Journal of Next Generation Sequencing & Applications

Blum and Oexle, Next Generat Sequenc & Applic
2014, 1:1
DOI: 10.4172/2469-9853.1000105

Letter to Editor Open Access

Clinical Interpretation and Implications of Whole Genome Sequencing

Hubert E Blum[1*] **and Konrad Oexle**[2]

[1]*Department of Medicine, University Medical Center Freiburg, Germany*

[2]*Institute of Human Genetics, Technical University Munich, Germany*

[*]***Corresponding author:*** *Hubert E Blum, Professor of Medicine, Department of Medicine*

University Medical Center Freiburg, Germany, Tel: 0049-761-270-18116; Fax: 0049-761-270-18117; E-mail: hubert.blum@uniklinik-freiburg.de

Rec Date: Apr 29, 2014, **Acc date:** June 26, 2014, **Pub date:** June 28, 2014

Introduction

In 2013, 60 years after the discovery of the double-helical structure of the DNA by James Watson and Francis Crick and 10 years after the completion of the sequence of the human genome, the first market authorization of a high throughput ('next-generation') sequencer (Illumina MiSeqDx) was granted by the American Food and Drug Administration. Next-generation sequencing (NGS) allows to effectively performing whole genome sequencing (WGS) that may lead to the introduction of new genome-based analyses into clinical practice [1]. WGS is one of very few recent developments in medicine which are characterized by the attributes 'better', 'faster' and 'cheaper'. With the most recent development of NGS technologies (i.e., Illumina HiSeqX-Ten), complete sequencing of the human genome is feasible within a day for less than US $ 1'000 (not including costs for curation). Several companies offer NGS solutions and compete for a market share. Technologies differ with respect to their miniaturization (e.g., beads, DNA clusters, DNA nanoballs, single molecule sequencing, nanopores). Most still involve sequence-specific nucleic acid hybridization and amplification. Differences exist in the average reading length as well as in the rate, reproducibility, and cost of sequencing. NGS always involves non-directed mass sequencing. Restriction of sequencing on a genomic region, a panel of genes, or the exome (all protein-coding exons), for instance, can be achieved by a capturing tool based on hybridizing oligonucleotides that define the sequence area. The application of NGS for the analysis of RNAs ('transcriptome') or epigenetic modifications ('methylome') is straightforward. For biomedical scientists as well as for clinicians NGS promises nearly unlimited possibilities to search for genetic variants with clinical relevance. This includes the detection of constitutional as well as of somatic mutations as causes of monogenic diseases, the genome-wide assignment of risk scores for polygenic diseases, the identification of tumor- or metastasis-specific mutations in individual tumor patients as well as the determination of gene expression profiles (signatures') with prognostic and/or therapeutic relevance [2,3]. Another example is pharmacogenomics that may contribute to a personalized therapy, allowing choosing for the individual patient the appropriate drug and/or the appropriate dose.

With simplified access to and the reduced cost of WGS or Whole Exome Sequencing (WES), commercialization of these analyses as well as their increasing application in medicine is to be expected. WGS and WES have not only the potential to answer specific disease-related questions [4] but at the same time may identify clinically relevant genetic variants that are not related to the clinical indication for DNA sequencing. From the identification of unexpected variants, ethical issues may arise with respect to their communication to the patient, parents or others [5]. While these questions are very important, our comment primarily focuses on the utility, reliability, resource needs and other aspects of WGS/WES.

In a recently published study, the WGS technology was evaluated in 12 healthy volunteers [4], using the Illumina/Solexa sequencing principle (paired-end reversible terminator massively parallel sequencing of microscopically amplified DNA clusters) on an HiSeq 2000 platform and, for comparison, the nanoarray technology of Complete Genomics Inc. (rolling circle amplification of DNA fragments followed by massively parallel sequencing of arrayed DNA nanoballs using probe-anchor ligation). Initial quality control and genotype annotation followed standardized procedures. The sequence data obtained were scrutinized for small sets of variants associated with cardiovascular risk or clinical drug response as well as a large set of variants likely or known to be involved in inherited diseases. The latter set was subjected to a two-step evaluation process. At first, an automated algorithm (STMP, 'Sequence To Medical Phenotypes') prioritized potentially pathogenic, previously known or novel variants according to the usual indicators, such as allele frequency (more common variants are less likely to be pathogenic), functional class, and evidence of evolutionary constraints. The STMP software selected 90-127 variants per individual which included possibly disease-causing variants and variants with the possible implication of a carrier status (8-18 per individual). These variants then were curated by a multidisciplinary team of specialists including genetic counselors, physician/information technology specialists, and molecular pathologists. The study addressed the concordance of the two sequencing platforms (99-100% for single nucleotide variants, 53-59% for insertion/deletion variants), coverage (i.e., readability) of disease-causing genes (incomplete in 10-19% of the genes), validity of the automated evaluation process, i. e., the final number of reportable (2-6 per individual) and 'actionable' variants [5], respectively (1 individual with a clearly actionable result, 4 individuals with variants that are disease-causing according to the 'Human Gene Mutation Database', HGMD, but were reclassified by the curation team as 'reportable with unclear significance'), the correlation between the curators (moderate), the resources needed for sequencing and curation (median time 54 minutes for each curated variant, resulting in median cost of about US $ 15'000 for sequencing, including curation), and the cost of clinical follow-up (less than US $ 1'000 per individual). The authors arrived at several conclusions that will be discussed in the following in the context of clinical aspects of WGS/WES analyses.

1. The authors concluded that incomplete coverage of a considerable fraction of disease-causing genes reduces the sensitivity of WGS/WES analyses. Depending on the NGS platform used, 10-19% of disease-causing genes did not meet the threshold of base readability accepted by the authors. Without giving a detailed definition of their accepted threshold, it seems to be set at 1%. Thus, the sequences of most of the

disease-causing genes were in fact readable. While a readability of 99% of a gene is a problem in single gene analysis, it is less problematic in genome-wide analyses where sensitivity correlates with disease-associated and not with gene-associated variants. Further, the overall coverage of the WGS analyses can be increased, thereby closing some sequence gaps while at the same time increasing the cost of sequencing.

2. The authors observed a low reproducibility of the identification of insertions and deletions between the platforms (<33%) but did not assess the relation between reproducibility and size of the variants. The low reproducibility is indeed very important because the insertion/deletion variants are likely to cause a loss of gene function. The technological problem underlying the low reproducibility still awaits a satisfactory solution. While the detection of insertions and deletions in WGS indeed may not yet be reliable enough, the concomitant moderate reduction in WGS sensitivity does not preclude its diagnostic application in patients. In fact, there are classes of mutations, e. g., repeat expansions, that cause Huntington's and various other neurological diseases, which currently cannot be detected by NGS technologies. To expect a 100% sensitivity of WGS analyses is and probably will remain unrealistic.

3. The automated sequence interpretation STMP software (see above) results in a high number of seemingly relevant variants (median 108 per individual) that require intensive curation by specialists but still leave an uncertainty in many cases. With this observation, the authors confirmed the findings of previous larger studies [6-9]. The lack of specificity in automated interpretation may be due to (i) false annotations in data bases, (ii) annotation of scientific errors resulting from different analytical problems, ascertainment biases in penetrance estimates, ethnicity effects, or irreproducibility of rare mutations, (iii) insufficient sophistication and power of interpretation algorithms in correctly predicting the effects of (novel) variants on an individual genetic background, and (iv) to the unclear definition of what is a disease [9,10]. The authors correctly concluded that the lack of specificity impairs the validity of reporting incidental findings of genome-wide analyses. It has to be noted, however, that the above-mentioned limitations in principle also affect single gene analyses by conventional Sanger sequencing. They rarely cause mistakes of these analyses because the probability of identifying a disease-causing variant is usually high if single gene analyses are performed based on a clinical finding. For the same reason, the specificity of WGS/WES analyses is improved if the search area can be restricted by clinical information [3] or by focussing on de novo mutations. Nevertheless, novel guidelines for demonstrating a causal relationship with sequence variants [10] must be followed in order to improve the interpretation of WGS/WES findings in individual patients as well as in the general population.

4. In their 12 healthy study participants the authors identified only 1 variant (frame-shifting deletion of BRCA1) that resulted in a therapeutic or prophylactic action. In a previous exome study [8], including 500 healthy Europeans, the prevalence of clinically relevant genetic variants in adults was 3.4% (determined in a subset of 52/56 'actionable' genes defined by the American College of Medical Genetics and Genomics [5]). Thus, the yield of WGS in unselected individuals is low. On the other hand, 11/12 study participants had 1 or more variants that affected the use or the dosing of a drug.

5. Importantly, significant time is required (about 100 hours per individual) for professional curation after automated evaluation of WGS data by the STMP software. This greatly contributes to the estimated overall cost of about US $ 15'000 per WGS. In principle, there are potentials for process optimization. For instance, 574/1'300 variants identified by the STMP software for professional curation, were listed in the HGMD data base. Among these, 441 were not considered 'disease causing' or 'likely disease causing'. The elimination of these 441 variants would have significantly reduced the time for professional curation without much loss in sensitivity (Table 3) [4]. Overall, the study clearly demonstrated that cost of sequencing is becoming a less important while data curation and interpretation is the major factor of the overall cost of WGS analyses.

In summary, the WGS study [4] confirmed known problems of WGS/WES analyses in unselected/healthy individuals: limited sensitivity for some classes of mutations such as structural variants, high false-positive rates in automated interpretation, high time requirement for professional curation, and a relatively low impact on medical care. In part, this also affects the diagnostic application of WGS/WES in patients [3] as long as there is no *in vitro or in silico* patient- or disease-specific restriction of the sequence analyses. Therefore, there is a great need for the development of restriction strategies that increase the probability of the detection of pathogenic variants, for the improvement of the molecular understanding of genetic effects, interactions and penetrance [9], for further sophistication of effect predicting software, for improvements and quality control of public data bases of mutations and other genetic variants associated with human diseases, for incentives for sharing genotypic and phenotypic data as well as for large-scale NGS genotyping of population cohorts [10].

References

1. Collins FS, Hamburg MA (2013) First FDA authorization for next-generation sequencer. N Engl J Med 369: 2369-2371.
2. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, et al. (2013) Mutational landscape and significance across 12 major cancer types. Nature 502: 333-339.
3. Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, et al. (2013) Clinical whole-exome sequencing for the diagnosis of mendelian disorders. N Engl J Med 369: 1502-1511.
4. Dewey FE, Grove ME, Pan C, Goldstein BA, Bernstein JA, et al. (2014) Clinical interpretation and implications of whole-genome sequencing. JAMA 311: 1035-1045.
5. Green RC, Berg JS, Grody WW, Kalia SS, Korf BR, et al. (2013) ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. Genet Med 15: 565-574.
6. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, et al. (2012) A systematic survey of loss-of-function variants in human protein-coding genes. Science 335: 823-828.
7. Xue Y, Chen Y, Ayub Q, Huang N, Ball EV, et al. (2012) Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. Am J Hum Genet 91: 1022-1032.
8. Dorschner MO, Amendola LM, Turner EH, Robertson PD, Shirts BH, et al. (2013) Actionable, pathogenic incidental findings in 1,000 participants' exomes. Am J Hum Genet 93: 631-640.
9. Cooper DN, Krawczak M, Polychronakos C, Tyler-Smith C, Kehrer-Sawatzki H (2013) Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. Hum Genet 132: 1077-1130.
10. MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, et al. (2014) Guidelines for investigating causality of sequence variants in human disease. Nature 508: 469-476.