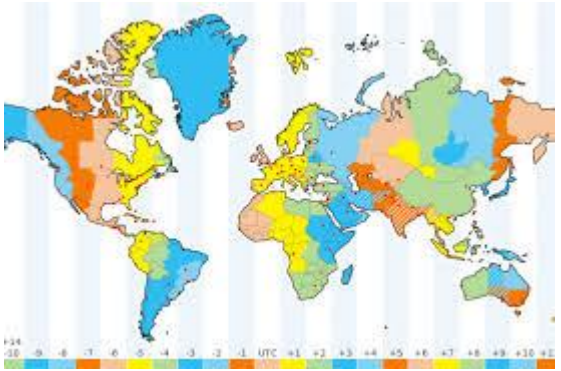
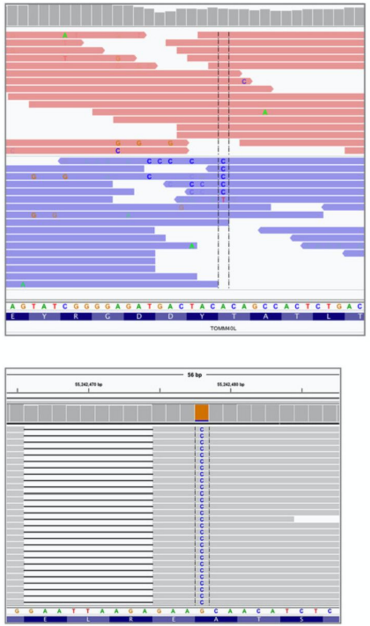
# SV filtering

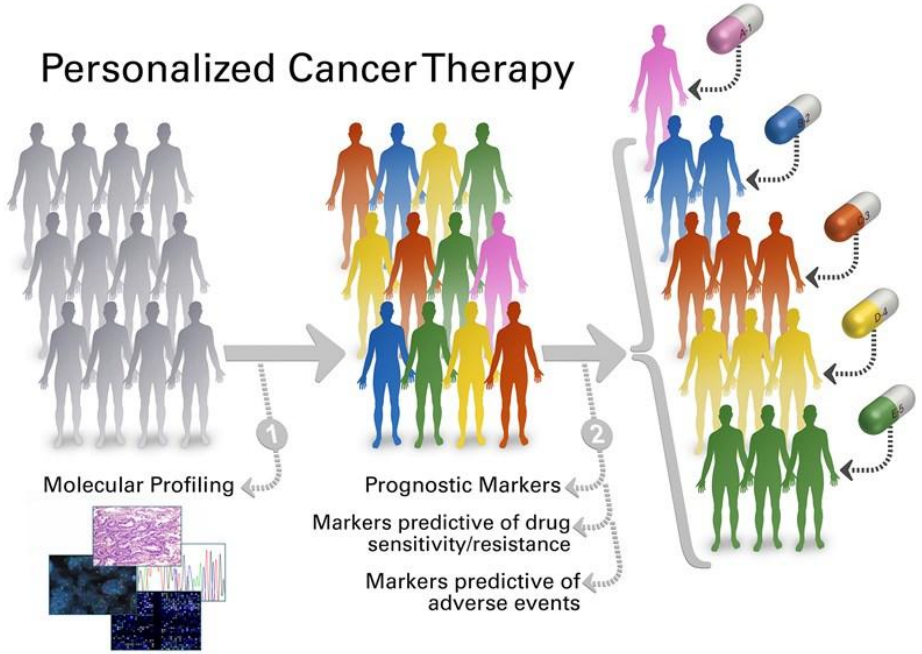
Evan Biederstedt
Daniel Cameron
Ann McCartney
Jim Havrilla
Divya Kalra
Michael Khayat
Jingwen Ren
Najeeb Syed
Angad Jolly



# Motivation

- Clinical genomics relies on short-read data, requiring extensive manual review (IGV)
- Problem: short-read SV calling has both high FPs and FNs, for somatic SVs (Sedlazeck et al, 2018)






https://cancerres.aacrjournals.org/content/77/21/e31

# Current strategies don’t work!

- Ignoring the problem **DOESN’T WORK**
  - SVs compose greater source of genomic diversity compared to SNPs!
    Chaisson et al, Nature volume 517, pages608–611(2015)
- “Ensemble approach” **DOESN’T WORK**
  - Cameron et al, Nat Commun. 2019; 10: 3240.

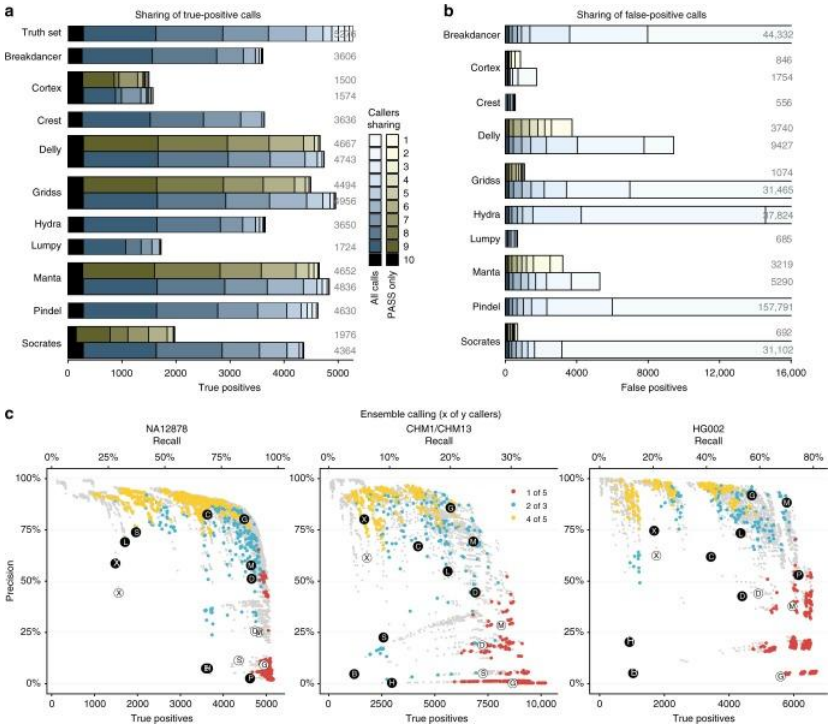



Figure 3: Simple ensemble-based calling **does not reliably improve** performance.

a) Agreements for SV callers in NA12878

b) Agreement between callers for false positives

# Current strategies don’t work!

- Crying **DOESN’T WORK (though it’s cathartic)**

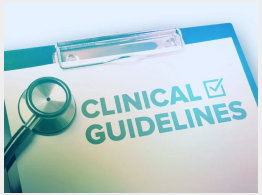
**^ sad Fritz :(**

# GOALS

## 1

## GUIDELINES

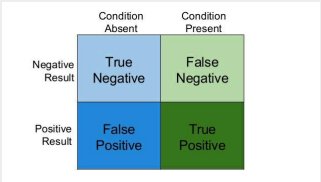
Supply guidelines for the community to aid SV assessment and filtration process




## 2

## FALSE POSITIVE EXPLORATION

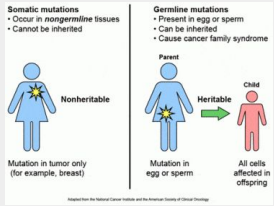
Determine the false positive rate of short-read SV callers (with benchmarks and long-read data)

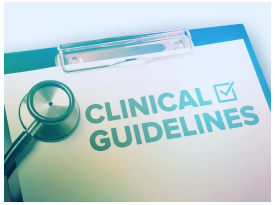



## 3

## GERMLINE FILTRATION

Generate suitable parameters for germline vs somatic filtration






# 1) GUIDELINES

Supply guidelines for the community to aid SV assessment and filtration process

**VISUALISATION**

- IGV & Samplot

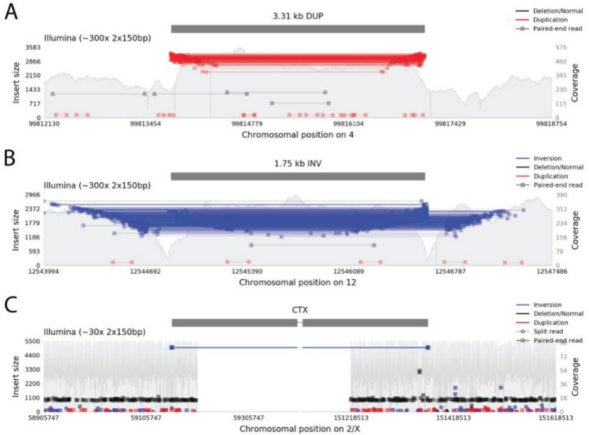



**Figure 2. Samplot images of duplication, inversion, and translocation variants**. **A)** A duplication variant plotted by Samplot with Illumina short-read sequencing evidence. Reads plotted in red have large insert sizes and inverted pair order (reverse strand followed by forward strand instead of forward followed by reverse), indicating potential support for a duplication. **B)** An inversion variant, with Illumina sequencing evidence. Reads plotted in blue have large insert sizes and same-direction pair alignments (both reads on forward strand, or both on reverse strand). **C)** A translocation variant, with Illumina sequencing. Discordant pairs align to each breakpoint. The blue color of the reads and extremely large insert sizes of these grouped discordant pairs indicate a large inverted translocation.

**FEATURES based on FPs**

- Investigate features associated with FPs by specific SV caller, e.g. mappability, repeat regions, read depth
- Intuition to write an algorithm to filter out *obvious* FPs.

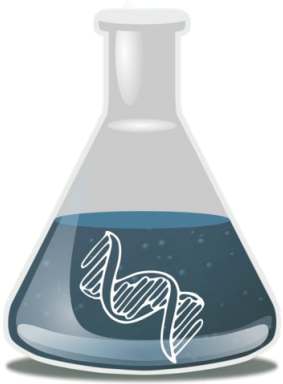
**<u>MANUAL CURATION</u>**

# 2) FALSE POSITIVES

Determine the false positive rate of short-read SV callers (with benchmarks and long-read data)

| | Condition Absent | Condition Present |
|---|---|---|
| Negative Result | True Negative | False Negative |
| Positive Result | False Positive | True Positive |

**Specific short-read SV callers have very specific behavior**

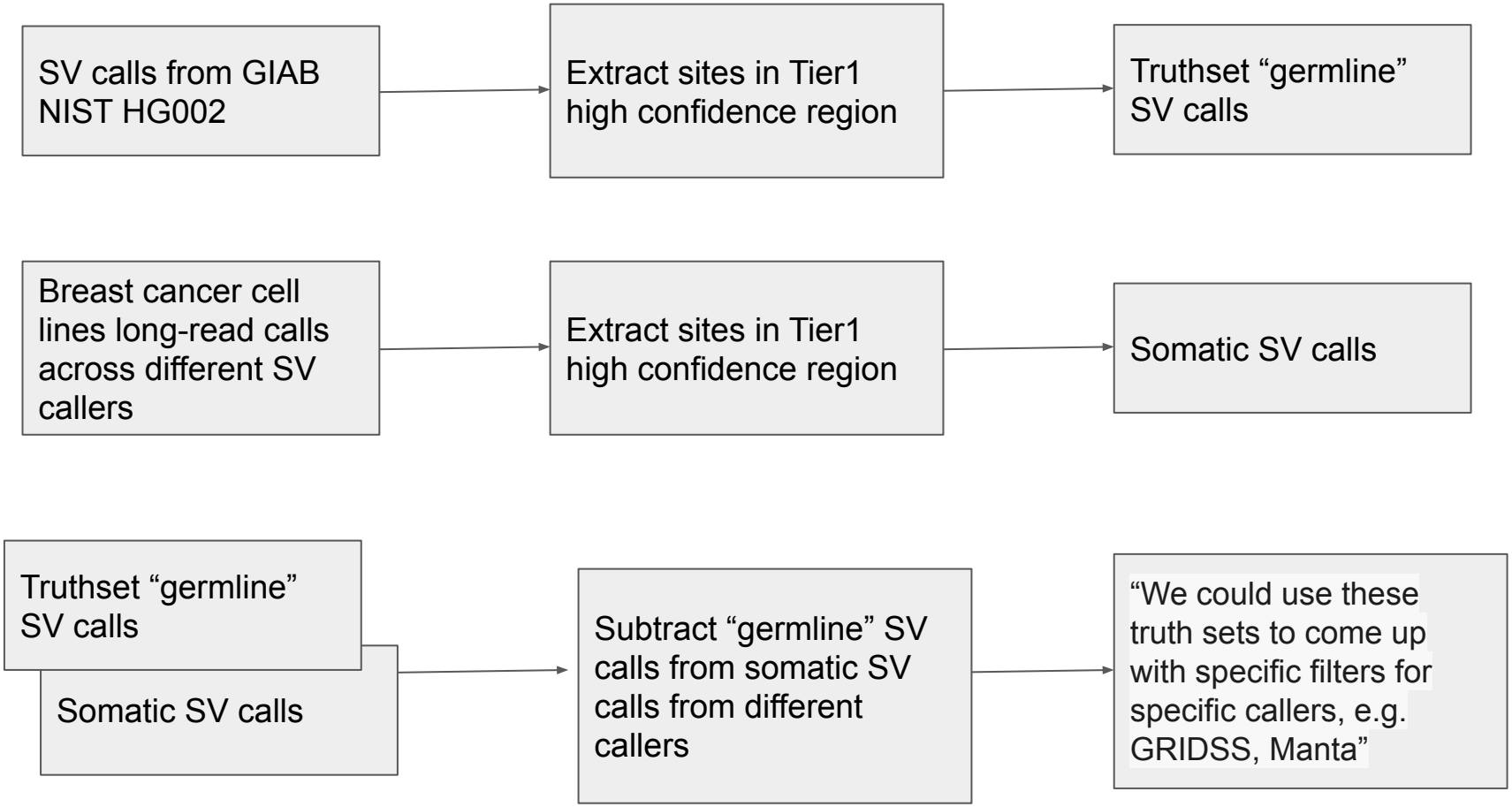


**Using HG002 & COLO829**

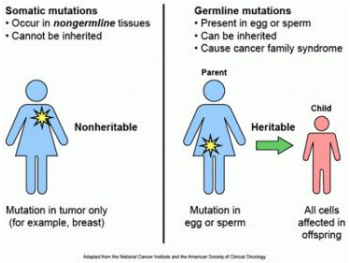
- **Both HG002 and COLO829 have truth sets!**
- Manually curate FPs & FNs with samplot
- Investigate features associated with FPs by specific SV caller, e.g. mappability, repeat regions, read depth
- Intuition to write an algorithm to filter out *obvious* FPs.

**Also:** Use targeted GRIDSS
VCF -> +/- some regions -> pulls all extra reads and do assembly to check accuracy of breakpoint

# WORKFLOW


SV calls from GIAB NIST HG002
Extract sites in Tier1 high confidence region
Truthset “germline” SV calls
Breast cancer cell lines long-read calls across different SV callers
Extract sites in Tier1 high confidence region
Somatic SV calls
Truthset “germline” SV calls
Somatic SV calls
Subtract “germline” SV calls from somatic SV calls from different callers
“We could use these truth sets to come up with specific filters for specific callers, e.g. GRIDSS, Manta”

# 3) GERMLINE FILTRATION




**FILTRATION GUIDELINES TO REDUCE FPs FROM GERMLINE**

FEATURES:
-TUMOUR VS NORMAL READ DEPTH
-FRAGMENT SIZE
-MAPPABILITY
-REPEAT UNITS
-ALIGNMENT STATISTICS