“Sequencing cancer genomes”

Phil Stephens

Wellcome Trust Sanger Institute
Breast, Lung, Colorectal cancers etc.....

Exposure
UV light
Tobacco smoke

Genomic instability
Genetic Instability

Pre-cancerous lesion
\textit{in situ} cancer
Invasive cancer
Metastatic cancer

\sim 10 \textbf{‘Driver mutations’}

1000s \textbf{‘Passenger Mutations’}
Cancer genome harbour multiple classes of mutation

- **Base Substitutions**
  - Tumour
  - Blood

- **Insertions deletions**
  - Tumour
  - Blood

- **Copy number changes**
  - MYC
  - Copy number
  - Chromosome 8
  - 127.6 Mb - 129.1 Mb

- **Rearrangements**
  - Interchromosominal
  - Intrachromosominal
Illumina Genome Analyser
Sequencing by synthesis

Five years of data per day vs capillary sequencing

Detect all classes of mutation in a single experiment

Screen the entire genome for mutations

Screen targeted regions: Coding exons
Comprehensive catalogue of somatic mutations from a malignant melanoma genome

- 33,345 somatic substitutions
- 88% sensitivity
- 97% specificity
- 31 Somatic rearrangements
- 41 Copy number changes
- 36 Small insertions & deletions
Substitution mutation spectrum

**Total Mutations**

- C>T common due to UV light exposure
- C>A 2%

**Early Mutations**

- C>T 82%
- C>A 2%

**Late Mutations**

- C>T 53%
- C>A 19%

P < 0.0001
Distribution of somatic substitutions

Somatic mutations in genic footprints

Expected = 40%
Observed = 29%
Transcription-coupled repair

Stalled RNA polymerase

Bulky adduct on transcribed strand

mRNA
Strand bias: Somatic C>T substitutions

Number of mutations

P < 0.0001
Strand bias: ALL somatic substitutions

Number of mutations

P < 0.0001
Transcription coupled repair only accounts for only a third of the deficit of mutations over protein coding gene footprints.
Novel ‘expression linked’ repair mechanism

Effect of strand of mutation: p<0.0001

Effect of expression level: p<0.0001

C>T non-transcribed strand

C>T transcribed strand

Expression levels (log₂)

Mutation prevalence (per Mb)
Novel ‘expression linked’ repair mechanism

Effect of strand of mutation: $p<0.0001$

Effect of expression level: $p<0.0001$

C>T non-transcribed strand

C>T transcribed strand

Mutation prevalence (per Mb)

Expression levels (log$_2$)
Feedback for our first cancer genome
Objections to sequencing cancer genomes:

1) “Cancer is too heterogeneous for sequencing to uncover therapeutically relevant mutations”.

2) “All of the major cancer genes/pathways have already been discovered”

3) “Sequencing large numbers of cancer genomes is not the best use of the technology, and will have very limited clinical use”

4) “Sequencing is unlikely to provide insights into metastasis, the dominant cause of mortality in cancer”
Objections to sequencing cancer genomes:

1) “Cancer is too heterogeneous for sequencing to uncover therapeutically relevant mutations”.

2) “All of the major cancer genes/pathways have already been discovered”

3) “Sequencing large numbers of cancer genomes is not the best use of the technology, and will have very limited clinical use”

4) “Sequencing is unlikely to provide insights into metastasis, the dominant cause of mortality in cancer”
BRAF is an oncogene 50% of Melanoma

BRAF is not important Melanoma

BRAF is not a good drug target

Melanoma is too complex to treat effectively
Selective inhibitor of BRAF

(Plexxicon 4032)

Before

15 days after

Courtesy of Dr Grant McArthur
Similar results are being seen with:

- **PARP inhibitors**: BRCA1/BRCA2 null breast cancer
- **EGFR inhibitors**: EGFR mutated lung cancer
- **ABL inhibitors**: BCR/ABL positive CML
Objections to sequencing cancer genomes:

1) “Cancer is too heterogeneous for sequencing to uncover therapeutically relevant mutations”.

2) “All of the major cancer genes/pathways have already been discovered”

3) “Sequencing large numbers of cancer genomes is not the best use of the technology, and will have very limited clinical use”

4) “Sequencing is unlikely to provide insights into metastasis, the dominant cause of mortality in cancer”
Cancer Exome Sequencing

Investigate all protein coding exons/miRNA’s simultaneously

- Protein coding genes 21,416
- Micro RNAs 1664

Can process large numbers of samples

High sensitivity in primary tumours (normal contamination)
18 samples from a common epithelial tumour

9/18 had a truncating mutation in a novel TSG

136/388 follow up tumours had truncating mutations in same gene

Potentially ‘druggable’ using a cheap, approved, existing cancer drug
Cancer Exome Sequencing

3 × ER$^\text{Negative}$ Breast cancers

25 × ER$^\text{Positive}$ Breast cancers
Numbers of somatic substitutions per sample

ER negative

Mean 75, (range 29-100)
Numbers of somatic substitutions per sample

Mean 19.7, (range 1-45)
## Somatic subs in known breast cancer genes

<table>
<thead>
<tr>
<th>Sample</th>
<th>Gene</th>
<th>Mutation</th>
</tr>
</thead>
<tbody>
<tr>
<td>PD3995a</td>
<td>AKT1</td>
<td>E17K</td>
</tr>
<tr>
<td>PD3995a</td>
<td>NF1</td>
<td>G-1T</td>
</tr>
<tr>
<td>PD3994a</td>
<td>PIK3CA</td>
<td>N345K</td>
</tr>
<tr>
<td>PD3989a</td>
<td>PIK3CA</td>
<td>E545K</td>
</tr>
<tr>
<td>PD3856a</td>
<td>PIK3CA</td>
<td>H1047R</td>
</tr>
<tr>
<td>PD3857a</td>
<td>PIK3CA</td>
<td>H1047R</td>
</tr>
<tr>
<td>PD3888a</td>
<td>PIK3CA</td>
<td>H1047R</td>
</tr>
<tr>
<td>PD3983a</td>
<td>PIK3CA</td>
<td>H1047R</td>
</tr>
<tr>
<td>PD3985a</td>
<td>PIK3CA</td>
<td>H1047R</td>
</tr>
<tr>
<td>PD3992a</td>
<td>PIK3CA</td>
<td>H1047R</td>
</tr>
<tr>
<td>PD3996a</td>
<td>PTEN</td>
<td>Y27D</td>
</tr>
<tr>
<td>PD3991a</td>
<td>TP53</td>
<td>G245S</td>
</tr>
<tr>
<td>PD4002a</td>
<td>TP53</td>
<td>H179Y</td>
</tr>
<tr>
<td>PD3987a</td>
<td>TP53</td>
<td>Y220C</td>
</tr>
<tr>
<td>PD3986a</td>
<td>TP53</td>
<td>G+1A</td>
</tr>
<tr>
<td>PD3985a</td>
<td>TP53</td>
<td>R306X</td>
</tr>
</tbody>
</table>
Somatic indels in known breast cancer genes

<table>
<thead>
<tr>
<th>Sample</th>
<th>Gene</th>
<th>Mutation</th>
</tr>
</thead>
<tbody>
<tr>
<td>PD3849a</td>
<td>CDH1</td>
<td>V193X</td>
</tr>
<tr>
<td>PD3984a</td>
<td>MAP2K4</td>
<td>V151X</td>
</tr>
<tr>
<td>PD3992a</td>
<td>MAP2K4</td>
<td>I81X</td>
</tr>
<tr>
<td>PD3989a</td>
<td>PTEN</td>
<td>L370X</td>
</tr>
<tr>
<td>PD3995a</td>
<td>GATA3</td>
<td>N352X</td>
</tr>
<tr>
<td>PD3988a</td>
<td>GATA3</td>
<td>N352X</td>
</tr>
<tr>
<td>PD4004a</td>
<td>GATA3</td>
<td>Read through</td>
</tr>
</tbody>
</table>
638 somatic substitutions
38 small INDELS
Mutations in a ‘novel pathway’
in >60% of ER⁺ breast cancer
Several potential new susceptibility alleles are being further evaluated.

**Exome sequencing & susceptibility alleles**

- **CHEK2**
  - 
- **BRCA1**
  - W411X
  - 4 bp deletion

Several potential new susceptibility alleles are being further evaluated.
Objections to sequencing cancer genomes:

1) “Cancer is too heterogeneous for sequencing to uncover therapeutically relevant mutations”.

2) “All of the major cancer genes/pathways have already been discovered”

3) “Sequencing large numbers of cancer genomes is not the best use of the technology, and will have very limited clinical use”

4) “Sequencing is unlikely to provide insights into metastasis, the dominant cause of mortality in cancer”
Rearrangement characterisation
Breast cancer karyotype

73 Chromosomes, 37 structural abnormalities
Breast cancer HCC38 (Triple neg)

238 somatic structural variants

Patterns of variation
What are these structural variants doing?
164 kb tandem duplication

Copy number

Chromosome 3 position (Mb)

SLC26A6/PRKAR2A in frame fusion gene

SLC26A6
(Exons 1-17)

PRKAR2A
(Exons 4-11)

FISH confirmation of tandem duplication

Genomic PCR

bps

600

400

200

164 kb tandem duplication
RT-PCR

Predicted 914 amino acid fusion protein

MGLADASGPRTQALLSATQAMDLRRRDYHMERPLLNLQEHLEELGRWGSAFTPRTWQRTWLQCSRARAYALLLQLHLPVLVWLPVRDPWVLDGLLSSGL
SAIMQLQPQLAYALLAALPVPFGLYSSFYPVFIYFLFSGRSKHISVGTFAVMSVMVGSVTESLAPQALKDNSMINETARDAARVQVASTLSVLGLQVGGLIH
FGFVVTYLSLVTYTTAAAVQFVSQKLFVGLHLSSGSPILTYTVLECWKLPQSKVGTVTAAVAVGVLVLVVVLKLLNDKLQQLQQILMPMPGEELTLIGAT
GISYGMGLKHRFEVDVGNIPAGLVPVAPNQLFSKLVGSAFTIAVAVFAIAISLGKIFALRHGYRVDSQELVALGLSLNLIGGIFCFQPVSCSMRSLSVQEST
GGNSQVAGAISLFLILLIVKLLGLFHLDPKAVLAIIIVNLGLMLRLQSLDMRSWLANRADLLIWLVFTATIILNLDLGLVLVAVIFSLLLVVVRTQMPHYSGQ
VPDTDIYRDVAEYSEAKEVRGKVFSSATVYFANAEYSDDALKQRCGVDVDLISQKKKLKQLKQEQLKLQKQEEKLQAAKPGASVINSNVTSLEDMDR
SNVNEDECKMVIHPKTDQRCRLQECACKDILFKNLQEQLSQVLDMFERIKADEHVIDQGDGGDNFYYIERQTYDLVTKDNQTRSVQYGDNRSFGEL
ALMYNTPRAAVTSEGSLWGLDRVFRRRIIVKNAAKKRMFESFIESVPLKSLEVSTEMKIDVIGEIKYKDDGERIITQGKSHOWFIIIESGEVSLIRTSKSN
KDGNQVEVIIARCHKGQYFGEALVHTNKPRAASAYAVGVDVCLKLMDVQAFFERLLGPCMDIMKRNHSHYEELVLMKFSSVDLGNLGGQStop
Five expressed in frame fusion genes

2 generated by tandem duplications

3 generated by large inversions
Potential applications in healthcare
Selective inhibitor of BRAF V600E
(Plexxicon 4032)

Before

15 days after

Courtesy of Dr Grant McArthur
Personalised Haematology

Accurate ascertainment of tumour burden in response to chemotherapy

Risk stratification

Therapy duration

Prediction of relapse
Solid tumour rearrangement screens

~two hundred tumours and identified at least one somatically acquired rearrangement in sample (range 1-245)
Identify tumour-specific rearrangements from cancer DNA

Design & test quantitative assay for rearrangements

Extract DNA from plasma (10-20mL blood sample)

Quantify tumour DNA

2 weeks ~£2,500

1 week £250

½ day £50

½ day £50
Detecting 1 copy of tumour genome

DNA quantity / reaction (pg)

Real-time PCR quantification (Ct)

Amount of DNA in one human cell

Rearr 1, reaction 1
Rearr 1, reaction 2
Rearr 2, reaction 1
Rearr 2, reaction 2
Potential healthcare applications

• Monitoring tumour response to therapy in real-time
  – Reduce toxicity, prevent drug wastage

• Identifying disease relapse before clinically evident
  – Pre-emptive therapy

• Choosing intensity of adjuvant therapy based on risk stratification

• Surrogate marker of cell kill in early phase clinical trials
Objections to sequencing cancer genomes:

1) “Cancer is too heterogeneous for sequencing to uncover therapeutically relevant mutations”.

2) “All of the major cancer genes/pathways have already been discovered”

3) “Sequencing large numbers of cancer genomes is not the best use of the technology, and will have very limited clinical use”

4) “Sequencing is unlikely to provide insights into metastasis, the dominant cause of mortality in cancer”
Dynamics of metastasis in pancreatic cancer

Patient 1

PCR genotype multiple metastases

Patient 5

PCR genotype multiple metastases

Patient 9

PCR genotype multiple metastases
Dynamics of metastasis in pancreatic cancer

- Shared in all metastases
- Partially shared
- Unique to index metastasis
Inter-individual differences

- **Private**
- **Partially shared**
- **Omnipresent**

Bar chart showing the distribution of different categories for each individual.
Phylogenetic relationships between metastases within same patient

Rearrangements 1-56
Evidence for convergent evolution
Organ-specific signature
Dynamics of genomic instability and metastasis

Primary tumour
Capacity to initiate metastasis
Secondary metastasis
Tertiary metastasis

Molecular time
Conclusions

1) “Although cancer is heterogeneous, sequencing is an effective tool to uncover therapeutically relevant mutations”.

2) “There are major cancer genes and pathways yet to be discovered”

3) “Sequencing large numbers of cancer genomes has massive clinical utility and will transform the way that cancer is treated and patients are managed”

4) Sequencing can generate ‘dramatic insights’ into modes and mechanisms of metastasis
Mike Stratton
Andy Futreal
Peter Campbell
Dave McBride
Meng-lay Lin
Patrick Tarpey
Stuart Mclaren
Keiran Raine
Dave Jones
Ignacio Varela
Laura Mudie
Lucy Stebbings
Catherine Leroy
Jon Teague
Mike Quail
Dan Turner
Harold Swerdlow
David Bentley
Keira Cheetham
Mark Ross