Critical Points for an Accurate Human Genome Analysis

Stefan J. White [1*], Jeroen F.J. Laros [1,2,12*], Egbert Bakker[2#], Anne Cambon-Thomsen[3#], Martin Eden[4#], Samantha Leonard[3#], Hanns Lochmüller[5#], Gert Matthijs[6#], Christopher Mattocks[7#], Simon Patton[8#], Katherine Payne[4#], Hans Scheffer[9#], Erica Souche[6#], Ellen Thomassen[1#], Rachel Thompson[5#], Jan Traeger-Synodinos[11#], Steven Van Vooren[10#], Bart Janssen[12#] and Johan T. den Dunnen [1,2#].

[1] Dept. of Human Genetics and [2] Clinical Genetics, Leiden University Medical Center, Netherlands; [3] Inserm and Université Toulouse III Paul Sabatier, UMR 1027, Toulouse, France; [4] University of Manchester, UK; [5] Newcastle University, UK; [6] Catholic University Leuven, Belgium; [7] Wessex regional Genetic Laboratory, Salisbury, UK; [8] Central Manchester Univ. Hospitals Foundation Trust, (EMQN), UK; [9] UMC St. Radboud Nijmegen, Netherlands; [10] Agilent Technologies, Lexington, MA, USA; [11] Medical Genetics, National and Kapodistrian University of Athens, Greece, [12] GenomeScan, Leiden, Netherlands

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/humu.23238.

#) Members of 3Gb-TEST consortium:

* Joint first authors

The 3Gb-TEST project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 602269.

Corresponding Author:

Prof. Johan T. den Dunnen

Depts. of Human Genetics and Clinical Genetics

Leiden University Medical Center

The Netherlands

ddunnen@humgen.nl

Abstract

Next-generation sequencing is radically changing how DNA diagnostic laboratories operate. What started as a single-gene profession is now developing into gene panel sequencing and whole exome and genome sequencing (WES/WGS) analyses. With further advances in sequencing technology and concomitant price reductions, whole genome sequencing (WGS) will soon become the standard and be routinely offered.

Here we focus on the critical steps involved in performing WGS, with a particular emphasis on points where WGS differs from WES, the important variables that should be taken into account, and the quality control measures that can be taken to monitor the process.

The points discussed here, combined with recent publications on guidelines for reporting variants, will facilitate the routine implementation of WGS into a diagnostic setting.

Key words: Whole genome sequencing, Whole exome sequencing, Genetic variation, Bioinformatics, Next generation sequencing

Introduction

Next-generation sequencing (NGS) is being rapidly implemented into the diagnostic setting. Due to the cost of whole genome sequencing (WGS) the first tests were limited to selected regions of the genome, and later either gene panels or whole exome sequencing (WES). These approaches have clearly demonstrated the value of this type of analysis, simplifying genetic screening and identifying a number of new disease genes. They have also given the first indications of the complicating factors that are part of genomic screening, such as unsolicited findings.

There have been several publications suggesting guidelines for reporting NGS-based findings (Green, et al., 2013; Hehir-Kwa, et al., 2015; Rehm, et al., 2013; Weiss, et al., 2013), and a comprehensive report regarding the implementation of exome sequencing in the clinic has also been published (Matthijs, et al., 2015). Many of the points addressed by (Matthijs, et al., 2015) are also applicable to whole genome sequencing, but the greater size and complexity of the human genome mean that there are additional issues to be considered (Table 1). Using (Matthijs, et al., 2015) as a baseline, we here focus on the different technical aspects involved in sequencing a complete human genome; from DNA sample source and preparation to variant calling and description (Table 2), but we are also fully aware of the other clinical, ethical and professional dimensions of such a technology (Julia, et al., 2016). A key consideration, addressed elsewhere, is the need for robust processes of health technology assessment (Payne, et al., 2017, in press).

Sample source

The first step in sequencing a genome is to obtain a DNA sample. There are many possible sources, each of which can influence the findings. For routine clinical purposes a venous blood sample is usually obtained, with genomic DNA isolated from nucleated cells such as lymphocytes. Depending on the isolation method used, such samples may also contain cell-free DNA, usually from dead cells (and would therefore be expected to have the same DNA as found in the nucleated cells). If the sample is obtained during pregnancy, there will be fetal DNA (actually derived from the placenta) present in the blood. Indeed, a recent report showed that it was possible to sequence the fetal genome solely by analysing circulating cell-free DNA from the mother (Kitzman, et al., 2012).

Likewise, if the person providing the sample has a cancer, then cells/DNA from a tumour may have been shed into the circulation. The proportion of different cell types within a blood sample can vary considerably (Whitney, et al., 2003) and it is important to note that B and T lymphocytes will have undergone somatic rearrangements at their respective cell receptor genes. Genomic DNA isolated from these sources may appear to contain deletions at these loci, if compared to DNA from non-lymphocytic cells.

When analysing a tumour, the issue of sample purity is a critical issue, as the biopsy may contain varying amounts of contaminating non-tumour tissue. In addition, it is well recognised that tumours are heterogenous, and may contain clonal subtumours within the sample (Marusyk and Polyak, 2010).

It is also possible for non-human DNA to be present in a sample. Low invasive methods for obtaining DNA such as buccal swabs or spit kits are often used (Freeman, et al., 1997). However, food material may contaminate DNA derived from a buccal swab or saliva sample. Additionally, bacteria such as *Streptococcus parasanguinis* are commonly found in the oral cavity, and may be a significant contributor to a DNA sample isolated from this source (Mahfuz, et al., 2013). Potential contaminations in a specific genomic DNA sample may derive from infections of the individual (viral, bacterial, etc.) or any cell culturing when performed. This is not necessarily an issue for WES, as the enrichment step will enrich for human sequences and therefore automatically reduce the level of contamination, but most WGS protocols will sequence all DNA present in a sample, irrespective of source.

Although cultured cells are an easy source of DNA, it is known that culturing can lead to genomic instability (Adey, et al., 2013; Maitra, et al., 2005; Narva, et al., 2010). In addition, depending on the stringency of the culture conditions, contaminants might be present like mycoplasma, yeast, etc. It is also possible that the cell line in question becomes contaminated with another cell line, or even overgrown by a different cell line completely. There have been numerous reports demonstrating that a significant proportion of cell lines are not in fact what they are thought to be (American Type Culture Collection Standards Development Organization Workgroup, 2010; Capes-Davis, et al., 2010; Chatterjee, 2007), and many journals now require evidence that cell lines used in a report are correct. It has now become a service of a number of biobanks to verify and authenticate cell lines used in research projects, prior to publication submission.

Another potential issue is the replication state of the cells. If DNA is isolated from proliferating cells, then DNA from early replicating regions will be over-represented in comparison with sequence from late replicating regions. This is especially important when looking at possible copy number variants or mosaic changes (Koren, et al., 2014).

Reports have shown it is possible to sequence the genome of a single cell (Navin, et al., 2011; Wang and Navin, 2015) and even DNA and RNA from the same cell (Dey, et al., 2015). Such sequences will never be complete, however, as part of the genome will be lost during preparation and sequencing. It is also important to note that every cell has its "own" genome, even when isolated from the same tissue within a single individual, due to *de novo* variation introduced during growth, derived from DNA replication and/or DNA damage repair. WGS analysis of individual cells has shown significant variation, ranging from large CNVs (Cai, et al., 2015; Knouse, et al., 2016) to LINE1 retrotranspositions (Evrony, et al., 2015).

DNA isolation

Ideally, clean and high molecular weight DNA will be obtained, and this will usually be the case from freshly obtained samples. Older DNA samples or those isolated from archived material may well be degraded or contaminated. DNA degradation is rarely random, meaning that genomic regions will be unequally represented. This will be reflected in non-

uniform coverage of specific regions (e.g. chromosome ends, AT/GC rich regions, etc.), and may preclude sequence analysis with long read instruments. In fact the data obtained can be used to monitor the extent of degradation, when present.

Sequencing Platforms

Although there are now a number of different NGS platforms available (reviewed in (Reuter, et al., 2015)), only a few are amenable for routine human WGS. These can be loosely divided into short read (<1kb read length) and long read (>1 kb) sequencers. The current market leaders for short read sequencers are the Illumina systems. Illumina sequencing chemistry is based on sequencing-by-synthesis, using nucleotides linked to fluorescently labelled terminators (Bentley, et al., 2008). Paired end sequencing allows both ends of the molecule to be read, theoretically doubling the amount of sequence produced and potentially allowing the identification of structural variants through discordant mapping of the two ends.

Of the sequencers that can routinely produce reads >1kb in length, the Pacific Biosciences RSII is the most commonly used for human WGS. Sequencing takes place in individual wells on a Single Molecule, Real-time (SMRT) cell. Using this approach it is possible to produce sequence reads >50kb in length, although with a much lower output and higher error rate than the Illumina. The error rate can be compensated for with sufficient read depth, as the errors are unbiased. This allows for regions not amenable to sequencing with Illumina chemistry e.g. homopolymer stretches or loci with high or low GC content, to be sequenced.

Several human genomes sequenced with this technology have been described (Chaisson, et al., 2015a; Zook, et al., 2016).

Another example of a long read sequencer is the Oxford Nanopore system. This technology reads the nucleotides in a DNA strand by measuring the change in electrical current as the DNA molecule passes through a pore in a membrane. Advantages over other sequencing platforms are the low startup cost, portability, and speed from sample to data. Primarily used for smaller genomes (Loman, et al., 2015; Quick, et al., 2016) the first human genome completely sequenced with using this technology was announced at the end of 2016.

Sample preparation

The method of DNA isolation used determines whether specific contaminants may be present or not. When blood is centrifuged cells can be separated from cell-free DNA and other methods may be used to reduce the amount of DNA derived from viruses or bacterial contaminants. Sample preparation mostly involves steps that depend on the sequencing platform to be used, but generally involves some form of DNA fragmentation followed by the attachment of linker sequences to facilitate the sequencing reaction. Again, fragmentation is not random, and depends on different factors including GC content, genomic location, etc. Currently popular methods using enzymatic fragmentation (transposases) are known to demonstrate some bias due to e.g. GC content (Lan, et al., 2015; Marine, et al., 2011). In addition, the size of the DNA fragments generated will influence possibilities for haplotype construction.

A PCR step is often used during library preparation, particularly when only small amounts of starting DNA are available. However, this step will also introduce biases. A study looking at the effect of GC content found that there was reduced abundance of DNA fragments with extreme GC% (<10% and >60%) (Aird, et al., 2011). This could be ameliorated to a degree, by optimising the PCR conditions with regards to temperature and DNA polymerase. Another study also showed that DNA polymerase choice can be critical in reducing bias (Dabney and Meyer, 2012).

It is also possible to prepare a genomic DNA sample for sequencing using an amplification-free protocol. In one such approach, adapters containing the sequences necessary for attachment to the Illumina flow cell are ligated onto the fragmented DNA (Kozarewa, et al., 2009). Amplification during cluster generation on the flow cell enriches for sequences containing the correct adaptors. This amplification step is inherent to the Illumina sequencing system, and it is only single molecule sequencing platforms like Pacific Biosciences (Eid, et al., 2009) and Oxford Nanopore (Clarke, et al., 2009) that do not include some type of amplification. As every amplification step introduces biases, amplification should be restricted to a minimum (Aird, et al., 2011; Kozarewa, et al., 2009; van Dijk, et al., 2014).

Read alignment

The two main variables in read alignment are the choice of aligner and the choice of reference genome. Many different algorithms have been developed for aligning short read

sequences (reviewed in (Ye, et al., 2015)), with Bowtie/Bowtie2 (Langmead and Salzberg, 2012; Langmead, et al., 2009) and BWA (Li and Durbin, 2009) amongst the most popular.

The reference sequence used for mapping has significant consequences, especially with regards to repeat regions, unassigned sequences (unplaced contigs), and different haplotypes (present as haplotype chromosomes or alternative alleles). Before a transition is made to a new human genome build (e.g. hg19 to hg38) existing data should be analysed relative to the old and new reference, and all differences should be understood.

Metrics of all sorts and from all stages in the pipeline can and should be stored in a QC database. The aim of such a database is to gather information about the distribution of these metrics, in order to automatically find outliers. If, for example, the GC content of all fastq files is stored, it can be noted that the GC content on average will be 40%, with a very small deviation (less than 1%). Any sample that has a GC content that significantly deviates from this distribution should be set aside for further scrutiny. Likewise, the Transition/Transversion ratio (Tr/Ti) can be used as a quality control for single nucleotide polymorphisms (SNPs), which is typically higher for exons compared to introns, and higher for synonymous SNPs compared to non-synonymous SNPs (Wang, et al., 2015).

There are several important stages in the pipeline where QC metrics are gathered. In general one should try to capture a distribution of a metric rather than one value. For example, do not store the average insert size of the library, but rather the distribution of insert sizes (using 100 bp bins).

Short read sequencers, e.g. Illumina HiSeq, produce sequence reads that are not sufficiently long to be unambiguously mapped to certain repetitive sequences in the genome, especially when they contain expanded (disease-associated) alleles. This has consequences for the analysis of e.g. CNVs/SVs in these regions, and for trinucleotide repeat disorders. Not only are repeat expansions like CGA and CGG extremely sensitive to PCR-bias, when there is insufficient unique flanking sequence present, reads will not be mapped with high confidence and it is unlikely that variants will be called reliably.

Given the difficulties observed with read mapping, and the associated consequences for accurate variant calling, it might be advantageous to make a reference genome that has been optimised for diagnostic purposes. This diagnostic reference genome should be modified at all sites where read mapping is not optimal for reliable variant detection. Such regions include those underlying trinucleotide repeat expansion disorders, where the ideal reference genome would contain an extended repeat, ensuring the mapping of all reads and thereby accurate repeat-length scoring. Similarly, repeated segments of genes such as *DRD4* (48bp unit) and *PRNP* (48bp unit), can be extended to improve variant detection and allele sizing. Larger duplicated sequences can be reduced to a single copy, ensuring that variants will not be missed because they are randomly distributed over the different copies. Linking variants based on this clinical reference genome to the standard reference genome build (e.g. hg38) can be achieved using a simple genomic coordinate translation table.

For quality purposes a number of specific genomic regions should be selected and used the generate quality metrics, as well as for determining coverage and the ability to detect

variants. These can be based on specific criteria, e.g. high (around the promoter of *EGFR* (Obradovic, et al., 2013)) and low (around exon 2 of *DMD* (White, et al., 2002)) GC content, repeats (variable number of tandem repeat (VNTR) polymorphisms within the *DRD4* gene (Van Tol, et al., 1992) or the *MUC1* gene (Kirby, et al., 2013)) and include sequences that are/are not included in exome analysis (exonic and intronic/intergenic regions respectively). Loci not expected to be covered using the applicable sequencing approach should also be included.

Small sequence variants (SNPs/Indels/STRs)

Many different algorithms have been described for calling variants in NGS data (reviewed in (Nielsen, et al., 2011)). Comparative studies have shown that no single algorithm can detect all variants, that there is <100% overlap between algorithms, and that indels are especially difficult to detect (Cornish and Guda, 2015; Liu, et al., 2013). Analysis of short tandem repeats (Press, et al., 2014) are complicated by the high mutation rate due to polymerase slippage. It has been shown that non-PCR amplification is better for STR analysis when compared to an amplification-based protocol, with a 9-fold reduced error rate (Fungtammasan, et al., 2015).

Structural Variation

There are many different types of structural variation (SV), including deletions, duplications and amplifications (collectively known as copy number variation, or CNV), insertions, transpositions, and translocations (Alkan, et al., 2011). There have been several approaches described for CNV analysis of WES data (Fromer, et al., 2012; Wu, et al., 2012) but all suffer from highly variable coverage derived from several steps in the process (incl. GC-percentage, PCR-bias, protocol changes, capture probes, capture efficiency, etc.). To address the variability in coverage, tools that are designed to do CNV calling on WES data either use reference sets or call CNVs in a population. An example of such a tool is cn.MOPS (Klambauer, et al., 2012), which aims to explain the coverage distribution per position within a set of samples by a mixture of Poisson distributions. If more than one distribution is needed to explain the observed variation, a (common) copy number variation is detected. This approach works well if the copy number variation is relatively common in the input dataset. Another well known tool, XHMM (Miyatake, et al., 2015), uses a combination of principal component analysis normalisation and a hidden markov model to detect CNVs. A relatively large number of samples (at least 50) should be used to reliably call CNVs within this set. Finally, WISECONDOR, a tool originally developed for non-invasive prenatal testing (NIPT) (Straver, et al., 2014) has successfully been used to detect CNVs in WES data. This tool detects violations of correlated coverage of bins within one sample. A reference set is used to establish the correlation, the CNV detection itself is done per sample. For all of the approaches described above, either a batch of samples, or a reference set is needed to cope with the high variability of coverage within a WES data set. Additionally, there is usually no possibility of determining the exact breakpoint, meaning that the exact nature of

duplications will usually be unclear and other types of SV, e.g. inversions will not be detected. Using WGS data, especially when using amplification-free protocols, gives a more uniform coverage, making SV detection more reliable and sensitive. In addition, WGS samples should contain the unique breakpoint sequences, instrumental for resolving the identity and exact borders of the SV.

In contrast to SNPs, a relatively low level of sequence coverage is sufficient to detect many types of SV. Aneuploidies can be detected in cell-free DNA as part of NIPT (Brady, et al., 2015), with <10 million reads (<1x average coverage) being sufficient for whole chromosome aneuplodies (Chiu, et al., 2008; Fan, et al., 2008). Multiallelic CNVs (mCNV) are specific genomic regions that can be present in a range of different copy numbers, and are difficult to accurately genotype (Cantsilieris and White, 2012). A report by Handsaker et al. (Handsaker, et al., 2015) modified a previously published read depth approach (Handsaker, et al., 2011) to identify and genotype >1000 mCNV loci in samples from the 1000 genome project.

Unmapped Reads (Dustbin analysis)

Reads that do not initially map to the reference may still be informative. This may be due to a read containing multiple variants, spanning a deletion/duplication/inversion breakpoint, insertion of non-reference sequence (viral, repeat), or representing a sequence not present in the reference genome. Different analytical approaches need to be applied for each of these possibilities. Not all reads are mapped perfectly or with only a small number of

mismatches. Different types of unmapped reads include: (i) split reads, (ii) discordant reads, (iii) unmapped mates, (iv) soft clipped reads. Split reads are direct evidence of translocations, but are very hard to find as they require rather clean breakpoints on both ends of the structural variation event. Pindel (Ye, et al., 2009) was designed for this specific use case. It uses one of the reads in a read pair as an anchor. Fragments of the other read are then located in the vicinity of the anchor read. This method is able to detect deletions, inversions, translocations and, to a certain extent, novel insertions. A related technique named BreakDancer (Chen, et al., 2009) was developed to exploit the presence of discordant reads. Especially in combination with mate pair sequencing, this technique is suitable for breakpoint detection of larger structural variation events, even in the absence of clean breakpoints.

Specific algorithms have been developed for identifying inversion breakpoints in PacBio sequence data, by reanalysing reads that did not initially align (Chaisson, et al., 2015a).

Unmapped mates and soft clipped reads are indirect evidence of large structural variations that may be translocations, but also large insertions, tandem duplications, etc. Although analysing these reads by themselves is not sufficient to draw any conclusions about the nature of the structural variation, it does indicate that there may be an aberration. Also, in combination with CNV calling it may exclude or support certain types of events.

Some read pairs cannot be mapped to the reference genome at all. There are several possible reasons for this, e.g. the reference sequence contains too many copies of this sequence, or the reference sequence does not contain the sequence of interest. The first

case is easy to detect from the alignment file, however there are multiple potential causes for the latter case:

- The sequence is sample specific, or the population that this sample comes from significantly differs from the reference.

- The sequence comes from a different organism (contamination, an infection, etc.).

To classify these read pairs, it is possible to use a BLAST-like approach to see if there is any significant enrichment for known pathogens. In practice, high throughput tools like Kraken (Wood and Salzberg, 2014) or Centrifuge (Kim, et al., 2016) are used for rapid classification of unmapped reads. Both tools use an index based strategy (in the case of Kraken based on *k*-mers, in the case of Centrifuge based on Burrows-Wheeler transform and the Ferragina-Manzini index).Both techniques have successfully been used for general metagenomics projects, but also for pre-filtering of WES and WGS data. This may be necessary when the sample was obtained via a buccal swab, especially when looking for *de novo* variants. Any reads that are classified as human can be assembled into a contig. This contig may be aligned again to the reference sequence, as this occasionally reveals a previously undetected structural variant, or it can be set aside for further analysis.

De novo assembly

There are several issues with assembling a genome through alignment to a reference. The reference genome will contain gaps, due to difficulties in assembling complex regions such

as repeat structures. In addition, homology between repeats means that unambiguous alignment will not always be possible. Routine sequence analysis will not necessarily identify on which allele a specific variant is located (Snyder, et al., 2015). Whether variants are in cis (on the same allele) or in trans (on different alleles on the two chromosomes) can have important clinical consequences. Two deleterious variants in a gene associated with a recessive condition may have no effect if they are in *cis*, but are likely to be disease-causing if in *trans* ("compound heterozygosity").

Most studies attempting to associate copy number variations of a specific locus use the sum of the different alleles, rather than measuring each allele separately (White, 2015). Although different haplotypes representing the range of copy numbers have been generated for several loci e.g. the amylase locus (Carpenter, et al., 2015), this information is not routinely available in typically used reference genomes. For WGS data, one option to obviate many of these problems is to perform a *de novo* assembly (Chaisson, et al., 2015b). Although this is more complex and computationally intensive, it will produce a more complete genome. If short read technology is used, a combination of paired end and mate pair libraries of different sizes is costly, but advantageous, and it will still not be sufficient to completely assemble a genome. A study used the PacBio system to sequence previously uncharacterised regions of the human genome (Chaisson, et al., 2015a). Mapping long reads to the ends of gaps and assembling from these points allowed the generation of >1 Mb of previously unmapped sequence.

Independent of what sequencing technology will predominate in the future, it can be anticipated that *de novo* assembly will eventually be the standard approach for genome sequencing.

Variant reporting

Ultimately, whether a mapping or a *de novo* assembly approach is used, the sequence will be compared to a reference to detect and call variants. The HGVS recommendations for the description of sequence variants are widely accepted standards for how each type of variant should be reported. It should be noted however that the standard output from NGS is not HGVS, but typically a VCF file, a semi-standard. The problem is that in VCF the same variant can be reported in different ways. When in HGVS a one nucleotide deletion is reported as g.12345678del, in VCF it may appear as position:12345677 refGC sampleG, or position:12345678 refCT sampleT. In addition, NGS software calls deletions on the 5' site of repeated sequences while the HGVS recommends the 3' rule. Needless to say this causes errors when tools or users annotate these variants and perform database searches using previous reports, and they do not realize that one variant may be reported using different formats. Although tools are available to cope with this problem (e.g. Mutalyzer (Wildeman, et al., 2008) and the Description Extractor (Vis, et al., 2015) they are not yet frequently used. Even when such tools do not, as stated, give correct HGVS descriptions the errors made are consistent and the same description is generated each time.

Each variant may be described in more than one manner e.g. with reference to a genomic position, as well as within a reference sequence for a specific gene. Although it may be possible to predict the effect of a coding variant on the mRNA or protein, or a non-coding variant on mRNA splicing, it should be stated whether this is a prediction or if there is supporting experimental evidence.

Suggestions for what is required in a clinical WGS report have been published (McLaughlin, et al., 2014). Especially when the purpose of the sequence analysis is to provide a genetic diagnosis, it is important to specify which regions have not been covered sufficiently to report any sequence variants (Brownstein, et al., 2014). This should include (i) regions known to not be covered using the technology implemented for sequencing e.g. repetitive regions, extreme GC percentages and (ii) regions that did not achieve sufficient coverage to allow variants to be called.

It may be that the WGS is being performed for non-clinical reasons e.g. out of general interest or to identify "nice-to-know" variants. In such a case it is essential that there are clear guidelines regarding what types of genes/variants will be analysed, as well as what counselling (if any) will be provided when reporting the findings.

Variant analysis of WGS data.

Even if the entire genome is sequenced, it may be that only a subset of loci is screened for variants. Largely due to biases introduced during the enrichment step, WGS provides more

consistent coverage of the exonic sequences than WES for the same depth of sequence (Meynert, et al., 2014). This can be compensated for by increasing the WES sequence depth, but this increases the cost of the assay further.

A study by Gilissen et al. used WGS to screen 50 intellectual disability (ID) cases where no causative variant had previously been identified with microarray CNV analysis and whole exome sequence (WES) analysis (Gilissen, et al., 2014). A comparison of variants identified by WGS and/or WES (only at loci covered by the exome capture kit) yielded >10x more variants identified by WGS only, when compared to those identified by WES only. For both approaches the differences were primarily due to differences in sequence coverage, with WGS giving a more uniform coverage. The discordance between WES and WGS was also described in Belkadi et al. (Belkadi, et al., 2015), who also showed that the false positive rate was higher in WES compared to WGS. Interestingly, the accuracy of indel calling was similar for both, demonstrating the inherent difficulty in calling indel variants.

A report by Sun et al., (Sun, et al., 2015) focussing only on 500 genes previously implicated in ID, found that WES detected all variants identified by WGS in nine samples, and that 99% of the 500 genes were covered to a sufficient depth. These findings demonstrate that the effective efficiency of WES vs WGS depends on the purpose of the analysis. However, as sequencing becomes cheaper it will eventually be more cost-effective, from a laboratory perspective, for WGS to be the default option for genetic analysis, irrespective of how much of the genome will subsequently be analysed.

A large study reported by Taylor et al used WGS to analyse 156 individual cases (Taylor, et al., 2015). The proportion of cases for which a causative variant could be identified varied according to condition, with the highest success rate (8/14) observed for trios (index and both parents). For Mendelian disorders a disease-causing variant was identified in 23/68 cases, with an overall study result of 33/156.

Assigning non-coding variants to a specific disease faces several challenges. In some cases, large *de novo* deletions or duplications upstream or downstream of a given gene can be linked to a condition with high certainty, e.g. *SOX9* with disorders of sex development, campomelic dysplasia, and / or craniofacial disorders (Gordon, et al., 2009; Kleinjan and van Heyningen, 2005). In most cases, however, it is not always immediately clear which gene is affected by a given non-coding variant.

The study of (Taylor et al. 2015) was only able to link non-coding variants to a condition in two cases. In one, a SNV in the 5' UTR of the *EPO* gene, identified in two unrelated families, was the only rare exonic variant in an identical by descent, 8Mb interval. The role of EPO in red blood cell development makes it a compelling candidate for the erythrocytosis seen in these families.

The other was a complex rearrangement near the *SOX3* gene, consisting of a deletion combined with an insertion of part of chromosome 2, identified in a case with X-linked hypoparathyroidism. *SOX3* is known to be involved in parathyroid development, providing a link between the affected locus and the condition. This example also highlights a strength of WGS, as the insertion would not have been detected with exome or microarray analysis.

Several consortia, such as the Encyclopedia of DNA Elements (ENCODE) project (Consortium, 2012), and Roadmap Epigenomics Mapping Consortium (Roadmap Epigenomics, et al., 2015), have generated genomewide data sets from multiple cell types that allow the identification of genomic regions that have regulatory potential. Even with this information, it is still challenging to predict the effect of a SNV. Different approaches have been described for using these data to predict functional effects (Boyle, et al., 2012; Lee, et al., 2015), however in the vast majority of cases it is necessary to combine these predictions with other biological information to make a link with disease.

Conclusions

Although routine WGS, as a laboratory activity, is rapidly becoming feasible from a financial viewpoint, there is currently no single sequencing technology able to generate complete, fully haplotype resolved, human genomes. The most accurate genome assemblies to date have used a combination of short and long range sequencers, along with other genomic assembly technologies such as optical mapping (Hastie, et al., 2013). Initiatives such as Genome In A Bottle are assembling high quality, haplotype aware, diploid reference genomes. These genomes can serve as superior references for mapping studies, and the approaches used can be applied more broadly for further genome-based studies. Ultimately, however, it will require the development of a technology that can sequence single DNA molecules, hundreds of kilobases in length, before complete genomes can routinely be generated, as well as considerable efforts to achieve the best benefit for patients regarding

their informed choices and relevant medical service. As a common endeavour, taking the time to share data, clinical findings and experiences with the implementation and use of different policies for these issues, whenever it is possible and appropriate to do so, is necessary. Cooperation between both doctor- patient and between research teams will be the key to achieving the successful use of this powerful tool, which offers many more challenges than the technical ones we have outlined here.

Acknowledgements

We thank all the people that contributed (Speakers/Participants) in the many 3Gb-TEST organised meetings/events listed at the consortium website: 3Gb-TEST.eu/meetings/.

Disclosure of Conflicts of Interest

Stefan Van Vooren is employed by Agilent Technologies. Bart Janssen is employed by GenomeScan B.V.

References

Adey A, Burton JN, Kitzman JO, Hiatt JB, Lewis AP, Martin BK, Qiu R, Lee C, Shendure J. 2013. The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. Nature 500:207-11.

Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A. 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol 12:R18.

Alkan C, Coe BP, Eichler EE. 2011. Genome structural variation discovery and genotyping. Nat Rev Genet 12:363-76.

American Type Culture Collection Standards Development Organization Workgroup ASN. 2010. Cell line misidentification: the beginning of the end. Nat Rev Cancer 10:441-8.

Belkadi A, Bolze A, Itan Y, Cobat A, Vincent QB, Antipenko A, Shang L, Boisson B, Casanova JL, Abel L. 2015. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. Proc Natl Acad Sci U S A 112:5473-8.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53-9.

Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, Karczewski KJ, Park J, Hitz BC, Weng S, Cherry JM, Snyder M. 2012. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res 22:1790-7.

Brady P, Brison N, Van Den Bogaert K, de Ravel T, Peeters H, Van Esch H, Devriendt K, Legius E, Vermeesch JR. 2015. Clinical implementation of NIPT - technical and biological challenges. Clin Genet.

Brownstein CA, Beggs AH, Homer N, Merriman B, Yu TW, Flannery KC, DeChene ET, Towne MC, Savage SK, Price EN, Holm IA, Luquette LJ et al. 2014. An international effort towards developing standards for best practices in analysis, interpretation and reporting of clinical genome sequencing results in the CLARITY Challenge. Genome Biol 15:R53.

Cai X, Evrony GD, Lehmann HS, Elhosary PC, Mehta BK, Poduri A, Walsh CA. 2015. Single-cell, genome-wide sequencing identifies clonal somatic copy-number variation in the human brain. Cell Rep 10:645.

Cantsilieris S, White SJ. 2012. Correlating multiallelic copy number polymorphisms with disease susceptibility. Hum Mutat 34:1-13.

Capes-Davis A, Theodosopoulos G, Atkin I, Drexler HG, Kohara A, MacLeod RA, Masters JR, Nakamura Y, Reid YA, Reddel RR, Freshney RI. 2010. Check your cultures! A list of cross-contaminated or misidentified cell lines. Int J Cancer 127:1-8.

Carpenter D, Dhar S, Mitchell LM, Fu B, Tyson J, Shwan NA, Yang F, Thomas MG, Armour JA. 2015. Obesity, starch digestion and amylase: association between copy number variants at human salivary (AMY1) and pancreatic (AMY2) amylase genes. Hum Mol Genet 24:3472-80.

Chaisson MJ, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, Antonacci F, Surti U, Sandstrom R, Boitano M, Landolin JM, Stamatoyannopoulos JA et al. 2015a. Resolving the complexity of the human genome using single-molecule sequencing. Nature 517:608-11.

Chaisson MJ, Wilson RK, Eichler EE. 2015b. Genetic variation and the de novo assembly of human genomes. Nat Rev Genet 16:627-40.

Chatterjee R. 2007. Cell biology. Cases of mistaken identity. Science 315:928-31.

Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, Shi X, Fulton RS et al. 2009. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods 6:677-81.

Chiu RW, Chan KC, Gao Y, Lau VY, Zheng W, Leung TY, Foo CH, Xie B, Tsui NB, Lun FM, Zee BC, Lau TK et al. 2008. Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma. Proc Natl Acad Sci U S A 105:20458-63.

Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, Bayley H. 2009. Continuous base identification for single-molecule nanopore DNA sequencing. Nat Nanotechnol 4:265-70.

Consortium EP. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489:57-74.

Cornish A, Guda C. 2015. A Comparison of Variant Calling Pipelines Using Genome in a Bottle as a Reference. Biomed Res Int 2015:456479.

Dabney J, Meyer M. 2012. Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. Biotechniques 52:87-94.

Dey SS, Kester L, Spanjaard B, Bienko M, van Oudenaarden A. 2015. Integrated genome and transcriptome sequencing of the same cell. Nat Biotechnol 33:285-9.

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K et al. 2009. Real-time DNA sequencing from single polymerase molecules. Science 323:133-8.

Evrony GD, Lee E, Mehta BK, Benjamini Y, Johnson RM, Cai X, Yang L, Haseley P, Lehmann HS, Park PJ, Walsh CA. 2015. Cell lineage analysis in human brain using endogenous retroelements. Neuron 85:49-59.

Fan HC, Blumenfeld YJ, Chitkara U, Hudgins L, Quake SR. 2008. Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. Proc Natl Acad Sci U S A 105:16266-71.

Freeman B, Powell J, Ball D, Hill L, Craig I, Plomin R. 1997. DNA by mail: an inexpensive and noninvasive method for collecting DNA samples from widely dispersed populations. Behav Genet 27:251-7.

Fromer M, Moran JL, Chambert K, Banks E, Bergen SE, Ruderfer DM, Handsaker RE, McCarroll SA, O'Donovan MC, Owen MJ, Kirov G, Sullivan PF et al. 2012. Discovery

and statistical genotyping of copy-number variation from whole-exome sequencing depth. Am J Hum Genet 91:597-607.

Fungtammasan A, Ananda G, Hile SE, Su MS, Sun C, Harris R, Medvedev P, Eckert K, Makova KD. 2015. Accurate typing of short tandem repeats from genome-wide sequencing data and its applications. Genome Res 25:736-49.

Gilissen C, Hehir-Kwa JY, Thung DT, van de Vorst M, van Bon BW, Willemsen MH, Kwint M, Janssen IM, Hoischen A, Schenck A, Leach R, Klein R et al. 2014. Genome sequencing identifies major causes of severe intellectual disability. Nature 511:344-7.

Gordon CT, Tan TY, Benko S, Fitzpatrick D, Lyonnet S, Farlie PG. 2009. Long-range regulation at the SOX9 locus in development and disease. J Med Genet 46:649-56.

Green RC, Berg JS, Grody WW, Kalia SS, Korf BR, Martin CL, McGuire AL, Nussbaum RL, O'Daniel JM, Ormond KE, Rehm HL, Watson MS et al. 2013. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. Genet Med 15:565-74.

Handsaker RE, Korn JM, Nemesh J, McCarroll SA. 2011. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. Nat Genet 43:269-76.

Handsaker RE, Van Doren V, Berman JR, Genovese G, Kashin S, Boettger LM, McCarroll SA. 2015. Large multiallelic copy number variations in humans. Nat Genet 47:296-303.

Hastie AR, Dong L, Smith A, Finklestein J, Lam ET, Huo N, Cao H, Kwok PY, Deal KR, Dvorak J, Luo MC, Gu Y et al. 2013. Rapid genome mapping in nanochannel arrays for highly

complete and accurate de novo sequence assembly of the complex Aegilops tauschii genome. PLoS One 8:e55864.

Hehir-Kwa JY, Claustres M, Hastings RJ, van Ravenswaaij-Arts C, Christenhusz G, Genuardi M, Melegh B, Cambon-Thomsen A, Patsalis P, Vermeesch J, Cornel MC, Searle B et al. 2015. Towards a European consensus for reporting incidental findings during clinical NGS testing. Eur J Hum Genet 23:1601-6.

Julia S, Bertier G, Cambon-Thomsen A. 2016. Quand l'anticipation devient plurielle : la complexité des données génomiques à l'épreuve des pratiques professionnelles. Revue Française d'Éthique Appliquée 2:19-28.

Kim D, Song L, Breitwieser FP, Salzberg SL. 2016. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res 26:1721-1729.

Kirby A, Gnirke A, Jaffe DB, Baresova V, Pochet N, Blumenstiel B, Ye C, Aird D, Stevens C, Robinson JT, Cabili MN, Gat-Viks I et al. 2013. Mutations causing medullary cystic kidney disease type 1 lie in a large VNTR in MUC1 missed by massively parallel sequencing. Nat Genet 45:299-303.

Kitzman JO, Snyder MW, Ventura M, Lewis AP, Qiu R, Simmons LE, Gammill HS, Rubens CE, Santillan DA, Murray JC, Tabor HK, Bamshad MJ et al. 2012. Noninvasive whole-genome sequencing of a human fetus. Sci Transl Med 4:137ra76.

Klambauer G, Schwarzbauer K, Mayr A, Clevert DA, Mitterecker A, Bodenhofer U, Hochreiter S. 2012. cn.MOPS: mixture of Poissons for discovering copy number variations in

next-generation sequencing data with a low false discovery rate. Nucleic Acids Res 40:e69.

Kleinjan DA, van Heyningen V. 2005. Long-range control of gene expression: emerging mechanisms and disruption in disease. Am J Hum Genet 76:8-32.

Knouse KA, Wu J, Amon A. 2016. Assessment of megabase-scale somatic copy number variation using single-cell sequencing. Genome Res 26:376-84.

Koren A, Handsaker RE, Kamitaki N, Karlic R, Ghosh S, Polak P, Eggan K, McCarroll SA. 2014. Genetic variation in human DNA replication timing. Cell 159:1015-26.

Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ. 2009. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. Nat Methods 6:291-5.

Lan JH, Yin Y, Reed EF, Moua K, Thomas K, Zhang Q. 2015. Impact of three Illumina library construction methods on GC bias and HLA genotype calling. Hum Immunol 76:166-75.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357-9.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25.

Lee D, Gorkin DU, Baker M, Strober BJ, Asoni AL, McCallion AS, Beer MA. 2015. A method to predict the impact of regulatory variants from DNA sequence. Nat Genet 47:955-61.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754-60.

Liu X, Han S, Wang Z, Gelernter J, Yang BZ. 2013. Variant callers for next-generation sequencing data: a comparison study. PLoS One 8:e75619.

Loman NJ, Quick J, Simpson JT. 2015. A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat Methods 12:733-5.

Mahfuz I, Cheng W, White SJ. 2013. Identification of Streptococcus parasanguinis DNA contamination in human buccal DNA samples. BMC Res Notes 6:481.

Maitra A, Arking DE, Shivapurkar N, Ikeda M, Stastny V, Kassauei K, Sui G, Cutler DJ, Liu Y, Brimble SN, Noaksson K, Hyllner J et al. 2005. Genomic alterations in cultured human embryonic stem cells. Nat Genet 37:1099-103.

Marine R, Polson SW, Ravel J, Hatfull G, Russell D, Sullivan M, Syed F, Dumas M, Wommack KE. 2011. Evaluation of a transposase protocol for rapid generation of shotgun high-throughput sequencing libraries from nanogram quantities of DNA. Appl Environ Microbiol 77:8071-9.

Marusyk A, Polyak K. 2010. Tumor heterogeneity: causes and consequences. Biochim Biophys Acta 1805:105-17.

Matthijs G, Souche E, Alders M, Corveleyn A, Eck S, Feenstra I, Race V, Sistermans E, Sturm M, Weiss M, Yntema H, Bakker E et al. 2015. Guidelines for diagnostic next-generation sequencing. Eur J Hum Genet.

McLaughlin HM, Ceyhan-Birsoy O, Christensen KD, Kohane IS, Krier J, Lane WJ, Lautenbach D, Lebo MS, Machini K, MacRae CA, Azzariti DR, Murray MF et al. 2014. A systematic approach to the reporting of medically relevant findings from whole genome sequencing. BMC Med Genet 15:134.

Meynert AM, Ansari M, FitzPatrick DR, Taylor MS. 2014. Variant detection sensitivity and biases in whole genome and exome sequencing. BMC Bioinformatics 15:247.

Miyatake S, Koshimizu E, Fujita A, Fukai R, Imagawa E, Ohba C, Kuki I, Nukui M, Araki A, Makita Y, Ogata T, Nakashima M et al. 2015. Detecting copy-number variations in whole-exome sequencing data using the eXome Hidden Markov Model: an 'exome-first' approach. J Hum Genet 60:175-82.

Narva E, Autio R, Rahkonen N, Kong L, Harrison N, Kitsberg D, Borghese L, Itskovitz-Eldor J, Rasool O, Dvorak P, Hovatta O, Otonkoski T et al. 2010. High-resolution DNA analysis of human embryonic stem cell lines reveals culture-induced copy number changes and loss of heterozygosity. Nat Biotechnol 28:371-7.

Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, Cook K, Stepansky A, Levy D, Esposito D, Muthuswamy L, Krasnitz A et al. 2011. Tumour evolution inferred by single-cell sequencing. Nature 472:90-4.

Nielsen R, Paul JS, Albrechtsen A, Song YS. 2011. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet 12:443-51.

Obradovic J, Jurisic V, Tosic N, Mrdjanovic J, Perin B, Pavlovic S, Djordjevic N. 2013. Optimization of PCR conditions for amplification of GC-Rich EGFR promoter sequence. J Clin Lab Anal 27:487-93.

Press MO, Carlson KD, Queitsch C. 2014. The overdue promise of short tandem repeat variation for heritability. Trends Genet 30:504-12.

Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, Bore JA, Koundouno R, Dudas G, Mikhail A, Ouedraogo N, Afrough B et al. 2016. Real-time, portable genome sequencing for Ebola surveillance. Nature 530:228-32.

Rehm HL, Bale SJ, Bayrak-Toydemir P, Berg JS, Brown KK, Deignan JL, Friez MJ, Funke BH, Hegde MR, Lyon E, Working Group of the American College of Medical G, Genomics Laboratory Quality Assurance C. 2013. ACMG clinical laboratory standards for next-generation sequencing. Genet Med 15:733-47.

Reuter JA, Spacek DV, Snyder MP. 2015. High-throughput sequencing technologies. Mol Cell 58:586-97.

Roadmap Epigenomics C, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, Amin V et al. 2015. Integrative analysis of 111 reference human epigenomes. Nature 518:317-30.

Snyder MW, Adey A, Kitzman JO, Shendure J. 2015. Haplotype-resolved genome sequencing: experimental methods and applications. Nat Rev Genet 16:344-58.

Straver R, Sistermans EA, Holstege H, Visser A, Oudejans CB, Reinders MJ. 2014. WISECONDOR: detection of fetal aberrations from shallow sequencing maternal plasma based on a within-sample comparison scheme. Nucleic Acids Res 42:e31.

Sun Y, Ruivenkamp CA, Hoffer MJ, Vrijenhoek T, Kriek M, van Asperen CJ, den Dunnen JT, Santen GW. 2015. Next-generation diagnostics: gene panel, exome, or whole genome? Hum Mutat 36:648-55.

Taylor JC, Martin HC, Lise S, Broxholme J, Cazier JB, Rimmer A, Kanapin A, Lunter G, Fiddy S, Allan C, Aricescu AR, Attar M et al. 2015. Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. Nat Genet 47:717-26.

van Dijk EL, Jaszczyszyn Y, Thermes C. 2014. Library preparation methods for next-generation sequencing: tone down the bias. Exp Cell Res 322:12-20.

Van Tol HH, Wu CM, Guan HC, Ohara K, Bunzow JR, Civelli O, Kennedy J, Seeman P, Niznik HB, Jovanovic V. 1992. Multiple dopamine D4 receptor variants in the human population. Nature 358:149-52.

Vis JK, Vermaat M, Taschner PE, Kok JN, Laros JF. 2015. An efficient algorithm for the extraction of HGVS variant descriptions from sequences. Bioinformatics 31:3751-7.

Wang J, Raskin L, Samuels DC, Shyr Y, Guo Y. 2015. Genome measures used for quality control are dependent on gene function and ancestry. Bioinformatics 31:318-23.

Wang Y, Navin NE. 2015. Advances and applications of single-cell sequencing technologies. Mol Cell 58:598-609.

Weiss MM, Van der Zwaag B, Jongbloed JD, Vogel MJ, Bruggenwirth HT, Lekanne Deprez RH, Mook O, Ruivenkamp CA, van Slegtenhorst MA, van den Wijngaard A, Waisfisz Q, Nelen MR et al. 2013. Best practice guidelines for the use of next-generation sequencing applications in genome diagnostics: a national collaborative study of Dutch genome diagnostic laboratories. Hum Mutat 34:1313-21.

White S. 2015. Counting copy number and calories. Nat Genet 47:852-3.

White S, Kalf M, Liu Q, Villerius M, Engelsma D, Kriek M, Vollebregt E, Bakker B, van Ommen GJ, Breuning MH, den Dunnen JT. 2002. Comprehensive detection of genomic duplications and deletions in the DMD gene, by use of multiplex amplifiable probe hybridization. Am J Hum Genet 71:365-74.

Whitney AR, Diehn M, Popper SJ, Alizadeh AA, Boldrick JC, Relman DA, Brown PO. 2003. Individuality and variation in gene expression patterns in human blood. Proc Natl Acad Sci U S A 100:1896-901.

Wildeman M, van Ophuizen E, den Dunnen JT, Taschner PE. 2008. Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. Hum Mutat 29:6-13.

Wood DE, Salzberg SL. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 15:R46.

Wu J, Grzeda KR, Stewart C, Grubert F, Urban AE, Snyder MP, Marth GT. 2012. Copy Number Variation detection from 1000 Genomes Project exon capture sequencing data. BMC Bioinformatics 13:305.

Ye H, Meehan J, Tong W, Hong H. 2015. Alignment of Short Reads: A Crucial Step for Application of Next-Generation Sequencing Data in Precision Medicine. Pharmaceutics 7:523-41.

Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25:2865-71.

Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, Weng Z, Liu Y, Mason CE, Alexander N, Henaff E, McIntyre AB et al. 2016. Extensive sequencing of seven human genomes to characterize benchmark reference materials. Sci Data 3:160025.

Table 1. Critical Differences between WES and WGS

	WES	WGS
Sample source	Enrichment step likely to reduce impact of contaminating DNA	All DNA in sample can be sequenced, appropriate care should be taken
Library prep	Enrichment step will introduce biases	No enrichment leads to more even coverage
Platform	Short reads sufficient	Longer reads preferable
Variant detection	Amplification/enrichment step will lead to unequal genomic representation; CNV calling relies on read depth	More even genomic coverage; structural variation analysis may allow breakpoint detection
Sequence alignment	Compare to reference genome	Can perform de novo assembly; allows for more accurate genome construction
	Focus on reads aligning to exome	Whole genome analysis, increased complexity due to repeats etc

Variant reporting	Focus on coding/splice site variation	More complex variation, genome-wide

Table 2. A summary of points to consider when performing whole genome sequencing

Aspect	**Points**
DNA Sample	*Source* • Blood - circulating DNA from e.g. tumour or fetus (pregnant female) can be isolated along with cellular DNA • Buccal swab - can be contaminated with non-human material • Cell line - Genetic variation introduced during culturing • Mosaicism - lymphocytes undergo rearrangements; neurons affected by LINE1 retrotranspositions; new variants will be introduced with each cell division due to replication errors • Cell cycle - Late replicating DNA will be under-represented compared to early replicating DNA, can influence copy number analysis *Isolation* • DNA quality / integrity - Isolation with e.g. phenol can influence downstream enzymatic reactions • Severely fragmented DNA is unsuitable for long read sequencing; sites of degradation will be non-random, significantly affecting coverage *Sample preparation*

	• Enzymatic or mechanical fragmentation will be non-random and introduce different biases • To reduce biases, especially GC% extremes, PCR amplification during sample preparation should be minimized (ideally avoided)
Sequencing	*Type of sequencer* • Short read sequencers generate many reads, but the read length can make it difficult to unambiguously align reads, especially in repetitive sequences • Long read sequencers e.g. PacBio, generate relatively fewer reads, but the longer reads are easier to align and assemble
Alignment/Assembly	*Quality control* • Important to check for duplicate reads, coverage vs. GC%, completeness of coverage etc. *Alignment* • Choice of aligner, as well as specific conditions chosen during alignment, will impact final results • Reference genome (build, presence / absence of different haplotypes) will affect the completeness and accuracy of alignment. • Some loci cannot be unambiguously mapped, especially with short read sequencing

	• Unmapped reads can be due to genetic variation incompatible with standard genomic alignment, e.g. certain types of structural variation • Aligning to a reference will never lead to a complete genome *De novo assembly* • Is more accurate, but also more time consuming and complex • With current sequencing technology it is not possible to completely assemble a human genome
Variant Detection/Reporting	*Variant calling* • No single algorithm can reliably detect all variant types, e.g. multiallelic copy number variation; specific types require a focused analysis approach • Sequencing platform used and sample preparation method will determine which types of variation can/can not be reliably called *Variant reporting* • Variants should be reported using standardised nomenclature • Unless experimentally confirmed, it should be clear that changes at RNA or protein level are predictions only • VCF files can report the same variant in different ways, and care

	should be taken when comparing files from different sources and/or when querying databases