Population Sequencing

\[ C = 2-7 \times 10^3 \text{ to } 10^6 \]
Population Sequencing

$G_1, \ldots, G_N; \quad G_i = g_{i1} \ldots g_{in}; \quad g_{ij} \in \{0, 1, 2\}$

$P_1, \ldots, P_N; \quad P_i: [ p_{ijg} = \text{Prob}(g_{ij} = g \mid \text{data}) ]$
Population Sequencing

When C is high (>30x),

\[
\text{Prob}(g_{ij} = g \mid \text{data}) \sim
\]

\[
\text{Prob}(g_{ij} = g \mid \text{reads mapping on } (i, j))
\]

fast & easy

When C is low,

\[
\text{Prob}(g_{ij} = g \mid \text{data}) \text{ needs to leverage LD:}
\]

positions \(j' \neq j\) in all individuals

in principle, intractable

\[
G_1, \ldots, G_N; \quad G_i = g_{i1} \ldots g_{in}; \quad g_{ij} \in \{0, 1, 2\}
\]

\[
P_1, \ldots, P_N; \quad P_i : \quad [ p_{ijg} = \text{Prob}(g_{ij} = g \mid \text{data}) ]
\]
HMM-based models

- Li and Stephens 2003

Given $m$ reference haplotypes, and a target sample,
Find the most likely path of haplotype pairs
$m^2$ states, $m^4$ transitions per position
Evolution of a local haplotype
Evolution of haplotypes, recombination
Evolution of haplotypes, recombination
Informative Neighbors

\[ (R_{\text{ref}}, R_{\text{alt}}) = \sum_{\{\text{target}, \text{nbrs}\}} (r_{\text{ref}}, r_{\text{alt}}) = (20, 0) \]
Informative Neighbors

\[ (R_{\text{ref}}, R_{\text{alt}}) = \sum_{\text{target, nbrs}} (r_{\text{ref}}, r_{\text{alt}}) = (11, 9) \]

In terms of linkage disequilibrium:

Target SNP

\( \kappa \)-”nearest” neighbors
How to pick k nearest neighbors fast

Let

$S_i = \{ \text{samples covering minor allele} \}$

$S_i' = \{ \text{read counts of minor allele} \}$

$S_i = \{1, 2, 3, 10\}$

$S_j = \{1, 3, 4\}$

$S_i' = \{1, 2, 3, 3, 3, 10\}$

$S_j' = \{1, 3, 3, 3, 3, 3, 4, 4\}$

$\text{Sim}_1(i, j) = (S_i \cap S_j) / (S_i \cup S_j)$

$\text{Sim}_2(i, j) = (S_i' \cap S_j') / (S_i' \cup S_j')$

$\text{Sim}_3(i, j) = ((S_i' \cap S_j') / (S_i' \cup S_j'))^2$

Correlation Coefficient:

$r^2 = (p_{AB} - p_A p_B)^2 / p_A p_B p_a p_b$

Caveat: need **genotyping, phasing**
Genetic distance between NNs

common SNPs

rare SNPs
Overview of Reveel

Reveel:

1. Identify candidate polymorphic sites

2. Calculate k nearest neighbors
   - Jaccard indices $\text{Sim}_1, \text{Sim}_2, \text{Sim}_3$

3. Initialize $G^{(0)}$

4. Summarization/Maximization
   \[
   p^{(n+1)}_{ijg} = \text{Prob}(g_{ij} = g \mid G^{(n)} \text{ data}) \\
   g^{(n+1)}_{ijg} = \text{argmax} \ p^{(n+1)}_{ijg}
   \]

5. Recalculate k nearest neighbors
   - Approximate Correlation Coefficient (Schaid 2004)

6. Summarization/Maximization

7. Recalculate k nearest neighbors
   - Approximate CC, Entropy

8. Summarization/Maximization

\[ G_1, \ldots, G_N; \ G_i = g_{i1} \ldots g_{in}; \ g_{ij} \in \{0, 1, 2\} \]
\[ P_1, \ldots, P_N; \ P_i : [ p_{ijg} = \text{Prob}(g_{ij} = g \mid \text{ data}) ] \]
Overview of Reveel

**Reveel:**

1. **Identify candidate polymorphic sites**
2. Calculate k nearest neighbors
   - Jaccard indices $\text{Sim}_1$, $\text{Sim}_2$, $\text{Sim}_3$
3. Initialize $G^{(0)}$
4. Summarization/Maximization
   
   \[
   p^{(n+1)}_{ijg} = \text{Prob}(g_{ij} = g \mid G^{(n)}, \text{data})
   \]
   
   \[
   g^{(n+1)}_{ijg} = \text{argmax} \ p^{(n+1)}_{ijg}
   \]
5. Recalculate k nearest neighbors
   - Approximate Correlation Coefficient (Schaid 2004)
6. Summarization/Maximization
7. Recalculate k nearest neighbors
   - Approximate CC, Entropy
8. Summarization/Maximization

**Candidate Polymorphic site**

Essentially, pos’n j where some individuals have at least 2 reads with same minor allele
Overview of Reveel

**Reveel:**

1. Identify candidate polymorphic sites
2. Calculate k nearest neighbors
   - Jaccard indices $\text{Sim}_1, \text{Sim}_2, \text{Sim}_3$
3. **Initialize $G^{(0)}$**
4. Summarization/Maximization
   \[ p^{(n+1)}_{ijg} = \text{Prob}(g_{ij} = g \mid G^{(n)}, \text{data}) \]
   \[ g^{(n+1)}_{ijg} = \text{argmax} \ p^{(n+1)}_{ijg} \]
5. Recalculate k nearest neighbors
   - Approximate Correlation Coefficient (Schaid 2004)
6. Summarization/Maximization
7. Recalculate k nearest neighbors
   - Approximate CC, Entropy
8. Summarization/Maximization

At each position j,

Use sum of read counts at j and its nearest neighbors
Overview of Reveel

Reveel:

1. Identify candidate polymorphic sites
2. Calculate k nearest neighbors
   • Jaccard indices $\text{Sim}_1, \text{Sim}_2, \text{Sim}_3$
3. Initialize $G^{(0)}$
4. **Summarization/Maximization**
   \[ p^{(n+1)}_{ijg} = \text{Prob}(g_{ij} = g \mid G^{(n)}, \text{data}) \]
   \[ g^{(n+1)}_{ijg} = \text{argmax} \ p^{(n+1)}_{ijg} \]
5. Recalculate k nearest neighbors
   • Approximate Correlation Coefficient (Schaid 2004)
6. Summarization/Maximization
7. Recalculate k nearest neighbors
   • Approximate CC, Entropy
8. Summarization/Maximization

\[ G_1, \ldots, G_N; \quad G_i = g_{i1} \ldots g_{in}; \quad g_{ij} \in \{0, 1, 2\} \]
\[ P_1, \ldots, P_N; \quad P_i : [ p_{ijg} = \text{Prob}(g_{ij} = g \mid \text{data}) ] \]

\[ p^{(n+1)}_{ijg} = P(g_{ij} = g \mid G^{(n)}, \text{reads}) \]
\[ \sim P(g_{ij} = g \mid g_{k\text{NN}}, \text{reads}) \]
\[ = P(\text{reads} \mid g_{ij} = g) \ P(g_{ij} = g \mid g_{k\text{NN}}) \]

\[ P(g_{ij} = g \mid g_{k\text{NN}}) = \]

Let $C_0, C_1, C_2 = \# \text{ samples matching } i \text{ in } k\text{NN}, \text{ with } j^{\text{th}} \text{ genotype pos'n } = 0, 1, 2$

\[ P(g_{ij} = g \mid g_{k\text{NN}}) = C_g / (C_0 + C_1 + C_2) \]
Overview of Reveel

**Reveel:**

1. Identify candidate polymorphic sites
2. Calculate k nearest neighbors
   - Jaccard indices $\text{Sim}_1$, $\text{Sim}_2$, $\text{Sim}_3$
3. Initialize $G^{(0)}$
4. Summarization/Maximization
   \[ p^{(n+1)}_{ijg} = \text{Prob}(g_{ij} = g \mid G^{(n)}, \text{data}) \]
   \[ g^{(n+1)}_{ijg} = \arg\max p^{(n+1)}_{ijg} \]
5. Recalculate k nearest neighbors
   - Approximate Correlation Coefficient (Schaid 2004)
6. Summarization/Maximization
7. Recalculate k nearest neighbors
   - Approximate CC, Entropy
8. Summarization/Maximization

Correlation Coefficient:
\[ r^2 = \frac{(p_{AB} - p_A p_B)^2}{p_A p_B p_a p_b} \]

Caveat: need **genotyping**, **phasing**

Schaid 2004:
\[ D = \frac{1}{N} \left( 2N_{AABB} + N_{AAbb} + N_{AaBB} + \frac{1}{2}N_{AaBb} \right) - 2p_A p_B \]
A faster alternative:
\[ D = \frac{1}{2} \text{Sim}_1(i, j) + p_A p_B (p_A + p_B - \frac{1}{2} p_A p_B - 2) \]
Overview of Reveel

Reveel:

1. Identify candidate polymorphic sites
2. Calculate k nearest neighbors
   • Jaccard indices Sim_1, Sim_2, Sim_3
3. Initialize G^{(0)}
4. Summarization/Maximization
   \[ p^{(n+1)}_{ijg} = \text{Prob}(g_{ij} = g \mid G^{(n)}, \text{data}) \]
   \[ g^{(n+1)}_{ijg} = \text{argmax} \ p^{(n+1)}_{ijg} \]
5. Recalculate k nearest neighbors
   • Approximate Correlation Coefficient (Schaid 2004)
6. Summarization/Maximization
7. Recalculate k nearest neighbors
   • Approximate CC, Entropy
8. Summarization/Maximization

Identify the sites at which decision trees formed by kNNs have high entropy

Reduce entropy by replacing one or more kNNs
Simulations

- Simulated data set: ~ low-coverage 1KGP
  - 2,535 samples, 1-Mbp region: Chr 20 (43-44 Mbp) of GRCh37
  - Using COSI to simulate variations
  - Reads same positions, lengths, qualities as reads in 1KGP
  - Injecting sequencing base errors, mapping with BWA

<table>
<thead>
<tr>
<th>Sample size</th>
<th>100</th>
<th>500</th>
<th>1000</th>
<th>2535</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reveel</td>
<td>1.8</td>
<td>14.6</td>
<td>47.4</td>
<td>273</td>
</tr>
<tr>
<td>Reveel+Beagle</td>
<td>3.1</td>
<td>25.3</td>
<td>71.2</td>
<td>526</td>
</tr>
<tr>
<td>Reveel-lite</td>
<td>1.5</td>
<td>7.8</td>
<td>21.7</td>
<td>145</td>
</tr>
<tr>
<td>SNPTools+Beagle</td>
<td>8.2</td>
<td>217</td>
<td>1089</td>
<td>&gt;5days</td>
</tr>
<tr>
<td>GATK+Beagle</td>
<td>13.4</td>
<td>388</td>
<td>1806</td>
<td>&gt;5days</td>
</tr>
<tr>
<td>glfMultiples+Thunder</td>
<td>307</td>
<td>2736</td>
<td>6120</td>
<td>~15 days</td>
</tr>
</tbody>
</table>
## Performance on 1000 Genomes Data

<table>
<thead>
<tr>
<th>population</th>
<th># of samples in 1KGP</th>
<th># of samples in HapMap3</th>
<th>population</th>
<th># of samples in 1KGP</th>
<th># of samples in HapMap3</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACB</td>
<td>96</td>
<td>0</td>
<td>ASW</td>
<td>66</td>
<td>50</td>
</tr>
<tr>
<td>BEB</td>
<td>86</td>
<td>0</td>
<td>CDX</td>
<td>99</td>
<td>0</td>
</tr>
<tr>
<td><strong>CEU</strong></td>
<td>99</td>
<td><strong>90</strong></td>
<td>CHB</td>
<td>103</td>
<td>94</td>
</tr>
<tr>
<td>CHS</td>
<td>108</td>
<td>0</td>
<td>CLM</td>
<td>94</td>
<td>0</td>
</tr>
<tr>
<td>ESN</td>
<td>99</td>
<td>0</td>
<td>FIN</td>
<td>99</td>
<td>0</td>
</tr>
<tr>
<td>GBR</td>
<td>92</td>
<td>0</td>
<td>GIH</td>
<td>106</td>
<td><strong>93</strong></td>
</tr>
<tr>
<td>GWD</td>
<td>113</td>
<td>0</td>
<td>IBS</td>
<td>107</td>
<td>0</td>
</tr>
<tr>
<td>ITU</td>
<td>103</td>
<td>0</td>
<td>JPT</td>
<td>104</td>
<td><strong>97</strong></td>
</tr>
<tr>
<td>KHV</td>
<td>101</td>
<td>0</td>
<td>LWK</td>
<td>101</td>
<td><strong>90</strong></td>
</tr>
<tr>
<td>MSL</td>
<td>85</td>
<td>0</td>
<td>MXL</td>
<td>67</td>
<td><strong>56</strong></td>
</tr>
<tr>
<td>PEL</td>
<td>86</td>
<td>0</td>
<td>PJL</td>
<td>96</td>
<td>0</td>
</tr>
<tr>
<td>PUR</td>
<td>105</td>
<td>0</td>
<td>STU</td>
<td>103</td>
<td>0</td>
</tr>
<tr>
<td><strong>TSI</strong></td>
<td>108</td>
<td><strong>96</strong></td>
<td>YRI</td>
<td>109</td>
<td>103</td>
</tr>
</tbody>
</table>

Compared on a 5-Mbp region on chromosome 20 (43-48 Mbp)
**Performance on 1000 Genomes Data**

**HapMap 3 benchmark**

769 individuals from 9 populations

- \( AF \geq 5\% \), 2686 SNPs
- \( 5\% > AF \geq 1\% \), 368 SNPs
- \( AF < 1\% \), 32 SNPs

<table>
<thead>
<tr>
<th>Population</th>
<th>Genotyping acc of Reveal w/ beagle</th>
<th>glfMultiples+thunder</th>
<th>SNPTools+beagle</th>
<th>gatk+beagle</th>
</tr>
</thead>
<tbody>
<tr>
<td>YRI</td>
<td>99.55%</td>
<td>0.02%</td>
<td>-0.27%</td>
<td>-0.31%</td>
</tr>
<tr>
<td>TSI</td>
<td>99.65%</td>
<td>-0.04%</td>
<td>-0.2%</td>
<td>-0.29%</td>
</tr>
<tr>
<td>MXL</td>
<td>99.15%</td>
<td>-0.05%</td>
<td>-0.31%</td>
<td>-0.35%</td>
</tr>
<tr>
<td>LWK</td>
<td>99.39%</td>
<td>-0.05%</td>
<td>-0.32%</td>
<td>-0.42%</td>
</tr>
<tr>
<td>JPT</td>
<td>99.69%</td>
<td>-0.01%</td>
<td>-0.09%</td>
<td>-0.16%</td>
</tr>
<tr>
<td>GIH</td>
<td>99.62%</td>
<td>-0.01%</td>
<td>-0.18%</td>
<td>-0.31%</td>
</tr>
<tr>
<td>CHB</td>
<td>99.63%</td>
<td>0.02%</td>
<td>-0.07%</td>
<td>-0.11%</td>
</tr>
<tr>
<td>CEU</td>
<td>99.67%</td>
<td>-0.03%</td>
<td>-0.1%</td>
<td>-0.22%</td>
</tr>
<tr>
<td>ASW</td>
<td>99.06%</td>
<td>-0.02%</td>
<td>-0.43%</td>
<td>-0.43%</td>
</tr>
</tbody>
</table>
Performance on 1000 Genomes Data

- **SNP discovery**
  - Discover 171,734 likely polymorphic sites in 26 populations
  - Benchmark: Complete Genomics data

<table>
<thead>
<tr>
<th>Method</th>
<th>false positive rate</th>
<th>sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reveel</td>
<td>0.021%</td>
<td>95.80%</td>
</tr>
<tr>
<td>Reveel+Beagle</td>
<td>0.020%</td>
<td>95.92%</td>
</tr>
<tr>
<td>Reveel-lite</td>
<td>0.021%</td>
<td>95.80%</td>
</tr>
<tr>
<td>GATK+Beagle</td>
<td>0.035%</td>
<td>95.62%</td>
</tr>
<tr>
<td>glfMultiples+Thunder</td>
<td>0.037%</td>
<td>95.80%</td>
</tr>
<tr>
<td>SNPTools+Beagle</td>
<td>0.035%</td>
<td>95.66%</td>
</tr>
<tr>
<td>GotCloud (w/ filters)</td>
<td>0.007%</td>
<td>91.29%</td>
</tr>
<tr>
<td>Integrated (w/ filters)</td>
<td>0.011%</td>
<td>95.89%</td>
</tr>
</tbody>
</table>
Performance on 1KGP Trios

- **SNP discovery**
  - Discover 171,734 likely polymorphic sites in 26 populations
  - Benchmark: 1KGP Pilot2 Trios

<table>
<thead>
<tr>
<th>Method</th>
<th>false positive rate</th>
<th>sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reveel</td>
<td>0.031%</td>
<td>97.06%</td>
</tr>
<tr>
<td>Reveel+Beagle</td>
<td>0.031%</td>
<td>97.53%</td>
</tr>
<tr>
<td>Reveel-lite</td>
<td>0.031%</td>
<td>97.06%</td>
</tr>
<tr>
<td>GATK+Beagle</td>
<td>0.040%</td>
<td>97.38%</td>
</tr>
<tr>
<td>glfMultiples+Thunder</td>
<td>0.048%</td>
<td>98.17%</td>
</tr>
<tr>
<td>SNPTools+Beagle</td>
<td>0.044%</td>
<td>96.81%</td>
</tr>
<tr>
<td>GotCloud (w/ filters)</td>
<td>0.011%</td>
<td>91.46%</td>
</tr>
<tr>
<td>Integrated (w/ filters)</td>
<td>0.023%</td>
<td>98.32%</td>
</tr>
</tbody>
</table>
Performance on 1000 Genomes Data

- **Genotyping**
  - Benchmark: Complete Genomics samples

R: Reveel; RL: Reveel-lite; R+B: Reveel+Beagle; g+T: glfMultiples+Thunder; S+B: SNPTools+Beagle; G+B: GATK+Beagle
Read Clouds for Cancer
Genome Diploid Sequencing
Shortcomings of Short Reads

- Unphased genotypes
- No variants detected in high-fidelity repeats (6% of genome)
- Low-accuracy structural variants
- Challenging regions such as HLA
1. Sample DNA is sheared into fragments of about 10 kbp

2. Fragments are diluted and placed into 384 wells

3. Fragments are amplified through long-range PCR, cut into short fragments and barcoded

4. Short fragments are pooled together and sequenced
10x System

Massively Parallel Partitioning

10X Instrument & Reagents

Read Clouds (“linked reads”)

Phased 60Kb deletion
Read Clouds

Synthetic Long Reads (SLR):
\[ C_R \geq 50x \]

Read Clouds:
\[ C_R < 1x \]

Coverage = \( C_F C_R \)
Identifying Variation in Segmental Duplications

~180 Mbp of human in almost exact repeats
Novel variation impossible to detect with short reads alone
Single unique nucleotides (SUNs) in segdups can be leveraged by read clouds
Candidate Cloud and Alignment Generation

- Use Bowtie2 to align each well separately to the reference
MRF-based Realignment

Generative model of read cloud generation of set of reads $R$

$$P(R) = \sum_{\{\text{Underlying molecule set } M\}} P(M) \ P(R \mid M)$$

Markov Random Field (MRF)
- Read alignment quality
- Mate pairs biased to map together
- Reads biased to form clouds

Optimize with Simulated Annealing
- Move a read
- Move a pair of reads
- Move a cloud of reads
Validation on Simulations

Baseline: Bowtie2

Naive: Naive policy of using alignments to abbreviated reference

Oracle: Pick the true alignment among Bowtie2 alignments

Previously Dark Regions

<table>
<thead>
<tr>
<th>Element Class</th>
<th>frequency (%)</th>
<th>illuminated (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>all</td>
<td>100.0</td>
<td>90.6</td>
</tr>
<tr>
<td>annotated</td>
<td>88.4</td>
<td>90.5</td>
</tr>
<tr>
<td>segdup</td>
<td>43.4</td>
<td>93.8</td>
</tr>
<tr>
<td>LINE</td>
<td>35.2</td>
<td>88.3</td>
</tr>
<tr>
<td>SINE</td>
<td>14.2</td>
<td>92.7</td>
</tr>
<tr>
<td>gene</td>
<td><strong>7.0</strong></td>
<td><strong>95.5</strong></td>
</tr>
<tr>
<td>LTR</td>
<td>6.3</td>
<td>92.3</td>
</tr>
<tr>
<td>Simple repeat</td>
<td>4.6</td>
<td>85.9</td>
</tr>
<tr>
<td>Satellite</td>
<td>2.3</td>
<td>88.1</td>
</tr>
<tr>
<td>Low complexity</td>
<td>1.6</td>
<td>89.0</td>
</tr>
</tbody>
</table>
• Normal (FFPE)
  • Sequenced by:
    – Shotgun (40X)

• IDC (Fresh Frozen)
  • Sequenced by:
    – Shotgun (40X)
    – Moleculo (78X)
    \[
    C_F=43, \ C_R=1.8
    \]

Align reads
Create clouds
Realign reads
Call Variants (GATK)
Phase

(a) Read cloud sizes
(b) Read Cloud Genome Coverage
97% of segdups mappable

Multiplex PCR 346 SNVs, 94% validated

Results on IDC Sample - SNVs

Novel SNVs within 6% dark genome:
131,957 Heterozygous, 65,529 Homozygous
Phasing – an MCMC approach

Sequence reads
Reference Genome

Genotype

Haplotypes
## Germline

<table>
<thead>
<tr>
<th>Het</th>
<th>Hom</th>
<th>Somatic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,948,144</td>
<td>1,266,460</td>
<td>4,689</td>
</tr>
</tbody>
</table>

## Diploid Genome

### $h_1$

### $h_2$
Move 1: assign a cloud to a haplotype
Move 2: unwind a switch error
Move 3: flag/unflag clouds as mixed

\[ h_1 \]

\[ h_2 \]
Separate somatic haplotypes
- Local phase N50

- Switch Errors

- Statistical phase N50
<table>
<thead>
<tr>
<th></th>
<th>Local</th>
<th>LD</th>
<th>Aneuploidy</th>
</tr>
</thead>
<tbody>
<tr>
<td>N50</td>
<td>80 Kbp</td>
<td>533 Kbp</td>
<td>26 Mbp</td>
</tr>
</tbody>
</table>
Moleculo on NA12878 - SNVs

- 3 lanes of Moleculo sequencing by Illumina
  - $C_F = 21x$; $C_R = 0.8x$ on average
- Within 6% repeat DNA, where coverage is sufficient:
  - **50,314** novel mutations (35,092 heterozygous)
  - 9,651 homozygous, 99.5% validated ($\geq 2$ SLR reads)
  - 24,333 heterozygous, 92.0% validated

Antonnaci, Eichler et al.
986kbp BACs within SDs of high structural variation:
- RFA calls 301 novel SNVs, of which:
  - 126 homozygous, 97.6% in BACs
  - 175 heterozygous, 53% in BACs
### 10X on NA12878 – Phasing

<table>
<thead>
<tr>
<th>Mean Coverage</th>
<th>63X</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_F, C_R$</td>
<td>200x, 0.3x</td>
</tr>
<tr>
<td>N50 Phase Block</td>
<td>20.6Mb</td>
</tr>
<tr>
<td>SNP Short Switch</td>
<td>0.3%</td>
</tr>
<tr>
<td>SNP Long Switch</td>
<td>0.001%</td>
</tr>
<tr>
<td>% SNPs Phased</td>
<td>97.0%</td>
</tr>
<tr>
<td>% Genes Fully Phased (&lt;100kb)</td>
<td>94.4%</td>
</tr>
</tbody>
</table>
10X on NA12878 – SNV detection

RFA-10X
• 10X team, with help from Alex Bishara
  – Fast version of RFA, part of 10X suite

SNVs: 4,474 k (1,666 k hom)
Recovered: 238 k (39 k hom)

Baseline: Illumina Platinum Genomes, 200x

<table>
<thead>
<tr>
<th></th>
<th>Moleculo SLR</th>
<th>Eichler BACs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Region Overlap</td>
<td>130,870</td>
<td>749</td>
</tr>
<tr>
<td>Homozygous Overlap</td>
<td>99.6%</td>
<td>100%</td>
</tr>
<tr>
<td>Heterozygous Overlap</td>
<td>88.2%</td>
<td>52.9%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>% of SNVs</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Segmental Duplications</td>
<td>54%</td>
</tr>
<tr>
<td>LINE</td>
<td>22.3%</td>
</tr>
<tr>
<td>SINE</td>
<td>9.5%</td>
</tr>
<tr>
<td>Genes</td>
<td>7.8%</td>
</tr>
<tr>
<td>LTR</td>
<td>7.5%</td>
</tr>
<tr>
<td>Satellite</td>
<td>5.5%</td>
</tr>
<tr>
<td>Simple repeats</td>
<td>1.3%</td>
</tr>
<tr>
<td>Low complexity</td>
<td>0.4%</td>
</tr>
</tbody>
</table>
SMN1 Gene: Associated w/ Spinal Muscular Atrophy
SMN1 & SMN2: 500K w/ >99% identity
Variants in Clinically Relevant Genes

PMS2: DNA repair (associated w/ Lynch Syndrome)
Variants in Clinically Relevant Genes

PMS2: DNA repair (associated w/ Lynch Syndrome)

- PacBio
- Standard Illumina
- Linked-Reads