Paediatr Croat. 2013;57:295-300
PREGLED / *REVIEW*
www.paedcro.com

Next generation sequencing – towards translation into clinical practice

Aleš Maver, Borut Peterlin*

Despite substantial efforts, the cause of a significant proportion of mendelian disorders is still unknown. Recently, significant advances in determination of nucleotide sequence present an opportunity to investigate the genetic material down to a single nucleotide resolution on the whole exome or genome scale. This opens complete possibilities for deciphering genetic etiology of genetic disorders and, most importantly, expanding the field of genetic studies from limited cases of diseases occurring in large families to smaller familial configurations, making detection of causative genes possible even in trios of parents with a single affected proband. Despite the development in both the technology of next generation sequencing and increases in computational power for complex data analysis, the interpretation of results, making sense of genome-scale information in medical setting and conveying this information to the patient, have been lagging behind. In addition, novel ethical questions have opened, especially those related to reporting of incidental findings. These issues are the key points for discussion prior to setting the protocols and guidelines for implementing the whole exome and genome sequencing into clinical practice. This overview will initially give a brief introduction to the technical background behind the next generation sequencing and methodological steps involved. Basic introduction to some of the key steps in data analysis of the next generation sequencing data and challenges encountered thereby will also be given, based on our experiences at the Centre for Mendelian Genomics in Slovenia.

Keywords: base sequence; genetic diseases, inborn

INTRODUCTION

Less than 4 decades ago, in 1977, *Fred Sanger* and *Alan R. Coulson* presented a method for determination of nucleic acid sequences by a fluorescence-based method, currently known as Sanger sequencing. Utilizing this method, it was possible to determine the sequence of smaller segments of DNA, smaller genomes, and after a considerable 10-year effort, in 2003 the human genome sequence was also determined using it. In 2005, however, an alternative approach to Sanger sequencing was developed, now known as the next generation sequencing (NGS). The main advantage of this approach, i.e. massive parallelization, has allowed for rapid and concurrent sequencing of billions of DNA fragments in minute volumes, thus highly reducing the *per* base price of sequencing (1). The resultant increase in the throughput of NGS has enabled sequencing of the whole human genome in a period of 1 day instead of the 10-year period required to sequence the human genome using conventional Sanger sequencing. This has made whole genome sequencing (WGS) a practical possibility also in the setting of medical genetics, not only due to reductions of resources required to perform the re-sequencing, but also opening the possibility for detection of new causes of human inherited disorders, even in cases where this was not previously tractable by conventional methods. Using the NGS technology, it is now possible to interrogate whole human genome or exome for single nucleotide variants (SNV) and smaller insertion-deletions (indels) potentially causing genetic disorders. In addition, recent advances in bioinformatic analyses of NGS data have extended this potential

* Clinical Institute of Medical Genetics, Department of Obstetrics and Gynecology, University Medical Center Ljubljana, Šlajmerjeva 3, Ljubljana 1000, Slovenia

Correspondence to:
Professor Borut Peterlin, MD, PhD, Institute of Medical Genetics, Department of Gynecology and Obstetrics, University Medical Centre Ljubljana, 1000 Ljubljana, Slovenia, E-mail: borut.peterlin@guest.arnes.si

Primljeno/Received: 19. 9. 2013., Prihvaćeno/Accepted: 10. 10. 2013

also to reliable detection of causal structural variation in the human genome, including segmental deletions, duplications and in some cases translocation events (2, 3).

NEXT GENERATION SEQUENCING

With classical Sanger method of sequencing referred to as "first-generation" sequencing approach, the methods that utilize the novel, massively parallel paradigm are now referred to as "next-generation" sequencing approaches (1). Although several different NGS methods have been developed initially, four main approaches have been used most frequently in recent years. These include sequencing by synthesis (Illmina/Solexa), sequencing by ligation (SOLiD), semiconductor sequencing (Ion Torrent) and pyrosequencing (454). These approaches do differ in the way in which detection of the sequence of short reads is performed: Illumina, SOLiD and 454 technologies use fluorescence detection, whereas Ion Torrent relies on detection of small conductance fluctuations during sequencing reactions (1).

Although the approaches differ in technological aspects, there are two main common features of all methods, i.e. compartmentalization of sequencing reactions and their massive parallelization. Briefly, the input DNA material is initially fragmented into short fragments and adaptors allowing the sequencing to proceed are added to the fragments. Afterwards the reads are fixed or compartmentalized on a flow cell and clonally amplified. In this way, a multitude (up to billions) of compartmentalized reactions can proceed in parallel in a single sequencing run. The results generated are sequences of short reads, in most cases up to 300 base pairs in length, which will be aligned to the reference genome in the following steps. Importantly, because *per* base accuracy of NGS is importantly lower than in Sanger sequencing, each base should be sequenced several times to allow for reliable detection of variants, which is also known as sequencing coverage or sequencing depth. The coverage should also be sufficient to allow for sufficient representation of reads from each of the alleles in the case of heterozygosity. Generally, for most applications today, a minimal required coverage is 50×, meaning that each base in the targeted regions will be sequenced 50 times on average.

FOCUSING ON GENOMIC REGIONS OF INTEREST

Despite the price of sequencing *per* base pair has dropped in orders of magnitude in the last decade with NGS, sequencing of whole genomes is still not economically feasible in a considerable proportion of cases. Using the DNA sample directly to perform NGS, all the regions of the genome would on average be equally represented in the final output. However, in most of the settings, especially in medical genetics, one is predominantly concerned about variants occurring in the coding segments of the human genome where the probability of phenotypic implications for the phenotype under investigation is highest. Additionally, the interpretation of variants found across the non-coding proportions of the genome still remains prohibitively complex. For this reason, methods for focusing the sequencing on the selected parts of the genome have been developed, most commonly referred to as target selection approaches. In this way, the sequences of interest are either selectively amplified or enriched by hybridization with other fragments are washed away (4).

In this manner, almost any target sequence can be captured and sequenced selectively, depending on the definition of the target. In NGS sequencing for clinical application, two main directions of such approach have recently been established: (1) targeted sequencing of specific genes, related to clinical disorders; and (2) whole human exome sequencing, which allows for whole exome re-sequencing, enabling reading subset of all coding regions in the human genome. The two approaches are inherently different, with the first approach allowing for rapid and low cost genetic diagnosis of highly genetically heterogeneous disorders, while the whole exome approach allows for open identification of the cause of disease, including the possibility of identifying novel alterations leading to human diseases with yet unidentified causes.

BIOINFORMATIC PROCESSING OF THE RESULTING DATA

The result of NGS is a collection of short read sequences, along with respective base quality scores for each base pair. Below we outline a bioinformatic pipeline we have implemented at the Center of Mendelian Genomics in Slovenia and have been using to analyze NGS data and for sake of clarity, we divide these analyses into three stages (as presented in Figure 1).

In the primary stage, the sequenced reads are inspected and filtered based on their quality scores and they are then positioned onto the reference genome using the available alignment algorithms (most commonly used are Burrows-Wheeler aligner BWA, bowtie and BFAST aligners, but there are many more available, each with its set of advantages and disadvantages).

In the secondary stage, several measures are taken to allow most specific and sensitive calling of SNVs and indels, which include removal of duplicate reads, realignment around indels, base quality and variant recalibration. All these steps

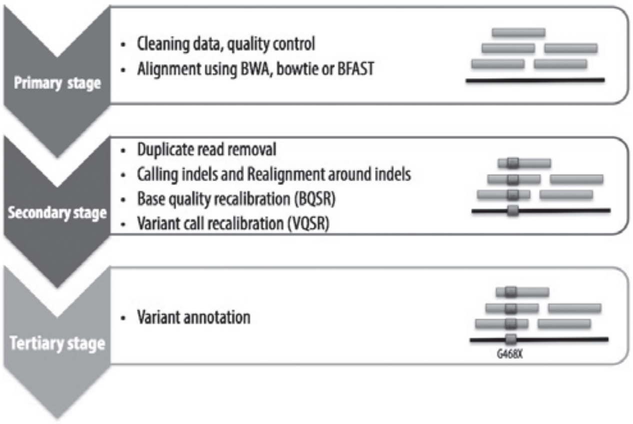
Primary stage
• Cleaning data, quality control
• Alignment using BWA, bowtie or BFAST
Secondary stage
• Duplicate read removal
• Calling indels and Realignment around indels
• Base quality recalibration (BQSR)
• Variant call recalibration (VQSR)
Tertiary stage
• Variant annotation
G468X

FIGURE 1. Bioinformatic analyses of data, the flow of analyses from sequence reads towards results ready for interpretation

are performed aiming to remove as many sequencing artifacts as possible and ensure proper variant calling even in regions where coverage is low or sequencing quality poor.

Usually variant calling is followed by extensive annotation, where we are utilizing sources from a multitude of databases on natural variation in the human genome (dbSNP, ESP6500, 1000 genomes), databases of functional effect predictions (dbNSFP) and tools for functional predictions (SIFT, Polyphen-II, MutationTaster), scores for genomic conservation (GERP and PhyloP). Finally, this is supplemented with data from databases containing clinically relevant information on the gene- and mutation- phenotype relations (including OMIM, Human Phenotype Ontology, ClinVar and HGMD).

Despite using several sources of information for prioritization of variants, there is still no single approach to single out a particular causative variant from the background of 40,000 sequence variants found in the human exome. For this reason, innovative approaches for prioritization and manual investigation of results by a multidisciplinary team is required prior to decision of any variants suitable for clinical reporting.

CLINICAL APPLICATIONS OF NEXT GENERATION SEQUENCING

We envisage a wide range of applications for NGS in clinical practice with the approach to sequencing dependent on the specific clinical setting. As stated previously, each of the two main strategies of sequencing, i.e. gene panel sequencing and whole exome sequencing, has its own specific applications.

Due to the increasing knowledge of genetic heterogeneity of genetic disorders, the number of genes required to test has increased to over 50 in some cases (epilepsy, cardiomyopathy, intellectual disability, and others), making the prices of sequencing using conventional Sanger sequencing prohibitively expensive. In such cases, gene panel sequencing with NGS allows for sequencing of several targeted genes with high coverage and thus represents an economic alternative.

On the other hand, in a significant proportion of genetic disorders, not all causative genes are known and in some cases, the causative gene has yet to be identified. In these settings, WES is a plausible solution to surveying the complete set of human genes at once. In addition, WES can also be of benefit in cases where clinical presentation is unclear and such testing may be used as a screening method to identify relevant mutations and possibly assist in establishment of clinical diagnosis.

With decreasing sequencing costs, WGS is already becoming a potentially viable testing method to be used in clinical practice. While the information amount generated by sequencing of the whole human genome is superior to WES and coverage of genes more uniform (as it does not require target capture protocols), the large amount of information generated and the current lack of proper tools for reliable evaluation of the non-coding variation functional impact currently hinders its effective translation into clinical practice.

CHALLENGE OF INTERPRETATION

One can expect several different outcomes in the clinical use of NGS aiming to identify the underlying genetic cause of a disorder in patients and families. Firstly, either targeted or WES approach may reveal a mutation previously known to cause the phenotype of the patient. Secondly, a novel sequence variant in a gene where other mutations have already been related to the same phenotype may be identified. Thirdly, a novel sequence variant may be discovered in a novel gene not previously associated with the disorder under investigation. The complexity of interpretation increases progressively from the first to the third case. While reporting in the first case is straightforward, based on the previously available evidence base, the second case heavily depends on in-silico predictions of the variant's effect and pathogenicity. In the third case, one would be dependent on both the functional information related to gene carrying the variant, in addition to predictions of the variant's pathogenicity. Furthermore, in the latter case, identification of a potentially damaging variant in a gene with relevant function is not sufficient in most cases and further functional studies of the gene or replication of genotype-phenotype associations are needed prior to establishing the correlation between the identified gene and phenotype.

The other key element of interpretation is the strategy used for diagnostic exome sequencing (5). Several possibilities exist, including sequencing of a single patient, several members of the family, a trio of parents and an affected proband, and other family based designs. The approach to diagnosis is also significantly affected by the expected underlying mode of inheritance and the number of family members available for sequencing. Generally, sequencing of more patients enables easier and more straightforward interpretation of results, while also allowing for identification of novel genes related to genetic disorder under investigation. While sequencing in a single affected individual is also possible, this approach is in most cases limited to finding causative variant with known association to investigated disorder.

Finally, the interpretation should always be carried out taking into consideration the test sensitivity for detecting causative variants. Due to a variety of factors affecting the success of target enrichment and subsequent sequencing, some genes may not be sufficiently covered and some sequence variants may be missed, especially larger indels (greater than the length of sequencing reads). For this reason, the interpretation should be supported by technical information on the target coverage of the genes, giving an overview of which regions, possibly related to the patient's phenotype, have not been targeted sufficiently and should thus be considered in further testing with alternative approaches.

ETHICAL IMPLICATIONS OF WHOLE EXOME AND GENOME TESTING

Gathering information on the level of the complete coding sequence of the genes and the level of complete genomes is introducing substantially new challenges not only into the care of patients with genetic disorders but it increasingly affects other branches of medicine and touches upon health of whole populations (6). Previously, genetic tests mostly focused on detection of a single alteration, usually related to a single disorder under clinical investigation. Even in case of karyotyping, the resolution was limited enough (around 10 Mb) restricting the diagnostics to finding only key alterations leading to elucidation of a clinical trait under investigation.

Sequencing of the whole exomes and genomes, however, has allowed for investigation of the common and rare sequence and copy number variants, not only those directly related to the phenotype under investigation, but also the ones associated with other significant health conditions of the patient. Unsurprisingly, this aspect of WES and WGS has taken ethical consideration in human genetics to a new scale, presenting the need to examine and consider the issues of findings of unknown significance, unsolicited findings, revisable reporting and many others, before such testing can be successfully translated into clinical practice.

For these reasons, several associations and genetic societies across have already assembled various sets of guidelines for such setting, especially focusing on the issue of findings unrelated to the disorder a genomic screen was initially intended for, commonly termed as incidental or unsolicited findings. The proposed guidelines on this issue vary from those advising to actively identify genetic alterations whose discovery could be of immediate benefit to the patients, ranging to those recommendations suggesting a more conservative approach, with reporting only variants with relevance to the disorder under investigation. Currently, the most widely accepted guidelines are those established by the European Society of Human Genetics, issued in 2013, suggesting a thoughtful approach, stressing the importance of adequately informing the patients prior to testing, timely establishing protocols for reporting of incidental findings, revisable reporting and protecting patients' privacy and endorsing collaboration and information exchange among professionals in the field (7). Additionally, they advise gene panel approach to be used instead of WES/WGS whenever possible in an attempt to minimize the possibility of uncovering unwanted incidental findings. Interestingly, the American College of Medical Genetics and Genomics (ACMG) has also issued guidelines at the beginning of 2013, where search for incidental and actionable findings is endorsed in cases where this may directly benefit the patient. Reporting of findings related to untreatable conditions, however, is advised against, a point especially relevant for the case of untreatable adult onset dementia syndromes (8).

Despite these opposing views, there is a general consensus that mandatory genetic counseling should be provided prior to patient's decision to undergo testing, with specific emphasis on explaining the complexity of the test, the possibility on identification of variants whose impact to the phenotype may not be definitely explained at present, and also the possibility of incidental findings.

TRANSLATION INTO CLINICAL PRACTICE

To ensure efficient and active translation of NGS into clinical practice, we have recently established a Center for Mendelian Genomics (CMG) at the Clinical Institute of Medical Genetics in Ljubljana. CMG offers two main approaches towards identification of causative mutations in patients with heterogeneous genetic disorders: gene panel re-sequencing and whole exome re-sequencing. While the first approach is intended to identify causative mutations in disorders with well-defined genetic etiology, whole exome sequencing is geared towards identification of new genes causing mendelian diseases.

We have defined the protocol for inclusion of patients into the diagnostic process, which begins by evaluation of applications for sequencing by a multidisciplinary team of clinical geneticists in tight collaboration with bioinformatics team and molecular biologists. Once the decision on the suitable approach is made, genetic counseling is offered to the patients and with tailed informed consent requested, which require the patients to be fully informed on the possibility of identification of variants of unknown significance, incidental findings and revisable reporting. After sequencing data are processed using our custom pipeline, results are analyzed by a bioinformatician and evaluated by a multidisciplinary team at CMG before releasing final reports.

With the establishment of CMG, we aim not only to constitute a national, but also regional center for diagnosis of mendelian disorders, facilitating more effective diagnosis in these patients.

CONCLUSIONS

In conclusion, next generation sequencing has opened an entirely new field in genomic medicine, allowing for a comprehensive and unified approach to detecting sequence and copy number variations in human genetic disorders. Despite the power of this approach, its translation should be done carefully and in collaboration with professionals with multidisciplinary background, bringing medical, computational and legal knowledge into the interpretation and decision making process. Finally, the ethical implications should be considered and measures introduced to maximize the benefit for the patients.

NOVČANA POTPORA*/FUNDING*
Nema/*None*

ETIČKO ODOBRENJE*/ETHICAL APPROVAL*
Nije potrebno/*None*

DOPRINOSI AUTORA*/DECLARATION OF AUTHORSHIP*
Maver A. – manuscript conception and design, drafting of the manuscript, writing paper, literature search, figure design
Peterlin B. - manuscript conception and design, drafting of the manuscript, writing paper, literature search, critical revision

SUKOB INTERESA*/CONFLICT OF INTEREST*
Autori su popunili *the Unified Competing Interest form na www.icmje.org/coi_disclosure.pdf (dostupno na zahtjev)* obrazac i izjavljuju: nemaju potporu niti jedne organizacije za objavljeni rad; nemaju financijsku potporu niti jedne organizacije koja bi mogla imati interes za objavu ovog rada u posljednje 3 godine; nemaju drugih veza ili aktivnosti koje bi mogle utjecati na objavljeni rad.*/All authors have completed the Unified Competing Interest form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare: no support from any organization for the submitted work; no financial relationships with any organizations that might have an interest in the submitted work in the previous 3 years; no other relationships or activities that could appear to have influenced the submitted work.*

REFERENCES

1. Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010;11:31-46.
http://dx.doi.org/10.1038/nrg2626
PMid:19997069
2. de Ligt J, Boone PM, Pfundt R, et al. Detection of clinically relevant copy number variants with whole-exome sequencing. Hum Mutat. 2013;34:1439-48.
http://dx.doi.org/10.1002/humu.22387
PMid:23893877
3. Chen W, Kalscheuer V, Tzschach A, et al. Mapping translocation breakpoints by next-generation sequencing. Genome Res. 2008;18:1143-9.
http://dx.doi.org/10.1101/gr.076166.108
PMid:18326688 PMCid:PMC2493403
4. Mamanova L, Coffey AJ, Scott CE, et al. Target-enrichment strategies for next-generation sequencing. Nat Methods. 2010;7:111-8.
http://dx.doi.org/10.1038/nmeth.1419
PMid:20111037
5. Gilissen C, Hoischen A, Brunner HG, Veltman JA. Disease gene identification strategies for exome sequencing. Eur J Hum Genet. 2012;20:490-7.
http://dx.doi.org/10.1038/ejhg.2011.258
PMid:22258526 PMCid:PMC3330229
6. Tabor HK, Berkman BE, Hull SC, Bamshad MJ. Genomics really gets personal: how exome and whole genome sequencing challenge the ethical framework of human genetics research. Am J Med Genet A. 2011;155A:2916-24.
http://dx.doi.org/10.1002/ajmg.a.34357
PMid:22038764
7. van El CG, Cornel MC, Borry P, et al. Whole-genome sequencing in health care: recommendations of the European Society of Human Genetics. Eur J Hum Genet. 2013;21:580-4.
http://dx.doi.org/10.1038/ejhg.2013.46
PMid:23676617
8. Green RC, Berg JS, Grody WW, et al. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. Genet Med. 2013;15:565-74.
http://dx.doi.org/10.1038/gim.2013.73
PMid:23788249

SAŽETAK

Sekvencioniranje druge generacije – prenošenje u kliničku praksu

Aleš Maver, Borut Peterlin

Uzrok značajnog broja mendelijanskih bolesti i dalje je nepoznat usprkos znatnim naporima da ga se razjasni. Nedavna značajna postignuća u određivanju sekvence nukleotida pružaju mogućnost ispitivanja genetskog materijala sve do rezolucije jednog nukleotida na razini cijelog eksoma ili genoma. To otvara sve mogućnosti za odgonetavanje genetske etiologije genetskih poremećaja i, što je najvažnije, za širenje područja genetskih studija s ograničenih slučajeva bolesti koji se javljaju u velikim obiteljima na manje obiteljske konfiguracije, omogućavajući time otkrivanje uzročnih gena čak i u roditeljskom triu s jednim zahvaćenim probandom. Usprkos razvoju tehnologije sekvencioniranja druge generacije i sve većoj kompjutorskoj snazi za složene analize podataka, zaostaje interpretacija rezultata kako bi podaci na razini genoma imali smisla u medicinskom okruženju i mogli se prenijeti bolesniku. Uz to, otvaraju se nova etička pitanja, poglavito ona koja se tiču izvješćivanja o slučajnim nalazima. Ovo su ključna pitanja koja treba raspraviti prije negoli se utvrde protokoli i smjernice za primjenu sekvencioniranja cijelog eksoma i genoma u kliničkoj praksi. Ovaj pregled daje kratak uvod u tehničke osnove i korake sekvencioniranja druge generacije, te u neke od ključnih koraka u analizi podataka dobivenih sekvencioniranjem druge generacije, kao i u izazove koji se pritom susreću, a sve to zasnovano na našim vlastitim iskustvima u Centru za mendelijansku genomiku u Sloveniji.

Ključne riječi: bazno sekvenciranje; genetske bolesti, prirođene