Sequencing Cancer Genomes and Transcriptomes

Dr. Richard Moore

Sequencing Group Leader

Genome Sciences Centre, BC Cancer Agency
Capacity at the GSC

- 11 Illumina GA II
- 2 AB SOLiD
- 8 3730 xls

- Paired and Single end reads
- Production 51 bases, Tech D upto 101

- Library Types
  Whole Genome Shotgun
  miRNA
  RNA seq / WTSS
  ChIP
  Bisulfite
  MRE
  Small RNA
  MeDIP
  SAGE
Total 1.6 Trillion bases
GSC Utilisation of UHTS technology

- 26 UHTS publications
  - Including Genome research and Nature Methods
- 1.6 trillion bases produced
- >1200 libraries
Examples of Current Projects

• Lobular Breast Cancer
• Adenocarcinoma - personal genomics
• Follicular Lymphoma

• Diffuse Large Bcell Lymphoma
• Ovarian Cancer
• Lung Cancer
• Acute Lymphoblastic Leukaemia
• Neuroblastoma
• Prostate Cancer
• Oligodendrioma
• Colo-rectal Carcinoma
Cancer prevalence

Life time probability of:

<table>
<thead>
<tr>
<th></th>
<th>developing</th>
<th></th>
<th>dying</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>%</td>
<td>one in:</td>
<td>%</td>
<td>one in:</td>
</tr>
<tr>
<td>Females</td>
<td>39.3</td>
<td>2.5</td>
<td>24.1</td>
<td>4.2</td>
</tr>
<tr>
<td>Males</td>
<td>44.5</td>
<td>2.2</td>
<td>28.5</td>
<td>3.5</td>
</tr>
</tbody>
</table>

Canadian Cancer Society Statistics, 2008
Why sequence cancer genomes and transcriptomes?

- To identify cancer driver mutations and pathways.
  - Challenges: passengers vs. drivers, undiscovered SNPs, changes selected for in culture, genetic heterogeneity of cell populations...

- To identify targets for development of new therapies.

- To improve diagnostic precision and prognostic accuracy.
  - E.g. “breast cancer” describes several diseases.

- To match patients to treatments.
  - Optimize treatment modalities based on individual genes and genomes

- To understand differences in treatment response.
  - Outright failure
  - Remission / relapse (treatment resistance).

- To manage treatment failure.
  - Alternative existing therapies?
Next-generation sequencing of cancer patient samples

- Depth-dependent sensitivity to rare events; e.g. genomic changes in cellular sub-populations (tumor initiating cells?).
  - Enabled by massively redundant “single molecule” sequencing, in contrast to technologies that integrate signal from many molecules.
Lobular Breast Cancer
Lobular Breast Cancer

**Primary Tumor**
- DNA extraction library construction sequencing
  - MAQ alignment
  - 8.04 x 10^9 raw reads
  - 4.98 x 10^9 mapped reads
  - CNAs
    - Table S7
  - SNVs
    - Table S1
  - Indels
    - Table S2
  - FISH
    - Table S8
  - Affx SNP 6
  - Sanger sequencing

**Pleural Effusion Tumor**
- DNA extraction library construction sequencing
  - MAQ alignment
  - 8.89 x 10^9 raw reads
  - 7.91 x 10^9 mapped reads
  - CNAs
    - Table S7
  - SNVs
    - Table S1
  - Indels
    - Table S2
  - FISH
    - Table S8
  - Sanger sequencing

**RGSS-PE**
- Expression
  - Table S9
- SNVs
  - Table S1
- Indels
  - Table S2
- Alt splicing
  - Table S4
- Sanger sequencing

**Sequence data**
- Mapped data
- Biological features
- Validation assays
- Combined inference from multiple data sources
- Validated events

**Progression events**
- Unequal allelic expression
  - Table S5
- RNA edits
  - Table 4

**Results**
- 291 novel germline SNVs
  - 19 somatic point mutations
  - 6 confirmed CNAs
- 160 novel germline SNVs
  - 11 somatic point mutations
# Samples and libraries

<table>
<thead>
<tr>
<th></th>
<th>WGSS-PE</th>
<th>WTSS-PE</th>
<th>WGSS-PRI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of reads</td>
<td>889,392,298</td>
<td>182,532,650</td>
<td>804,148,860</td>
</tr>
<tr>
<td>Total nucleotides (Gb)</td>
<td>34.419</td>
<td>7.108</td>
<td>35.511</td>
</tr>
<tr>
<td>Number of aligned reads</td>
<td>790,665,100</td>
<td>160,919,484</td>
<td>497,521,910</td>
</tr>
<tr>
<td>Aligned nucleotides (Gb)</td>
<td>30.599</td>
<td>6.266</td>
<td>21.971</td>
</tr>
<tr>
<td>Estimated error rate</td>
<td>0.015</td>
<td>0.013</td>
<td>0.019</td>
</tr>
<tr>
<td>Estimated depth (non-gap regions)</td>
<td>9.809</td>
<td>N/A</td>
<td>6.776</td>
</tr>
<tr>
<td>Canonically aligned reads</td>
<td>721,929,588</td>
<td>109,093,616</td>
<td>439,409,649</td>
</tr>
<tr>
<td>Percent reads aligned canonically</td>
<td>91.31</td>
<td>67.79</td>
<td>88.32</td>
</tr>
<tr>
<td>Unaligned reads</td>
<td>98,727,198</td>
<td>21,613,166</td>
<td>306,626,950</td>
</tr>
<tr>
<td>Mean read length (bp)</td>
<td>38.7</td>
<td>38.94</td>
<td>44.16</td>
</tr>
</tbody>
</table>
Summary

• Alternate splicing shown in the estrogen signalling pathway

• 25 NS somatic mutations

• 16 only in the metastatic tumour

• 2 RNA editing somatic NS mutations

• Novel amplicon region in the Insulin receptor
Follicular Lymphoma

BAC re-sequencing of large scale rearrangements
Rearrangement confirmation and clone sequencing pipeline

- Clones with candidate FPP rearrangements
  - Select clones for each rearrangement
  - BES + Alternate fingerprint
  - Rearrangements confirmed by ≥2 clones (721 total)
  - Select a representative clone(s) for each rearrangement

- Data processing and analysis
  - Sequence each pool on a single lane on a flow cell (Illumina GA2)
  - Sequencing library created for each pool

- Divide into pools of 8 clones
  - 95 pools
  - 727 clones

- Group clones by patient

- 691 of the 721 total rearrangements were sampled
- All 253 distinct rearrangements were sampled in at least one patient
  ⇒ Not all recurrent rearrangements sampled in all patients

© BC Cancer Agency Branch
Summary of validated large-scale somatic events

We have identified and validated 52 somatic and 66 germline breakpoints, corresponding to 38 and 41 distinct events. Examples of somatic mutations include:

<table>
<thead>
<tr>
<th>Event type</th>
<th>Chromosome band(s)</th>
<th>Event size (kb)</th>
<th>Patient ID</th>
<th>Junction feature</th>
<th>Affected gene(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEL</td>
<td>1p22.3</td>
<td>558</td>
<td>8</td>
<td>Microhomology</td>
<td>LPAR3, MCOLN2, MCOLN3, WDR63, SYDE2, C1orf52, BAG and last exon of BCL10</td>
</tr>
<tr>
<td>DEL</td>
<td>9p24.1</td>
<td>251</td>
<td>10</td>
<td>Blunt</td>
<td>PTPRD</td>
</tr>
<tr>
<td>DEL</td>
<td>9p21.3</td>
<td>124</td>
<td>16</td>
<td>Microhomology</td>
<td>CDKN2A</td>
</tr>
<tr>
<td>DEL</td>
<td>2p11.2</td>
<td>193-412</td>
<td>8, 13, 16, 17</td>
<td>All features</td>
<td>AbParts, IGK</td>
</tr>
<tr>
<td>DEL</td>
<td>22q11.22</td>
<td>112-861</td>
<td>6, 17</td>
<td>Sequence additions</td>
<td>AbParts, IGL</td>
</tr>
<tr>
<td>DEL</td>
<td>14q32.33</td>
<td>395</td>
<td>19</td>
<td>Sequence additions</td>
<td>AbParts, IGH</td>
</tr>
<tr>
<td>TRX</td>
<td>t(14;18)(q32;q21)</td>
<td>NA</td>
<td>6, 8, 10, 13, 16</td>
<td>Sequence additions</td>
<td>IGH and BCL2</td>
</tr>
<tr>
<td>TRX</td>
<td>t(10;12)(q25.1;q23.1)</td>
<td>NA</td>
<td>10</td>
<td>Microhomology</td>
<td>SORCS1</td>
</tr>
<tr>
<td>TRX</td>
<td>t(6;7)(q15;q36.1)</td>
<td>NA</td>
<td>20</td>
<td>Microhomology</td>
<td>CDK5</td>
</tr>
<tr>
<td>TRX</td>
<td>t(1;3)(q23.3;p24.3)</td>
<td>NA</td>
<td>21</td>
<td>Blunt</td>
<td>RFTN1</td>
</tr>
<tr>
<td>INV</td>
<td>3q27.3</td>
<td>718/722</td>
<td>14, 21</td>
<td>All features</td>
<td>BCL6 and ST6GAL1</td>
</tr>
<tr>
<td>INV</td>
<td>6q25.2</td>
<td>80</td>
<td>21</td>
<td>Microhomology</td>
<td>OPRM1</td>
</tr>
<tr>
<td>DUP</td>
<td>10q22.2-10q23.32</td>
<td>18,126</td>
<td>19</td>
<td>Sequence additions</td>
<td>AP3M1 and BTAF1</td>
</tr>
<tr>
<td>RNG</td>
<td>4q12</td>
<td>137</td>
<td>20</td>
<td>Sequence additions</td>
<td>KIT</td>
</tr>
</tbody>
</table>

DEL, Deletion; TRX, Translocation; INV, Inversion; DUP, Duplication; RNG, Complex rearrangement; NA, Not applicable
Fork Stalling and Template Switching - a model for nucleotide additions at breakpoint junctions

A DNA Replication Mechanism for Generating Nonrecurrent Rearrangements Associated with Genomic Disorders

Jennifer A. Lee, Claudia M.B. Carvalho, and James R. Lupski

Department of Molecular and Human Genetics
Department of Pediatrics
Baylor College of Medicine, Houston, TX, 77030, USA

Texas Children’s Hospital, Houston, TX, 77030, USA

Correspondence: jjlupski@bcm.tmc.edu
DOI: 10.1016/j.cell.2007.11.037

Cell. 2007 Dec 28;131(7):1235-47
Concluding remarks

- To date we have identified 38 distinct somatic events from 20 primary follicular lymphoma patient tumour samples.
- 40% of the rearrangement events assayed are acquired in the tumour samples.
- Having sequence level resolution of breakpoints is informing on mechanistic insights.
- For the first time in cancer, to our knowledge, we have demonstrated that sequence additions at rearrangement breakpoints are consistent with the FoSTeS model of DNA replication.
Personal cancer genomics:

Adenocarcinoma, pulmonary metastases - non-responsive to treatment (erlotnib)
Questions

• In this case (rare tumor, no standard chemotherapy options), can next generation genome analysis provide clues as to the expression and mutational profiles of known drug targets?

• Can next generation genome analysis provide insight into the apparent resistance of the tumor to erlotinib, despite apparent amplification of EGFR?
## Samples and libraries

<table>
<thead>
<tr>
<th>Type</th>
<th>Source</th>
<th>Amount</th>
<th>Total reads</th>
<th>% aligned to genome</th>
<th>Sequence yield total (base-pairs)</th>
<th>Sequence yield aligned (base-pairs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNA-seq</td>
<td>Normal: Blood</td>
<td>10ug Total RNA</td>
<td>85,795,100</td>
<td>72.9</td>
<td>3,603,394,200</td>
<td>2,625,754,824</td>
</tr>
<tr>
<td>Genome Shotgun</td>
<td>Normal: Blood</td>
<td>10ug gDNA</td>
<td>40,793,289</td>
<td>83.7</td>
<td>1,713,318,138</td>
<td>1,434,873,300</td>
</tr>
<tr>
<td>RNA-seq lite (cDNA amp)</td>
<td>Cancer: Lung Biopsy</td>
<td>10ng Total RNA</td>
<td>27,897,5099</td>
<td>69.2</td>
<td>11,716,954,158</td>
<td>8,104,150,782</td>
</tr>
<tr>
<td>RNA-seq lite (cDNA amp)</td>
<td>Cancer: Lung Biopsy</td>
<td>10ng Total RNA</td>
<td>431,797,605</td>
<td>69.7</td>
<td>18,135,499,410</td>
<td>12,645,184,944</td>
</tr>
<tr>
<td>RNA-seq lite (RNA amp)</td>
<td>Cancer: Lung Biopsy</td>
<td>2ng Total RNA</td>
<td>6,328,326</td>
<td>66.3</td>
<td>265,789,692</td>
<td>176,282,652</td>
</tr>
<tr>
<td>Genome Shotgun</td>
<td>Cancer: Lymph Node FFPE</td>
<td>10ug gDNA</td>
<td>2,942,711,791</td>
<td>87.8</td>
<td>123,593,895,222</td>
<td>108,551,254,728</td>
</tr>
</tbody>
</table>

**Totals:**

Total reads: 3,786,401,210
Sequence yield total (base-pairs): 159,028,850,820
Sequence yield aligned (base-pairs): ~84%
Data analysis

• Transcriptome libraries:
  - Digital gene expression profiling, comparing tumor read counts across ~23 in-house WTSS libraries.
  - Mutation detection.

• Genome libraries:
  - Digital karyotyping (copy number analysis), comparing tumor genome sequence to PB ("normal") genome sequence and the Yoruban sequence.
  - Mutation detection.

• Integration with drug bank:
  - Relate mutations, copy number alterations, and gene expression data to known drugs.
Drug Bank

- [www.drugbank.ca](http://www.drugbank.ca)
- 4,408 drugs in drugbank for 4535 targets.
- 1,359 FDA-approved drugs for 1613 targets
- 189 "cancer" drugs for 291 targets.
Intersection of drug bank with genomic features and cancer pathways

- 30 amplified genes in cancer pathways.
- 76 deleted genes in cancer pathways.
- ~400 up- and ~400 down-regulated genes.
  - 19 of these in known cancer pathways
- Single base changes (transcriptome):
  - 303 non-synonymous point mutations in 177 genes
  - 4 genes in cancer pathways
- Other:
  - 233 novel coding SNPs,
  - 25 genes affected by indels
  - 126 genomic deletions spanning exons for 83 genes
  - No intersections with known cancer pathways
Sunitinib was chosen because it targets the most number of expressed genes in the tumour. It inhibits RET, PDGFRA, CSF1R.

Sorafenib is also able to inhibit a mutant RET and B-RAF, and was thus selected as secondary treatment in case the patient did not respond or acquired resistance to sunitinib.
Integrating pathways, mutations and gene expression

- **PDGFR**
- **EGFR**
- **RET**
- **TNFR**
- **SHC**
- **GRB**
- **SOS**
- **PI3K**
- **PTEN**
- **AKT**
- **mTOR**
- **BAD**
- **eIf4EBP**
- **p70S6K**
- **eIf4E**
- **Rb1**
- **AP-1**
- **TRAF2**
- **TRAF1**
- **JNK**
- **MEK**
- **ERK**

- **amplified**
- **LOH deleted**
- **mutated**
- **upregulated**
- **downregulated**

**Pathways and Processes**

- **Apoptosis**
- **Growth**
- **Proliferation**
Response to Sunitinib

After 29 days on Erlotinib tumours increased in size by >20%
After 28 days on Sunitinib tumour reduced in size by >20%
Infectious Agents and Cancer
Genomic Subtraction For Infectious Agent Detection

Tumour Biopsy

- Total RNA extraction
  - PolyA selection
  - Ribominus depletion

- cDNA fragment library construction 150-300bp

- Sequencing on Illumina GA
10 million to 100 million reads
Map unmapped reads to human genome (hg18)
Map unmapped reads to infectious agent database
Pool hits to same infectious agent and rank candidates by number of hits and type
Assemble viral reads

Infectious Agent specific reads

All Reads
Transcriptome Subtracted reads
Genome subtracted reads
De-novo assembly of unmapped reads reveals viral RNA and integration site

- HeLa is derived from a cervical cancer positive for HPV infection
- De-novo assembly of unmapped reads reveals HPV-derived sequences
- A few HPV contigs contain human sequence (reveal integration site of virus)
Acknowledgements

Bilenky, Misha
Birol, Inanc
Chuah, Eric
Corbett, Richard
Griffith, Obi
Hirst, Martin & team
Holt, Rob
Jones, Steve
Li, Yvonne
Marra, Marco
Moore, Richard & team
Morin, Ryan
Morozova, Olena
Pugh, Trevor
Roscoe, Robyn & team
Shah, Sohrab
Thiessen, Nina
Varhol, Richard & team
Zhao, Yongjun & team

Cancer patients and research advocates
Aparicio, Sam
Huntsman, David
Jones, Steven
Laskin, Janessa

Funding:
Genome Canada
Genome British Columbia
BC Cancer foundation
CIHR
NCI (USA)
NCI (Canada)
Western Economic Diversification