‘Next-gen’
Y chromosome sequencing: progress and promise

2014 International Genetic Genealogy Conference
Washington, D.C., August 16, 2014

Greg Magoon
gmagoon@aerodyne.com
Why do we care about ‘Y’?

an American family Bible from 1859; David Ball: www.davidball.net
The Y chromosome as a “Genetic Family Bible”

- Y chromosome is inherited paternally
  - No recombination*
  - Small variations from generation to generation arise from essentially random [germline] mutations

- Genetic testing and comparison of Y chromosomes can help determine where they have been, and when they have been there; useful in:
  - Genetic genealogy
  - Genetic anthropology
  - Forensics

*No recombination means that genetic material does not mix with the other gender chromosome.
The spectrum of Y chromosome variation

- Various mutation events produce different types of genetic sequence variants
  - SNPs (single nucleotide polymorphisms)
    - single-base substitution
    - typically arise from point mutations
  - INDELs (insertions and deletions)
  - MNPs (multiple nucleotide polymorphisms)
    - e.g. AT > CC
  - STRs (short tandem repeats)
  - SVs (structural variants)
    - CNVs (copy number variations)
      - Large scale deletions, duplications and insertions
    - Inversions
    - Substitutions
    - Translocations
What is ‘next-gen’ sequencing (NGS)?

• High-throughput sequencing approaches that are alternatives to conventional Sanger sequencing
• Has become much more cost-effective over the last ~5 years
• Much of this talk is based around Illumina NGS platforms (based on Solexa technology)
• But there are many others:
  – Ion Torrent
  – Pacific Biosciences
  – Oxford Nanopore
  – Complete Genomics
  – SOLiD
  – 454
  – Helicos
  – …
Why ‘next-gen’ sequencing for ‘Y’?

• Previously: analysis has relied on Y-STRs and a relatively small number of known SNP markers
• NGS data can be used to identify new genetic markers (and determine results for known markers) across wide regions of Y
• Identify significant numbers of novel SNP markers
  – SNPs are relatively easy to test and identify with NGS data
  – Point mutations tend to be very stable (not prone to back mutation)
  – Opportunity for very precise haplogroup determination
• Allows for accurate, high-resolution phylogenies
  – Unlocks the potential of the Y chromosome to provide a comprehensive record of paternal ancestry ranging from a few generations back to hundreds of thousands of years ago
  – “Molecular clock”: Potential for more accurate age estimates
How is it done?

DNA

fragmentation, target enrichment, and amplification

DNA “library”
(many fragments hundreds of bp long)

sequencing
(read ~100 bp from each end of the fragment)

sequence data
for millions of “reads”, each ~100 bp long

ATCGCGCGTAG...
GGTTGATCAAA...
GTGGCTGACGA...

• Library construction is key
NGS data analysis

Data analysis is non-trivial:
• Gigabytes of data
• No direct information on where reads come from on genome
• Data can contain errors or be of low quality

The traditional pipeline:
  sequence data (FASTQ format)
  mapping reads to reference sequence
  aligned sequence data (BAM format)
  variant calling
  list of differences from reference genome

<table>
<thead>
<tr>
<th></th>
<th>chromosome</th>
<th>position</th>
<th>variant</th>
</tr>
</thead>
<tbody>
<tr>
<td>chrY</td>
<td>2574932</td>
<td>T</td>
<td>A</td>
</tr>
<tr>
<td>chrY</td>
<td>3137270</td>
<td>G</td>
<td>T</td>
</tr>
<tr>
<td>chrY</td>
<td>4472711</td>
<td>AG</td>
<td>A</td>
</tr>
</tbody>
</table>

Unique aspects for Y chromosome:
• Haploid, non-recombining chromosome
• Many repetitive and highly similar sequences
• Opportunity to perform phylogenetic analysis to inform the “Y tree”
  – Detailed comparisons between results for different individuals are essential
Sources of NGS Y data

• Public datasets
  – Individual sequences (James Watson, ancient genomes, etc.)
  – 1000 Genomes Project
  – Personal Genome Project

• Restricted-access research datasets
  – Population studies
  – Clinical sequencing studies

• Commercial products
  – Full Genomes Corporation (FGC)
  – Big Y
But, some words of caution...

- Many false positives and misleading results
  - Need to separate the wheat from the chaff
- Unstable sites
- Reference confusion
How can we use all these data?

**Generation of high-resolution *a priori* Y-chromosome phylogenies using “next-generation” sequencing data**

Gregory R Magoo, Raymond H Banks, Christian Rottensteiner, Bonnie E Schrack, Vincent O Tilroe, Terry Robb, Andrew J Grierson

doi: 10.1101/000802

**ABSTRACT**

An approach for generating high-resolution *a priori* maximum parsimony Y-chromosome (chrY) phylogenies based on SNP and small INDEL variant data from massively-parallel short-read (next-generation) sequencing data is described; the tree-generation methodology produces annotations localizing mutations to individual branches of the tree, along with indications of mutation placement uncertainty in cases for which "no-calls" (through lack of mapped reads or otherwise) at particular
NGS Y-tree construction methods

• Key features of the approach:
  – *A priori* approach, not based on existing tree
  – Careful pre-filtering of SNP and INDEL variants
  – Uses an iterative reweighting approach
  – Ancestral state reconstruction is used, to place individual mutations on the branches of the tree
  – Takes missing data (‘no-calls’) into account
The tree construction “pipeline”

“next-gen” sequencing data
1292 individual BAM files
300+ GB (mapped to Y)

variant calling (SAMtools)

variant call file (VCF)
437,000+ variants; 500 million+ variant calls
8 GB

variant call post processing and filtering; conversion into PHYLIP input format

PHYLIP input file
45,427 variants (excluding singletons); 58 million+ variant calls
58 MB

iterative tree generation with PHYLIP and postprocessing of results

tree with mutations placed on individual branches
120,000+ variants (with singletons); 2,285 tree branches
10 MB
A high-resolution Y-tree...
...with a few new insights into the branching of the “tree trunk”
Tree results can also inform sequence stability and erroneous results.

*consistency with known phylogeny can be a powerful indicator of marker reliability*
Sources of error in next-gen sequencing

DNA

library construction

sequencing

mapping

variant calling

variant list

- DNA may be altered or biased
  - PCR stutter
- Miscalling of bases
- Skipping or rereading bases
- Inaccurate or inadequate reference sequence
- Misalignment of reads
- Ambiguous alignment of reads
- Overlooking true variants
- Identification of false positive variants
There are important differences between NGS Y tests

|                        |  
|------------------------|---
|                        | **Y chromosome capture** | **whole-genome sequencing** |
|                        | Big Y | FGC Y Prime | FGC Y Elite | low-coverage (Illumina) | high-coverage (Illumina) | high-coverage (Illumina, longer read) | high-coverage (CGI) |
| “callable” sites, chrY (bp) | 8.4 million | 13.3 million | 14.0 million | 3.0 million | 13.9 million | 15.7 million | 10.2 million |
| “known” Y-SNP sites with calls (out of 47,042) | 36,124 | 43,817 | 46,135 | 40,011 | 46,490 | 45,715 |
| “novel” Y-SNP sites with calls (out of 6,867) | 4,109 | 6,403 | 6,771 | 5,694 | 6,801 | 6,751 |

a point of reference
Advanced analysis: structural variation

- Copy-number variation analysis can be performed to identify large duplications and deletions
- These variants appear to be phylogenetically useful in some cases:
  - A ~7,500 bp duplication shared by a small group, including an STR that had gone from one copy to two copies
  - Large deletion in a subgroup of R1b-U106 (approx. with Z5/Z8 markers)
- There are likely many more examples like this, spanning a large spectrum of age, prevalence, and stability
Future opportunities

• DNA sampling
• Next-gen sequencing technology
  – Lower cost
  – Longer read technologies
  – Epigenetic information
• Data analysis
  – Reference sequence refinement
    • “Build 38”
    • Reconstruction of ancestral sequence?
    • Taking into account common structural variants?
  – Refinement of the Y-tree
  – Mutation rates (→ better age estimates)
• Insights into the broader human genome
  – Better understanding of mutation mechanisms and rates
  – Guide testing and refinement of NGS data analysis pipelines
Odds and ends

• Sharing BAM files
  – Google Drive
    • allows files up to 10 GB
    • can easily control sharing

• Visualization tools:
  – IGV (Broad Institute)
  – SAMtools
  – BAMview
Odds and ends: FGC

- FGC project initiative
  - Send your thoughts, ideas, etc. to FGCprojectadvisors mailing list (join via goo.gl/7hQvXq)
- whole-genome sequencing (pilot: $1850)
- FGC original Y product now named Y Elite
  - Y coverage comparable to whole-genome sequencing
  - Cost: $999
- FGC new product: Y Prime
  - Most comprehensive Y sequencing product besides Y Elite
  - Y Prime introductory sale, $589
    - use coupon code YPRIME through August 31
- Donate your third-party BAM files to Full Genomes Y chromosome comparison database
  - e-mail file link to fgcfilesharing@gmail.com
- www.fullgenomes.com ; sales@fullgenomes.com
Acknowledgements

• Prof. Andy Grierson, Univ. of Sheffield
• Thomas Krahn, YSEQ.net
• Vince Tilroe, David Reynolds, Rich Rocca, Ray Banks, Bonnie Schrack, Chris R., Peter Op den Velde Boots, Terry Robb, Charles Moore, Raymond Wing, Wayne Kauffman, Mike Maddi, and many others
• Full Genomes Corporation
  – Justin Loe
  – Leon Kull
• FGC customers and BAM file donors