

Genotyping structural variants in TOPMed using pangenome graphs

Jean Monlong
February 12-13, 2020



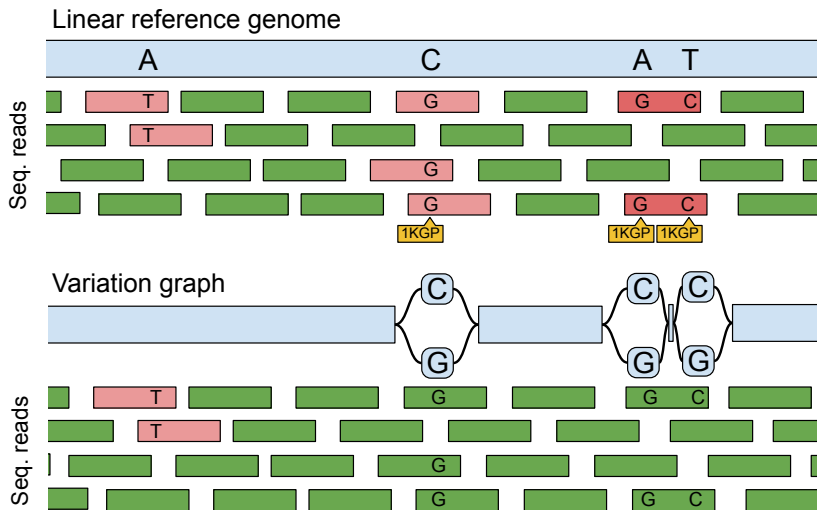
UNIVERSITY OF CALIFORNIA

SANTA CRUZ

Genomics
Institute

GSP-TOPMED ANALYSIS WORKSHOP

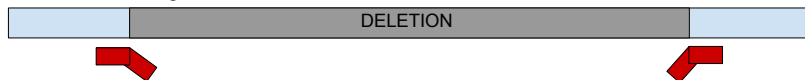
Pangenome graphs and variant-aware read mapping



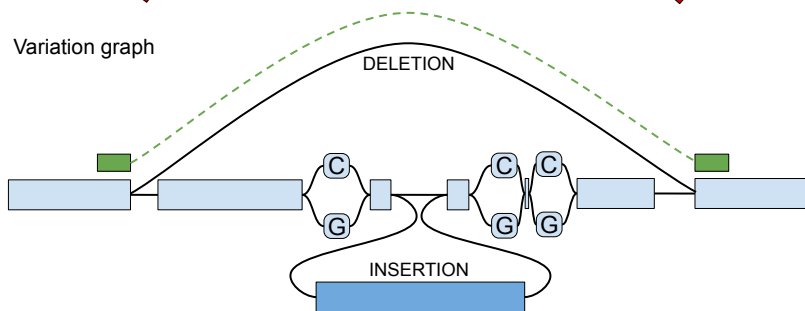
Mapping reads across structural variants

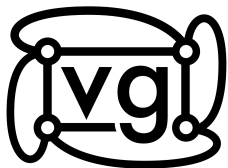
Structural variants (SVs) are genomic variants larger than 50 bp, e.g. insertions, deletions, inversions translocations.

Linear reference genome



Variation graph





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

<https://github.com/vgteam/vg>

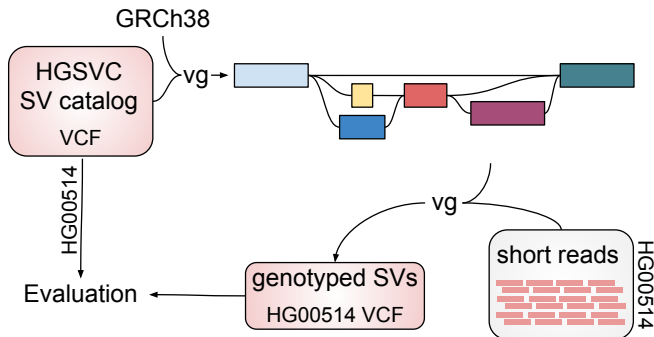
Garrison et al. Nature Biotech 2018

vg can **genotype structural variants from short-read sequencing datasets** starting from public SV catalogs or de novo assemblies.

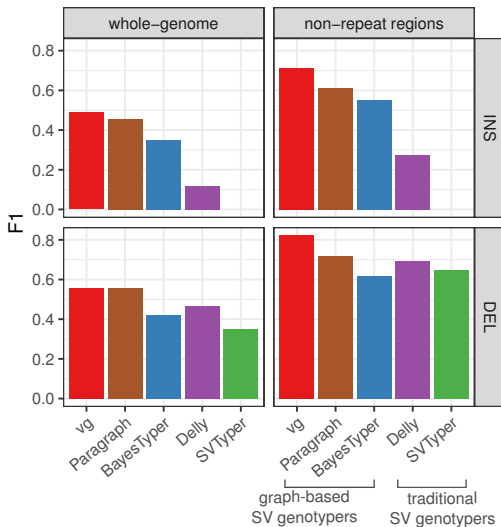
Hickey et al. bioRxiv 2019, in press at Genome Biology

Genotyping SVs from long-read sequencing studies

Ref.	Project	Samples
Chaisson et al. 2019	Human Genome Structural Variation Consortium (HGSVC)	3
Audano et al. 2019	SVPOP	15
Zook et al. 2019	Genome in a Bottle (GIAB)	1



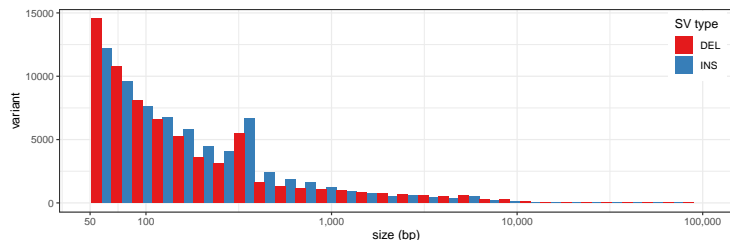
SV genotyping accuracy for deletions and insertions



Non-repeat regions: regions not overlapping segmental duplications or simple repeats

Combined SV catalogs from 3 long-read studies

Ref.	Project	Samples
Chaisson et al. 2019	Human Genome Structural Variation Consortium (HGSVC)	3
Audano et al. 2019	SVPOP	15
Zook et al. 2019	Genome in a Bottle (GIAB)	1



71K deletions and 70K insertions include most of the common deletions and insertions in the population.

760 TOPMed samples genotyped in 5 days



BioData **CATALYST**



GEN3
DATA COMMONS



- ◆ Using **BioData Catalyst** as an alpha user.
- ◆ Workflow in **Dockstore**.
- ◆ TOPMed data imported from **Gen3**.
- ◆ Genotyping and exploratory analysis on **Terra** using workflows and notebooks.
- ◆ ~\$12 per sample (soon <\$4 with new read mapper).

TOPMed data available in Gen3

I selected the MESA cohort and exported the CRAM files to Terra.

NIH National Heart, Lung, and Blood Institute

BioData CATALYST
Powered by Gen3

Submit Data

Documentation

MONLONG

Logout

Dictionary

Exploration

Query

Workspace

Profile

DataFiles

Filters

Medical HistoryDiagnosisSubject

Collapse all

Project id

☐ parent-WHI_HMB-IRB_117676

☐ parent-WHI_HMB-IRB-NPU_25538

☐ parent-ARIC_HMB-IRB_15608

☐ parent-FHS_HMB-IRB-MDS_13132

☐ topmed-COPDGene_HMB-MDS_10277

☐ parent-COPDGene_HMB_10099

☐ topmed-WHI_HMB-IRB_9001

☐ parent-MESA_HMB_7440

☐ parent-CHS_HMB-IRB-MDS_5353

☒ topmed-MESA_HMB4879

☐ topmed-JHS_HMB-IRB_4036

Download

Export All to Terra

Export to PFB

Export to Workspace

Projects1

Subjects4,879

Annotated Sex

no data100%


Race

no data100%

Showing 1 - 20 of 4879 subjects

Project id	Data Format	Race	Annotated Sex	Ethnicity	BP Diastolic	HDL	LDL
topmed-MESA_HMB	CRAMVCF						
topmed-MESA_HMB	CRAMVCF						

WDL workflow for vg in Dockstore



Search Organizations Docs

jmonlong

Workflows

github.com/vgteam/vg_wdl/vg_map_call_sv_cram:svpack

Last Modified: 13 days ago

genome-graph genomics genotyping structural-variant sv variation-graph

Info Launch Versions Files Tools

Workflow Information

Source Code: github.com/vgteam/vg_wdl:svpack

TRS: [#workflow/github.com/vgteam/vg_wdl/vg_map_call_sv_cram](#)

Workflow Path: [/workflows/vg_map_call_sv_cram.wdl](#)

Test File Path: [/params/vg_map_call_sv_cram_test.inputs.json](#)

Checker Workflow: n/a

Descriptor Type: WDL

DOI: n/a

Workflow Version Information

svpack

DOI: n/a

Author: Jean Monlong

E-mail: jmonlong@ucsc.edu

Export as ZIP

Description:

Read mapping and SV genotyping using vg. It takes a CRAM file and graphs containing the structural variants to genotype. The XG and GCSA graph indexes as required, as well as the original VCF used to create the graphs. It studies the reads to find variants and reports them as structural variants. It outputs VCF files. The CRAM

Launch with

DNASTack »

DNAnexus »

Terra »

AnVIL »

Recent Versions

svpack Jan 28, 2020

See all versions

Source Repositories

GitHub

Collections

Structural Variant Calling

vg - Variation Graphs Toolkit

Add to mv collection

Genotyping and analysis on Terra

NIH

National Heart Lung and Blood Institute

BioData CATALYST
Powered by Terra

DASHBOARD

DATA

NOTEBOOKS

WORKFLOWS

JOB HISTORY

1

← Back to list

Workflow Statuses
✓ Succeeded: 96

Workflow Configuration
bdc1at-team-calcium-alpha-users/vg_map

Data Entity
vg_map_call_sv_cram_2020-02-09T15-3
aligned_reads_390samples_b3_set

Submitted by
jmonlong@ucsc.edu
Feb 9, 2020, 7:34 AM

Submission ID
fa927b84-760d-497d-ab83-5a24bc...

Total Run Cost
\$1,152.33

Call Caching
Enabled

Completion status

▼

	Data Entity	Last Changed	Status	Run Cost	Message
View	02d6ae72-8c23-4...	Feb 9, 2020, 8:48 P...	✓ Succeeded	\$14.48	
View	02e15f6c-2962-41...	Feb 9, 2020, 7:41 P...	✓ Succeeded	\$11.93	
View	03a83563-3153-4...	Feb 9, 2020, 8:08 P...	✓ Succeeded	\$12.59	
View	09a08c17-2dd0-4...	Feb 9, 2020, 7:38 P...	✓ Succeeded	\$11.34	
View	09fbc05c-875b-4c...	Feb 9, 2020, 8:10 P...	✓ Succeeded	\$10.77	
View	0b195937-d9d6-4...	Feb 9, 2020, 7:29 P...	✓ Succeeded	\$11.53	
View	106502c4-a0a9-4...	Feb 9, 2020, 8:33 P...	✓ Succeeded	\$11.30	
View	10af4506-7d1b-4d...	Feb 9, 2020, 6:30 P...	✓ Succeeded	\$11.94	
View	112247cc-5576-4f...	Feb 9, 2020, 6:43 P...	✓ Succeeded	\$12.00	
View	13596085-9c6d-4...	Feb 9, 2020, 8:55 P...	✓ Succeeded	\$15.73	

NIH

National Heart Lung and Blood Institute

BioData CATALYST
Powered by Terra

PREVIEW (READ-ONLY)

EDIT

PLAYGROUND MODE

⋮

PCA

```
In [5]: pca.o = prcomp(gt.nat)
pc.df = tibble(sample=rownames(pca.o$x), PC1=pca.o$x[,1], PC2=pca.o$x[,2], PC3=pca.o$x[,3],
sdev.df = tibble(sdev=pca.o$sdev, PC=1:length(pca.o$sdev))
```

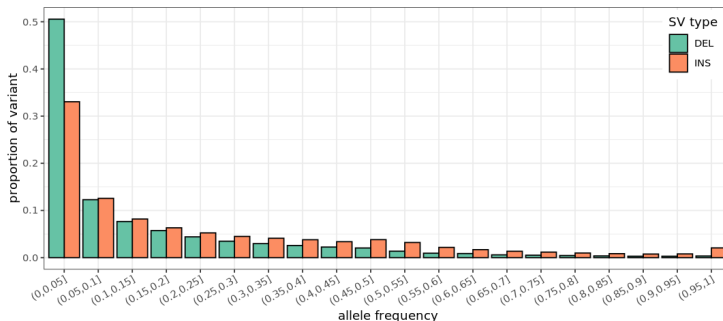
The scatterplot using the first 2 PCs and coloring with the third PC.

```
In [6]: ggplot(pc.df, aes(x=PC1, y=PC2, colour=PC3)) +
  geom_point(alpha=.8) +
  scale_colour_gradientn(colors=c('indianred2', 'darkblue')) +
  theme_bw()
sdev.df %>% filter(PC>21) %>%
  ggplot(aes(x=PC, y=sdev)) +
  geom_bar(stat='identity') +
  theme_bw()
```

SV genotyped in 760 diverse genomes



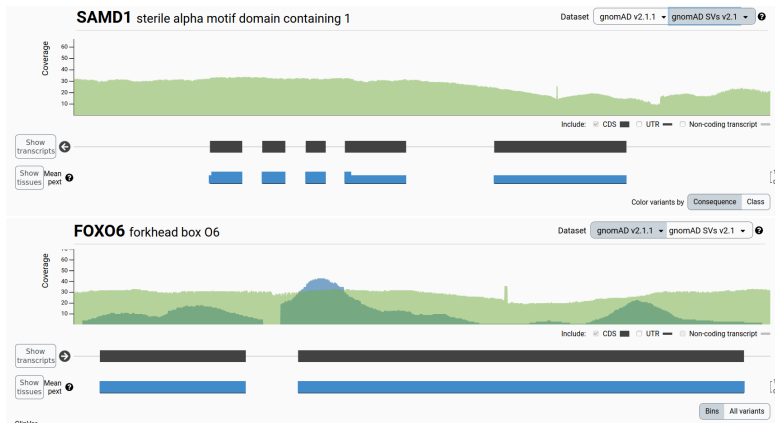
Frequency estimates



- ◆ Insertions slightly more frequent than deletions...
- ◆ ...especially for larger variants.
- ◆ Hundreds of fixed SVs, especially insertions.

Fixed insertions

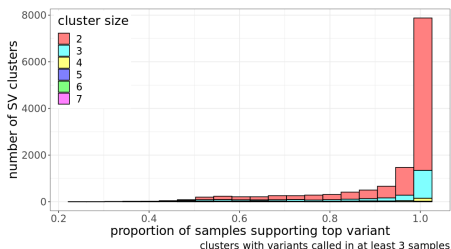
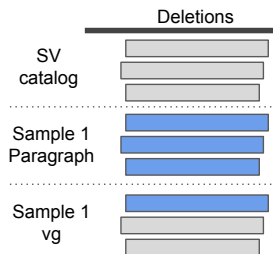
- ◆ 736 insertions with allele frequency >0.99 .
- ◆ Two repeat expansions in coding regions of SAMD1 and FOXO6.



Screenshots from <https://gnomad.broadinstitute.org/>

Fine-tuning breakpoints of deletions

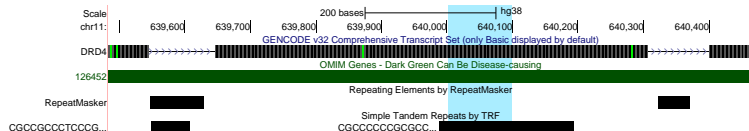
- Although sequence-resolved, many deletions are extremely similar and likely near-duplicates of the same real deletion.



- In >9K clusters, the 760 samples supported mostly one variant.

Coding deletions with fine-tuned breakpoints

- ◆ 95 of the fine-tuned deletions overlap coding regions.
- ◆ Two near-duplicated deletions overlapped DRD4 gene.
 - ◆ Within long short tandem repeat...
 - ◆ 96 bp or 97 bp deletion?
 - All samples supported the 96 bp deletion.
 - ◆ Known 2-copies version of the 48nt repeat (DRD4-2R).



Conclusions

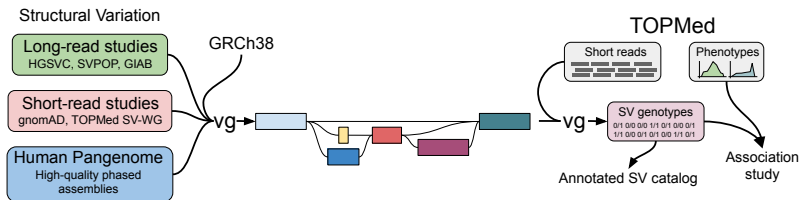
- ◆ The vg toolkit can integrate and genotype SVs.
- ◆ 760 TOPMed samples genotyped in 5 days using the BioData Catalyst ecosystem.
- ◆ SV catalog from long-read studies annotated with frequencies and better breakpoint resolution.

Conclusions

- ◆ The vg toolkit can integrate and genotype SVs.
- ◆ 760 TOPMed samples genotyped in 5 days using the BioData Catalyst ecosystem.
- ◆ SV catalog from long-read studies annotated with frequencies and better breakpoint resolution.

Future directions

- ◆ Documented workflows for the BioData Catalyst community (and GSP through NHGRI AnVIL).
- ◆ More SVs genotyped in more TOPMed samples for association studies.



Acknowledgment

vg Team

Benedict Paten

Glenn Hickey

David Heller

Adam Novak

Erik Garrison

Jouni Siren

Jordan Eizenga

Charles Markello

Xian Chang

Robin Rounthwaite

Jonas Sibbesen

Eric T. Dawson

BioData Catalyst Team

Beth Sheets (talk to her!)

Michael Baumann

Brian Hannafious



UNIVERSITY OF CALIFORNIA
SANTA CRUZ

Genomics
Institute

EMBL



European Molecular
Biology Laboratory

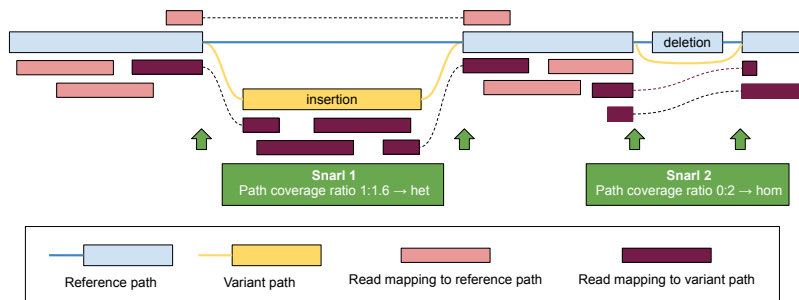


National Heart, Lung,
and Blood Institute

BioData

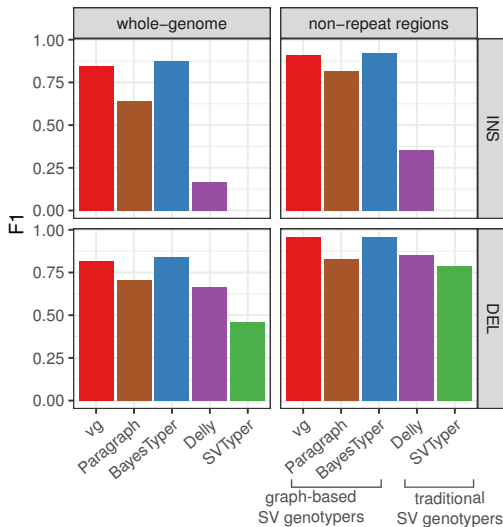
CATALYST

Genotyping variants in vg



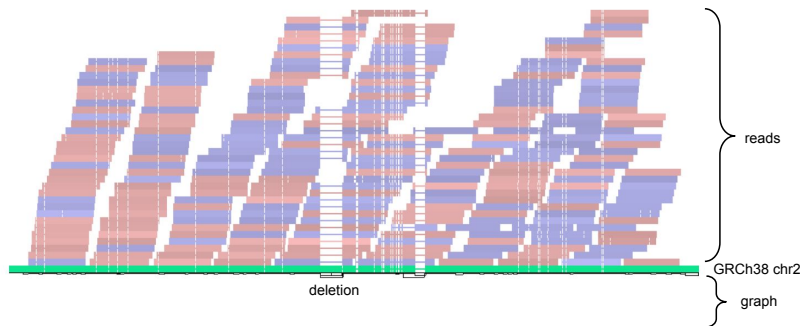
- ◆ Genotyping is based on the path coverage.
- ◆ A snarl is a variant site in the graph, a “bubble”.

Results on HGSVC - Simulated reads



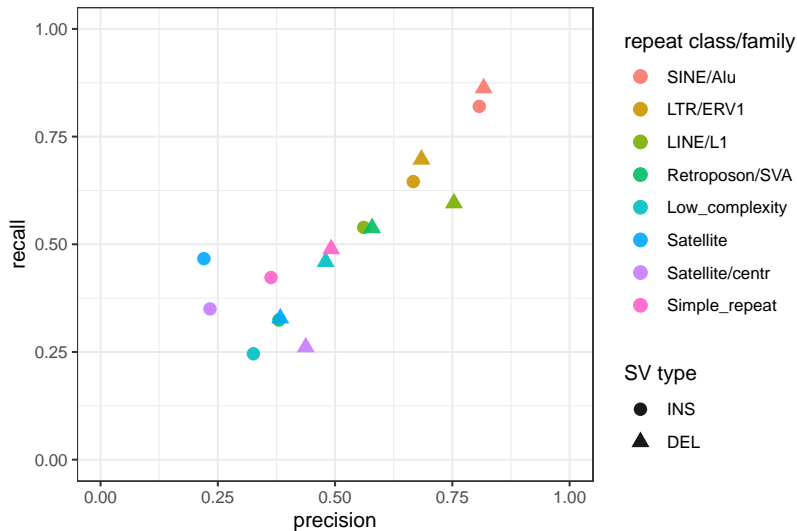
Non-repeat regions: regions not overlapping segmental duplications or simple repeats

Deletion correctly genotyped by vg



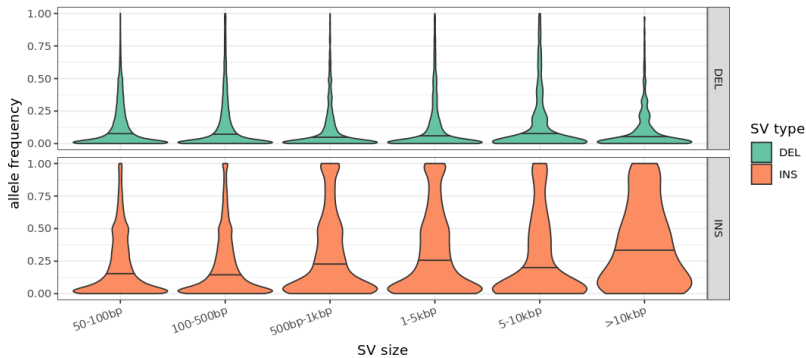
51 bp homozygous deletion in the 3' UTR of the LONRF2 gene.

Simple repeat/low complexity regions are challenging

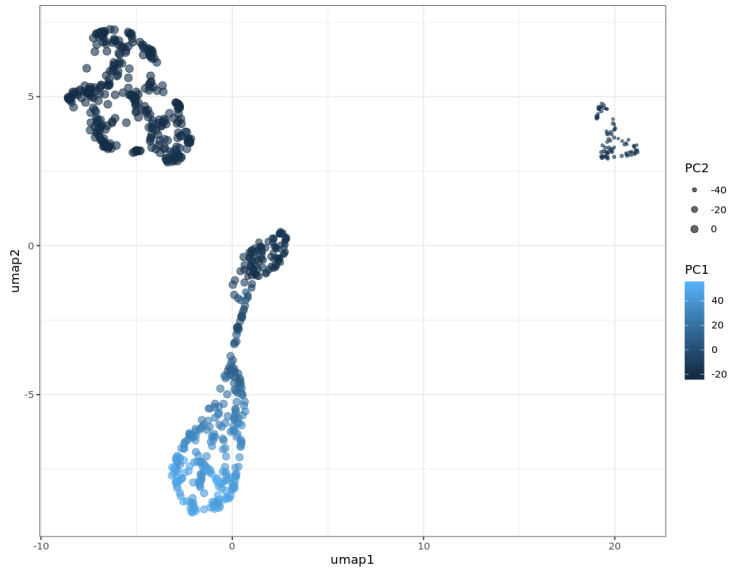


SV sequence annotated with RepeatMasker. Class assigned if covered $\geq 80\%$ by a repeat element.

Frequency distribution vs variant size



UMAP



Genotype quality and samples with genotype calls

