Strategies for Exome and Genome Sequence Data Analysis in Disease Gene Discovery Projects

Peter N. Robinson[1,2,3,†], Peter Krawitz[1,2,3] and Stefan Mundlos[1,2,3]

1. *Institute for Medical Genetics and Human Genetics, Charité-Universitätsmedizin Berlin*
2. *Berlin-Brandenburg Center for Regenerative Therapies (BCRT), Charité-Universitätsmedizin Berlin*
3. *Max-Planck-Institut für Molekulare Genetik, Berlin*

† *Correspondence to Dr. med. Peter N. Robinson, Institut für Medizinische Genetik und Humangenetik, Charité-Universitätsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin, Germany; peter.robinson@charite.de*

Abstract

In whole-exome sequencing, target capture methods are used to enrich the sequences of the coding regions of genes from fragmented total genomic DNA, followed by massively parallel, "next-generation" sequencing of the captured fragments. Since its introduction in 2009, whole-exome sequencing has been successfully used in several disease-gene discovery projects, but the analysis of whole-exome sequence data can be challenging. In this overview, we present a summary of the main computational strategies that have been applied to identify novel disease genes in whole-exome data, including intersect filters, the search for *de novo* mutations, and the application of linkage mapping or inference of identity by descent in family studies.

Keywords: Next Generation Sequencing, Exome, Genome, Disease Gene, Bioinformatics
Conflict of interest: None

The identification of Mendelian disease genes has long been a major focus of human genetics. Until recently, most efforts at disease-gene identification involved positional cloning, that is, linkage analysis to identify a genomic interval usually spanning approximately 0.5–10 cM and containing up to about 300 genes. Sequencing large numbers of genes was time-consuming and expensive, and international efforts based primarily on positional cloning strategies had identified less than 2,000 disease genes by the year 2009 (corresponding to something less than 4,000 diseases, since some genes are associated with multiple diseases). A large number of Mendelian diseases remain for which no disease gene has been identified yet, and presumably there are many more unnamed monogenic diseases that will be found in coming years.

The initial demonstration by Sarah Ng and colleagues that whole exome sequencing can be used to identify disease genes in 2009 (1) can probably be regarded as the beginning of a revolution in human genetics, and many reports of novel disease genes discovered using whole-exome sequencing have been published in the subsequent two years. However, whole-exome sequencing (WES) is not a panacea, and researchers need to carefully consider how to design WES experiments for disease-gene discovery to avoid frustration. This article will review the major strategies that have been applied to discover novel disease genes with whole-exome sequencing and discuss some of the pitfalls of the methodology.

Whole-Exome Sequencing (WES)

To date, the great majority of mutations identified in human hereditary diseases have been located in the coding sequences of genes. It seems possible that mutations in non-coding sequences are more common than is currently appreciated and have been rarely detected because of various technical and experimental biases.

This is an Accepted Article that has been peer-reviewed and approved for publication in the *Clinical Genetics*, but has yet to undergo copy-editing and proof correction. Please cite this article as an "Accepted Article"; doi: 10.1111/j.1399-0004.2011.01713.x

However, the fact that the great majority of disease-causing mutations characterized to date have been located in or around exons strongly suggested that it would be useful in disease-gene discovery projects to concentrate sequencing efforts on the approximately 1% of the human genome that codes for protein sequences in order to avoid the additional cost and complexity of whole-genome sequencing (WGS). However, as the costs of WGS continue to fall and our ability to interpret variation in non-coding sequences improves, it seems likely that WGS will replace WES in many settings.

Current methods for enriching exonic sequences all work in principle in the same fashion. Oligonucleotide probes are constructed to hybridize to ("capture") the target sequences from fragmented total genomic DNA. Common linkers or adaptors are used as primers to amplify the target sequences in a single PCR reaction, and the unwanted sequences are discarded. A number of companies are offering target capture methods for whole-exome sequencing (2). The methods typically aim to capture all exonic and flanking sequences and may also include probes to target microRNA and other sequences of interest. Several reviews on the technical aspects of target capture methods have appeared recently (3-5). Current commercial offerings are compatible with the three major next-generation sequencing platforms by Illumina, Roche, and Applied Biosystems. The kits tend to comprise exons from the consensus coding sequence (CCDS) project (6), which currently comprises 176,266 exons from 18,409 genes, as well as additional sequences.

WES Strategies

One of the main challenges for disease-gene discovery by WES lies in the sheer number of variants found in individual exomes. It has been reported that each genome carries 165 homozygous protein-truncating or stop loss variants in genes representing a diverse set of pathways (7). This means that the mere finding of a sequence variant that appears to be a pathogenic mutation cannot be taken as proof that the change is causally related to the disease being investigated, and integrative computational analysis is required that takes into account phenotypes, prioritizes sequence variants, and makes use of information from multiple databases, in order to use the data from WES in a medical context. Additionally, the choice of which samples to sequence and what type of bioinformatic analysis to apply will depend on the clinical situation. The following sections intend to provide an intuitive introduction to the relevant issues and pointers to the literature.

Intersection Filtering

Although the numbers have varied between different publications, which presumably reflects differences in the technologies and analysis strategies, typically, an individual exome is found to have 20–30,000 variants as compared with the genomic reference sequence. Up to roughly 10,000 of these variants are predicted to lead to nonsynonymous amino acid substitutions (missense mutations), alterations of conserved splice site residues, or represent small insertions or deletions ("NS/SS/I"). Depending on the ethnic background of the proband and other factors, up to about 90% of these variants can be found in databases of common variants such as dbSNP (8), the 1000 Genomes project (9), and in-house exome databases. Based on the assumption that variants that are common in the population are not likely to be the cause of rare Mendelian diseases, such variants are typically filtered out before further analysis. Similarly, variants that are computationally predicted to be benign are typically removed from further analysis based on the results of algorithms that estimate the pathogenicity of missense and other variants (10-12). It should be noted that computational algorithms have relatively high rates of false-positive and false-negative predictions (13,14). Although it is difficult to give an exact numerical value, it is likely that the false-negative and false-positive rates are at least 20% for WES data.

This kind of filtering has been used successfully in several projects in which multiple individuals with a given disease were sequenced. Following removal of common variants and those not predicted to be pathogenic, only those genes are considered that show rare and potentially pathogenic sequence variants in all (or most) sequenced individuals (1;15-17). For autosomal dominant disorders, each candidate gene must show at least one such change per individual, and for autosomal recessive disorders, candidate genes must have either homozygous or compound heterozygous mutations.

The assumption is that each exome or genome contains numerous sequence variants unrelated to the disease being studied and that are not removed by the filtering steps described above (thus, these variants can be regarded as false-positive calls). Under the assumption that these variants are distributed at random in the population, if we examine the intersection between a sufficient number of multiple unrelated individuals, only the disease gene itself will show mutations in all individuals. Imagine for the sake of argument that 5% of all genes show rare, potentially pathogenic sequence variants in all individuals. If we sequence a single individual, then 1000 genes will remain as candidates after we filter as above. If we sequence a second individual and examine only those genes with variants in both individuals, then 5% of 1000 or 50 candidates will remain. After we sequence a third individual, less than one gene is predicted to have a variant in all three individuals just by chance, and only the true disease gene will remain.

Of course, this strategy is highly susceptible to false-negative and false-positive results if applied naively. For instance, in a study on 10 individuals with Kabuki syndrome, the only gene that was found to have at least one NS/SS/I in all ten sequenced individuals was the *MUC16* gene, which codes for a protein with 22,152 amino acids that provides a protective, lubricating barrier against particles and infectious agents at mucosal surfaces. It soon became apparent that this was a false positive result that might be related to the extremely large size of the coding sequence and the resultant higher chance of an unrelated sequence variant being present. On the other hand, it is possible that a mutation is located in a poorly covered exon and thus escapes detection (typically, a reasonable coverage can be achieved for up to about 90% of the sequenced exome using current targeting technologies; thus, if several of the mutations amongst the sequenced individuals are located in poorly covered exons, the candidate gene would falsely be removed from further consideration). Alternatively, a mutation might not be a typical NS/SS/I variant and thus might be have been mistakenly removed. For instance, a mutation such as c.6354C>T, a silent mutation in exon 51 of the fibrillin-1 gene that induces exon skipping (18), would not be identified by current filtering strategies. Additionally, although most point mutations in human hereditary disease identified to date have been located in or near exons, point mutations in distant enhancers and other regulatory elements have been associated with hereditary diseases (e.g., ref. 19), and such mutations would for the most part not be detectable using current enrichment strategies. Finally, genetically heterogeneous disorders can be missed by this approach, because different genes could be involved in individual patients of the study group.

De novo mutations

Many consultations in medical genetics clinics deal with isolated cases of mental retardation, multiple congenital anomalies, or other diseases. Unless an etiological diagnosis can be made, it is not formally possible to know whether the manifestations in the patient are related to an autosomal recessive disorder, an oligogenic or otherwise multifactorial disease, environmental factors, or to a *de novo* (spontaneous) mutation. Recent results suggest that the role of *de novo* mutations in such situations may have been underappreciated. The per generation mutation rate in humans has been estimated at between 7.6 x 10^9 and 2.2 x 10^8, or roughly one in a hundred million positions in the haploid genome, which corresponds to 0.86 *de novo* amino-acid altering mutations per newborn (20.21). Therefore, for heterogeneous diseases in which mutations in one of any of a number of genes can cause the disease, it seemed reasonable to hypothesize that *de novo* mutations might be more common than previously believed and to use an analysis strategy that in which case-parent trios are sequenced in order to identify potentially pathogenic, *de novo* changes in the exome sequences of the affected children (22,23).

The pioneering work of Vissers and colleagues on this topic describes how exome sequencing was performed on ten trios (affected child and healthy parents). After ruling out copy number variations (CNVs) by array CGH analysis, exome sequences were obtained and subjected to a bioinformatic pipeline to exclude common variants and those predicted not to be pathogenic. Then, variants were sought that were present only in an affected child but not in the parents. This led to the identification of convincing candidate mutations in seven of the ten trios (22).

Family-Based Filtering Strategies: Homozygosity Mapping and Linkage Approaches

Pierce and coworkers examined a non-consanguineous family in which two sisters had Perrault syndrome, a recessive disorder characterized by ovarian dysgenesis in females, sensorineural deafness, and neurological manifestations. WES of only one of the sisters revealed exactly one gene with two rare variants both predicted to be pathogenic: *HSD17B4*, which encodes 17β-hydroxysteroid dehydrogenase type 4 (24). However, the experience of most labs involved in WES projects suggests that it is extremely difficult to narrow down the search to exactly one gene based on only a single exome sequence. For this reason, classical approaches such as homozygosity mapping (25) and linkage analysis (26) have been used to exclude irrelevant parts of the exome or genome prior to the application of other computational filters.

For instance, Bilgüvar and colleagues examined a small consanguineous kindred with two sibs having microcephaly and other brain malformations. Whole-genome genotyping was used to identify shared homozygous segments that together made up 80 cM. The analysis of WES data concentrated on these regions, and a novel frameshift mutation in *WDR62* was identified. To confirm *WDR62* as the disease gene, Sanger sequencing was used to identify *WDR62* mutations in other kindreds (27). Similar approaches using homozygosity mapping or linkage analysis to narrow down the candidate regions have been used successfully in a number of WES and related studies to identify or confirm novel disease genes (28-35).

IBD Inference from WES/WGS Data

Roach and coworkers performed an analysis of genetic inheritance in a family quartet by WGS, using a Hidden Markov Model (HMM) to model the Mendelian inheritance states at each reference position. There are four possible states of inheritance, depending on whether the two children shared alleles from both parents, from only the mother or only the father, or shared none. Both children in this family had two recessive disorders, Miller syndrome and primary ciliary dyskinesia, as had previously been characterized by exome sequencing (16). The inheritance states in the family quartet as inferred by the HMM were observed in large contiguous blocks, allowed the location of the candidate genes for both of these Mendelian disorders to be narrowed down to only four (21).

The authors of this review developed an HMM-based algorithm to infer chromosomal regions that are identical-by-descent (IBD) based only on the (potentially noisy) exome sequences of the affected siblings. In consanguineous families, affected individuals share two IBD haplotypes inherited from a single common ancestor (homozygosity by descent, HBD). The disease gene is located somewhere within the HBD haplotype block, which is the basis of homozygosity mapping (25). In the general case in which the parents are not consanguineous, each affected sibling inherits the same haplotype from each parent. Such chromosomal regions are referred to as IBD=2 (Figure 1)

Figure 1: In autosomal recessive disorders, the disease gene must bu located in a chromosomal region in which the paternal and maternal haplotypes are both identical by descent (IBD=2).

Our algorithm uses a non-homogeneous hidden Markov model that employs local recombination rates to identify IBD=2 chromosomal regions in children of consanguineous or non-consanguineous parents solely based on genotype data of siblings derived from high-throughput sequencing platforms. Inference of IBD=2 regions can be used to identify the chromosomal regions that are compatible with the inheritance patterns of a recessive monogenic disorder and can be combined with previous methods for filtering out common variants and for predicting potentially pathogenic sequence changes as described above. This approach was first successfully used in identifying *PIGV* as the disease gene in Hyperphosphatasia with Mental Retardation syndrome (36,37).

Conclusions

It has been less than two years since the first publication by Sarah Ng and coworkers on the application of WES for disease-gene identification (1), but it already seems clear that the field of human genetics has entered a new era, in which we can hope to quickly elucidate the molecular basis of most remaining Mendelian disorders and to radically improve our ability to perform timely and accurate diagnostics for persons with rare diseases. The trend towards cheaper and higher-throughput DNA sequencing has been proceeding substantially more rapidly than predicted even by “Moore’s law“ for computing hardware, according to which the number of transistors on a chip roughly doubles every two years. The primary challenge in diagnostics in human genetics is likely to shift from the mere identification of sequence variants to the interpretation of the variants, and bioinformatics will play a key role at all levels of data analysis and interpretation. Currently, labs involved in many WES projects, including our own, report success rates in identifying novel disease genes of at most about 50%. There are many potential reasons for failure in WES projects, some of which have been mentioned in this short overview. As we move forward, a number of things will be needed to achieve the full promise of WES for disease-gene discovery and later on for routine diagnostics (38).

Improvements of the software used for the analysis of WES/WGS data are sorely needed, and continued algorithmic development will be required as target capture technologies continue to evolve. As soon as whole genome sequencing becomes economically feasible in the way that WES is already, algorithms for reliably capturing structural variation (39) and for interpreting variants in non-coding conserved sequences will become essential.

Standard protocols and ontologies will be required for interoperability of databases; a major problem in studying rare diseases, is that one can never be entirely certain that a given gene is in fact the sought after disease gene until a second unrelated individual or family is described with a mutation in the same gene and a comparable phenotype. Standard ontologies for describing the human phenotype (40-42) combined with other standards for reporting mutations, classifying diseases, and storing the genotypes of WES or WGS data will be an essential component to allow communication and interoperability. At present, there is no comprehensive database of human mutations and phenotypes that can be used for the interpretation of WES/WGS data, although efforts are underway in the community. Such a database could be used to connect groups at different locations each of which has identified single individuals or families with mutations in a novel candidate for a rare disease and thus help to accelerate efforts to identify the remaining disease genes in the human genome.

Acknowledgement

This work was supported by grants from Bundesministerium für Bildung und Forschung (BMBF, project number 0313911).

References

1. Ng SB, Turner EH, Robertson PD, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature 2009: 461: 272–276.

2. Garber K. Fixing the front end. Nat Biotechnol 2008 : 26: 1101–1104.

3. Turner EH, Ng SB, Nickerson DA, et al. Methods for genomic partitioning. Annu Rev Genomics Hum Genet 2009 : 10: 263–284.

4. Summerer D. Enabling technologies of genomic-scale sequence enrichment for targeted high-throughput sequencing. Genomics 2009 : 94: 363–368.

5. Mamanova L, Coffey AJ, Scott CE, et al. Target-enrichment strategies for next-generation sequencing. Nat Methods 2010 : 7: 111–118.

6. Pruitt KD, Harrow J, Harte RA, et al. The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. Genome Res 2009 : 19: 1316–1323.

7. Pelak K, Shianna KV, Ge D, et al. The characterization of twenty sequenced human genomes. PLoS Genet 2010 : 6: e1001111.

8. Sayers EW, Barrett T, Benson DA, et al. Database resources of the National Center for Biotechnology

Information. Nucleic Acids Res 2011 : 39 (Database issue): D38–D51.

9. 1000 Genomes Project Consortium, Durbin RM, Abecasis GR, et al. A map of human genome variation from population-scale sequencing. Nature 2010 : 467: 1061–1073.

10. Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. Nucleic Acids Res 2002 : 30: 3894–3900.

11. Binkley J, Karra K, Kirby A, et al. ProPhylER: a curated online resource for protein function and structure based on evolutionary constraint analyses. Genome Res 2010 : 20: 142–154.

12. Schwarz JM, Rödelsperger C, Schuelke M, et al. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods 2010 : 7: 575–576.

13. Wei Q, Wang L, Wang Q, et al. Testing computational prediction of missense mutation phenotypes: functional characterization of 204 mutations of human cystathionine beta synthase. Proteins 2010 : 78: 2058–2074.

14. Mathe E, Olivier M, Kato S, et al. Computational approaches for predicting the biological effect of p53 missense mutations: a comparison of three sequence analysis based methods. Nucleic Acids Res 2006 : 34: 1317–1325.

15. Ng SB, Bigham AW, Buckingham KJ, et al. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat Genet 2010 : 42: 790–793.

16. Ng SB, Buckingham KJ, Lee C, et al. Exome sequencing identifies the cause of a mendelian disorder. Nat Genet 2010 : 42: 30–35.

17. Hoischen A, van Bon BWM, Gilissen C, et al. De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. Nat Genet 2010 : 42: 483–485.

18. Liu W, Qian C, Francke U. Silent mutation induces exon skipping of fibrillin-1 gene in Marfan syndrome. Nat Genet 1997 : 16: 328–329.

19. Lettice LA, Hill AE, Devenney PS, et al. Point mutations in a distant sonic hedgehog cis-regulator generate

a variable regulatory output responsible for preaxial polydactyly. Hum Mol Genet 2008: 17: 978–985.

20. Lynch M. Rate, molecular spectrum, and consequences of human mutation. Proc Natl Acad Sci U S A 2010 : 107: 961–968.

21. Roach JC, Glusman G, Smit AFA, et al. Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing. Science 2010 : 328: 636–639.

22. Vissers LELM, de Ligt J, Gilissen C, et al. A de novo paradigm for mental retardation. Nat Genet 2010, : 42: 1109–1112.

23. Robinson PN. Whole-exome sequencing for finding de novo mutations in sporadic mental retardation. Genome Biol 2010 : 11: 144.

24. Pierce SB, Walsh T, Chisholm KM, et al. Mutations in the DBP-deficiency protein HSD17B4 cause ovarian dysgenesis, hearing loss, and ataxia of Perrault Syndrome. Am J Hum Genet 2010 : 87: 282–288.

25. Lander ES, Botstein D. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. Science 1987 : 236: 1567–1570.

26. Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. Science 2008, 322 (5903):881–888.

27. Bilgüvar K, Oztürk AK, Louvi A, et al. Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. Nature 2010 : 467: 207–10.

28. Rehman AU, Morell RJ, Belyantseva IA, et al. Targeted capture and next-generation sequencing identifies C9orf75, encoding taperin, as the mutated gene in nonsyndromic deafness DFNB79. Am J Hum Genet 2010 : 86: 378–388.

29. Nikopoulos K, Gilissen C, Hoischen A, et al. Next-generation sequencing of a 40 Mb linkage interval reveals TSPAN12 mutations in patients with familial exudative vitreoretinopathy. Am J Hum Genet 2010 : 86: 240–247.

30. Volpi L, Roversi G, Colombo EA, et al. Targeted next-generation sequencing appoints c16orf57 as clericuzio-type poikiloderma with neutropenia gene. Am J Hum Genet 2010 : 86: 72–76.

31. Lalonde E, Albrecht S, Ha KCH, et al. Unexpected allelic heterogeneity and spectrum of mutations in

Fowler syndrome revealed by next-generation exome sequencing. Hum Mutat 2010, 31 (8):918–923.

32. Johnston JJ, Teer JK, Cherukuri PF, et al. Massively parallel sequencing of exons on the X chromosome identifies RBM10 as the gene that causes a syndromic form of cleft palate. Am J Hum Genet 2010 : 86: 743–748.

33. Wang JL, Yang X, Xia K, et al. TGM6 identified as a novel causative gene of spinocerebellar ataxias using exome sequencing. Brain 2010 : 133: 3510–3518.

34. Walsh T, Shahin H, Elkan-Miller T, et al. Whole exome sequencing and homozygosity mapping identify mutation in the cell polarity protein GPSM2 as the cause of nonsyndromic hearing loss DFNB82. Am J Hum Genet 2010 : 87: 90–94.

35. Musunuru K, Pirruccello JP, Do R, et al. Exome sequencing, ANGPTL3 mutations, and familial combined hypolipidemia. N Engl J Med 2010 : 363: 2220–2227.

36. Krawitz PM, Schweiger MR, Rödelsperger C, et al. Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome. Nat Genet 2010 : 42: 827–829.

37. Rödelsperger C, Krawitz P, Bauer S, et al. Identity-by-descent filtering of exome sequence data for disease-gene identification in autosomal recessive disorders. Bioinformatics 2011 : 27: 829–836.

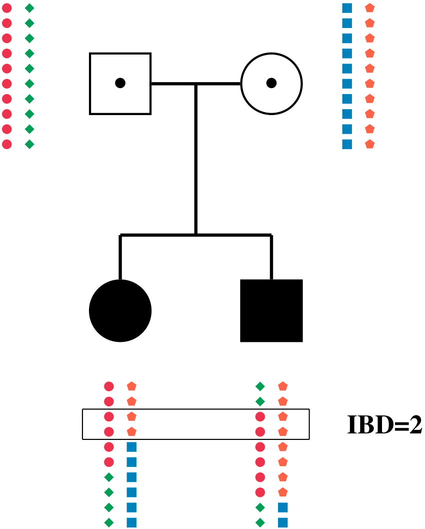
38. Lindblom A, Robinson PN. Bioinformatics for human genetics: promises and challenges. Hum Mutat 2011 : 32: 495–500.

39. Medvedev P, Stanciu M, Brudno M. Computational methods for discovering structural variation with next-generation sequencing. Nat Methods 2009 : 6: S13–S20.

40. Robinson PN, Mundlos S. The Human Phenotype Ontology. Clin Genet 2010 : 77: 525–534.

41. Köhler S, Schulz MH, Krawitz P, et al. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. Am J Hum Genet 2009 : 85: 457–464.

42. Robinson PN, Köhler S, Bauer S, et al. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. Am J Hum Genet 2008 : 83: 610–615.


IBD=2