1-sequencing

This process extract variants in the sample. We'll focus on illumina sequencer, it's short read type can handle double stranded DNA of length 200–600 base pairs (bp) and 0.1% error rate. It use synthesis sequencing:

DNA fragments that are loaded onto a glass plate in the instrument. Fragments bind and form spots on the plate. During sequencing, a molecule complementary to the bound fragment is constructed with one nucleotide with a fuorescent tag being added in each sequencing cycle. There are different fuorescent dyes for each of the A, C, G, T nucleotides. Photographs are taken after each cycle and the intensity of each of the four possible fuorescences at the spot is obtained from the images. The sequence of (color, intensity) values detected at a certain spot is used to infer the sequence of bases that were added and thereby decipher the overall sequence of the fragment at the spot. This process of base-calling is complex, since there can be interference in the signal from neighboring spots and because of errors that build up as the sequencing process we use basecallers .

The base quality is expressed as a Phred score and corresponds to (−10log10(p)) where p is the probability that the base in the read is an incorrect call ,it's usually in the 0–40 range and scores above 30 are considered good. Many basecallers use a model-based approach where the biases in the platform are modeled. However, some basecallers use a machine learning approach where a model is trained for each cycle based on a control sample with a known reference genome which is (ideally) run along with all the other samples.

2-Alignment

The first step in this process is to map each read to its possible location on the reference genome with algorithms. Complexities happen because of differences between the read and the reference (differences between the sample and the reference genome, errors introduced in the sequencing process, repetitive or highly similar reigons)

The confidence is expressed as a mapping quality which is typically in the 0–7 range.

3-Detection of Substitution and Indel Variants

Base Quality Score Recalibration (using a Phred score that ranges from 0 to 40. If the observed error rates among Phred 30 quality bases are far lower than 1/1000, then it is possible that the base qualities have been underestimated by the basecaller.so BQSR offers an opportunity to modify the base qualities.)

Local realignment (attempt to correct these issues in two steps: - Regions with a high concentration of reads with insertions and deletions are identified by scanning across the alignments. - For each candidate region, haplotypes (short genomic sequences) that best explain the observed read sequences are generated and the reads are realigned against these haplotypes.)

Bayesian Variant Calling (to filter artifacts variant out, useful biases are: - Strand bias: computes if the variant is supported by reads in both directions. Variants with high strand bias are typically the result of sequencing errors. - Base quality bias: computes if the variant bases are of significantly lower quality than wild-type bases. Variants with high base quality bias are typically the result of sequencing errors. -Tail distance bias: computes if the variant bases come disproportionately from the ends of the reads. Variants with high tail distance bias may be false positives because of sequencing errors, unclipped alignments, incorrect realignment, etc. - Mapping quality bias: computes if the reads with the variant bases have significantly lower mapping quality when compared to the reads with wild-type bases. Variants with high mapping quality can be the result of alignment artifacts.)

Variant Filtering (There are two approaches to eliminating the false positives variant calls: -Hard filter approach: in this approach, thresholds for the various metrics mentioned above are fine-tuned by hand and variants satisfying the threshold conditions for all the metrics are considered true variants. -Machine learning approach: in this approach, models are trained with examples of true and false positive variants and their attributes. The trained model is used to either make a binary call on each variant or to assign the variant an overall quality score.)

conclusion

In this article, we've provided an summary of the data flow that takes place during rare disease diagnosis. Our rare disease diagnostics test, based on the 5000 gene Illumina TruSight One panel, shows a diagnostic yield of~40d which is on par with other groups using WES. However, considering the rapid rate65 at which new gene–disease associations are discovered, it's possible that diagnostics laboratories like ours will eventually move from smaller targeted panels to WES to avoid the prospect of missing important genes.

The case for routine WGS within the clinical context is a smaller amount clear. during a recent meta-analysis, the diagnostic utility of WES didn't differ significantly from that of WGS66. However, if one were to put aside the storage and computational costs, WGS may eventually become the quality practice. WGS data are more uniform and permit for far more sensitive detection of CNVs, SVs, and repeat expansions. Some early adopters of clinical WGS have had promising results67, 68. WGS by itself isn't a remedy , since most knowledge about rare diseases remains targeting the exome which represents less than 2% of the whole genome. Variant prioritization will become challenging, since there will be little information within the knowledge bases about the many variants which will be detected in each sample. WGS of trios— proband, mother, father—is the sole solution to rapidly filter the variants to spot the variants that segregate with the phenotype. Trio WGS has been successfully deployed within the neonatal and pediatric medical care setting where detailed phenotyping isn't possible and time is of the essence69. A broader usage of WGS will eventually enable the invention of novel regulatory elements scattered across the vast non-coding regions of the human genome and yield fundamental insights into disease mechanisms. However, the data analysis challenges are signifcant and it will take highly skilled teams of bioinformaticians, computer scientists, data scientists, variant scientists and experimental biologists working closely with patients and clinicians to push the envelope.