PacBio Long Read Sequencing and Structural Analysis of a Breast Cancer Cell Line

W. Richard McCombie

Disclosures

Orion Genomics – Founder and Shareholder
Cancer epigenetics and plant genomics

Previously Compensated Speaker for Illumina, Inc.

Previously Compensated Speaker for Pacific Biosciences, Inc.
Expressed genes, Alu repeats and polymorphisms in cosmids sequenced from chromosome 4p16.3

W.R. McCloskie¹, A. Martin-Gallardo¹, J.D. Gocayne², M. FitzGerald³, M. Dubnick³, J.M. Kelley¹, L. Camilla¹, L.T. Liu¹, S. Wallace¹, S. Trapp¹, D. Tague¹, W.L. Whaley¹, S. Cheng¹, I. Giusti², A.-M. Franchard³, A. Poole³, H. Lehrach³, F.S. Collins³, A. R. Kerns⁎², C. Field² & I.C. Verstre³

The sequences of three cosmids (90 kilobases) from the Huntington’s disease region in chromosome 4p16.3 have been determined. A 30,837 base overlap of DNA sequences from two individuals was found to contain 72 DNA sequence polymorphisms, an average of 2.3 polymorphisms per kilobase (kb). The assembled 58 kb contig contains 62 Alu repeats and eleven predicted exons representing at least three expressed genes that encode previously unidentified proteins. Each of these genes is associated with a CpG island. The structure of one of the new genes, hda1-1, has been determined by characterizing cDNAs from a placental library. This gene is expressed in a variety of tissues and may encode a novel housekeeping gene.
Evolution of genome assemblies

• Initial references – very high quality – extremely expensive

• Period of lower quality Sanger assemblies (~2001-2007)

• Next gen assemblies (short read) – 2007- now

• Third generation – long read assemblies -2013/2014 –now – what can we do currently?
Her2 amplified breast cancer

Breast cancer

• About 12% of women will develop breast cancer during their lifetimes
• ~230,000 new cases every year (US)
• ~40,000 deaths every year (US)

Her2 amplified breast cancer

• 20% of breast cancers
• 2-3X recurrence risk
• 5X metastasis risk

Statistics from American Cancer Society and Mayo Clinic.
SK-BR-3

Most commonly used Her2-amplified breast cancer cell line

Often used for pre-clinical research on Her2-targeting therapeutics such as Herceptin (Trastuzumab) and resistance to these therapies.

(Davidson et al, 2000)
High quality sequencing with Pac Bio

• Just long read data
• High coverage
• Map back to human reference
• *De novo* assembly
• Characterize the view of the SKBR3 genome presented by different assemblies as well as determine “truth”
• Ongoing project
PacBio read length distribution

- Mean: 9kb
- 72.6X coverage
- 49.3X coverage over 10kb
- 12.0X coverage over 20kb
- Max: 71kb
Improving SMRTcell Performance

- **November 2014**: mean: 6.2kb, yield: 213 Mbp/SMRT cell
- **December 2014**: mean: 8.3kb, yield: 620 Mbp/SMRT cell
- **January 2015**: mean: 9.7kb, yield: 900 Mbp/SMRT cell
- **February 2015**: mean: 11.3kb, yield: 1031 Mbp/SMRT cell
Genome-wide alignment coverage

Genome-wide coverage averages around 54X
Coverage per chromosome varies greatly as expected from previous karyotyping results
PacBio

Her2

Chr 17: 83 Mb

8 Mb
PacBio and Illumina coverage values are highly correlated but Illumina shows greater variance because of poorly mapping reads.
PacBio
67X @ 10kb

Illumina
120X @ 100bp

Repeats
21-mers

Her2

8 Mb
Structural variant discovery with long reads

1. Alignment-based split read analysis: Efficient capture of most events
   BWA-MEM + Lumpy

2. Local assembly of regions of interest: In-depth analysis with base-pair precision
   Localized HGAP + Celera Assembler + MUMmer

3. Whole genome assembly: In-depth analysis including novel sequences
   DNAnexus-enabled version of Falcon

Total Assembly: 2.64Gbp   Contig N50: 2.56 Mbp   Max Contig: 23.5Mbp
Green arrow indicates an inverted duplication. False positive and missing Illumina calls due to mis-mapped reads (especially low complexity).
Confirmed both known gene fusions in this region
Confirmed both known gene fusions in this region
Joint coverage and breakpoint analysis to discover underlying events
By comparing the proportion of reads that are spanning or split at breakpoints we can begin to infer the history of the genetic lesions.

1. Healthy diploid genome
2. Original translocation into chromosome 8
3. Duplication, inversion, and inverted duplication within chromosome 8
4. Final duplication from within chromosome 8
## SKBR3 Oncogene Analysis

### Known missense mutation in p53: **R175H**

<table>
<thead>
<tr>
<th>Gene</th>
<th>Reference Sequence</th>
<th>Illumina Sequence</th>
<th>PacBio Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Arg</strong></td>
<td>ATCTGAGCAGCGCTCATGGTGCGGCGGCTCAACCTCCGTCATGTGCTTGACTGCTT</td>
<td>ATCTGAGCAGCGCTCATGGTGCGGCGGCTCAACCTCCGTCATGTGCTTGACTGCTT</td>
<td>ATCTGAGCAGCGCTCATGGTGCGGCGGCTCAACCTCCGTCATGTGCTTGACTGCTT</td>
</tr>
<tr>
<td><strong>His</strong></td>
<td></td>
<td><strong>T</strong></td>
<td><strong>T</strong></td>
</tr>
</tbody>
</table>

### Oncogene amplifications

<table>
<thead>
<tr>
<th>Oncogene</th>
<th>Approximate Fold Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>ErbB2 (Her2/neu)</td>
<td>≈20X</td>
</tr>
<tr>
<td>MYC</td>
<td>≈27X</td>
</tr>
<tr>
<td>MET</td>
<td>≈8X</td>
</tr>
</tbody>
</table>

### Known Gene fusions

<table>
<thead>
<tr>
<th>Gene Fusion</th>
<th>Confirmed by PacBio reads?</th>
</tr>
</thead>
<tbody>
<tr>
<td>TATDN1-GSDMB</td>
<td>Yes</td>
</tr>
<tr>
<td>RARA-PKIA</td>
<td>Yes</td>
</tr>
<tr>
<td>ANKHD1-PCDH1</td>
<td>Yes</td>
</tr>
<tr>
<td>CCDC85C-SETD3</td>
<td>Yes</td>
</tr>
<tr>
<td>SUMF1-LRRFIP2</td>
<td>Yes</td>
</tr>
<tr>
<td>WDR67 (TBC1D31)-ZNF704</td>
<td>Yes</td>
</tr>
<tr>
<td>DHX35-ITCH</td>
<td>Yes</td>
</tr>
<tr>
<td>NFS1-PREX1</td>
<td>Yes *read-through transcription</td>
</tr>
<tr>
<td>CYTH1-EIF3H</td>
<td>Yes *nested inside 2 translocations</td>
</tr>
</tbody>
</table>

---

**Genetic Lesion History Analysis Underway**
Her2+ Breast Cancer Reference Genome

http://schatzlab.cshl.edu/data/skbr3/

Releasing all data pre-publication to accelerate breast cancer research

Available today under the Toronto Agreement:
• Fastq & BAM files of aligned reads
• Interactive Coverage Analysis with BAM.IOBIO

Available soon
• Whole genome assembly and methylation analysis
• Comparison to single cell analysis of >100 individual cells
Acknowledgments

Maria Nattestad
Sara Goodwin
Timour Baslan
James Gurtowski
Melissa Kramer
Panchajanya Deshpande
Senem Mavruk Eskipehlivan
Eric Antoniou
James Hicks
Michael Schatz

Karen Ng
Timothy Beck
Yogi Sundaravadanam
John McPherson

LUMPY technical support: Ryan Layer and Aaron Quinlan
Thank you!

May the force be with you!

http://schatzlab.cshl.edu/data/skbr3/