Hum Genet (2016) 135:603–614
DOI 10.1007/s00439-016-1658-6

REVIEW

Novel bioinformatic developments for exome sequencing

Stefan H. Lelieveld[1] · Joris A. Veltman[2,3] · Christian Gilissen[2]

Received: 1 February 2016 / Accepted: 15 March 2016 / Published online: 13 April 2016

Abstract With the widespread adoption of next generation sequencing technologies by the genetics community and the rapid decrease in costs per base, exome sequencing has become a standard within the repertoire of genetic experiments for both research and diagnostics. Although bioinformatics now offers standard solutions for the analysis of exome sequencing data, many challenges still remain; especially the increasing scale at which exome data are now being generated has given rise to novel challenges in how to efficiently store, analyze and interpret exome data of this magnitude. In this review we discuss some of the recent developments in bioinformatics for exome sequencing and the directions that this is taking us to. With these developments, exome sequencing is paving the way for the next big challenge, the application of whole genome sequencing.

Introduction

Bioinformatics has been central to the analysis and interpretation of exome sequencing data. Initial bioinformatics challenges concerned quality control; short read-mapping (Langmead et al. 2009; Li and Durbin 2009), variant calling (Albers et al. 2011; Li et al. 2009; McKenna et al. 2010), and variant annotation (Jager et al. 2014; Liu et al. 2013; Ng et al. 2009; Yang and Wang 2015). Most of these challenges have now been tackled to a degree that bioinformatic workflows are available to analyze and interpret exomes in a standard fashion and provide workable results (DePristo et al. 2011; Pabinger et al. 2014). Some of the original hurdles have simply become less relevant with the progression of technology giving rise to more and higher quality sequence data and longer sequence reads (e.g., the ambiguous alignment of very short sequencing reads). Nevertheless, quality control of exome sequencing data still remains a necessity to guarantee reliable downstream results. This task has now become fairly routine through the development of several software packages that facilitate the assessment of standard quality control measures for exome sequencing (Li et al. 2009; McKenna et al. 2010; Okonechnikov et al. 2016; Quinlan and Hall 2010).

With the widespread adoption of next generation sequencing (NGS) technologies by the genetics community and the rapid decrease in costs per base, exome sequencing has become a standard within the repertoire of genetic experiments for both research and diagnostics (Neveling et al. 2013; Yang et al. 2013). Although whole genome sequencing represents the ultimate genetic experiment, exomes still offer advantages in terms of costs, speed and ease of data storage and analysis. The steady increase of sequencing capacity and the widespread application of exome sequencing has allowed the sequencing of thousands of individuals and studies with hundred thousands of exomes are already in progress (Fu et al. 2013; Lohmueller et al. 2013; The Deciphering Developmental Disorders Study 2015; Walter et al. 2015). As an example, the

✉ Christian Gilissen
christian.gilissen@radboudumc.nl

[1] Department of Human Genetics, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Geert Grooteplein 10, 6525 GA Nijmegen, The Netherlands

[2] Department of Human Genetics, Donders Centre for Neuroscience, Radboudumc, Geert Grooteplein 10, 6525 GA Nijmegen, The Netherlands

[3] Department of Clinical Genetics, GROW-School for Oncology and Developmental Biology, Maastricht University Medical Centre, Universiteitssingel 50, 6229 ER Maastricht, The Netherlands

Exome Aggregation Consortium (ExAC) collected a dataset of over 60,000 individuals and will grow even larger in the nearby future (Lek et al. 2015). This scale at which exome data are now being generated has given rise to novel challenges in bioinformatics to store, analyze and interpret exome data of this magnitude (Stephens et al. 2015). In this review we will discuss some of the recent developments in bioinformatics for exome sequencing. We have summarized some of the tools that we believe may be of interest to the reader in Table 1.

More data, more storage

With growing datasets, simply storing data becomes a challenge that all laboratories will at some point need to face. Sequencing instruments typically generate FASTQ files containing all individual sequencing reads. After alignment the resulting reads are stored in the Sequence Alignment/Map (SAM) format that describes where sequence reads are mapped onto the reference genome. SAM files are usually compressed into the binary SAM (BAM) format that reduces the file size 3–4 times (Li et al. 2009). The BAM format is currently the de facto standard format for aligned reads and can be used by a large variety of downstream analysis and visualization tools (Li et al. 2009; Quinlan and Hall 2010; Thorvaldsdottir et al. 2013). Genomic variants that are subsequently identified based on the BAM file are then stored in a variants call format (VCF) (Danecek et al. 2011). The typical size of a single exome BAM file is within the range of Gigabytes whereas the VCF file is usually no more than 100 MB.

Storing less

The most straightforward method for reducing data storage needs is by simply storing less data, or by removing data as soon as possible. As an example, sequencing instruments currently only store raw images of the sequencing process for a limited time for trouble-shooting after which they are discarded. Similarly, many labs no longer keep the original raw sequencing reads (FASTQ file) after alignment since modern sequence aligners also include reads in the BAM file that are not aligned to the genome. This adds a little bit to the size of the BAM files but there is no longer any need for storing FASTQ files, since raw reads can now be regenerated from the alignment files by tools like Picard (http://picard.sourceforge.net.) and SAMtools (Li et al. 2009). This potentially reduces storage requirements by half. In addition to this, several clinical guidelines have been proposed that allow diagnostic laboratories to remove the alignment files after 1 or 2 years (Rehm et al. 2013; Weiss et al. 2013). However, although VCF files contain the primary result of the experiment it is worthwhile to keep BAM files for future analysis since they contain much more information than VCF files, for example the identification of CNVs (Krumm et al. 2012), somatic mutations (Lindhurst et al. 2011), and mitochondrial DNA variation (Samuels et al. 2013). It is not uncommon that reanalysis of FASTQ or BAM files can identify additional variants that were initially missed (Zighelboim et al. 2014).

Compression

An alternative to the straightforward removal of large files to save space is data-compression. This has already been introduced for raw sequence files that are now by default compressed with gzip.

Although the SAM/BAM format is convenient in the sense that it contains almost all information of the original reads and all details about the alignment in an intuitive fashion, it was not designed for efficient storage (Li et al. 2009). Since BAM files are already in binary format, ordinary compression algorithms cannot significantly reduce their size. However, specialized compression tools use various techniques to further reduce the size of BAM files. First of all non-essential information, e.g. read identifiers, can be removed. Secondly, the majority of the exome will be the same as the reference genome and can be stored more efficiently: Reference-based compression encodes reads based on a reference sequence and stores only positions that differ from the reference sequencing (Hsi-Yang Fritz et al. 2011; Kingsford and Patro 2015). For regions where there are no differences to the reference genome, only coordinates and depth information are retained. Lastly, individual base quality scores (or Q scores) are typically encoded as PHRED-like scores within a range of 0–40 (Ewing and Green 1998). These quality scores are used to optimize read-mapping and variant calling. However, the scale of quality scores is very fine-grained and encoding Q scores into bins reduces storage space (Hach et al. 2012; Ochoa et al. 2013). Binning quality scores often results in compression with some loss of information (lossy compression), where the original quality scores lose precision during compression. The lost precision does, however, not necessarily result in significant loss of accuracy for variant calling (Yu et al. 2015).

Based on these approaches, alternative formats such as Goby (Campagne et al. 2013), SlimGene (Kozanitis et al. 2011), CRAM (Hsi-Yang Fritz et al. 2011) and DEEZ (Hach et al. 2014), have been introduced that attempt to keep as much of the original information yet at a lower cost of disk space than BAM. In particular, the CRAM format has gained a lot of traction. Compression of a BAM file to CRAM format with the Scramble tool resulted in file reductions of 38–55 % with a compression time of a few minutes (Bonfield 2014). CRAM compression has already been applied to

Table 1 Overview of some of the novel bioinformatics tools related to the storage, analysis or interpretation of exome sequencing data

Name	Description	Website
Data-compression		
CRAMtools	Framework to compress BAM files into CRAM format	https://github.com/enasequence/cramtools
Scramble	C implementation of CRAM to compress BAM into CRAM format for faster encoding	http://sourceforge.net/projects/staden/files/io_lib/
TABIX	Tool to index and query bgzip-compressed VCF formatted files, available via SAMtools	http://sourceforge.net/projects/samtools/files/tabix/
Genotype query tools	Toolset to compress and query VCF files. Designed to compress large-scale cohorts	https://github.com/ryanlayer/gqt
Cloud tools		
CloudBurst	Cloud-based parallel read-mapping algorithm to map sequence reads to a reference	http://sourceforge.net/projects/cloudburst-bio/
Cloud aligner	Cloud-based Hadoop MapReduce-based approach mapping of sequence reads	http://cloudaligner.sourceforge.net/
Crossbow	Cloud-computing software tool that combines read-mapping and the SNP genotyping	http://bowtie-bio.sourceforge.net/crossbow/index.shtml
VAT	Variant Annotation Tool (VAT) is a Cloud-based platform to functionally annotate variants	http://vat.gersteinlab.org/
Mercury	A whole exome sequencing analysis workflow deployed In the Amazon Web Services (AWS) cloud	https://www.hgsc.bcm.edu/software/mercury
Variant prioritization tools		
CADD	Combined 63 annotations into one meta-score (C score) for the entire genome based on a SVM	http://cadd.gs.washington.edu/
Eigen	Spectral approach to the functional annotation of genetic variants in coding and non-coding regions.	http://www.columbia.edu/~ii2135/eigen.html
DANN	DANN used the same feature set and training data as CADD to train a deep neural network (DNN).	https://cbcl.ics.uci.edu/public_data/DANN/
FitCons	Predictions of pathogenicity for the entire genome based on evolutionary conservation and functional data	http://compgen.cshl.edu/fitCons/
SPANR/SPIDEX	Trained a model optimized for the prioritization of splice site variants with a deep learning approach	http://www.deepgenomics.com/spidex
HAL	Prioritization of splice site variants based on their effect of (alternative) RNA splicing	http://splicing.cs.washington.edu
PHIVE	Analysis of exome variants by computing phenotype similarity between human disease phenotypes and phenotype information from knockout experiments in model organisms	http://www.sanger.ac.uk/resources/databases/exomiser
RVIS	The Residual Variation Intolerance Score or RVIS is a gene based score to prioritize disease genes based on intolerant to genetic variation	http://genic-intolerance.org/
CNV detection		
CoNIFER	Detects rare CNVs in exome data based on sequence read-depth	http://conifer.sourceforge.net/
XHMM	Uses principal-component analysis (PCA) to normalize exome read-depth and a hidden Markov model (HMM) to detect CNVs	https://atgu.mgh.harvard.edu/xhmm/
Codex	Normalization and CNV calling procedure for whole exome sequencing data	http://www.bioconductor.org/packages/devel/bioc/html/CODEX.html
Data sharing		
ExAC	60,706 unrelated individuals sequenced as part of various disease-specific and population genetic studies	http://exac.broadinstitute.org/
DECIPHER	Database containing data from 18,533 patients who have given consent for broad data-sharing	https://decipher.sanger.ac.uk/

Table 1 continued

Name	Description	Website
Café variome	Platform to share genetic variant and phenotype data on a global scale	http://www.cafevariome.org/
GeneMatcher	Online platform designed to connect clinicians and researchers from around the world who share an interest in the same gene or genes	https://genematcher.org/
RD-connect	Platform that links up data used in rare disease research into a central resource for researchers worldwide	http://rd-connect.eu/
PhenomeCentral	Repository for secure data-sharing targeted to clinicians and scientists working in the rare disorder community	https://www.phenomecentral.org/
MatchMaker Exchange	Platform enabling matching of cases with similar phenotypic and genotypic profiles though a number of databases	http://www.matchmakerexchange.org/
Phenotypes		
Phenotips	A software tool for collecting and analyzing phenotypic information for patients with genetic disorders	https://phenotips.org/
PhenoDB	A software tool to store and analyze standardized phenotypic information	http://phenodb.net
Phenominer	A tool to extract structured phenotypes from text	http://phenominer.mml.cam.ac.uk/

tackle storage-capacity problems in large databases such as Sequence Read Archive (SRA) (Cochrane et al. 2011) and the 1000 Genomes project (http://www.1000genomes.org). The CRAM format is now supported by some widely used genomic analysis tools such as SAMtools, Picard and GATK (Li et al. 2009; McKenna et al. 2010). With the increasing support for the CRAM format, it may well replace the use of BAM files in the near future.

With ever growing datasets containing variants of thousands of individuals, it becomes worthwhile to compress the relatively small VCF files as well. The Tabix format offers a convenient compression format for large VCF files. This reduces file sizes roughly 3–5 times, and also supports indexing to perform efficient querying of genome positions (Li 2011). Some common resources are already available in Tabix format such as dbSNP (NCBI Resource Coordinators 2015) and Combined Annotation-Depended Depletion (CADD) scores (Kircher et al. 2014). Another option is to use the recently published Genotype Query Tools (GQT) to index and compress large number of VCF files. This tool also facilitates fast querying. GQT was used to compress the Exome Aggregation Consortium (ExAC) VCF file, consisting of 9.36 million exonic variants for 60,706 individuals, from 14.1 TB to 28 GB (Layer et al. 2016).

All in all, the growing need to reduced storage space is leading to new data formats for alignment and variant files and smarter algorithms to query these efficiently.

Cloud-based solutions

Compression of data is an easy and efficient way to reduce storage needs, but in the end the reduction in data sizes is limited. An alternative is to store large amounts of genomics data in the cloud. Cloud storage offers several out-of-the-box advantages to local storage: it is scalable, has default access control policies, protects against data loss, allows for auditing, data encryption, easy sharing, and automation by programmable interfaces (Fusaro et al. 2011; Stein 2010). Currently, there are multiple commercial providers of cloud services of which Amazon Web Services (AWS; https://aws.amazon.com/), Microsoft Azure (https://azure.microsoft.com) and Google cloud platform (https://cloud.google.com) are the largest. In addition there are non-profit organizations offering cloud-computing solutions such as Open Cloud Consortium (http://occ-data.org/).

Cloud storage is based on a "pay as you go" monetary model whereby one only pays for used storage that can be expanded ad hoc. Although cloud storage itself is relatively inexpensive with less than $100 for storing 1 TB of data per month, there are some additional costs to consider (Shanahan et al. 2014). While transferring data into the cloud storage is usually free of costs, analyzing and downloading data from the cloud can be relatively expensive. For example, downloading 1 TB of data from the cloud costs approximately $120 per TB (Shanahan et al. 2014). This makes it worthwhile not only to keep data in the cloud but also to perform the analysis there and only download smaller result files. Special software is, however, needed to make efficient use of the scalability of the cloud-computing platform. Currently there are already a variety of tools for cloud-based mapping of sequence reads, (Nguyen et al. 2011; Schatz 2009), genotyping (Gurtowski et al. 2012), variant annotation (Habegger et al. 2012) as well as complete cloud-based exome sequencing pipelines

(Liu et al. 2014; Reid et al. 2014). Fusaro et al. showed that the alignment of the entire genome (4 billion paired reads, 35 pb long) of a person in 48 h costing approximately $48 of cloud resources (Fusaro et al. 2011). According to Stein et al. the International Cancer Genome Consortium (ICGC) analyzed 500 genomes in the cloud for a price of $18 per sample whereas the authors estimate this would require $200 on standard computer systems (Stein et al. 2015).

Data stored in the cloud can also provide a solution for effective public and private data-sharing. For example, the Amazon Web Services (AWS) contains 1000 Genomes Project data (Clarke et al. 2012) and accommodates 1200 whole genome sequences of the ICGC. In addition, data from Ensembl and GenBank are being hosted in AWS and data transfer between AWS instances is free of charge (Fusaro et al. 2011). Furthermore, the US National Cancer Institute is exploring how the cloud could facilitate a cost-effective platform to store and share large amounts of tumor data (https://cbiit.nci.nih.gov/ncip/nci-cancer-genomics-cloud-pilots).

The uptake of cloud-based solutions by academic and non-academic hospital laboratories has been slow, likely because of practical concerns, unfamiliarity, as well as ethical and legal concerns of storing patient DNA data in the cloud (Dove et al. 2015). Although data storage in the cloud is relatively inexpensive, transferring vast quantities of sequencing data via the Internet from and into the cloud may be a considerable cost and a time-consuming process due to low network bandwidth (Schatz et al. 2010; Stein 2010). In addition, moving genetic data of patients to a third-party server introduces issues concerning security and privacy (Greenbaum et al. 2011). This has limited the use of cloud-based storage solutions for most clinical NGS applications so far. However, given the advantages and a future of routine genome sequencing, it may well be unavoidable that all genomics data end up in the cloud for analysis and for patients and their physicians to access.

Variant identification

To some extent, challenges for calling variants have become less urgent with improved exome enrichment assays, increasing sequence quality and read length and reduced sequencing prices, allowing for higher coverage sequencing of the exome in individual patients. Whereas early comparisons of whole exome capture kits showed that around 80 % of the human protein coding sequence regions were captured at a minimal coverage of 20× (Parla et al. 2011), current exome capture kits and sequencing at minimal 100× median average coverage capture more than 95 % of the coding regions with a minimal coverage of 20× (Lelieveld et al. 2015). Due to the increased coverage and improved sequencing quality for modern exomes, variant calling has become more reliable. Several studies have even demonstrated the identification of somatic mutations for rare syndromes (Lindhurst et al. 2011; Poduri et al. 2013; Sato et al. 2014), which is only possible with high coverage. These improvements in exome quality have led some laboratories to reconsider the validation of sequencing variants by the gold standard Sanger sequencing. A recent validation study found that all single nucleotide variants with a Genome Analysis Toolkit (GATK) (McKenna et al. 2010) quality score above 500 were confirmed by Sanger sequencing and estimated that only validating variants with a quality score below this threshold would reduce the Sanger confirmation workload by 70–80 % (Strom et al. 2014). Overall the significant improvements in exome sequencing quality may indeed eliminate the need for validation of high quality variants. However, the detection of SNVs in NGS data has not been fully resolved and results from different variant callers remain inconsistent (O'Rawe et al. 2013; Pabinger et al. 2014; Zook et al. 2014). In addition, small insertions/deletion (indels) are still particularly problematic to identify accurately (Jiang et al. 2015b). Highly accurate genotypes across the genome of a single individual as for example provided by the "Genome in a Bottle Consortium" may help resolve these issues in the future (Zook et al. 2014).

Detection of copy number variants

From SNVs attention has moved towards the identification of other types of variants in exome sequencing data. In particular, the identification of copy number variation (CNV) from exome data poses an attractive possibility as CNVs are an important cause of disease (Zhang et al. 2009). Genomic microarray platforms such as the SNP-array and Array CGH are the de facto standard to detect CNVs (Miller et al. 2010), whereas whole genome sequencing will likely be the preferred platform for the detection and characterization of CNVs as well as other structural variants (Gilissen et al. 2014). In contrast to microarrays and whole genome sequencing, exome sequencing targets only 1–2 % of the protein coding regions of the genome. The sparse and fragmented nature of exome data makes it more difficult to identify CNVs and methods rely mostly on depth-of-coverage approaches. For these approaches a normalized read count in a genomic window of a single individual is compared to that of other exomes. Normalization of read counts is required to counteract technical issues such as poor read mappability, GC bias, and batch effects between sequencing experiments (Teo et al. 2012). Many different algorithms have been devised based on read-depth methods, such as CODEX, Convex, Conifer, and XHMM (Amarasinghe et al. 2013; Fromer et al.

2012; Jiang et al. 2015a; Krumm et al. 2012). Comparisons of CNV algorithms for exome data have shown that none of the algorithms performed well in all situations and that the resolution is limited to at least three exome targets (de Ligt et al. 2013; Fromer et al. 2012; Krumm et al. 2012). Although this does not equal the sensitivity of high-resolution microarrays, it is comparable to that of medium resolution microarrays that are still commonly used. Studies describing the large-scale application of CNV detection only based from exome data are, however, still limited (Poultney et al. 2013), which may perhaps hint at some of the underlying difficulties to obtain robust CNV calls from exome data. The possibility to detect copy number variants in exome data is, however, a great benefit of exome sequencing that should not be ignored as CNVs contribute significantly to disease. The identification of other types of structural variants such as inversions, and the accurate prediction of CNV breakpoint remains challenging and whole genome sequencing is likely needed for this (Meienberg et al. 2015).

Variant interpretation

Sequencing the protein coding regions of a patient typically yields tens of thousands of variants of which the majority is likely to be benign and only one or perhaps two variants contribute to the disease phenotype (Bamshad et al. 2011; Gilissen et al. 2012). The most effective way of distinguishing benign from pathogenic variants is based on using population frequencies of variants. For this approach all variants occurring in the population at higher frequencies than the disease prevalence are considered as benign. Databases with collections of exome variants of individuals without clear disease phenotypes have therefore been tremendously helpful to prioritize variants in Mendelian disease. This has given rise to several initiatives for large-scale variant databases with exome data (Fu et al. 2013; Tryka et al. 2014). The largest of these, thus far, is the Exome Aggregation Consortium (ExAC) database, containing variants of more than 60,000 exomes (Lek et al. 2015). These large databases are instrumental to the interpretation of future exomes for Mendelian disease gene identification. In addition, a need for population-specific databases of variation will remain, especially for those populations that are poorly represented in the large public databases (Tennessen et al. 2012). Interpretation based on population frequency information from databases should be done with care because of the possibility of false positives (MacArthur and Tyler-Smith 2010), founder mutations (Gunel et al. 1996), somatic or tissue-specific variants (Acuna-Hidalgo et al. 2015).

Coding mutations

Although accurate population frequencies are a necessity for the interpretation of exomes, this will only reduce the number of possible candidate mutations to a couple of hundred or so (Gilissen et al. 2012). Further prioritization of pathogenic variants remains a challenging task, in particular for missense variants. Various tools such as Polyphen2 (Adzhubei et al. 2013), SIFT (Kumar et al. 2009) and PhyloP (Pollard et al. 2010), have long been used in the pathogenicity assessment of these protein coding variants based on evolutionary conservation. Unfortunately, these prioritization methods lack specificity and sensitivity to sufficiently reduce the large number of candidate mutations from exome sequencing on their own (Gilissen et al. 2012). This becomes even more apparent when considering the prioritization of non-coding variation from whole genome sequencing experiments.

However, in the last few years, novel tools have been published that are expected to offer better sensitivity and specificity compared to the traditional prioritization tools. The availability of genome-wide functional annotations of coding and non-coding variants in combination with algorithmic improvements resulted in novel tools adapted to prioritize both coding and non-coding variants. These novel tools can broadly be divided into two groups. The first group focuses on the prediction of deleterious variation by computing a functional meta-score based on integrating a variety of genome-wide annotations. Combined Annotation-Depended Depletion (CADD) is the most well-known example of such a framework that applies a support vector machine (SVM) to integrate 63 sources of functional and evolutionary data into a relative pathogenicity score (Kircher et al. 2014). Eigen (Ionita-Laza et al. 2016) and DANN (Quang et al. 2015) are other examples using different algorithmic approaches to combine large varieties of annotations into one pre-computed meta-score trained to distinguish between benign and deleterious variants. Fitness consequence (FitCons) (Gulko et al. 2015) is different in the sense that it compares patterns of divergence between the human population and primates to assess functional sites that emerged quite recently. The second group of tools attempts to specifically predict non-coding regulatory variants. DeepSEA (Zhou and Troyanskaya 2015) and DeltaSVM (Lee et al. 2015) are examples of such tools and were trained on a variety of annotations of non-coding annotations mainly derived from the ENCODE project (The ENCODE Project Consortium 2012). Notably, the DeepSEA method was based on a Deep learning algorithm, which is a form of machine learning that is increasingly being applied to biological problems (Alipanahi et al. 2015; Rusk 2016).

In spite of the potential of these tools, it remains unclear how well they perform in clinical practice because independent validation studies for these novel tools are still lacking. Moreover, such studies are hampered by a lack of sufficient validation data that have not already been used in the development of the prediction software or original benchmark (Grimm et al. 2015). Van der Velde et al. demonstrated the practical utility of CADD for the interpretation of variants. The authors applied CADD to a set of 2210 variants that were manually assessed by an expert panel and found that, beside a relatively small number of discrepancies in favor of the expert, CADD scores proved valuable for the prioritization of pathogenic variants. (van der Velde et al. 2015). However, a recent validation of CADD and other prediction tools using in vivo mouse models, found that about half of the assessed mutations that were predicted to be deleterious had little impact on the clinical phenotype (Miosge et al. 2015). This once again highlights the importance of functional validation of potential pathogenic variants.

Splice site mutations

Due to the increased read lengths, exome sequencing typically captures a large part of the extended splice site at sufficient coverage for variant identification. However, mutations in the extended splice site are typically excluded during the prioritization step because variation within these regions is more prevalent but also more difficult to interpret. Existing algorithms for splice sites such as MaxEntScan (Eng et al. 2004) and NNSplice (Reese et al. 1997) were not designed to offer predictions for the large numbers of variants from exome sequencing (Jian et al. 2014). Like for coding variants recent developments in algorithms have improved the ability to interpret this type of variants.

The SPANR (splicing-based analysis for variants) tool is another example of a "deep learning" computational model scoring the effect of variants on the mRNA-splicing. The SPANR model is trained on 1393 sequence features extracted from 10,689 alternatively spliced exons and their corresponding mRNA expression levels in 16 human tissues and offers predictions up to 300 bp within the intron (Xiong et al. 2015). The authors found that SPANR correctly predicted the direction of change in expression of the exon for 73 out of 99 (74 %) splice site mutations. Another novel splice site prediction tool called hexamer additive linear (HAL) is a model, trained on nearly two million synthetic alternatively spliced mini genes, to predict the effect of 5′ and 3′ mutations on exon skipping (Rosenberg et al. 2015). In a set of 286 variants within three genes (*CTFR*, *BRCA2* and *SMN2*) the prediction accuracy ranged from 86 to 90 %. These improvements in splice site prediction programs may open up new avenues for the interpretation of variants in exomes.

Gene prioritization

For the interpretation of exome data it is not sufficient to only determine whether a variant is likely to impair normal gene function, but also whether the function of a mutated gene is actually relevant for the disease (MacArthur et al. 2012). Two novel approaches for the interpretation of gene function have gained a lot of traction.

Phenotypic interpretation of variants in exomes (PHIVE) is an algorithm that computes phenotype similarity between human disease phenotypes and phenotype information from knockout experiments in model organisms. This gene-level information is then combined with variant pathogenicity predictions and thereby achieves better rankings of pathogenic variants in exome data (Robinson et al. 2014). A totally different approach to predict deleteriousness for genes is based on the use of population variation to determine how intolerant genes are to normal variation. Two studies independently showed that human disease genes are much more intolerant to genetic variation than other genes (Khurana et al. 2013; Petrovski et al. 2013). Several studies have already successfully used this approach to prioritize genes with likely pathogenic mutations (Allen et al. 2013; Gilissen et al. 2014).

Overall, algorithm development has leveraged the availability of genome-wide datasets such as exome sequencing project (ESP) (Fu et al. 2013), encyclopedia of DNA elements (ENCODE) (The ENCODE Project Consortium 2012) and the International Mouse Phenotype Consortium (IMPC) (Brown and Moore 2012) to provide improved pathogenicity predictions for variants and to cope with exome-sized variant datasets. These novel algorithms represent our first steps to the next big challenge, the interpretation of non-coding variation from whole genome sequencing experiments. In the meanwhile, technologies for high-throughput functional assays are under development that may produce the high-throughput functional validations needed to improve in silico variant predictions (Findlay et al. 2014).

Finding recurrent mutations

Besides the interpretation of variants and genes, progress has also been made in the approaches to provide proof of pathogenicity for novel candidate genes. While functional proof of pathogenicity remains crucial, it is time-consuming and expensive to obtain, and requires specific expertise. An additional layer of evidence for pathogenicity of a mutation in a candidate disease gene can be obtained by identifying multiple patients with mutations in the same gene and a similar phenotype. Two different approaches for finding recurrently mutated candidate genes have emerged, depending on whether the disorder is either rare and monogenic or more common and genetically heterogeneous.

Genotype-centric approach for common genetically heterogeneous disorders

For genetically heterogeneous disorders, it is not possible to select specific subsets of patients based on their phenotype to perform a targeted analysis of the candidate gene. Therefore, screening of a large cohort of patients for additional mutations within the same candidate gene is typically performed (de Ligt et al. 2012; O'Roak et al. 2012). When costs allow, it is even more efficient to immediately screen the entire cohort by exome sequencing, rather than start with a small number of selected samples (Neale et al. 2012; O'Roak et al. 2011; The Deciphering Developmental Disorders Study 2015). In such a set-up, however, the probability of random findings becomes very large and rigorous statistics are required. Statistical methods do not only protect against potential false positive findings but are also able to take into account factors like reduced penetrance, modifiers, and multigenic effects (MacArthur et al. 2014). The first large-scale exome sequencing studies already relied on different statistical approaches based on estimates of genome-wide mutation rates to identify genes enriched for de novo mutations (Neale et al. 2012; O'Roak et al. 2011). An improved statistical framework was proposed by Samocha et al. (2014) which was first applied by the DDD project which performed large-scale trio sequencing of 1133 trios with developmental disorders (The Deciphering Developmental Disorders Study 2015). After identifying de novo coding mutations in this cohort, a statistical approach was used based on estimates of the gene specific mutation rate to identify 12 novel genes that were enriched for de novo mutations. The same group also used a novel statistical framework for the identification of recessive genes in a cohort of 4125 families with developmental disorders (Akawi et al. 2015). In this case a model was constructed to estimate the probability of drawing n unrelated families with similar biallelic genotypes by chance from the general population. Estimates of population allele frequencies for rare loss-of-function and missense variants were obtained from the Exome Aggregation Consortium data set (Lek et al. 2015). Although in both studies the cohorts are considered to be very large, statistical power was still limited and the authors emphasize that this should motivate data-sharing through international databases.

Phenotype-centric approach for rare monogenic disorders

For many Mendelian diseases the phenotype is very rare, and individual groups do not have more than a few cases making it impossible to perform large-scale screening. The alternative is then to take a phenotype-centric approach where one finds additional patients with the same, or similar, distinct phenotype. Once more patients have been identified with overlapping phenotypes specific testing of candidate genes can be performed. Alternatively, there is the opposite approach in which first patients with matching genotypes are identified and final evidence of pathogenicity comes from the matching of patient phenotypes. In both cases additional evidence is obtained not just by the common genotype, but also by the shared specific phenotype of patients with mutations. This approach is now facilitated by various data-sharing initiatives for rare diseases such as DECIPHER (Bragin et al. 2014), Café Variome (Lancaster et al. 2015), GeneMatcher (Sobreira et al. 2015), RD-connect (Thompson et al. 2014), and PhenomeCentral (Buske et al. 2015). See Brookes and Robinson for an overview of data-sharing initiatives and databases (Brookes and Robinson 2015). The matchmaker exchange is a recent initiative to integrate the information from all of these databases by providing a single interface for queries together with match-making algorithms (Philippakis et al. 2015). A nice example of a phenotype-centric approach is a recent paper on the identification *RSPRY1* by which the authors identified an additional case with the same phenotype using the Care4Rare Canada matchmaker (Faden et al. 2015). This should hopefully inspire more researchers to contribute to these databases and facilitate the identification of the genetic cause for their patients. By contributing these data to public databases they do not only become available to researchers and physicians but also to the patients themselves (Chong et al. 2015; Kirkpatrick et al. 2015). A nice illustration of this is the case of Massimo Damiani who suffered from an unclassified form of leukoencephalopathy and whose parent's efforts resulted in the genetic diagnosis (Lambertson et al. 2015). The authors argue that these patient-led efforts have the potential to increase the value of matchmaking networks.

Structured phenotypes

The probability of success for matchmaking increases with the availability of good phenotype information. A long-standing challenge with phenotype descriptions is the lack of standardization. This presents several problems such as the use of different clinical nomenclature for similar phenotypes, the uncertainty whether phenotypes are absent or not assessed, and the fact that it is unclear how phenotypes are related to each other, which makes it difficult to perform computational analyses (Kohler et al. 2014). For some years these issues have been tackled by the introduction of standardized phenotype vocabularies and ontologies. Several ontologies have been developed but one of the most used is the human phenotype ontology (HPO)

(Kohler et al. 2014). HPO currently consists of more than 250,000 phenotypic annotations (Groza et al. 2015). The practical benefits of using HPO have been demonstrated by the development of novel tools that facilitate the prioritization of exome variants (as discussed in the previous section), but also by a recent study of the DDD project. Akawi et al. used structured phenotype information to statistically quantify the phenotypic similarity of patients with developmental delay for which rare mutations were identified in the same gene (Akawi et al. 2015). Although the added value of the integrated phenotypes in the statistical assessment was limited, this will likely improve when phenotype information becomes more comprehensive. Obtaining comprehensive structured phenotypes, however, is difficult and time-consuming. The DDD project mandated the availability of phenotype information in HPO format for all of their samples (Firth and Wright 2011). Such criteria are not easily imposed for most other projects and several tools have been developed to encourage and facilitate the use of phenotype information. PhenoDB (Hamosh et al. 2013) and PhenoTips (Girdea et al. 2013) are platforms that allow clinicians to enter, store and analyze structured phenotypic data. Phenominer is a tool able to extract phenotype contexts from simple text to identify relationships between human diseases described in OMIM and literature (Collier et al. 2015). In the future even the actual measuring of phenotypes may be automated leading to more robust and objective phenotypes that will also take less time of physicians to administrate (Oellrich et al. 2015), and allowing bioinformaticians to use these data for interpretation of exome variants.

Conclusions

Here we have discussed some of ongoing bioinformatic developments that have the potential to impact the way we currently analyze and interpret exome data. It is clear that many developments in bioinformatics are still needed with respect to exome sequencing and that this is still a very active field of development. This requires a high degree of flexibility and adaptiveness from those working in this field. Especially since new challenges are already on the horizon with the anticipated large-scale application of whole genome sequencing.

References

Acuna-Hidalgo R et al (2015) Post-zygotic point mutations are an underrecognized source of de novo genomic variation. Am J Hum Genet 97:67–74. doi:10.1016/j.ajhg.2015.05.008

Adzhubei I, Jordan DM, Sunyaev SR (2013) Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet 7:7–20. doi:10.1002/0471142905.hg0720s76

Akawi N et al (2015) Discovery of four recessive developmental disorders using probabilistic genotype and phenotype matching among 4,125 families. Nat Genet 47:1363–1369. doi:10.1038/ng.3410

Albers CA, Lunter G, MacArthur DG, McVean G, Ouwehand WH, Durbin R (2011) Dindel: accurate indel calls from short-read data. Genome Res 21:961–973. doi:10.1101/gr.112326.110

Alipanahi B, Delong A, Weirauch MT, Frey BJ (2015) Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol 33:831–838. doi:10.1038/nbt.3300

Allen AS et al (2013) De novo mutations in epileptic encephalopathies. Nature 501:217–221. doi:10.1038/nature12439

Amarasinghe KC, Li J, Halgamuge SK (2013) CoNVEX: copy number variation estimation in exome sequencing data using HMM. BMC Bioinf 14(Suppl 2):S2. doi:10.1186/1471-2105-14-S2-S2

Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J (2011) Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet 12:745–755. doi:10.1038/nrg3031

Bonfield JK (2014) The scramble conversion tool. Bioinformatics 30:2818–2819. doi:10.1093/bioinformatics/btu390

Bragin E, Chatzimichali EA, Wright CF, Hurles ME, Firth HV, Bevan AP, Swaminathan GJ (2014) DECIPHER: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation. Nucleic Acids Res 42:D993–D1000. doi:10.1093/nar/gkt937

Brookes AJ, Robinson PN (2015) Human genotype-phenotype databases: aims, challenges and opportunities. Nat Rev Genet 16:702–715. doi:10.1038/nrg3932

Brown SD, Moore MW (2012) The International Mouse Phenotyping Consortium: past and future perspectives on mouse phenotyping. Mamm Genome 23:632–640. doi:10.1007/s00335-012-9427-x

Buske OJ et al (2015) PhenomeCentral: a portal for phenotypic and genotypic matchmaking of patients with rare genetic diseases. Hum Mutat 36:931–940. doi:10.1002/humu.22851

Campagne F, Dorff KC, Chambwe N, Robinson JT, Mesirov JP (2013) Compression of structured high-throughput sequencing data. PloS One 8:e79871. doi:10.1371/journal.pone.0079871

Chong JX et al (2015) Gene discovery for Mendelian conditions via social networking: de novo variants in KDM1A cause developmental delay and distinctive facial features. Genet Med. doi:10.1038/gim.2015.161

Cochrane G, Karsch-Mizrachi I, Nakamura Y (2011) The international nucleotide sequence database collaboration. Nucleic Acids Res 39:D15–D18. doi:10.1093/nar/gkq1150

Clarke L et al (2012) The 1000 Genomes Project: data management and community access. Nat Meth 9:459–462. doi:http://www.nature.com/nmeth/journal/v9/n5/abs/nmeth.1974.html#supplementary-information

Collier N, Groza T, Smedley D, Robinson PN, Oellrich A, Rebholz-Schuhmann D (2015) PhenoMiner: from text to a database of phenotypes associated with OMIM diseases. Database (Oxford) 2015. doi:10.1093/database/bav104

Danecek P et al (2011) The variant call format and VCFtools. Bioinformatics 27:2156–2158. doi:10.1093/bioinformatics/btr330

de Ligt J et al (2012) Diagnostic exome sequencing in persons with severe intellectual disability. N Engl J Med 367:1921–1929. doi:10.1056/NEJMoa1206524

de Ligt J et al (2013) Detection of clinically relevant copy number variants with whole-exome sequencing. Hum Mutat 34:1439–1448. doi:10.1002/humu.22387

DePristo MA et al (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43:491–498. doi:10.1038/ng.806

Dove ES, Joly Y, Tasse AM, Knoppers BM (2015) Genomic cloud computing: legal and ethical points to consider. Eur J Hum Genet 23:1271–1278. doi:10.1038/ejhg.2014.196

Eng L et al (2004) Nonclassical splicing mutations in the coding and noncoding regions of the ATM Gene: maximum entropy estimates of splice junction strengths. Hum Mutat 23:67–76. doi:10.1002/humu.10295

Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8:186–194

Faden M et al (2015) Identification of a recognizable progressive skeletal dysplasia caused by RSPRY1 mutations. Am J Hum Genet 97:608–615. doi:10.1016/j.ajhg.2015.08.007

Findlay GM, Boyle EA, Hause RJ, Klein JC, Shendure J (2014) Saturation editing of genomic regions by multiplex homology-directed repair. Nature 513:120–123. doi:10.1038/nature13695

Firth HV, Wright CF (2011) The deciphering developmental disorders (DDD) study. Dev Med Child Neurol 53:702–703. doi:10.1111/j.1469-8749.2011.04032.x

Fromer M et al (2012) Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. Am J Hum Genet 91:597–607. doi:10.1016/j.ajhg.2012.08.005

Fu W et al (2013) Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 493:216–220. doi:10.1038/nature11690

Fusaro VA, Patil P, Gafni E, Wall DP, Tonellato PJ (2011) Biomedical cloud computing with Amazon Web Services. PLoS Comput Biol 7:e1002147. doi:10.1371/journal.pcbi.1002147

Gilissen C, Hoischen A, Brunner HG, Veltman JA (2012) Disease gene identification strategies for exome sequencing. Eur J Hum Genet 20:490–497. doi:10.1038/ejhg.2011.258

Gilissen C et al (2014) Genome sequencing identifies major causes of severe intellectual disability. Nature 511:344–347. doi:10.1038/nature13394

Girdea M et al (2013) PhenoTips: patient phenotyping software for clinical and research use. Hum Mutat 34:1057–1065. doi:10.1002/humu.22347

Greenbaum D, Sboner A, Mu XJ, Gerstein M (2011) Genomics and privacy: implications of the new reality of closed data for the field. PLoS Comput Biol 7:e1002278. doi:10.1371/journal.pcbi.1002278

Grimm DG et al (2015) The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. Hum Mutat 36:513–523. doi:10.1002/humu.22768

Groza T et al (2015) The human phenotype ontology: semantic unification of common and rare disease. Am J Hum Genet 97:111–124. doi:10.1016/j.ajhg.2015.05.020

Gulko B, Hubisz MJ, Gronau I, Siepel A (2015) A method for calculating probabilities of fitness consequences for point mutations across the human genome. Nat Genet 47:276–283. doi:10.1038/ng.3196

Gunel M et al (1996) A founder mutation as a cause of cerebral cavernous malformation in Hispanic Americans. N Engl J Med 334:946–951. doi:10.1056/NEJM199604113341503

Gurtowski J, Schatz MC, Langmead B (2012) Genotyping in the cloud with crossbow. Curr Protoc Bioinf 39:15.3.1–15.3.15 doi:10.1002/0471250953.bi1503s39

Habegger L et al (2012) VAT: a computational framework to functionally annotate variants in personal genomes within a cloud-computing environment. Bioinformatics 28:2267–2269. doi:10.1093/bioinformatics/bts368

Hach F, Numanagic I, Alkan C, Sahinalp SC (2012) SCALCE: boosting sequence compression algorithms using locally consistent encoding. Bioinformatics 28:3051–3057. doi:10.1093/bioinformatics/bts593

Hach F, Numanagic I, Sahinalp SC (2014) DeeZ: reference-based compression by local assembly. Nat Methods 11:1082–1084. doi:10.1038/nmeth.3133

Hamosh A, Sobreira N, Hoover-Fong J, Sutton VR, Boehm C, Schiettecatte F, Valle D (2013) PhenoDB: a new web-based tool for the collection, storage, and analysis of phenotypic features. Hum Mutat 34:566–571. doi:10.1002/humu.22283

Hsi-Yang Fritz M, Leinonen R, Cochrane G, Birney E (2011) Efficient storage of high throughput DNA sequencing data using reference-based compression. Genome Res 21:734–740. doi:10.1101/gr.114819.110

Ionita-Laza I, McCallum K, Xu B, Buxbaum JD (2016) A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat Genet 48:214–220. doi:10.1038/ng.3477

Jager M, Wang K, Bauer S, Smedley D, Krawitz P, Robinson PN (2014) Jannovar: a java library for exome annotation. Hum Mutat 35:548–555. doi:10.1002/humu.22531

Jian X, Boerwinkle E, Liu X (2014) In silico prediction of splice-altering single nucleotide variants in the human genome. Nucleic Acids Res 42:13534–13544. doi:10.1093/nar/gku1206

Jiang Y, Oldridge DA, Diskin SJ, Zhang NR (2015a) CODEX: a normalization and copy number variation detection method for whole exome sequencing. Nucleic Acids Res. doi:10.1093/nar/gku1363

Jiang Y, Turinsky AL, Brudno M (2015b) The missing indels: an estimate of indel variation in a human genome and analysis of factors that impede detection. Nucleic Acids Res 43:7217–7228. doi:10.1093/nar/gkv677

Khurana E, Fu Y, Chen J, Gerstein M (2013) Interpretation of genomic variants using a unified biological network approach. PLoS Comput Biol 9:e1002886. doi:10.1371/journal.pcbi.1002886

Kingsford C, Patro R (2015) Reference-based compression of short-read sequences using path encoding. Bioinformatics 31:1920–1928. doi:10.1093/bioinformatics/btv071

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J (2014) A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46:310–315. doi:10.1038/ng.2892

Kirkpatrick BE et al (2015) GenomeConnect: matchmaking between patients, clinical laboratories, and researchers to improve genomic knowledge. Hum Mutat 36:974–978. doi:10.1002/humu.22838

Kohler S et al (2014) The human phenotype ontology project: linking molecular biology and disease through phenotype data. Nucleic Acids Res 42:D966–D974. doi:10.1093/nar/gkt1026

Kozanitis C, Saunders C, Kruglyak S, Bafna V, Varghese G (2011) Compressing genomic sequence fragments using SlimGene. J Comput Biol J Comput Mol Cell Biol 18:401–413. doi:10.1089/cmb.2010.0253

Krumm N et al (2012) Copy number variation detection and genotyping from exome sequence data. Genome Res 22:1525–1532. doi:10.1101/gr.138115.112

Kumar P, Henikoff S, Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 4:1073–1081. doi:10.1038/nprot.2009.86

Lambertson KF, Damiani SA, Might M, Shelton R, Terry SF (2015) Participant-driven matchmaking in the genomic era. Hum Mutat 36:965–973. doi:10.1002/humu.22852

Lancaster O et al (2015) Cafe Variome: general-purpose software for making genotype-phenotype data discoverable in restricted or open access contexts. Hum Mutat 36:957–964. doi:10.1002/humu.22841

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25. doi:10.1186/gb-2009-10-3-r25

Layer RM, Kindlon N, Karczewski KJ, Quinlan AR (2016) Efficient genotype compression and analysis of large genetic-variation data sets. Nat Methods 13:63–65. doi:10.1038/nmeth.3654

Lee D, Gorkin DU, Baker M, Strober BJ, Asoni AL, McCallion AS, Beer MA (2015) A method to predict the impact of regulatory variants from DNA sequence. Nat Genet 47:955–961. doi:10.1038/ng.3331

Lek M et al (2015) Analysis of protein-coding genetic variation in 60,706 humans. bioRxiv. doi:10.1101/030338

Lelieveld SH, Spielmann M, Mundlos S, Veltman JA, Gilissen C (2015) Comparison of exome and genome sequencing technologies for the complete capture of protein-coding regions. Hum Mutat 36:815–822. doi:10.1002/humu.22813

Li H (2011) Tabix: fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics 27:718–719. doi:10.1093/bioinformatics/btq671

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754–1760. doi:10.1093/bioinformatics/btp324

Li H et al (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079. doi:10.1093/bioinformatics/btp352

Lindhurst MJ et al (2011) A mosaic activating mutation in AKT1 associated with the Proteus syndrome. N Engl J Med 365:611–619. doi:10.1056/NEJMoa1104017

Liu X, Jian X, Boerwinkle E (2013) dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. Hum Mutat 34:E2393–E2402. doi:10.1002/humu.22376

Liu B et al (2014) Cloud-based bioinformatics workflow platform for large-scale next-generation sequencing analyses. J Biomed Inform 49:119–133. doi:10.1016/j.jbi.2014.01.005

Lohmueller KE et al (2013) Whole-exome sequencing of 2000 Danish individuals and the role of rare coding variants in type 2 diabetes. Am J Hum Genet 93:1072–1086. doi:10.1016/j.ajhg.2013.11.005

MacArthur DG, Tyler-Smith C (2010) Loss-of-function variants in the genomes of healthy humans. Hum Mol Genet 19:R125–R130. doi:10.1093/hmg/ddq365

MacArthur DG et al (2012) A systematic survey of loss-of-function variants in human protein-coding genes. Science 335:823–828. doi:10.1126/science.1215040

MacArthur DG et al (2014) Guidelines for investigating causality of sequence variants in human disease. Nature 508:469–476. doi:10.1038/nature13127

McKenna A et al (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297–1303. doi:10.1101/gr.107524.110

Meienberg J et al (2015) New insights into the performance of human whole-exome capture platforms. Nucleic Acids Res 43:e76. doi:10.1093/nar/gkv216

Miller DT et al (2010) Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. Am J Hum Genet 86:749–764. doi:10.1016/j.ajhg.2010.04.006

Miosge LA et al (2015) Comparison of predicted and actual consequences of missense mutations. Proc Natl Acad Sci USA 112:E5189–E5198. doi:10.1073/pnas.1511585112

NCBI Resource Coordinators (2015) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 43:D6–17. doi:10.1093/nar/gku1130

Neale BM et al (2012) Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature 485:242–245. doi:10.1038/nature11011

Neveling K et al (2013) A post hoc comparison of the utility of sanger sequencing and exome sequencing for the diagnosis of heterogeneous diseases. Hum Mutat 34:1721–1726. doi:10.1002/humu.22450

Ng SB et al (2009) Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461:272–276. doi:10.1038/nature08250

Nguyen T, Shi W, Ruden D (2011) CloudAligner: a fast and full-featured MapReduce based tool for sequence mapping. BMC Res Notes 4:171. doi:10.1186/1756-0500-4-171

Ochoa I, Asnani H, Bharadia D, Chowdhury M, Weissman T, Yona G (2013) QualComp: a new lossy compressor for quality scores based on rate distortion theory. BMC Bioinf 14:187. doi:10.1186/1471-2105-14-187

Oellrich A et al (2015) The digital revolution in phenotyping. Brief Bioinform. doi:10.1093/bib/bbv083

Okonechnikov K, Conesa A, Garcia-Alcalde F (2016) Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics 32:292–294. doi:10.1093/bioinformatics/btv566

O'Rawe J et al (2013) Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Med 5:28. doi:10.1186/gm432

O'Roak BJ et al (2011) Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. Nat Genet 43:585–589. doi:10.1038/ng.835

O'Roak BJ et al (2012) Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. Science 338:1619–1622. doi:10.1126/science.1227764

Pabinger S et al (2014) A survey of tools for variant analysis of next-generation genome sequencing data. Brief Bioinf 15:256–278. doi:10.1093/bib/bbs086

Parla JS, Iossifov I, Grabill I, Spector MS, Kramer M, McCombie WR (2011) A comparative analysis of exome capture. Genome Biol 12:R97. doi:10.1186/gb-2011-12-9-r97

Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB (2013) Genic intolerance to functional variation and the interpretation of personal genomes. PLoS Genet 9:e1003709. doi:10.1371/journal.pgen.1003709

Philippakis AA et al (2015) The Matchmaker Exchange: a platform for rare disease gene discovery. Hum Mutat 36:915–921. doi:10.1002/humu.22858

Poduri A, Evrony GD, Cai X, Walsh CA (2013) Somatic mutation, genomic variation, and neurological disease. Science 341:1237758. doi:10.1126/science.1237758

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A (2010) Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res 20:110–121. doi:10.1101/gr.097857.109

Poultney CS et al (2013) Identification of small exonic CNV from whole-exome sequence data and application to autism spectrum disorder. Am J Hum Genet 93:607–619. doi:10.1016/j.ajhg.2013.09.001

Quang D, Chen Y, Xie X (2015) DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics 31:761–763. doi:10.1093/bioinformatics/btu703

Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842. doi:10.1093/bioinformatics/btq033

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

Rehm HL et al (2013) ACMG clinical laboratory standards for next-generation sequencing. Genet Med 15:733–747. doi:10.1038/gim.2013.92

Reid JG et al (2014) Launching genomics into the cloud: deployment of Mercury, a next generation sequence analysis pipeline. BMC Bioinf 15:30. doi:10.1186/1471-2105-15-30

Robinson PN et al (2014) Improved exome prioritization of disease genes through cross-species phenotype comparison. Genome Res 24:340–348. doi:10.1101/gr.160325.113

Rosenberg AB, Patwardhan RP, Shendure J, Seelig G (2015) Learning the sequence determinants of alternative splicing from millions of random sequences. Cell 163:698–711. doi:10.1016/j.cell.2015.09.054

Rusk N (2016) Deep learning. Nat Meth 13:35. doi:10.1038/nmeth.3707

Samocha KE et al (2014) A framework for the interpretation of de novo mutation in human disease. Nat Genet 46:944–950. doi:10.1038/ng.3050

Samuels DC, Han L, Li J, Quanghu S, Clark TA, Shyr Y, Guo Y (2013) Finding the lost treasures in exome sequencing data. Trends Genet 29:593–599. doi:10.1016/j.tig.2013.07.006

Sato Y et al (2014) Recurrent somatic mutations underlie corticotropin-independent Cushing's syndrome. Science 344:917–920. doi:10.1126/science.1252328

Schatz MC (2009) CloudBurst: highly sensitive read mapping with MapReduce. Bioinformatics 25:1363–1369. doi:10.1093/bioinformatics/btp236

Schatz MC, Langmead B, Salzberg SL (2010) Cloud computing and the DNA data race. Nat Biotechnol 28:691–693. doi:10.1038/nbt0710-691

Shanahan HP, Owen AM, Harrison AP (2014) Bioinformatics on the cloud computing platform Azure. PLoS One 9:e102642. doi:10.1371/journal.pone.0102642

Sobreira N, Schiettecatte F, Valle D, Hamosh A (2015) GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. Hum Mutat 36:928–930. doi:10.1002/humu.22844

Stein LD (2010) The case for cloud computing in genome informatics. Genome Biol 11:207. doi:10.1186/gb-2010-11-5-207

Stein LD, Knoppers BM, Campbell P, Getz G, Korbel JO (2015) Data analysis: create a cloud commons. Nature 523:149–151. doi:10.1038/523149a

Stephens ZD et al (2015) Big Data: astronomical or genomical? PLoS Biol 13:e1002195. doi:10.1371/journal.pbio.1002195

Strom SP, Lee H, Das K, Vilain E, Nelson SF, Grody WW, Deignan JL (2014) Assessing the necessity of confirmatory testing for exome-sequencing results in a clinical molecular diagnostic laboratory. Genet Med 16:510–515. doi:10.1038/gim.2013.183

Tennessen JA et al (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337:64–69. doi:10.1126/science.1219240

Teo SM, Pawitan Y, Ku CS, Chia KS, Salim A (2012) Statistical challenges associated with detecting copy number variations with next-generation sequencing. Bioinformatics 28:2711–2718. doi:10.1093/bioinformatics/bts535

The Deciphering Developmental Disorders Study (2015) Large-scale discovery of novel genetic causes of developmental disorders. Nature 519:223–228. doi:10.1038/nature14135

The ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74. doi:10.1038/nature11247

Thompson R et al (2014) RD-Connect: an integrated platform connecting databases, registries, biobanks and clinical bioinformatics for rare disease research. J Gen Intern Med 29(Suppl 3):S780–S787. doi:10.1007/s11606-014-2908-8

Thorvaldsdottir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinf 14:178–192. doi:10.1093/bib/bbs017

Tryka KA et al (2014) NCBI's database of genotypes and phenotypes: dbGaP. Nucleic Acids Res 42:D975–D979. doi:10.1093/nar/gkt1211

van der Velde KJ et al (2015) Evaluation of CADD scores in curated mismatch repair gene variants yields a model for clinical validation and prioritization. Hum Mutat 36:712–719. doi:10.1002/humu.22798

Walter K et al (2015) The UK10K project identifies rare variants in health and disease. Nature 526:82–90. doi:10.1038/nature14962

Weiss MM et al (2013) Best practice guidelines for the use of next-generation sequencing applications in genome diagnostics: a national collaborative study of Dutch genome diagnostic laboratories. Hum Mutat 34:1313–1321. doi:10.1002/humu.22368

Xiong HY et al (2015) RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. Science 347:1254806. doi:10.1126/science.1254806

Yang H, Wang K (2015) Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. Nat Protoc 10:1556–1566. doi:10.1038/nprot.2015.105

Yang Y et al (2013) Clinical whole-exome sequencing for the diagnosis of mendelian disorders. N Engl J Med 369:1502–1511. doi:10.1056/NEJMoa1306555

Yu YW, Yorukoglu D, Peng J, Berger B (2015) Quality score compression improves genotyping accuracy. Nat Biotechnol 33:240–243. doi:10.1038/nbt.3170

Zhang F, Gu W, Hurles ME, Lupski JR (2009) Copy number variation in human health, disease, and evolution. Annu Rev Genomics Hum Genet 10:451–481. doi:10.1146/annurev.genom.9.081307.164217

Zhou J, Troyanskaya OG (2015) Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods 12:931–934. doi:10.1038/nmeth.3547

Zighelboim I, Mutch DG, Knapp A, Ding L, Xie M, Cohn DE, Goodfellow PJ (2014) High frequency strand slippage mutations in CTCF in MSI-positive endometrial cancers. Hum Mutat 35:63–65. doi:10.1002/humu.22463

Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, Salit M (2014) Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat Biotechnol 32:246–251. doi:10.1038/nbt.2835