phgfoundation
making science work for health

January 2014

Authors
Dr Leila Luheshi
Dr Sobia Raza

Contact
Dr Leila Luheshi
leila.luheshi@phgfoundation.org

Clinical whole genome analysis: delivering the right diagnosis

Whole genome analysis has the potential to transform genetic diagnostic services for families with rare inherited diseases. If this goal is to be achieved, it is vital that healthcare practitioners, clinical scientists and commissioners work together to identify and overcome the challenges to the delivery of diagnostic quality, clinically actionable genetic test results from genome scale data.

Whole genome analysis (WGA) presents an unprecedented opportunity to diagnose rare genetic diseases, and may be a step forward for the thousands of patients for whom existing genetic testing has not provided a diagnosis. WGA may also reduce the lengthy quest for a diagnosis for many rare disease patients by circumventing the need for multiple expensive and often invasive tests and in the process offer cost efficiencies. Whilst the UK National Health Service (NHS) is not yet ready to use whole genome data in regular clinical practice, this situation could be changed rapidly through the co-ordinated efforts of all stakeholders. Implementation will require a methodical approach to achieve both reliable and measurable quality of sequence data, and the significant improvements of the evidence base on which analysis and interpretation depends.

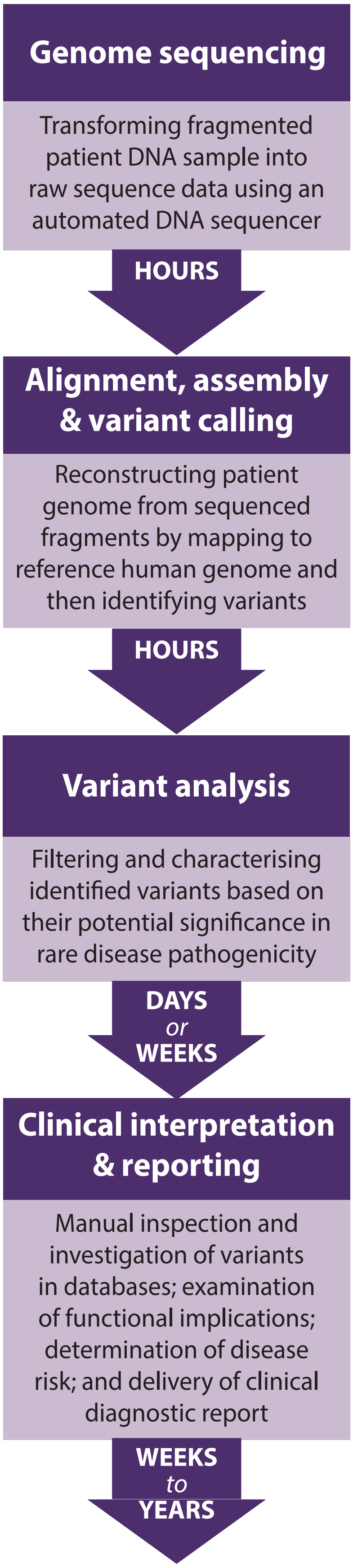
This briefing note summarises the procedures involved in converting genome scale DNA sequence data into a clinically actionable diagnostic report and highlights the critical steps where further development and standardisation are required before WGA can enter into routine practice in a clinical diagnostic setting.

Diagnostic whole genome analysis - what is involved?

Next generation sequencing technology is routinely used to read an entire human genome at a cost of a few thousand pounds in as little as 24 hours. This is achieved by breaking the 3 billion base pairs of the human genome into millions of tiny fragments, which can be repeatedly sequenced in parallel. The accurate *in silico* assembly of a complete genome from these millions of fragmented DNA sequences, and the eventual detection (with high levels of sensitivity and specificity) of clinically relevant variants, are hampered by the potential for error, variation and uncertainty at every stage.

Proportion of disease-associated mutations a recent study found to be incorrectly described in the published literature[1]

43%
mismatch
between variants identified from a single sample processed by five common alignment and variant calling pipelines[2]



Genome sequencing

Transforming fragmented patient DNA sample into raw sequence data using an automated DNA sequencer

HOURS

Alignment, assembly & variant calling

Reconstructing patient genome from sequenced fragments by mapping to reference human genome and then identifying variants

HOURS

Variant analysis

Filtering and characterising identified variants based on their potential significance in rare disease pathogenicity

DAYS *or* **WEEKS**

Clinical interpretation & reporting

Manual inspection and investigation of variants in databases; examination of functional implications; determination of disease risk; and delivery of clinical diagnostic report

WEEKS *to* **YEARS**

Critical steps in whole genome analysis:

1. Generating enough high quality raw DNA sequence

Analytical sensitivity and specificity depends on the quality and quantity of raw DNA sequence fragments produced, from which each patient's genome is then assembled. If quality or quantity is insufficient (*e.g.* due to sequencing errors, gaps in coverage, or low sequencing depth), potentially significant variants will go undetected (low sensitivity) and 'errors' could be misinterpreted as significant (low specificity). Current genetic diagnostic tests aim for 99% analytic sensitivity and specificity, and WGA cannot at present meet this standard as it does not yet detect all classes of disease causing variants.

2. Accurately assembling the patient's genome and identifying variants

The process of assembling (mapping) a genome sequence from millions of short DNA fragments is variable and error prone, depending on the performance of the statistical algorithms used for this task and the quality of the reference genome maps used to guide assembly. Some types of disease-causing genomic variation such as insertions, deletions and inversions are particularly hard to deal with during genome assembly due to the challenges of placing them correctly on the reference genome map (*i.e.* they cannot be uniquely positioned).

3. Filtering and prioritising potentially pathogenic variants

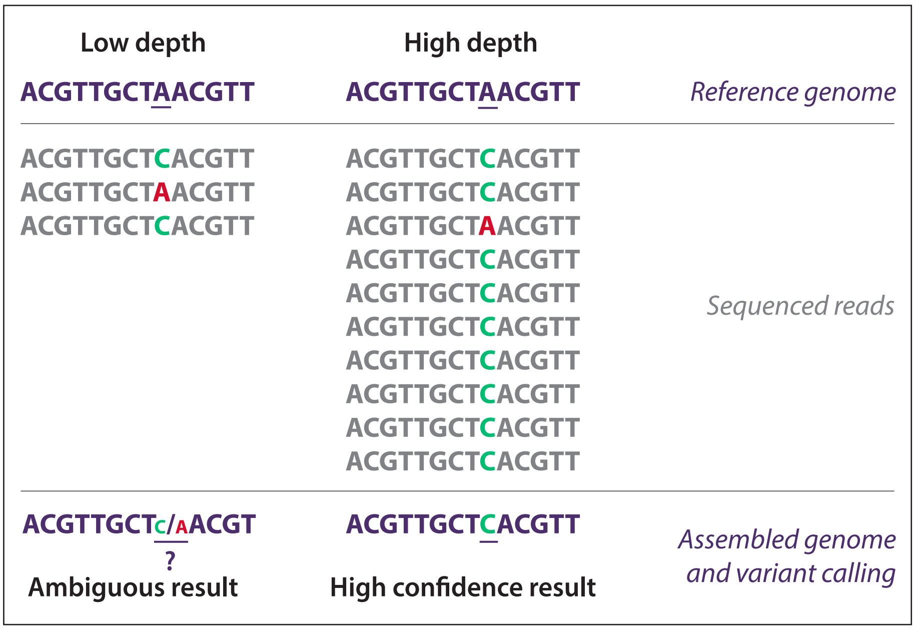
The average human genome has 3-4 million variants. Deciphering which are harmless and which could cause disease requires extensive filtering to create a priority list of the variants most likely to be pathogenic in that individual. This is mostly done automatically by bespoke computer programmes using existing databases of genomic data to filter out variants that are common (hence unlikely to explain a rare disease) and those unlikely to have effects relevant to the clinical picture of the patient. However, the published data (genomic and functional) on which the databases and algorithms used in these steps depends is known to be unreliable and incomplete. Scientific rigour is therefore required to ensure non-functional or normal variants are not interpreted as being causal, and results can vary based on data sources used.

4. Interpreting, validating and reporting clinically significant variants

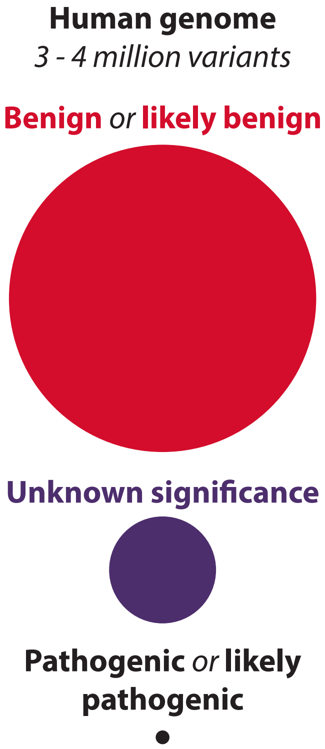
Given the numerous potential sources of error and uncertainty in the automated analyses, manual curation of the final candidate list of potentially pathogenic variants is essential to produce a reliable clinical diagnostic report. Most of the time and expense in genome analysis accrues at this stage, where evidence to support pathogenicity of each candidate variant is carefully evaluated to exclude false positives based on the patient's detailed clinical features. Experienced clinicians and scientists must use judgement and knowledge to decide whether there is sufficient high quality evidence to support the definitive identification of a genomic variant as pathogenic. For any potential pathogenic variants they will also have to validate the accuracy of the original variant call by resequencing the relevant genetic region using Sanger sequencing (the current gold standard). In some situations, cascade testing, segregation analysis and functional studies may also be required to clarify the potential pathogenic credentials of variants.

Depth of coverage in genome sequencing affects result reliability

Genomic fragments are sequenced multiple times inside a modern sequencing machine to provide overlapping sequence 'reads'. A greater number of reads at a given point (vertical coverage) and across the entire genome (horizontal coverage) provides greater statistical confidence that a given DNA base in a sequence has been inferred correctly, and can therefore reduce the time and substantial cost needed to perform secondary validations in downstream variant identification.



Low depth	High depth	
ACGTTGCTAACGTT	**ACGTTGCTAACGTT**	*Reference genome*
ACGTTGCTCACGTT ACGTTGCTAACGTT ACGTTGCTCACGTT	ACGTTGCTCACGTT ACGTTGCTCACGTT ACGTTGCTAACGTT ACGTTGCTCACGTT ACGTTGCTCACGTT ACGTTGCTCACGTT ACGTTGCTCACGTT ACGTTGCTCACGTT ACGTTGCTCACGTT ACGTTGCTCACGTT	*Sequenced reads*
ACGTTGCTC/AACGT ? **Ambiguous result**	**ACGTTGCTCACGTT** **High confidence result**	*Assembled genome and variant calling*



Recognising the risks to patients and the NHS

Examining the steps in whole genome analysis reveals that the fundamental challenges to its use in a clinical diagnostic setting are in ensuring raw genome data themselves are of sufficient quality and completeness, and that the validity and accuracy of clinical interpretation (step 4, left) is not undermined by an unreliable evidence base (steps 1-3, left).

It is our view that using whole genome data in clinical diagnostic services within the NHS without first addressing these fundamental issues of diagnostic quality poses potentially unacceptable risks to patient safety, and quality of care. These risks include:

- **Incorrect diagnosis (false positive or negative), leading to inappropriate patient care and decision making and threatening patient safety.**
- **Failure to provide a conclusive diagnosis for the patient and a continuation of their diagnostic odyssey.**
- **Inappropriate use of NHS resources.**

Impact of switching from unmatched to ethnically matched reference genome used for sequencing[3]

Taking steps towards delivering the correct diagnosis

None of the challenges to accurate whole genome sequencing are insurmountable, but overcoming these will take time, resources and a methodical approach that should include the following:

1. **Determining minimum standards for sequence generation**

   The question of how much high quality DNA sequence must be generated to achieve the required diagnostic sensitivity for whole genome analysis must be established empirically based on the sensitivity for detecting different types of genetic variants (step 1).

2. **Establishing best practices in bioinformatics**

   There are no established 'best practice' protocols for the computational analysis of genomes (steps 2-4) and it currently relies largely on unregulated academic software, bespoke pipelines and incomplete databases. Necessarily these will vary to some extent depending on the disease being analysed, but minimum standards and mechanisms to ensure consistency and accuracy will be required.

3. **Improving the quality of the evidence base**

   Consideration must be given to improving the quality and completeness of the databases and other scientific evidence on which whole genome analysis depends (step 4). This will depend on establishing appropriate standards and measures, and significantly improving the sharing of genomic and clinical data both nationally and internationally.

4. **Building infrastructure**

   Dramatically improved computing infrastructure is required within or easily accessible to the NHS to enable clinical scientists to undertake genome scale variant analysis, clinical interpretation and to share findings and data.

Although we focus on the diagnosis of rare inherited diseases, the obstacles and targets outlined above are very similar if not identical to those in other clinical applications of WGA. The challenges surrounding each analysis step will be reviewed in further detail as part of the PHG Foundation's work programme on clinical whole genome analysis, and recommendations made to address these.

Glossary

Genome
Complete genetic information encoded by a person's DNA

Reference genome map
This is a single, internationally agreed, composite human genome sequence against which all others can be compared. The reference genome is used as a scaffold to guide assembly of individual patient genomes and then identify how the patient and reference genomes vary. Reference genomes are incomplete and require periodic updating and improving.

Sequencing depth of coverage
A measure of the number of times the same DNA target is sequenced by the technology; increasing the depth can substantially reduce the error rate.

Variant
Either a point or region in a sequenced genome that varies when compared to a reference genome map. Variants can be single DNA point (base) changes or larger deviations such as insertions or deletions of multiple adjoining bases. Every human has millions of variants in their genome, ranging from those that are common in the population to those that are very rare.

References

1. Bell CJ *et al*. Carrier testing for severe childhood recessive diseases by next-generation sequencing. Sci Transl Med 3, 65ra4 (2011).
2. O'Rawe J *et al*. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Med 5, 28 (2013).
3. Dewey FE *et al*. Phased whole-genome genetic risk in a family quartet using a major allele reference sequence. PLoS Genet 7, e1002280 (2011).

For more information about the PHG Foundation, visit:

www.phgfoundation.org