Trio Analysis: A Reproducible Example

Andrew Borgman

VARI BBC

September 12, 2013

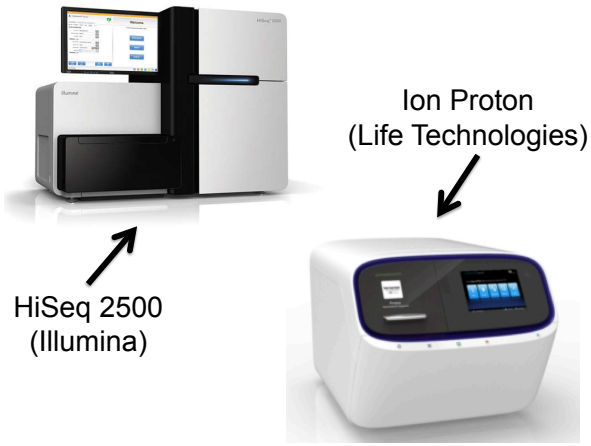
Overview

1. Next Generation Sequencing Technologies
2. Trio Analysis
3. Hometown Translational Opportunity
4. Reproducible Research
5. Questions?

Overview

NGS Platforms

- Massively parallel DNA sequencing
- Rapidly decreasing costs coupled with increasing yields
  - Moore's law joke
- Trend toward sequencing centers
  - Bringing manufacturing efficiency & scalability to NGS
- $1,000 genome?
  - Not yet (~$4,500 for 30x human)



So you have 1,000,000,000 reads...

Whole Genome Sequencing (WGS)

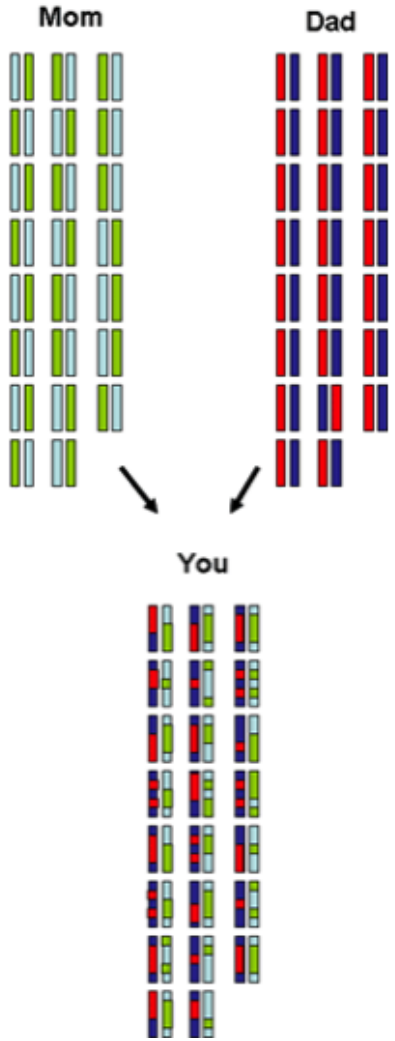
- Map all DNA reads and align to reference genome (Alignment)
- See where your reads differ from reference (Variant Calling)
  - Try to detangle sequencing errors from true mutations
- Facilitates genome-wide scan for mutations
  - Single nucleotide variants (A -> T)
  - Small insertions or deletions (GATTACA -> GACA)
  - Structural Variants (part of chr 7 is now in chr 3)
- Assess importance and impact of mutations
  - Annotation, prioritization, etc.

What's all the fuss about?

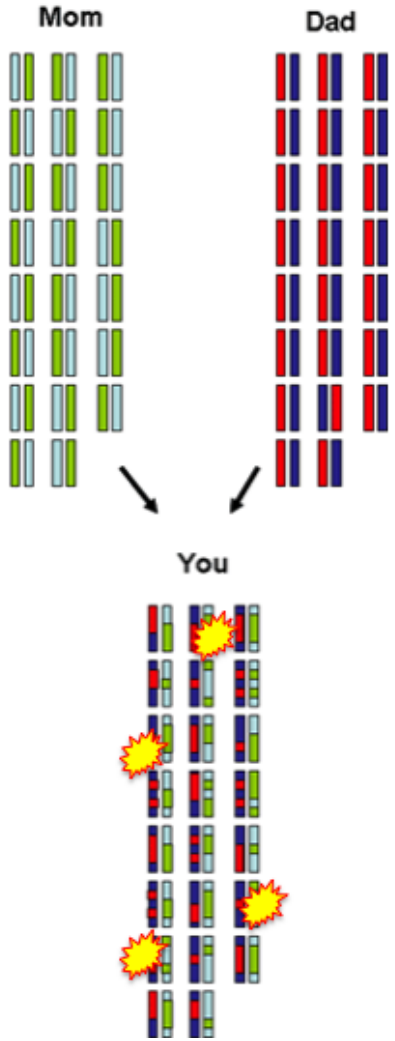
- WGS provides individual mutational profile
  - Genetic risk assessment and preventative therapies
- Cheap enough for population-scale studies
  - Huge consortium efforts
  - GTEX, TCGA, etc.
- One step closer to personalized medicine

Overview

Where do you come from?



Where do you come from?



Trio Analysis Concepts

- Prioritizing sea of variants from sequencing study is hard
  - ~3-4 million variants per sequencing run
  - Which ones do I care about?
- Trio Analysis: NGS design to increase detection power
- Perform WGS/WES on parents and affected offspring
  - Use resulting variant calls for trio-informed analysis
- Extend to multiple families for greater power

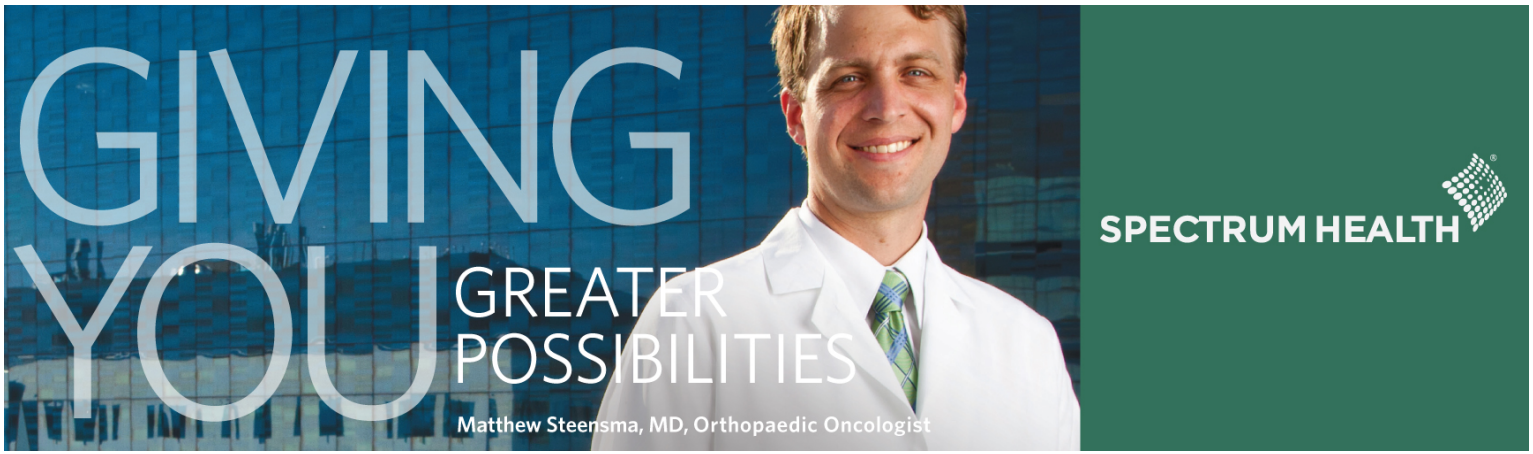
Trio Analysis Advantages

- Heritable diseases
  - Guide analysis w/ inheritance pattern
  - Variants in offspring should be seen in parents
  - Dominant, recessive, etc.
- For non-heritable diseases
  - Looking for variants not seen in unaffected parents
  - Identification of *de novo* mutations
  - Can use siblings as additional filter
- Improves variant calling procedures
  - Probabilistic trio-aware variant calling
- Used for both rare & common diseases
  - Inform on genetic disease etiologies
  - Confirm/formulate diagnosis for syndrome

Overview

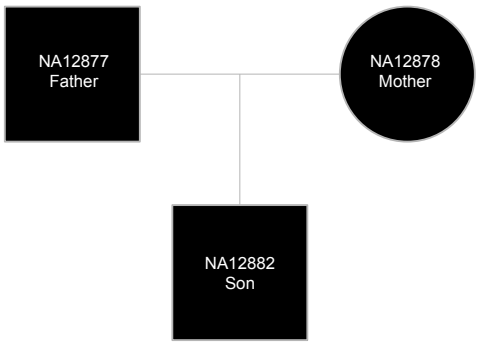
Supporting treatment across the street

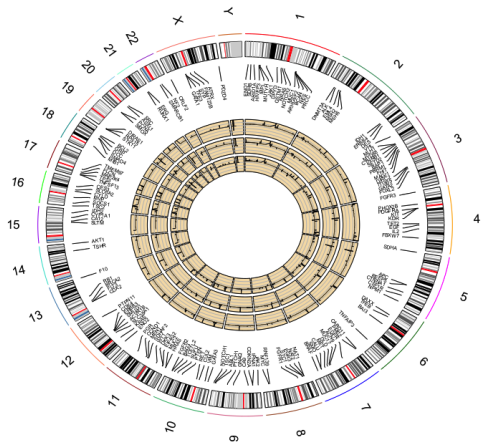
- Collaboration with Matt Steensma's lab
- Patient w/ rare, genetically uncharacterized disease
  - 16 individuals in U.S. w/ similar disease
  - No clear inheritance pattern
- Part of his *Outliers* initiative



Our contributions

- Consulting on NGS study design
- Forming collaborations with MSU Genomics Core
- Implementing trio analysis workflow in-house
  - Based on MIT/Broad's best practices
  - Variant calling, filtering and annotating
  - SV/CNV detection
- Trained with Illumina "Platinum Genomes"





Overview

Want to see how I did this?

- Code for trio analysis is on GitHub
  - Look here: https://github.com/borgmaan/gvsu-symposium
  - Presentation code is there too
- Learning/Thinking about open science initiatives
  - See xenophobia project on GitHub [https://github.com/borgmaan/xenophobia]
  - Internally advocating for "open notebook science"

Ram *Source Code for Biology and Medicine* 2013, **8**:7
http://www.scfbm.org/content/8/1/7

SOURCE CODE FOR BIOLOGY AND MEDICINE

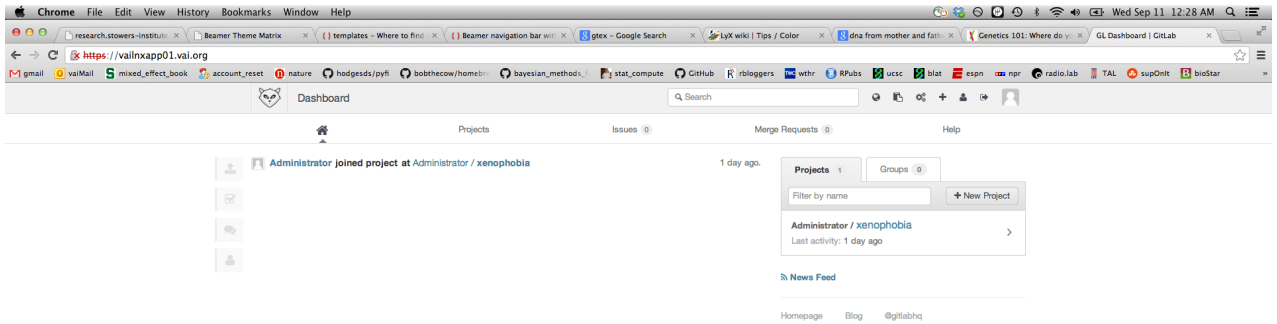
BRIEF REPORTS **Open Access**

Git can facilitate greater reproducibility and increased transparency in science

Karthik Ram

In-house reproducibility efforts

- Implementing version control practices in house
  - Git backed project tracking
- Using GitLab for an internally hosted GitHub
- Just began working on this
- No experiences to share yet...



Why are we trying it?

- Provides high level of integrity in analysis
  - Results can always be reproduced by typing 'make'
  - All figures and tables created directly from data
  - Everything is tracked; no more uncertainty
- Should be providing "publication quality analysis"
  - Our duty as a core service to VARI investigators
  - Publications are asking for data and code; have it ready
- We need more structure
  - Current structure is OK, but it is not effective for code sharing
  - More defined project frameworks == reuse & efficiency
  - I think Git/Make approach would help

Thanks

- Grand Valley State University
  - Thanks for hosting and inviting!
  - Biostatistics Alum '13
- Van Andel Research Institute
  - Bioinformatics & Biostatistics Core
  - Mark Neff, Lab of Canine Genetics and Genomics

Questions?