Principles of Next generation sequencing Technology

Abid Sharif
East Midlands Regional Molecular Genetics Service
Outline of talk

• NGS – What is it
• Chemistry of NGS
• Different platforms
• Work flow
• Target enrichment
• Data Analysis- Bioinformatics
• Data Storage
• Interpretation
Introduction

• Next generation sequencing is also known as massively-parallel sequencing
• It enables large amount sequencing to be performed in a single assay
• To fully appreciate the advantages NGS affords we need to remind ourselves of life before NGS
Sanger sequencing - chain termination method
Sanger sequencing - chain termination method
Sanger sequencing - chain termination method

3,000,000,000 bp human genome
1000 bp in one well
384 wells in one run
10 runs per day

782 days = 2 years !!!

Human genome sequencing consortium generated 3 Gb at the cost of approximately US $3 billion and took 13 years.
Next generation sequencing/Massively parallel sequencing

- Next generation sequencing/Massively parallel sequencing allows millions of sequencing reactions to be carried out in parallel.
- It allows multiplexing for different patient samples.
- Sequencing and detection take place simultaneously.
- Sequencing is of clonally amplified DNA templates which has been amplified from a single fragment.
- Disadvantage: read lengths shorter than Sanger. This is due to the signal to noise ratio which increases with read length.
- Error rates for individual NGS reads is greater.
- Accuracy in the NGS is achieved by sequencing a given region multiple times e.g. 30 times (reads) and a consensus sequence is generated.
- Hence the initial advantage is off set by needing to sequence the same region so many times but despite this cost savings can still be made.
- To assemble, align and analyse NGS data requires an adequate number of overlapping reads.
Definitions and Terminologies

- Read length – The number of bases sequenced in a fragment
- Capture efficiency
- Paired end sequencing
- Read depth - How many times has a base been sequenced?
NGS Workflow

Enrichment

Amplification of the Sequencing library

Sequencing

Source: Heart © 2012 BMJ Publishing Group Ltd & British Cardiovascular Society
Common library preparation method

- **Fragmentation**
  - Sonication
  - Nebulization
  - Shearing

- **End-Repair**
  - Size selection
  - Blunt-end or A-overhang

- **Adaptor Ligation**
  - Ligation
  - Purification
Target enrichment: Multiplex PCR

Gene of interest

Perform multiplex PCR-based target enrichment

Pool amplicons

Prepare library

One library per sample
Target enrichment: Fluidigm access array

48 primers x 48 samples with unique barcodes
2304 unique products
Target enrichment: Microdroplet PCR

RainDance Microdroplet PCR

- Forward- and reverse-targeting primer droplets
- Genomic DNA and associated PCR reagent droplets
- Microfluidic chip
- Merging area
- Microdroplet PCR in a single tube

Reported 84% of capture efficiency
Target Enrichment: PCR - Truseq custom amplicon

Use DesignStudio to create custom oligo capture probes flanking each region of interest.

CAT probes hybridize to flanking regions of interest in unfragmented gDNA.

Extension/Ligation between Custom Probes across regions of interest.

PCR adds indices and sequencing primers.

Uniquely tagged amplicon library ready.
Target Enrichment: Enrichment by circularisation of selector probes

HaloPlex Protocol

Step 1.

1. Digest and denature sample DNA
   Restriction enzymes are used to fragment DNA.

Step 2.

2. Hybridize oligonucleotide probe library
   Each probe is designed to hybridize to both ends of a targeted DNA restriction fragment, guiding the targeted fragments to form circular DNA molecules. The probes also contain a method-specific sequencing motif that is incorporated during the circularization.

3. Purify and ligate targets
   HaloPlex probes are biotinylated so targeted fragments can be retrieved using magnetic streptavidin beads. Target-probe complexes are closed by ligation to ensure that only perfectly hybridized fragments are circularized.

Step 3.

4. Amplify targeted fragments with PCR
   Only circular DNA targets are amplified, providing an enriched amplification product that is ready for sequencing. Sample barcodes are introduced during amplification for precise tracking.

The HaloPlex PCR protocol combines target enrichment and library preparation in a single workflow. The current protocol takes two days to complete however, alterations in the Haloplex protocol can reduce the turnaround time to less than one day.
Target enrichment: Hybridisation based methods
Enrichment by Nextera:
Nextera: Rapid fragmentation and ligation with adapters
Nextera® Rapid Capture Custom Enrichment

A. Prepare sample

B. Denature double-stranded DNA library (for simplicity, adapters and indexes not shown)

C. Hybridize biotinylated probes to targeted regions

D. Enrich using streptavidin beads

E. Elute from beads
Clonal amplification

Bridge amplification (Illumina)

- Takes place on the sequencing instrument (flow cell). The surface of the flow cell is densely coated with primers that are complementary to the primers attached to the DNA library fragments.
clonal amplification

Emulsion PCR

Aqueous bubbles immersed in oil form microscopic reaction entities for each individual capture bead. Clonal amplification of the initial template molecules into colonies is required because the technology currently used is not sensitive enough to detect the incorporation of a nucleotide in the sequencing reaction.
Sequencing Chemistries

- Sequencing by synthesis (Ion Torrent, Illumina)
- Sequencing by oligo ligation detection (SOLiD)
- Others: Single molecule sequencing real time sequencing, nanopore technology
Sequencing chemistry-Ion torrent

- Ion Semiconductor Sequencing
- Detection of hydrogen ions during the polymerization DNA
- Sequencing occurs in microwells with ion sensors
- No modified nucleotides
- No optics
During sequencing, the four bases (A, T, G, and C) are introduced one at a time during the run. A nucleotide complementary to the base on the template is incorporated into the growing genome strand by DNA polymerase. The incorporation of each nucleotide is measured as a change in voltage and the number of nucleotides incorporated is determined from the increased magnitude of the voltage. Problem with homopolymer detection.
Illumina: sequencing by synthesis using reversible terminator

Only one nucleotide labelled with a fluorophore is incorporated at each cycle. After incorporation the array is imaged, terminator moiety cleaved and fluorescent label is cleaved off and removed. Fresh nucleotides added and the cycle is repeated.
# Different Illumina machines

<table>
<thead>
<tr>
<th></th>
<th>MiniSeq System</th>
<th>MiSeq Series</th>
<th>NextSeq Series</th>
<th>HiSeq Series</th>
<th>HiSeq X Series*</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Maximum Output</strong></td>
<td>7.5 Gb</td>
<td>15 Gb</td>
<td>120 Gb</td>
<td>1500 Gb</td>
<td>1800 Gb</td>
</tr>
<tr>
<td><strong>Maximum Reads per Run</strong></td>
<td>26 million</td>
<td>26 million²</td>
<td>400 million</td>
<td>5 billion</td>
<td>6 billion</td>
</tr>
<tr>
<td><strong>Maximum Read Length</strong></td>
<td>2 × 150 bp</td>
<td>2 × 300 bp</td>
<td>2 × 150 bp</td>
<td>2 × 150 bp</td>
<td>2 × 150 bp</td>
</tr>
<tr>
<td><strong>Run Time</strong></td>
<td>4–24 hours</td>
<td>4–55 hours</td>
<td>12–30 hours</td>
<td>&lt;1–3.5 days (HiSeq 3000/HiSeq 4000) 7 hours–6 days (HiSeq 2500)</td>
<td>&lt;3 days</td>
</tr>
<tr>
<td><strong>Benchtop Sequencer</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td><strong>System Versions</strong></td>
<td>MiniSeq System for low-throughput targeted DNA and RNA sequencing</td>
<td>MiSeq System for targeted and small genome sequencing</td>
<td>MiSeq FGx System for forensic genomics</td>
<td>NextSeq 500 System for everyday genomics</td>
<td>HiSeq 3000/HiSeq 4000 System for production-scale genomics</td>
</tr>
<tr>
<td></td>
<td>MiniSeq DX System for molecular diagnostics</td>
<td>NextSeq 550 System for both sequencing and cytogenomic arrays</td>
<td>HiSeq 2500 Systems for large-scale genomics</td>
<td>HiSeq X Five System for production-scale whole-genome sequencing</td>
<td>HiSeq X Ten System for population-scale whole-genome sequencing</td>
</tr>
</tbody>
</table>
## Ion Torrent Sequencers

### Comparison Table

<table>
<thead>
<tr>
<th></th>
<th>Ion Torrent PGM</th>
<th>Ion Torrent Proton</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Chip Type</strong></td>
<td>PGM 314</td>
<td>PGM 316</td>
</tr>
<tr>
<td><strong># of sensors</strong></td>
<td>1.3M</td>
<td>6.3M</td>
</tr>
<tr>
<td><strong>Total output</strong></td>
<td>10-40Mb</td>
<td>100-400Mb</td>
</tr>
<tr>
<td><strong>Run time</strong></td>
<td>1-2 hrs</td>
<td>1-2 hrs</td>
</tr>
<tr>
<td><strong>Read length</strong></td>
<td>up to 400bp</td>
<td>~200bp</td>
</tr>
<tr>
<td><strong>Total reads</strong></td>
<td>up to 0.6M</td>
<td>up to 3M</td>
</tr>
</tbody>
</table>
Validation

- **Sequencing Platform validation:** system provides reliable analysis across the genomic regions of interest targeted by the test.

- **Test validation:** system correctly identifies disease-associated and other variants in targeted regions of the genome.

- **Informatics pipeline validation:** algorithms reliably analyse platform data to produce an accurate sequence.
Validation

- Ensure correct region of interest
- Use of reference material
- Use of controls
- Ensure adequate coverage - sequence depth >30 reads
- Decide what to do about regions of low coverage - Fill in gaps by sanger sequencing or redesign the probes?

- **Analytical sensitivity:** assay detect a sequence variation when present this value (false negative rate).
- **Analytical specificity:** probability that assay will not detect sequence variation(s) when none are present (false positive rate).
Trusight hereditary cancer panel

<table>
<thead>
<tr>
<th>Gene</th>
<th>Gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>FANCC</td>
<td>SDHB</td>
</tr>
<tr>
<td>FANCD2</td>
<td>SDHC</td>
</tr>
<tr>
<td>FANCE</td>
<td>SDHD</td>
</tr>
<tr>
<td>FANCF</td>
<td>SLX4</td>
</tr>
<tr>
<td>FANCG</td>
<td>SMAD4</td>
</tr>
<tr>
<td>FANCI</td>
<td>SMARC1B1</td>
</tr>
<tr>
<td>FANCL</td>
<td>STK11</td>
</tr>
<tr>
<td>FANCM</td>
<td>SUFU</td>
</tr>
<tr>
<td>FH</td>
<td>TMEM127</td>
</tr>
</tbody>
</table>

Miseq

23 samples +blank

Bioinformatics: JSI software

Confirmation of variants by Sanger sequencing
Gene panels

### TruSight Cancer Panels

(Illumina Catalog No. TG-141-1002).

<table>
<thead>
<tr>
<th>Breast and Ovarian Cancer</th>
<th>Primary</th>
<th>Extended</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRCA1</td>
<td>ATM</td>
<td>BRCA1</td>
</tr>
<tr>
<td>BRCA2</td>
<td>BAP1</td>
<td>BRCA2</td>
</tr>
<tr>
<td>TP53</td>
<td>CDH1</td>
<td>CHEK2</td>
</tr>
<tr>
<td></td>
<td>PALB2</td>
<td>APC</td>
</tr>
<tr>
<td></td>
<td>RAD51C</td>
<td>MTPYH</td>
</tr>
<tr>
<td></td>
<td>MSH2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MSH6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PMS1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PMS2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RB1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>STK11</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PP1M1D</td>
<td></td>
</tr>
<tr>
<td></td>
<td>NBN</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Colorectal Cancer</th>
<th>Primary</th>
<th>Extended</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLH1</td>
<td>APC</td>
<td></td>
</tr>
<tr>
<td>MSH2</td>
<td>BMPR1A</td>
<td></td>
</tr>
<tr>
<td>MSH6</td>
<td>EPCAM</td>
<td></td>
</tr>
<tr>
<td>PMS1</td>
<td>MLH1</td>
<td></td>
</tr>
<tr>
<td>PMS2</td>
<td>MSH2</td>
<td></td>
</tr>
<tr>
<td>PMS2</td>
<td>MSH6</td>
<td></td>
</tr>
<tr>
<td>APC</td>
<td>MTPYH</td>
<td></td>
</tr>
<tr>
<td>MUTYH</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<p>| | | |
|       |          |          |
|       |          |          |
|       |          |          |</p>
<table>
<thead>
<tr>
<th>Colorectal Cancer</th>
<th>Primary</th>
<th>Extended</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLH1</td>
<td>APC</td>
<td></td>
</tr>
<tr>
<td>MSH2</td>
<td>BMPR1A</td>
<td></td>
</tr>
<tr>
<td>MSH6</td>
<td>EPCAM</td>
<td></td>
</tr>
<tr>
<td>PMS1</td>
<td>MLH1</td>
<td></td>
</tr>
<tr>
<td>PMS2</td>
<td>MSH2</td>
<td></td>
</tr>
<tr>
<td>PMS2</td>
<td>MSH6</td>
<td></td>
</tr>
<tr>
<td>APC</td>
<td>MTPYH</td>
<td></td>
</tr>
<tr>
<td>MUTYH</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

|       |          |          |
|       |          |          |
|       |          |          |

[Image of gene panels]
## Tumor Actionable Mutations panel V2

Mutations with confirmed prognostic or diagnostic importance for solid tumor

<table>
<thead>
<tr>
<th>Genes covered</th>
<th>Actionable mutations detected</th>
<th>Indication</th>
<th>Clinical relevance</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRAF</td>
<td>V600 mutations</td>
<td>Melanoma, Colon cancer, NHL, GIST, NSCLC</td>
<td>Venamafenib recommended if mutation detected</td>
</tr>
<tr>
<td>EGFR</td>
<td>Kinase Domain Mutations, Exon 19 Deletions, Exon 20 Insertions</td>
<td>NSCLC</td>
<td>Tested before therapy for recurrence or metastases</td>
</tr>
<tr>
<td>IDH1</td>
<td>R132 mutations</td>
<td>CNS</td>
<td>Mutated gene is a low risk feature in these tumors</td>
</tr>
<tr>
<td>IDH2</td>
<td>R172 mutations</td>
<td>CNS</td>
<td>Mutated gene is a low risk feature in these tumors</td>
</tr>
<tr>
<td>KIT</td>
<td>A582, Y503, F506, W557, V559, V560, L576, K642, N522 mutations</td>
<td>AML, GIST</td>
<td>Tested if imatinib therapy begun for unresectable or metastatic disease</td>
</tr>
<tr>
<td>KRAS</td>
<td>G12, G13, A59, Q61, K17, A146 mutations</td>
<td>NSCLC, Colon</td>
<td>EGFR antibody drugs avoided if mutated</td>
</tr>
<tr>
<td>NRAS</td>
<td>G12, G13, A59, Q61, K17, A146 mutations</td>
<td>Colon</td>
<td>EGFR antibody drugs avoided if mutated</td>
</tr>
</tbody>
</table>

Testing considered for patients with primary disease especially high-risk tumors
Secondary or acquired mutations can be associated with resistance during long-term treatment
Supports GIST diagnosis
Advantages of gene panel

- Process more patient samples at a time
- Not sequential testing
- More genes in one test
  - Higher diagnostic rate
- Cost savings
Disadvantages

Å Needs to be sufficient number of referrals - Batch
Å In a clinical diagnostic laboratory many referrals for a vast number of disorders
Å Design and validate panel for each type of referral/ syndrome - costly and time consuming
Clinical exome sequencing

- Large number of genes would cover all types of referrals
- Large target area - can run fewer patient sample per run
- Requires high capacity sequencer.
- Only analyse set of genes relevant to reason for referral
- Many different panels of genes can be analysed pertinent to the reason for referral (virtual panel)
- Need higher capacity sequencing instrument to be cost effective
- Some genes may be poorly covered.
Potential limitations with NGS

- Homopolymer tracks: (sequencing chemistry-alignment issues)
- Complex rearrangements
- Copy number variation?
- GC bias: coverage may be low
- Sanger validation?
Bioinformatics from data to sequence
Bioinformatics:

- Bioinformatician: develop an in-house analysis pipeline
- Outsource:
- Analysis software: licence
Data storage

- Negotiate with IT more space on server
- Archive - outsource
- Resequencing
Interpreting variants

À Recent technology has increased our capacity to sequence large numbers of genes
À Our ability to interpret sequence changes lags far behind
À Hence many of variants detected by sequencing are reported as variants of unknown clinical significance
À Nevertheless there are established procedures for investigating the clinical significance of variants and this is carried out by appropriately qualified and experienced staff
Interpreting variants

- Important aspects of clinical scientists role
- Best practice guidelines available
- Need to follow a strict procedure to arrive at a conclusion-documentation
- Appreciation of the limitation of tools available
- Decision should be based on evidence from multiple sources
- Requires critical evaluation of all information available
First step: Correctly naming the variant

- Descriptions of sequence variants should use HGVS (Human Genome Variation Society) nomenclature.
- [https://mutalyzer.nl/](https://mutalyzer.nl/) can be used to support correct nomenclature.
- HGVS recommendations for the description of sequence variants are designed to be stable, meaningful, memorable and unequivocal.
- However every now and then small modifications are made in the guidelines to remove small inconsistencies and/or to clarify confusing conventions and to add any previously uncovered sequence alterations.
- Variants should therefore be described with respect to a HGVS nomenclature [http://www.hgvs.org/mutnomen/](http://www.hgvs.org/mutnomen/) and with respect to a reference DNA sequence specified by an accession number and a version e.g. BRCA2 NM_000059.3:c.7977-1G>C
Investigating variants of unknown clinical significance

Which types of variants?

- Missense changes
  - Amino acid changes of unknown effect
  - The effect depends on the chemical difference between old/normal and new amino acid
  - Location of the change: is it in the active site?
  - Location: Is it in the non-functional part but important in maintaining the protein structure/stability. May lead to the degradation of protein or in correct localisation of the protein or may make the protein less or more flexible, which may lead to difference in specificity. May give a loss of function or a gain of function.

- Intronic variants (not donor/acceptor site)

- Synonymous (silent) changes
  - Variant does not change the amino acid
  - May create a cryptic splice site
  - May disrupt Exon Splice Enhancer/Silencer sequences
Investigating variants of unknown clinical significance

- Need to accumulate information/evidence to make a decision on the likelihood of a variant being pathogenic or not
- No one piece of evidence should be used alone
- ACGS Best Practice Guidelines 2013
  - Practice Guidelines for the Evaluation of Pathogenicity and the Reporting of Sequence Variants in Clinical Molecular Genetics.
Lines of evidence

- Mutation databases
- Presence/absence in SNP databases
- Testing matched controls
- Literature search
- Co-occurrence in trans with a known pathogenic mutation
- Co-segregation with the disease in a family
- Loss of heterozygosity in tumour (loss of the wt allele)
- Occurrence of a new variant with sporadic incidence of the Disease
- Species Conservation
- In silico prediction software programs
Steps to investigate unknown sequence changes continued

**Literature search:** Need to read any publications critically. Many of these will be from researchers. Need to decide if sufficient evidence for clinical use.

- **Multiple search engines**
  - Yahoo, PubMed, Google Scholar
- **May show evidence of functional studies**
  - Requires knowledge of all possible functions of the protein
  - Requires a suitable assay
  - Usually performed only in research laboratories
  - Just because the variant in question leads to reduction or abolition of function does not necessarily mean that it would lead to the phenotype in question e.g. cancer
Steps to investigate unknown sequence changes

- Co-segregation with disease in the family
  - Limited by partial penetrance
  - The disorder may be late onset
  - Phenocopies: individuals displaying similar symptoms
  - Requires samples from enough family members.
  - Is the variant linked to unidentified pathogenic variant on the same allele? Necessitates functional assay on the observed sequence variant.

- Co-occurrence in trans with a known pathogenic mutation
  - Need to determine if on opposite chromosome (in trans) by testing parents or other relatives
  - Principle: 2 mutations may be embryonic lethal e.g. BRCA1
  - Principle: 2 mutations in homozygous state or in compound heterozygote state may give distinct phenotype e.g. BRCA either embryonic lethal or leads to Fanconi’s anaemia.
  - Depends on penetrance of heterozygote genotypes which is impossible to determine.

- Occurrence of a new variant concurrent with the disease
  - Requires parental samples
  - Partial penetrance
  - Imprinting
Steps to investigate unknown sequence changes

**Species conservation**

- Principle: the more conserved an amino acid is then the less likely an amino acid change will be tolerated.
- Requires reference sequence alignments. Requires careful design and may need to vary with gene type.
- Recommended to include at least five mammalian homologues over 300 million years.
- Need to cover an appropriate amount of evolutionary time.
Variant Report

ALAMUT RESULTS

Report

Warning: This report is based on knowledge and data that are not firmly established. Consequently, medical decisions must not be made on the basis of this report.

GNAQI Variation

Class 3: Unknown pathogenicity
Transition from C to T in exon 6
Missense substitution
Arg at position 209 is changed to Cys

HGVS v2.0 Nomenclature

cDNA Level: NM_138756.2:c.625C>T

gDNA Level: Chr18:GRCh37.p6:6370674C>T

p.Arg209Cys

Pathogenicity clues

- Moderately conserved nucleotide (phyloP: 3.60 [-14.1,6.4])
- Highly conserved amino acid, up to C. elegans (considering 10 species)
- Large physicochemical difference between Arg and Cys (Grantham dist.: 180 [0-215])
- This variant is in protein domains
  - Guanine nucleotide binding protein (G-protein), alpha subunit
  - Small GTPase superfamily, ARF/SAR type
  - Align GYGQ, C60 (QY: 0.00 - GD: 1.93)
  - SIFT: Deleterious (score 0, median: 0.32)
  - MutationTaster: disease causing (p-value: 1)

Occurrences

<table>
<thead>
<tr>
<th>Patient</th>
<th>Family</th>
<th>RNA Analysis</th>
<th>Phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>N1205936</td>
<td></td>
<td></td>
<td>Global developmental delay, deeply set eyes, motor delay, cerebral palsy, thick upper lip, thin, long nose, persistent fist, pectus, absence of the ribs, genital anomalies</td>
</tr>
</tbody>
</table>

©Interactive Biosoftware - Created by Alamut Visual v:2.7.1 on 20/01/2016
Tumour Pathology and Immunochemistry

- If details of tumour phenotype differ between clearly BRCA mutations and non-carriers then it should be possible to use tumour phenotype to help assess UVs e.g. BRCA1
- Requires good quality tumour.
<table>
<thead>
<tr>
<th>Class</th>
<th>Wording to include within reports</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><strong>Not</strong> pathogenic “Common” polymorphism and therefore not reported</td>
</tr>
<tr>
<td>2</td>
<td><strong>Unlikely</strong> to be pathogenic Diagnosis not confirmed molecularly</td>
</tr>
<tr>
<td>3</td>
<td><strong>Uncertain</strong> pathogenicity <strong>Does not</strong> confirm or exclude diagnosis</td>
</tr>
<tr>
<td>4</td>
<td><strong>Likely</strong> to be pathogenic <strong>Consistent</strong> with the diagnosis</td>
</tr>
<tr>
<td>5</td>
<td>Predicted to be <strong>pathogenic</strong> This result <strong>confirms</strong> the diagnosis</td>
</tr>
</tbody>
</table>
Thank you