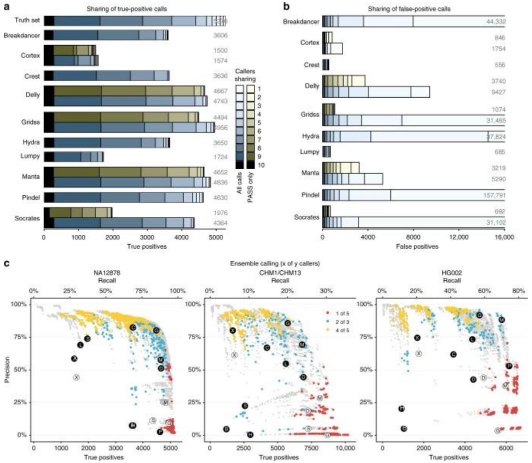
CleanSV

Daniel Cameron, WEHI
Ann McCartney, NIH
Jim Havrilla, CHOP
Divya Kalra, Baylor College of Medicine
Michael Khayat, Baylor College of Medicine
Jingwen Ren, University of Southern California
Najeeb Syed, Sidra Medicine
Evan Biederstedt, HMS
Angad Jolly, Baylor College of Medicine

What is the problem?



1) Clinical bioinformaticians *need* high quality calls for patient care!

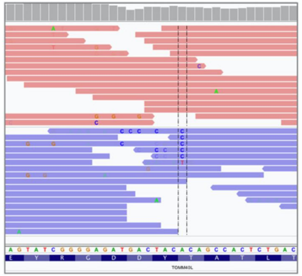
    e.g. gene fusions are common (successful!) targets for cancer therapies

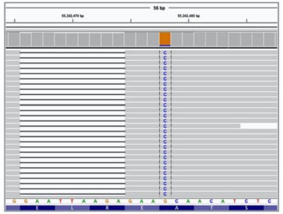
2) However, ***High*** FP + FN rates for short-read SV callers

3) Ensemble methods **do not work!**

(Cameron et al, Nat Commun. 2019; 10: 3240)

4) Therefore, the community has a pressing need for quality filters to remove FPs, especially for somatic SVs. Otherwise, SV calling will remain dismissed as “unreliable” for clinical care in precision oncology.





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

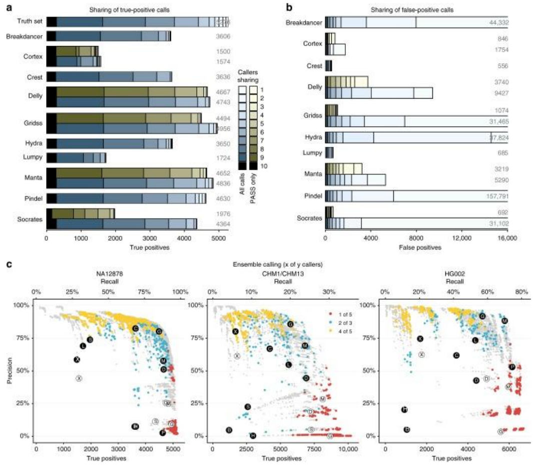
What is the problem?

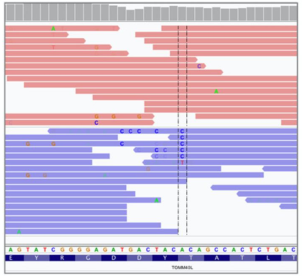
1) Clinical bioinformaticians *need* high quality calls for patient care!

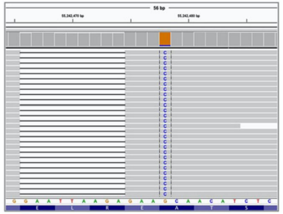
2) However, ***High*** FP + FN rates for short-read SV callers

3) Ensemble methods **do not work!**

(Cameron et al, Nat Commun. 2019; 10: 3240)







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

Sad Fritz

Feeling ***AWESOME*** Fritz

GOALS

1

GENERATE FP+FN against benchmarks

HG002 (germline with GIAB Truthset)

COLO829 (germline+somatic with Truthset generated from Jose Espejo Valle-Inclan. (2020))

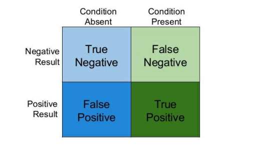
DELLY

MANTA

GRIDSS

2

ANNOTATE FP+FNs



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

GRCh37 GENERATED TRACKS:

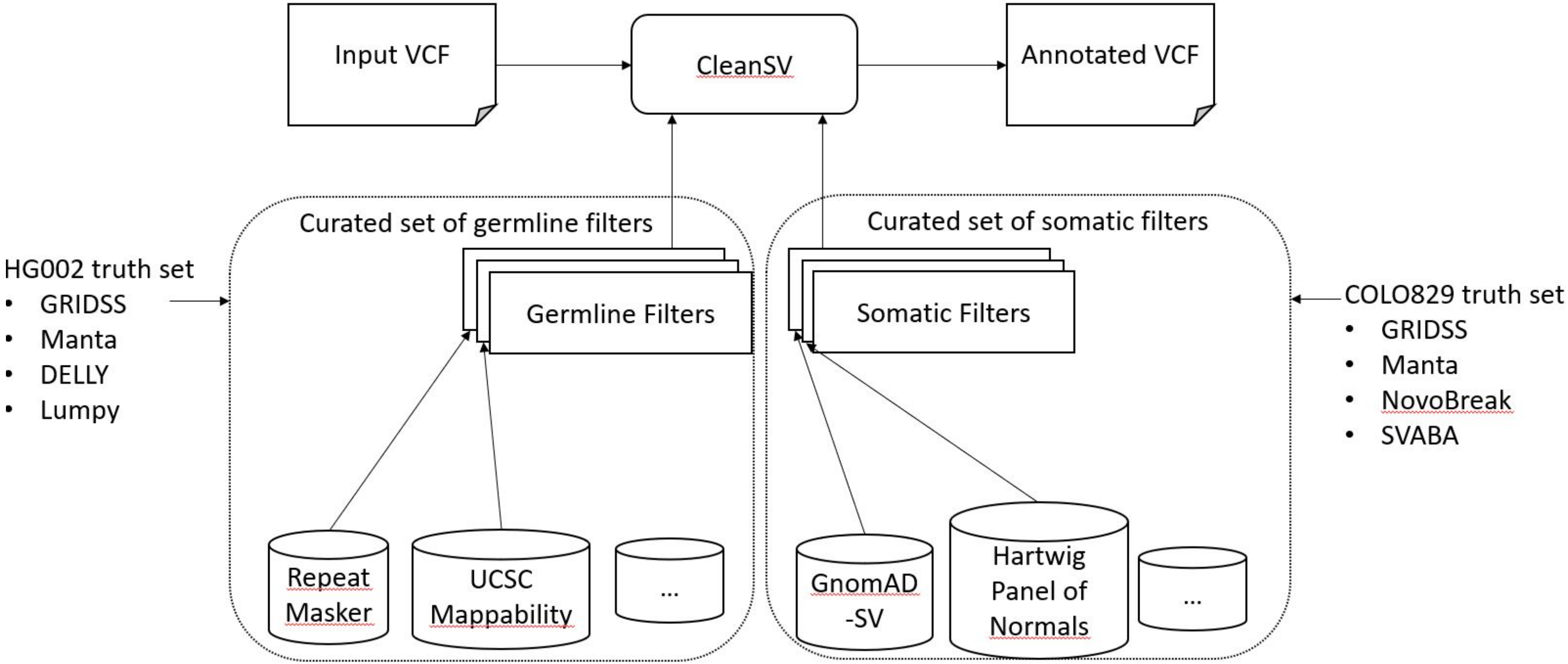
REPEAT/DUSTMASKER
GC CONTENT
MAPPABILITY
MITO CONTAMINATION
SEGMENTAL DUPLICATIONS
SIMPLE REPEATS
MICROSATELLITES

3

CLEAN VCFs USING CALLER SPECIFIC FILTERS

CleanSV

WORKFLOW



Input VCF

CleanSV

Annotated VCF

HG002 truth set

- GRIDSS
- Manta
- DELLY
- Lumpy

Curated set of germline filters

Germline Filters

Repeat Masker

UCSC Mappability

...

Curated set of somatic filters

Somatic Filters

GnomAD -SV

Hartwig Panel of Normals

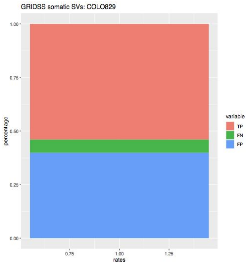
...

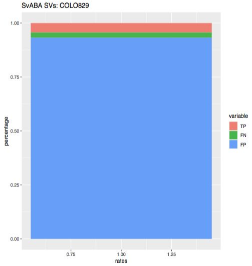
COLO829 truth set

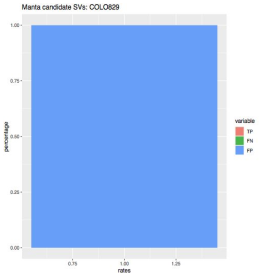
- GRIDSS
- Manta
- NovoBreak
- SVABA

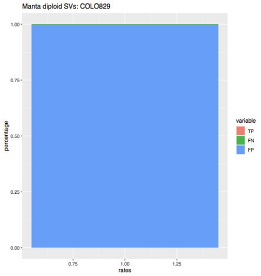
GOAL 1: SV CALLER FP+ FN, COLO829

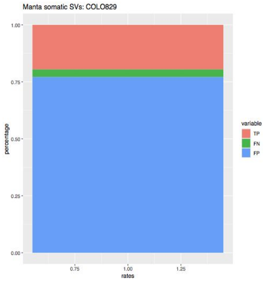
PERCENTAGES

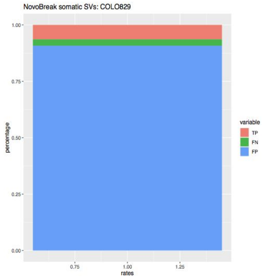












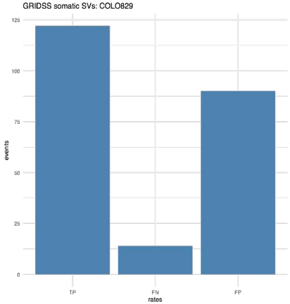
variable

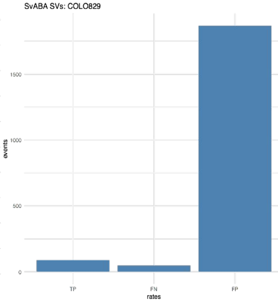
🟥 TP

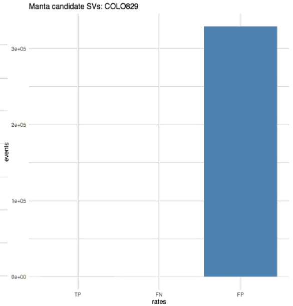
🟩 FN

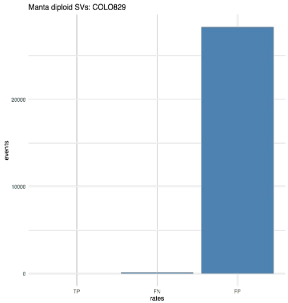
🟦 FP

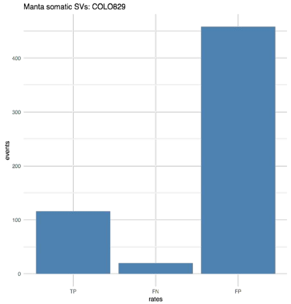
RATES

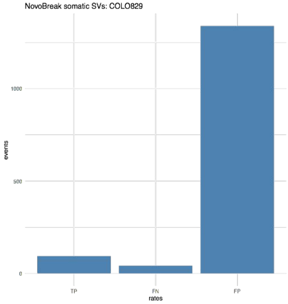








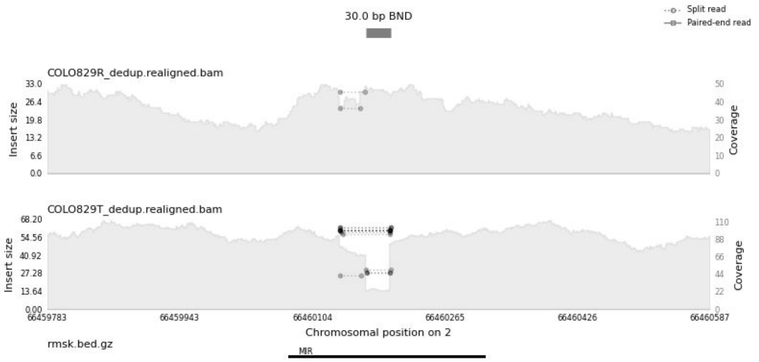


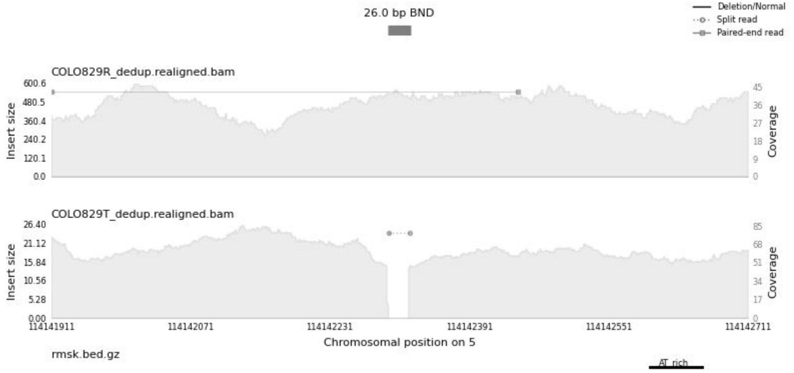


GRIDSS SvABA (unfiltered) Manta Manta Novobreak

GOAL 2: ANNOTATIONS OF FP + FN

Sample GRIDDSS FP annotations on COLO829



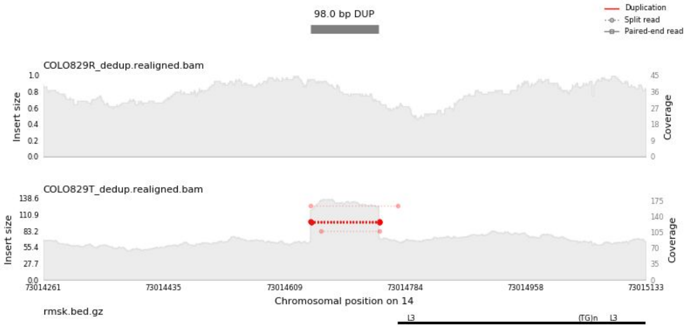


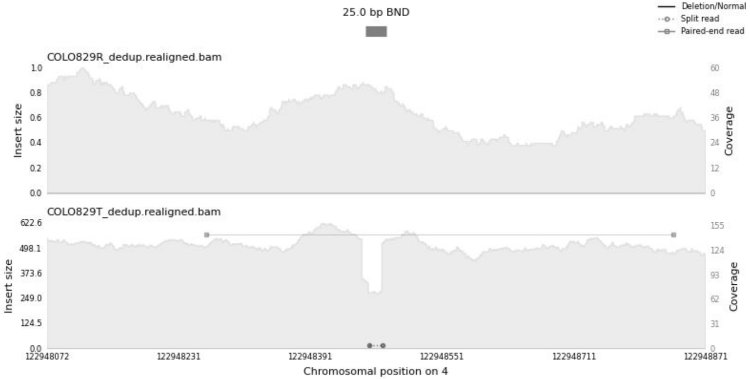
REAL DELETION IN TUMOR (read evidence, coverage)

REAL DELETION IN TUMOR (coverage dip, weak read evidence)

GOAL 2: ANNOTATIONS OF FP + FN

Sample MANTA FP annotations on COLO829



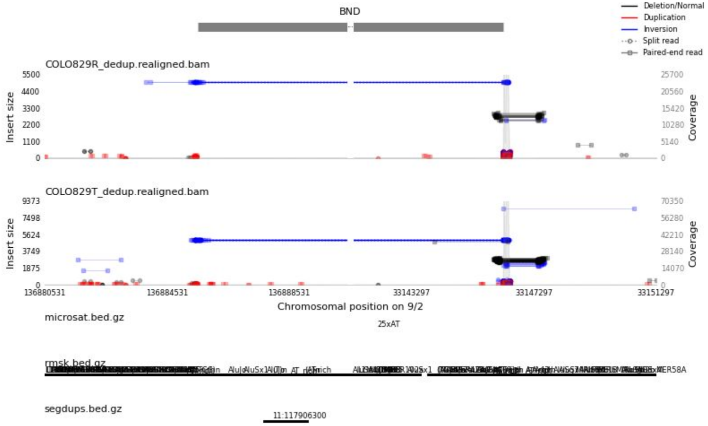


REAL DUPLICATION IN TUMOR (coverage, read evidence)

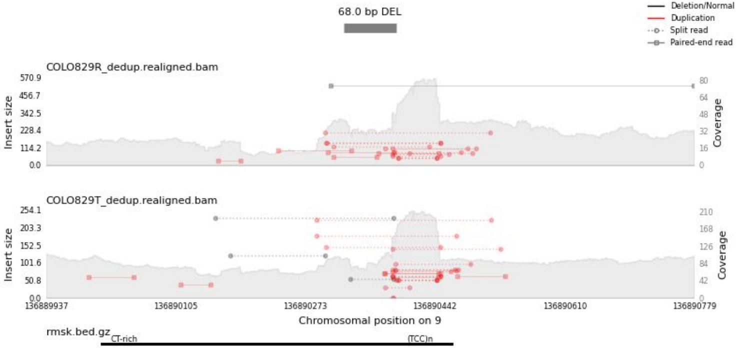
REAL DELETION IN TUMOR (coverage dip, weak read evidence)

GOAL 2: ANNOTATIONS OF FP + FN

Sample DELLY FP annotations on COLO829



INVERTED TRANSLOCATION IN TUMOR AND NORMAL (insert size evidence in both)

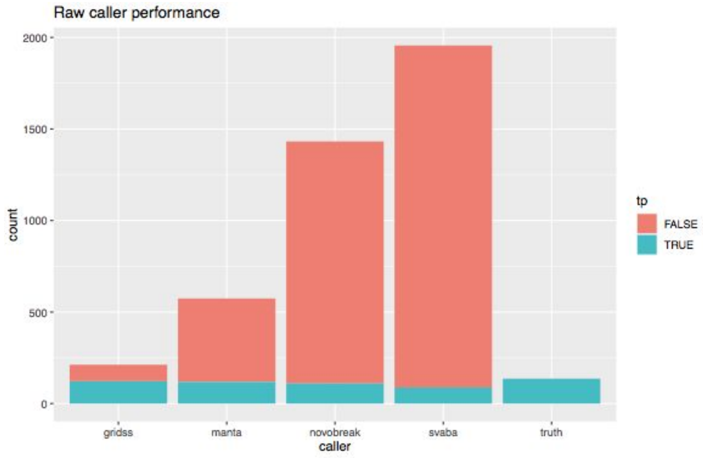


REAL DUPLICATION IN TUMOR AND NORMAL (coverage, read evidence)

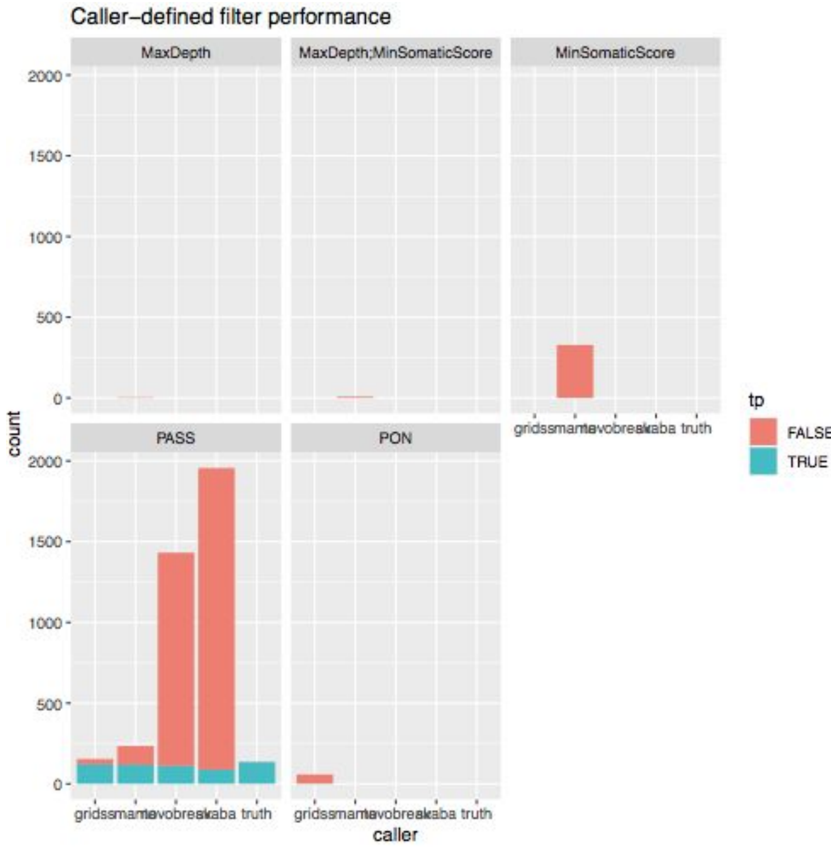
GOAL 3: CleanSV

Construction of filters for somatic callers

COLO829



TP = green
FP = red



GOAL 3: CleanSV

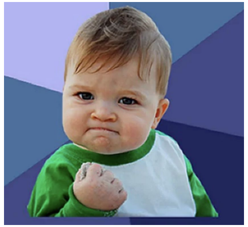
Hartwig MEDICAL FOUNDATION

Construction of filters for somatic callers

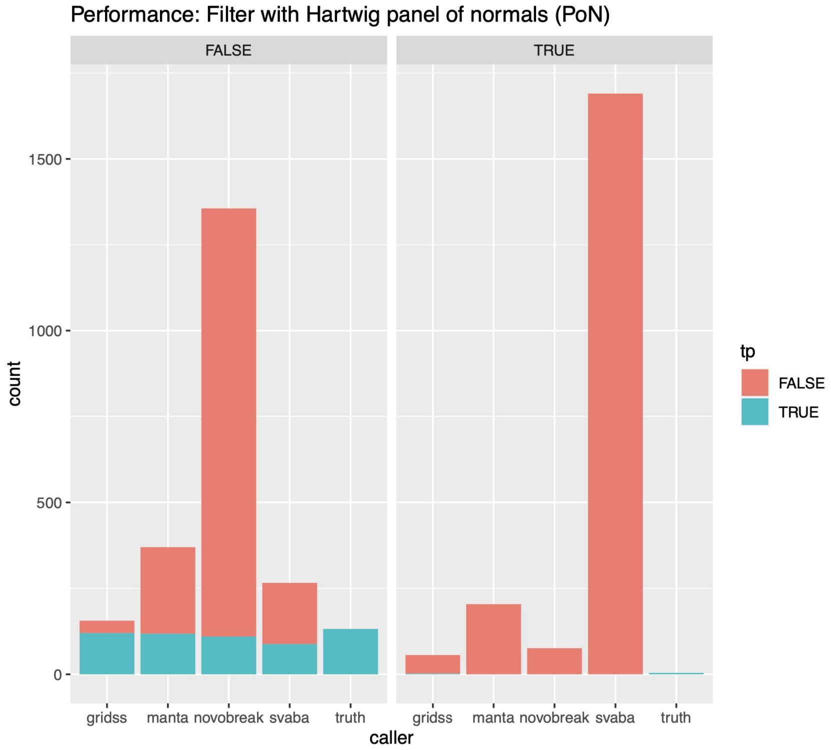
COLO829 Truth Set

Panel of Normals created with WGS Hartwig Foundation freshly sequenced

First pass shows the number of germline SVs we are able to filter from the somatic SV dataset!



Success!


Performance: Filter with Hartwig panel of normals (PoN)

FUTURE

CleanSV

- Continue to revise filters by specific SV caller and release
  - Include more features such as read depth, tumor purity/ploidy, better PoN
- Explicitly write out guidelines for researchers to do manual FP curation
- Apply to current large consortia datasets!