Implementing high-throughput cancer predisposition gene testing using Illumina HiSeq2500

Nazneen Rahman MD PhD
Head of Cancer Genetics
Institute of Cancer Research, London

Open and tweetable
@rahman_nazneen
Two types of cancer gene

Genes in which germline mutations confer high/mod clinically important risks of cancer are called **Cancer Predisposition Genes**
> 100 cancer predisposition genes have clinical utility

Improved diagnosis

Optimised management and follow-up

Tailored therapies

Information for relatives
Cancer prevention

Cost efficiency

Strong clinical and economic rationale for greater genetic testing of cancer predisposition genes
We need to test more genes in more people
A Wellcome Trust funded, cross-disciplinary, translational initiative to develop the assays, informatics, clinical infrastructure, education, ethics and evaluation that will allow implementation of (germline) cancer genetic testing into routine clinical care of cancer patients and their relatives.

www.mcgprogramme.com
MCG Workstreams

1. Technical
2. Analytical
3. Interpretation
4. Implementation
5. Education & Engagement
6. Evaluation & Ethics

Genetic Testing
Genetic testing

• A genetic test involves three components
  – Data generation (sequencing)
  – Data analysis
  – Data interpretation

• Analogy of a book: Need all the letters (generation), in the right order with correct punctuation etc (analysis), then need to read and understand its meaning (interpretation).
How can one do CAPPA?

- Genome
  - All genes (1%)
    - 100 genes (0.01%)
      - 1 gene
Why we are using a targeted panel

• Need high-throughput (1000s of tests).
• Needs to be *at least* equivalent to existing tests
  – Every base in gene must be covered (50x).
  – Every mutation class has to be detected.
  – *V* high sensitivity and specificity
  – Robustly provide both positive and negative results
  – Consistent turnaround times (<8 weeks)
• Cost-effective (from patient to report)
• Likely change to exome/genome in time.
TruSight Cancer Panel

3881 biotinylated 80 bp oligonucleotides (probes) that hybridise to specific DNA targets

<table>
<thead>
<tr>
<th>97</th>
<th>Genes/gene regions</th>
</tr>
</thead>
<tbody>
<tr>
<td>260</td>
<td>Cancer GWAS SNPs</td>
</tr>
<tr>
<td>24</td>
<td>Fingerprinting SNPs</td>
</tr>
</tbody>
</table>

\[
egin{align*}
97 & \quad \text{Genes/gene regions} \\
260 & \quad \text{Cancer GWAS SNPs} \\
24 & \quad \text{Fingerprinting SNPs}
\end{align*}
\]

\[
\begin{align*}
& 1449 \text{ exons} \\
& 287 \text{ SNPs} \\
& 456 \text{ KB}
\end{align*}
\]

1736 targets

0.01% of the genome
TruSight target:
CDS exon plus 60 bp flanking intronic sequence
Targeting content – BRCA2
Workflow optimisation

Off-the-shelf workflow: plexing 12 samples
Higher throughput workflow 48 samples with no loss of performance (every base minimum of 50x).

HiSeq2500 Rapid, 200 cycle run:
24 flow cells / month
48 lanes in total
192 samples per run
576 samples per week / per instrument
96 sample pipeline

DAY 1
• Quantify input DNA

DAY 2
• Tagmentation
• Add index tags
• Quantify and QC library

DAY 3-4
• Pool 48 samples x2
• Hybridise to target probes
• Wash away unselected DNA

DAY 5
• Quantify and QC enriched library
• Dilute enriched library

DAY 6
• Sequence

Low template requirement
Simple lab process
Low failure rate
Evaluation of performance

• 378/384 samples passed
• 1449 exonic targets
• Minimum threshold of 50x high quality coverage per base of a target

In 95% of samples processed, 95% of TruSight target bases will reach 50x high quality coverage
- 8/1449 targets failed in every sample
- 27/1449 targets failed in >50% of samples
- 71/1449 targets failed in at least 5% of samples
- 38/95 genes >50x quality coverage at every base

<table>
<thead>
<tr>
<th>TARGET</th>
<th>Target_BP</th>
<th>Lowest_BPlength_&lt;50x QCOV</th>
<th>Median_BPlength_&lt;50x QCOV</th>
<th>Highest_BPlength_&lt;50x QCOV</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDKN1C_2</td>
<td>840</td>
<td>223</td>
<td>361.5</td>
<td>539</td>
</tr>
<tr>
<td>CEBPA</td>
<td>1097</td>
<td>212</td>
<td>338</td>
<td>596</td>
</tr>
<tr>
<td>FANCE_1</td>
<td>268</td>
<td>24</td>
<td>109.5</td>
<td>148</td>
</tr>
<tr>
<td>GATA2_4</td>
<td>662</td>
<td>11</td>
<td>341</td>
<td>532</td>
</tr>
<tr>
<td>NSD1_1</td>
<td>947</td>
<td>7</td>
<td>303.5</td>
<td>633</td>
</tr>
<tr>
<td>PHOX2B_1</td>
<td>536</td>
<td>93</td>
<td>142</td>
<td>260</td>
</tr>
<tr>
<td>RHBDF2_1</td>
<td>440</td>
<td>168</td>
<td>323</td>
<td>394</td>
</tr>
<tr>
<td>RUNX1_1</td>
<td>496</td>
<td>152</td>
<td>329</td>
<td>416</td>
</tr>
</tbody>
</table>
Gene region **BRCA2**
Gene CEBPA
Probe optimisation

• 168/1449 exonic targets can be optimised by probe addition.
• Suboptimal targets generally associated with terminal drop-off, gapping in probe spacing, or presence of a repeat sequence.
• A few genes (CEBPA) and gene regions not well captured by TruSight.
BRCA1 and BRCA2 in practice

384 samples

3 sample failures

15 Sanger amplicon ‘fill-ins’ required

Median coverage 1000x
Minimum per base coverage 50x
Also have to pass base quality / mapping quality filters
Patient Sample

**CAPPA**
Cancer Predisposition and Pharmacogenetic Analysis

**GAMA**
Genetic Analysis and Mutation Annotation

**CIGMA**
Clinical Impact of Genetic Mutation Analysis

Data Generation

Data Analysis

Data Interpretation

Clinical Report
Software Development
GAMA workflow

GAMA.sh
- Begun by TGLclinical
- Creates analysis folder and scripts from templates with create_jobs.R
- Sends email notification

FASTQ_creation.sh
- Runs CASAVA based on TGL's SampleSheet.csv
- Sorts samples by fastq size into
- Sends email when finished

BAM_creation_$i.sh
- Alignment by Stampy
- BRCA coverage evaluation

SmallVariant_creation_$i.sh
- Variant calling by Platypus, SAVANT annotation
- BRCA test outputs by Sanger_creation_$i.R
- Sends email if all variant calling is finished

LargeVariant_creation_$i.sh
- Run modified ExomeDepth for each pool
- Sends email if all variant calling is finished

Dedicated HPC cluster
- 8 nodes
- 12 cores per node

Elise Ruark
GAMA workflow

GAMA.sh
Begun by TGLclinical
Creates analysis folder and scripts from templates with create_jobs.R
Sends email notification

FASTQ_creation.sh
Runs CASAVA based on TGL's SampleSheet.csv
Sorts samples by fastq size into
Sends email when finished

BAM_creation_${i}.sh
Alignment by Stampy
BRCA coverage evaluation

SmallVariant_creation_${i}.sh
Variant calling by Platypus, SAVANT annotation
BRCA test outputs by Sanger_creation_${i}.R
Sends email if all variant calling is finished

LargeVariant_creation_${i}.sh
Run modified ExomeDepth for each pool
Sends email if all variant calling is finished

Dedicated HPC cluster
8 nodes
12 cores per node

Elise Ruark
96 samples – 8 hours
100% sensitivity and specificity for BRCA
BRCA1 and BRCA2 in practice

384 samples

77 small variants to validate (by Sanger)

10 large variants to validate (by MLPA)

Validate all pathogenic mutations
Rare variants
Repeats and validations take ~3 days
CAPP - Cancer Predisposition and Pharmacogenetic Analysis

GAMA - Genetic Analysis and Mutation Annotation

CIGMA - Clinical Impact of Genetic Mutation Analysis

Patient Sample

Data Generation

Data Analysis

Data Interpretation

Clinical Report
CAPP A
Cancer Predisposition and Pharmacogenetic Analysis

GAMA
Genetic Analysis and Mutation Annotation

CIGMA
Clinical Impact of Genetic Mutation Analysis

Data Generation
Data Analysis
Data Interpretation
Clinical Report
Analysis indicates that this patient is heterozygous for the BRCA1 sequence variant c.2521C>T (p.Arg841Trp). This variant is listed in the BIC database as being of unknown clinical significance, the literature that describe it is contradictory, and in silico analyses gave inconclusive results, therefore the pathogenicity of this variant cannot be determined.

**Summary:** A variant of unknown clinical significance has been identified in BRCA1 in this patient
Current interpretation

• Clinical labs undertake highly intensive analysis of individual variant to decide if pathogenic.
• Do not take into account context of variant discovery.
• Frequently classify as a “variant of uncertain significance” (VUS)
  →implications unclear to clinician and patient
  →Ad hoc, inconsistent in management
  →Often assumed ‘guilty until proven innocent’
Most variants in cancer predisposition genes are not pathogenic

- We have LOTS of mutations/variants in our genomes.
- Pathogenic mutations (high risk for cancer) are rare.
- Evaluation of TruSight panel in 1000 UK population
  → 117 variants per individual, 3-4 rare.
  → Most (>80%) are rare (<1% of the population)
  → >10% of population has a rare BRCA variant
- Collectively rare variants are common!
- Most variants (>99%) are NOT pathogenic.
- All variants should be ‘innocent until proven guilty’
CIGMA aspirations

- Triage variants into clear clinical management categories.
- Fully informed, evidence-based and comprehensive.
- High-throughput with automated classifications for >95% of variants.
- Dynamic and iteratively improve
For each variant
CAPPA
GAMA
Multiple variants
For each variant
Data inputs
Decision tree
Clear statement about clinical pathogenicity
Output classes
Data inputs for variant classification

- Case variant data
- Genetic analyses of case data
- Control variant data
- *In silico* prediction
- Functional assays
Data inputs for variant classification

Gene level data

- Mechanism of pathogenicity
- Gene variability index
- Gene structure/function
- Clinical gene-phenotype relationship
  - Frequency of phenotype
  - Attribution of gene for phenotype
  - Penetrance of gene for phenotype
### CIGMA database for BRCA genes

<table>
<thead>
<tr>
<th>CIGMA for BRCA1/BRCA2</th>
</tr>
</thead>
<tbody>
<tr>
<td>• <strong>All</strong> possible base substitution mutations in BRCA1/BRCA2 generated</td>
</tr>
<tr>
<td>• case mutational data (3 sources: HGMD, LOVD, BIC)</td>
</tr>
<tr>
<td>• multifactorial genetic analysis data (2 sources: IARC, Easton et al)</td>
</tr>
<tr>
<td>• control mutational data (3 sources: ICR-1958, EVS, 1KG)</td>
</tr>
<tr>
<td>• in silico data (7 sources: SIFT, PolyPhen, MAPP, Align-GVGD, MutationTaster, NN-Splice, Max-ENT)</td>
</tr>
</tbody>
</table>
Output classes for clinical management

Class 1 - Manage as pathogenic

Class 2 - Manage as not pathogenic

[Class 3 - Bespoke management]
Current classification of rare variants

- **Pathogenic**
- **Likely pathogenic**
- **Variant of Uncertain Significance**
- **Unlikely pathogenic**
- **Non-pathogenic**
CIGMA: clinical-level classifications

- Manage as pathogenic
- Manage as not pathogenic

Rare variants/mutations detected in 143 BRCA1/BRCA2 full diagnostic tests
CIGMA: clinical-level classifications

- Manage as pathogenic
- Manage as not pathogenic

Bespoke management

Rare variants/mutations detected in 143 BRCA1/BRCA2 full diagnostic tests
TruSight cancer pipeline

96 samples
14 day turnaround

DNA extraction
TruSight
GAMA
CIGMA
Sanger / MLPA confirmations
Clinical report

Days 1-2
Days 3-8
Day 9
Days 10-13
Day 14
Acknowledgements

Many people are involved in the MCG consortium
Nazneen Rahman (programme director)
Daniel Riddell (programme manager)
Shazia Mahamdallie, Elise Ruark, Clare Turnbull, Helen Hanson, Angela George, Ingrid Slade (workstream coordinators)
UK Cancer Genetics Consultation Group
WTCHG - Gerton Lunter, Chris Holmes, Andrew Rimmer, Márton Münz, Anna Fowler

Programme Committee and Advisory Board