In pursuit of perfect personal genomes

Michael Schatz

Feb 13, 2018
AGBT Informatics
“Without a doubt, this is the most important, most wondrous map ever produced by humankind.”

June 26, 2000
• The “reference” doesn’t represent *any* human
• Your sample may contain unique genes, gene structures, and other sequences not in the reference
• Mapping short reads to the reference can bias the results
• The reference can limit analysis of how genome variant impact regulation and expression or allele-specific features

• De novo assembly, while greatly improved, is still slow, demanding and unpredictable

“Without a doubt, this is the most important, most wondrous map ever produced by humankind.”

June 26, 2000
Reference Guided Assembly

1. **High quality reference**
   - Contig N50 over 1Mbp
   - Scaffold N50 over 10Mbp
   - High Quality Gene Annotation
   - Your sample is sufficiently similar (~99%)

2. **Sample specific data**
   - **SNPs and Indels**: Illumina-based (PE/10X)
   - **Structural Variants**: Long PacBio/ONT
   - **Phasing Data**: 10X and/or HiC; trios

**Comparative Genome Assembly ("AMOScmp")**
Reference Guided Assembly

1. **High quality reference**
   - Contig N50 over 1Mbp
   - Scaffold N50 over 10Mbp
   - High Quality Gene Annotation
   - Your sample is sufficiently similar (~99%)

2. **Sample specific data**
   - **SNPs and Indels**: Illumina-based (PE/10X)
   - **Structural Variants**: Long PacBio/ONT
   - **Phasing Data**: 10X and/or HiC; trios

*Data requirements similar to de novo, but less demanding, more accurate, and more predictable*
CrossStitch

https://github.com/schatzlab/crossstitch

In collaboration with Sedlazeck, Gingeras, Guigo, Ring, & Gerstein labs
CrossStitch
https://github.com/schatzlab/crossstitch

HQ Reference

A -> C
Call SNPs + Indels

A/C -> A|C
Phase SNPs + Indels

GAT -> XXX
Call SVs

GAT/XXX -> GAT|XXX
Phase SVs & Local Assembly

Phased SNPs, Indels & SVs.vcf

Assemble Diploid Sequence

my.mat.fa
my.pat.fa
CrossStitch
https://github.com/schatzlab/crossstitch

A -> C
Call SNPs + Indels

A/C -> A|C
Phase SNPs + Indels

GAT -> XXX
Call SVs

GAT/XXX -> GAT|XXX
Phase SVs & Local Assembly

Phased SNPs, Indels & SVs.vcf

Assemble Diploid Sequence

my.mat.fa
my.pat.fa

HQ Reference
Phasing Results

Phase Block 1

Unphased

Phase Block 2

Read length (log10 bp)

N50 Phase block size (log10 bp)

Short Reads: ~1kbp

Linked Reads: ~10Mbp

Long Reads: ~500kbp

Hi-C: 100Mbp+

10X + Hi-C: 145Mbp N50 😊
(Healthy human, varies by sample)
CrossStitch

https://github.com/schatzlab/crossstitch

HQ Reference -> Call SNPs + Indels

A -> C -> A/C -> A|C
Call SNPs + Indels
Phase SNPs + Indels

GAT/XXX -> GAT|XXX
Phase SVs & Local Assembly

Phased SNPs, Indels & SVs.vcf

Assemble Diploid Sequence

my.mat.fa
my.pat.fa
NGMLR + Sniffles

NGMLR: Convex gap penalty to balance frequent small sequencing errors with larger SVs
Sniffles: Scan within and between split reads to accurately find SVs (Ins, Del, Dup, Inv, Trans)
Mendelian concordance >95%, experimental validation also very high

Accurate detection of complex structural variations using single molecule sequencing
Illumina data

Truncated reads:

Missing pairs
Truncated reads:

Insertion detected by long reads

Missing pairs
SVs in a typical healthy human

### Sniffles calls

<table>
<thead>
<tr>
<th></th>
<th>All SVs (50bp+)</th>
<th>Large SVs (10kbp+)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deletions</td>
<td>7,389</td>
<td>164</td>
</tr>
<tr>
<td>Duplications</td>
<td>1,284</td>
<td>139</td>
</tr>
<tr>
<td>Insertions</td>
<td>8,382</td>
<td>4</td>
</tr>
<tr>
<td>Inversions</td>
<td>229</td>
<td>116</td>
</tr>
<tr>
<td>Translocations</td>
<td>170</td>
<td>170</td>
</tr>
<tr>
<td><strong>All</strong></td>
<td><strong>17,454</strong></td>
<td><strong>593</strong></td>
</tr>
</tbody>
</table>

### Translocation in Ribbon

*Ribbon: Visualizing complex genome alignments and structural variation*

Nattestad et al. (2016) *bioRxiv* doi: http://dx.doi.org/10.1101/082123
Long-read genome sequencing identifies causal structural variation in a Mendelian disease
CrossStitch
https://github.com/schatzlab/crossstitch

HQ Reference

A -> C
Call SNPs + Indels

A/C -> A|C
Phase SNPs + Indels

GAT -> XXX
Call SVs

GAT/XXX -> GAT|XXX
Phase SVs & Local Assembly

Phased SNPs, Indels & SVs.vcf

my.mat.fa

Assemble Diploid Sequence

my.pat.fa
Hybrid Phasing and Local Assembly

**Phased Short Read Variants**

**Phased Long Reads and Structural Variants**

*Phase SVs*: Determine the haplotype of each read and each SV

*Local Assembly*: Refine sequence of insertions, resolve complex nested variants
CrossStitch
https://github.com/schatzlab/crossstitch

- A -> C
  - Call SNPs + Indels

- A/C -> A|C
  - Phase SNPs + Indels

- GAT -> XXX
  - Call SVs

- GAT/XXX -> GAT|XXX
  - Phase SVs & Local Assembly

- Phased SNPs, Indels & SVs.vcf

- Assemble Diploid Sequence

HQ Reference

my.mat.fa

my.pat.fa
Assembling a “Perfect” Personalized Diploid Genome

Phased SNPs, Indels & SVs.vcf

G A T T A C A

G A T T A C A
Assembling a “Perfect” Personalized Diploid Genome

Phased SNPs, Indels & SVs.vcf

sub inversion

G C T G T A A

G C T T T A C A A A

sub insertion del

G C T T T A C A A A - A
Assembling a “Perfect” Personalized Diploid Genome

Phased SNPs, Indels & SVs.vcf

Stitching based on AlleleSeq pipeline enhanced for SVs (Rozowsky et al, 2011)
• Maintains a mapping from reference to personal genome coordinates for liftover

Using 10X + HiC + PacBio, assemble nearly perfect diploid human genomes
• Phased diploid genome can be aligned or aligned against
**Improved mapping of functional data**

- Typically 10k – 100k additional mapped RNA-seq reads per sample; mappability more complicated

**Expression of deleted genes and promoters**

- Heterozygous or homozygous deletions of genes and promoters often show reduced expression

**SVs intersecting eQTLs**

- Deletions overlapping a SNP eQTL affects the expression of the target gene; further analysis in progress
Reference-quality Genomes without de novo assembly

Why should we assemble perfect personal genomes?
• Pathogenic and other important variants might be missed
• Improved mapping, fixes “differential” expression, allele-specific
• Explore interplay between variation, regulation, and expression

Multiple sequencing technologies & approaches needed
• >20x coverage PacBio/ONT: Best Resolution of SVs
• >20x coverage 10X/HIC: Best Phasing
• Trio or Population-based phasing also possible to reduce costs

We have just begun to explore the universe of variants present
• Also need to push these ideas into single cell and population scale analysis

http://schatz-lab.org
Acknowledgements

Schatz Lab
Mike Alonge
Amelia Bateman
Charlotte Darby
Han Fang
Michael Kirsch
Sam Kovaka
Laurent Luo
Srividya
Ramakrishnan
T. Rhyker
Ranallo-Benavide
*Your Name Here*

Baylor Medicine
Fritz Sedlazeck

CSHL
Gingeras Lab
McCombie Lab

GRC
Roderic Guido
Alessandra Breschi
Anna Vlasova

University of Vienna
Arndt von Haeseler
Philipp Rescheneder

DNA nexus
Maria Nattestad

PacBio
Greg Concepcion

ENCODE Partners
Berstein Lab
Gerstein Lab
Myers Lab
Ren Lab
Snyder Lab
Stam Lab
Wold Lab
+ All ENCODE Members

NSF
National Human Genome Research Institute
U.S. Department of Energy
ALFRED P. SLOAN FOUNDATION
Thank you

http://schatz-lab.org
@mike_schatz
SVs using short, long, and linked reads

Main Diagonal
• Calls per tool

Outer triplets
• Concordance by Technology

Inner triplets
• Concordance by Assembly
• Concordance by Mappers

Overall:
• Lonnnnnnng reads give the most variants with the best concordance 😊
**Expression & Regulation**

*Foundation for mapping functional data*
- Discover novel genes and gene fusions
- Analyze differential expression in CNVs
- Discover new regulatory regions, allele-specific binding and expression

**Population Genetics**

*Framework for GWAS of Structural Variations*
- Many GWAS SNPs appear to be in linkage with SVs that are the likely functional variant
- Resequencing key individuals with phenotype data available

**Tumor Progression**

*Chromosome instability in breast cancer*
- 10X, PacBio and Oxford Nanopore sequencing of breast cancer samples from Northwell Health
- Cell lines, patient tissues, and patient-derived organoids
Analysis in progress…

• Construct personal genome and personal annotation for all individuals

• Expression changes due to SVs overlapping functional elements, i.e. enhancers, eQTLs SNPs and short indel analysis

• Novel transcription elements in insertions

• Chimeric transcripts in reference and personal genomes

• Allele specific expression and binding

• Integrate other functional assays to perform tissue specific analysis, i.e. smallRNAs, RAMPAGE, ChiP-seq

• … and many more …