Tools for Annotating the Human Genome

Mark B Gerstein
Yale

Slides at Lectures.GersteinLab.org
(See Last Slide for References & More Info.)
Technologies used by the CEGS for Interrogating the Human Genome, over the past decade

**Tiling Arrays**
- Application in a variety of contexts:
  - Transcription Mapping

**Massively Parallel Sequencing**
- DNA binding (inc. chromatin struc.)
- Replication
- Structural Variation

---

800 bp
36mer

PCR Products

Oligonucleotide Array

5 Mb

AGTTACACCTAAGA...
CTTGAATGCCGAT...
GTCATTCCGCAAT...

b02
b04

‘02
‘04
‘06
Projects

- Array Technologies
  - Transcript Mapping: identifying 3’ ends & correlating with DNA binding
  - Array-based approaches for measuring structural variation
  - Methods for optimizing tiling arrays -- in particular, tools for analyzing cross hybridization
  - Human transcription factor DNA microarray

- Sequencing Technologies
  - Methods for ChIP-sequencing technologies
  - Next generation technology for genome characterization – RNA-Seq
  - New sequencing-based methods for analyzing structural variation in the genome – e.g. Paired-End Mapping (PEM)

- Methods for annotating the human genome based on the results of the above experiments
The Cost of DNA Sequencing is Dropping Rapidly: ~10 fold each Year!

Human Genome Sequencing Cost <$4K
The Cost of DNA Sequencing is Dropping Rapidly: ~10 fold each Year!

Human Genome Sequencing Cost <$4K
The Cost of DNA Sequencing is Dropping Rapidly: ~10 fold each Year!

Human Genome Sequencing Cost <$4K
The Cost of DNA Sequencing is Dropping Rapidly: ~10 fold each Year!
The Cost of DNA Sequencing is Dropping Rapidly: ~10 fold each Year!

Human Genome Sequencing Cost < $4K
Dropping Cost of Sequencing is Faster than Moore’s Law

[Sboner et al. Genome Biol. ('11)]
Outline

• Tools for finding & characterizing SVs
  – RD : MSB + CNVnator
  – SR : SRiC, AGE & BreakSeq
• AlleleSeq : Integrating Variation & Func. Genomics
• Platform Comparison
• Test Sample Project Integrating the Technologies
• Tools for Selection : ncVAR
• Tools for RNAseq
  – RSeqTools, FusionSeq, ACT
Main Steps in Genome Resequencing

[Snyder et al. Genes & Dev. ('10), in press]
Methods to Find SVs

1. Paired ends

2. Split read

3. Read depth (or aCGH)

4. Local Reassembly

[Snyder et al. Genes & Dev. (‘10), in press]
Different Approaches Work Differently on Different Events

**Deletions**

<table>
<thead>
<tr>
<th>Method</th>
<th>Indel size (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Split-read analysis</td>
<td>&gt;1 bp</td>
</tr>
<tr>
<td>RP (fosmid)</td>
<td>&gt;8 kb</td>
</tr>
<tr>
<td>RP (454)</td>
<td>&gt;3 kb</td>
</tr>
<tr>
<td>RP (Solexa/SOLiD)</td>
<td>&gt;0.1 kb</td>
</tr>
<tr>
<td>hr-aCGH</td>
<td>&gt;0.5 kb</td>
</tr>
<tr>
<td>dbSNP</td>
<td>1–28 bp</td>
</tr>
</tbody>
</table>

**Insertions**

<table>
<thead>
<tr>
<th>Method</th>
<th>Indel size (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Split-read analysis</td>
<td>1–250 bp</td>
</tr>
<tr>
<td>RP (Solexa/SOLiD)</td>
<td>100–250 bp</td>
</tr>
<tr>
<td>hr-aCGH</td>
<td>&gt;0.5 kb</td>
</tr>
<tr>
<td>RP (454)</td>
<td>2–3 kb</td>
</tr>
<tr>
<td>RP (fosmid)</td>
<td>8–40 kb</td>
</tr>
</tbody>
</table>

[Zhang et al. ('11) *BMC Genomics*]
MSB+CNVnator: Read-Depth Segmentation
Array Signal

Read depth

Individual genome

Reads

Reference genome

Mapping

Counting mapped reads

Read depth signal

Zero level

[Urban et al. ('06) PNAS; Wang et al. Gen. Res. ('09); Abyzov et al. Gen. Res. ('11)]
Mean-shift-based (MSB) segmentation: no explicit model

- For each bin attraction (mean-shift) vector points in the direction of bins with most similar RD signal
- No prior assumptions about number, sizes, haplotype, frequency and density of CNV regions
- Not Model-based (e.g. like HMM) with global optimization, distr. assumption & parms. (e.g. num. of segments).
- Achieves discontinuity-preserving smoothing
- Derived from image-processing applications

[Abzyov et al. Gen. Res. ('11)]
Intuitive Description of MSB

Objective: Find the densest region
Distribution of identical billiard balls

Adapted from S. Ullman et al. “Advanced Topics in Computer Vision,” www.wisdom.weizmann.ac.il/~vision/courses/2004_2
Example of Application of CNVnator to RD data

NA12878, Solexa 36 bp paired reads, ~30x coverage
CNV size distribution

[Abyzov et al. Gen. Res. ('11)]

NA12878, Solexa 36 bp paired reads, ~30x coverage
CNVnator is very accurate

<table>
<thead>
<tr>
<th></th>
<th>CEPH trio</th>
<th>Yoruba trio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>F</td>
</tr>
<tr>
<td>Coverage by mapped reads</td>
<td>~24X</td>
<td>~28X</td>
</tr>
<tr>
<td>Bin size</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Strength for CNV discovery</td>
<td>4.8</td>
<td>4.7</td>
</tr>
<tr>
<td>Strength for CNV discovery</td>
<td>5.4</td>
<td>5.3</td>
</tr>
<tr>
<td>(after GC correction)</td>
<td></td>
<td></td>
</tr>
<tr>
<td># of all calls</td>
<td>3678</td>
<td>3615</td>
</tr>
<tr>
<td># of q0 filtered calls</td>
<td>2352</td>
<td>2223</td>
</tr>
<tr>
<td># of q0 filtered calls, larger</td>
<td>738</td>
<td>737</td>
</tr>
<tr>
<td>and excluding chromosomes X and Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>concordant with M</td>
<td>-</td>
<td>343</td>
</tr>
<tr>
<td>concordant with F</td>
<td>343</td>
<td>-</td>
</tr>
<tr>
<td>concordant with C</td>
<td>471</td>
<td>488</td>
</tr>
<tr>
<td>concordant with M or F</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>FDR for all calls</td>
<td>19%</td>
<td>16%</td>
</tr>
<tr>
<td>FDR for q0 filtered calls</td>
<td>13%</td>
<td>8%</td>
</tr>
<tr>
<td>FDR corrected for reference</td>
<td>6%</td>
<td>3%</td>
</tr>
<tr>
<td>individual bias in CGH</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proportion of calls with</td>
<td>9%</td>
<td>8%</td>
</tr>
<tr>
<td>incorrect breakpoints</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Estimated sensitivity</td>
<td>96%</td>
<td>(90%)</td>
</tr>
</tbody>
</table>
RD works well on a variety of sequencing platforms

[Abyzov et al. Gen. Res. ('11)]
Split Read:
SRiC+AGE
+Breakseq
Split-read Analysis

Reference

Target Genome

Read

Breakpoint

Deletion

Insertion

Target Genome

Reference

Read

Breakpoint

Breakpoint

Breakpoint

Zhang et al. Submitted

More: Breakpoint Assembly

Alt: BreakSeq

BreakSeq
Deletions are the Easiest to Identify

Simple SVs

Complex SVs

[Zhang et al. ('11) BMC Genomics]
SRiC: Split Read Pipeline

Special Pipeline for Pseudogenes: A Abyzov

[Zhang et al. ('11) BMC Genomics]
SV Detection and Genotyping

“BreakSeq” leverages the junction library to detect previously known SVs at nucleotide-level from short-read sequenced genome, which can hardly be achieved by methods such as split-read.

Reference Genome

Alternative Junctions of an Insertion

Junction A

Alternative Junction of a Deletion

Junction C

Read

or

Read

Map reads onto

Library of SV breakpoint junctions

Junctions can be put on a chip

60 bp

60 bp

60 bp

* Read overlaps <10 bp to one side of the breakpoint is discarded and read matches also to the reference genome is classified as non-unique match

[Lam et al., (‘10) Nat. Biotech.]
SV Breakpoint Library

[Image: Pie chart showing percentages of different breakpoint libraries: Levy 35%, Wheeler 31%, Korbel 10%, Mills 5%, Kim 3%, Perry 1%, Tuzun 2%]

[Image: Diagram showing the process of generating junction sequences from SV deletion (or insertion) and resulting in the Library of SV breakpoint junctions]

[Lam et al., ('10) Nat. Biotech.]
SVs with sequenced breakpoints

[1KG Project >20,000]

[Published BreakSeq Library]

[Lam et al., ('10) Nat. Biotech.]
Validation for Identified SVs

48 positive outcomes out of 49 PCRs that were scored in NA12891:
98% PCR validation rate (for low and high-support events)
12 amplicons sequenced in NA12891: all breakpoints confirmed

<table>
<thead>
<tr>
<th>Personal genome (ID)</th>
<th>Ancestry</th>
<th>High support hits (&gt;4 supporting hits)</th>
<th>Total hits (incl. low support)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA18507*</td>
<td>Yoruba</td>
<td>105</td>
<td>179</td>
</tr>
<tr>
<td>YH*</td>
<td>East Asian</td>
<td>81</td>
<td>158</td>
</tr>
<tr>
<td>NA12891 [1000 Genomes Project, CEU trio]</td>
<td>European</td>
<td>113</td>
<td>219</td>
</tr>
</tbody>
</table>

[Lam et al., ('10) Nat. Biotech.]
SR Calibration
Using Simulation to Parameterize SRiC: Deletions Easier than Insertions

[Zhang et al. ('11) BMC Genomics]
Using Simulation to Parameterize SRiC: Coverage & Read Length

Deletion

<table>
<thead>
<tr>
<th>Read length (bp)</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>400</th>
<th>800</th>
</tr>
</thead>
<tbody>
<tr>
<td>False positives</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>False negatives</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True positives, deletion length (bp)</td>
<td>1</td>
<td>2</td>
<td>6</td>
<td>11</td>
<td>51</td>
</tr>
<tr>
<td>All length</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Insertion

<table>
<thead>
<tr>
<th>Read length (bp)</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>400</th>
<th>800</th>
</tr>
</thead>
<tbody>
<tr>
<td>False positives</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>False negatives</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True positives, insertion length (bp)</td>
<td>1</td>
<td>2</td>
<td>6</td>
<td>11</td>
<td>21</td>
</tr>
<tr>
<td>All length</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A

Different read length

Percentage of called deletions

B

Different read length

Percentage of called insertions

C

Different coverage

Percentage of called deletions

D

Different coverage

Percentage of called insertions

[Zhang et al. ('11) BMC Genomics]
What is the problem?

[Abyzov & Gerstein ('11) Bioinfo.]
More problems
(homology around breakpoints)

Where are breakpoints?

[Byzov & Gerstein (’11) Bioinfo.]
Problem definition

• Given scoring scheme (match, mismatch, gap open, gap extend) find an optimal alignment of two sequences (i.e. with highest score) where **ONE gap is **NOT** penalized**

• Each sequence end aligns independently and one **need to find an optimal place to introduce a non-penalized gap**

[Abyzov & Gerstein ('11) Bioinfo.]
AGE Alignment with Gap Excision

[Abyzov & Gerstein ('11) Bioinfo.]
Mechanism Assignment Pipeline
SV Mechanism Classification

NAHR

Highly similar with minor offset

Single RETRO

Multiple RETRO

[Laughlin et al., (‘10) *Nat. Biotech.*]
SV Mechanism Classification

1 kb ≤ SV ≤ 1 Mbp

Has flanking sequences

yes

Has extensive coverage by VNTR regions

yes

Annotate SV and flanking regions by RepeatMasker

yes

Extract a window at each breakpoint and align the two sequences

no

Two sequences share high similarity; Homologous regions have minor offsets, correct orientations and span the breakpoints

yes

NAHR

no

Unclassified

SV region covered by a single TE

no

Potential processed pseudogene and other ambiguous cases

yes

Has a poly-A tail and TSD

SV region covered by multiple successive TEs

no

NHR

yes

Annotated as fragments from a single TE

MTEI

no

STEI

report


[2] Other 69%

[3] Reported active L1 16%

[4] Putative novel active L1 13%

[5] Potential processed pseudogene 1%

[6] NHR 28%

[7] MTEI 5%

[8] STEI 6%

[9] VNTR 2%

[10] Other ambiguous 2%
SV Ancestral State Analysis

Inferring Insertion according to Ancestral State

Inferring Deletion according to Ancestral State

Region in Reference Genome inferring Deletion State

Region in Reference Genome inferring Insertion State

Junction A

Junction C

Junction B

1000 bp

Syntenic Primate Region inferring Insertion State

Syntenic Primate Region inferring Deletion State

SV Junction Library

[10]

[Lam et al., (‘10) Nat. Biotech.]
SV Insertion Traces

NAHR-based insertions involve nearby sequences.

NHR- and RT-based insertions are mostly inter-chromosomal.

[Lam et al., ('10) Nat. Biotech.]
Breakpoint Features Analysis

[Image: Diagram showing SVs vs. Telomeres and Breakpoints]

[Lam et al., '10 Nat. Biotech.]
AlleleSeq

Allele-Specific Binding & Expression
Inferring Allele Specific Binding/Expression using Actual Sequence Reads

RNA/ChIP-Seq Reads
ACTTTGATAGCGTCAATG
CTTTGATAGCGTCAATGC
CTTTGATAGCGTCAACGC
TTGACAGCGTCAATGCAC
TGATAGCGTCAATGCACG
ATAGCGTCAATGCACGTC
TAGCGTCAATGCACGTCG
CGTCAACGCACGTCGGGA
GTCAATGCACGTCGAGAG
CAATGCACGTCGAGAGTT
AATGCACGTCGAGAGTTG
TGCACGTTGGGAGTTGGC

10 x T
2 x C

Haplotypes with a Heterozygous Polymorphism

Interplay of the annotation and individual sequence variants
Many Technical Issues in Determining ASE/ASB: Reference Bias
(naïve alignment against reference)

[Rozowsky et al., MSB (in press, '11)]
Construction of a Personal Diploid Genome & Transcriptome

Reference Genome

Paternal Haplotype

Equivalence map

Personal Genome

Maternal Haplotype

vcf2diploid

Genotyping, Phasing, Filtering

Personal Variants

SVs

Indels

SNPs

Reference [TGGAAGAAGACCCGTTT...]

Deletion [TGGAAGAAGACCCGTTT...]

SNP [TGGAAGAAGACCCGTTT...]

Insertion [TGGAAGAAGACCCGTTT...]

Personal Haplotype [TGGAAGCGAGTTT...]

[Rozowsky et al., MSB (in press,‘11)]
Align reads to paternal haplotype
Align reads to maternal haplotype
Align reads to paternal splice-junction library
Align reads to maternal splice-junction library
Compare to find best alignment
Counts over het SNPs to determine allele specificity
Filter SNPs in CNVs using read-depth
Overlap ASB SNPs with TF binding sites
Overlap ASE SNPs with gene annotation
Report ASB and ASE SNPs with significance in VCF format

[Rozowsky et al., MSB (in press,'11)]
Specific ENCODE/1000G Data Sets

<table>
<thead>
<tr>
<th>Data</th>
<th>Number of reads (millions)</th>
<th>Number of mapped reads (millions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNA-Seq</td>
<td>393.9</td>
<td>164.7</td>
</tr>
<tr>
<td>Pol II ChIP-Seq</td>
<td>128 (33)</td>
<td>69.5 (13.2)</td>
</tr>
<tr>
<td>Pol III ChIP-Seq</td>
<td>12</td>
<td>7.5</td>
</tr>
<tr>
<td>cMyc ChIP-Seq</td>
<td>125</td>
<td>65.5</td>
</tr>
<tr>
<td>Max ChIP-Seq</td>
<td>79</td>
<td>46.1</td>
</tr>
<tr>
<td>JunD ChIP-Seq</td>
<td>133</td>
<td>72.5</td>
</tr>
<tr>
<td>cFos ChIP-Seq</td>
<td>84</td>
<td>30.4</td>
</tr>
<tr>
<td>NFkB ChIP-Seq</td>
<td>62</td>
<td>35.5</td>
</tr>
<tr>
<td>CTCF ChIP-Seq</td>
<td>46</td>
<td>26.4</td>
</tr>
</tbody>
</table>

- GM12878 is the immortalized lymphoblastoid cell-line from NA12878, the daughter in one of the deeply sequenced 1000G trios

[Rozowsky et al., MSB (in press,'11)]
Reference Bias Revisited

Assessing Reference Bias for GM12878 RNA-Seq data using Naïve reference mapping vs Modified reference mapping vs NA12878 mapping

[Rozowsky et al., MSB (in press,’11)]
Mapping Comparison

Chip-Seq Reads (Pol II & CTCF)

- Bowtie
  - Reference genome
  - Paternal haplotype
  - Maternal haplotype

Unique best mapping

Haplotype A
- Equivalent positions
  - Equivalent mapping

Haplotype B
- Equivalent positions
  - Different mapping
  - Not mapped

[Rozowsky et al., MSB (in press, '11)]
## Comparing Mapped Reads for Pol II ChIP-Seq

<table>
<thead>
<tr>
<th>Haplotype</th>
<th>Mapped reads</th>
<th>Equivalently mapped reads in</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Reference</td>
</tr>
<tr>
<td>Reference</td>
<td>69,086,591</td>
<td>68,942,501 (99.79%)</td>
</tr>
<tr>
<td>Paternal</td>
<td>(+0.3%) 69,296,783</td>
<td>68,942,501 (99.49%)</td>
</tr>
<tr>
<td>Maternal</td>
<td>(+0.4%) 69,394,995</td>
<td>69,034,357 (99.48%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Differently mapped reads in</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

## Comparing Peaks for Pol II ChIP-Seq

<table>
<thead>
<tr>
<th>Pol II Peaks</th>
<th>Maternal</th>
<th>Paternal</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maternal</td>
<td>1.000</td>
<td>0.977</td>
<td>0.958</td>
</tr>
<tr>
<td>Paternal</td>
<td>1.000</td>
<td>0.951</td>
<td></td>
</tr>
<tr>
<td>Reference</td>
<td></td>
<td></td>
<td>1.000</td>
</tr>
</tbody>
</table>

[Rozowsky et al., MSB (in press,'11)]
Allele-Specific Expression & Binding

<table>
<thead>
<tr>
<th>Genomic element</th>
<th>Number of Elements Accessible for Allele-Behavior</th>
<th>Number with ASE or ASB</th>
<th>Fraction with Allele-Specific Behavior</th>
<th>Maternal</th>
<th>Paternal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chromosome X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genes</td>
<td>94</td>
<td>75</td>
<td>0.80</td>
<td>70</td>
<td>4</td>
</tr>
<tr>
<td>Novel TARs</td>
<td>149</td>
<td>75</td>
<td>0.50</td>
<td>70</td>
<td>1</td>
</tr>
<tr>
<td>Pol II Sites</td>
<td>110</td>
<td>48</td>
<td>0.44</td>
<td>47</td>
<td>1</td>
</tr>
<tr>
<td>TFs Sites Combined</td>
<td>259</td>
<td>40</td>
<td>0.15</td>
<td>28</td>
<td>10</td>
</tr>
<tr>
<td>Autosomes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genes</td>
<td>4,829</td>
<td>935</td>
<td>0.19</td>
<td>491</td>
<td>424</td>
</tr>
<tr>
<td>Splice Junctions</td>
<td>2,556</td>
<td>552</td>
<td>0.21</td>
<td>272</td>
<td>202</td>
</tr>
<tr>
<td>Novel TARs</td>
<td>9,238</td>
<td>860</td>
<td>0.09</td>
<td>386</td>
<td>363</td>
</tr>
<tr>
<td>Pol II</td>
<td>3,187</td>
<td>344</td>
<td>0.11</td>
<td>172</td>
<td>126</td>
</tr>
<tr>
<td>Pol III</td>
<td>46</td>
<td>2</td>
<td>0.04</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>CTCF</td>
<td>4,573</td>
<td>443</td>
<td>0.10</td>
<td>178</td>
<td>207</td>
</tr>
<tr>
<td>NFkB</td>
<td>1,300</td>
<td>56</td>
<td>0.04</td>
<td>22</td>
<td>27</td>
</tr>
<tr>
<td>cFos</td>
<td>378</td>
<td>36</td>
<td>0.10</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>Max</td>
<td>943</td>
<td>55</td>
<td>0.06</td>
<td>24</td>
<td>22</td>
</tr>
<tr>
<td>cMyc</td>
<td>1,542</td>
<td>36</td>
<td>0.02</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>JunD</td>
<td>313</td>
<td>25</td>
<td>0.08</td>
<td>15</td>
<td>6</td>
</tr>
</tbody>
</table>

~20% sites show ASE, ~10% show ASB; equal betw. M & P, except on X

[Rozowsky et al., MSB (in press, ’11)]
Allele-Specific Regulatory Network: coordination of ASE & ASB

<table>
<thead>
<tr>
<th>Single TF</th>
<th>Maternal Expression</th>
<th>Paternal Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maternal Regulation</td>
<td>81</td>
<td>22</td>
</tr>
<tr>
<td>Paternal Regulation</td>
<td>31</td>
<td>64</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Multiple TFs (MIM)</th>
<th>Maternal Expression</th>
<th>Paternal Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Both Maternal Regulation</td>
<td>40</td>
<td>0</td>
</tr>
<tr>
<td>Both Paternal Regulation</td>
<td>4</td>
<td>36</td>
</tr>
<tr>
<td>Mixed Regulation</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Single TF (SIM)</th>
<th>Both Maternal Expression</th>
<th>Both Paternal Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Both Maternal Regulation</td>
<td>2,840</td>
<td>224</td>
</tr>
<tr>
<td>Both Paternal Regulation</td>
<td>254</td>
<td>1,232</td>
</tr>
</tbody>
</table>

[Rozowsky et al., MSB (in press,’11)]
Platform Comparison
Correlation of Tiling Array and RNA-Seq Signals

- Raw signals correlate well
  Spearman = 0.90

- Blue region – probes potentially experiencing cross-hyb

[Agarwal et al., BMC Genomics ('10)]
**Seq vs Array: Sequencing Depth Required to Match Array**

- Using all 32M reads, AUC much higher for RNA-Seq (red) than array (black).
- At desired FPR of 0.05, RNA-Seq matches tiling array with 4M reads, about 1/2 lane.

[Agarwal et al., BMC Genomics ('10) ]
Genome Sequencing Reveals Many Variants
(3.7 M SNPs, 217K Indels and ~3K High confidence SVs)

• Complete Genomics: 35 b paired ends (150X)
• Illumina: 100 b paired ends (120X)
Exome-seq and WGS-specific detection

Putting it Together
HugeSeq: An Automatic Pipeline for Calling Variants

I. Mapping
- Reads
- Dividing Reads
  - Set 1
  - Set n
  - Gapped Alignment
    - BWA Mapping
    - BAM Generation
  - Aligned BAM 1
  - Aligned BAM n

II. Sorting
- Aligned BAM 1
- Aligned BAM n
- Sorting by Chromosomes
  - chr1 BAM
  - chrM BAM
- Cleanup
  - Duplicate Removal
  - Local Realignment
  - Base Recalibration
  - Cleaned chr1 BAM
  - Cleaned chrM BAM

III. Reduction
- Cleaned chr1 BAM
- Cleaned chrM BAM
- Variant Calling
  - SNP/Indel
    - GATK
    - SAMtools
  - SV/CNV
    - Findel (SR)
    - CNVnator (RD)
    - BreakDancer (RP)
    - BreakSeq (JM)
- Final
  - Combine & Merge
  - Functional Annotation
  - SNP/Indel (VCF)
  - SV/CNV (GFF)

Mapping → SNVs, Indels, SVs

Lam et al.
Personal “Omics” Profiling (POP)

- Genome and Epigenome
- Transcriptome (mRNA, miRNA, isoforms, edits)
- Proteome
- Cytokines
- Metabolome
- Autoantibody-ome
- Microbiome

Personal Omics
Personalized Medicine: Combine Genomic and Other Omic Information

Genomic

GGTTCCAAAAAGTTTATTGGATGCCGTTTCA
GTACATTTATCGTTTGCTTTGGATGCCCTA
ATTAAAAGTGACCCTTTCAAACTGAAATTC
ATGATAAACAAATGGATATCCTAGTGAT
AAAATTTGCAATRACTTCAAGCCAAATG
AAATTATCTATGGTAGACAAAAACATTGACC
AATTTGATAGATCTCCTGATATTATTG
GCGTTGACACAGCTGGTATATTTCAAGTG
ACAAGGCAAATTACTTGGACCGTAAATAGAT
TTTTGAGCTCAGCAAAAAGAAAAATGGA
AATTAATTTTGAAGTGCCATTGA...

1. Predict risk
2. Diagnose,
3. Monitor,
4. Treat, &
5. Understand Disease States
Personalized Medicine: Combine Genomic and Other Omic Information

Genomic

Transcriptomic, Proteomic

GGTTCCAAAGTGTATTGGATGCGGTTTCA
GTACATTATTGCTTTGGATGGCCTA
ATTTAAAGTGGACCGTACTTTCAAAACTGAAATTC
ATGATACACCAATGGATATCCTTTAGTGAT
AAAATTTGCGAGTACTTTCAAGCCAAATG
AAATTATCTATGTAGACAAACACATTGACC
AATTTCATATCGATCTTCTCTAAATTATCG
GCGTTAGACACAGTTGGTATATTTCAAGTG
ACAAGGACAATTACTTTGACCGTAATAGAT
TTTTGAGGCTACAGCAAAAAGAAAAATGGA
AATTAATTTGAAGTGCCATTGA...

1. Predict risk
2. Diagnose,
3. Monitor,
4. Treat, &
5. Understand Disease States
Follow One Person: 21 Month Time Course

Healthy (Day -123)

Common Cold (HRV Infection) (Day 0)

Recovery I (Day 4)

Recovery II (Day 21)

Healthy (Day 116)

Healthy (Day 185)

Healthy* (Day 186)

Healthy (Day 255)

Common Cold (RSV Infection) (Day -1'/289)

Common Cold (RSV Infection) (Day 0'/290)

Recovery I (Day 2'/292)

Recovery II (Day 4'/294)

Recovery III (Day 7'/297)

Recovery IV (Day 11'/301)

Recovery V (Day 17'/307)

Recovery VI (Day 21'/311)

Healthy* (Day 32'/322)

Healthy (Allergy)* (Day 39'/329)

Healthy* (Day 79'/369)

* Fasted.
Many SNVs are Expressed
Integrated Analysis of Proteome, Transcriptome, Metabolome Dynamics: Overall trend
SV/CNV: vs. DGV + 1KG

<table>
<thead>
<tr>
<th># Methods</th>
<th>Total</th>
<th>DGV + 1KG</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Three</td>
<td>479</td>
<td>463</td>
<td>96.7%</td>
</tr>
<tr>
<td>Two</td>
<td>2,326</td>
<td>2,091</td>
<td>89.9%</td>
</tr>
<tr>
<td>One</td>
<td>53,694</td>
<td>6,094</td>
<td>11.3%</td>
</tr>
<tr>
<td>Total</td>
<td>56,499</td>
<td>8,646</td>
<td>15.3%</td>
</tr>
</tbody>
</table>

* Preliminary results -- Methods: RP, SR, RD; Platform*
Medical Interpretation Pipeline
Curated List of Rare Variants
(All heterozygous)

Bolded Genes expressed in PBMC (RNA).
Disease risk profile from Varimed + Integration
GLUCOSE LEVELS
Conclusions

• Tools for SVs
  – RD : MSB + CNVnator
  – SR : SRiC, AGE & BreakSeq
    • Split reads can readily find deletions & to a lesser degree insertions
    • This process can be calibrated and precise breakpoints defined, which allow mechanisms (NAHR, NHR) to be suggested

• AlleleSeq
  – Allele-specific binding & expression are widespread (~10-20%) and coordinated
  – Measurement requires surmounting technical issues

• Platform Comparison:
  sequencing is accurate but there are differences

• Test Sample Project
  Integrating the Technologies
  – ASE & SVs in practice
  – Interesting test case: Genome other omics information can monitor disease risk that is actionable.
ncVAR
**ncVar** : framework for integrative analysis for genomic variations and non-coding elements

1000 Genomes genomic variations: SNPs, indels, and SVs

Combine datasets, annotating genomic variations within genomic elements

Data Integration

Non-coding elements: TF-binding sites, ncRNA, pseudogenes...

Subclasses of elements: proximal/distal TF motifs, highly/lowly expressed ncRNAs, old/young pseudogenes...

Element-aware aggregation plots

Compute population genetics metrics for each (sub)class of elements: diversity, divergence, excess of rare SNPs...

X-Y plots

Element-aware aggregation plots
SNP diversity and divergence

In 1000G pilot, CEU pop. All $p$-values < 0.05 by MK test

- CDS
- TF-binding peak and assoc.
- TF-binding motifs
- ncRNA
- Pseudogene (neutral reference)

Accelerated element

- intron
- 3'UTR
- 5'UTR
- CTCF
- STAT1
- NFkB
- NRSF
- c-Fos
- c-Jun
- c-Myc
- JunD
- PolII
- DNase
- tRNA
- miRNA
- rRNA
- snRNA
- snoRNA
- misc_RNA

Figure 2

Diversity ($\pi \times 1,000$) vs. Divergence ($D_{xy} \times 100$)

[Mu et al., NAR (2011, in press)]
Variant allele frequency spectrum

- CDS
- TF-binding peak and assoc.
- TF-binding motifs
- ncRNA
- Pseudogene (neutral reference)
- intron
- 3'UTR
- 5'UTR

Derived allele frequency of SV

Fraction of SV

Derived allele frequency of SNPs

Derived allele frequency of Indel
**Variant ε:**
excess of rare alleles (MAF<0.05) compared to neutral ref.

- **A**: Scatter plot showing ε of SNP (%) vs ε of Indel (%).
- **B**: Scatter plot showing ε of SNP (%) vs ε of SV (%).
- **C**: Scatter plot showing ε of SV (%) vs ε of Indel (%).
- **D**: Scatter plot showing fraction of indel vs ε of SV (%).

- **Legend**:
  - CDS
  - TF-binding peak and assoc.
  - TF-binding motifs
  - ncRNA
  - Pseudogene (neutral reference)
  - intron
  - 3’UTR
  - 5’UTR
  - CTCF
  - STAT1
  - NFκB
  - NRSF
  - c-Fos
  - c-Jun
  - c-Myc
  - JunD
  - PolII
  - DNase
  - tRNA
  - miRNA
  - rRNA
  - snRNA
  - snoRNA
  - misc_RNA

- **References**:
  [Mu et al., NAR (2011, in press)]
Comparison between subgroups of elements

[Mu et al., NAR (2011, in press)]
Element-aware aggregation plot for SNP and indel diversity within and around coding genes

Confidence intervals estimated by block bootstrapping
Red box blow-out: estimation by simple bootstrapping

[Mu et al., NAR (2011, in press)]
Element-aware agg. over TF-binding motifs
## SVs interacting with genomic elements
(enrichment wrt randomized control)

<table>
<thead>
<tr>
<th>Element</th>
<th>All SVs</th>
<th>NAHR</th>
<th>VNTR</th>
<th>NHR</th>
<th>TEI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Enrichment</td>
<td>$P$-value</td>
<td>Enrichment</td>
<td>$P$-value</td>
<td>Enrichment</td>
</tr>
<tr>
<td>Gene CDS</td>
<td>0.90</td>
<td>8.68E-20</td>
<td>1.13</td>
<td>4.98E-08</td>
<td>0.84</td>
</tr>
<tr>
<td>5'UTR Only</td>
<td>0.37</td>
<td>8.72E-85</td>
<td>0.68</td>
<td>1.94E-06</td>
<td>0.07</td>
</tr>
<tr>
<td>3'UTR Only</td>
<td>0.96</td>
<td>2.17E-01</td>
<td>1.03</td>
<td>3.86E-01</td>
<td>0.83</td>
</tr>
<tr>
<td>Intron Only</td>
<td>0.72</td>
<td>3.47E-03</td>
<td>1.06</td>
<td>3.90E-01</td>
<td>0.80</td>
</tr>
<tr>
<td>Whole Gene</td>
<td>1.02</td>
<td>7.60E-02</td>
<td>1.25</td>
<td>5.92E-13</td>
<td>0.91</td>
</tr>
<tr>
<td>Partial Gene</td>
<td>1.41</td>
<td>8.96E-03</td>
<td>1.92</td>
<td>1.72E-03</td>
<td>2.76</td>
</tr>
<tr>
<td></td>
<td>0.90</td>
<td>1.06E-20</td>
<td>1.12</td>
<td>3.54E-07</td>
<td>0.83</td>
</tr>
<tr>
<td>ncRNA</td>
<td>1.08</td>
<td>2.06E-01</td>
<td>1.21</td>
<td>1.25E-01</td>
<td>0.97</td>
</tr>
<tr>
<td>Whole ncRNA</td>
<td>1.03</td>
<td>3.94E-01</td>
<td>1.18</td>
<td>1.64E-01</td>
<td>0.76</td>
</tr>
<tr>
<td>Partial ncRNA</td>
<td>1.83</td>
<td>2.58E-02</td>
<td>1.73</td>
<td>2.17E-01</td>
<td>2.10</td>
</tr>
<tr>
<td>Motif</td>
<td>0.73</td>
<td>3.74E-13</td>
<td>0.87</td>
<td>3.86E-2</td>
<td>1.44</td>
</tr>
<tr>
<td>Whole Motif</td>
<td>0.73</td>
<td>5.58E-13</td>
<td>0.90</td>
<td>7.35E-02</td>
<td>1.39</td>
</tr>
<tr>
<td>Partial Motif</td>
<td>0.75</td>
<td>1.74E-01</td>
<td>0.00</td>
<td>4.66E-02</td>
<td>2.48</td>
</tr>
<tr>
<td>Pseudogene</td>
<td>1.24</td>
<td>1.11E-05</td>
<td>1.56</td>
<td>3.37E-07</td>
<td>1.54</td>
</tr>
<tr>
<td>Whole Pseudogene</td>
<td>1.51</td>
<td>1.15E-12</td>
<td>1.95</td>
<td>3.98E-13</td>
<td>2.50</td>
</tr>
<tr>
<td>Partial Pseudogene</td>
<td>0.93</td>
<td>2.39E-01</td>
<td>0.97</td>
<td>4.40E-01</td>
<td>1.05</td>
</tr>
</tbody>
</table>

SVs are shuffled in the whole genome.
Significant $P$-values (<0.05) in black and bold
Significant enrichments in green; Significant depletions in red.

[Mu et al., NAR (2011, in press)]
RNA-Seq pipeline
RSeqTools + MRF
Introduction

• The application of next-generation sequencing for functional genomics has generated large quantities of data

• Challenges:
  – Data sets are so large that they are difficult to share
  – Sequence information potentially contains sufficient information to identify the underlying individual

• Privacy concerns

[Habegger et al., Bioinformatics (in revision, '10)]
Mapped Read Format

- Mapped Read Format (MRF)
  - Compact data summary format for short and long read alignments
  - Enables the anonymization of confidential information
  - Still possible to carry out most functional genomics analyses

- Implemented a suite of tools (RSEQtools) that uses MRF to analyze RNA-Seq data sets
  - [http://rseqtools.gersteinlab.org/](http://rseqtools.gersteinlab.org/)

[Habegger et al., Bioinformatics (in revision, '10)]
MRF flat file

• MRF flat file consists of three components
  – Comment lines
  – Header line
  – Mapped reads

• Mapped reads are represented in terms of AlignmentBlocks
  – TargetName:Strand:TargetStart:TargetEnd:QueryStart:QueryEnd

[Habegger et al., Bioinformatics (in revision, '10)]
MRF flat file: example 1

# Example 1
AlignmentBlocks
# Example 2

*AlignmentBlocks*


Legend: TS = TargetStart, TE = TargetEnd, QS = QueryStart, QE = QueryEnd

[Habegger et al., Bioinformatics (in revision, '10)]
Anonymization of confidential sequence information

# Example 3

<table>
<thead>
<tr>
<th>AlignmentBlocks</th>
<th>Sequences</th>
<th>QueryID</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr4:-1277:1290:1:24</td>
<td>ATCCCGTTTACAATCGGCATATCA</td>
<td>1</td>
</tr>
<tr>
<td>chr16:+511:534:1:24</td>
<td>GGTGACCTGACGTTGACAAAACC</td>
<td>2</td>
</tr>
</tbody>
</table>

[Habegger et al., Bioinformatics (in revision, '10)]
FusionSeq
What are chimeric transcripts?

- Transcripts that are *not co-linear* in the genome space

- They can arise from:
  - genomic rearrangements, i.e. *gene fusions* (implicated in cancer)
  - post-transcriptional events, i.e. *trans-splicing*

**Example**: BCR-ABL gene fusion t(9;22)(q34;q11) schematic and FISH
Identification of fusion transcripts

- Traditional detection of fusion genes typically involves cytogenetic methods
- Some require a hypothesis about the genes involved in the fusion
- Next-generation sequencing can help to address this question, especially with:
  - *Paired-end* RNA-Seq: keeping connectivity information

mRNA fragment
Identification of fusion transcripts

- Traditional detection of fusion genes typically involves cytogenetic methods
- Some require a hypothesis about the genes involved in the fusion
- Next-generation sequencing can help to address this question, especially with:
  - Paired-end RNA-Seq: keeping connectivity information

CTTGGAAGC GTGCTATGAA
mRNA fragment

FusionSeq

Sboner A*, Habegger L* et al., FusionSeq: a modular framework for finding gene fusions by analyzing Paired-End RNA-Sequencing data: 2010, Genome Biology, in press
Fusion Detection Module

FusionSeq uses Mapped Read Format (MRF)

Habegger L*, Stoner A* et al., RSEQtools: A modular framework to analyze RNA-Seq data with a concise confidential format*. 2010; Bioinformatics, submitted
Not an ideal world: sources of errors

- **Mis-alignment**
  - Base caller errors
  - SNPs
  - RNA editing
  - Sequence similarity (paralogs, pseudogenes)
- **Random pairing of transcript fragments**
  - Sample preparation
- **Combination of mis-alignment and random pairing**
- **PCR amplification, gene annotation inconsistencies**

[Filtration Cascade module]

[Sboner et al., Genome Biol. (in press, '10)]
Scoring the candidates

- Supportive PE Reads per million mapped reads (SPER)
  - Normalized number of inter-transcript PE reads ($m_i$)
    \[
    SPER_i = \frac{m_i}{N_{mapped}} \cdot 10^6
    \]

- How good is the observed SPER compared with the expected SPER?
  - Difference of observed SPER and analytically computed SPER (DASPER)
    \[
    DASPER_i = SPER_i - \langle SPER_i \rangle
    \]
    \[
    \langle SPER \rangle = \frac{\langle m_{AB} \rangle}{N_{mapped}} \cdot 10^6
    \]
    \[
    \langle m_{AB} \rangle = P(A) \cdot P(B) \cdot N_{mapped} = \frac{m_A \cdot m_B}{N_{mapped}}
    \]
  - Ratio of observed SPER and empirically computed SPER (RESPER)
    \[
    RESPER_i = \frac{SPER_i}{\overline{SPER}}
    \]
    \[
    \overline{SPER} = \frac{1}{M} \cdot \sum_{j=1}^{M} SPER_j
    \]

[Sboner et al., Genome Biol. (in press, '10)]
Junction-Sequence Identifier

[Sbener et al., Genome Biol. (in press, '10)]
Junction-Sequence Identifier

[Sboner et al., Genome Biol. (in press, '10)]
### Results (I)

Applied to 8 samples with and without known ERG rearrangements

<table>
<thead>
<tr>
<th>Type</th>
<th>Sample ID</th>
<th>Fusion Candidate</th>
<th>RESPER</th>
</tr>
</thead>
<tbody>
<tr>
<td>intra</td>
<td>580_B</td>
<td>TMPRSS2-ERG</td>
<td>14.31</td>
</tr>
<tr>
<td>intra</td>
<td>1700_D</td>
<td>TMPRSS2-ERG</td>
<td>8.79</td>
</tr>
<tr>
<td>intra</td>
<td>106_T</td>
<td>TMPRSS2-ERG</td>
<td>3.97</td>
</tr>
<tr>
<td>inter</td>
<td>2621_D</td>
<td>SLC45A3-ERG</td>
<td>3.56</td>
</tr>
<tr>
<td>inter</td>
<td>1700_D</td>
<td>ERG-GMPR</td>
<td>2.05</td>
</tr>
<tr>
<td>read-through</td>
<td>1700_D</td>
<td>SLC16A8-BAIAP2L2</td>
<td>1.93</td>
</tr>
<tr>
<td>read-through</td>
<td>106_T</td>
<td>AK094188-AK311452</td>
<td>1.9</td>
</tr>
<tr>
<td>read-through</td>
<td>1700_D</td>
<td>ZNF473-FLJ26850</td>
<td>1.58</td>
</tr>
<tr>
<td>read-through</td>
<td>580_B</td>
<td>ZNF577-FLJ26850</td>
<td>1.58</td>
</tr>
<tr>
<td>read-through</td>
<td>1043_D</td>
<td>ZNF577-ZNF649</td>
<td>1.55</td>
</tr>
<tr>
<td>read-through</td>
<td>1700_D</td>
<td>CAMTA2-INCA1</td>
<td>1.35</td>
</tr>
<tr>
<td>inter</td>
<td>1700_D</td>
<td>HDAC5</td>
<td>1.29</td>
</tr>
<tr>
<td>read-through</td>
<td>1043_D</td>
<td>FLJ00248-LRCH4</td>
<td>1.27</td>
</tr>
<tr>
<td>read-through</td>
<td>1700_D</td>
<td>VMAC-CAPS</td>
<td>1.17</td>
</tr>
<tr>
<td>read-through</td>
<td>106_T</td>
<td>FLJ00248-LRCH4</td>
<td>1.16</td>
</tr>
<tr>
<td>cis</td>
<td>1043_D</td>
<td>AX747861-FLI1</td>
<td>1.13</td>
</tr>
<tr>
<td>read-through</td>
<td>106_T</td>
<td>TAGLN-AK126420</td>
<td>1.07</td>
</tr>
<tr>
<td>inter</td>
<td>580_B</td>
<td>PIGU-ALG5</td>
<td>1.07</td>
</tr>
<tr>
<td>inter</td>
<td>99_T</td>
<td>NDRG1-ERG</td>
<td>1.02</td>
</tr>
</tbody>
</table>

---

Sboner A*, Habegger L* et al., FusionSeq: a modular framework for finding gene fusions by analyzing Paired-End RNA-Sequencing data: 2010, Genome Biology, in press
Integration with ACT

ACT (Aggregation and Correlation Toolbox)

- Can also correlate signal tracks
- Saturation analysis

Genome-wide signal track, with gene annotations below
ACT: Aggregation and Correlation Toolbox for Analyses of Genomic Tracks

- Aggregation
- Correlation
- Saturation

Jee et al. Bioinformatics (in press)
ACT: Aggregation Analyses

ChIP-Seq Signal Over TSSs

Aggregate of human and worm Pol II ChIP-Seq Signals

Aggregate Array vs Sequencing Signal over Exons
ACT: Saturation Analyses

Saturation Plots for coding and non-coding Transcribed Regions in C.elegans
Outline

• Tools for SV
  – RD : MSB + CNVnator
  – SR : SRiC, AGE & BreakSeq
• AlleleSeq : Integrating Var. & Func. Genomics
• Platform Comparison
• Test Sample Project Integrating the Technologies
• Tools for Selection : ncVAR
• Tools for RNAseq
  – RSeqTools, FusionSeq, ACT
Acknowledgements


- The Personal Genome Sequencing Project: Rui Chen, George Mias, Hugo Lam, Jennifer Li-Pook-Than, Lihua Jiang, Konrad Karczewski, Maeve O’Huallachain, Manoj Hariharan, Suganthi, Sara Hillemenyer, Rajini Haraksingh, Elana Miriami, Lukas Habegger, Rong Chen, Joel Dudley, Frederick Dewey, Shin Lin, Teri Klein, Russ Altman, Atul Butte, Euan Ashley, Mark Gerstein, Kari Nadeau, Hua Tang, M Snyder