Mutation Calling: Benchmark 4

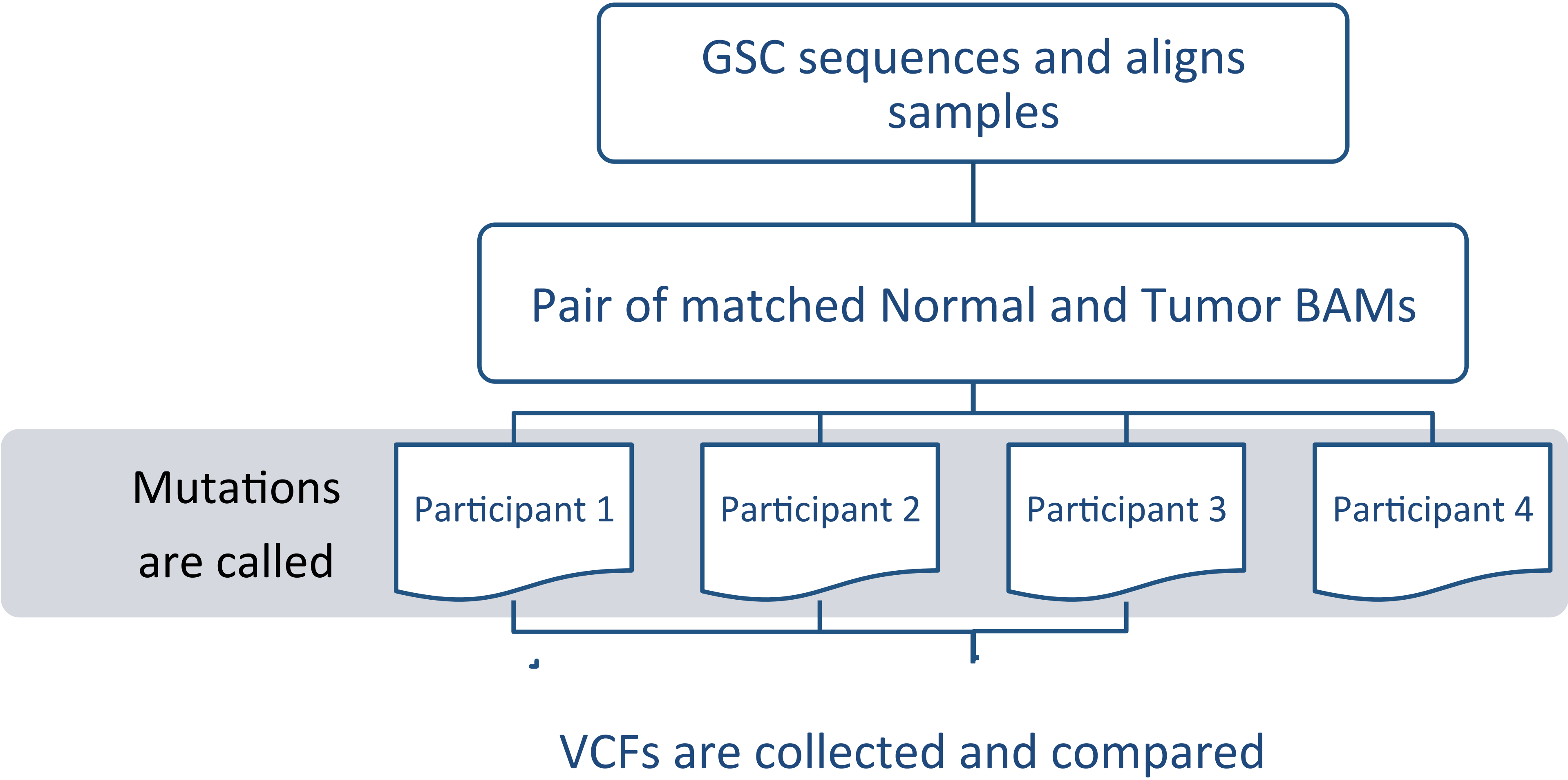
(call for participation)

Adam Ewing, UCSC

TCGA 2nd Annual Scientific Symposium

Baskin Engineering UC SANTA CRUZ

Mutation Calling Benchmark Process



GSC sequences and aligns samples

Pair of matched Normal and Tumor BAMs

Mutations are called

Participant 1

Participant 2

Participant 3

Participant 4

VCFs are collected and compared

Background / History

Mutation types:

SNV (single nucleotide variant)
INDEL (insertions and deletions < 100 bp)
SV (insertions, deletions, duplications, inversions > 100 bp)
CNV (copy number variation)

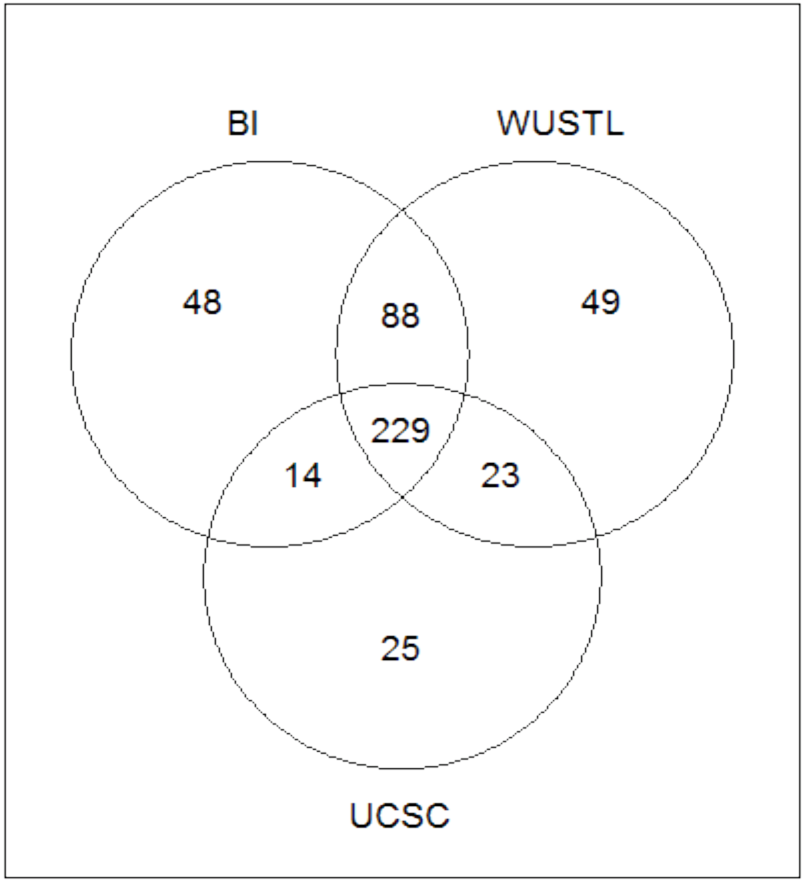
History:

- Benchmark 1: **SNVs** on six pairs of **whole genomes**
- Benchmark 2: **SNVs** on 14 **exomes**
- Benchmark 3: **SNVs** on 25 **exomes** with validation
- Benchmark 4: **SNVs**, **INDELs**, **SVs**, and **CNVs** on **whole genomes from cell lines**

Goals: why do another benchmark?

- TCGA must measure and set standards for the accuracy of mutation calls
- Evaluate performance on INDELs, SVs, CNVs
- This is a **controlled experiment:**
  - Simulate normal contamination
    - Mix tumor and normal cell line data
  - Simulate subclonal expansion
    - Mix spiked-in mutations
- **Wide participation:** cell line data is public

Goals: why do another benchmark?



Andrey Sivachenko, Broad Institute

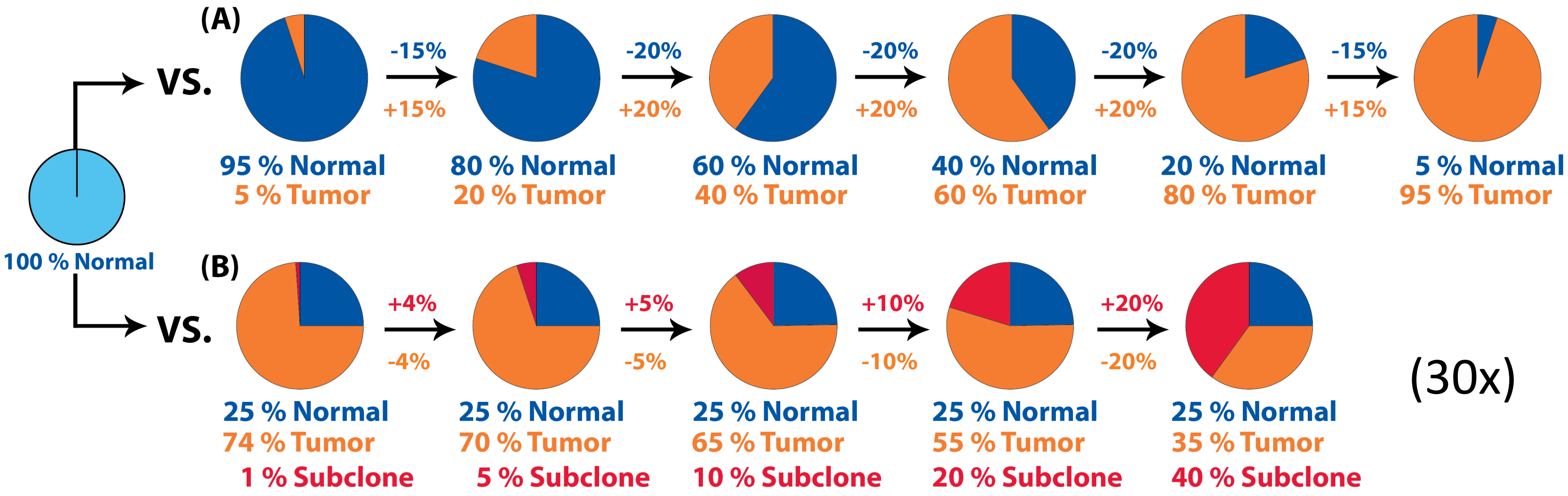
Cancer genomics depends on somatic mutation calls

Samples for benchmark 4: Cell lines

- HCC1143 / HCC1143 BL
- HCC1954 / HCC1954 BL
- Available from ATCC, sequenced at Broad:
  - HCC1143 (50x)
  - HCC1143 BL (60x)
  - HCC1954 (58x)
  - HCC1954 BL (71x)
- All data distributed through CGHub

Benchmark 4: Modeling heterogeneity

- Three parts to mutation calling exercise:
- **HCC1143** (50x) **vs. HCC1143 BL** (60x)
- **HCC1954** (58x) **vs. HCC1954 BL** (71x)
- Simulate **normal contamination** and **subclone expansion** for both:



100 % Normal

(A)

VS.

95 % Normal
5 % Tumor

-15%
+15%

80 % Normal
20 % Tumor

-20%
+20%

60 % Normal
40 % Tumor

-20%
+20%

40 % Normal
60 % Tumor

-20%
+20%

20 % Normal
80 % Tumor

-15%
+15%

5 % Normal
95 % Tumor

(B)

VS.

25 % Normal
74 % Tumor
1 % Subclone

+4%
-4%

25 % Normal
70 % Tumor
5 % Subclone

+5%
-5%

25 % Normal
65 % Tumor
10 % Subclone

+10%
-10%

25 % Normal
55 % Tumor
20 % Subclone

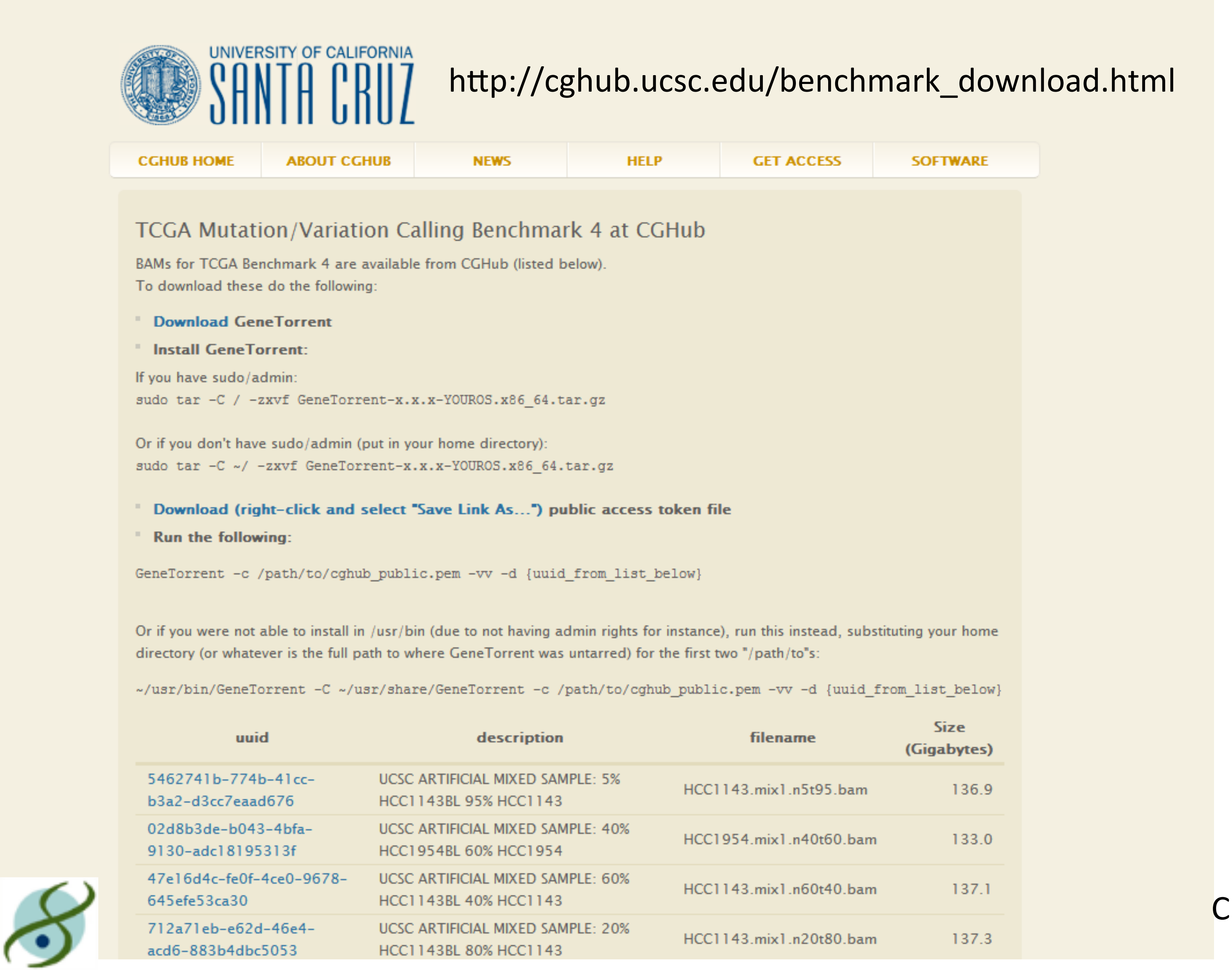
+20%
-20%

25 % Normal
35 % Tumor
40 % Subclone

(30x)

- Total: 28 .bam files, ~4.3 TB

Public .bams on CGHub



http://cghub.ucsc.edu/benchmark_download.html

TCGA Mutation/Variation Calling Benchmark 4 at CGHub

BAMs for TCGA Benchmark 4 are available from CGHub (listed below).
To download these do the following:

- **Download GeneTorrent**
- **Install GeneTorrent:**

If you have sudo/admin:

```
sudo tar -C / -zxvf GeneTorrent-x.x.x-YOUROS.x86_64.tar.gz
```

Or if you don't have sudo/admin (put in your home directory):

```
sudo tar -C ~/ -zxvf GeneTorrent-x.x.x-YOUROS.x86_64.tar.gz
```

- **Download (right-click and select "Save Link As...") public access token file**
- **Run the following:**

```
GeneTorrent -c /path/to/cghub_public.pem -vv -d {uuid_from_list_below}
```

Or if you were not able to install in /usr/bin (due to not having admin rights for instance), run this instead, substituting your home directory (or whatever is the full path to where GeneTorrent was untarred) for the first two "/path/to"s:

```
~/usr/bin/GeneTorrent -C ~/usr/share/GeneTorrent -c /path/to/cghub_public.pem -vv -d {uuid_from_list_below}
```

uuid	description	filename	Size (Gigabytes)
5462741b-774b-41cc-b3a2-d3cc7eaad676	UCSC ARTIFICIAL MIXED SAMPLE: 5% HCC1143BL 95% HCC1143	HCC1143.mix1.n5t95.bam	136.9
02d8b3de-b043-4bfa-9130-adc18195313f	UCSC ARTIFICIAL MIXED SAMPLE: 40% HCC1954BL 60% HCC1954	HCC1954.mix1.n40t60.bam	133.0
47e16d4c-fe0f-4ce0-9678-645efe53ca30	UCSC ARTIFICIAL MIXED SAMPLE: 60% HCC1143BL 40% HCC1143	HCC1143.mix1.n60t40.bam	137.1
712a71eb-e62d-46e4-acd6-883b4dbc5053	UCSC ARTIFICIAL MIXED SAMPLE: 20% HCC1143BL 80% HCC1143	HCC1143.mix1.n20t80.bam	137.3

Chris Wilks

New evaluation tools for VCF

- VCF is a successful standard
  - Existing VCF tools: e.g. VCFtools, GATK, PyVCF, etc.

Benchmark 4 is stimulating the creation of new tools:

- Bamsurgeon
- VCFcomparator
- LeftShiftBreakends
- VCF to MAF converter (Thanks to Sage!)

How to participate

(and/or get more information)

- **Everyone is welcome!!**
- Sign up at **poster 64**
- E-mail ewingad@soe.ucsc.edu

Mailing list: tcga-mutation@soe.ucsc.edu
(contact Chris Wilks: cwilks@soe.ucsc.edu)

Thanks!

David Haussler (UCSC/CGHub)

Singer Ma (UCSC)

Chris Wilks (UCSC/CGHub)

Mark Diekhans (UCSC/CGHub)

Su Yeon Kim (UC Berkley)

Gaddy Getz (Broad Institute)

Scott Carter (Broad Institute)

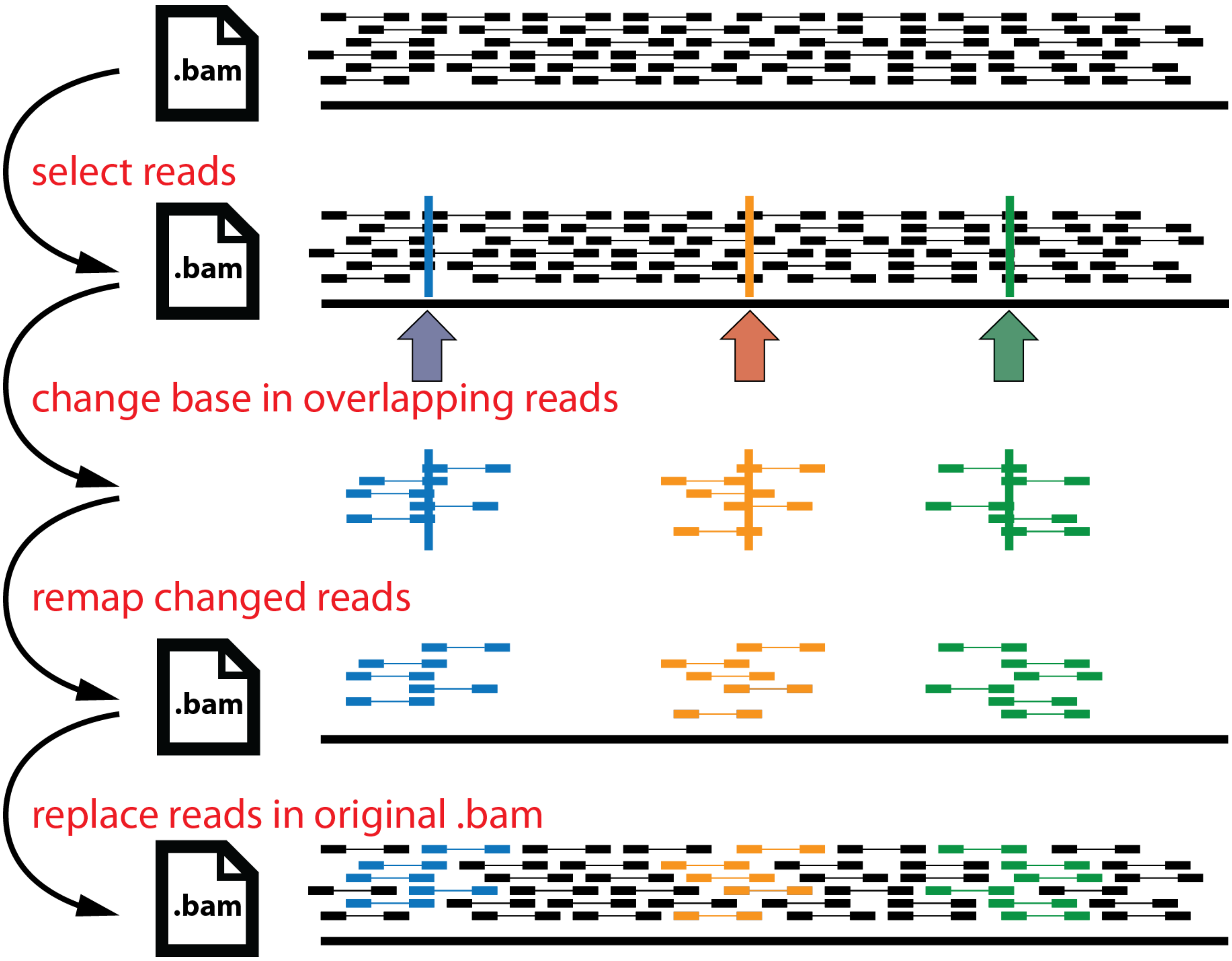
Andrey Sivachenko (Broad Institute)

Mara Rosenberg (Broad Institute)

UCSC Cancer Genomics Hu (CGHub)

UCSC Reconstruction & Cancer groups

How? (SNVs)



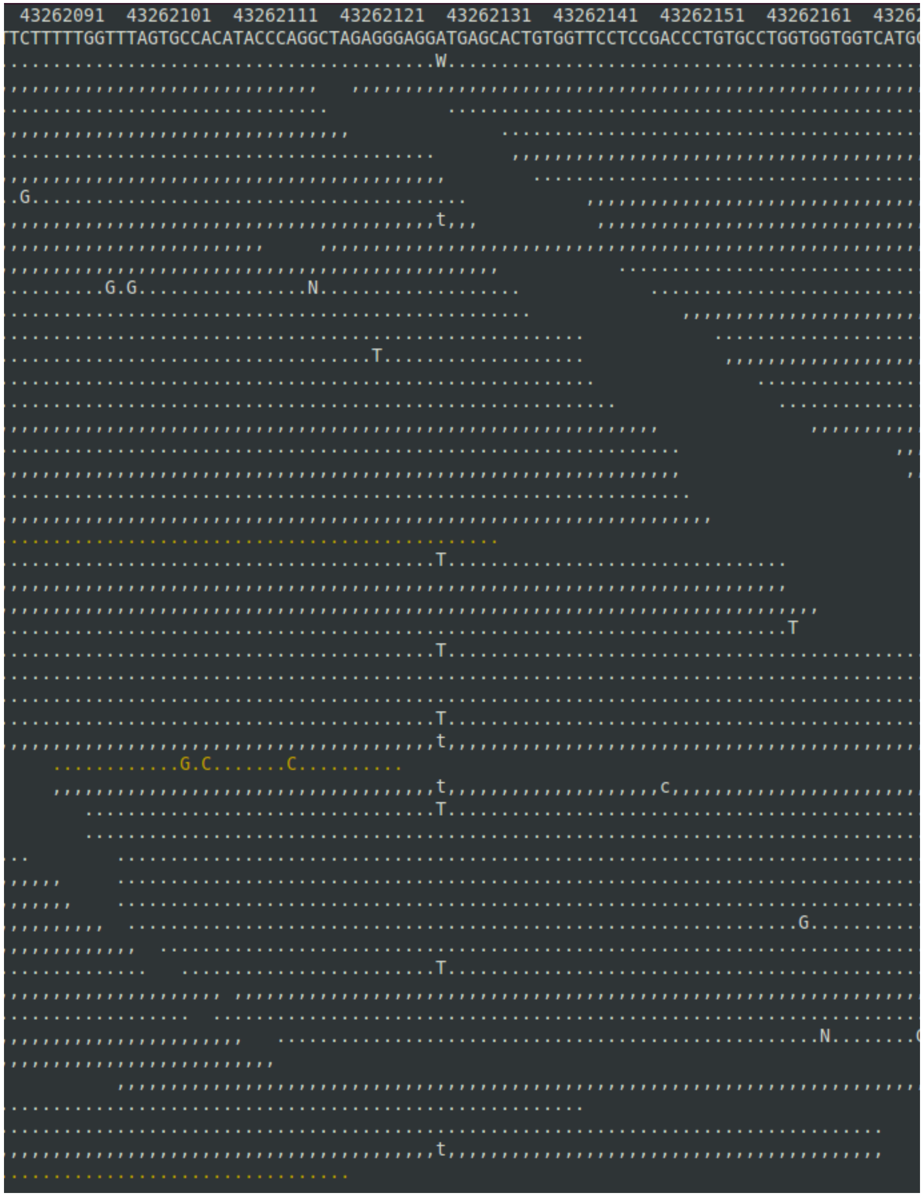
select reads

change base in overlapping reads

remap changed reads

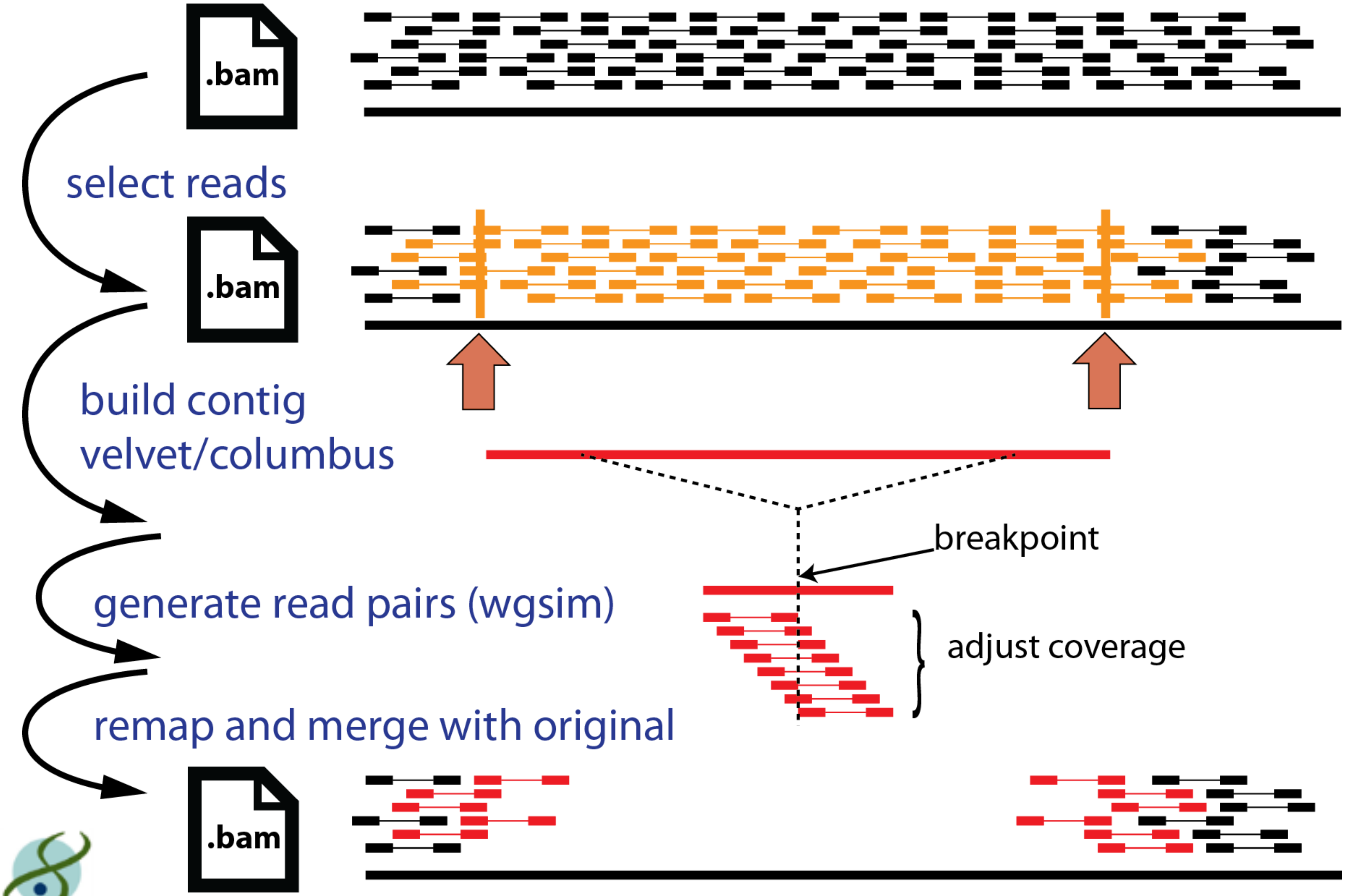
replace reads in original .bam

Good call (20% alt.)



```
 43262091  43262101  43262111  43262121  43262131  43262141  43262151  43262161  4326
TTCTTTTTGGTTTAGTGCCACATACCCAGGCTAGAGGGAGGATGAGCACTGTGGTTCCTCCGACCCTGTGCCTGGTGGTGGTCATG
```

How? (SVs)



select reads

build contig
velvet/columbus

generate read pairs (wgsim)

remap and merge with original

Examples: Deletion (50% MAF)