Next-generation tools for assaying genome variation

Carlos D. Bustamante

Department of Genetics

Stanford University School of Medicine

cdbustam@stanford.edu
COMPARING THE UNCOMPARABLE

The rarer a genetic variant is within a population, the less likely it is to be found in all ethnic groups. One hundred people were sampled from each population.

- **Europeans***
- **Europeans & Chinese†**
- **European & African‡**

<table>
<thead>
<tr>
<th>Degree to which variants are shared between populations (%)</th>
<th>Rare variants (1%)</th>
<th>Common variants (15%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>80</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>60</td>
<td>40</td>
<td>20</td>
</tr>
<tr>
<td>40</td>
<td>60</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>80</td>
<td>20</td>
</tr>
<tr>
<td>0</td>
<td>100</td>
<td>40</td>
</tr>
</tbody>
</table>

*Comparison of individuals of European descent in Utah and in Tuscany, Italy. † Han Chinese individuals from Beijing compared with Utah sample ‡ Yoruba individuals from Ibadan, Nigeria, compared with Utah sample.

Tools we’ve been developing

1. **GBS**: Genotype-by-sequencing
2. **WISC**: whole-in-solution genome capture
3. **RFMix**: Tool for admixture deconvolution from WGS and exome data
GBS specifically targets 2% of the genome for high-coverage sequencing

1. Genomic DNA
2. BpuEl, BsaXI, CspCl restriction sites
3. Digest w/ restriction enzymes
4. Ligate indexed sequencing adapters
5. Size-selection
6. Pool multiplexed samples
7. Paired-end Illumina sequencing
8.Mapped reads
## GBS is a low-cost solution for Population Genomics

<table>
<thead>
<tr>
<th>Method</th>
<th>Samples Per $10 K</th>
<th>Sequencing Coverage</th>
<th>SNPs Per Sample</th>
<th>Private variant Call Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exome Array (250k)</td>
<td>200</td>
<td>-</td>
<td>250 k</td>
<td>Does not assay private variants</td>
</tr>
<tr>
<td>Genotyping-by-sequencing (GBS)</td>
<td>60</td>
<td>20X</td>
<td>~75-100 k</td>
<td>&gt; 99 %</td>
</tr>
<tr>
<td>Whole-genome Array (1M)</td>
<td>15</td>
<td>-</td>
<td>1 M</td>
<td>Does not assay private variants</td>
</tr>
<tr>
<td>Exome Sequencing</td>
<td>10</td>
<td>100X</td>
<td>50 k</td>
<td>&gt; 99 %</td>
</tr>
<tr>
<td>Whole genome sequencing</td>
<td>10</td>
<td>~2-4X</td>
<td>3 M</td>
<td>~ 75 %</td>
</tr>
<tr>
<td>Whole genome sequencing</td>
<td>2-3</td>
<td>30X</td>
<td>3 M</td>
<td>&gt; 99 %</td>
</tr>
</tbody>
</table>
GBSTools

8 genomes sequenced by CG and GBS

Concordance

Read coverage bin

99%

95%

Sites per sample x 1000

Filter

Basic filters
GBS filters

Sunday, October 6, 13
GBS allows unbiased assaying of the Frequency Spectra

Expected GBS spectrum taken from 13514 sites and a subsample of 35/39 samples.
Summary of results

• GBS enables high coverage sequencing of 1-2 % of the genome at low cost.

• Concordance with known high-coverage whole-genome sequence data is high.

• Markers are randomly distributed, but having good genome assembly is critical for improving technology.

• GBS is well-suited for non-model organisms where arrays and exome capture tools are unavailable.
Tools we’ve been developing

1. **GBS**: Genotype-by-sequencing
2. **WISC**: whole-in-solution genome capture
3. **RFMix**: Tool for admixture deconvolution from WGS and exome data
Whole Genome In Solution Capture (WISC)

- Develop an a capture technology that was fast, cheap, and highly accurate
- Demonstrate utility on “real world” samples of interest
  - Four Iron Age and Bronze Age human teeth from Bulgaria
  - Bone samples from seven Peruvian mummies
  - Bronze Age hair sample from Denmark
- Intersect the data with 1,000 Genomes and see improved utility of capture for population genetics

Meredith Carpenter
Jason Buenrostro
Will Greenleaf

Sunday, October 6, 13
5,300-yr-old ‘iceman’


Otzi’s Male Lineage in Sardinia

L91 mutation distribution

Sunday, October 6, 13
Inferred Y-Chromosome Phylogeny

Poznik et al., Science (2013)
Ancient tooth
Extracted DNA
(~1% human, 99% environmental)
Sequencing library
Modern human DNA library
In vitro transcription (with biotin-dUTP)
Biotinylated RNA probes
T7 RNA polymerase promoter
DNA extraction
Library building
Extracted DNA
(~1% human, 99% environmental)
Sequencing library
indexed Illumina adapter
Hybridization RNA probes (bait) + ancient DNA library (pond)
Capture with streptavidin-coated magnetic beads
Sequencing of target molecules only
Indexed Illumina adapter
RNA adapter blocker
RNA adapter
WISC
Sunday, October 6, 13
Coverage is even across the genome

Carpenter et al. (submitted)
Up to 139-fold enrichment in reads mapping to the human genome
Capture increases the number of SNPs by 2- to 14-fold

<table>
<thead>
<tr>
<th>Library name</th>
<th># SNPs Pre-capture</th>
<th># SNPs Post-capture</th>
<th>Fold enrichment</th>
</tr>
</thead>
<tbody>
<tr>
<td>V2</td>
<td>5,281</td>
<td>40,583</td>
<td>7.7</td>
</tr>
<tr>
<td>P192-1</td>
<td>30,081</td>
<td>67,221</td>
<td>2.2</td>
</tr>
<tr>
<td>T2G2</td>
<td>597</td>
<td>4,068</td>
<td>6.8</td>
</tr>
<tr>
<td>K8</td>
<td>19,960</td>
<td>94,394</td>
<td>4.7</td>
</tr>
<tr>
<td>M4</td>
<td>5,115</td>
<td>40,340</td>
<td>7.9</td>
</tr>
<tr>
<td>NA39</td>
<td>14,751</td>
<td>40,048</td>
<td>2.7</td>
</tr>
<tr>
<td>NA40</td>
<td>9,119</td>
<td>129,872</td>
<td>14.2</td>
</tr>
<tr>
<td>NA41</td>
<td>4,621</td>
<td>26,118</td>
<td>5.7</td>
</tr>
<tr>
<td>NA42</td>
<td>73,266</td>
<td>147,243</td>
<td>2.0</td>
</tr>
<tr>
<td>NA43</td>
<td>1,553</td>
<td>6,337</td>
<td>4.1</td>
</tr>
<tr>
<td>NA47</td>
<td>1,279</td>
<td>9,383</td>
<td>7.3</td>
</tr>
<tr>
<td>NA50</td>
<td>217</td>
<td>3,062</td>
<td>14.1</td>
</tr>
</tbody>
</table>

Carpenter et al. (submitted)
Conclusions

• Developed a whole-genome capture method to enrich the endogenous contents of aDNA sequencing libraries
  • Pre-capture: 1.2% human
  • Post-capture: up to 59% human
  • Enrichment: 5X to 139X.
  • SNPs: ~50,000 per sample per 1 Million reads
• Capture majority of fragments present in the pre-capture library.
• Substantially increase the number of ancient samples amenable to NGS
Tools we’ve been developing

1. **GBS**: Genotype-by-sequencing
2. **WISC**: whole-in-solution genome capture
3. **RFMix**: Tool for admixture deconvolution from WGS and exome data
African diaspora
RFMix

Trained Random Forest

Random Forest Inference

CRF Smoothing

Inferred Local Ancestry
Local ancestry estimation in Latino populations

Ecuadorians

Native American
European
African
Local ancestry estimation in Latino populations

Ecuadorians

Colombians

Puerto Ricans

Dominicans

Native American

European

African

Sunday, October 6, 2013
Colonial History of Puerto Rico

- 1506 – Spanish colonists arrive
- 1508-1530 – Taíno population is decimated by war, disease, and forced labor primarily in gold mining.
- 1530-1640 – Importation of African slaves from Cape Verde increases dramatically for sugar trade
- 1640-1765 – Legal slave trade is greatly diminished
- 1765-1846 – New wave of African slaves imported primarily from South of the equator.
- 1873 – Slavery officially abolished in PR.
Tract length distribution informs history of admixture
Old vs. Recent African migration

Primarily “Older” Migrants

Primarily “Recent” Migrants

Number of tracts vs. Tract Length (cM)

- Anc.
- Model
- Data

EUR
AFR
NAT

Sunday, October 6, 13
Old vs. Recent African migration
Old vs. Recent African migration

“old” migrants (<= 50 cM)
Old vs. Recent African migration

“old” migrants (<= 50 cM)  
“recent” migrants (> 50 cM)
Take home message

1. Personal ancestry reconstruction (including detection of admixture tracts) is feasible on genome-wide scale
2. We should be able to do for all what we can do for Otzi!
3. African-Americans exhibit, on average, ~78% West African and 22% European ancestry with large variation
4. Hispanic-Latinos vary tremendously in admixture proportions
5. Tremendous fine-scale population structure even within countries
6. Need more fully sequenced human genomes
Acknowledgements

Funding: NIGMS, NHGRI, HLBI, MacArthur Foundation
Acknowledgements

High Accuracy on Sub-Continental Admixtures