Amplicon resequencing on Illumina: possibilities for clinical research

James Hadfield
Illumina June ‘10
CRUK CRI Genomics core

CRUK is the world’s leading independent cancer research organisation. Over 3000 CRUK supported scientists (plus 1500 doctors and nurses). £333M in 2008

Cambridge Research Institute
- Opened February 2007
- Translational research focus
- Strong core facilities including: Genomics, research focus
- Strong core facilities including: Genomics, Histopathology, Microscopy, Flow, Imaging, Proteomics

https://cruk.finchlab.com

James.Hadfield@cancer.org.uk  Amplicon resequencing on Illumina  June’10  Slide 2 of 17
The need for clinical resequencing

• Currently round 60 clinical tests and only about 10 with clinical utility in drug stratification are being run on around 5-10,000 cases per year in the UK.

  Acquired solid tumour tests: KRAS, BRAF, MSI, EGFR, c-KIT & PDGFRA (for GISTs), MGMT methylation & LoH

  Germline mutation tests: BRCA1, BRCA2, MLH1, MSH2, MSH6, PMS2, TP53, MUTYH, APC, PTEN, NF1, TSC1, TSC2

  Unproven clinical utility: PIK3CA, PTEN, N-RAS, H-RAS, TMPRSS fusions, ERCC2, RRM1, UGT1A1

• Potential: BrCA 54000, PrCa 36000, OvCa 6000, PaCa 7000 UK cases per year
• Over 100,000 cases where early clinical tests would be useful. Which tests and how?

• Difficulty in translating these tests from current accredited methods to next-gen:
  Sample collection, Sample Prep, Throughput, Analysis, Ethics

• Is next-gen the right approach?
  BRCA pilots by St James’s in Leeds suggest this would be cost competitive with current methods. Estimates of sequencing analysis are as low as £10 per sample/locus for all sequencing costs.
Genome capture

- Resequencing whole genomes is still too expensive – time and money
- There is a need to select specific genomic regions in large cohorts
- Multiple approaches with different strengths

<table>
<thead>
<tr>
<th></th>
<th>PCR</th>
<th>MIP</th>
<th>Array capture</th>
<th>Solution capture</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost</td>
<td>High</td>
<td>&lt;10 samples, high</td>
<td>Medium</td>
<td>&lt;10 samples,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&gt;10 samples, low</td>
<td></td>
<td>medium</td>
</tr>
<tr>
<td>Ease of use</td>
<td>Low</td>
<td>High</td>
<td>Medium</td>
<td>High</td>
</tr>
<tr>
<td>Mass DNA</td>
<td>8ug for 1Mb</td>
<td>20ng</td>
<td>10-15ug</td>
<td>3ug</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>99.5%</td>
<td>98%</td>
<td>98% of target</td>
<td>99.5% of target</td>
</tr>
<tr>
<td>Specificity</td>
<td>93%</td>
<td>98%</td>
<td>70%</td>
<td>80%</td>
</tr>
<tr>
<td>Uniformity</td>
<td>80%, 2 fold</td>
<td>60%, 10 fold</td>
<td>60%, 1.5 fold</td>
<td>60%, 1.5 fold</td>
</tr>
<tr>
<td>Reproducibility</td>
<td>Up to 100%</td>
<td>92%</td>
<td>95%</td>
<td>96%</td>
</tr>
</tbody>
</table>

Table adapted from: Mamanova et al. Target-enrichment strategies for next-generation sequencing. Nature Methods 2010

- Careful thought needs to be given to the needs of a specific project.
- Total sequencing project costs can still be very high, genome capture can be more expensive than the sequencing.
Array/In-solution capture

1. Define ROI
2. Design targets (>98%)
3. Synthesise probes
4. Illumina Library (unamplified)
5. Sequence ROI

Hybrid capture > 100,000 exons

Adapter-modified shotgun library

Solution hybridization

Array capture

Bead capture
PCR capture

direct-sequence PCR
*Long range PCR*
*Next-gen PCR*

multiple modalities allow users to choose what best fits their need

Amplicon size and number of samples need to ‘fit’ sequence output

Primer design, PCR validation, pooling, library prep

---

Define ROI

Design primers (>99.5%)

PCR amplify

Illumina

Library

Sequence ROI
PCR capture: IrPCR

TP53 the most commonly mutated cancer gene
A tumour suppressor gene with anti-proliferative cellular function.
May be predictive clinically e.g. poor prognosis.
Most mutations are SNPs resulting in mutant protein.
Number of mutations is high; SNPs, In-Dels, etc. Over 25,000 mutations in IARC TP53 database.
Hotspots vs the whole gene

Sanger sequencing is very labour intensive
Ahmed & Brenton et al: Driver mutations in TP53 are ubiquitous in high grade serous carcinoma of the ovary
J Pathol 2010; 221: 49–56

How to analyse 1000 samples?

Tested long range PCR sequencing capability
Amplify whole genes and not just exons
What to do with the data?
   Find mutations; SNPs, In-Dels, etc
   Validate (design PCR-sequencing assay)
   Alignment to genome or ROI?
   Depth of coverage, smoothness, primer spikes
PCR capture: lrPCR

33 samples multiplexed after lrPCR for library prep
Read density imbalance between amplicons: QC and normalisation
PCR primer and other coverage spikes
Alignment to genome vs ROI

![IGV genome browser](IGV.png)

1st prep
2nd prep
Next-gen PCR capture: Rain Dance

8 Tumour DNA’s amplified for 4500 amplicons, single lane per sample and pooled analysis

Figure from: Tewhey et al, Microdroplet based PCR enrichment for large-scale targeted sequencing. Nature Biotechnology 2009
Verified one loss of function truncating mutation

A Tumour suppressor gene with mutations found in gastric, breast, colorectal, thyroid and ovarian cancer. Loss of function is thought to contribute to progression in cancer by increasing proliferation, invasion, and/or metastasis. Somatic mutation in exon 12.
Next-gen PCR capture: Fluidigm

- Microfluidic parallel PCR plates
- Users retain flexibility over amplicon choice

- AccessArray: Amplicon generation, barcoding and tagging for direct sequencing
- Amplification of 96 170-580bp amplicons
- From 48 DNA samples (possibilities for population genetics)
- Producing Illumina ready sequencing ‘libraries’

- Future developments for multi-modal analysis: real-time PCR (CNV), sequence analysis (SNPs, InDels) from a single platform.

AccessArray £200

James.Hadfield@cancer.org.uk  Amplicon resequencing on Illumina  June’10  Slide 11 of 17
Next-gen PCR capture: Fluidigm

Access Array
Next-gen PCR capture: Fluidigm

Amplification of seven Cancer genes, designed for 454 sequencing
48 Cancer cell lines were run on a single AccessArray, 96 amplicons
Producing Illumina ready sequencing libraries
Pooled for ~2% of a PE72bp lane producing >1500 fold coverage
Only 72bp from each end of the amplicon and the first 20-25 are primer

<table>
<thead>
<tr>
<th>Gene</th>
<th>PCR bp</th>
<th>Ave Reads</th>
<th>Mutations</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRAF</td>
<td>410</td>
<td>3218</td>
<td>SNP: 5, 6, 21, 25, 26</td>
</tr>
<tr>
<td>ERBB4</td>
<td>405</td>
<td>2727</td>
<td></td>
</tr>
<tr>
<td>KRAS</td>
<td>399</td>
<td>3443</td>
<td>SNP: 4, 9</td>
</tr>
<tr>
<td>MET</td>
<td>236</td>
<td>6781</td>
<td>PTEN_8 to chr7, PTEN_9 to chr9</td>
</tr>
<tr>
<td>PIK3CA</td>
<td>466</td>
<td>2383</td>
<td></td>
</tr>
<tr>
<td>PTEN</td>
<td>445</td>
<td>2298</td>
<td></td>
</tr>
<tr>
<td>TP53</td>
<td>382</td>
<td>1585</td>
<td></td>
</tr>
</tbody>
</table>

25bp primer sequence
Next-gen PCR capture: Fluidigm

How to design Fluidigm/Illumina amplicon resequencing projects:
TP53: Designed primer sets for all Exons at 100, 150 and 200bp length
Amplify cell lines (known TP53 status), mixed samples (sensitivity), FFPE and WGA (clinical utility)
Single sequencing run PE150bp
How do overlapping reads improve error rate estimates

A novel method of Sanger sequencing template prep!

How does run type affect quality of results?

<table>
<thead>
<tr>
<th>Samples</th>
<th>Amplicons</th>
<th>Coverage</th>
<th>Cost/sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>25M</td>
<td>£1405</td>
</tr>
<tr>
<td>8</td>
<td>60</td>
<td>52000x</td>
<td>£212</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>10000x</td>
<td>£37</td>
</tr>
<tr>
<td>50</td>
<td>500</td>
<td>1000x</td>
<td>£90</td>
</tr>
<tr>
<td>500</td>
<td>50</td>
<td>1000x</td>
<td>£7.30</td>
</tr>
<tr>
<td>10000</td>
<td>BRCA</td>
<td>50x</td>
<td>£4.33</td>
</tr>
</tbody>
</table>
Is there a clinical need?
What are the clinical requirements; samples types, single-plex vs multip-plex, turnaround, analysis, ethics, cost, centralisation vs distributed

Oncomap
COSMIC
Cancer Gene Census
Current UK molecular tests
DMuDB

CRUK
NHS
EU (whom)
Technology providers (Illumina, etc)
Patients
Governments

TP53 IARC database
http://www.genet.sickkids.on.ca
Communities

http://pathogenomics.bham.ac.uk/hts

Thanks to Nick Loman!
Acknowledgements

CRI Genomics
Sarah Moffatt
Claire Fielding
Sara Sayalero
Nik Matthews (LRI)
Christian Bourne (MRC hub)

CRI Bioinformatics
Matt Eldridge
Rory Stark
Kevin Howe
Ben Davis

Duncan Odom Group
Michael Wilson
Dominic Schmidt
Gord Brown

Illumina
Mark Gibbs

Fluidigm
Andy May