Genotyping structural variants in human cohorts using pangenome graphs

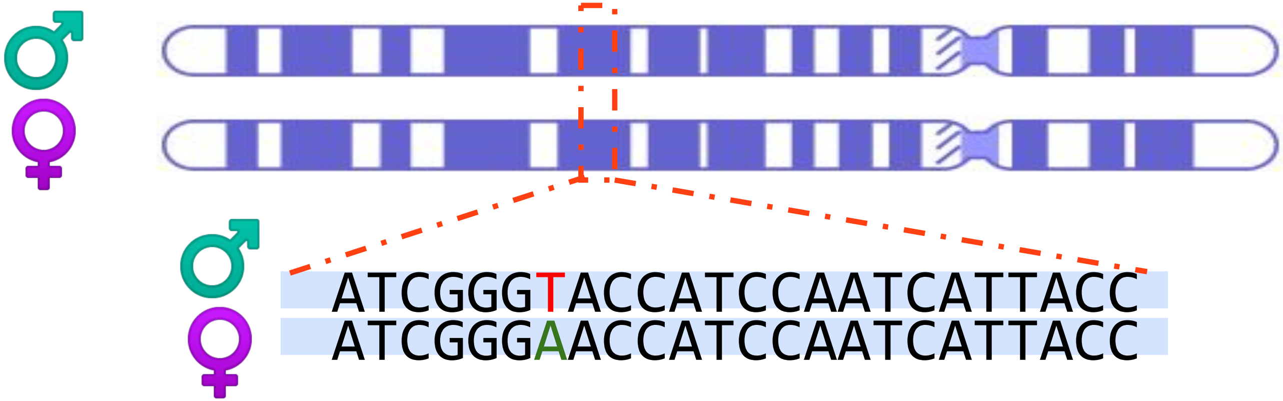
Jean Monlong
Post-doctoral Scholar
Computational Genomics Lab

UNIVERSITY OF CALIFORNIA SANTA CRUZ | Genomics Institute

Overview

- Background: sequencing, genotypes, structural variations
- Pangenome analysis with the vg toolkit
- Genotyping structural variants across thousands of genomes
- Next: pangenomes from de novo assemblies

Humans are diploid: 2 copies of the genome



♂ ATCGGGTACCATCCAATCATTACC

♀ ATCGGGAACCATCCAATCATTACC

Our genome is comprised of a paternal and a maternal "haplotype". Together, they form our "**genotype**"

From https://github.com/quinlan-lab/applied-computational-genomics

Types of genetic variation

ctc**c**gag
ctc**t**gag

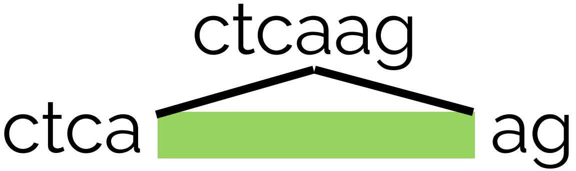
Single-nucleotide polymorphisms (**SNPs**)

“DNA spelling mistakes”

ctc**--**ag
ctc**tg**ag

Insertion-deletion polymorphisms (**INDELs**)

“extra or missing DNA”



ctcaag

ctca ag

Structural variants (**SVs**)

“Large blocks of extra, missing or rearranged DNA (>50bp)”

From https://github.com/quinlan-lab/applied-computational-genomics

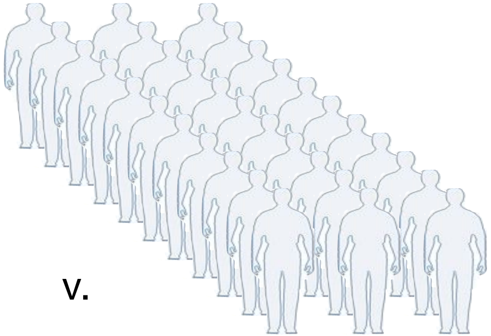
Why do we care?

Understanding the relationship between genetic variation and traits or disease phenotypes

Complex diseases (multiple genes contribute to risk)

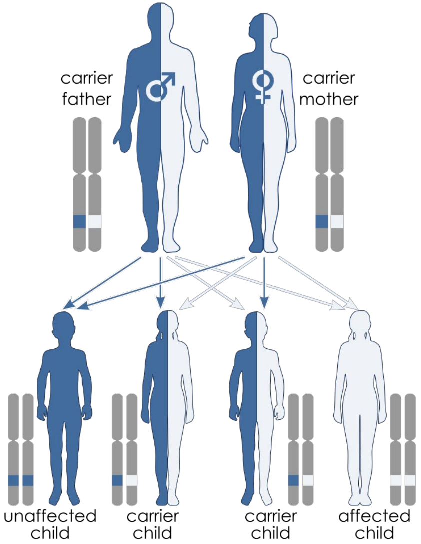
Cases
(have disease)

v.



Controls
(no disease)

Rare diseases



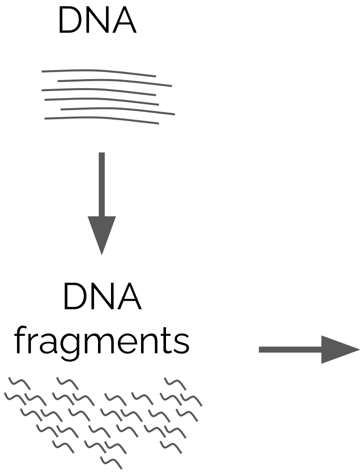
■ *Unaffected*

□ *Affected*

◧ *Carrier*

From https://github.com/quinlan-lab/applied-computational-genomics

Genome sequencing



Illumina

PacBio

Oxford Nanopore

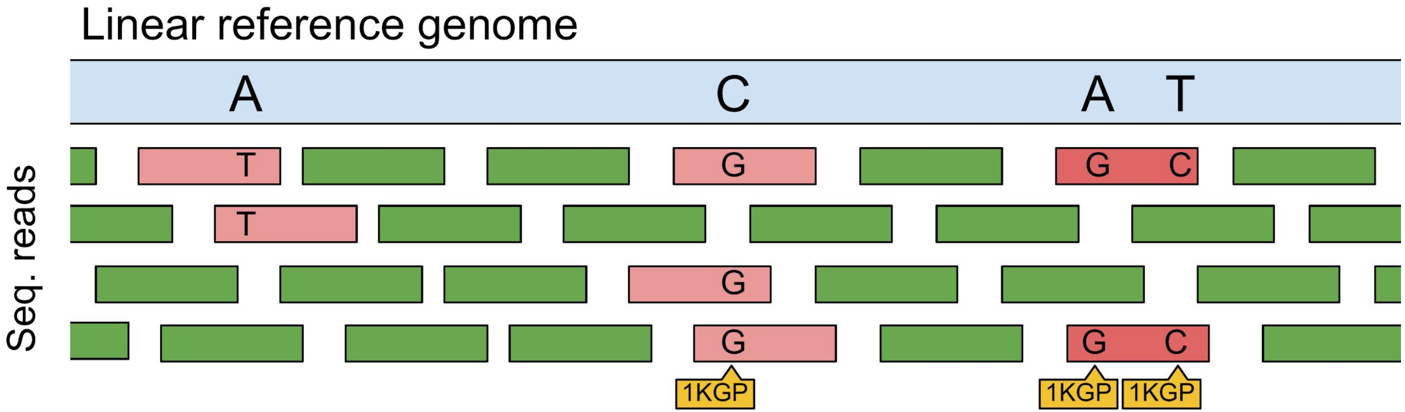
Short reads: 150-250bp

FASTQ file

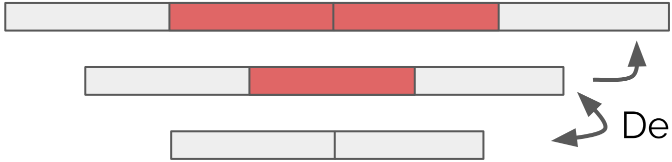
```
@ERR903030.219 HWI-D00574:82:C6L01ANXX:3:1101:3953:1913/1
AGCTCTTATTATTTTGAAATATGTCCCATCAATACCTAATTTATTGAGAGTTTTTAGCATGAAGGGTTG
+
<<A0?FGGGGGGGGGGGEGGGGGGGGGGGGGG:CFCGDFGG>FG1F@0:FGGGGGF@>FG1=F@FFG
@ERR903030.220 HWI-D00574:82:C6L01ANXX:3:1101:3863:1914/1
ACCATGAAAGACAGGTGTTAGAATCAGTACAAGAAGCAACAGGGAGCCATTGCATTTTGAGCATTTG
+
<ABBGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
@ERR903030.221 HWI-D00574:82:C6L01ANXX:3:1101:3906:1914/1
GATGGGGTTTCACATTGGCCAGGCTGGTCTCAAACTCCTAACCTCAAGGGATCCACCCACCTCGGCCT
+
<<AAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGFGGGGGGGGGGGGGGG
@ERR903030.222 HWI-D00574:82:C6L01ANXX:3:1101:3833:1922/1
TACTGAAGAATCAGTGTCAGTTTTGTTAGTTGTGATAATGACATTCTGCTAAGCTAAAGTATAGAGGG
+
=<BBGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGEGGGFG
@ERR903030.223 HWI-D00574:82:C6L01ANXX:3:1101:3942:1927/1
AAAAAATGAACTAAAAATGCATTAAAGACCAAATGTAATACCTAAAAATGTAAAACTTTTAGAAGGA
+
=ABBGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
@ERR903030.224 HWI-D00574:82:C6L01ANXX:3:1101:3920:1929/1
ATTCCATTCGATTCCATTCGATGATTCCATTGGATTCCATTCGATGATGATTCCATTCAAATCCATTCGA
+
=ABBGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGC>FFGGGEF@EDFGCEGFGGGGGGGG
...
...
```

Long reads: 10,000s-100,000s bp

Traditional read mapping & variant calling

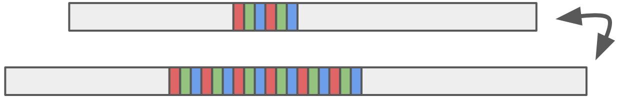


Structural Variants (SVs)

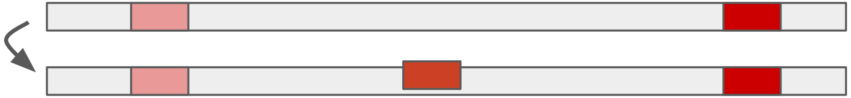


Tandem duplication

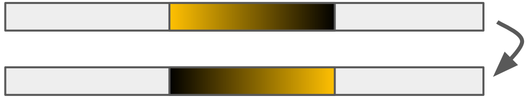
Deletion/Insertion



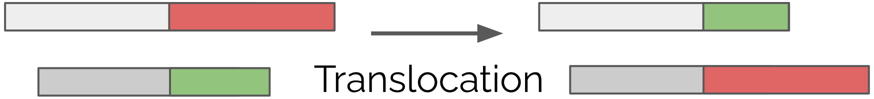
Short tandem repeat contraction/expansion



Mobile element insertion

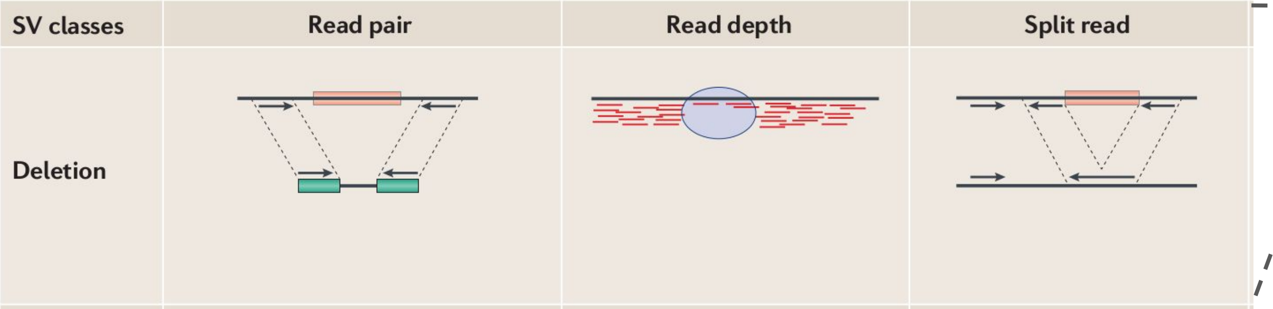


Inversion

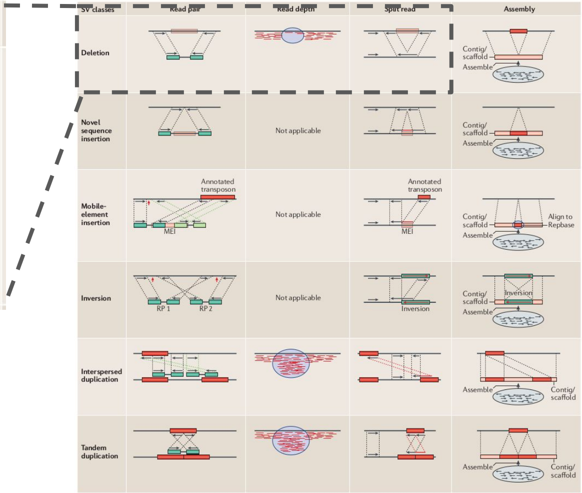


Translocation

Traditional structural variant calling



SV classes	Read pair	Read depth	Split read
Deletion			

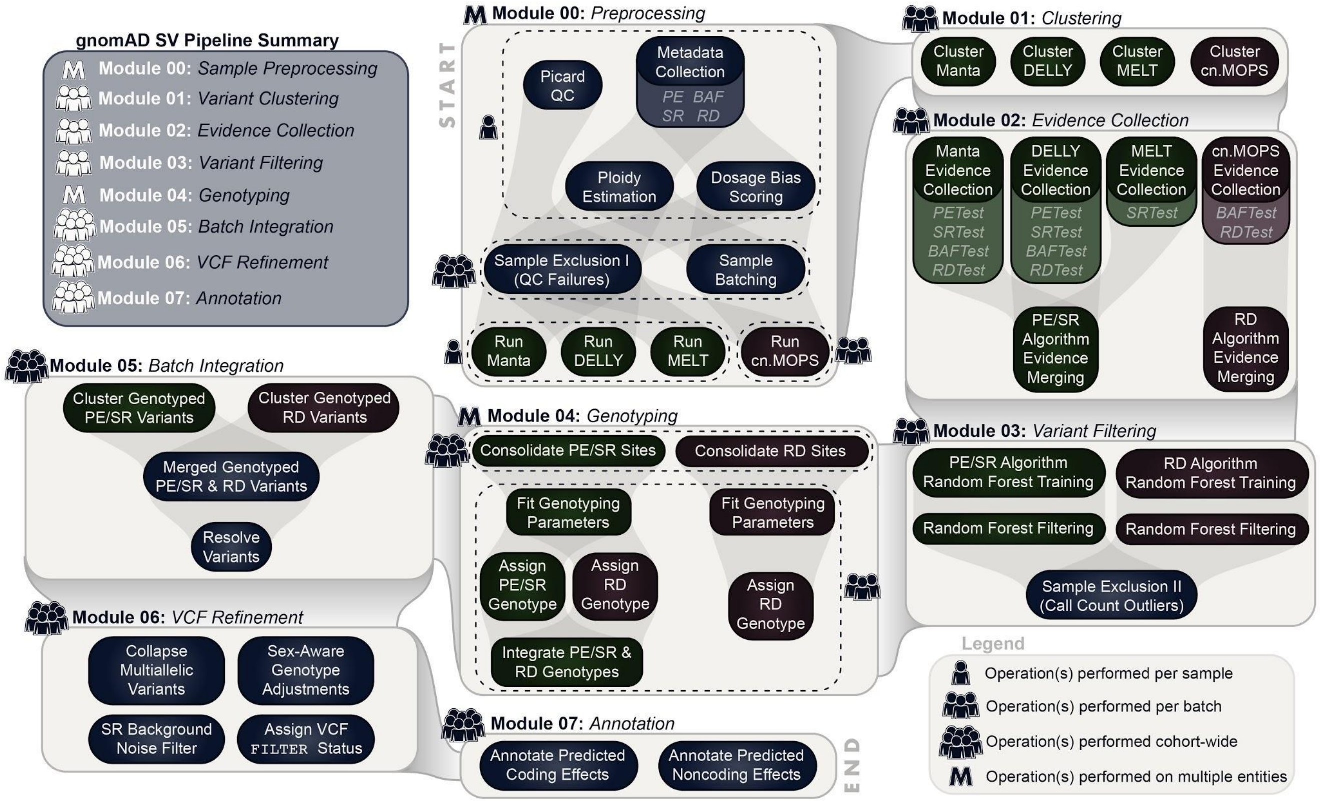


SV classes	Read pair	Read depth	Split read	Assembly
Deletion				
Novel sequence insertion		Not applicable		
Mobile-element insertion		Not applicable		
Inversion		Not applicable		
Interspersed duplication				
Tandem duplication				

Many algorithms often specialized for size range or variant type.

Alkan et al Nat. Rev. Genetics 2011

Example: gnomAD-SV discovery pipeline



gnomAD SV Pipeline Summary

- M **Module 00:** *Sample Preprocessing*
- **Module 01:** *Variant Clustering*
- **Module 02:** *Evidence Collection*
- **Module 03:** *Variant Filtering*
- M **Module 04:** *Genotyping*
- **Module 05:** *Batch Integration*
- **Module 06:** *VCF Refinement*
- **Module 07:** *Annotation*

START

M **Module 00:** *Preprocessing*

Picard QC

Metadata Collection

PE BAF SR RD

Ploidy Estimation

Dosage Bias Scoring

Sample Exclusion I (QC Failures)

Sample Batching

Run Manta

Run DELLY

Run MELT

Run cn.MOPS

Module 01: *Clustering*

Cluster Manta

Cluster DELLY

Cluster MELT

Cluster cn.MOPS

Module 02: *Evidence Collection*

Manta Evidence Collection

PETest SRTest BAFTest RDTest

DELLY Evidence Collection

PETest SRTest BAFTest RDTest

MELT Evidence Collection

SRTest

cn.MOPS Evidence Collection

BAFTest RDTest

PE/SR Algorithm Evidence Merging

RD Algorithm Evidence Merging

Module 03: *Variant Filtering*

PE/SR Algorithm Random Forest Training

RD Algorithm Random Forest Training

Random Forest Filtering

Random Forest Filtering

Sample Exclusion II (Call Count Outliers)

M **Module 04:** *Genotyping*

Consolidate PE/SR Sites

Consolidate RD Sites

Fit Genotyping Parameters

Fit Genotyping Parameters

Assign PE/SR Genotype

Assign RD Genotype

Assign RD Genotype

Integrate PE/SR & RD Genotypes

Module 05: *Batch Integration*

Cluster Genotyped PE/SR Variants

Cluster Genotyped RD Variants

Merged Genotyped PE/SR & RD Variants

Resolve Variants

Module 06: *VCF Refinement*

Collapse Multiallelic Variants

Sex-Aware Genotype Adjustments

SR Background Noise Filter

Assign VCF `FILTER` Status

Module 07: *Annotation*

Annotate Predicted Coding Effects

Annotate Predicted Noncoding Effects

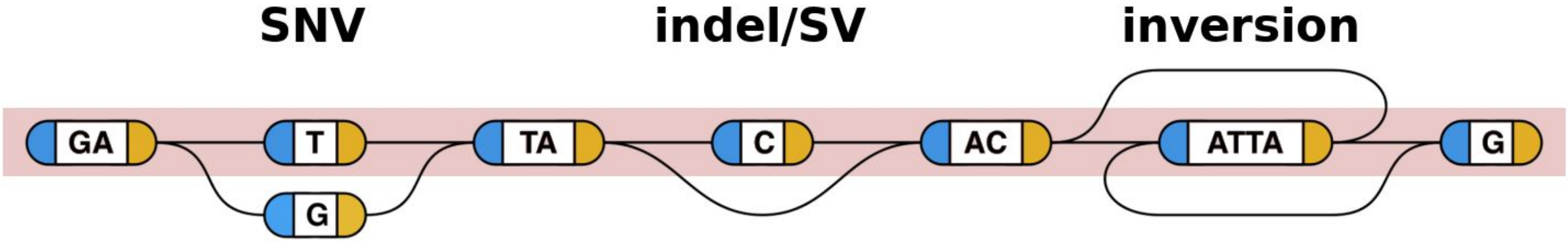
END

Legend

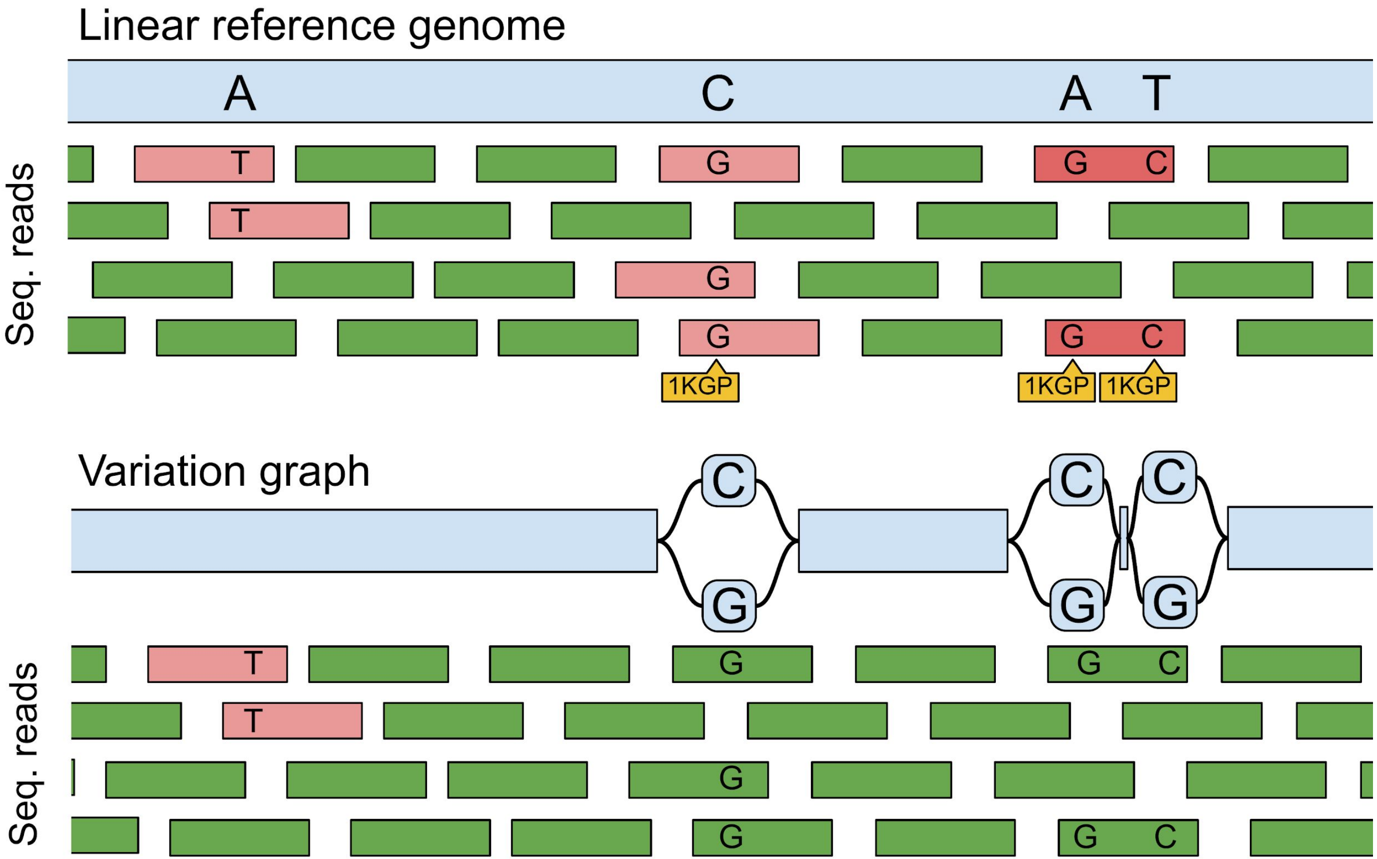
- Operation(s) performed per sample
- Operation(s) performed per batch
- Operation(s) performed cohort-wide
- M Operation(s) performed on multiple entities

Variation Graphs / Genome graphs / Pangenomes

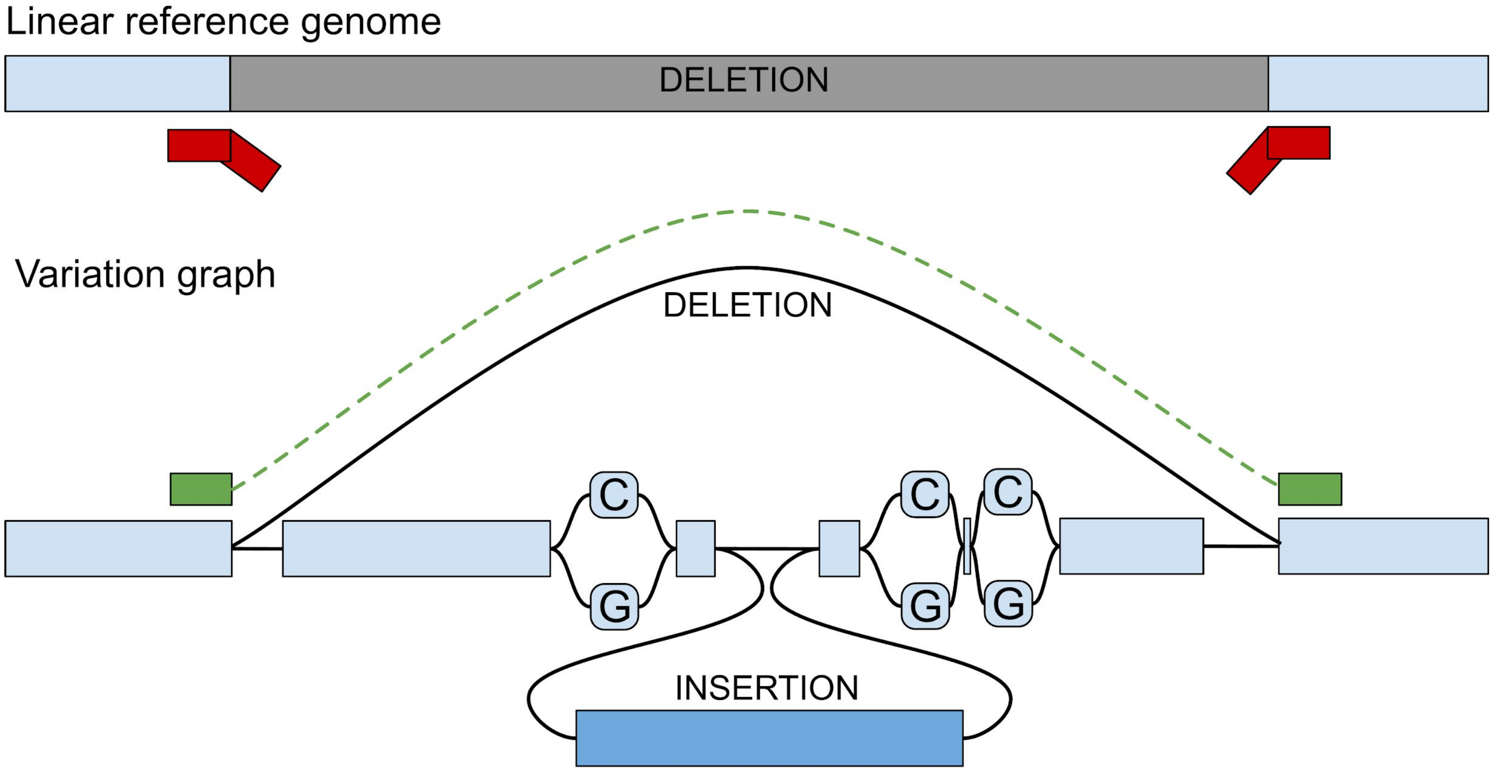
An approach to incorporating information on human diversity into the genomic reference.



Sequencing reads map better on variation graphs



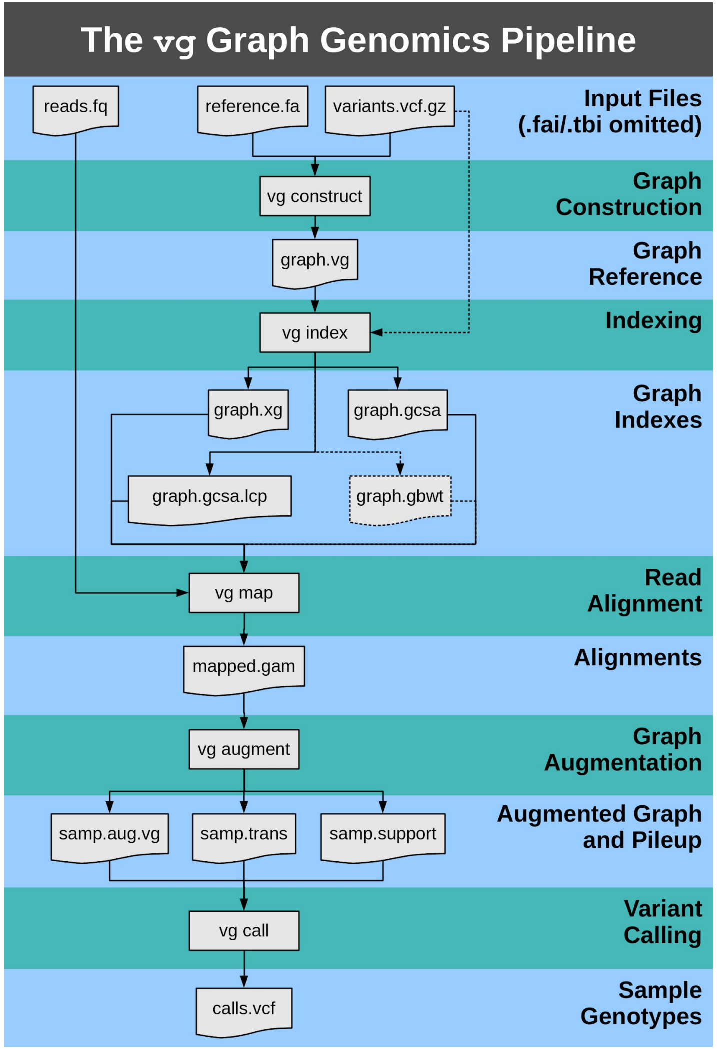
Structural Variants (SVs)



vg

is a complete,
open source solution
for graph construction,
read mapping,
and variant calling.

https://github.com/vgteam/vg



The vg Graph Genomics Pipeline

Input Files (.fai/.tbi omitted): reads.fq, reference.fa, variants.vcf.gz

Graph Construction: vg construct

Graph Reference: graph.vg

Indexing: vg index

Graph Indexes: graph.xg, graph.gcsa, graph.gcsa.lcp, graph.gbwt

Read Alignment: vg map

Alignments: mapped.gam

Graph Augmentation: vg augment

Augmented Graph and Pileup: samp.aug.vg, samp.trans, samp.support

Variant Calling: vg call

Sample Genotypes: calls.vcf

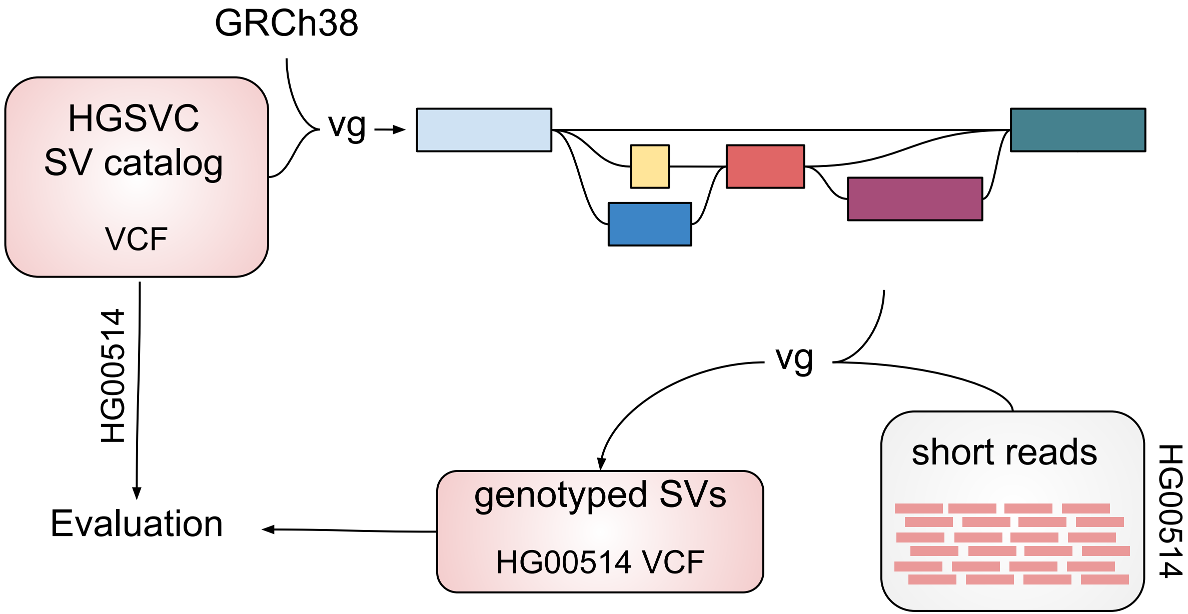
SV pangenomes + short reads -> genotypes

Genotype known SVs from public catalogs in short-read datasets using vg.

1. Test genotyping performance and compare with existing methods
   - Hickey et al. Genome Biology 2020
2. Genotype SVs in a large number of individuals
   - Sirén et al. bioRxiv 2020
3. Find associations between SVs and phenotypes/diseases

Long-read sequencing studies as truth-set

HGSVC sequenced 3 genomes with PacBio sequencing and discovered ~60K SVs

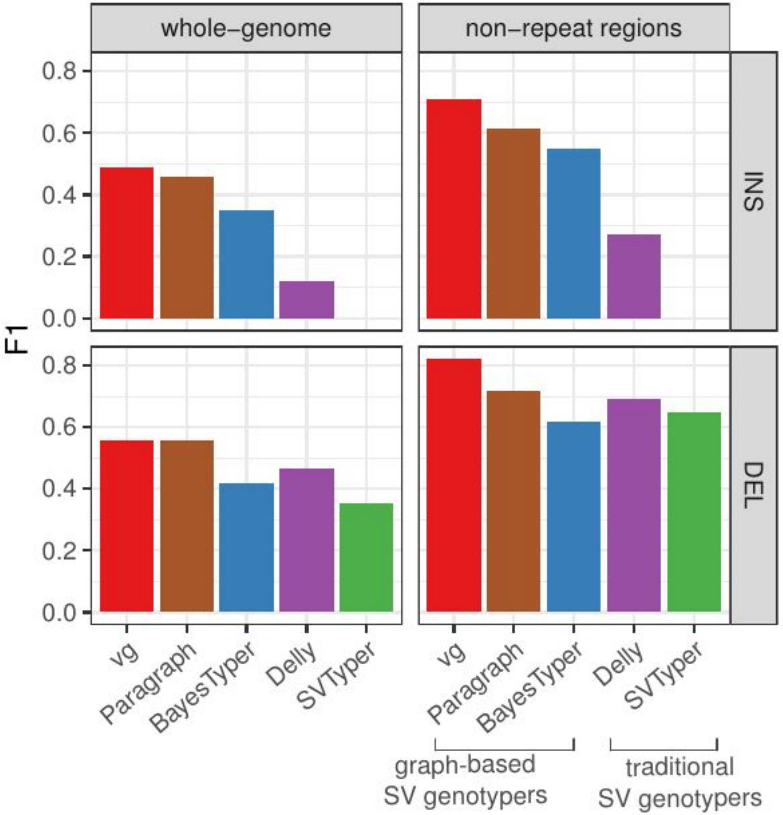


Hickey et al. Genome Biology 2020

vg is better at genotyping SVs

All graph-based methods in general work better.

Especially for insertions



Hickey et al. Genome Biology 2020

Genotype SVs in TOPMed samples

“The goal of the TOPMed program is to generate scientific resources that will improve the understanding of heart, lung, blood, and sleep disorders and advance precision medicine”

-> 100,000s of genomes sequenced with short-reads

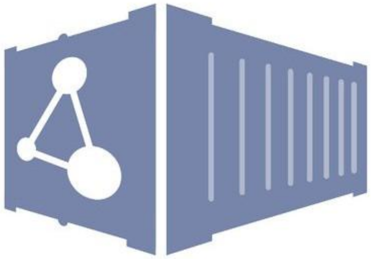
Bring the tools to the data with the BioData Catalyst ecosystem

NIH
National Heart, Lung,
and Blood Institute

BioData
CATALYST

Dockstore + Gen3 + Terra

- Using BioData Catalyst as a Fellow
- WDL workflow in **Dockstore**.
- TOPMed data imported from **Gen3**.
- Genotyping and exploratory analysis on **Terra**



In January, read mapping with vg was slow

Aligning short reads and genotyping SVs cost ~$12 per sample

In January, read mapping with vg was slow

Aligning short reads and genotyping SVs cost ~$12 per sample

Then, vg giraffe was finalized and it's blazingly fast!

Minimizer index (fast seeding), distance index (fast clustering), haplotype index (fast recombination avoidance).

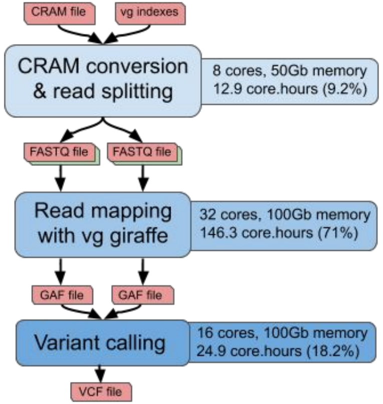
Sirén et al. bioRxiv 2020

Now my workflow costs ~$1.2 per sample!

It’s Giraffe time!

Aka time to genotype SVs across lots of samples

Pangenome with structural variants from 3 long-read sequencing studies: ~15 genomes (SVPOP, HGSVC, GIAB)



CRAM file

vg indexes

CRAM conversion & read splitting

8 cores, 50Gb memory
12.9 core.hours (9.2%)

FASTQ file

FASTQ file

Read mapping with vg giraffe

32 cores, 100Gb memory
146.3 core.hours (71%)

GAF file

GAF file

Variant calling

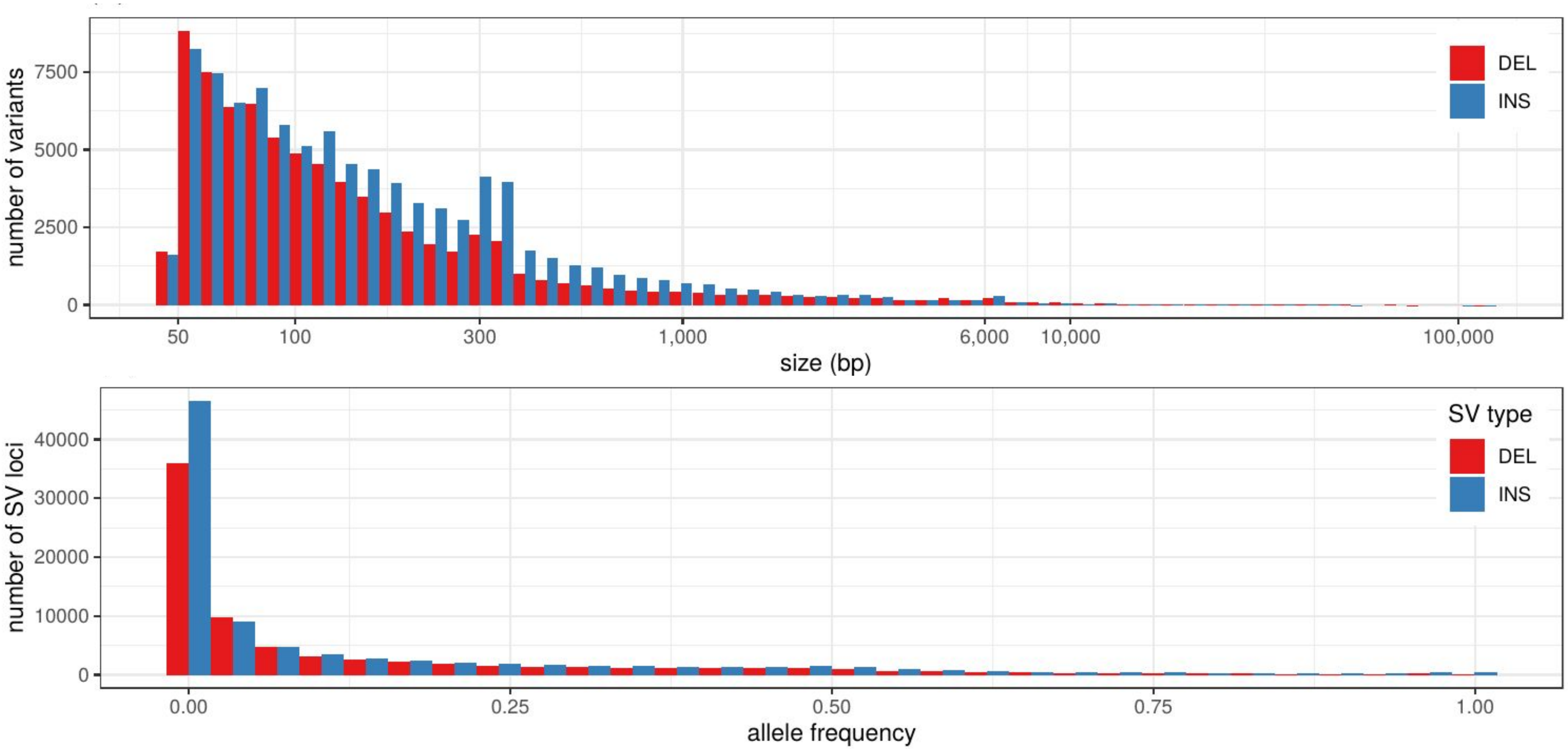
16 cores, 100Gb memory
24.9 core.hours (18.2%)

VCF file

2,000 MESA samples: 4 days, $1.11 per sample
3,202 1KGP samples: 6 days, $1.56 per sample

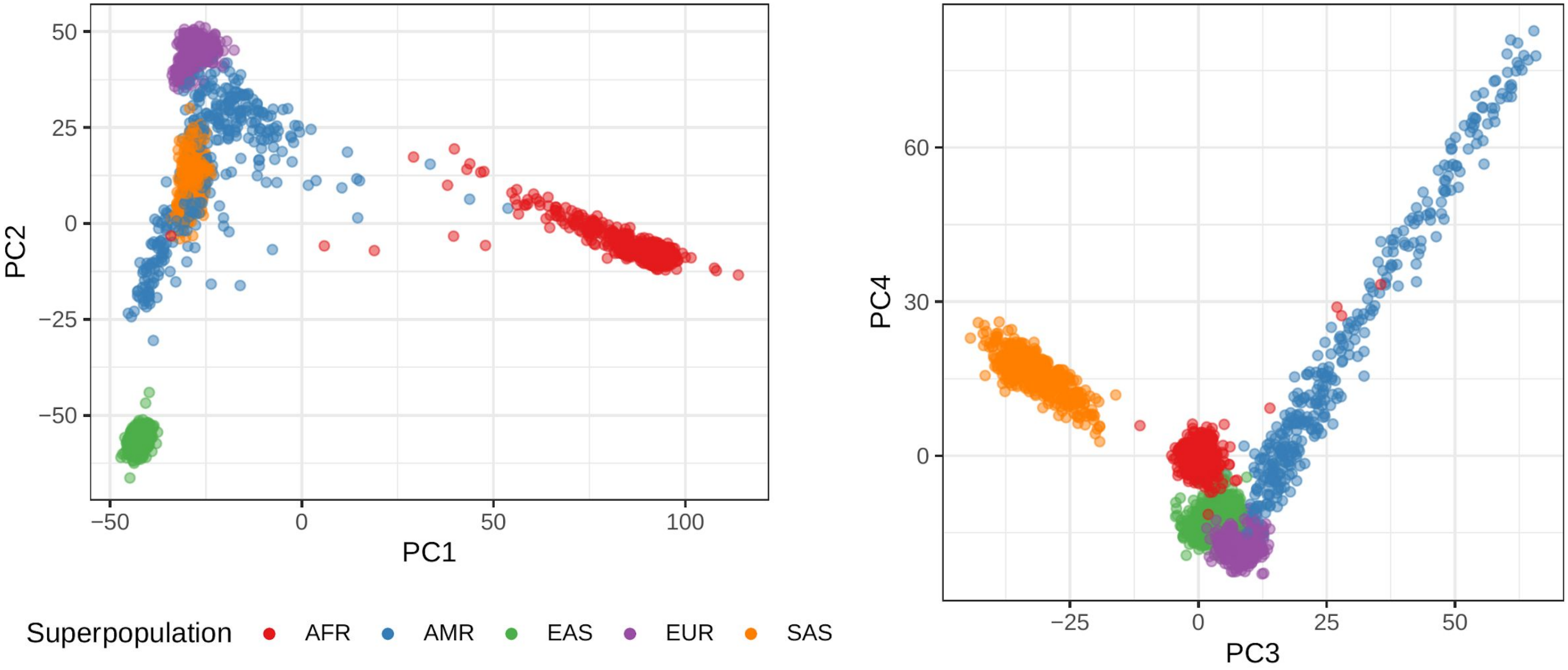
Philipp Bayer @PhilippBayer · Dec 6
Replying to @JMonlong
I'll be very disappointed if this gif isn't used in their talks!

Structural variant frequencies across thousands of diverse individuals



SV genotyped: 1.7 million alleles clustered in **167 thousand SV loci**
89.4% shorter than 500 bp; 83.9% in repeat-rich regions.
67-93% missing from gnomAD-SV or 1000GP SV catalog

Allele frequencies in diverse populations

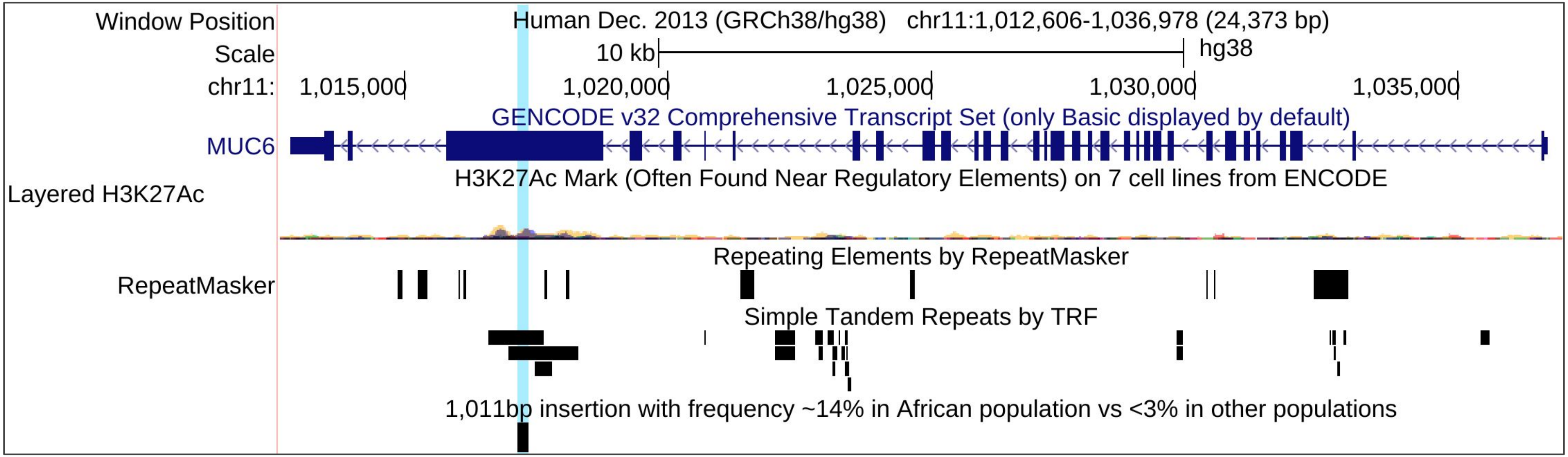


Valuable information for variant annotation

Example: 1,011bp coding “insertion”

- Short tandem repeat expansion.
- Common in the African super-population.

- Missing from large SV databases (gnomAD-SV, 1000GP)

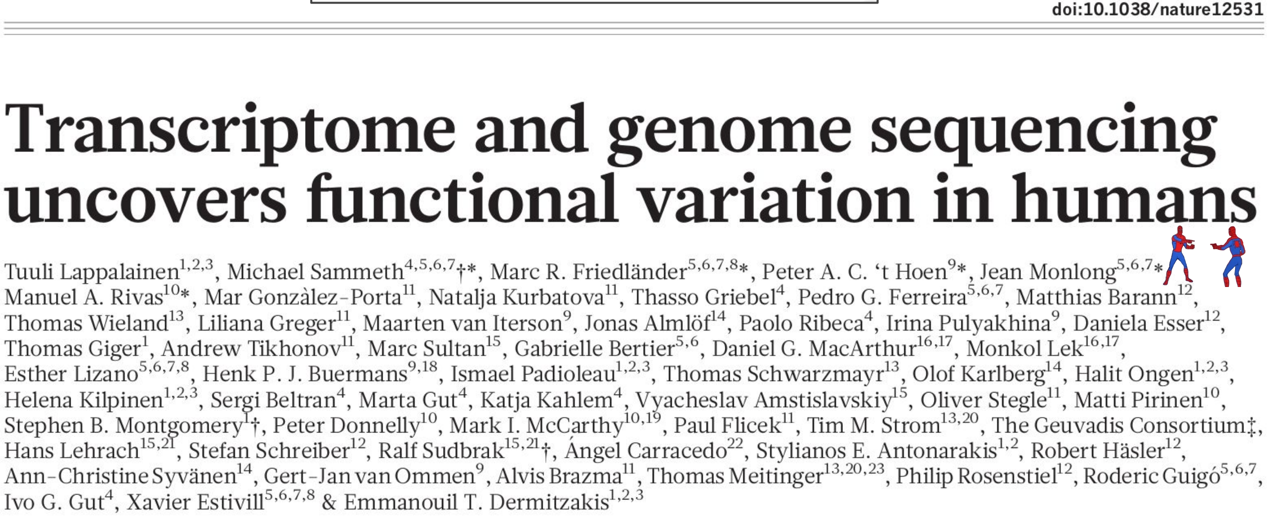


Window Position
Human Dec. 2013 (GRCh38/hg38) chr11:1,012,606-1,036,978 (24,373 bp)
Scale
10 kb
hg38
chr11:
1,015,000
1,020,000
1,025,000
1,030,000
1,035,000
GENCODE v32 Comprehensive Transcript Set (only Basic displayed by default)
MUC6
H3K27Ac Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
Repeating Elements by RepeatMasker
RepeatMasker
Simple Tandem Repeats by TRF
1,011bp insertion with frequency ~14% in African population vs <3% in other populations

1000 Genomes Project + Geuvadis

A subset of 445 samples have expression data (RNA-seq) publicly available.

GEUVADIS CONSORTIUM

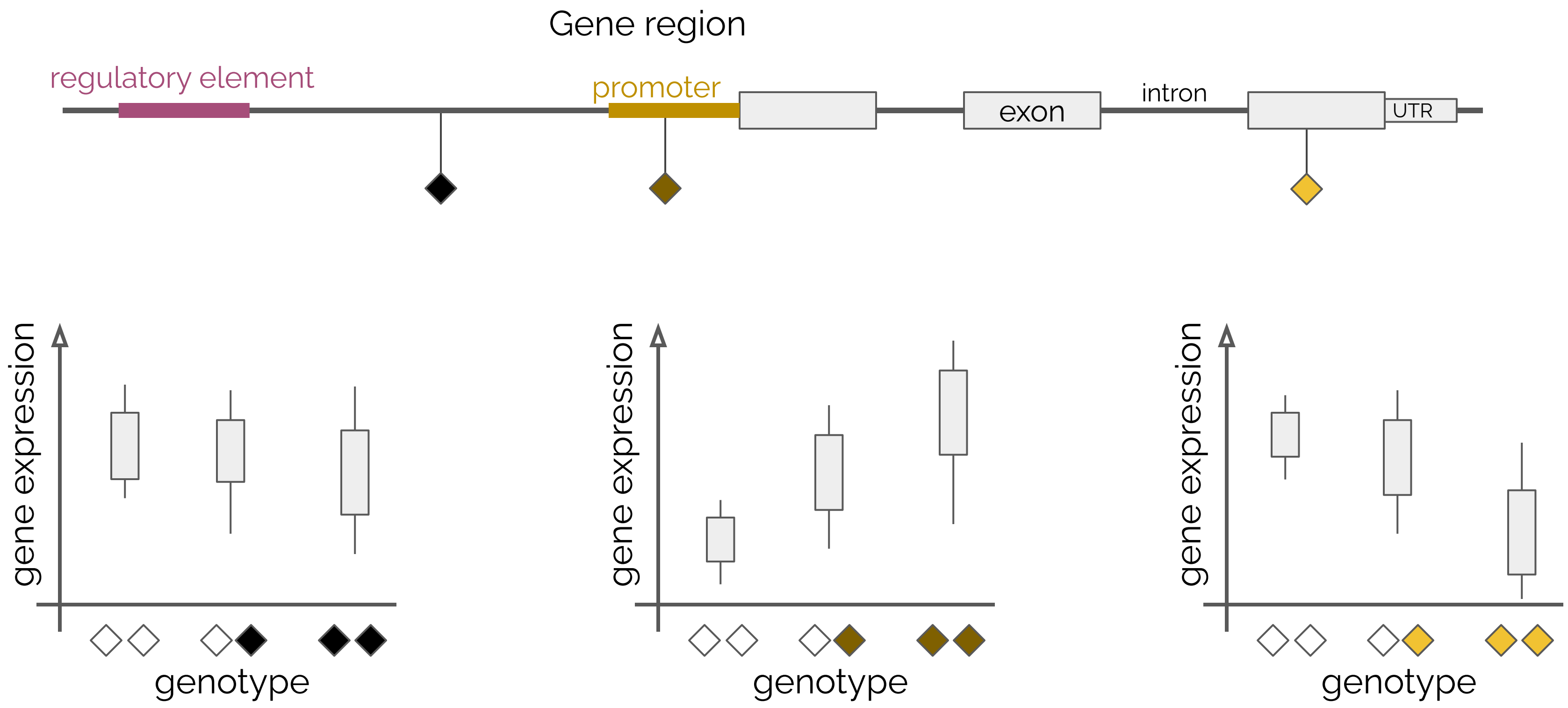


doi:10.1038/nature12531

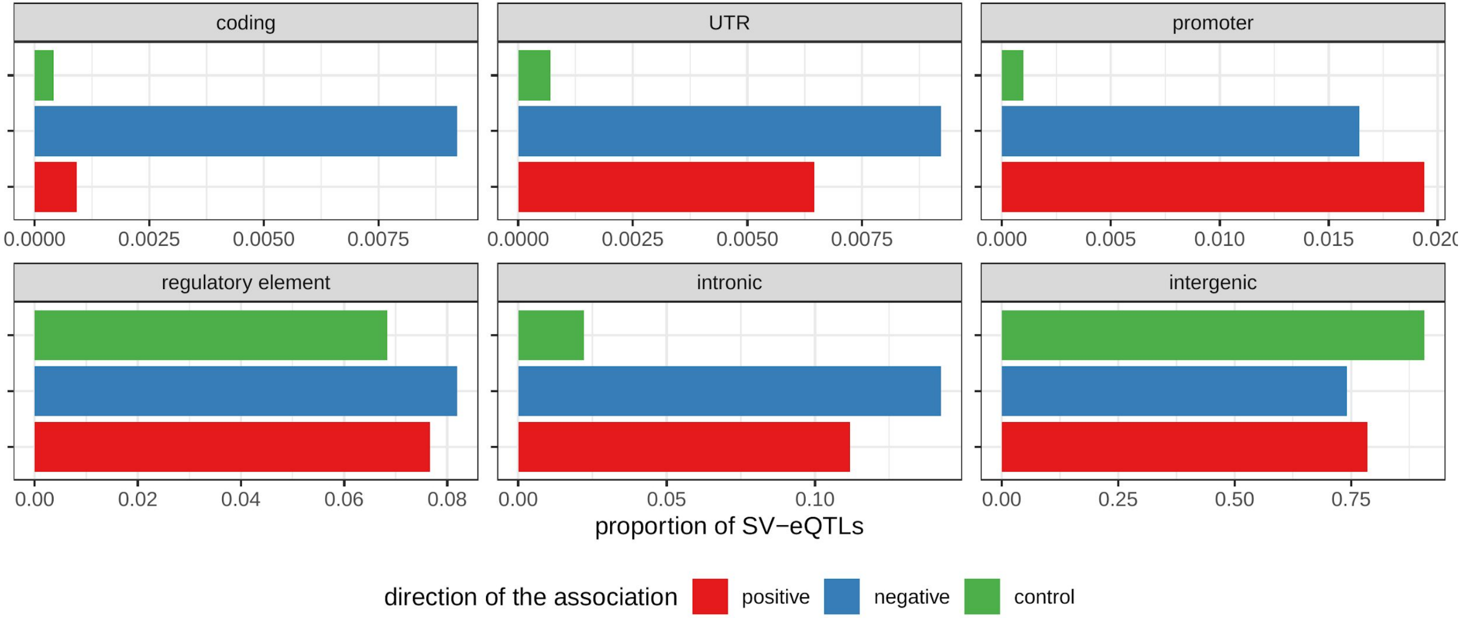
Transcriptome and genome sequencing uncovers functional variation in humans

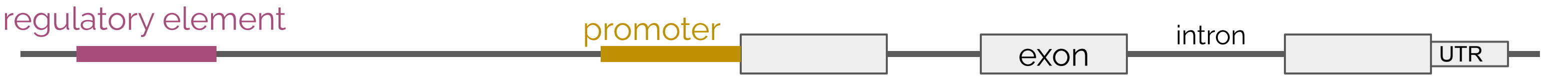
Tuuli Lappalainen[1,2,3], Michael Sammeth[4,5,6,7]†*, Marc R. Friedländer[5,6,7,8]*, Peter A. C. 't Hoen[9]*, Jean Monlong[5,6,7]*, Manuel A. Rivas[10]*, Mar Gonzàlez-Porta[11], Natalja Kurbatova[11], Thasso Griebel[4], Pedro G. Ferreira[5,6,7], Matthias Barann[12], Thomas Wieland[13], Liliana Greger[11], Maarten van Iterson[9], Jonas Almlöf[14], Paolo Ribeca[4], Irina Pulyakhina[9], Daniela Esser[12], Thomas Giger[1], Andrew Tikhonov[11], Marc Sultan[15], Gabrielle Bertier[5,6], Daniel G. MacArthur[16,17], Monkol Lek[16,17], Esther Lizano[5,6,7,8], Henk P. J. Buermans[9,18], Ismael Padioleau[1,2,3], Thomas Schwarzmayr[13], Olof Karlberg[14], Halit Ongen[1,2,3], Helena Kilpinen[1,2,3], Sergi Beltran[4], Marta Gut[4], Katja Kahlem[4], Vyacheslav Amstislavskiy[15], Oliver Stegle[11], Matti Pirinen[10], Stephen B. Montgomery[1]†, Peter Donnelly[10], Mark I. McCarthy[10,19], Paul Flicek[11], Tim M. Strom[13,20], The Geuvadis Consortium‡, Hans Lehrach[15,21], Stefan Schreiber[12], Ralf Sudbrak[15,21]†, Ángel Carracedo[22], Stylianos E. Antonarakis[1,2], Robert Häsler[12], Ann-Christine Syvänen[14], Gert-Jan van Ommen[9], Alvis Brazma[11], Thomas Meitinger[13,20,23], Philip Rosenstiel[12], Roderic Guigó[5,6,7], Ivo G. Gut[4], Xavier Estivill[5,6,7,8] & Emmanouil T. Dermitzakis[1,2,3]

Expression Quantitative Trait Locus (eQTL)



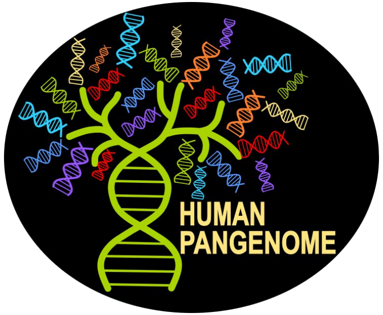
~2,000 SV-eQTLs in the Geuvadis dataset



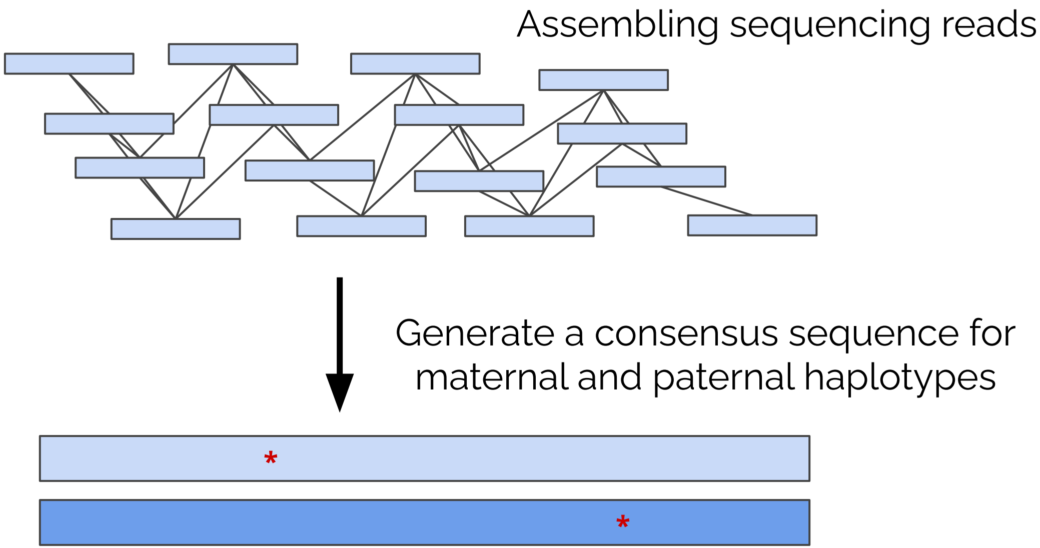


Next: pangenomes from de novo assemblies

The Human Pangenome Reference Consortium (HPRC) will produce phased de novo assemblies for >300 diverse individuals.

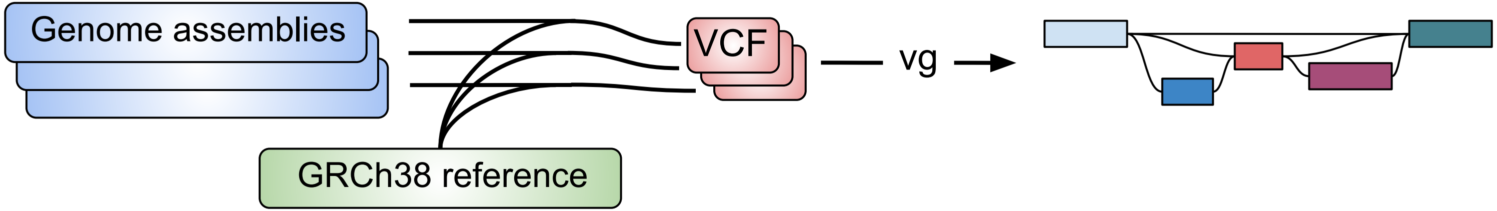


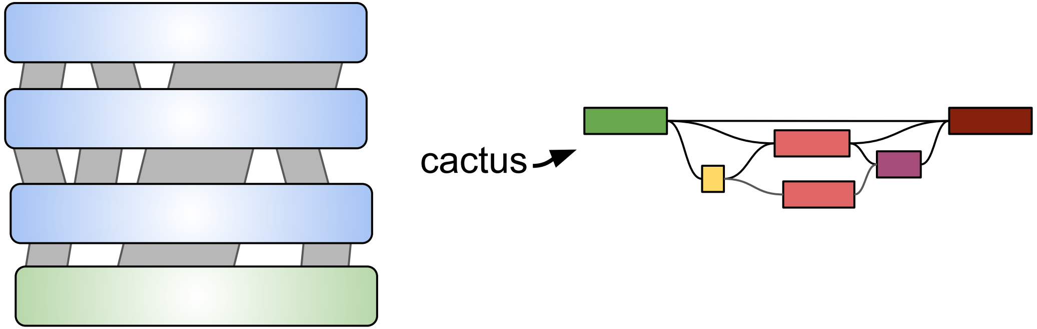
https://humanpangenome.org/


Assembling sequencing reads

Generate a consensus sequence for maternal and paternal haplotypes

Different strategies to construct pangenomes

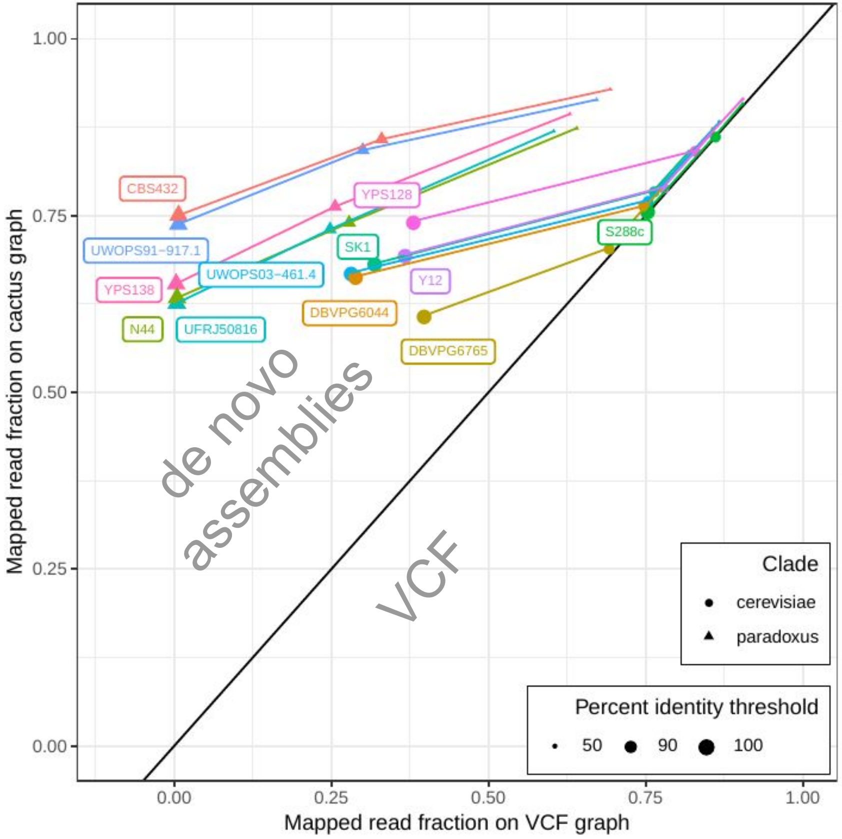




Graph from de novo assemblies

Experiment with 12 yeast strains.

- better read mapping.
- SV better supported by reads.



Hickey et al. Genome Biology 2020

SV pangenomes + short reads -> genotypes

Genotype known SVs from public catalogs in short-read datasets using vg.

1. Test genotyping performance and compare with existing methods
   ○ Hickey et al. Genome Biology 2020

2. Genotype SVs in a large number of individuals
   ○ Sirén et al. bioRxiv 2020

3. Find associations between SVs and phenotypes/diseases

*Build HPRC pangenomes from de novo assemblies and repeat

Acknowledgements

Benedict Paten

Glenn Hickey

Adam Novak

Erik Garrison

Jordan Eizenga

Jouni Siren

David Heller

Jonas Sibbesen

Xian Chang

Charles Markello

Yohei Rosen

Robin Rounthwaite

Susanna Morin

Beth Sheets

Michael Baumann

Brian Hannafious

UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute

NIH National Heart, Lung, and Blood Institute

BioData CATALYST