; Ye, X.C.; Zhang, L.-H.; Bhavsar, A.P.; Lee, J.J.Y.; Drögemöller, B.I.; et al. Exome Sequencing and the management of neurometabolic disorders. *N. Engl. J. Med.* **2016**, *374*, 2246–2255.
26. Posey, J.E.; Harel, T.; Liu, P.; Rosenfeld, J.A.; James, R.A.; Coban Akdemir, Z.H.; Walkiewicz, M.; Bi, W.; Xiao, R.; Ding, Y.; et al. Resolution of disease phenotypes resulting from multilocus genomic variation. *N. Engl. J. Med.* **2017**, *376*, 21–31.
27. Balci, T.B.; Hartley, T.; Xi, Y.; Dyment, D.A.; Beaulieu, C.L.; Bernier, F.P.; Dupuis, L.; Horvath, G.A.; Mendoza-Londono, R.; Prasad, C.; et al. Debunking Occam's razor: Diagnosing multiple genetic diseases in families by whole-exome sequencing. *Clin. Genet.* **2017**, *92*, 281–289.
28. Wen, X.-Y.; Tarailo-Graovac, M.; Brand-Arzamendi, K.; Willems, A.; Rakic, B.; Huijben, K.; Da Silva, A.; Pan, X.; El-Rass, S.; Ng, R.; et al. Sialic acid catabolism by *N*-acetylneuraminate pyruvate lyase is essential for muscle function. *JCI Insight* **2018**, *3*, doi:10.1172/jci.insight.122373.
29. Pérez-Torras, S.; Mata-Ventosa, A.; Drögemöller, B.; Tarailo-Graovac, M.; Meijer, J.; Meinsma, R.; van Cruchten, A.G.; Kulik, W.; Viel-Oliva, A.; Bidon-Chanal, A.; et al. Deficiency of perforin and hCNT1, a novel inborn error of pyrimidine metabolism, associated with a rapidly developing lethal phenotype due to multi-organ failure. *Biochim. Biophys. Acta Mol. Basis Dis.* **2019**, doi:10.1016/j.bbadis.2019.01.013.
30. Armour, C.M.; Smith, A.; Hartley, T.; Chardon, J.W.; Sawyer, S.; Schwartzentruber, J.; Hennekam, R.; Majewski, J.; Bulman, D.E.; FORGE Canada Consortium; et al. Syndrome disintegration: Exome sequencing reveals that Fitzsimmons syndrome is a co-occurrence of multiple events. *Am. J. Med. Genet. A.* **2016**, *170*, 1820–1825.
31. Sass, J.O.; Gemperle-Britschgi, C.; Tarailo-Graovac, M.; Patel, N.; Walter, M.; Jordanova, A.; Alfadhel, M.; Barić, I.; Çoker, M.; Damli-Huber, A.; et al. Unravelling 5-oxoprolinuria (pyroglutamic aciduria) due to bi-allelic *OPLAH* mutations: 20 new mutations in 14 families. *Mol. Genet. Metab.* **2016**, *119*, 44–49.
32. DIDA | DIDA is a novel database that provides for the first time detailed information on genes and associated genetic variants involved in digenic diseases, the simplest form of oligogenic inheritance. Available online: http://dida.ibsquare.be/ (accessed on Feb 21, 2019).
33. Gazzo, A.M.; Daneels, D.; Cilia, E.; Bonduelle, M.; Abramowicz, M.; Van Dooren, S.; Smits, G.; Lenaerts, T. DIDA: A curated and annotated digenic diseases database. *Nucleic Acids Res.* **2016**, *44*, D900-907.
34. Lee, J.E.; Silhavy, J.L.; Zaki, M.S.; Schroth, J.; Bielas, S.L.; Marsh, S.E.; Olvera, J.; Brancati, F.; Iannicelli, M.; Ikegami, K.; et al. *CEP41* is mutated in Joubert syndrome and is required for tubulin glutamylation at the cilium. *Nat. Genet.* **2012**, *44*, 193–199.

35. Schäffer, A.A. Digenic inheritance in medical genetics. *J. Med. Genet.* **2013**, *50*, 641–652.
36. Kim, A.; Savary, C.; Dubourg, C.; Carré, W.; Mouden, C.; Hamdi-Rozé, H.; Guyodo, H.; Douce, J.L.; FREX Consortium; GoNL Consortium; et al. Integrated clinical and omics approach to rare diseases: novel genes and oligogenic inheritance in holoprosencephaly. *Brain J. Neurol.* **2018**.
37. Lek, M.; Karczewski, K.J.; Minikel, E.V.; Samocha, K.E.; Banks, E.; Fennell, T.; O'Donnell-Luria, A.H.; Ware, J.S.; Hill, A.J.; Cummings, B.B.; et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **2016**, *536*, 285–291.
38. Dewey, F.E.; Murray, M.F.; Overton, J.D.; Habegger, L.; Leader, J.B.; Fetterolf, S.N.; O'Dushlaine, C.; Van Hout, C.V.; Staples, J.; Gonzaga-Jauregui, C.; et al. Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science* **2016**, *354*.
39. Chen, R.; Shi, L.; Hakenberg, J.; Naughton, B.; Sklar, P.; Zhang, J.; Zhou, H.; Tian, L.; Prakash, O.; Lemire, M.; et al. Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases. *Nat. Biotechnol.* **2016**, *34*, 531–538.
40. Tarailo-Graovac, M.; Zhu, J.Y.A.; Matthews, A.; van Karnebeek, C.D.M.; Wasserman, W.W. Assessment of the ExAC data set for the presence of individuals with pathogenic genotypes implicated in severe Mendelian pediatric disorders. *Genet. Med.* **2017**, *12*, 1300.
41. Wenger, A.M.; Guturu, H.; Bernstein, J.A.; Bejerano, G. Systematic reanalysis of clinical exome data yields additional diagnoses: Implications for providers. *Genet. Med.* **2017**, *19*, 209–214.
42. Short, P.J.; McRae, J.F.; Gallone, G.; Sifrim, A.; Won, H.; Geschwind, D.H.; Wright, C.F.; Firth, H.V.; FitzPatrick, D.R.; Barrett, J.C.; et al. De novo mutations in regulatory elements in neurodevelopmental disorders. *Nature* **2018**, *555*, 611–616.
43. Guéant, J.-L.; Chéry, C.; Oussalah, A.; Nadaf, J.; Coelho, D.; Josse, T.; Flayac, J.; Robert, A.; Koscinski, I.; Gastin, I.; et al. *APRDX1* mutant allele causes a MMACHC secondary epimutation in *cblC* patients. *Nat. Commun.* **2018**, *9*, 67.
44. Falkenberg, K.D.; Braverman, N.E.; Moser, A.B.; Steinberg, S.J.; Klouwer, F.C.C.; Schlüter, A.; Ruiz, M.; Pujol, A.; Engvall, M.; Naess, K.; et al. Allelic Expression imbalance promoting a mutant *PEX6* allele causes Zellweger spectrum disorder. *Am. J. Hum. Genet.* **2017**, *101*, 965–976.
45. Ece Solmaz, A.; Onay, H.; Atik, T.; Aykut, A.; Cerrah Gunes, M.; Ozalp Yuregir, O.; Bas, V.N.; Hazan, F.; Kirbiyik, O.; Ozkinay, F. Targeted multi-gene panel testing for the diagnosis of Bardet Biedl syndrome: Identification of nine novel mutations across *BBS1*, *BBS2*, *BBS4*, *BBS7*, *BBS9*, *BBS10* genes. *Eur. J. Med. Genet.* **2015**, *58*, 689–694.
46. Saudi Mendeliome Group. Comprehensive gene panels provide advantages over clinical exome sequencing for Mendelian diseases. *Genome Biol.* **2015**, *16*, 134.
47. Matthews, A.M.; Tarailo-Graovac, M.; Price, E.M.; Blydt-Hansen, I.; Ghani, A.; Drögemöller, B.I.; Robinson, W.P.; Ross, C.J.; Wasserman, W.W.; Siden, H.; et al. A de novo mosaic mutation in *SPAST* with two novel alternative alleles and chromosomal copy number variant in a boy with spastic paraplegia and autism spectrum disorder. *Eur. J. Med. Genet.* **2017**, *60*, 548–552.
48. Ragotte, R.J.; Dhanrajani, A.; Pleydell-Pearce, J.; Del Bel, K.L.; Tarailo-Graovac, M.; van Karnebeek, C.; Terry, J.; Senger, C.; McKinnon, M.L.; Seear, M.; et al. The importance of considering monogenic causes of autoimmunity: A somatic mutation in *KRAS* causing pediatric Rosai-Dorfman syndrome and systemic lupus erythematosus. *Clin. Immunol.* **2017**, *175*, 143–146.

49. Boycott, K.M.; Rath, A.; Chong, J.X.; Hartley, T.; Alkuraya, F.S.; Baynam, G.; Brookes, A.J.; Brudno, M.; Carracedo, A.; den Dunnen, J.T.; et al. International cooperation to enable the diagnosis of all rare genetic diseases. *Am. J. Hum. Genet.* **2017**, *100*, 695–705.
50. Deciphering Developmental Disorders Study. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **2015**, *519*, 223–228.
51. Tarailo-Graovac, M.; Wasserman, W.W.; Van Karnebeek, C.D.M. Impact of next-generation sequencing on diagnosis and management of neurometabolic disorders: Current advances and future perspectives. *Expert Rev. Mol. Diagn.* **2017**, *17*, 307–309.
52. Van Kuilenburg, A.B.P.; Tarailo-Graovac, M.; Meijer, J.; Drogemoller, B.; Vockley, J.; Maurer, D.; Dobritzsch, D.; Ross, C.J.; Wasserman, W.; Meinsma, R.; et al. Genome sequencing reveals a novel genetic mechanism underlying dihydropyrimidine dehydrogenase deficiency: A novel missense variant c.1700G>A and a large intragenic inversion in *DPYD* spanning intron 8 to intron 12. *Hum. Mutat.* **2018**, *39*, 947–953.
53. Gilissen, C.; Hehir-Kwa, J.Y.; Thung, D.T.; van de Vorst, M.; van Bon, B.W.M.; Willemsen, M.H.; Kwint, M.; Janssen, I.M.; Hoischen, A.; Schenck, A.; et al. Genome sequencing identifies major causes of severe intellectual disability. *Nature* **2014**, *511*, 344–347.
54. Alfares, A.; Aloraini, T.; Subaie, L.A.; Alissa, A.; Qudsi, A.A.; Alahmad, A.; Mutairi, F.A.; Alswaid, A.; Alothaim, A.; Eyaid, W.; et al. Whole-genome sequencing offers additional but limited clinical utility compared with reanalysis of whole-exome sequencing. *Genet. Med.* **2018**, *20*, 1328.
55. Lionel, A.C.; Costain, G.; Monfared, N.; Walker, S.; Reuter, M.S.; Hosseini, S.M.; Thiruvahindrapuram, B.; Merico, D.; Jobling, R.; Nalpathamkalam, T.; et al. Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. *Genet. Med.* **2018**, *20*, 435–443.
56. Ostrander, B.E.P.; Butterfield, R.J.; Pedersen, B.S.; Farrell, A.J.; Layer, R.M.; Ward, A.; Miller, C.; DiSera, T.; Filloux, F.M.; Candee, M.S.; et al. Whole-genome analysis for effective clinical diagnosis and gene discovery in early infantile epileptic encephalopathy. *NPJ Genom. Med.* **2018**, *3*, 22.
57. Ishiura, H.; Doi, K.; Mitsui, J.; Yoshimura, J.; Matsukawa, M.K.; Fujiyama, A.; Toyoshima, Y.; Kakita, A.; Takahashi, H.; Suzuki, Y.; et al. Expansions of intronic TTTCA and TTTTA repeats in benign adult familial myoclonic epilepsy. *Nat. Genet.* **2018**, *50*, 581–590.
58. Nakagawa, H.; Fujita, M. Whole genome sequencing analysis for cancer genomics and precision medicine. *Cancer Sci.* **2018**, *109*, 513–522.
59. Rhoads, A.; Au, K.F. PacBio Sequencing and Its Applications. *Genom. Proteom. Bioinform.* **2015**, *13*, 278–289.
60. Loose, M.W. The potential impact of nanopore sequencing on human genetics. *Hum. Mol. Genet.* **2017**, *26*, R202–R207.
61. Laver, T.; Harrison, J.; O'Neill, P.A.; Moore, K.; Farbos, A.; Paszkiewicz, K.; Studholme, D.J. Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomol. Detect. Quantif.* **2015**, *3*, 1–8.
62. Leggett, R.M.; Clark, M.D. A world of opportunities with nanopore sequencing. *J. Exp. Bot.* **2017**, *68*, 5419–5429.
63. Chiu, C.-Y.; Su, S.-C.; Fan, W.-L.; Lai, S.-H.; Tsai, M.-H.; Chen, S.-H.; Wong, K.-S.; Chung, W.-H. Whole-genome sequencing of a family with hereditary pulmonary alveolar proteinosis identifies a rare structural variant involving *CSF2RA*/*CRLF2*/*IL3RA* gene disruption. *Sci. Rep.* **2017**, *7*, 43469.

64. Borràs, D.M.; Vossen, R.H.A.M.; Liem, M.; Buermans, H.P.J.; Dauwerse, H.; van Heusden, D.; Gansevoort, R.T.; den Dunnen, J.T.; Janssen, B.; Peters, D.J.M.; et al. Detecting *PKD1* variants in polycystic kidney disease patients by single-molecule long-read sequencing. *Hum. Mutat.* **2017**, *38*, 870–879.
65. Cretu Stancu, M.; van Roosmalen, M.J.; Renkens, I.; Nieboer, M.M.; Middelkamp, S.; de Ligt, J.; Pregno, G.; Giachino, D.; Mandrile, G.; Espejo Valle-Inclan, J.; et al. Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nat. Commun.* **2017**, *8*, 1326.
66. Tavares, E.; Tang, C.Y.; Vig, A.; Li, S.; Billingsley, G.; Sung, W.; Vincent, A.; Thiruvahindrapuram, B.; Héon, E. Retrotransposon insertion as a novel mutational event in Bardet-Biedl syndrome. *Mol. Genet. Genom. Med.* **2018,** doi: 10.1002/mgg3.521.
67. Cowley, M.J.; Liu, Y.-C.; Oliver, K.L.; Carvill, G.; Myers, C.T.; Gayevskiy, V.; Delatycki, M.; Vlaskamp, D.R.M.; Zhu, Y.; Mefford, H.; et al. Reanalysis and optimisation of bioinformatic pipelines is critical for mutation detection. *Hum. Mutat.* **2018**. *40*, 374–379.
68. Miao, H.; Zhou, J.; Yang, Q.; Liang, F.; Wang, D.; Ma, N.; Gao, B.; Du, J.; Lin, G.; Wang, K.; et al. Long-read sequencing identified a causal structural variant in an exome-negative case and enabled preimplantation genetic diagnosis. *Hereditas* **2018**, *155*, 32.
69. Merker, J.D.; Wenger, A.M.; Sneddon, T.; Grove, M.; Zappala, Z.; Fresard, L.; Waggott, D.; Utiramerur, S.; Hou, Y.; Smith, K.S.; et al. Long-read genome sequencing identifies causal structural variation in a Mendelian disease. *Genet. Med.* **2018**, *20*, 159–163.
70. Sanchis-Juan, A.; Stephens, J.; French, C.E.; Gleadall, N.; Mégy, K.; Penkett, C.; Shamardina, O.; Stirrups, K.; Delon, I.; Dewhurst, E.; et al. Complex structural variants in Mendelian disorders: Identification and breakpoint resolution using short- and long-read genome sequencing. *Genome Med.* **2018**, *10*, 95.
71. Mizuguchi, T.; Toyota, T.; Adachi, H.; Miyake, N.; Matsumoto, N.; Miyatake, S. Detecting a long insertion variant in *SAMD12* by SMRT sequencing: Implications of long-read whole-genome sequencing for repeat expansion diseases. *J. Hum. Genet.* **2019**, *64*, 191–197.
72. Narzisi, G.; Schatz, M.C. The challenge of small-scale repeats for indel discovery. *Front. Bioeng. Biotechnol.* **2015**, *3*, 8.
73. Goodwin, S.; McPherson, J.D.; McCombie, W.R. Coming of age: Ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **2016**, *17*, 333–351.
74. Dolzhenko, E.; van Vugt, J.J.F.A.; Shaw, R.J.; Bekritsky, M.A.; van Blitterswijk, M.; Narzisi, G.; Ajay, S.S.; Rajan, V.; Lajoie, B.R.; Johnson, N.H.; et al. Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res.* **2017**, *27*, 1895–1903.
75. Hannan, A.J. Tandem repeats mediating genetic plasticity in health and disease. *Nat. Rev. Genet.* **2018**, *19*, 286–298.
76. De Koning, A.P.J.; Gu, W.; Castoe, T.A.; Batzer, M.A.; Pollock, D.D. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet.* **2011**, *7*, e1002384.
77. Tarailo-Graovac, M.; Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinform.* **2009**, *25*, 4–10.
78. Mousavi, N.; Shleizer-Burko, S.; Gymrek, M. Profiling the genome-wide landscape of tandem repeat expansions. *bioRxiv* **2018,** doi:10.1101/361162.
79. Sudmant, P.H.; Rausch, T.; Gardner, E.J.; Handsaker, R.E.; Abyzov, A.; Huddleston, J.; Zhang, Y.; Ye, K.; Jun, G.; Fritz, M.H.-Y.; et al. An integrated map of structural variation in 2,504 human genomes. *Nature* **2015**, *526*, 75–81.

80. Weischenfeldt, J.; Symmons, O.; Spitz, F.; Korbel, J.O. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat. Rev. Genet.* **2013**, *14*, 125–138.
81. Bergman, C.M.; Quesneville, H. Discovering and detecting transposable elements in genome sequences. *Brief. Bioinform.* **2007**, *8*, 382–392.
82. Tarailo-Graovac, M.; Drögemöller, B.I.; Wasserman, W.W.; Ross, C.J.D.; van den Ouweland, A.M.W.; Darin, N.; Kollberg, G.; van Karnebeek, C.D.M.; Blomqvist, M. Identification of a large intronic transposal insertion in *SLC17A5* causing sialic acid storage disease. *Orphanet J. Rare Dis.* **2017**, *12*, 28.
83. Gross, A.M.; Ajay, S.S.; Rajan, V.; Brown, C.; Bluske, K.; Burns, N.J.; Chawla, A.; Coffey, A.J.; Malhotra, A.; Scocchia, A.; et al. Copy-number variants in clinical genome sequencing: Deployment and interpretation for rare and undiagnosed disease. *Genet. Med.* **2018**, doi:10.1038/s41436-018-0295-y.
84. Roller, E.; Ivakhno, S.; Lee, S.; Royce, T.; Tanner, S. Canvas: Versatile and scalable detection of copy number variants. *Bioinformatics* **2016**, *32*, 2375–2377.
85. Ivakhno, S.; Roller, E.; Colombo, C.; Tedder, P.; Cox, A.J. Canvas SPW: Calling de novo copy number variants in pedigrees. *Bioinformatics* **2018**, *34*, 516–518.
86. Becker, T.; Lee, W.-P.; Leone, J.; Zhu, Q.; Zhang, C.; Liu, S.; Sargent, J.; Shanker, K.; Mil-Homens, A.; Cerveira, E.; et al. FusorSV: An algorithm for optimally combining data from multiple structural variation detection methods. *Genome Biol.* **2018**, *19*, 38.
87. Antaki, D.; Brandler, W.M.; Sebat, J. SV2: Accurate structural variation genotyping and de novo mutation detection from whole genomes. *Bioinformacs* **2018**, *34*, 1774–1777.
88. Abyzov, A.; Urban, A.E.; Snyder, M.; Gerstein, M. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* **2011**, *21*, 974–984.
89. Rausch, T.; Zichner, T.; Schlattl, A.; Stütz, A.M.; Benes, V.; Korbel, J.O. DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **2012**, *28*, i333–i339.
90. Calabrese, C.; Simone, D.; Diroma, M.A.; Santorsola, M.; Guttà, C.; Gasparre, G.; Picardi, E.; Pesole, G.; Attimonelli, M. MToolBox: A highly automated pipeline for heteroplasmy annotation and prioritization analysis of human mitochondrial variants in high-throughput sequencing. *Bioinformatics* **2014**, *30*, 3115–3117.
91. Layer, R.M.; Chiang, C.; Quinlan, A.R.; Hall, I.M. LUMPY: A probabilistic framework for structural variant discovery. *Genome Biol.* **2014**, *15*, R84.
92. Chen, X.; Schulz-Trieglaff, O.; Shaw, R.; Barnes, B.; Schlesinger, F.; Källberg, M.; Cox, A.J.; Kruglyak, S.; Saunders, C.T. Manta: Rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **2016**, *32*, 1220–1222.
93. Ebler, J.; Schönhuth, A.; Marschall, T. Genotyping inversions and tandem duplications. *Bioinformatics* **2017**, *33*, 4015–4023.
94. Liang, Y.; Qiu, K.; Liao, B.; Zhu, W.; Huang, X.; Li, L.; Chen, X.; Li, K. Seeksv: An accurate tool for somatic structural variation and virus integration detection. *Bioinformatics* **2017**, *33*, 184–191.
95. Kim, S.; Scheffler, K.; Halpern, A.L.; Bekritsky, M.A.; Noh, E.; Källberg, M.; Chen, X.; Kim, Y.; Beyter, D.; Krusche, P.; et al. Strelka2: Fast and accurate calling of germline and somatic variants. *Nat. Methods* **2018**, *15*, 591–594.
96. Ye, K.; Schulz, M.H.; Long, Q.; Apweiler, R.; Ning, Z. Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **2009**, *25*, 2865–2871.

97. Ye, K.; Guo, L.; Yang, X.; Lamijer, E.-W.; Raine, K.; Ning, Z. Split-read indel and structural variant calling using PINDEL. *Methods Mol. Biol.* **2018**, *1833*, 95–105.
98. Wala, J.A.; Bandopadhayay, P.; Greenwald, N.F.; O'Rourke, R.; Sharpe, T.; Stewart, C.; Schumacher, S.; Li, Y.; Weischenfeldt, J.; Yao, X.; et al. SvABA: Genome-wide detection of structural variants and indels by local assembly. *Genome Res.* **2018**, *28*, 581–591.
99. E pluribus unum. *Nat. Methods* **2010**, *7*, 331–331.
100. Jain, M.; Koren, S.; Miga, K.H.; Quick, J.; Rand, A.C.; Sasani, T.A.; Tyson, J.R.; Beggs, A.D.; Dilthey, A.T.; Fiddes, I.T.; et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **2018**, *36*, 338–345.
101. Scherer, S. *A short guide to the human genome*; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, USA, 2008; ISBN 978-0-87969-791-4.
102. Ameur, A.; Che, H.; Martin, M.; Bunikis, I.; Dahlberg, J.; Höijer, I.; Häggqvist, S.; Vezzi, F.; Nordlund, J.; Olason, P.; et al. De novo assembly of two Swedish genomes reveals missing segments from the human grch38 reference and improves variant calling of population-scale sequencing data. *Genes* **2018**, *9*, 486.
103. Magi, A.; D'Aurizio, R.; Palombo, F.; Cifola, I.; Tattini, L.; Semeraro, R.; Pippucci, T.; Giusti, B.; Romeo, G.; Abbate, R.; et al. Characterization and identification of hidden rare variants in the human genome. *BMC Genom.* **2015**, *16*, 340.
104. Dewey, F.E.; Chen, R.; Cordero, S.P.; Ormond, K.E.; Caleshu, C.; Karczewski, K.J.; Whirl-Carrillo, M.; Wheeler, M.T.; Dudley, J.T.; Byrnes, J.K.; et al. Phased whole-genome genetic risk in a family quartet using a major allele reference sequence. *PLoS Genet.* **2011**, *7*, e1002280.
105. Novak, A.M.; Hickey, G.; Garrison, E.; Blum, S.; Connelly, A.; Dilthey, A.; Eizenga, J.; Elmohamed, M.A.S.; Guthrie, S.; Kahles, A.; et al. Genome Graphs. *bioRxiv* **2017**, doi:10.1101/101378.
106. Smigielski, E.M.; Sirotkin, K.; Ward, M.; Sherry, S.T. dbSNP: A database of single nucleotide polymorphisms. *Nucleic Acids Res.* **2000**, *28*, 352–355.
107. NHLBI Trans omics for precision medicine. Available online: https://www.nhlbiwgs.org/ (accessed on Jan 7, 2019).
108. Iranome. Available online: http://www.iranome.com/about (accessed on Jan 7, 2019).
109. Lencz, T.; Yu, J.; Palmer, C.; Carmi, S.; Ben-Avraham, D.; Barzilai, N.; Bressman, S.; Darvasi, A.; Cho, J.H.; Clark, L.N.; et al. High-depth whole genome sequencing of an Ashkenazi Jewish reference panel: Enhancing sensitivity, accuracy, and imputation. *Hum. Genet.* **2018**, *137*, 343–355.
110. Oleksyk, T.K.; Brukhin, V.; O'Brien, S.J. Putting Russia on the genome map. *Science* **2015**, *350*, 747.
111. Oleksyk, T.K.; Brukhin, V.; O'Brien, S.J. The Genome Russia project: Closing the largest remaining omission on the world Genome map. *GigaScience* **2015**, *4*, 53.
112. Silent Genomes Project. Available online: https://www.bcchr.ca/silent-genomes-project (accessed on Jan 7, 2019).
113. Stenson, P.D.; Mort, M.; Ball, E.V.; Evans, K.; Hayden, M.; Heywood, S.; Hussain, M.; Phillips, A.D.; Cooper, D.N. The Human Gene Mutation Database: Towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum. Genet.* **2017**, *136*, 665–677.
114. Pawliczek, P.; Patel, R.Y.; Ashmore, L.R.; Jackson, A.R.; Bizon, C.; Nelson, T.; Powell, B.; Freimuth, R.R.; Strande, N.; Shah, N.; et al. ClinGen Allele Registry links information about genetic variants. *Hum. Mutat.* **2018**, *39*, 1690–1701.

115. Phan, L.; Hsu, J.; Tri, L.Q.M.; Willi, M.; Mansour, T.; Kai, Y.; Garner, J.; Lopez, J.; Busby, B. dbVar structural variant cluster set for data analysis and variant comparison. *F1000Research* **2016**, *5*, 673.

116. Preste, R.; Vitale, O.; Clima, R.; Gasparre, G.; Attimonelli, M. HmtVar: A new resource for human mitochondrial variations and pathogenicity data. *Nucleic Acids Res.* **2018,** *47*, D1202–D1210.

117. Köhler, S.; Carmody, L.; Vasilevsky, N.; Jacobsen, J.O.B.; Danis, D.; Gourdine, J.-P.; Gargano, M.; Harris, N.L.; Matentzoglu, N.; McMurry, J.A.; et al. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Res.* **2018**.

118. Haendel, M.A.; Chute, C.G.; Robinson, P.N. Classification, ontology, and precision medicine. *N. Engl. J. Med.* **2018**, *379*, 1452–1462.

119. Sobreira, N.; Schiettecatte, F.; Valle, D.; Hamosh, A. GeneMatcher: A matching tool for connecting investigators with an interest in the same gene. *Hum. Mutat.* **2015**, *36*, 928–930.

120. Buske, O.J.; Girdea, M.; Dumitriu, S.; Gallinger, B.; Hartley, T.; Trang, H.; Misyura, A.; Friedman, T.; Beaulieu, C.; Bone, W.P.; et al. PhenomeCentral: A portal for phenotypic and genotypic matchmaking of patients with rare genetic diseases. *Hum. Mutat.* **2015**, *36*, 931–940.

121. Philippakis, A.A.; Azzariti, D.R.; Beltran, S.; Brookes, A.J.; Brownstein, C.A.; Brudno, M.; Brunner, H.G.; Buske, O.J.; Carey, K.; Doll, C.; et al. The Matchmaker Exchange: A platform for rare disease gene discovery. *Hum. Mutat.* **2015**, *36*, 915–921.

122. Consortium, U.; Walter, K.; Min, J.L.; Huang, J.; Crooks, L.; Memari, Y.; McCarthy, S.; Perry, J.R.B.; Xu, C.; Futema, M.; et al. The UK10K project identifies rare variants in health and disease. *Nature* **2015**, *526*, 82–90.

123. Splinter, K.; Adams, D.R.; Bacino, C.A.; Bellen, H.J.; Bernstein, J.A.; Cheatle-Jarvela, A.M.; Eng, C.M.; Esteves, C.; Gahl, W.A.; Hamid, R.; et al. Effect of genetic diagnosis on patients with previously undiagnosed disease. *N. Engl. J. Med.* **2018**, *379*, 2131–2139.

124. Collins, F.S. 2005 William Allan Award address. No longer just looking under the lamppost. *Am. J. Hum. Genet.* **2006**, *79*, 421–426.