GA4GH CONNECT

Code of Ethics and Community Conduct

GA4GH CONNECT

Help us create an inclusive environment

- Treat everyone with **respect**, **empathy**, **consideration**, and **professionalism**
- **Consider other points of view**, **eliminate your own biases**
- **Acknowledge others' contributions**
- Respect GA4GH and venue **policies and rules**.

The following will NOT be tolerated:

- **Abuse** or **harassment** in any form, or threats thereof
- **Violating boundaries** when previously communicated
- **Photographing or recording** others without their consent
- **Disrespectful communication** and any other behaviours that lead to **hostile environment**

Read the **full CECC** on our website

ga4gh.org/code-of-conduct

Have questions or want to report? Contact **Safe GA4GH Officer Paula Brantner** at

conduct@ga4gh.org

Read the **Guidelines for Respectful Behaviour**

bit.ly/ga4gh-respectful-engagement

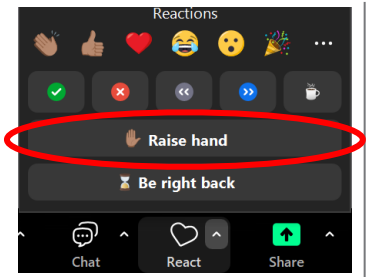
GA4GH CONNECT

In Person

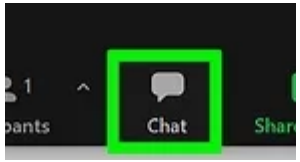
Please use the **handheld microphone** closest to you. If there isn't one nearby, simply **raise your hand**, and a GA4GH staff member will bring one to you.

Virtual

Use the raise hand function to request the floor—wait to be called on before speaking.



Use the chat function for relevant comments



Beatrice will be your in-person representative, relaying questions and comments from virtual attendees.

"My name is <*name*> from <*affiliation*>"

GA4GH CONNECT

The Revolution is Now!

Pangenome Structural Variant Reporting and Representation

Melissa Cline, Karen Miga, Benedict Paten, Glenn Hickey, Heng Li, Nathan Salomonis, Alex Wagner

The pangenome: what is it, exactly?

Benedict Paten

Human Reference Genome

The current human reference genome (GRCh38) is the cornerstone of human genomics

It is a proxy to a universal coordinate system for human genetics

It originally cost $3B and took an act of congress

Released in 2001, it has been refined over 20 years

Originally it was built to represent the euchromatin, it is still incomplete..

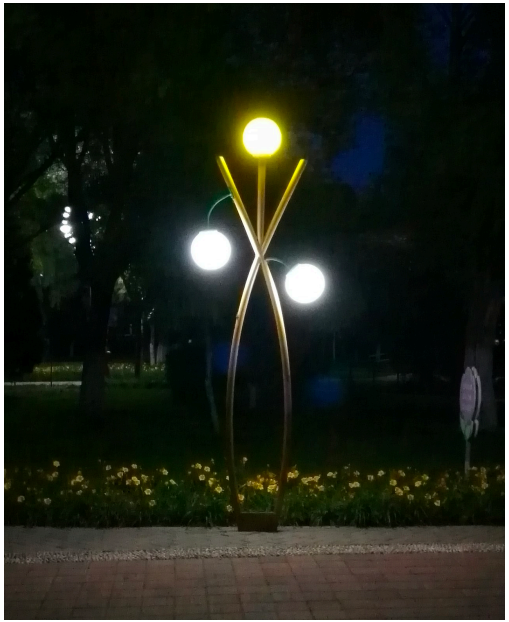
The First Complete, Haploid Human Genome

20 years after the human genome Karen Miga (UCSC), Adam Phillippy (NHGRI), et al. released the first complete assembly of a haploid human g

TELOMERE-TO-TELOMERE CONSORTIUM

Missing Polymorphic Sequence

- There are megabases of commonly polymorphic euchromatic sequence missing from any individual genome
- As a result, no single reference assembly, even a complete one, is optimal for all people, because any reference creates a bias away from the missing sequence



Reference bias is an observational bias, aka streetlamp effect: it is harder to find something not in the reference.

Call to Action: A Human Pangenome Reference

HUMAN PANGENOME

NIH NHGRI

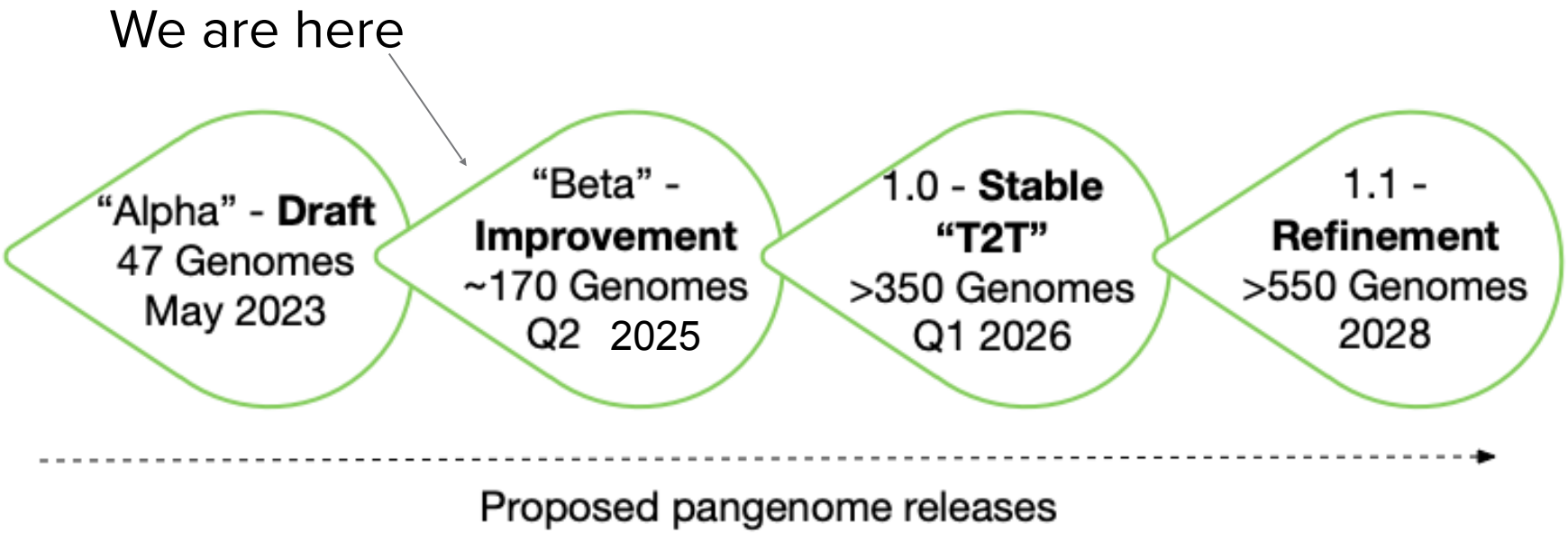
- Better representation of sequence diversity in the human population (>350 diverse humans)
- Comprehensive, public map of genome variation
- New reference data structure and nucleate and foster a new ecosystem of pangenome tools

Now: A Draft Human Pangenome Reference*

- 47 phased, diploid genome assemblies (~1/7th of final cohort)
- Pangenome alignments, annotations
- Pangenome tools and applications

* Liao, Asri, Ebler, et al. A Draft Human Pangenome Reference, Nature, 2023

Soon: Release 2



New "beta" human pangenome release coming this summer!

Human Pangenome

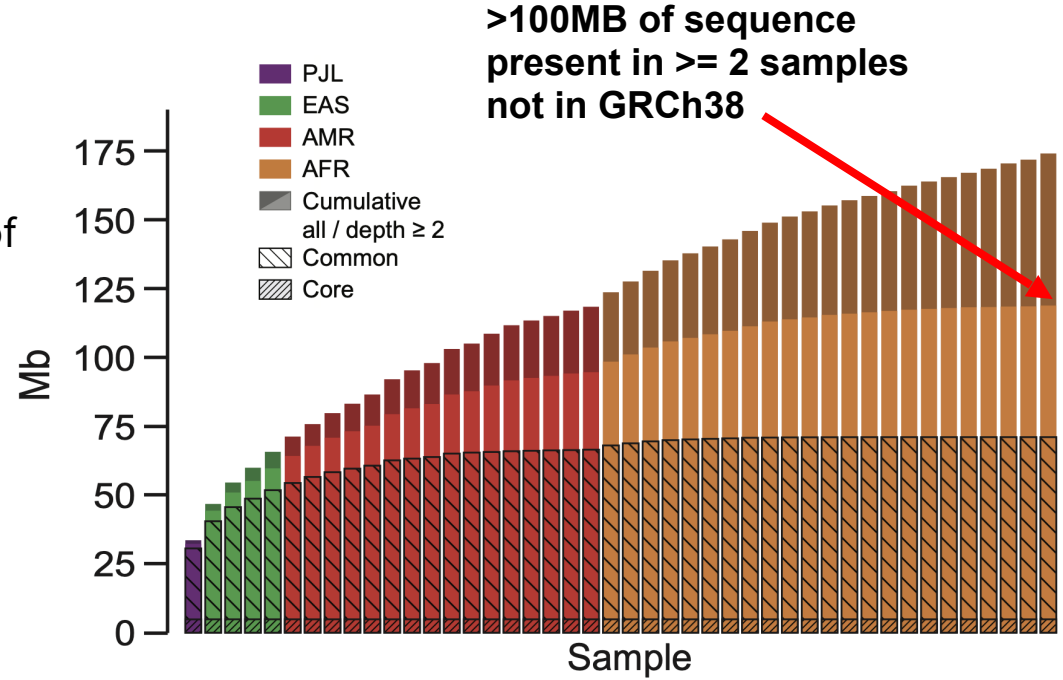
Defined by three As:

- **Assemblies**
  - Haplotype resolved (soon T2T), but also 37, 38, T2T-CHM13.
- **Alignment**
  - Provides canonical homology information
- **Annotations**
  - Genes, etc. Should be consistent with alignment

Goal: provide a comprehensive view of common human variation

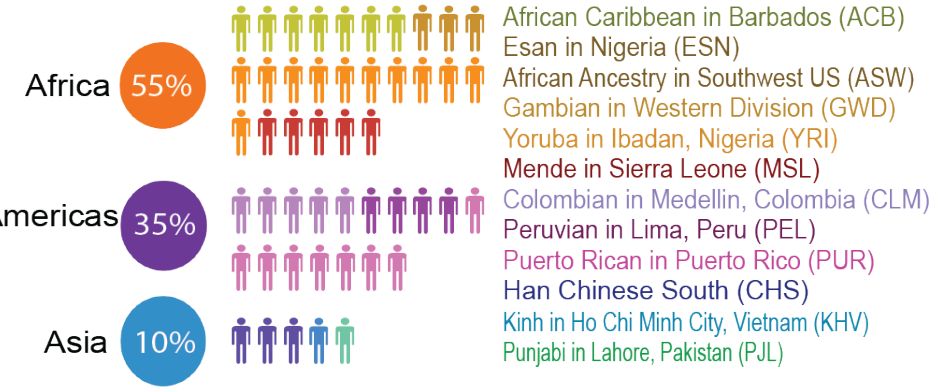
Pangenome: Adding Common, Polymorphic Sequence To The Reference

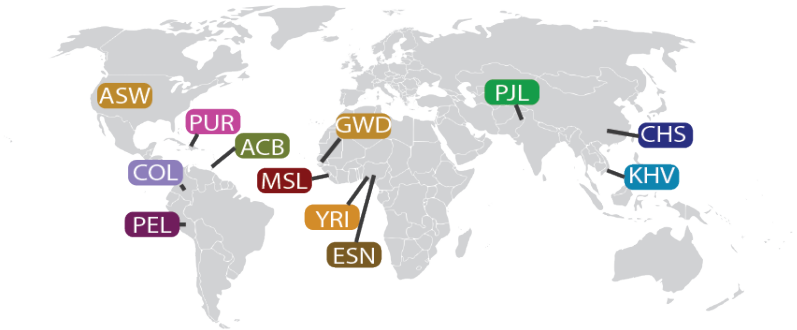
- T2T-CHM13 adds ~200MB of (principally) heterochromatin to reference (6-7%)
- Draft pangenome adds >100MB of common, polymorphic euchromatin (3-4%) (and a **lot** more heterochromatin)
- 0.6-4.4 Mb of additional genic sequences per haplotype compared to GRCh38 (38 gene CNVs/haplotype)



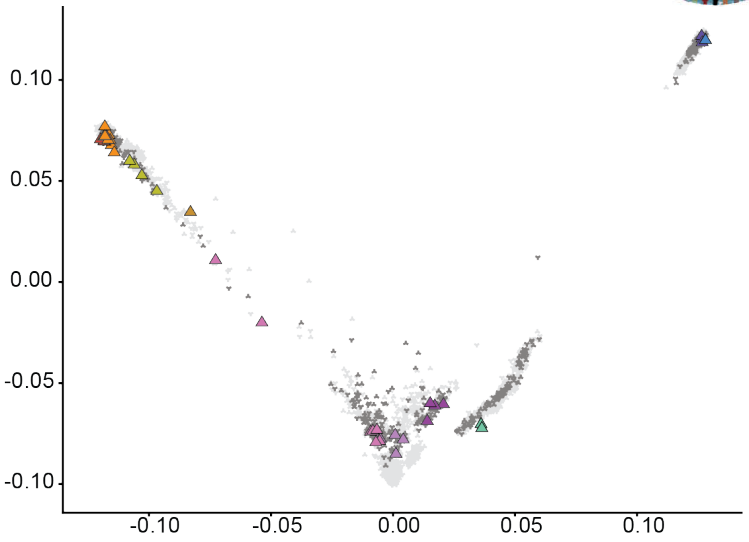
Population Representation and Sampling: Draft Selection

1000 Genomes Consortium Recruitment





Initial Sampling Efforts:



- [x] Cover genetic and geographic diversity
- [x] Availability of low passage cell lines
- [x] Availability of trios/parental data.

De novo assembly quantum leaps

- New sequencing technologies are leading to a dramatic improvement in contiguous assembly
- Haplotype resolution is now essential
- Simultaneously, computational efficiency of *de novo* assembly is being dramatically improved
- T2T will shortly be the standard

› Nat Biotechnol. 2023 Oct;41(10):1474-1482. doi: 10.1038/s41587-023-01662-6. Epub 2023 Feb 16.

Telomere-to-telomere assembly of diploid chromosomes with Verkko

Mikko Rautiainen [1], Sergey Nurk [1 2], Brian P Walenz [1], Glennis A Logsdon [3], David Porubsky [3], Arang Rhie [1], Evan E Eichler [3 4], Adam M Phillippy [5], Sergey Koren [6]

HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads

Sergey Nurk[1], Brian P Walenz[1], Arang Rhie[1], Mitchell R Vollger[2], Glennis A Logsdon[2], Robert Grothe[3], Karen H Miga[4], Evan E Eichler[5], Adam M Phillippy[1] and Sergey Koren[1,6]

Haplotype-resolved *de novo* assembly with phased assembly graphs

Haoyu Cheng[1,2], Gregory T Concepcion[3], Xiaowen Feng[1,2], Haowen Zhang[4], and Heng Li[1,2,*]

[1]Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA, USA
[2]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[3]Pacific Biosciences, Menlo Park, CA, USA
[4]College of Computing, Georgia Institute of Technology, Atlanta, GA, USA
[*]To whom correspondence should be addressed: hli@jimmy.harvard.edu

A fully phased accurate assembly of an individual human genome

David Porubsky, Peter Ebert, Peter A. Audano, Mitchell R. Vollger, William T. Harvey, Katherine M. Munson, Melanie Sorensen, Arvis Sulovari, Marina Haukness, Maryam Ghareghani, Human Genome Structural Variation Consortium, Peter M. Lansdorp, Benedict Paten, Scott E. Devine, Ashley D. Sanders, Charles Lee, Mark J.P. Chaisson, Jan O. Korbel, Evan E. Eichler, Tobias Marschall

doi: https://doi.org/10.1101/855049

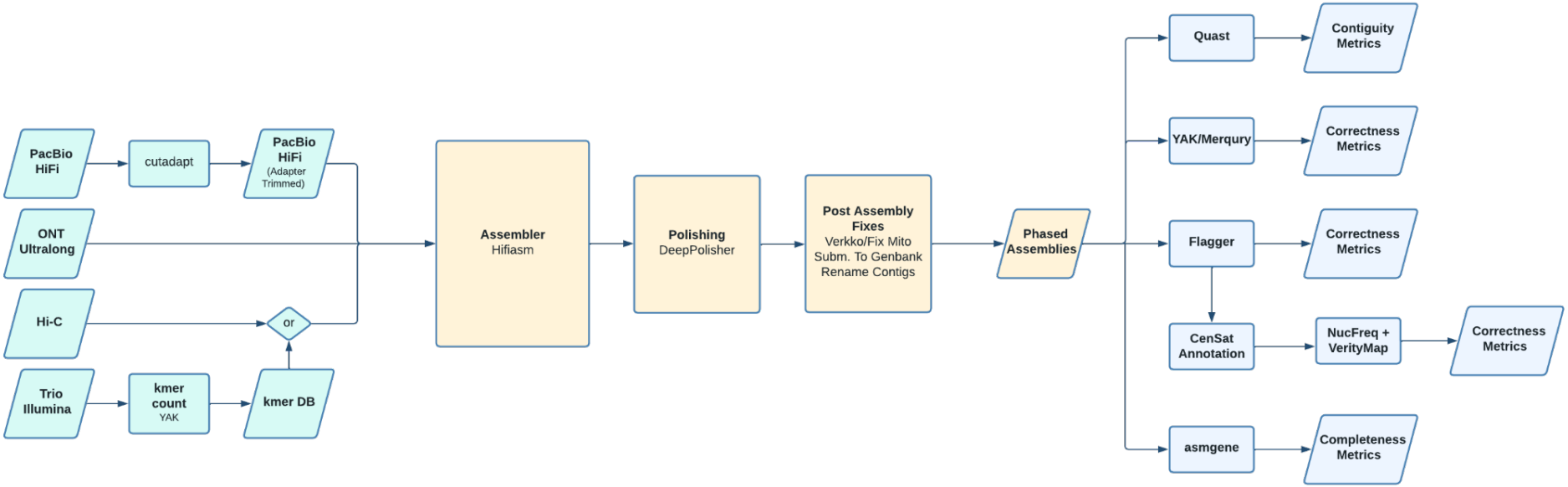
Article | Open Access | Published: 04 May 2020

Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes

Kishwar Shafin, Trevor Pesout, [...] Benedict Paten

Nature Biotechnology **38**, 1044–1053(2020) | Cite this article

R2 Pangenome - Assembly Pipeline & QC



- Both Hifiasm and Verkko now integrate both HiFi/Duplex + ONT UL reads
- Hi-C or Trio Illumina used for long-range phasing
- Lots of QC!

Julian Lucas
& the HPRC Assembly Working Group

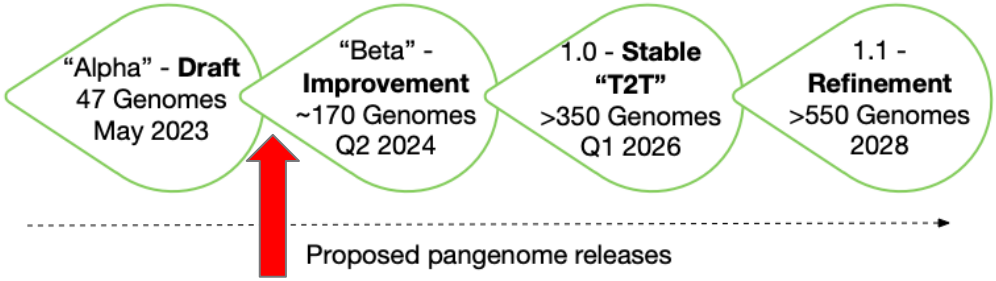
Release 2 Summary: A vital intermediate

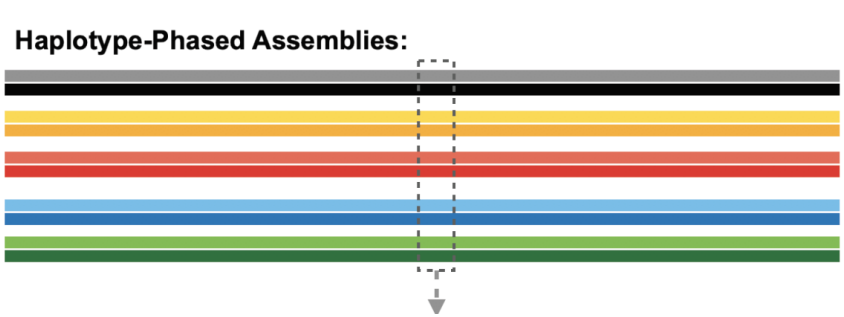
- 5x increase in samples (47 -> 233)
- 3x increase in NG50 (avg. 130 MB)
  - Many (not all) chromosomes T2T
- Approx 3x reduction in structural errors (best guess)
- Focus on alignment/representation of complex regions (e.g. centromeres)
- All samples have long-read transcriptome data(Ensembl and UCSC annotations shortly available)

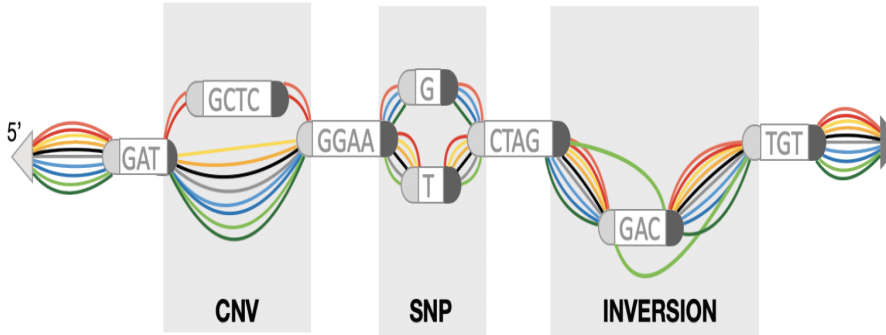
Variant representation on graphs (part 1)

Glenn Hickey

V2.0 HPRC Graphs







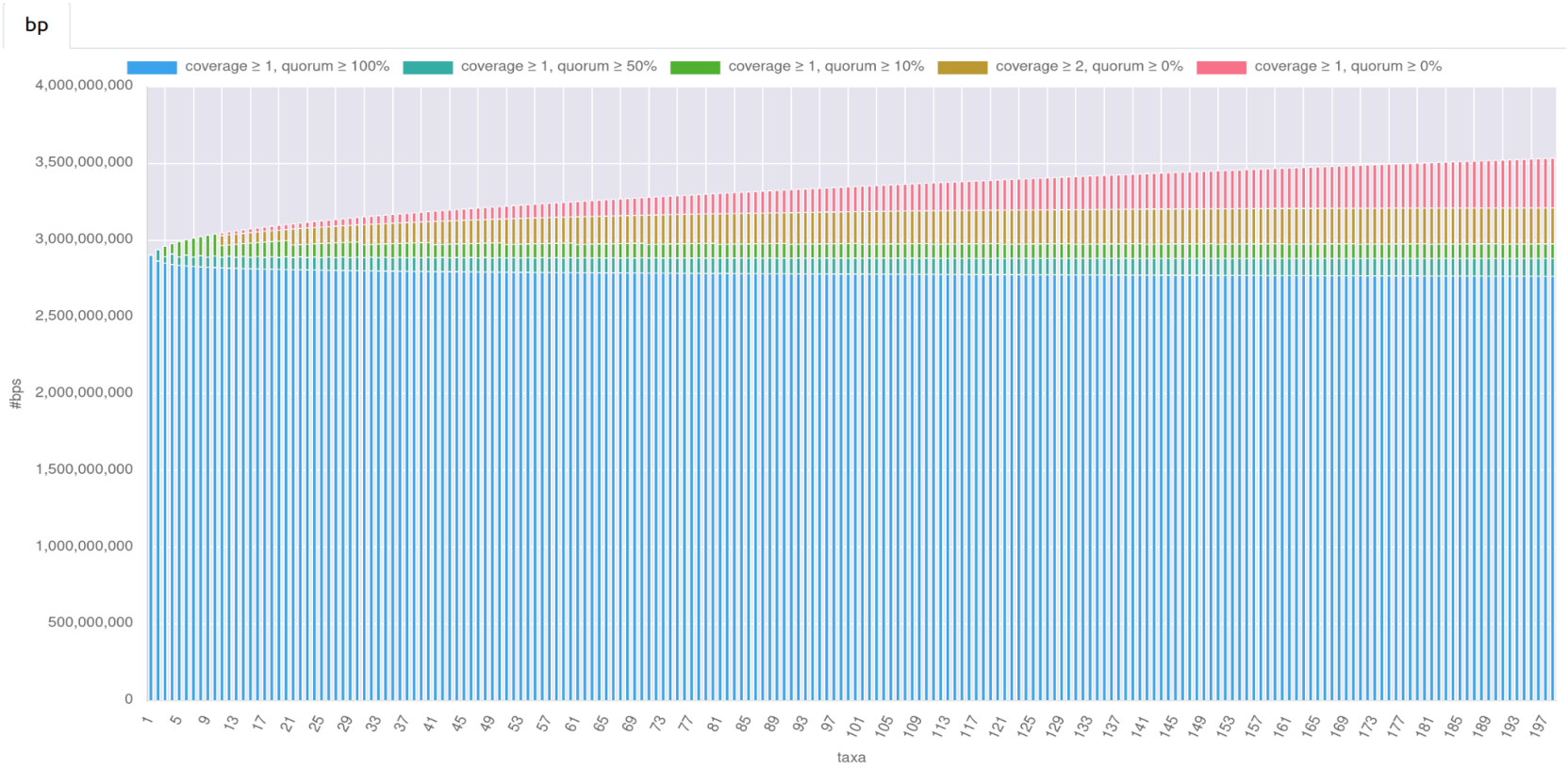
V2.0 HPRC Graphs (INPUT)

	V1.1	**V2.0**
Haplotypes	90	464
CHM13	V1.1 (+ GRCh38#chry)	V2.0
HG002	Held out	Included (held out graph also available)

V2.0 HPRC Graphs (STATS)

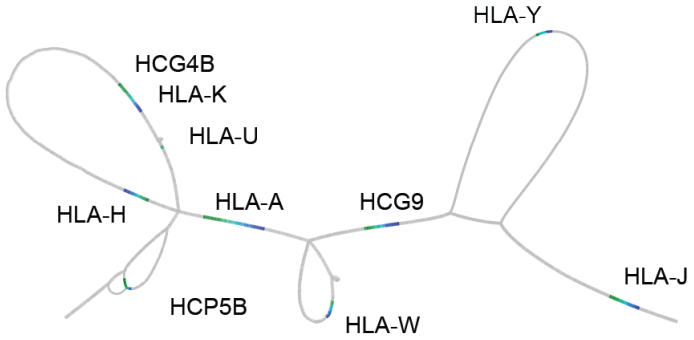
	hprc-v1.1-mc-chm13	hprc-v2.0-mc-chm13
Bases (eurchrom.)	3,338,032,439	3,585,175,934
Nodes	93,165,628	148,283,410
Edges	128,451,813	206,031,684
Total Path Length	256,747,067,932	1,316,880,465,123
File Size (GBZ)	3,632,228,024	5,748,718,968

V2.0 HPRC Graphs (Growth)



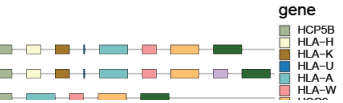
Pangenome Graph Variation

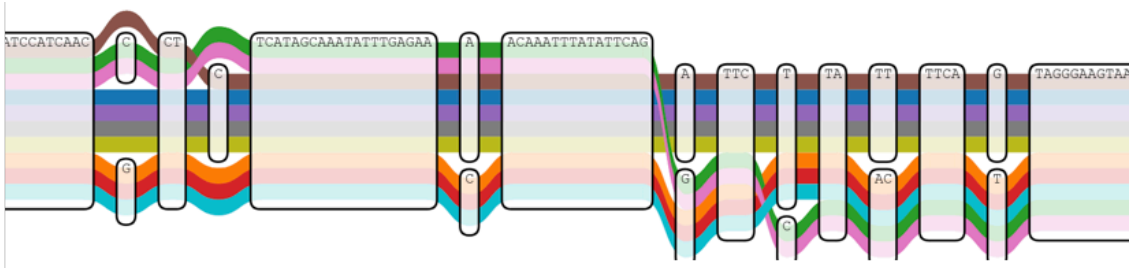
- Pangenome graphs elegantly encode genetic variation using *nodes*, *edges* and *paths*



F

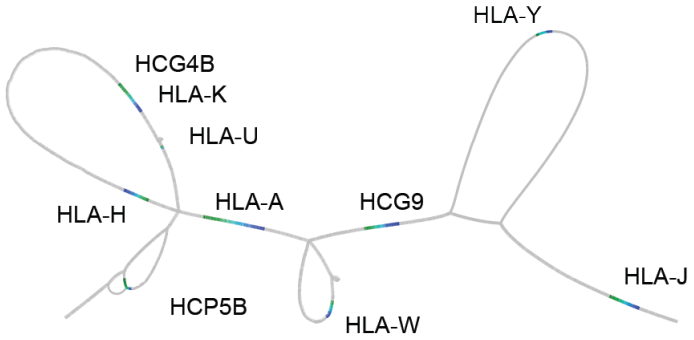
Count	Frequency	Haplotype name
57	0.63	HLA-A
25	0.28	HLA-Y ins
7	0.08	HLA-H/HCG4B/HLA-K/HLA-U del





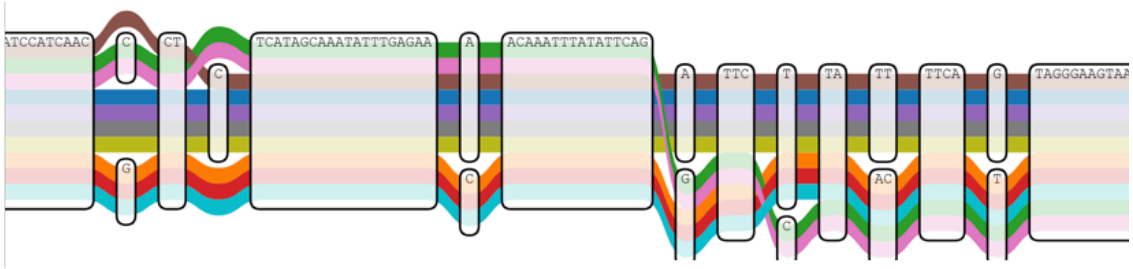
Pangenome Graph Variation

- Pangenome graphs elegantly encode genetic variation using *nodes*, *edges* and *paths*
- **But quantitative analysis of variation requires coordinates and sites**



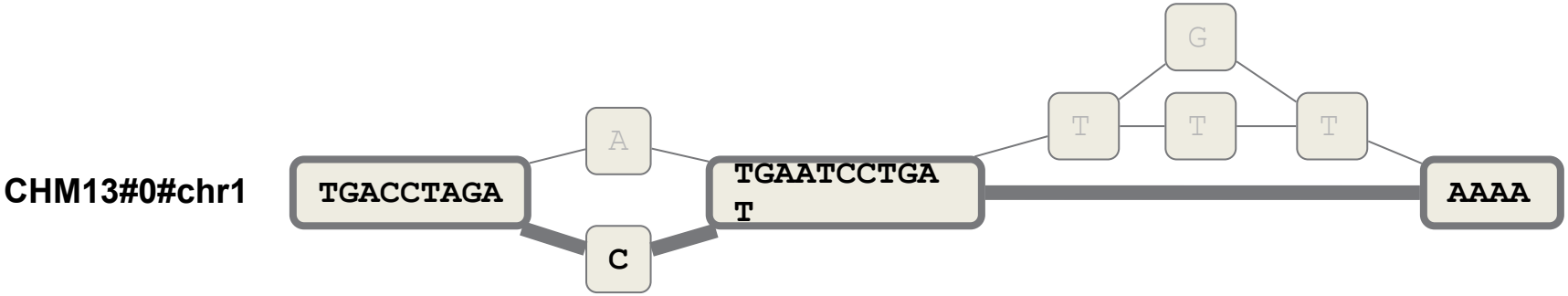
F

Count	Frequency	Haplotype name
57	0.63	HLA-A
25	0.28	HLA-Y ins
7	0.08	HLA-H/HCG4B/HLA-K/HLA-U del



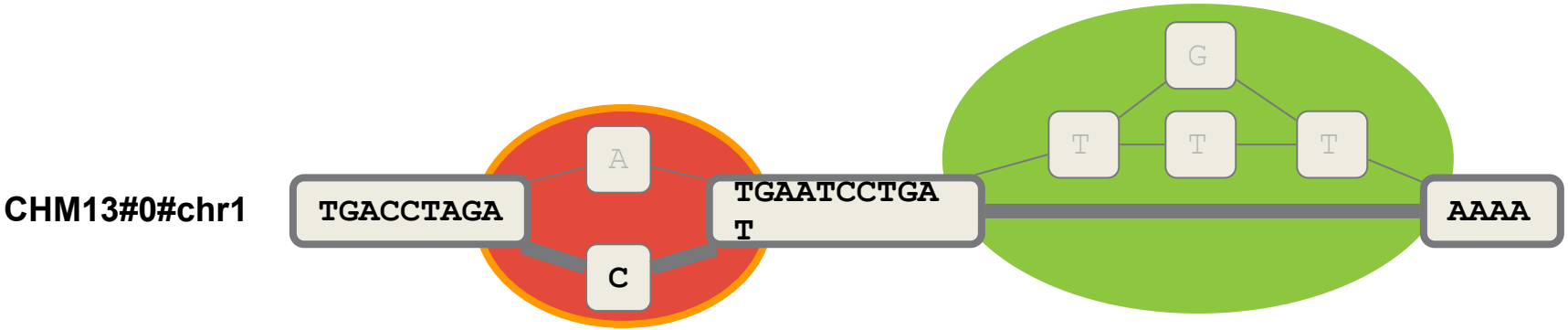
Coordinates and Bubbles

- Standard reference genomes included in HPRC graphs to serve as backbone / coordinate system
  - CHM13, GRCh38
  - Effective as genetic similarity results in largely “linear” graph structures



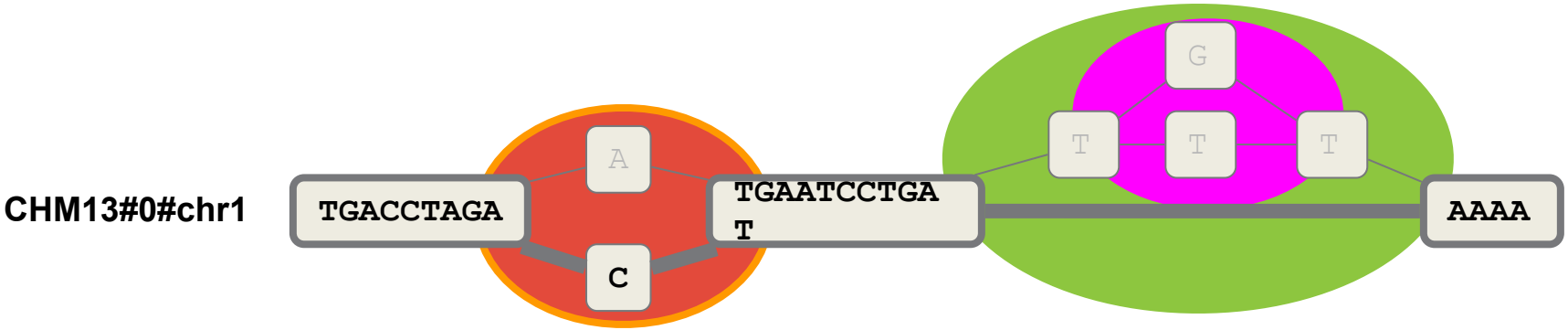
Coordinates and Bubbles

- Standard reference genomes included in HPRC graphs to serve as backbone / coordinate system
  - CHM13, GRCh38
  - Effective as genetic similarity results in largely “linear” graph structures
- Bubbles (aka snarls) are minimal sites of variation in graph



Coordinates and Bubbles

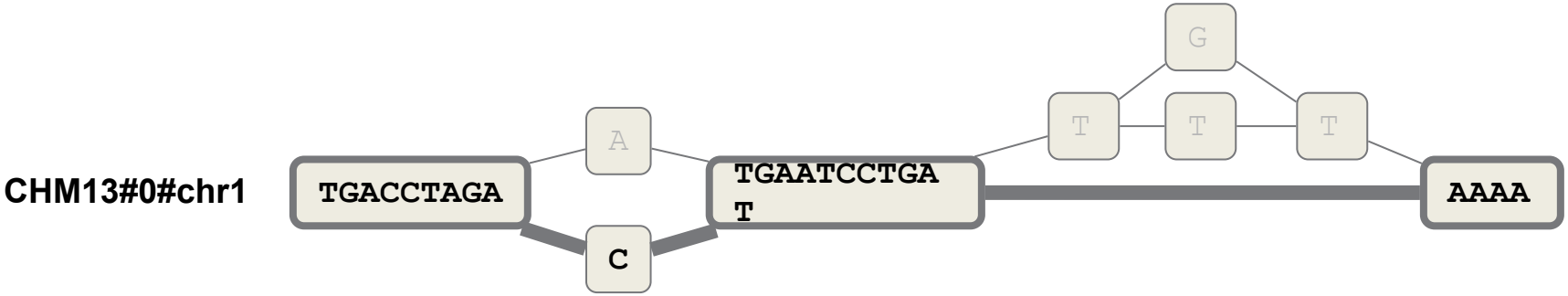
- Standard reference genomes included in HPRC graphs to serve as backbone / coordinate system
  - CHM13, GRCh38
  - Effective as genetic similarity results in largely “linear” graph structures
- Bubbles (aka snarls) are minimal sites of variation in graph
- Bubbles can be nested (but not otherwise overlap)



Graph Variants

- VCF “deconstructed” along chosen reference (vg tools)
- VCF site (line) for every bubble on reference
- ALT-allele for every alternate path spanning the bubble

```
#CHROM    POS    ID    REF    ALT     QUAL    FILTER    INFO                              FORMAT    1
chr1      10     .     C      A       99      .         AC=1;LEN=1;NA=1;NS=1;TYPE=snp     GT        1|0
chr1      21     .     T      TTTT,TTGT    99   .       AC=2;LEN=1;NA=1;NS=1;TYPE=ins     GT        1|1
```



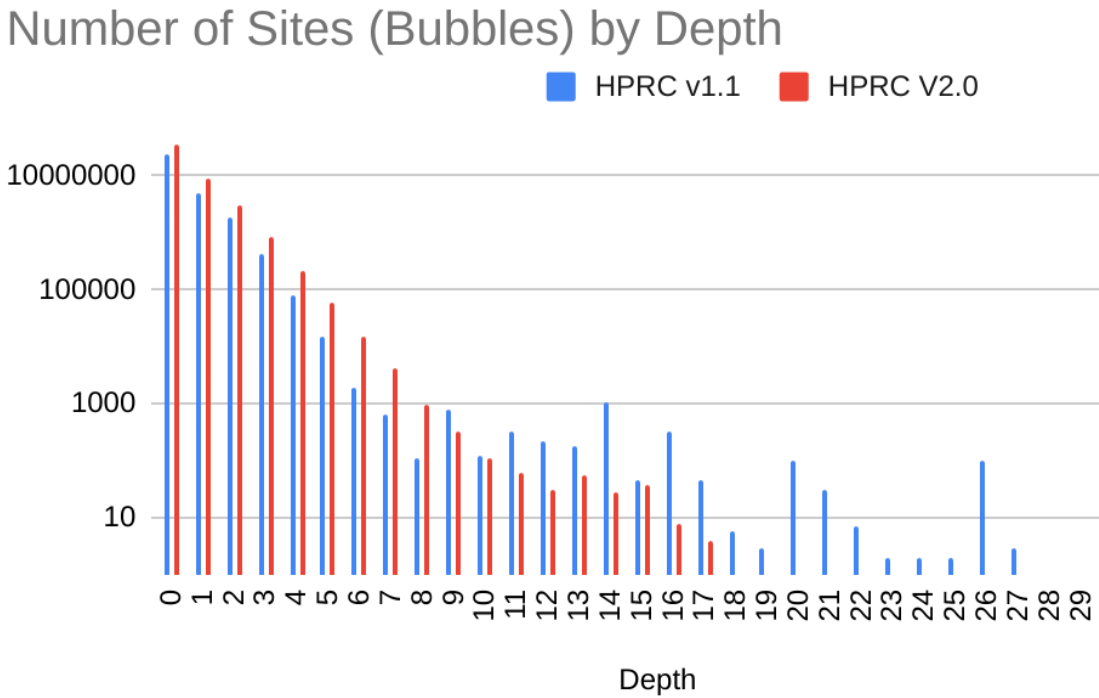
HPRC Variants

V1.1 (90 haplotypes):

- 23.5M sites on CHM13
- 7.4M nested sites

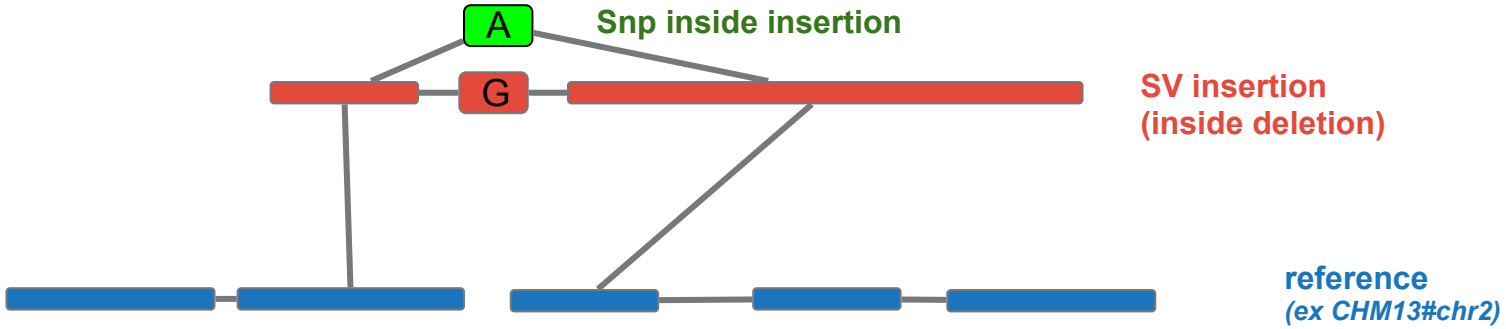
V2.0 (464 haplotypes):

- 35.9M sites on CHM13 (1.5x)
- 12.9M nested sites (1.7x)



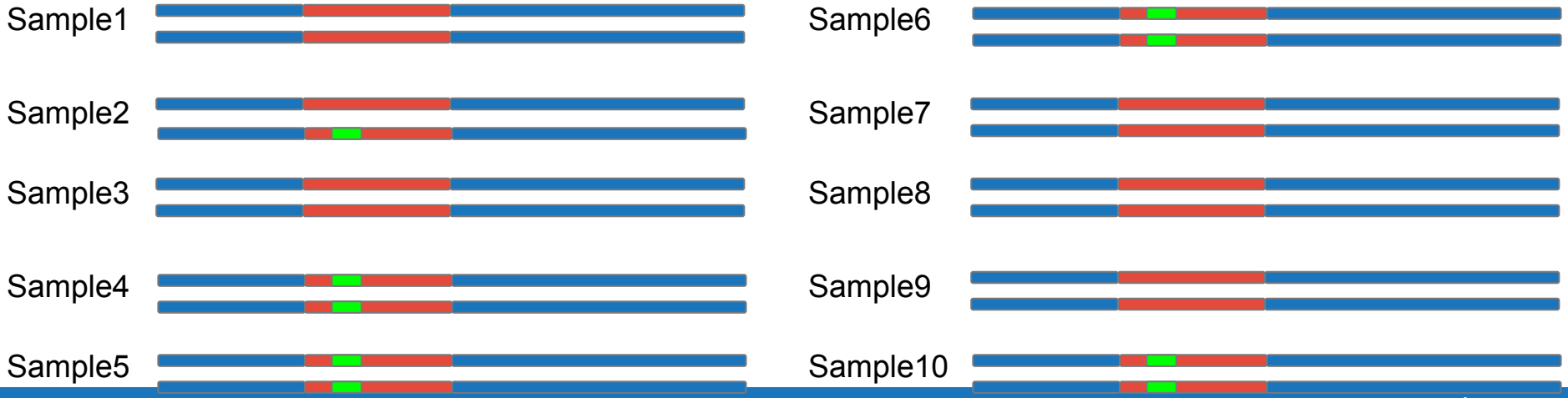
Nested Variation Mischaracterized

- Example currently considered as SV insertion with 2 (very similar) ALT alleles
  - Maybe ALT alleles get merged into one (ie with truvari)
  - Or maybe counted as two separate SV alleles (misleading)

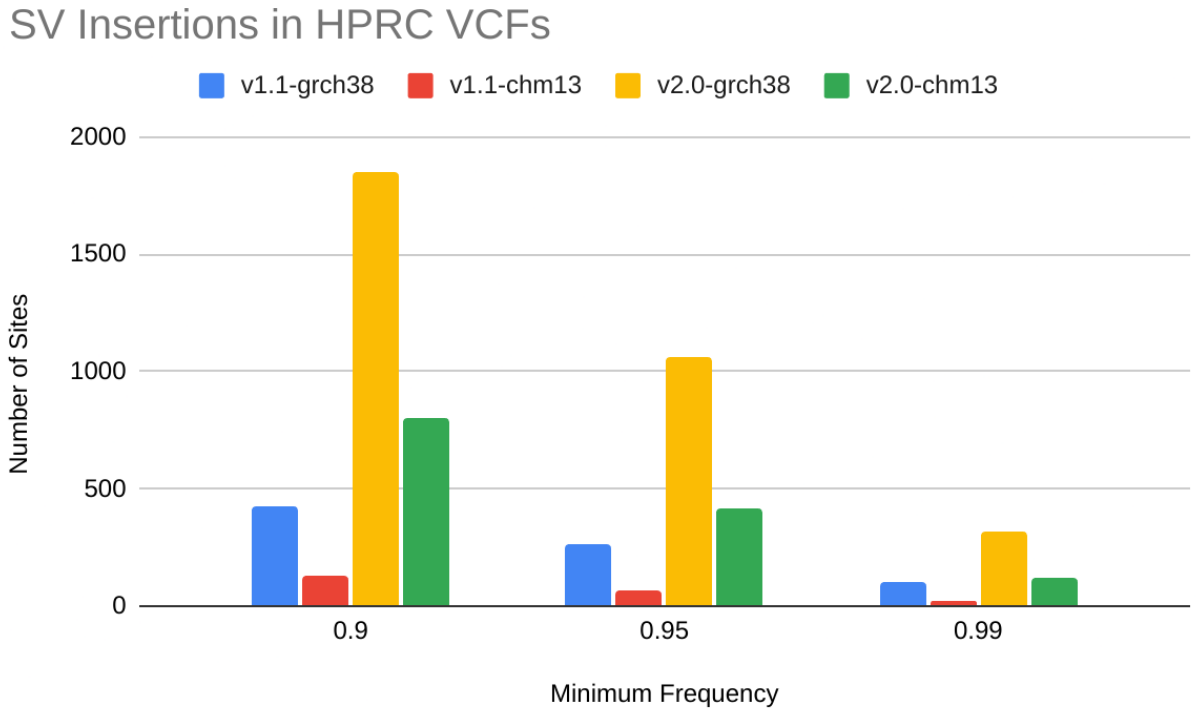


Nested Variation Mischaracterized

- And sometimes the nested variant is all that matters



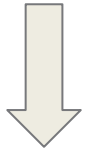
Common HPRC Insertions



Nested VCF

- Off-reference sites included (inspired by rGFA)
- Similar SV alleles merged together
- VCF tags used to record exact nesting relationship
- https://github.com/vgteam/vg/wiki/VCF-export-with-vg-deconstruct
- (Coming soon) Integration with PanGenie, vg call, giraffe-deepvariant

```
CHM13#chr1    21    .    T    TTTT,TTGT    99  .   .    GT    1|2
```



```
CHM13#chr1              21    .    T    TTTT    99  .   .    GT    1|1
HG00438##1#CM089176.1   21    .    T    G       99  .   .    GT    0|1
```

Thanks!

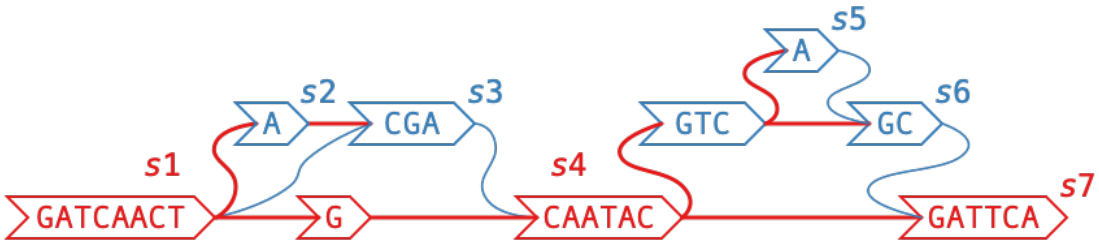
- Benedict Paten Lab (UCSC)
- vg team
- HPRC

Variant representation on graphs (part 2)

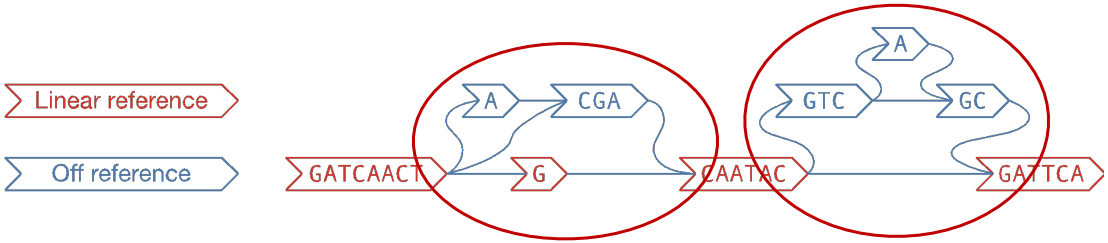
Heng Li

What is a variant, exactly?

- Necessary to report in VCF
- Discussions back to early GA4GH days
- Variants are not defined on graphs
- For years, we have taken a bubble as a variant
- But this is problematic



Current strategy: variant = bubble

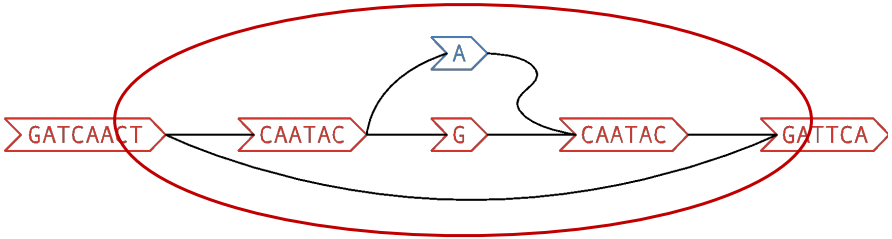



```
ctg   9  G  CGA,ACGA
ctg  15  C  CTCGC,CGTCAGC
```

Problems:

1. Multi-allelic
2. Hard to encode variants in long insertions

Long deletions



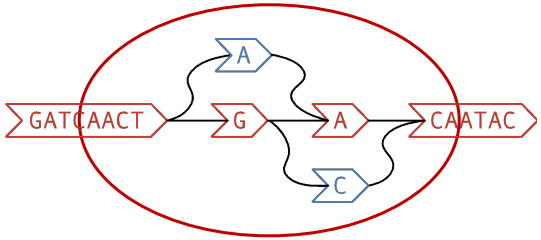

```
ctg  8  TCAATACGCAATAC  T,TCAATACGCAATAC
```

Problems:

1. Multi-allelic
2. Hard to encode variants in long insertions
3. Hide small variants in long deletions (vcfwave as a hack)

vcfwave ignores graph topology (inconsistent) and does pairwise alignment only (underpowered)

Interlocking variants



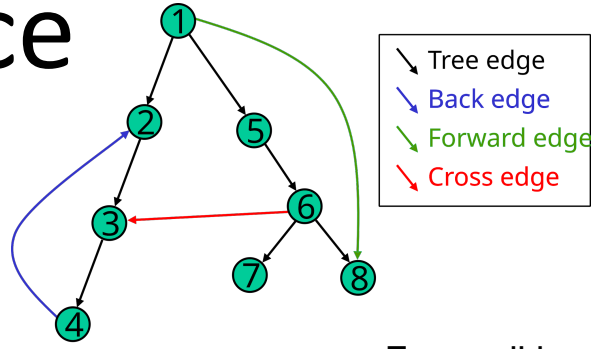

```
ctg  9  GA  AA,GC
```

Reference is the ancestral haplotype; the two SNPs arose independently

Problems:

1. Multi-allelic
2. Hard to encode variants in long insertions
3. Hide small variants in long deletions (vcfwave as a hack)
4. Merge interlocking variants

A new idea: tree reference

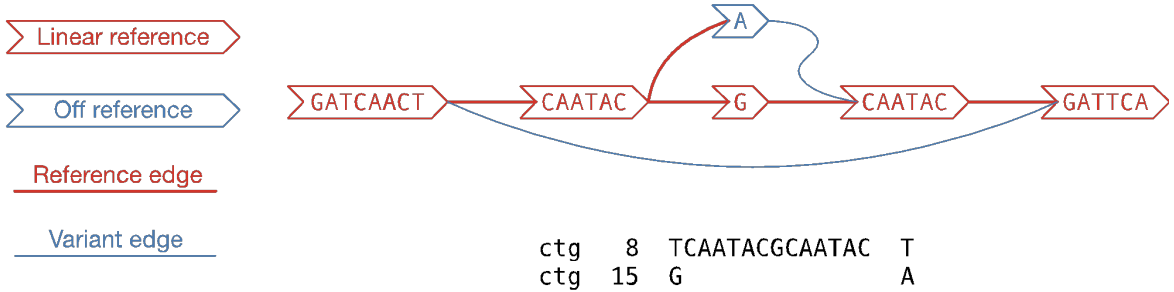


From wiki

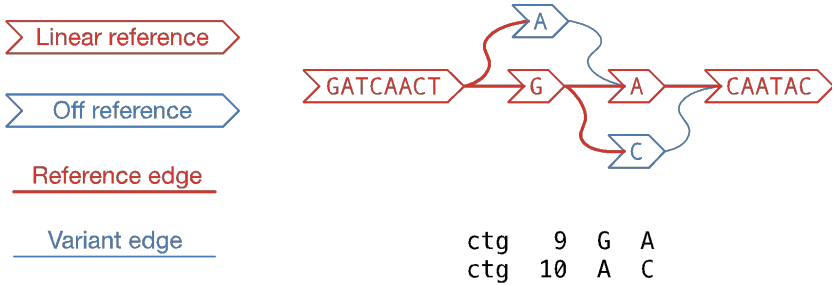
- DFS tree as the **reference**
  - Always include the linear reference
  - The reference can't be "collapsed"
- Edges not in the reference tree represent **variants**
- Based on digraph representation
  - Separate into forward and reverse subgraphs
  - Edges between the two are inversions edges
- Idea from Pouria Salehi Nowbandegani & Luke O'connor
  - Also Haoyang Hu and Shenghan Zhang

Reference tree: a long deletion

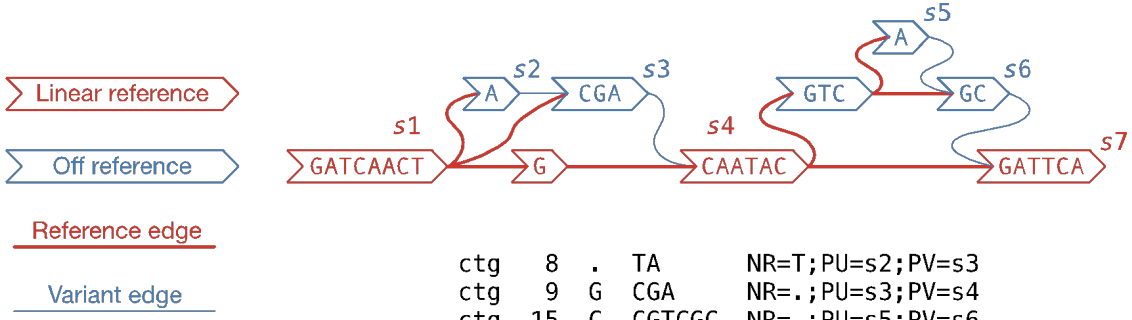
Reference edges form a reference tree; leftover edges are **variants**



Reference tree: interlocking variants

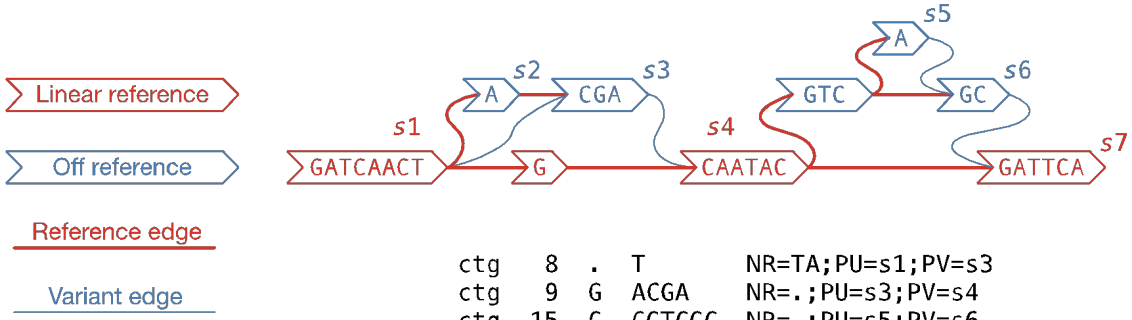


Reference tree: a more complex example



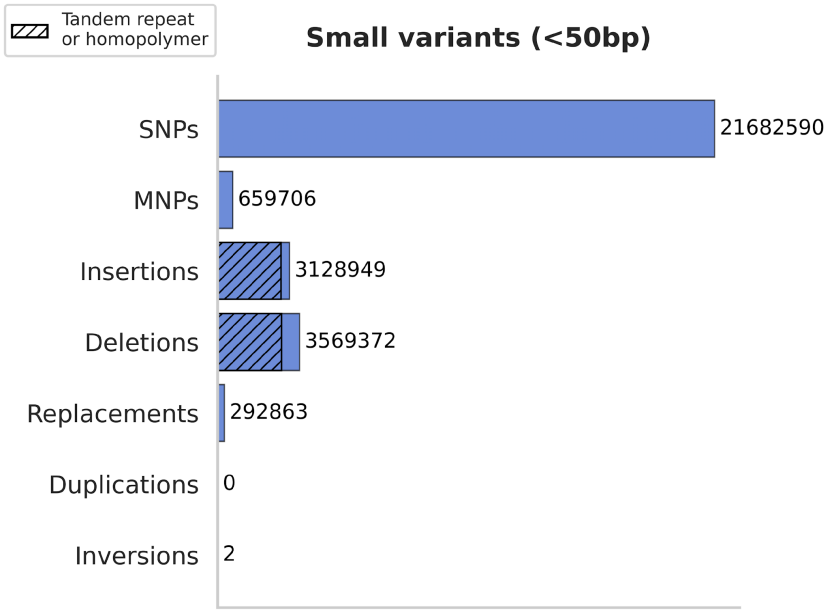

```
ctg    8  .  TA      NR=T;PU=s2;PV=s3
ctg    9  G  CGA     NR=.;PU=s3;PV=s4
ctg   15  C  CGTCGC  NR=.;PU=s5;PV=s6
ctg   15  .  CA      NR=C;PU=s6;PV=s7
```

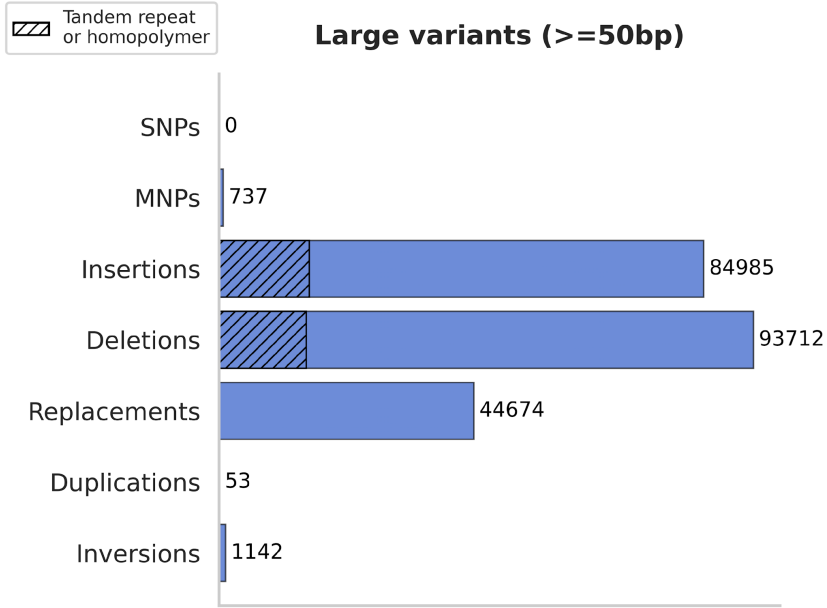
Reference tree is not unique



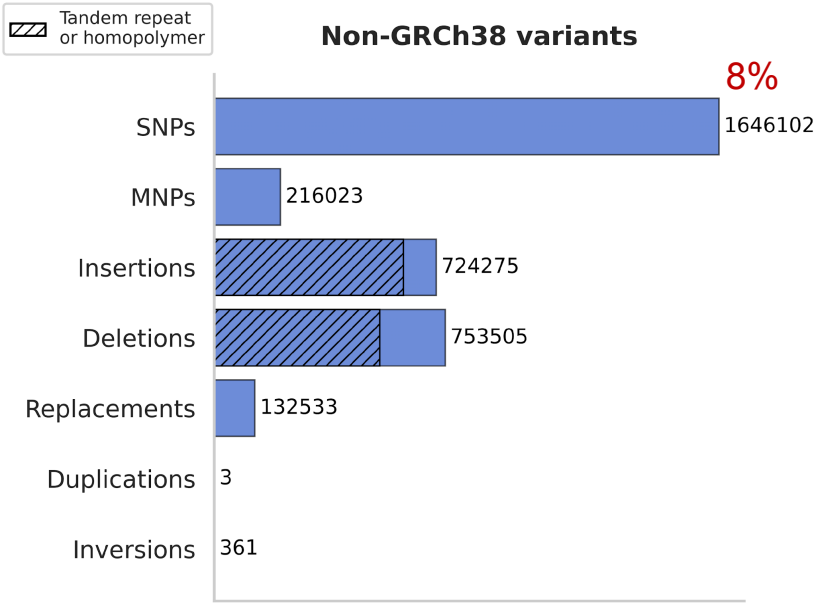

```
ctg    8  .  T       NR=TA;PU=s1;PV=s3
ctg    9  G  ACGA    NR=.;PU=s3;PV=s4
ctg   15  C  CGTCGC  NR=.;PU=s5;PV=s6
ctg   15  .  CA      NR=C;PU=s6;PV=s7
```

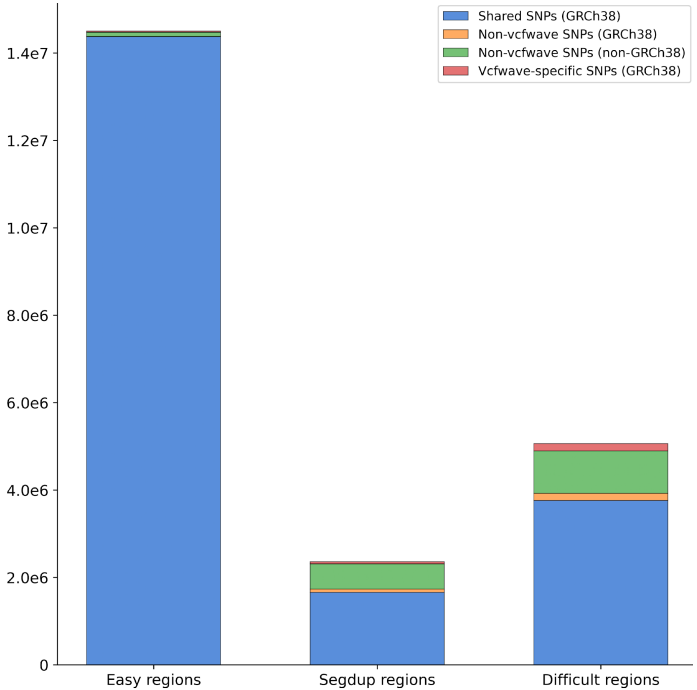
Number of variants





Variants not on GRCh38



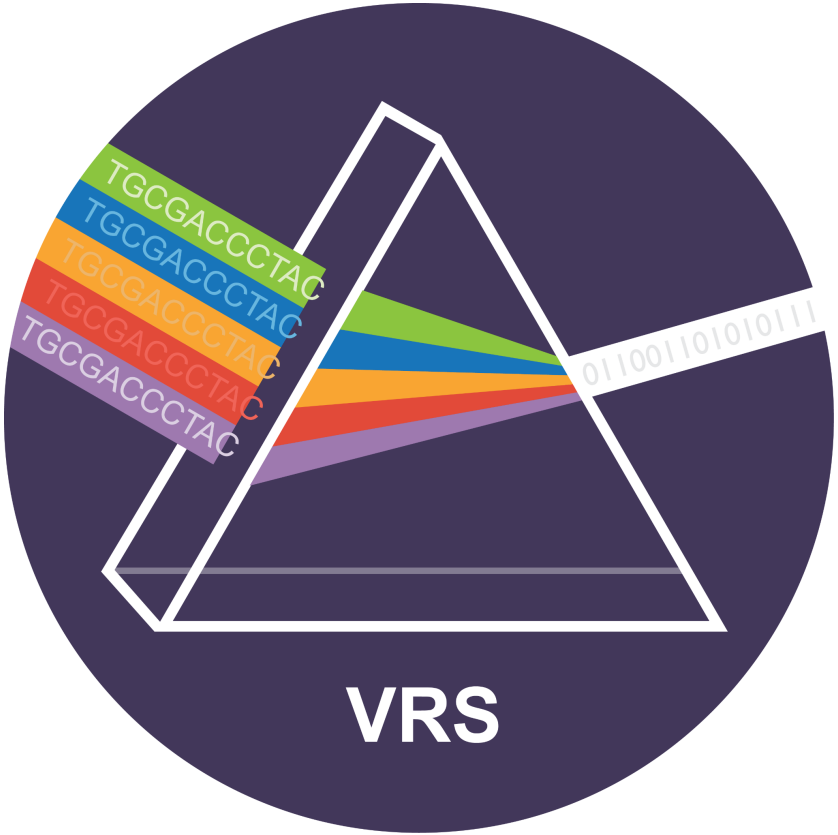


Summary of Part 2

- Reference tree: a new way to represent variants
  - Biallelic variants
  - Variants in long insertions or deletions
  - No interlocking
- Mostly linear-time algorithms
- Solving most problems in bubble representation

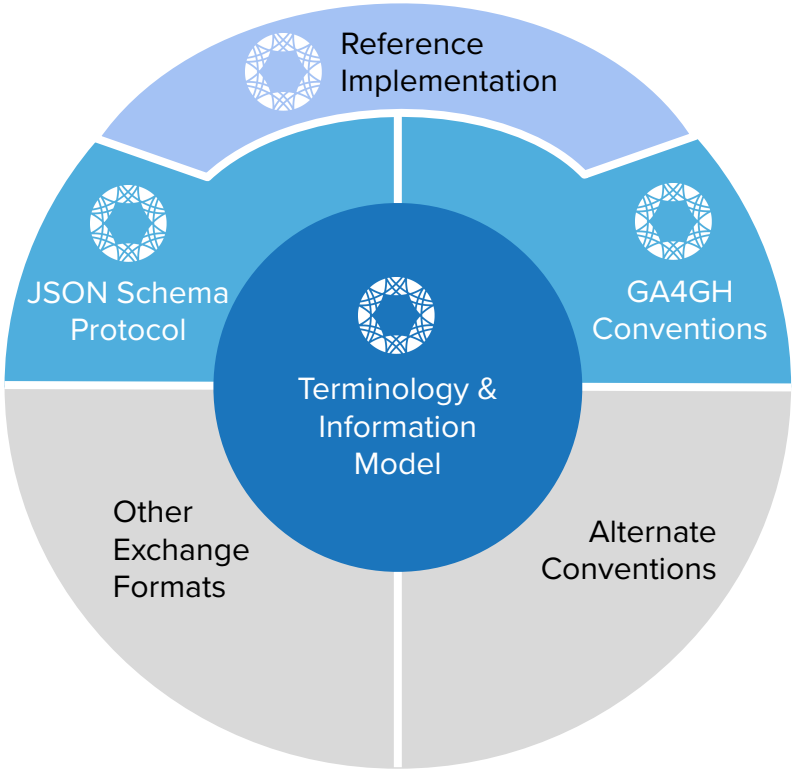
Current GA4GH Standards:

RefGet, Sequence Collections, and VRS


TGCGACCCTAC
TGCGACCCTAC
TGCGACCCTAC
TGCGACCCTAC
TGCGACCCTAC
VRS

The Variation Representation Specification

Global Alliance for Genomics & Health



VRS is...

...INTEROPERABLE ACROSS FORMATS

protobuf Protocol Buffers

XML

...SELF-DESCRIPTIVE

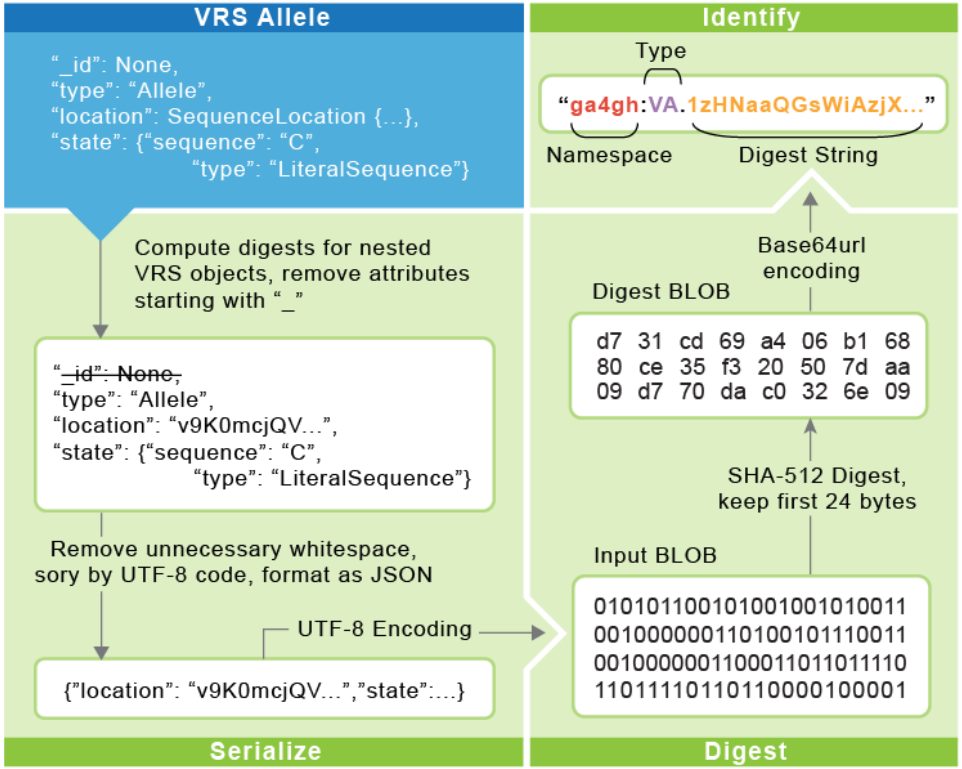
ACTG

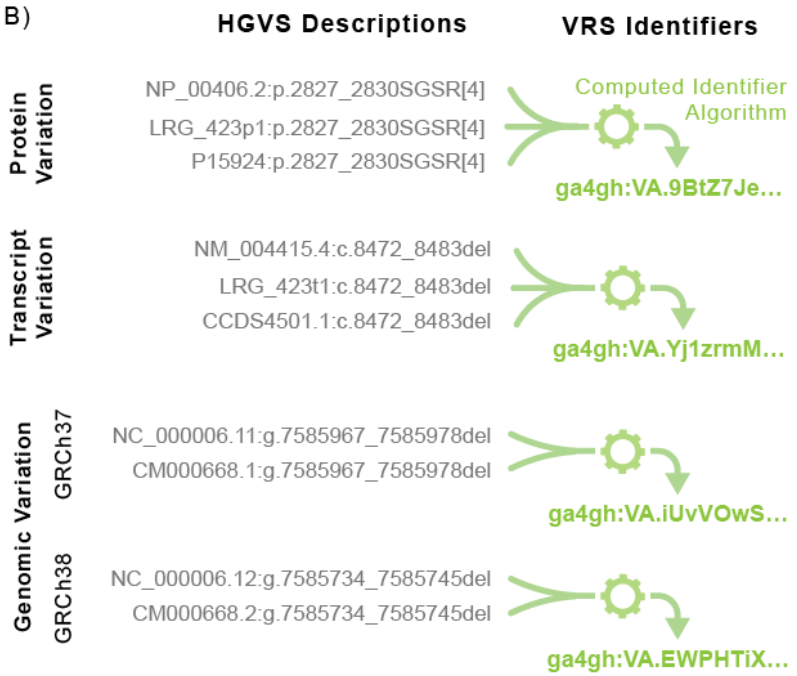
...GLOBALLY CONSISTENT AND DECENTRALIZED

NC_000006.12:g.7585734_7585745del
CM000668.2:g.7585734_7585745del
ga4gh:VA.EWPHTiX...

VRS Variants and Global Identifiers

Global Alliance for Genomics & Health





Wagner AH, et al. *Cell Genomics*. 2021

Design decision: inter-residue coordinates

Global Alliance for Genomics & Health

Insertion between AG in Sequence

Sequence		T		C		A		G		C	A	G		C		A		G		C		T	
Residue		1		2		3		4		5	6		7		8		9		10		11		12	
Inter-residue	0		1		2		3		4		5	6		7		8		9		10		11		12

These residue coordinates are interpreted to exclude associated sequence for an insertion event; inter-residue coordinates are unambiguous

Deletion/Substitution of AG in Sequence

Sequence		T		C		A		G		C	A	G		C		A		G		C		T	
Residue		1		2		3		4		5	6		7		8		9		10		11		12	
Inter-residue	0		1		2		3		4		5	6		7		8		9		10		11		12

The same residue coordinates are interpreted to include associated sequence for a deletion or substitution event; inter-residue coordinates remain unambiguous

Fully-Justified Normalization Captures Region of Shuffling Ambiguity

Normalization Example: In sequence TCAGCAGCT, replace CA at bases 5-6 with CAGCA

Actual location of variation is ambiguous due to the sequence context

(HGVS format: S:g.5_6delinsCAGCA)

$$TCAG\left[\frac{CA}{CAGCA}\right]GCT$$

Holmes, J.B., Moyer, E., Phan, L., Maglott, D. & Kattman, B. L.
SPDI: Data Model for Variants and Applications at NCBI.
bioRxiv 537449 (2019). doi:10.1101/537449

Global Alliance for Genomics & Health

MTOR c.7280T>C in VRS (minimal)

SequenceReference **describes the sequence**


```
{
    "id": "ga4gh:VA.j0DhKAqyn6YhEeWn45Id3dD5C5_4SHVW",
    "type": "Allele",
    "location": {
        "id": "ga4gh:SL.YNXEsQ3w00ociDyTclbgZdBYqf5ydFY_",
        "type": "SequenceLocation",
        "sequenceReference": {
            "type": "SequenceReference",
            "refgetAccession": "SQ.QheGYEnKbwNpM3LulbPTBQhyBSyZwuYm"
        },
        "start": 7400,
        "end": 7401
    },
    "state": {
        "type": "LiteralSequenceExpression",
        "sequence": "C"
    }
}
```


Retrieved from:
https://normalize.cancervariants.org/variation/normalize?q=MTOR c.7280T>C

Global Alliance for Genomics & Health

SequenceLocation describes a location on a sequence

SequenceReference describes the sequence

MTOR c.7280T>C in VRS (minimal)


```
{
    "id": "ga4gh:VA.j0DhKAqyn6YhEeWn45Id3dD5C5_4SHVW",
    "type": "Allele",
    "location": {
        "id": "ga4gh:SL.YNXEsQ3w00ociDyTclbgZdBYqf5ydFY_",
        "type": "SequenceLocation",
        "sequenceReference": {
            "type": "SequenceReference",
            "refgetAccession": "SQ.QheGYEnKbwNpM3LulbPTBQhyBSyZwuYm"
        },
        "start": 7400,
        "end": 7401
    },
    "state": {
        "type": "LiteralSequenceExpression",
        "sequence": "C"
    }
}
```


Retrieved from:
https://normalize.cancervariants.org/variation/normalize?q=MTOR c.7280T>C

Global Alliance for Genomics & Health

MTOR c.7280T>C in VRS (minimal)

***Allele* describes the sequence state at a location**

SequenceLocation describes a location on a sequence

SequenceReference describes the sequence


```
{
    "id": "ga4gh:VA.j0DhKAqyn6YhEeWn45Id3dD5C5_4SHVW",
    "type": "Allele",
    "location": {
        "id": "ga4gh:SL.YNXEsQ3w00ociDyTclbgZdBYqf5ydFY_",
        "type": "SequenceLocation",
        "sequenceReference": {
            "type": "SequenceReference",
            "refgetAccession": "SQ.QheGYEnKbwNpM3LulbPTBQhyBSyZwuYm"
        },
        "start": 7400,
        "end": 7401
    },
    "state": {
        "type": "LiteralSequenceExpression",
        "sequence": "C"
    }
}
```


Retrieved from:
https://normalize.cancervariants.org/variation/normalize?q=MTOR c.7280T>C

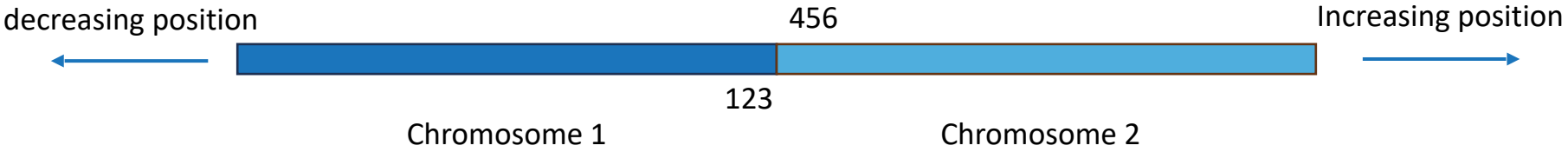
Global Alliance for Genomics & Health

MTOR c.7280T>C in VRS 2.0, annotated

Objects can include descriptive labels, text, etc.


```
{
    "id": "ga4gh:VA.j0DhKAqyn6YhEeWn45Id3dD5C5_4SHVW",
    "type": "Allele",
    "label": "MTOR c.7280T>C",
    "expressions": [
        {
            syntax: "hgvs.c",
            version: "21.0",
            value: "NM_004958.4:c.7280T>C"
        }
    ]
    "location": {
        "id": "ga4gh:SL.YNXEsQ3w00ociDyTclbgZdBYqf5ydFY_",
        "label": "NM_004958.4 sequence residue 7401"
        "type": "SequenceLocation",
        "sequenceReference": {
            "id": "refseq:NM_004958.4"
            "type": "SequenceReference",
            "refgetAccession": "SQ.QheGYEnKbwNpM3LulbPTBQhyBSyZwuYm",
            "label": "MTOR transcript NM_004958.4 (MANE Select)"
            "alphabet": "na"
        },
        "start": 7400,
        "end": 7401
        "sequence": "T",
    },
    "state": {
        "type": "LiteralSequenceExpression",
        "sequence": "C"
    }
}
```

Simple breakpoint example




```
id: simple_example
type: Adjacency
adjoinedSequences: [
  - type: SequenceLocation
    sequenceReference:
      refgetAccession: 7aI9u0vheriv13er,
      residueAlphabet: na,
      id: NC_000001.10
    end: 123
  - type: SequenceLocation
    sequenceReference:
      refgetAccession: hiu2gii33iv13er,
      residueAlphabet: na,
      id: NC_000002.11
    start: 456
```

1. Everything is a path through a graph – all “variants” embedded

SequenceReference only

```
{
  "id": "hpp:HPP_0000000073156.1"
  "type": "SequenceReference",
  "refgetAccession": "SQ.QheGYEnKbwNpM3LulbPTBQhyBSyZwuYm",
  "label": "HPP Haplotype 0000000073156, version 1",
  "alphabet": "na"
}
```

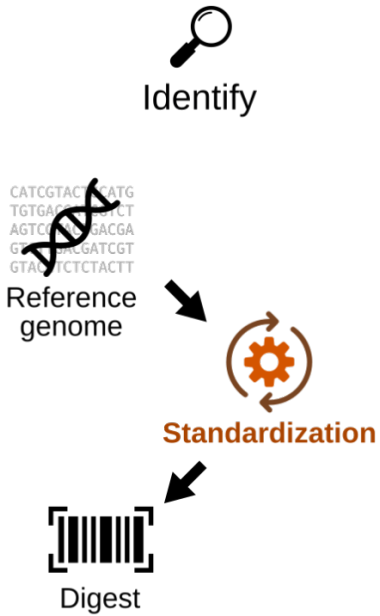
2. Variants are described with respect to a graph of major haplotypes


```
{
    "id": "ga4gh:VA.j0DhKAqyn6YhEeWn45Id3dD5C5_4SHVW",
    "type": "Allele",
    "location": {
        "id": "ga4gh:SL.YNXEsQ3w00ociDyTclbgZdBYqf5ydFY_",
        "type": "SequenceLocation",
        "sequenceReference": {
            "id": "hpp:HPP_0000000073156.1"
            "type": "SequenceReference",
            "refgetAccession": "SQ.QheGYEnKbwNpM3LulbPTBQhyBSyZwuYm",
            "label": "HPP Haplotype 0000000073156, version 1"
            "alphabet": "na"
        },
        "start": 7400,
        "end": 7401
        "sequence": "T",
    },
    "state": {
        "type": "LiteralSequenceExpression",
        "sequence": "C"
    }
}
```

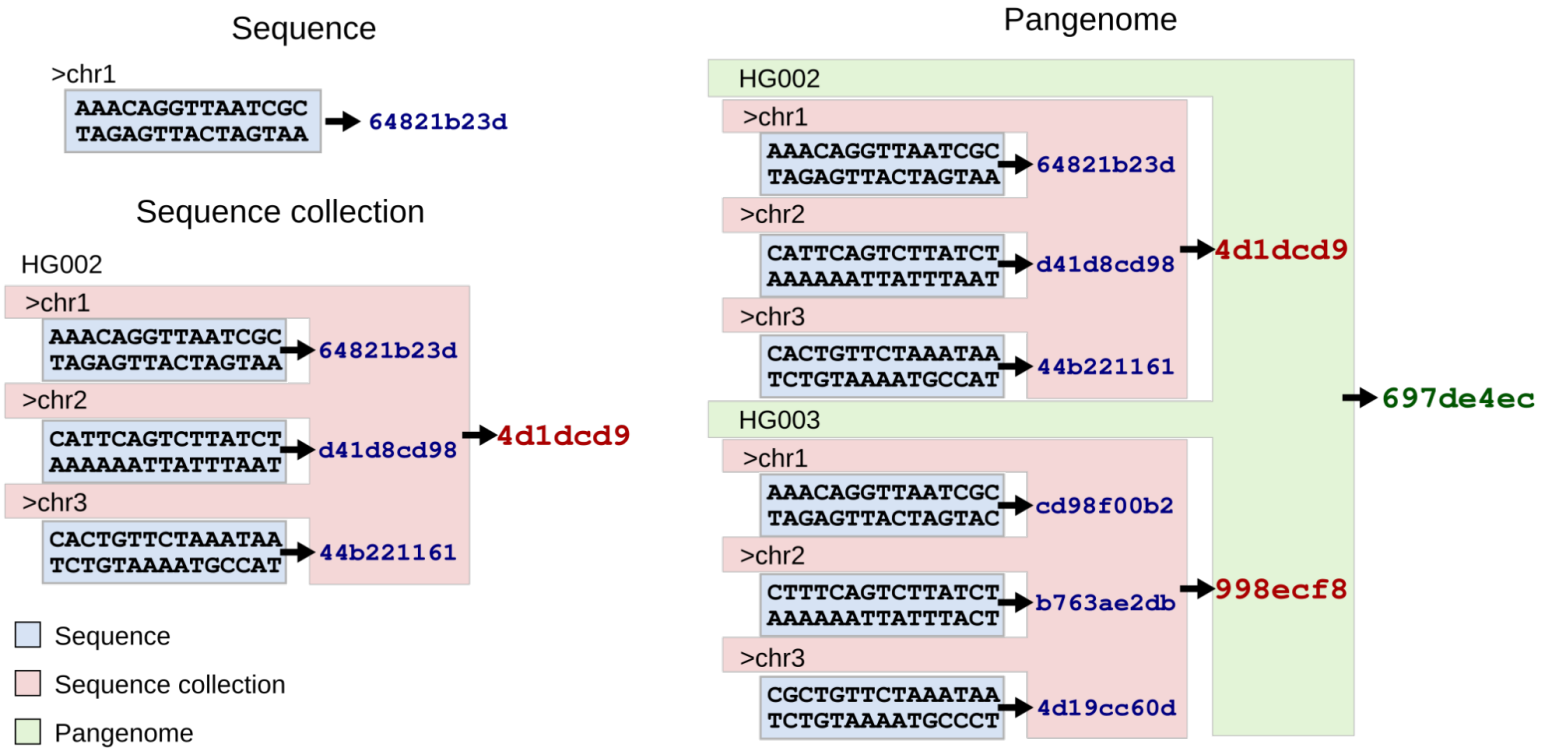
3. Variants are described as a coordinate on a graph

```
{
    "id": "ga4gh:VA.j0DhKAqyn6YhEeWn45Id3dD5C5_4SHVW",
    "type": "Allele",
    "location": {
        "id": "ga4gh:GL.YNXEsQ3w00ociDyTclbgZdBYqf5ydFY_",
        "type": "GraphLocation",
        "graphReference": {
            "id": "hpp:HPP_00013.1"
            "type": "GraphReference",
            "refgetAccession": "SG.QheGYEnKbwNpM3LulbPTBQhyBSyZwuYm",
            "label": "HPP Graph 13, version 1"
            "alphabet": "na"
        },
        "path": ...,
        "end": 7401
        "start": 7400
        "sequence": "T",
    },
    "state": {
        "type": "LiteralSequenceExpression",
        "sequence": "C"
    }
}
```

Sequence Collections: core functions



From Sequence Collections to Pangenomes



What tooling is required?

Are there ways to integrate VRS and Pangenome SV calling methods?

Who is interested in working in this collaborative space to meet these goals?

GA4GH CONNECT

Share your feedback
on the Connect meeting!

GA4GH CONNECT

Lunch time

1:00 PM to 2:00 PM

Second Floor Connector

Head right outside for lunch!