

Mendelian Disease Gene Identification and Diagnostics using Targeted Next Generation Sequencing

Daniel Trujillano^{1*}, Rami Abou Jamra¹ and Arndt Rolfs^{1,2}

¹Centogene AG, Rostock, Germany

²Albrecht-Kossel-Institute for Neuroregeneration, Medical University Rostock, Rostock, Germany

Email: daniel.trujillano@centogene.com

During the last few months we have observed for the first time since the introduction of the first massive parallel sequencers in 2007, that the cost of sequencing a human genome has not changed significantly (Figure 1) [Wetterstrand, 2015]. These numbers challenge a trend that has been maintained during years beating Moore's law, and indicates that in the medium term we should not expect sequencing to be significantly cheaper until the next technological revolution arrives. This can have profound consequences in the genomics field, which has heavily relied during the last years on the fact that each sequencing run was cheaper than the last. Most of the recent big genomic achievements have been based on brute force experiments, made possible by the rapid technological advances. The present change of cycle would require more ingenious and elegant ideas to keep publishing interesting findings after the genomics boom of the last decade. So, it is time to fully exploit the potential of the current sequencing technologies, which are likely to be static for a while. Moreover, it is also time to tone down the rhetoric around the advent of \$1000 human genomes and to start working seriously on the clinical translation of our research based on the technology that we currently have, not what we expect to have tomorrow [Hall, 2013].

Current next generation sequencing technologies have the potential to address the mismatch between promises and achievements in medical genetics, still present more than ten years since the human genome project was drafted. Shortly before the completion of the first human genome, Francis Collins, one of the leaders of the Human Genome Project predicted that by 2010 the genetic causes of most Mendelian diseases would have

been unveiled and therapies would be available for most of them, that disease gene associations for most of the common disorders would have been established, and that personalized preventive medicine would be a reality. Although some of his claims have come true, most of the post-genomic era promises are yet to be accomplished [Collins, 2010].

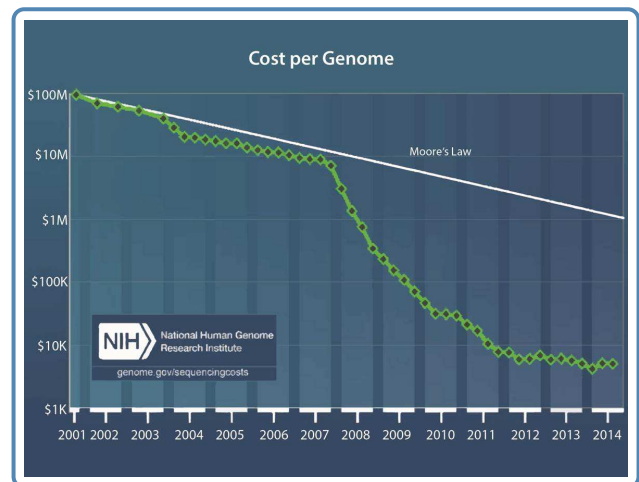


Figure 1 Evolution of the cost of sequencing a human genome. From [Wetterstrand, 2015].

Genomics as an enhanced approach to health-care has the potential to transform the quality of life worldwide, allowing the widespread implementation of more tailored medical care based on individual risk. It seems quite likely that whole human genome sequencing (and eventually, proteome, metabolome, microbiome, etc.) would be

a routine component of everyone's health record available to both patients and physicians for predictive and preventive healthcare purposes. This is poised to have a transforming effect in clinical practice, including diagnosis and decision-making for appropriate therapeutic procedures.

Identification of Mendelian disease genes by exome sequencing

During the last decades, positional cloning strategies have led to the discovery of several new insights in human genetics. However, when this approach is used for rare Mendelian disorders, the results often are not conclusive due to the lack of complete pedigrees, the unavailability of large family collections and marked locus heterogeneity. Thus, in small and non-consanguineous families, neither linkage analysis nor homozygosity mapping are likely to succeed in identifying the responsible gene of a given rare Mendelian disease. As a consequence of these limitations, the underlying genetic cause of more than three thousand disorders of Mendelian inheritance still remain to be determined (<http://www.ncbi.nlm.nih.gov/Omim/mimstats.html>).

In that respect, advances in next generation sequencing technologies, especially exome sequencing, represent an important milestone in genomics, providing an effective alternative for the discovery of candidate genes and mutations that underlie Mendelian disorders that have been resistant to conventional approaches, thanks to an unprecedented ability to identify rare variants. Next generation sequencing technologies have been a much awaited step forward from linkage mapping for Mendelian disease gene discovery, since it has enabled mapping of genes for monogenic traits in families with small pedigrees and even in as few samples as two unrelated individuals [Lalonde et al., 2010].

- **Next generation sequencing bioinformatics:**

Due to the sheer magnitude of the genomic information produced by the current next generation sequencing equipment (several hundred Gigabases can now be generated in just one sequencing run), the experimental bottleneck has shifted from data acquisition towards its correct storage and processing. Efficient methods to align millions of short-read sequences to the human genome (matching the short reads to a preexisting

reference genome) and the calling of variants (determination of the best guess for the genotype, or other sequence feature, at each aligned position) have been developed, allowing to access most of the reference genome and to align *de novo* sequences that are missing in the reference genome sequence. Since DNA sequence variants may vary from single nucleotides up to several kb, specific algorithms have been developed for single base substitutions, insertions/deletions and structural variants.

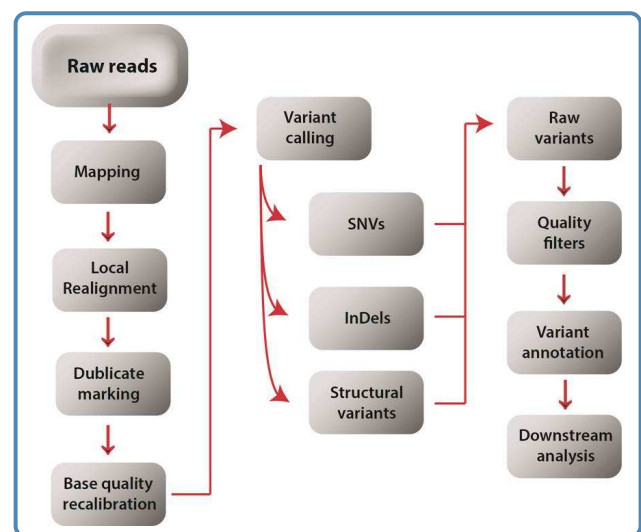


Figure 2 Overview of the bioinformatics analyses for NGS data. From the mapping of the raw sequencing reads to the annotation of the detected variants.

Once the variant calling process of a given sample or project is finished, the next task is the annotation of each detected sequence variant (Figure 2). During this process information regarding the alignment of the variant to a specific base position in a gene, the *in silico* assessment of the variant's potential to disrupt gene function ("pathogenicity") [Kumar et al., 2009; Schwarz et al., 2010], and the presence of the variant in databases such as dbSNP, 1000 Genome project, Exome Variant Server etc are gathered and recorded. Several annotation tools are available, such as the Genome Analysis Toolkit (GATK) [McKenna et al., 2010], SeattleSeq ([www://gvs.gs.washington.edu/](http://gvs.gs.washington.edu/)), or ANNOVAR [Wang et al., 2010], among many other.

- **Disease gene/variant filtering strategies:** Next generation sequencing technologies produce sheer numbers of genotype calls on the order of 10^4 per exome, 10^5 for the combined exomes of a small family, and 10^6 per genome. Thus, after data acquisition and variant calling the main challenge in the downstream analysis of next generation sequencing data is to “winnow” the list of variants to be able to differentiate known and potential novel disease-causing mutations (the “wheat”) from both technical artifacts and benign genetic variation (the “chaff”).

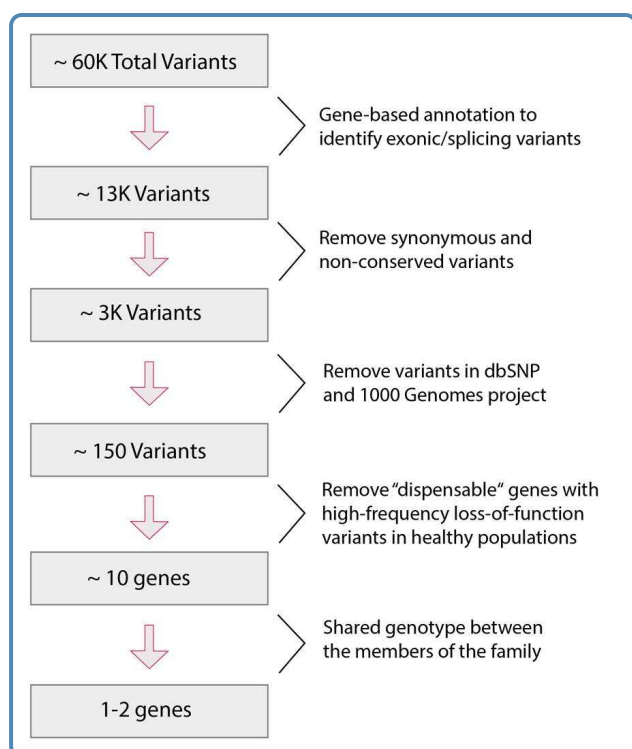


Figure 3 Variant prioritization process for Mendelian disorders.

Depending on the capture design and the depth of coverage, an average exome contains between five to ten thousand variant calls representing either non-synonymous substitutions in protein coding sequences, small InDels, or alterations of the canonical splice-site dinucleotides (NS/SS/I), being between 100 and 200 homozygous protein truncating or stop loss variants. Thus, the mere identification of an apparently causative variant cannot be taken as a proof that it is relevant to the disease being investigated, and additional variant

filtering and functional analyses are required to assign causativeness.

When exome sequencing is applied to Mendelian disorders, the filtering strategy is designed to highlight rare or *de novo*, high penetrance protein-modifying mutations responsible for a large phenotypic effect, as well as all variants previously associated with the disease. Thus, the downstream analyses are focused on the identification of very rare or novel loss-of-function mutations that introduce truncations to the encoded protein, i.e. nonsense and non-synonymous variants, splice acceptor and donor site mutations and coding InDels, anticipating that synonymous variants are far less likely to be pathogenic. This filtering strategy, which has been successfully applied in several studies, substantially reducing the list of candidate variants, making feasible their individual confirmation prior to expression and functional testing (Figure 3).

Another parameter that has to be taken in to account during variant/gene prioritization is the inheritance pattern, since for autosomal dominant disorders each gene must show at least one potentially causative variant per individual, whereas in autosomal recessive disorders, candidate genes must have either homozygous or compound heterozygous mutations (Figure 4) [Robinson et al., 2011].

All in all, the variant filtering strategy must be flexible enough to allow adjustment of all analytic parameters. But even more importantly, those performing the analysis must understand the rationale, procedures, and assumptions inherent in each step.

- **Exome sequencing example in a diagnostics setup:** Exome sequencing is a powerful tool and is often the only possibility to clarify the cause(s) of disorders. As an example, we recently performed exome sequencing in an 18 years old male index patient and his consanguineous parents. The index patient presented with short stature, bilateral nystagmus, cerebellar ataxia, and intellectual disability (Figure 5). Bioinformatic analysis and medical evaluation revealed a candidate mutation in the gene *KIF1C*. The variant is a missense, is rare in public databases and is predicted to be pathogenic by several pathogenicity prediction programs. Mutations in *KIF1C* lead to the autosomal recessive spastic ataxia 2 with horizontal nystagmus, distal muscle atrophy, cerebellar gait ataxia, tremor, spasticity of the lower limbs, cerebellar atrophy (in some pa-

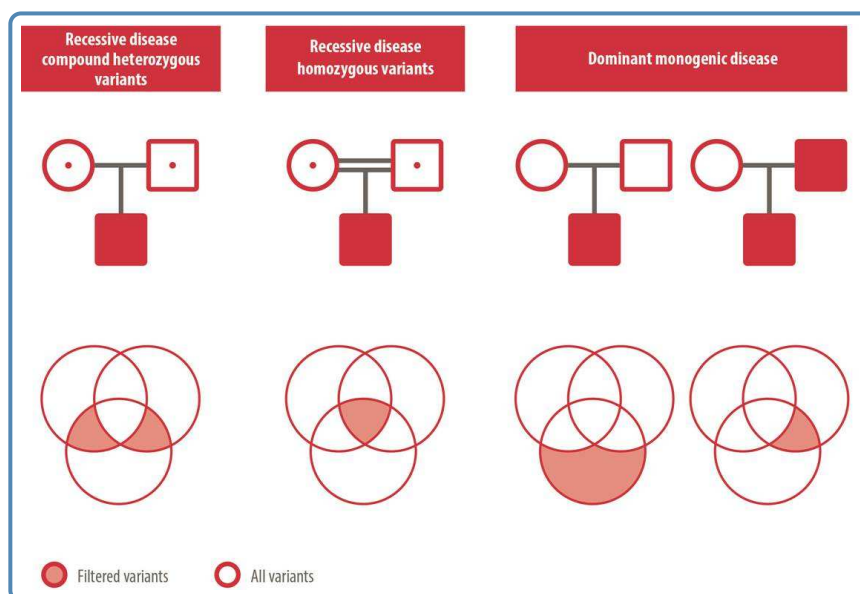


Figure 4 Strategies for Mendelian disease gene identification.

tients) and it has the onset in teenage years. Given the results of the exome sequencing, we concluded that the phenotypic spectrum of the patient can be clarified to a large extent via this variant; it clarifies the nystagmus and the cerebellar ataxia, as well as the intellectual disability. The etiology of short stature is still not clarified. If no exome sequencing was performed, the physician would follow the medical rule of finding one cause of a phenotypic spectrum, and would thus go for other differential diagnoses that include short stature. This would be, however, misleading, and only through analysis of the whole exome we were able to find that a symptom of the patient is not correlated to the rest of the phenotype. It may be possible that the phenotypic spectrum of mutations in *KIF1C* may get extended to short stature too.

- **Exome sequencing advantages and disadvantages:** Exome sequencing represents the most cost-effective alternative to whole genome sequencing for the discovery of highly penetrant rare variants because it involves a drastic reduction in the sequencing required. In fact, as opposed to whole genome sequencing, exome sequencing requires about 20-fold less (~5%) sequencing to achieve the same depth of coverage, which is translated into considerably less raw sequence and lower costs. Despite the inherent costs of genomic capture in addition to sequencing, according to

the list prices, the all-in cost of exome sequencing is roughly 10- to 20-fold less than for the whole genome. Also, exome sequencing requires less complicated analyses than whole genome sequencing, and the number of variants detected is up to two orders of magnitude lower as a consequence of only retrieving variants affecting the coding regions of the genome. This reduces data fatigue and simplifies the analyses for the identification of disease-causative variants.

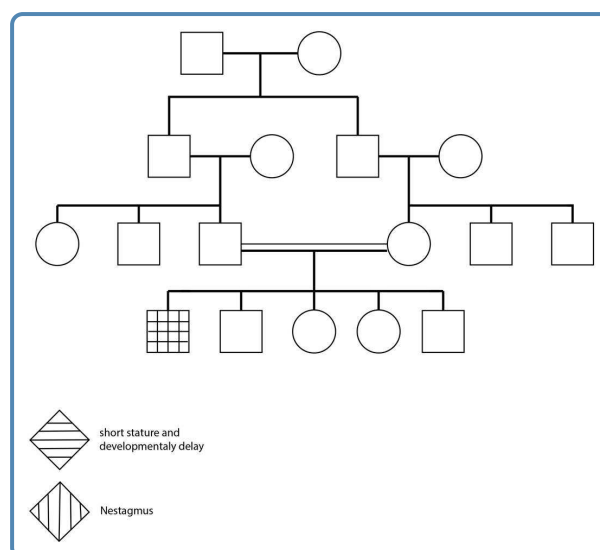


Figure 5 Pedigree of the studied family.

In addition to the technical limitations inherent to genomic enrichment, such as selection bias and uneven capture efficiency, the main limitation of the targeted resequencing approach is the impossibility to efficiently capture and sequence the repetitive and low-complexity, and GC-rich genomic sequences that are refractory to enrichment. However, the constant optimization of the capture and next generation sequencing chemistries will gradually close the capture gaps (mainly due to uniqueness constraints, homopolymer runs, ambiguous bases or other factors that are known to cause issues in either oligonucleotide synthesis or hybridization), and reduce enrichment variability between samples and targets.

Exome sequencing has proven its reliability for the identification of genetic variability underlying relatively simple, single-gene disorders. However, the step from rare monogenic and simple Mendelian disorders to more-complex multigenic disorders is going to be a challenging move. Exome sequencing studies done so far have to be considered as a starting point in the effort to apply these technologies to multigenic diseases. The extent of heterogeneity associated with common complex disorders will have to be mitigated with larger sample sizes and more sophisticated weighting of non-synonymous variants by predicted functional impact.

Concluding remarks

Currently, our ability to discover genetic variation in a patient genome is running far ahead of our ability to interpret that variation. The success of next generation sequencing for medical genetics hinges on the accuracy in distinguishing causal from benign alleles, which is the key challenge for interpreting DNA sequence data for diagnostics. Over the last three decades, PCR amplification of target regions followed by Sanger sequencing has been the gold standard for the identification of clinically relevant mutations in the terms of routine diagnostics. It offers great accuracy, at the expense of being laborious and costly, especially when it comes to the analysis of disorders of heterogeneous etiology for which multiple targets/genes might be tested in a stepwise fashion. Such disorders may require extensive screening of several genes, using different molecular approaches for every type of sequence variant being tested.

However, this rather costly, stepwise, and time-consuming technology will be gradually replaced

by next generation sequencing technologies, which offer higher throughput and scalability and, as a corollary, have reduced costs per sequenced nucleotide and shorter turnaround time. Given the current cost of targeted next generation sequencing of small genomic regions, it is inciting to use next generation sequencing approaches to screen these genes for diagnostics purposes.

The transition over the next years of next generation sequencing technologies from basic research to the routine detection of mutations in genetic loci with well documented diagnostic value will take advantage not only of the new benchtop next generation sequencing platforms which can be much more easily incorporated in the daily clinical practice, but also of automated workflows and simplified bioinformatics analyses able to generate medical report-like outputs adapted to clinical laboratories. However, the correct interpretation, storage, and dissemination of the large amount of the datasets generated remain a major challenge on the path of next generation sequencing to medical applications [Pop et al., 2008]. These challenges could be addressed with extensive exchange of data, information and knowledge between medical scientists, sequencing centers, bioinformatics networks and industry. Some genomic centers working in biomedicine have developed collaborative initiatives aiming at bringing everyone together to harmonize genomic medical research, set up standards in medical sequencing and review the current diagnostic standards according to the new insights gained from genomic and phenotypic data integration.

Genomics is making faster progress than any other area of biomedical research. Especially, the advances in the field of next generation sequencing development and applications, make this an exciting time for the study of how genetic variation affects health and disease. The ultimate game changer in clinical genetics will be the routine sequencing of individual genomes, but until this becomes feasible, targeted approaches are the more convenient interim solution. The standardization and further development of the methods described here will provide powerful and cost-effective techniques for the identification of causative variants of heritable disorders caused by known and unknown genes.

References

1. Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Large-Scale Genome Sequencing Program. <https://www.genome.gov/sequencingcosts/> (accessed April 2015)
2. Hall N. After the gold rush. *Genome Biol* 2013; 14, 115.
3. Collins F. Has the revolution arrived? *Nature* 2010; 464, 674-675.
4. Lalonde E, et al. Unexpected allelic heterogeneity and spectrum of mutations in Fowler syndrome revealed by next-generation exome sequencing. *Hum Mutat* 2010; 31, 918-923.
5. Kumar P, et al. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 2009; 4, 1073-1081.
6. Schwarz JM, et al. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* 2010; 7, 575-576.
7. McKenna A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010; 20, 1297-1303.
8. Wang K, et al. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010; 38, e164.
9. Robinson PN, et al. Strategies for exome and genome sequence data analysis in disease-gene discovery projects. *Clin Genet* 2011; 80, 127-132.
10. Pop M, et al. Bioinformatics challenges of new sequencing technology. *Trends Genet* 2008; 24, 142-149.