Stanford-Yale CEGS:

Integrated Technologies for Characterizing the Human Genome

Mark B Gerstein (co-PI)
Sherman Weissman (co-PI)
Yale

Mike Snyder (PI)
Stanford
Technologies used by the CEGS for Interrogating the Human Genome, over the past decade

Tiling Arrays

- Application in a variety of contexts:
  - Transcription Mapping
  - DNA binding (inc. chromatin struc.)
  - Replication
  - Structural Variation

Massively Parallel Sequencing

- AGTTCACCTAAGA…
- CTTGAATGCCGAT…
- GTCATTCCGCAAT…
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
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
Analysis of the Human Genome Using Integrated Technologies

• Technology for finding & characterizing SVs
  – Using split reads (SRiC, AGE & BreakSeq)
• Technology for allelic analysis, integrating variation & functional genomics (AlleleSeq)
• POP project, integrating many genomic technologies
  – Incorporating SVs & allelic analysis
  – Platform comparison
Analysis of the Human Genome Using Integrated Technologies

- Technology for finding & characterizing SVs
  - Using split reads (SRiC, AGE & BreakSeq)
- Technology for allelic analysis, integrating variation & functional genomics (AlleleSeq)
- POP project, integrating many genomic technologies
  - Incorporating SVs & allelic analysis
  - Platform comparison

Alexej Abyzov: Special pipeline for measuring pseudogene variability with SRs:

Maeve O’Huallachain: Analysis of SVs in somatic tissues

George Mias: Dynamical Whole omics Profiling
Main Steps in Genome Resequencing

[Snyder et al. Genes & Dev. ('10)]

Step 0: Generate Reads

Step 1: Call SNPs
- using uniquely and correctly mapped reads

Step 2: Find SVs
- with aberrant paired-end reads, split-reads, read-depth analysis and CGH array data

Step 3: Assemble New Sequences
- with split-, spanning- and misleading-reads

Step 4: Phasing
- mostly with paired-end reads
Methods to Find SVs

1. Paired ends

Reference

Genome

Sequenced paired-ends

Mapping

Reference

2. Split read

Reference

Genome

Read

Mapping

Reference

3. Read depth (or aCGH)

Reference

Genome

Reads

Read count

Zero level

Mapping

Reference

4. Local Reassembly

[Snyder et al. Genes & Dev. (’10)]
Different Approaches Work Differently on Different Events

**Deletions**

- **Split-read analysis**: > 1 bp
- **RP (fosmid)**: > 8 kb
- **RP (454)**: > 3 kb
- **RP (Solexa/SOLiD)**: > 0.1 kb
- **hr-aCGH**: > 0.5 kb
- **dbSNP**: 1–28 bp

**Insertions**

- **dbSNP**: 1–28 bp
- **RP (Solexa/SOLiD)**: 100–250 bp
- **hr-aCGH**: > 0.5 kb
- **RP (454)**: 2–3 kb
- **RP (fosmid)**: 8–40 kb
- **Split-read analysis**: 1–250 bp

[Zhang et al. ('11) *BMC Genomics*]
Split-read Analysis

Zhang et al. Submitted

More: Breakpoint AssemblyAlt: BreakSeq
Deletions are the Easiest to Identify

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

[Zhang et al. ('11) BMC Genomics]
“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.

* 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

[Chen et al., ('10) Nat. Biotech.]
SV Breakpoint Library

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

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

[Figure S3. [Zhang et al. ('11) BMC Genomics]]
Using Simulation to Parameterize SRiC: Coverage & Read Length

(A) Different read length

(C) Different coverage

(Zhang et al. ('11) BMC Genomics]
SV Ancestral State Analysis

Inferring Insertion according to Ancestral State

Inferring Deletion according to Ancestral State

Region in Reference Genome inferring Deletion State

Junction A

1000 bp

Junction C

Junction B

Syntenic Primate Region inferring Insertion State

SV Junction Library

Region in Reference Genome inferring Insertion State

Junction A

1000 bp

Junction C

Junction B

Syntenic Primate Region inferring Deletion State

[Jam et al., ('10) *Nat. Biotech.*]
Difficulties in defining breakpoints

[Abyzov & Gerstein ('11) Bioinfo.]

Optimal alignment

NW alignment

Small gap penalty

Large gap penalty

SW alignment

Optimal alignment

NW alignment

Small gap penalty

Large gap penalty

SW alignment

Optimal alignment

Local/global alignment at right

Local alignment at left

A)

B)

C)
AGE
Alignment with Gap Excision

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

[Abyzov & Gerstein ('11) Bioinfo.]
Using Precise Breakpoints to Assign Mechanism
SV Mechanism Classification

NAHR

Highly similar with minor offset

Single RETRO

Multiple RETRO

[By et al., ('10) Nat. Biotech.]
SV Mechanism Classification

1 kb ≤ SV ≤ 1 Mb

Has flanking sequences

yes

Annotate SV and flanking regions by RepeatMasker

Has extensive coverage by VNTR regions

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

MTEI

Annotated as fragments from a single TE

no

STEI

Other ambiguous

2%

RNTR

5%

NHR

45%

Other novel active L1

16%

Putative novel active L1

13%

Potential processed pseudogene

1%

Other

69%

Reported active L1

15%

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

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

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

RNA/ChIP-Seq Reads

ACTTTTGATAGCGTCAATG
CTTTTGATAGCGTCAATGC
CTTTTGATAGCGTCAACGC
TTGACAGCGTCAAATGCAC
TGATAGCGTCAAATGCACG
ATAGCGTCAAATGCACGTC
TAGCGTCAAATGCACGTCG
CGTCAACGCACGTCGGGA
GTCAATGCACGTCGAGAG
CAATGCACGTCGAGGAGTT
AATGCACGTCGAGGAGTTG
TGACGTTGGGAGTTGGC

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)

[Graph showing the distribution of ASE SNPs with frequency on the y-axis and the ratio of alternate counts to total counts on the x-axis.]

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

Reference Genome

\[\text{Reference Genome}\]

Paternal Haplotype

Personal Genome

Maternal Haplotype

\[\text{Personal Genome}\]

\[\text{Paternal Haplotype}\]

\[\text{Maternal Haplotype}\]

\[\text{Reference}\ TGGAGAGAACCGTTT...\]

\[\text{Deletion}\ TGG\_\_\_\_\_\_AACCGTTT...\]

\[\text{SNP}\ TGG\_\_\_\_\_\_AACCGTTT...\]

\[\text{Insertion}\ TGG\_\_\_\_\_\_AACCG\_\_\_TTT...\]

\[\text{Personal Haplotype}\ TGGAAAGGAGTTT...\]

Genotyping, Phasing, Filtering

\[\text{Genotyping, Phasing, Filtering}\]

vcf2diploid

\[\text{vcf2diploid}\]

Personal Variants

\[\text{Personal Variants}\]

SVs

\[\text{SVs}\]

Indels

\[\text{Indels}\]

SNPs

\[\text{SNPs}\]

\[\text{Construction of a Personal Diploid Genome & Transcriptome}\]

[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

[��owski et al., MSB (in press, '11)]
### Specific Data Sets Used

<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 NA12878 mapping

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

~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

![Network Motifs](image)

**Expression**
- Gene
- Novel TAR
- TF

**Binding**
- Maternal
- Paternal

<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)]
Putting it Together:
POP
HugeSeq: An Automatic Pipeline for Calling Variants

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

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

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

Mapping → SNVs, Indels, SVs

[Lam et al. (submitted)]
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

GGTTCCAAAAGTTTATTGGATGCCGTTTCA
GTACATTTATCGTTTGCTTTGGATGCCCTA
ATTAAAAGTGACCCTTTCAAAACTGAAATTC
ATGATACACCAATGGATATCCTTAGTCAAT
AAAATTTGCGAGTACTTTCAAGCCAAATG
AAATTATCTATGGTAGACAAAACATTTGAAC
AATTTCACTCATCGATCCTCCTGAAATTTTG
GCGTTACACAGGTGATATTTCAAGTG
ACAAGGACAAATTACTTGGACCGTAAATAGAT
TTTTGAGGCTCAGCAAAAAAAAGAATTTGA
AATTAATTTGAAAGTCCATTGA….

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

Genomic

Transcriptomic, Proteomic

- Predict risk
- Diagnose,
- Monitor,
- Treat, &
- 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)

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

→
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
SVs in Test Sample: Analyzed with SR, RP, RD

Concordant deletion calls between the detected SVs and SVs from the 1000 genomes project.

<table>
<thead>
<tr>
<th># Algorithms</th>
<th>Deletion calls</th>
<th>1bp overlapping</th>
<th>50% reciprocal overlapping</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;=3 algorithms</td>
<td>451</td>
<td>443</td>
<td>98.2%</td>
</tr>
<tr>
<td>&gt;=2 algorithms</td>
<td>1,594</td>
<td>1,472</td>
<td>92.3%</td>
</tr>
<tr>
<td>Any algorithm</td>
<td>19,809</td>
<td>5,468</td>
<td>27.6%</td>
</tr>
</tbody>
</table>

[Lam et al. (submitted)]
Platform Comparison
Genome Sequencing Reveals Many Variants (3.7 M SNPs, 217K Indels and ~3K Hi confidence SVs)

- Complete Genomics: 35 b paired ends (150X)
- Illumina: 100 b paired ends (120X)

[Lam et al. (submitted)]
Exome-seq and WGS-specific detection

Medical Interpretation Pipeline
Disease risk profile from Varimed + Integration
GLUCOSE LEVELS
Conclusions: Analysis of the Human Genome Using Integrated Technologies

- Technologies for SVs
  - SR : SRiC, AGE & BreakSeq
    - Split reads can readily find deletions & to a lesser degree insertions
    - SR can be calibrated and precise breakpoints defined, suggesting mechanisms (NAHR, NHR)

- Technology for allelic analysis, integrating variation & functional genomics(AlleleSeq)
  - Allele-specific binding & expression are widespread (~10-20%) and coordinated
  - Measurement requires surmounting reference bias

- Test Sample Project
  - Integrating the Technologies
    - ASE & SVs in practice
    - Platform Comparison: sequencing is accurate but there are differences
    - Interesting test case: other omics information can monitor disease risk that is actionable.
Acknowledgements

SRs: A Abyzov, ZD Zhang, H Lam, J Du, AE Urban, XJ Mu, AM Stütz, A Tanzer, PD Cayting, PM Kim, JO Korbel, M Snyder, M Gerstein

AlleleSeq: J Rozowsky, A Abyzov, J Wang, P Alves, D Raha, A Harmanci, J Leng, R Bjornson, Y Kong, N Kitabayashi, N Bhardwaj, M Rubin, M Snyder, M Gerstein

The Personal Genome Sequencing Project: Rui Chen, George Mias, Hugo Lam, Jennifer Li-Pook-Than, Lihua Jiang, Konrad Karczewski, Maeve O’Huallachain, Manoj Hariharan, Suganthy, Sara Hillemeenyer, 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