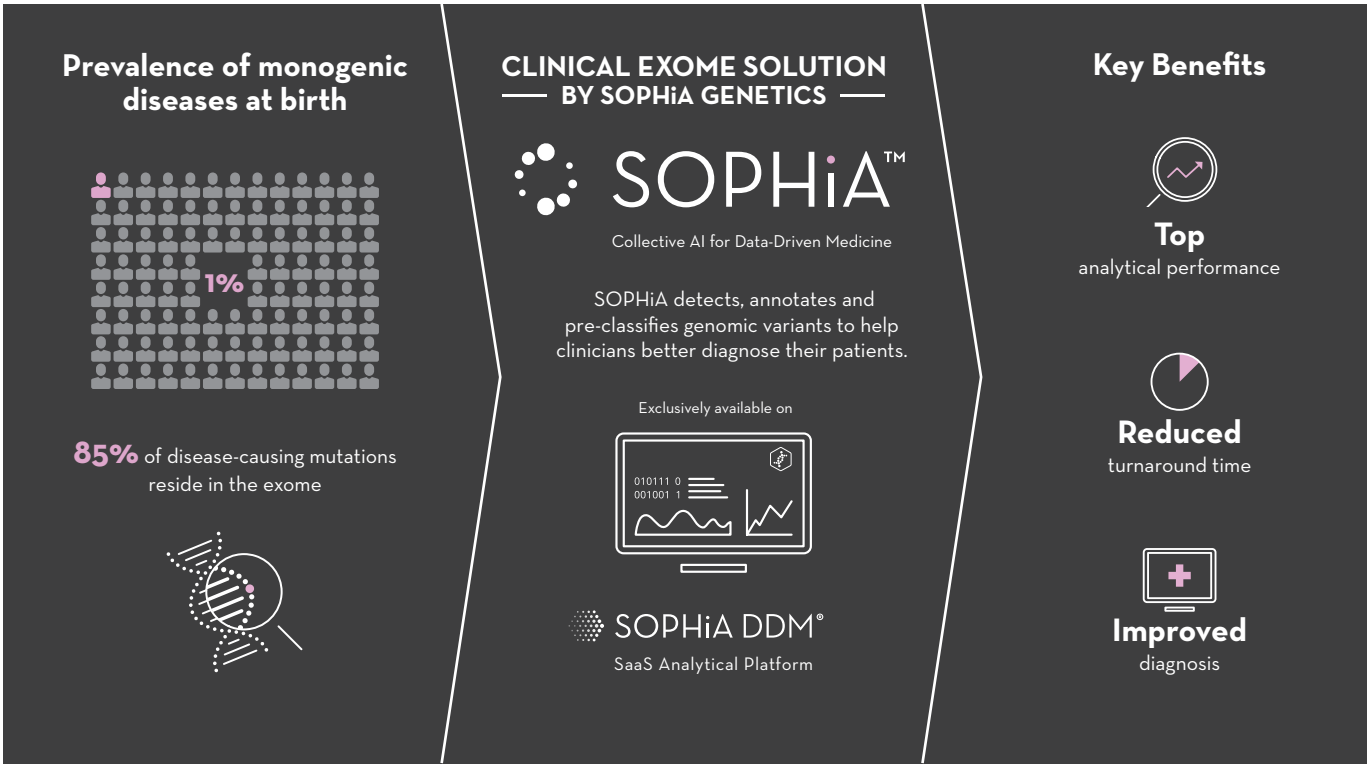
Large Panels

SOPHiA GENETICS®

Unlocking the Power of Clinical Exome Sequencing For Diagnostics



Introduction

Since the completion of the Human Genome Project in 2003[1,2], Next Generation Sequencing (NGS) technologies have begun to revolutionize medical practice.[3] NGS proved to be accurate and cost-effective and compensated for some of the technical weaknesses of older sequencing and genotyping approaches in elucidating a cause for Mendelian (monogenic) diseases.[4] The identification of the genetic basis of these diseases has been challenging and labor-intensive, until a few years ago.

Many disorders have a genetic component and besides monogenic diseases caused by the mutation of a single gene, an increasing number of genetic variants and polymorphisms are being identified as risk factors for complex pathologies.[5,6] The global prevalence of all monogenic diseases at birth is approximately 10/1000, affecting millions of people worldwide.[7] Typical inherited monogenic disorders include Huntington's disease and thalassemia, in addition to approximately 1,000 other inherited rare disorders.[8]

Although the human genome contains 3 billon base pairs (bp), only 1% of these correspond to the exome coding for proteins.[9,10] As such, 85% of disease-causing mutations reside in the exome.[11,12,13] Since the diagnostic rate of rare diseases is estimated at 25%, exome sequencing is becoming increasingly popular.

In recent years, exome sequencing has become a widely used method for identifying the molecular basis of genetic disorders across various medical specialties, when other alternative methods fail to detect causal gene mutations for diseases of Mendelian inheritance.[14,15]

Human exome: variant detection and data analysis

Exome sequencing is being widely adopted for variant detection related to disease-causing protein structural and functional changes.[16] Indeed, Copy Number Variations (CNVs), Single Nucleotide Variants (SNVs) and small Insertions or deletions (Indels) are detected with this approach.[17]

In order to identify causative alterations with a Mendelian pattern of inheritance among the huge amount of data generated, sequenced data is filtered.[5,16,18] Although most of these genetic variants are "passenger changes", few are considered to be "driver changes", directly correlated with the disease.[18]

To identify relevant variants, a comparison between affected and unaffected individuals is needed. It is also essential to predict the importance of these variants and their impact on protein function.

Today, more studies are being conducted, and more variants are being identified as causative of inherited diseases. These mutated genes can be found in specialized databases such as the Online Mendelian Inheritance in Man database (http://www.omim.org).

Comparison of exome sequencing to genome sequencing and targeted gene sequencing

Exome sequencing versus whole genome sequencing

The translation of whole human genome sequencing to clinical practice has been enabled, thanks to the Genome in a Bottle initiative developed by the National Institute of Standards and Technology (NIST).[19] Although whole genome sequencing is becoming more and more affordable, it comes associated with bigger challenges when compared to exome sequencing. The greater amount of data generated from whole genome sequencing makes the identification of driver genes more difficult and requires a higher capacity of computational analysis and data storage.

However, genome sequencing provides a comprehensive view of the genetic alterations present in the patient, including large genome reorganizations such as deletions, inversions, or translocations.[20] Such alterations cannot be detected by exome sequencing since they do not affect encoded proteins. Therefore, whole genome sequencing provides more information than exome sequencing at the expense of increased complexity and economic costs.

Exome sequencing versus whole genome sequencing

Targeted gene sequencing is a direct approach used when most of the genes involved in the disease are already known, or when only actionable genes are considered. This is for instance the case of some cancer types due to mutations in a reduced number of genes.[6] The main limitation of this sequencing technique is that it does not allow for the detection of gene mutations not previously related to the disease being investigated.

Challenges facing exome sequencing

Exome sequencing is used as an alternative approach in diagnostics when multigene panels fail to identify causal genes for rare diseases. Nonetheless, its applications are facing 3 main challenges.[21,22]

1) Data analysis

Data analysis forms a core component of NGS-based testing. It involves several steps to quality control raw sequence data and align it to the human genome reference sequence, from which variant calls are generated.[23]

Data analysis remains the most challenging part of exome sequencing. Overcoming technological limitations, caused by DNA enrichment methods and sequencing platforms, is necessary for a rapid and precise variant detection. Most of the technologies used for exome sequencing and other NGS tests introduce biases, making data analysis very complicated, thus

affecting the sensitivity, specificity, reproducibility and other analytical performance.

Overcoming such sources of bias starts by adopting a standardized approach of the analytical workflow from sample preparation to data generation and analysis. It is necessary to develop optimal techniques for DNA library preparation to ensure full coverage of all the exons.

Furthermore, given the large number of variants that exome sequencing produces, it is challenging to select the most relevant variants directly related to the disease being investigated. Filtering based on population frequencies or prediction scores leaves users with too many variants. Therefore, it is essential to develop filtering options to render the analysis more efficient.

Moreover, to pinpoint a small subset of functional variants, many annotation tools have been developed. Reporting CNVs, Indels and SNPs depends on the algorithms used, which can lead to a substantial discordance in the description of variants between the various annotation tools and databases. Thereby, a lot of information can be missed when the same variants are represented differently in different databases and resources. These issues highlight the urgent need to unify standards in variant annotation, with consistent reporting within the genomic context, to enable accurate data-driven care medicine.[24]

2) Incidental findings

A major concern of exome sequencing is related to the reporting of incidental findings, defined as results that are not related to the indication/disease being investigated, but may be of medical utility to the patient.

The American College of Medical Genetics and Genomics (ACMG) Working Group specified a set of disorders, the relevant associated genes and certain categories of variants that should be reported, based on a consensus-driven assessment of clinical validity and utility. They prioritized disorders where preventative measures and/or treatments were available and disorders in which individuals with pathogenic mutations might be asymptomatic for long periods of time. Typical examples are Hereditary Breast and Ovarian Cancer, Li-Fraumeni Syndrome, Arythmogenic Right Ventricular Cardiomyopathy, and more.[25]

According to the European Society of Human Genetics (ESHG), it should be decided, at the laboratory, institute or national level, whether patients are offered opt-in, opt-out options to get additional information besides the initial diagnostic result.[26]

3) Data privacy and storage

Given the increasing number of requests for access to patients' genomic data, privacy has developed into a vital challenge. Therefore, to prevent unsolicited use of personal information, genomic data confidentiality should be preserved. Privacy safeguards include data encryption, password protection, secure data transmission, audits of data transfer methods, and strategies against breeches and abuse of the data.[27]

Today, it is of high priority to develop a privacy-protecting technology for the storage and access to patients' genomic information worldwide.

As the costs of NGS decrease the data volumes increase, especially when dealing with large sequencing panels, requiring more storage space. But data storage costs can be prohibitive. It requires hardware maintenance and physical space and sometimes, older data must be moved or deleted to make some additional space for the newly generated ones.

Clinical Exome Solution™ (CES) by SOPHiA GENETICS

The Clinical Exome Solution (CES) by SOPHiA GENETICS bundles the analytical power of SOPHiA™, the collective Artificial Intelligence (AI) for Data-Driven Medicine, with a capture-based target enrichment kit and full access to SOPHiA DDM® platform. It consists of 116,355 individually designed probes and spans 11 megabytes (Mb) of target region covering more than 4,900 genes with known inherited disease-causing mutations.

Most of the technologies used in clinical exome sequencing do not provide full coverage of all the exons. Therefore, in order to generate high quality raw genomic data, SOPHiA GENETICS used a knowledge-driven kit design that leads to a high percentage of on-target reads and uniform coverage even through difficult templates (GC-rich regions), thus improving sensitivity with minimal impact on specificity. As a result, the first exons of all genes included in the clinical exome panel by SOPHiA GENETICS are always optimally covered.
Although clinical exome sequencing applications are facing some technical and ethical challenges, numerous hurdles are overcome by the CES solution:

1) Data analysis is eased by SOPHiA AI

SOPHiA has been exposed to over half a million of total variants that have been curated or confirmed by experts. This enabled

SOPHiA to learn how to solve complex patterns and reach unmatched levels of specificity and sensitivity in targeted and large panels such as that for the clinical exome.

SOPHiA overcomes biases, caused by DNA enrichment methods and sequencing platforms, and accurately detects, annotates and pre-classifies all types of genomic variants (SNPs, Indels and CNVs) generated by clinical exome sequencing, to help clinicians better diagnose their patients.

SOPHiA is particularly important for efficient variant annotation since it resolves the representation discordance between individual variant sources such as public databases (ClinVar, ExAC, dbSNP and more) literature and patient variants. This discordance prevents variants to be reliably matched and can therefore lead to a loss of information. SOPHiA features a database search engine (MOKA™) that ensures the identification and retrieval of matching variants regardless of their particular representations, such as Indels aligned differently or complex variants. Indeed, SOPHiA annotates variants in a robust way to help clinicians better interpret genomic data and achieve accurate and precise clinical care.

Given the large number of variants that exome sequencing produces, it is also challenging and time-consuming to select the most promising variant candidates that best explain the patient's phenotypes. To make the analysis more manageable, SOPHiA DDM offers features that allow to define and edit custom filters for efficient and dynamic analysis of exomes.

Additionally, SOPHiA promises a more effective and complementary strategy to identify relevant variants. This strategy compares variants of an affected individual with those of his/her mother and father and then uses Mendelian-consistent and Mendelian-inconsistent inheritance patterns as criteria to filter the variants. This approach may require testing different modes of inheritance (recessive, dominant, or de novo) successively or in combination, to discover the true disease-causing variants. And in certain cases, the analysis can be extended to other members of a family (more than two generations, siblings...).

2) Incidental findings are reduced

Reporting incidental findings is a major concern when large panels such as clinical exome are sequenced. They may be of high medical utility to the patient who can choose to restrict the interpretation of her/his analysis.

Therefore, it is possible to restrict the interpretation to sub-panels of genes, filter for variants in genes of interest and thus prevent the detection of incidental findings. Through SOPHiA DDM, users can specify a virtual panel at the analysis level, which corresponds to the consent restriction. For instance, if a patient chooses to only look at the genes involved in the disease addressed by her/his reason for referral, only these specific genes will be analyzed.

3) Data privacy and storage are secured

Data privacy is becoming an increasingly serious concern. Preventing unsolicited use of personal data requires not only data encryption but also strict definitions of data access privileges that enable selective retrieval of genomic data.

SOPHiA GENETICS is certified for the ISO 27001 for Information Security Management and guarantees the protection of personal data. It is also developing an advanced privacy-protecting technology for the storage and access to patients' genomic information worldwide, in collaboration with genomic data privacy and security experts from the Swiss Federal Institute of Technology Lausanne (EPFL), and biomedical researchers from Stanford University. This new technology, SECRAM (Selective retrieval on Encrypted and Compressed Reference-oriented Alignment Map), makes sure the privacy of genomic information is not compromised as patients' data go through various processing steps, from compression to storage and finally access by healthcare institutions.

Additionally, SECRAM will offer an effective space-saving solution for the storage of clinical genomic data.[28] "The currently available solutions were developed before the widespread usage of high-throughput technologies and do not consider effective protection when compressing genomic sequences; the current standard saves 34% storage on average in lossless compression. SOPHiA GENETICS' no-compromise solution uses 18% less storage while allowing for unprecedented levels of security in genomic data storage as well as selective retrieval", says Jean-Pierre Hubaux, Professor at EPFL.

Reference partner

BioAnalytica-Genotype S.A. in Athens offers Cytogenetics and Molecular Genetics tests for diagnosis, prognosis and therapy selection of haematological malignancies and solid tumors. The Molecular Genetics laboratory was looking for a NGS-based solution to improve the diagnosis of hereditary and rare disorders. They rapidly adopted the Clinical Exome Solution by SOPHiA GENETICS, thanks to its superior performance.

"Being a quality-oriented company, we give absolute priority to quality above everything else, and with SOPHiA GENETICS, we have found the partner to our expedition to Molecular Diagnostics that has no match in the world. We decided to adopt clinical exome sequencing to identify disease-causing variants for patients with hereditary and rare diseases. We tried with enormous success the CES solution by SOPHiA GENETICS, which combines the absolute superiority of probe capture design with the insurmountable data analysis, thanks to SOPHiA, the collective artificial intelligence for Data-Driven Medicine".

Dr. Pantelis Constantoulakis, Head of the Molecular Genetics Department at BioAnalytica-Genotypes S.A.

Conclusion

The widespread adoption of high throughput sequencing technologies toward clinical applications, and the progress made with the development of analytical solutions for genomics have increased the understanding of the human genome and its relation to diseases. Clinical exome sequencing is now being used for diagnostics when other sequencing methods fail to identify alterations in genes associated to hereditary and rare diseases.

The Clinical Exome Solution by SOPHiA GENETICS leads to a performance that reaches the standards required for clinical diagnostic testing.

With the Clinical Exome Solution by SOPHiA GENETICS, healthcare institutions can better diagnose their patients that are at high risk of developing for instance Mendelian disorders. This will ultimately enable the prescription of adapted treatments.

Moreover, SOPHiA GENETICS has built the World's Largest Clinical Genomics Community with more than hundreds of institutions participating in the democratization of Data-Driven Medicine. By adopting the CES, experts become members of the community, allowing them to anonymously and safely share knowledge for an improved variant interpretation.

In the coming years, most of the genetic variants associated to diseases will be discovered and reported to public databases, allowing clinical exome sequencing to be fully adopted in the medical practice.

Sensitivity	> 99%*
Precision	> 99%*
Reproducibility	> 99%
Average on-target rate	> 90%
Coverage uniformity	> 99%
Average percentage of target region > 50x	98%

Performance metrics are based on high confidence regions in a reference sample. Values have been calculated on a reference sample and 10 M fragments per sample (300 bp read length).

References

1: Exome sequencing explained: a practical guide to its clinical application. Eleanor G. Seaby; Reuben J. Pengelly; Sarah Ennis. Brief Funct Genomics. 2016 Sep;15(5):374-84. doi: 10.1093/bfgp/elv054. Epub 2015 Dec 9.

2: International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature 2004; 431 (7011): 931 – 45.

3: A framework for variation discovery and genotyping using next-generation DNA sequencing data. DePristo MA Banks E Poplin R et al. Nat Genet 2011; 43 (5): 491 – 8.

4: The diploid genome sequence of an individual human. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, et al. PLoS Biol. 2007;5: e254.

5: Exome sequencing and the genetic basis of complex traits. Kiezun A, Garimella K, Do R, et al. Nat Genet. 2012;44(6):623–630.

6: Exome sequencing: what clinicians need to know. Sastre L. Jounal of Advances in Genomics and Genetics. 2014; 4:15-27.

7: http://www.who.int/genomics/public/geneticdiseases/en/index2.html

8: Novel genomic techniques open new avenues in the analysis of monogenic disorders. Kuhlenbäumer G, Hullmann J, Appenzeller S. Hum Mutat. 2011;32(2):144–151.

9: Genomics: ENCODE explained. Ecker JR, Bickmore WA, Barroso I et al. Nature 2012; 489 (7414): 52 – 5.

10: The ENCODE (ENCyclopedia of DNA elements) project. Consortium EP. Science 2004; 306 (5696): 636 – 40.

11: Exome sequencing as a tool for Mendelian disease gene discovery. Bamshad MJ, Ng SB, Bigham AW et al. Nat Rev Genet 2011; 12 (11): 745 – 55.

12: Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Botstein D, Risch N. Nature Genet 2003; 33: 228 – 37.

13: What can exome sequencing do for you? Majewski J, Schwartzentruber J, Lalonde E et al. J Med Genet 2011; 48 (9): 580 – 9.

14: Targeted capture and massively parallel sequencing of 12 human exomes. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, et al. Nature. 2009; 461:272–6.

15: Clinical exome sequencing for genetic identification of rare Mendelian disorders. Lee H, Deignan JL, Dorrani N et al. JAMA. 2014 Nov 12;312(18):1880-7. doi: 10.1001/jama.2014.14604.

16: Exome sequencing: the sweet spot before whole genomes. Teer JK, Mullikin JC. Hum Mol Genet. 2010;19(R2): R145–R151.

17: Performance comparison of exome DNA sequencing technologies. Clark MJ, Chen R, Lam HY, et al. Nat Biotechnol. 2011;29(10):908–914.

18: Next generation sequencing in clinical medicine: Challenges and lessons for pathology and biomedical informatics. Gullapalli RR, Desai KV, Santana-Santos L, Kant JA, Becich MJ. J Pathol Inform. 2012; 3:40.

19: http://jimb.stanford.edu/giab/

20: Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Stephens PJ, Greenman CD, Fu B, et al. Cell. 2011;144(1):27–40.

21: Next generation sequencing in clinical medicine: Challenges and lessons for pathology and biomedical informatics. Gullapalli RR, Desai KV, Santana-Santos L, Kant JA, Becich MJ. J Pathol Inform. 2012;3: 40.

22: Challenges in medical applications of whole exome/genome sequencing discoveries. Marian AJ. Trends Cardiovasc Med. 2012;22(8):219–223.

23: Next-Generation Sequencing Informatics: Challenges and Strategies for Implementation in a Clinical Environment. Roy S, LaFramboise WA et al., Arch Pathol Lab Med. 2016 Sep;140(9):958-75. doi: 10.5858/arpa.2015-0507-RA. Epub 2016 Feb 22.

24: A variant by any name: quantifying annotation discordance across tools and clinical databases. Yen JL, Garcia S, Montana A et al., Genome Med. 2017 Jan 26;9(1):7. doi: 10.1186/s13073-016-0396-7.

25: ACMG Recommendations for Reporting of Incidental Findings in Clinical Exome and Genome Sequencing. Robert C. Green, Jonathan S. Berg, Wayne W. Grody et al. Genet Med. 2013 July; 15(7): 565–574. doi:10.1038/gim.2013.73.

26: Guidelines for diagnostic next-generation sequencing. Matthijs G, Souche E, Alders M, Corveleyn A et al. Eur J Hum Genet. 2016 Jan;24(1):2-5. doi: 10.1038/ejhg.2015.226. Epub 2015 Oct 28.

27: Ethical, legal, and social implications of incorporating genomic information into electronic health records. Hazin R., Brothers K.B., Malin B.A., Koenig B.A., Sanderson S.C.et al. Genet. Med. 2013; 15:810–816. doi: 10.1038/gim.2013.117.

28: A privacy-preserving solution for compressed storage and selective retrieval of genomic data. Huang Z[1], Ayday E[2], Lin H[1], Aiyar RS[3], Molyneaux A[4], Xu Z[4], Fellay J[5], Steinmetz LM[3,6], Hubaux JP[1]. Genome Res. 2016 Dec;26(12):1687-1696. Epub 2016 Oct 27.