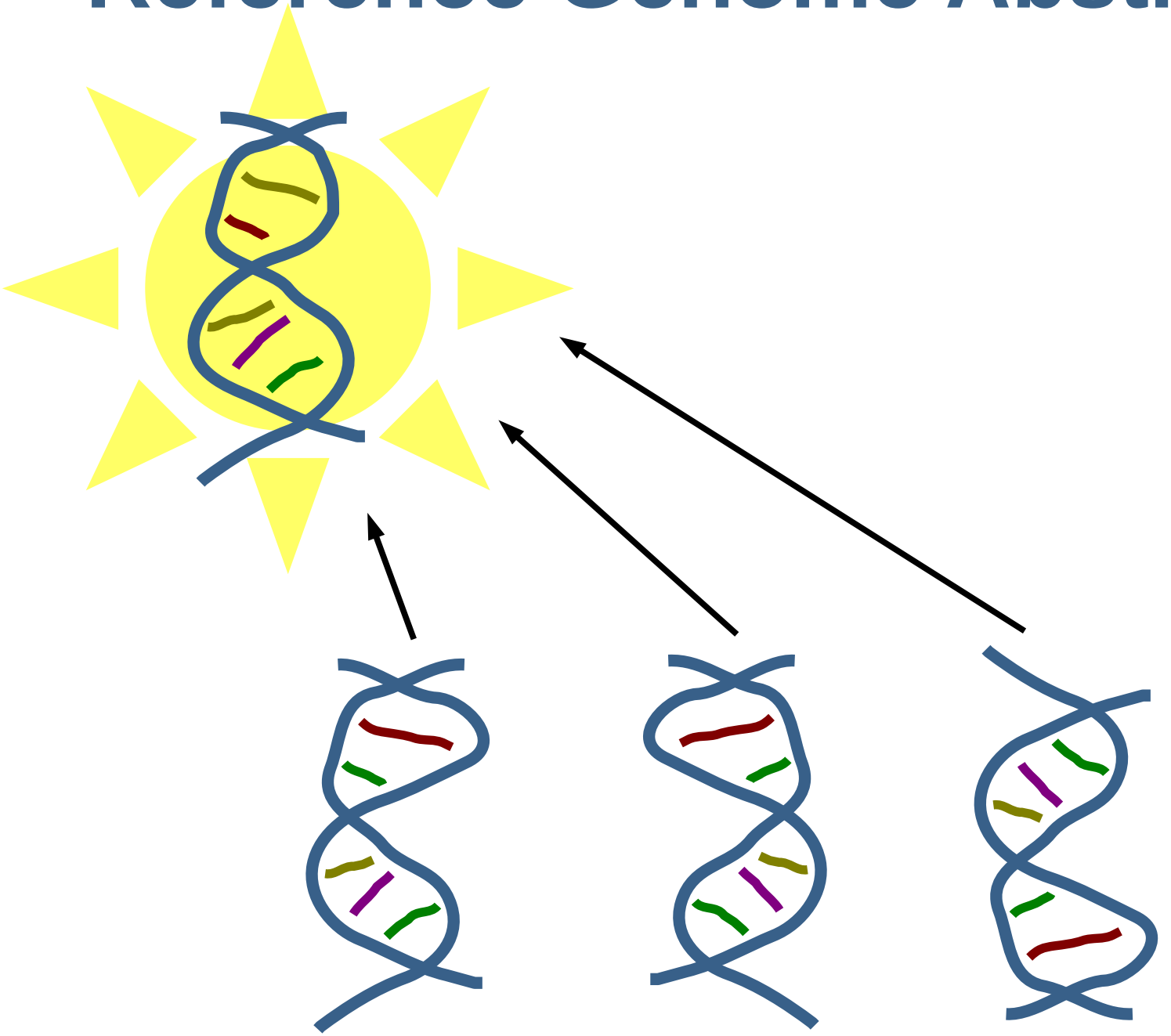


The Human Genome Variation Map

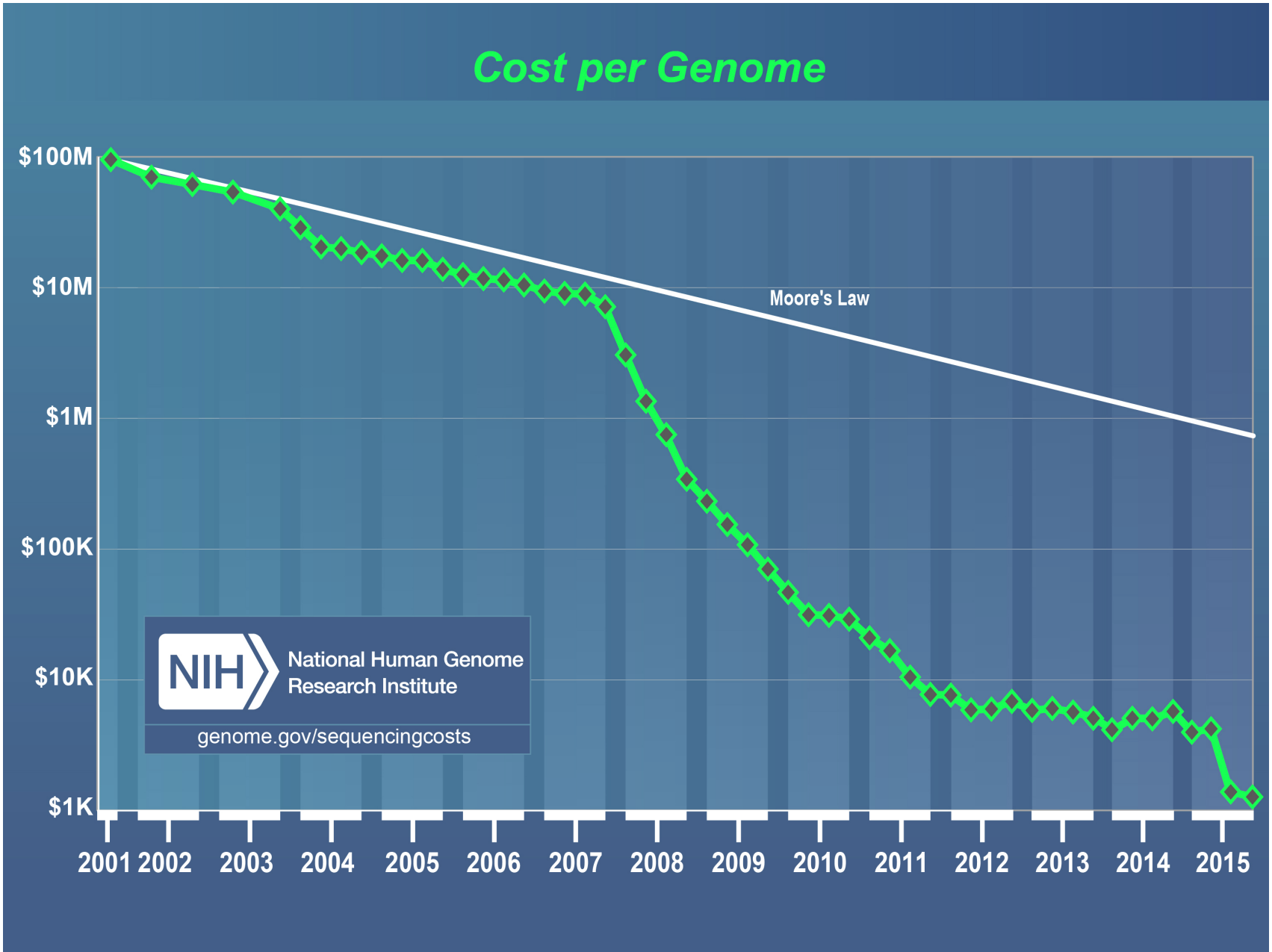
Dr. Adam Novak
Senior Software Engineer
UC Santa Cruz Genomics Institute

Agilent Technologies
Global Alliance for Genomics & Health
W. M. KECK FOUNDATION
NIH Big Data to Knowledge (BD2K)
ARCS ADVANCING SCIENCE IN AMERICA
SIMONS FOUNDATION
UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute

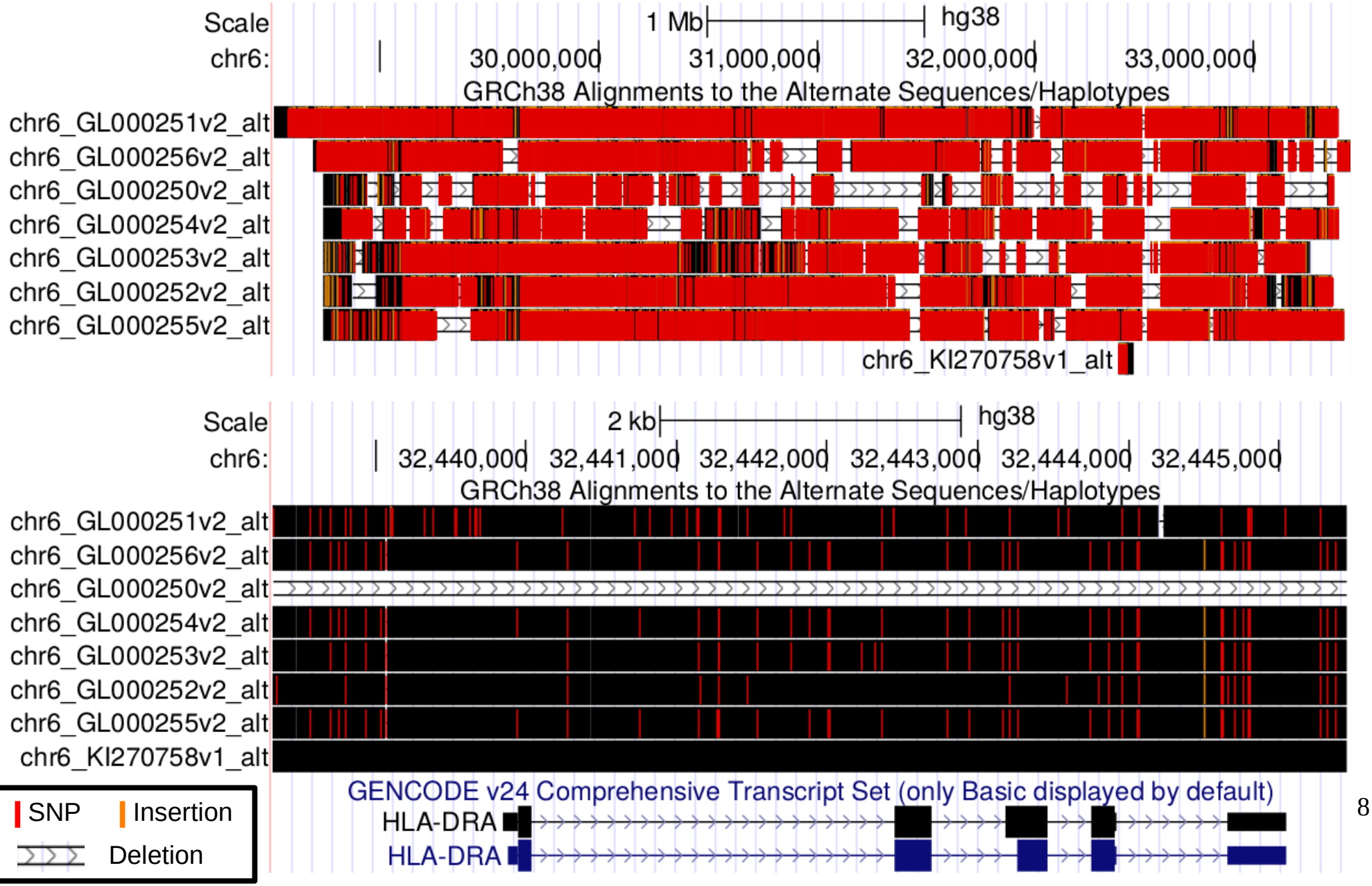
Reference Genome Abstraction



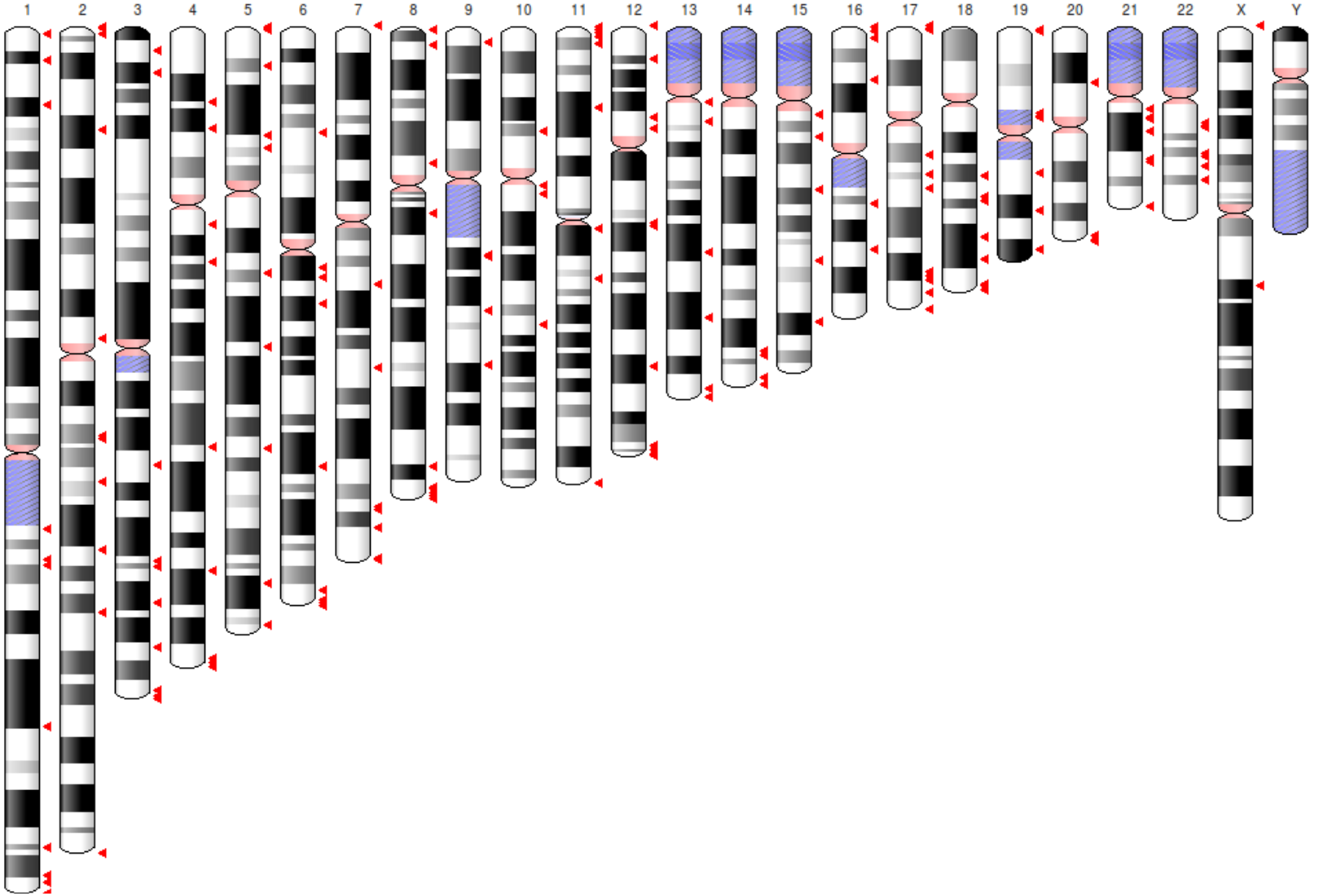
The O($1000) Genome



Some Regions are Highly Variable



187 Holes In The Abstraction



http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/

Proceedings of the National Academy of Sciences of the United States of America

PNAS

Deep sequencing of 10,000 human genomes

Amalio Telenti[a,b,1], Levi C. T. Pierce[a,c,1], William H. Biggs[a,1], Julia di Iulio[a,b], Emily H. M. Wong[a], Martin M. Fabani[a], Ewen F. Kirkness[a], Ahmed Moustafa[a], Naisha Shah[a], Chao Xie[d], Suzanne C. Brewerton[d], Nadeem Bulsara[a], Chad Garner[a], Gary Metzker[a], Efren Sandoval[a], Brad A. Perkins[a], Franz J. Och[a,c], Yaron Turpaz[a,d], and J. Craig Venter[a,b,2]

[a]Human Longevity Inc., San Diego, CA 92121;

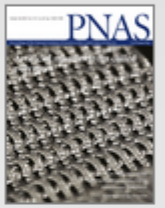
[b]J. Craig Venter Institute, La Jolla, CA 92037;

[c]Human Longevity Inc., Mountain View, CA 94041;

[d]Human Longevity Singapore Pte. Ltd., Singapore 138542

Contributed by J. Craig Venter, August 18, 2016 (sent for review July 1, 2016; reviewed by David B. Goldstein and Stephen W. Scherer)

This Issue



October 18, 2016
vol. 113 no. 42
Masthead (PDF)
Table of Contents

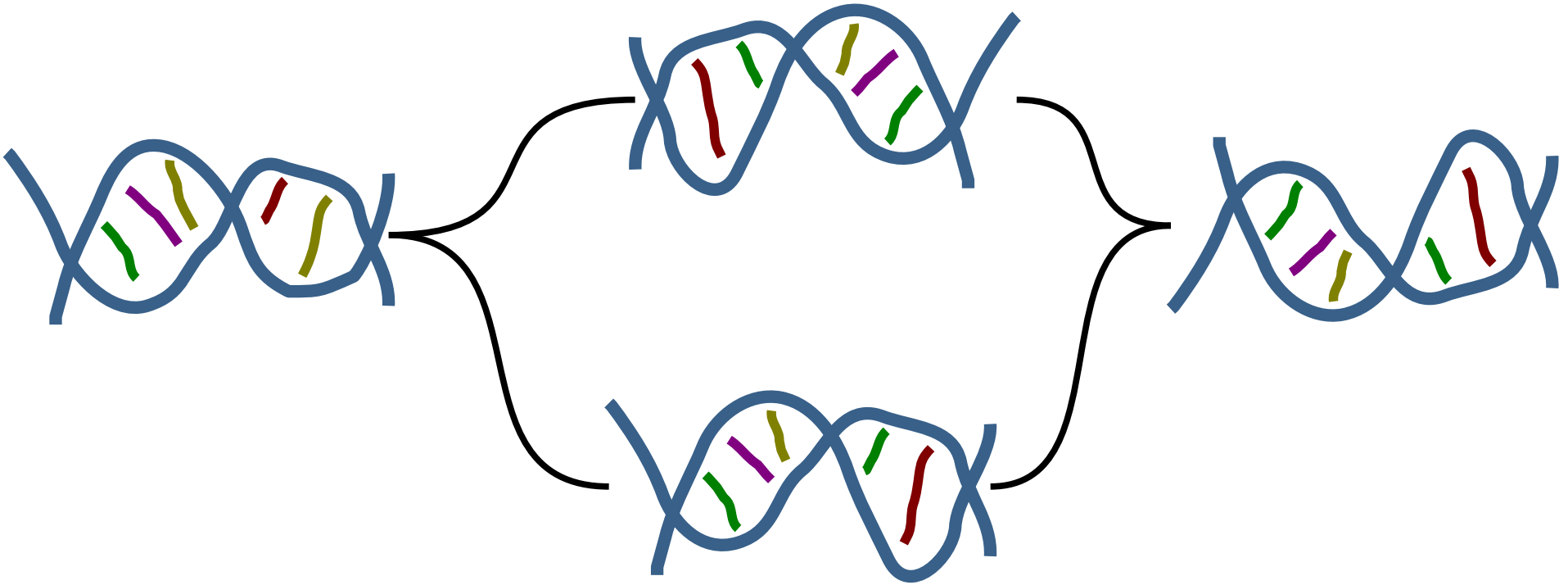
PREV ARTICLE NEXT ARTICLE

Don't Miss

PNAS Full-Text iOS App
Download the app for free from iTunes today!

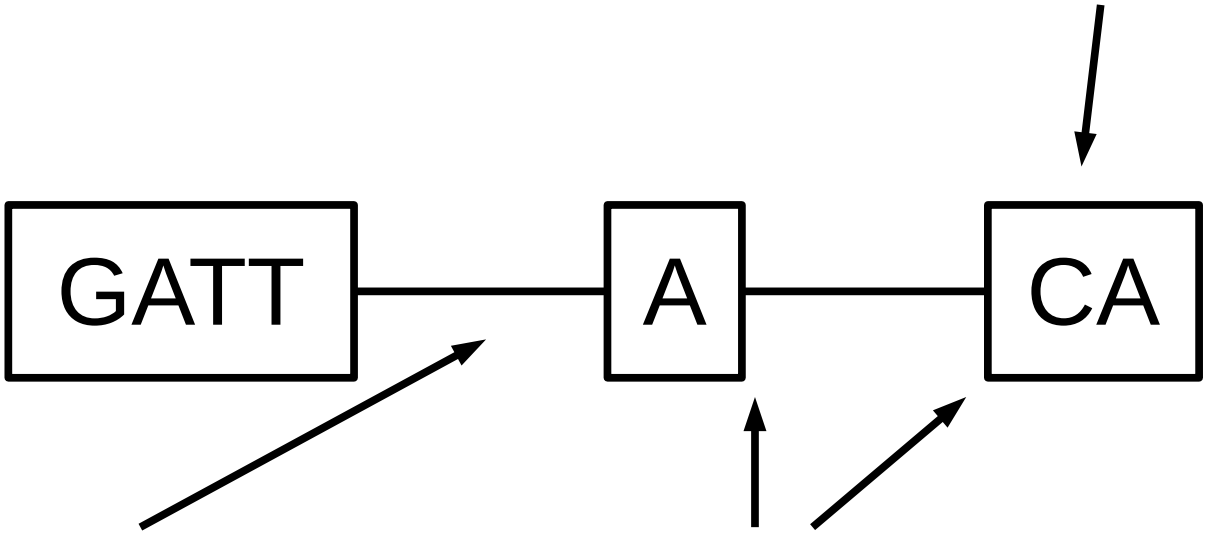
"each genome carries on average 0.7 Mb of sequence that is not found in the main build"

Add Variation to the Reference



Variation Graphs

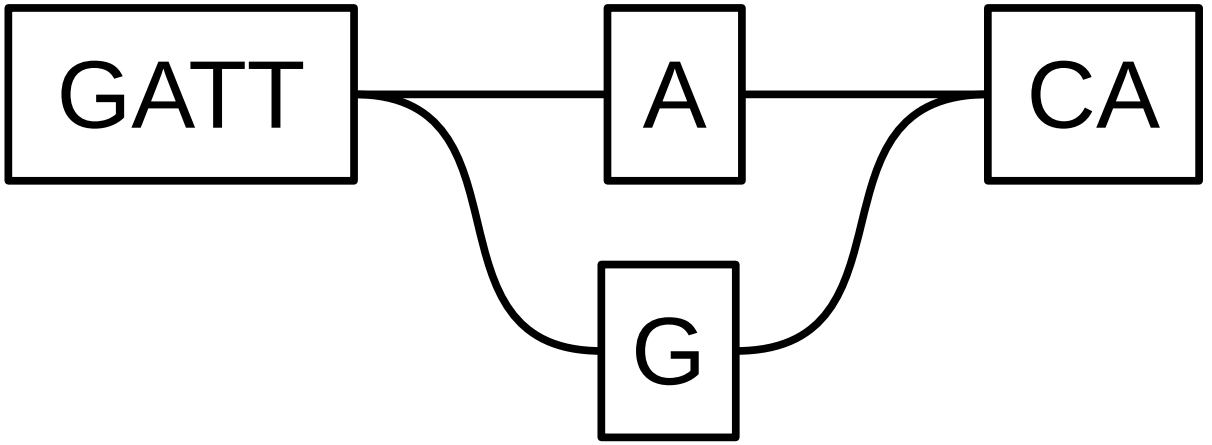
Pieces of sequence are **nodes**



Edges connect **sides** of nodes.

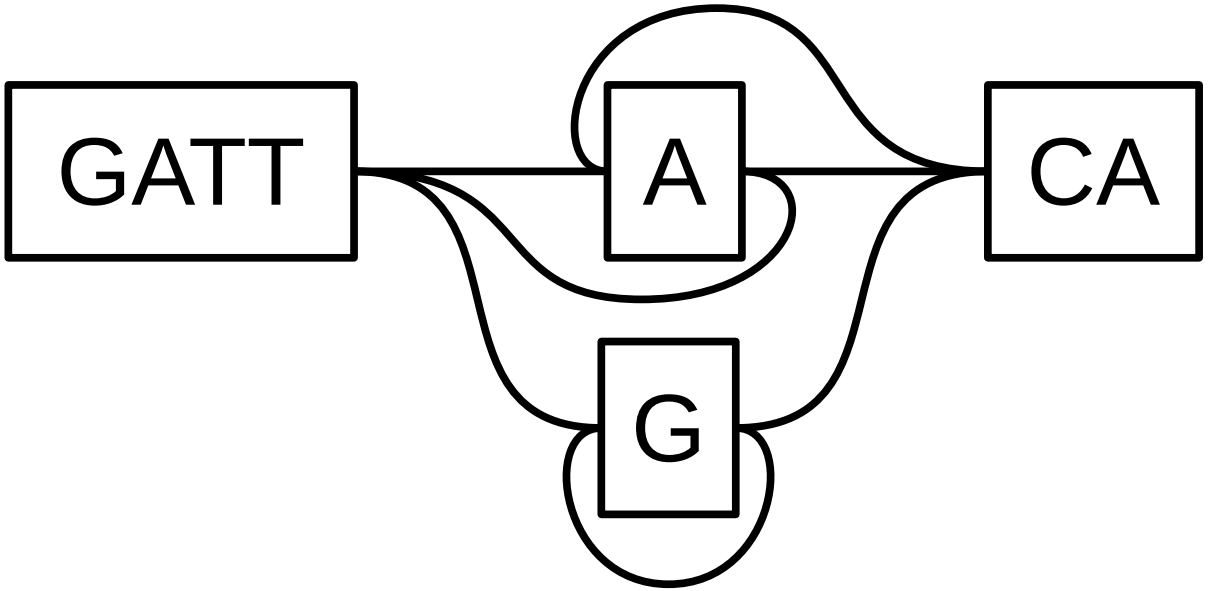
Variation Graphs

Parallel nodes represent alternatives



Expressive Power

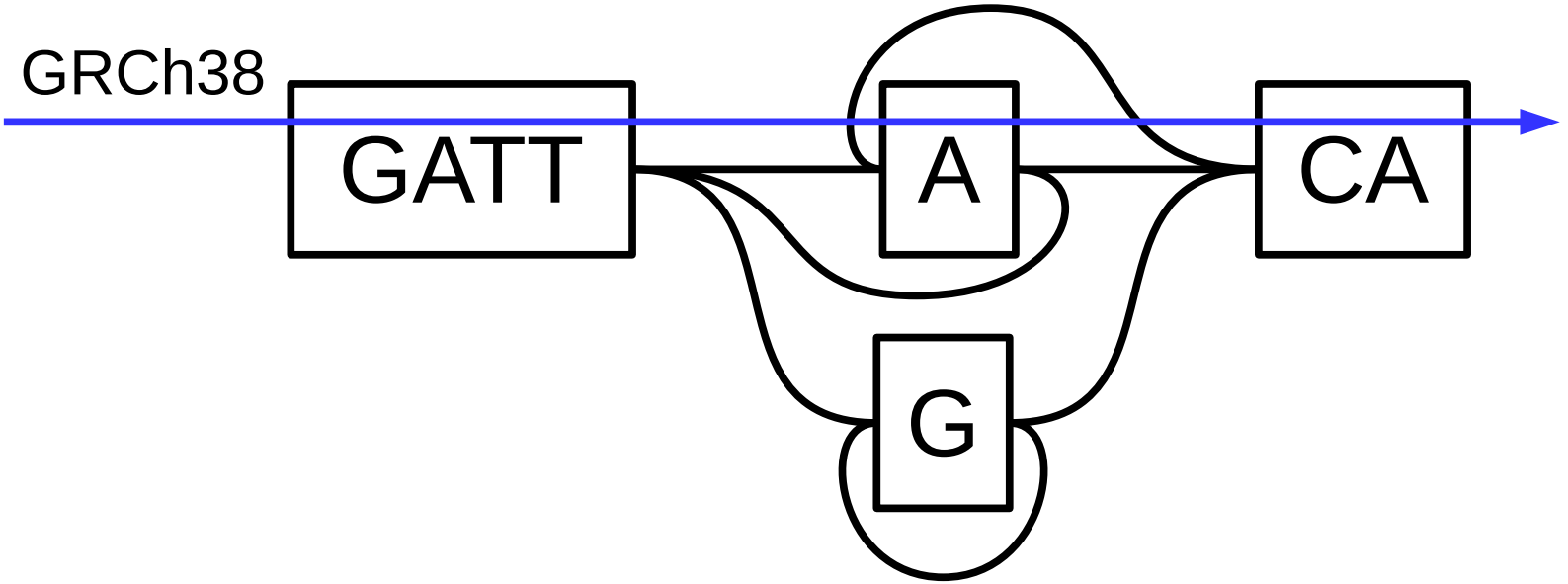
Inversions and duplications are allowed



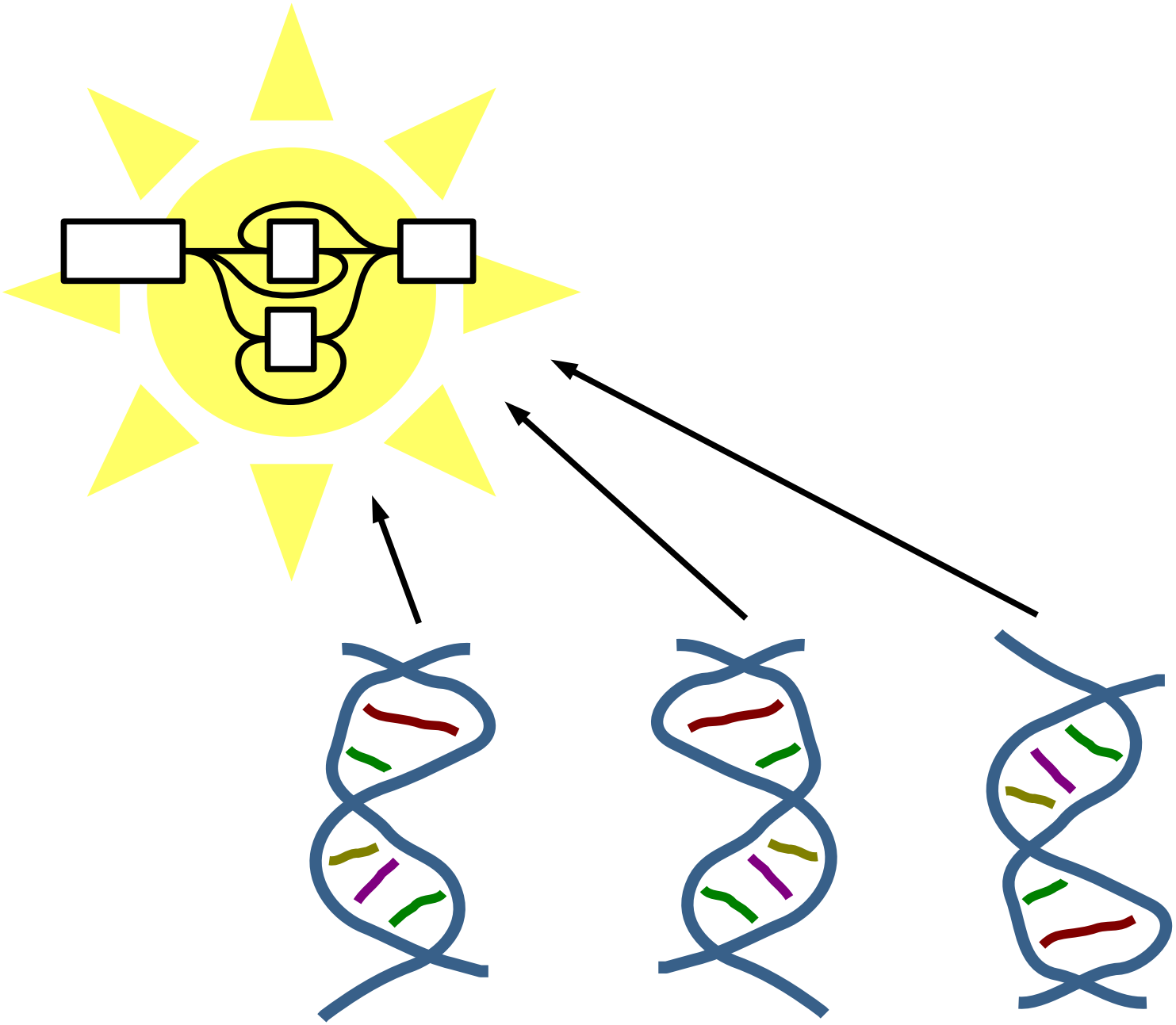
Not just a DAG

Backwards Compatibility

Embed the primary reference path



The Human Genome Variation Map



Eventual HGVM Goals

Build a whole-genome graph reference

Include all variants at 1% frequency in any population

HGVM Graph Strategy

Combine variation data from:

- GRCh38
- 1000 Genomes
- Simons Genome Diversity Project
- Illumina Platinum Genomes
- ...

Immediate Goals

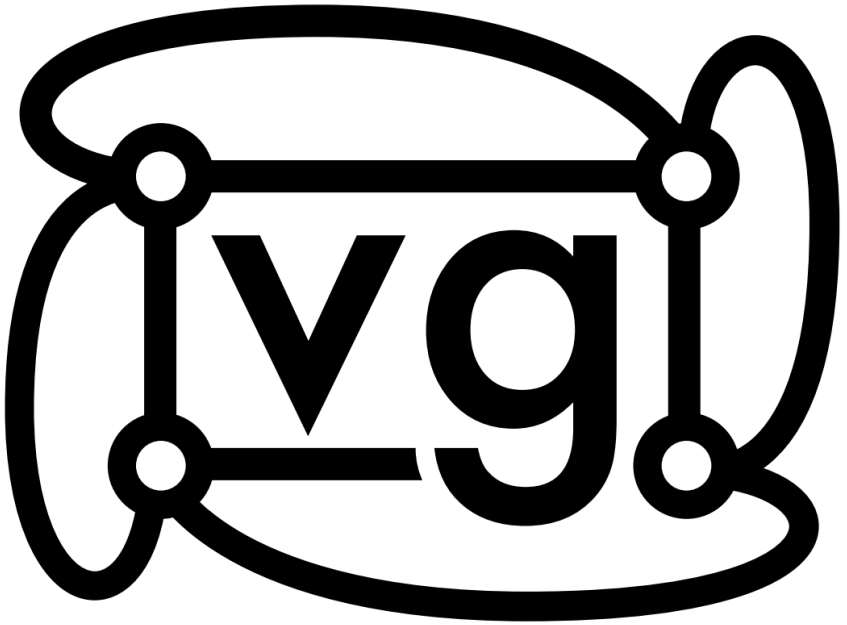
Construct a Chromosome 22 Variation Map

Combine data from multiple sources

Implement pipeline logic

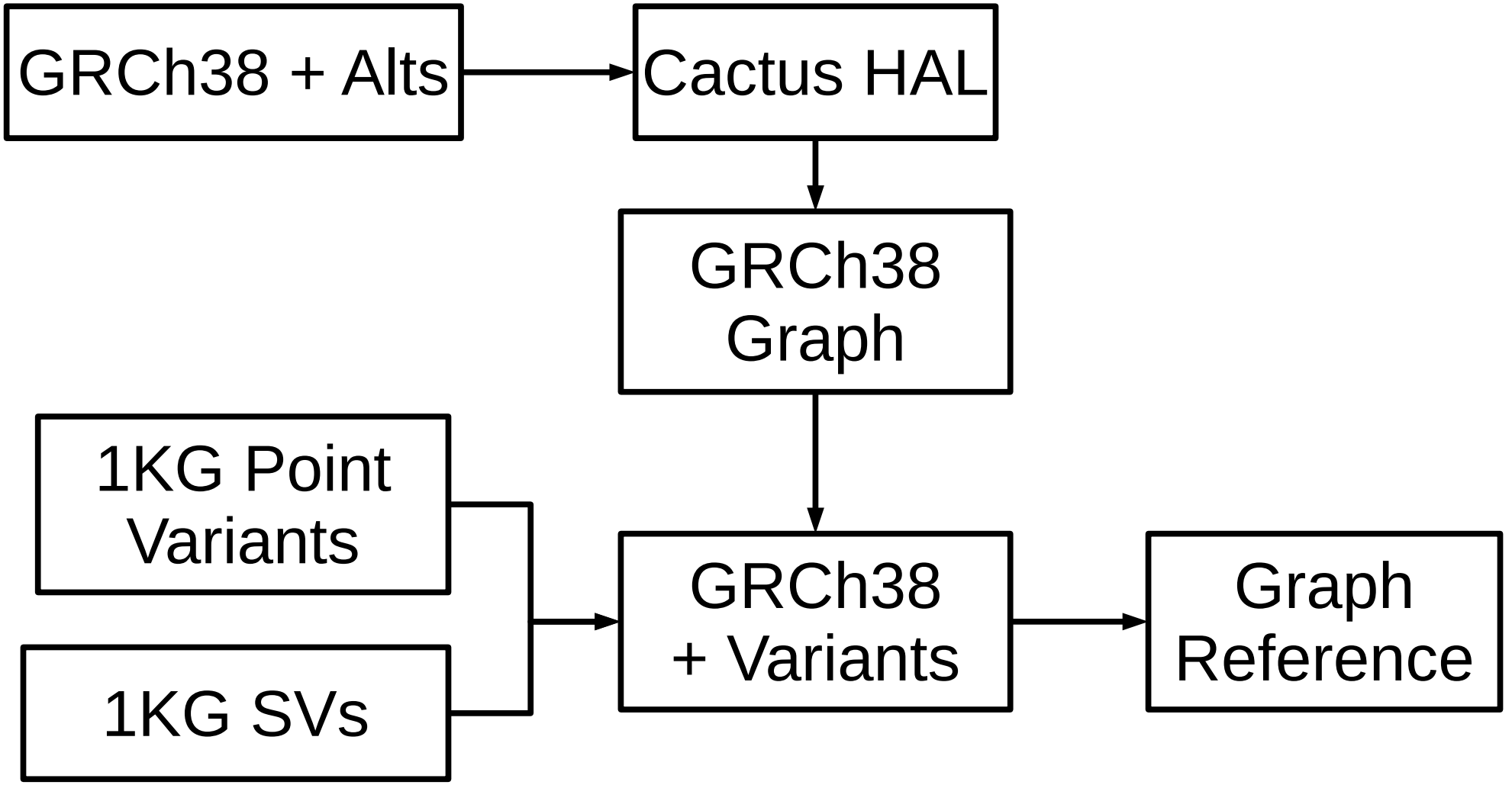
Identify toolchain shortcomings

Tools

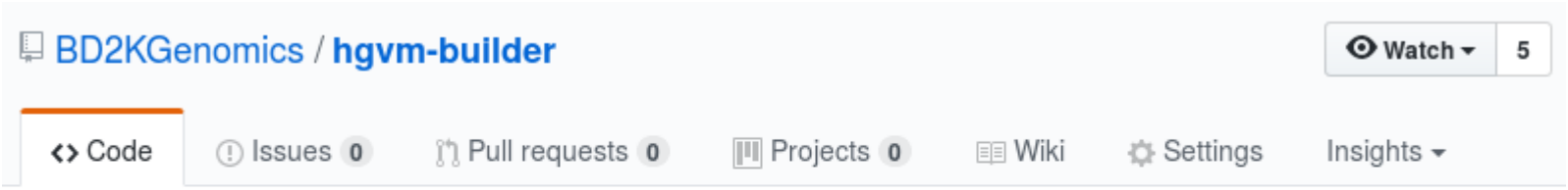

vg

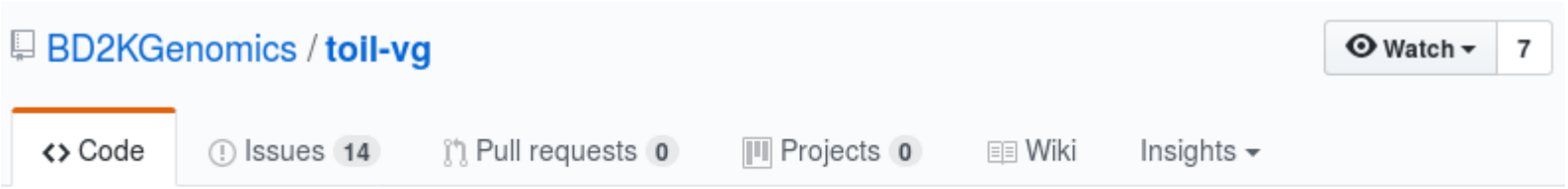

TOIL
Massively Parallel Workflows
(and fire-breathing dragon slugs)

Pipeline


GRCh38 + Alts
Cactus HAL
GRCh38
Graph
1KG Point
Variants
1KG SVs
GRCh38
+ Variants
Graph
Reference

Pipeline Engineering





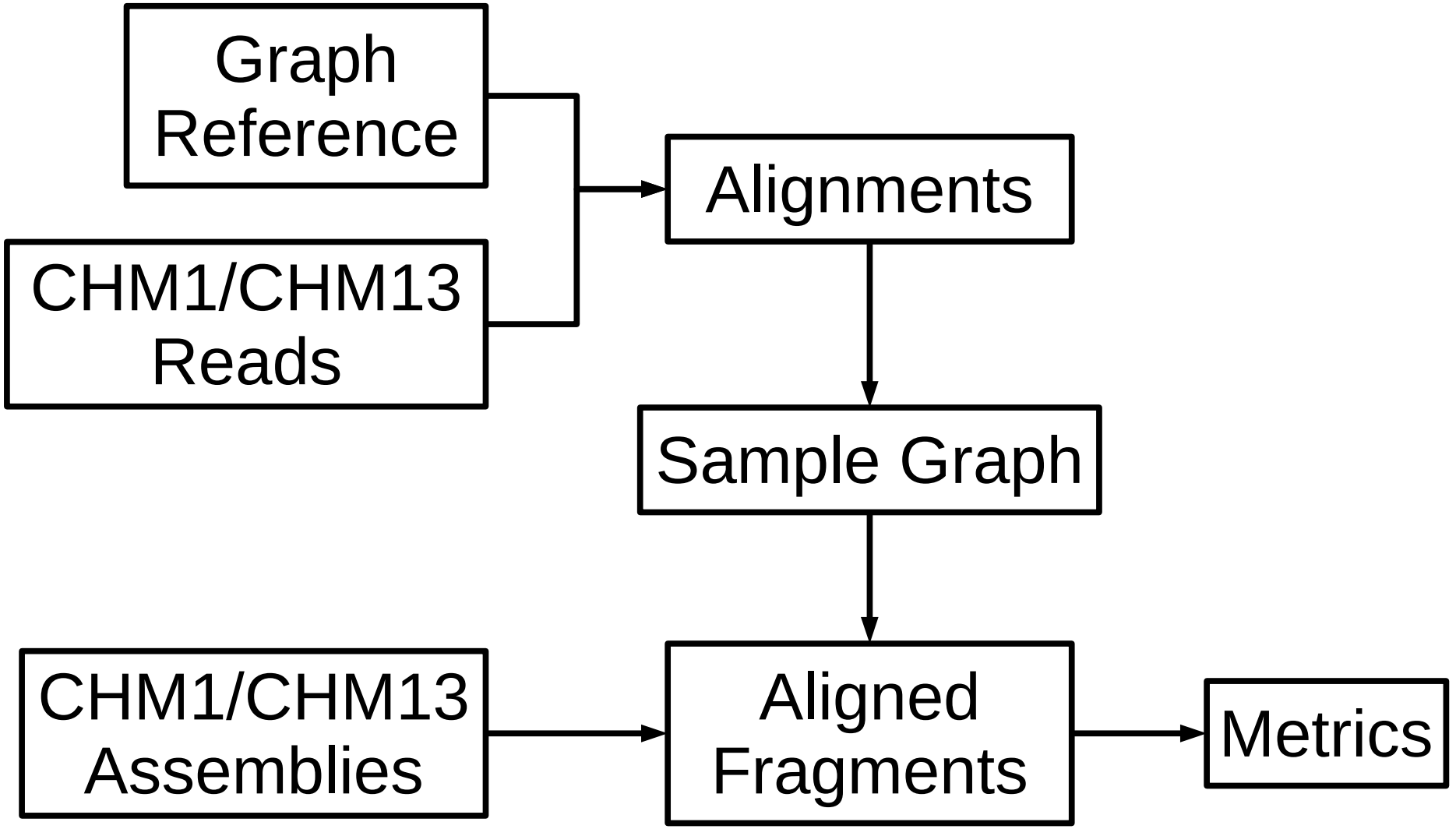
Packaging

Graph and alignment indexes

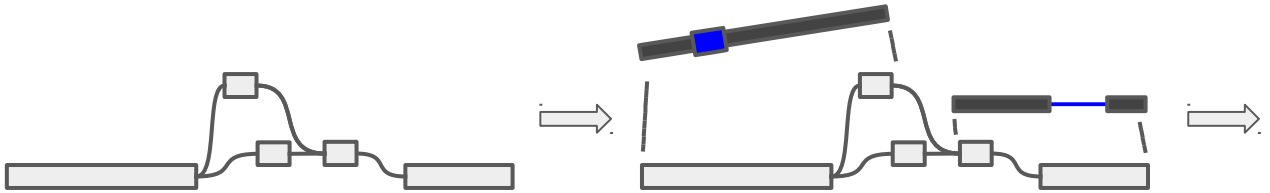
Includes a JSON manifest defining primary paths


```
{
  "hgvm_manifest_version": "0.1",
  "primary_paths": [
    "chr22_KI270737v1_random",
    "chr22",
    "chr22_KI270735v1_random",
    "chr22_KI270734v1_random",
    "chr22_KI270738v1_random",
    "chr22_KI270731v1_random",
    "chr22_KI270732v1_random",
    "chr22_KI270736v1_random",
    "chr22_KI270733v1_random",
    "chr22_KI270739v1_random"
  ],
  "uuid": "9ef69e94-a95f-455e-8fca-f705a334968a",
  "build_time": "2017-05-27 14:05:42"
}
```

Synthetic Diploid Evaluation



Assembly Fragment Alignment



Substitutions: 1 bp
Deletions: 2 bp

Sample Graph

Aligned Assembly Fragments

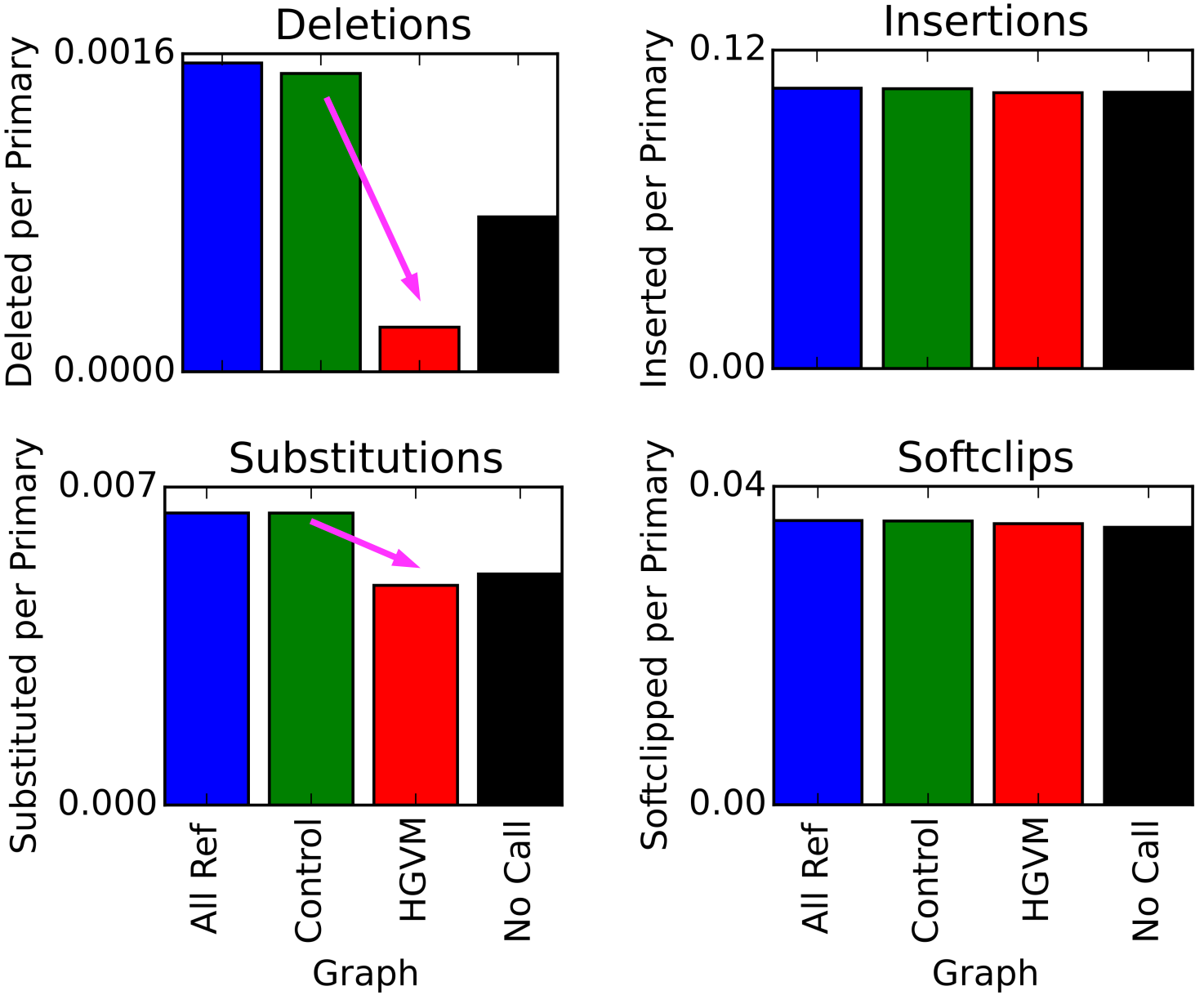
Edit Metrics

Synthetic Diploid Metrics

How many edits of each type are present in assembly fragments aligned to the sample graph?

Fewer edits = better variant calling

Chr22 Results



Future Work

Create read-aware variant caller that handles cycles

Explain synthetic diploid performance

Scale to whole genome

Acknowledgments

David Haussler, Josh Stuart, Beth Shapiro, Ed Green, Benedict Paten, Glenn Hickey, Jordan Eizenga, Yohei Rosen, Charles Markello, Sean Blum, Maciek Smuga-Otto, Kishwar Shafin, Erik Garrison, Jouni Sirén, Mike Lin, Joel Armstrong, CJ Ketchum, John Vivian, Arjun Rao, Eric Dawson, Toshiaki Katayama, Frank Nothaft, Lynn Brazil, Kelly Sauder, Tracie Tucker, Anna Henderson, Richard Novak, Cheryl Covino, The GA4GH, The ARCS Foundation, The Simons Foundation, The W.M. Keck Foundation, Agilent, Microsoft, Amazon, NIH

Questions