Benchmarking Somatic Variant Caller: MuTect and GATK Mutect2 on TCGA WES

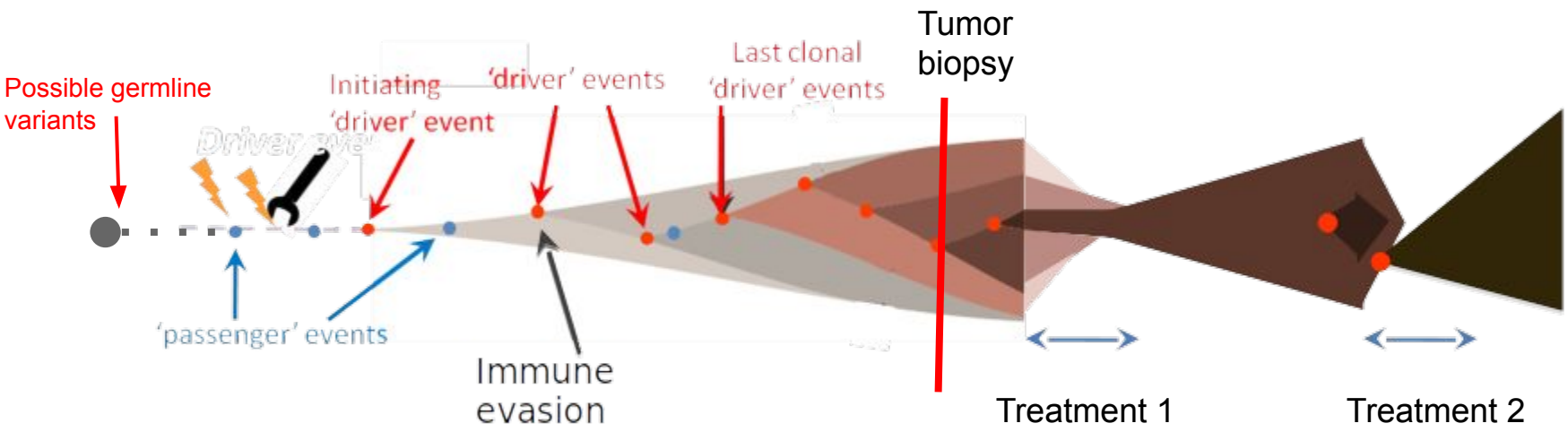
Call set comparison & Impact on driver gene discovery

Qing Zhang
MS candidate at Harvard
Supervisor: Dr Gad Getz

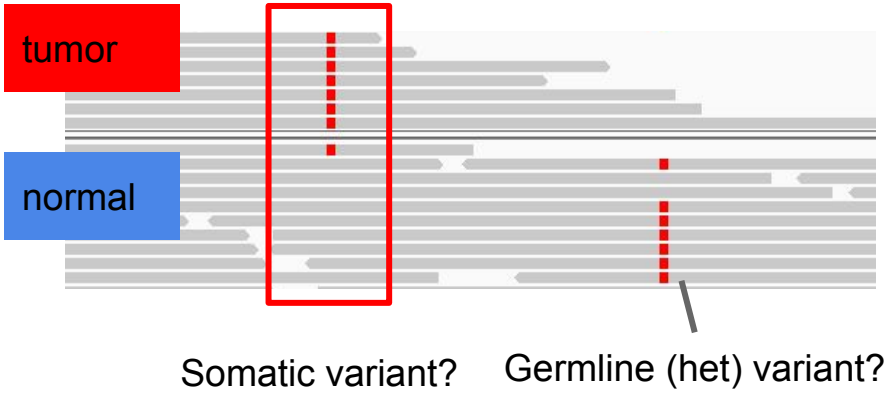
Outline

1. **Background**
   a. Why study mutations?
   b. Somatic variant caller: Mutect1 and GATK Mutect2
2. **Workflow**
3. **Results**
   a. Call set comparison
   b. Driver gene discovery
   c. Examples from manual review
4. **Conclusion**
5. **Perspective**
6. **Acknowledgement**

Identifying somatic mutations is crucial for cancer genomics



Somatic callers are algorithms to identify somatic mutations.



Challenges

- Allele fraction cannot be assumed from ploidy.
- Differentiate germline from somatic
- Low signal-to-noise ratio: sequencing error, mapping artifact, contamination

Campbell et al. Nature 467(7319), 2010

MuTect and GATK Mutect2

Mutect1 (**M1**)

Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples

Kristian Cibulskis[1], Michael S Lawrence[1], Scott L Carter[1], Andrey Sivachenko[1], David Jaffe[1], Carrie Sougnez[1], Stacey Gabriel[1], Matthew Meyerson[1,2], Eric S Lander[1,3,4] & Gad Getz[1,5]

\+ Haplotype assembly
\+ Bayesian somatic model

GATK Mutect2 (**M2**)

METHODS

Calling Somatic SNVs and Indels with Mutect2

David Benjamin*, Takuto Sato, Kristian Cibulskis, Gad Getz, Chip Stewart and Lee Lichtenstein

Benchmark included in this paper:

1. Synthetic truth set with BamSurgeon (DREAM challenge)
2. Normal mixtures to mimic tumor subclones
3. 60 TCGA Whole-Genome-Sequencing samples (WGS)

We would like test with **real data** in more **tumor types**.

MuTect1: Cibulskis et al. Nat. Biotechnol. (2013)
GATK Mutect2: Benjamin et al, 2019 preprint

Evaluating call sets

Good call sets:

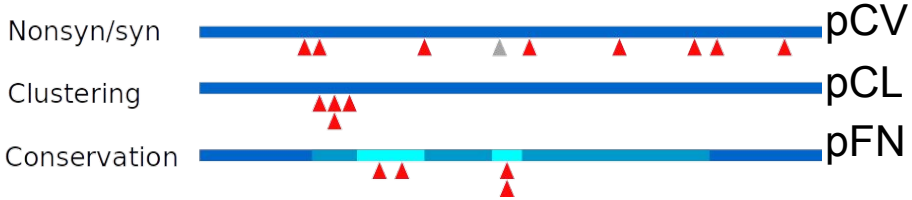
- Rediscover known drivers
- Produce mutation signatures that reflect known biological processes
- Model tumor phylogeny (inferred via clustering cancer cell fraction)
- Produce consensus tumor subtype

LETTER

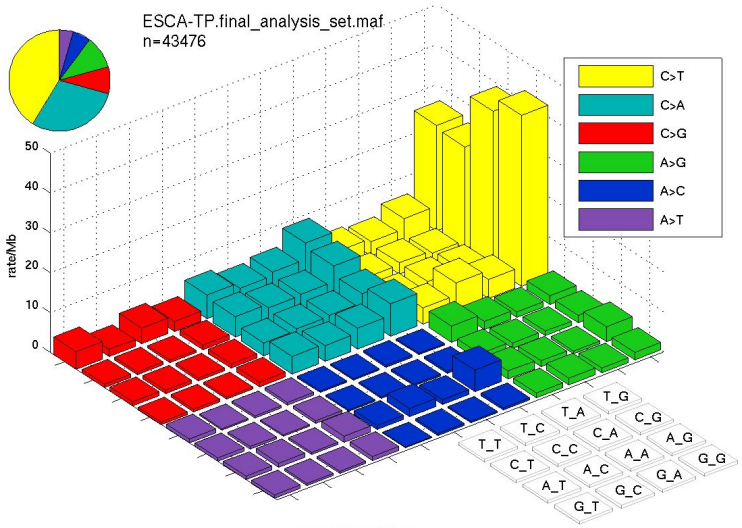
doi:10.1038/nature12213

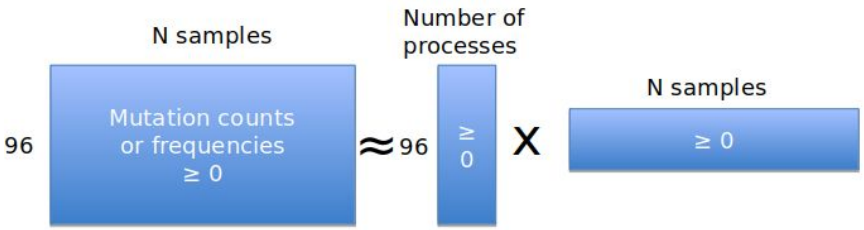
Mutational heterogeneity in cancer and the search for new cancer-associated genes

Three types of scores (MutSig suite)

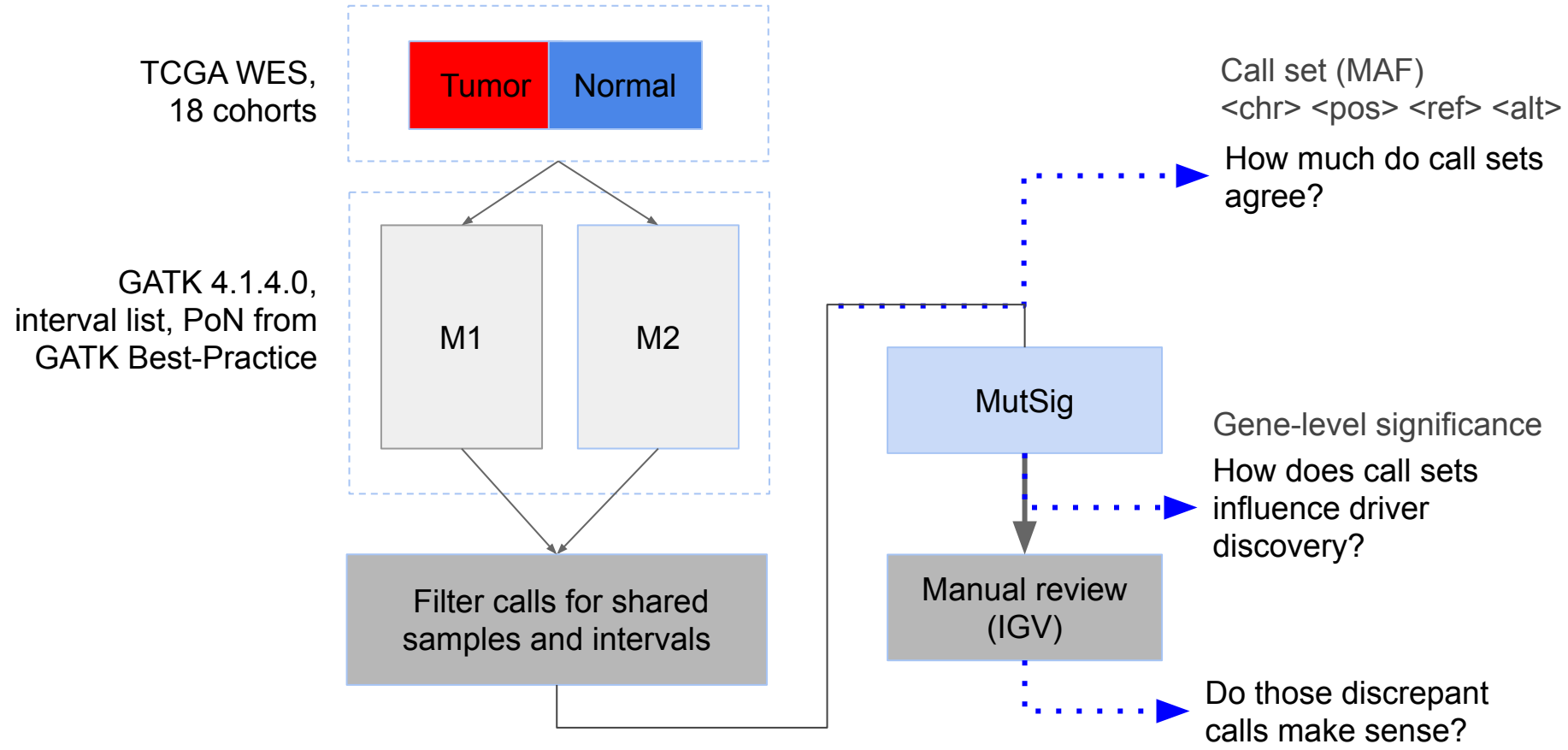


MutSig builds background model for gene-level significance, has been used to reduce false positives in driver discovery.





Workflow

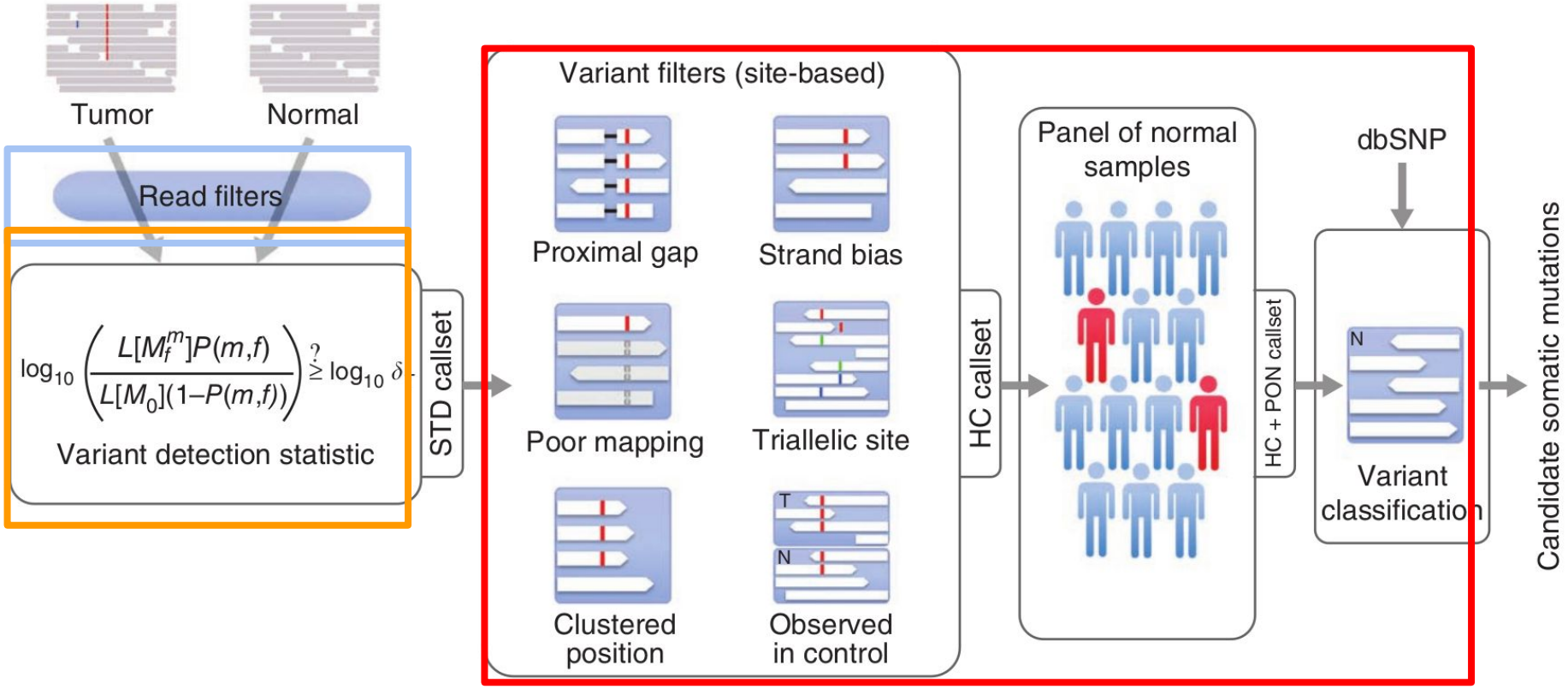


Snakemake pipeline to reproduce M2 calling is available at this github repo,
*Addition of `independent-mate` is based on this bug report

M1 logic

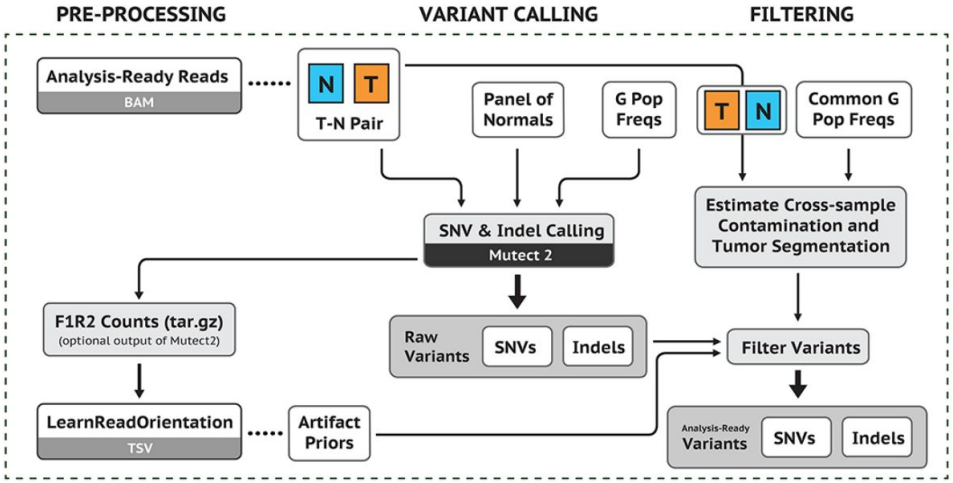
> Prefilter
> Somatic likelihood model
> Postfilter for artifacts

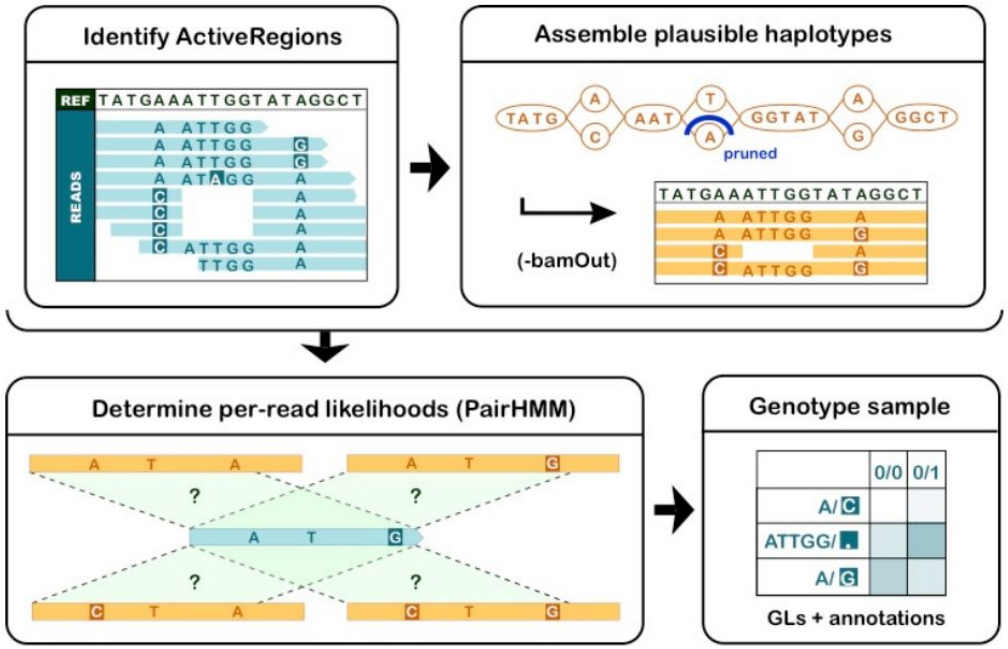
Panel-of-Normal(PoN)
Call mutation from normal as if they are tumor. It models recurrent background noise, eliminates germline calls.

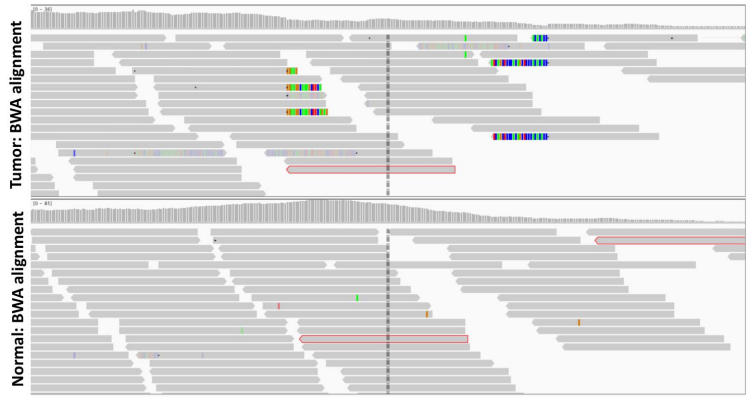


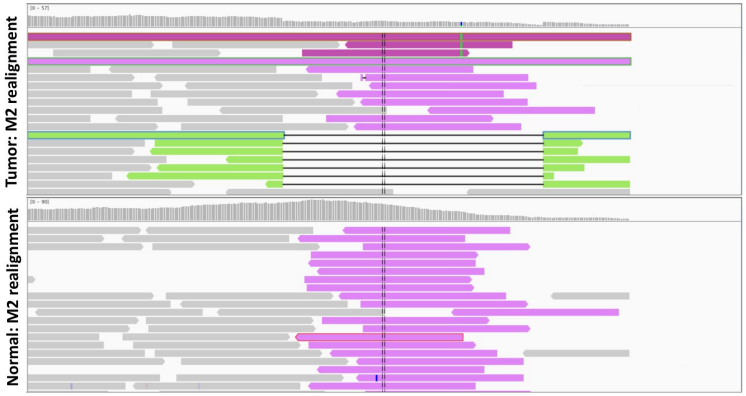
M2 logic

Prefilter >
Realignment >
Somatic model >
PostFiltering









Return: P(read r| allele a)

From GATK training slides

Somatic calling model

aims to differentiate real variant from sequencing error

Both produce the likelihood ratio of two models - alt exists at allele fraction f, or only ref exists. However, M2 models the allele set A, which accounts for multiallelic events.

M1

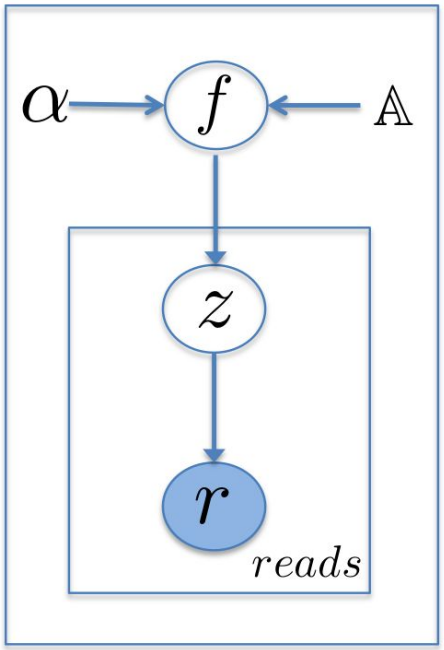
$$L(M_f^m) = P(\{b_i\} \mid \{e_i\}, r, m, f) = \prod_{i=1}^{d} P(b_i \mid e_i, r, m, f)$$

$$P(b_i|e_i,r,m,f) = \begin{cases} f\frac{e_i}{3} + (1-f)(1-e_i) & \text{if } b_i = r \\ f(1-e_i) + (1-f)\frac{e_i}{3} & \text{if } b_i = m \\ \frac{e_i}{3} & \text{otherwise} \end{cases}$$

$$LOD_T(m,f) = \log_{10}\left(\frac{L(M_f^m)P(m,f)}{L(M_0)(1-P(m,f))}\right) \geq \log_{10}\delta_T$$

$$LOD_T(m,f) = \log_{10}\left(\frac{L(M_f^m)}{L(M_0)}\right) \geq \log_{10}\delta - \log_{10}\left(\frac{P(m,f)}{(1-P(m,f))}\right) = \theta_T$$

M2



$$f \sim \text{Dirichlet}(\alpha)$$

$$z|f \sim \text{Categorical}(f)$$

$$p(r|z_{ra}) = l_{ra}$$

$$\ell_{ra} \equiv P(\text{read } r|\text{allele } a)$$

from PairHMM

$$\log\frac{p(\mathbb{R}|\mathbb{A}_{alt})}{p(\mathbb{R}|\mathbb{A}_{ref})} > \delta = 3.0$$

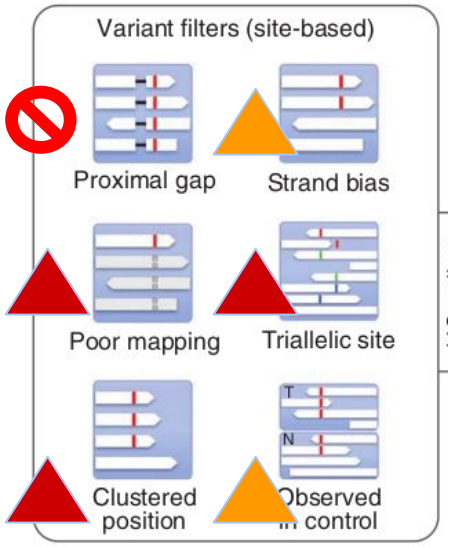
then emit variant

Cibulskis et al. Nat. Biotechnol (2013)

Postfilters: Hard filters vs Probabilistic filters

Hard filters (remove calls by a preset cutoff)

Probabilistic filters, models P(error) and final threshold is determined by F score



M1 filters are all **hard filters** with heuristic cutoffs.

Example: “proximal gap” filter

Remove false positives caused by nearby misaligned small insertion and deletion events. Reject candidate site if there are ≥ 3 reads with insertions in an 11-base-pair window centered on the candidate mutation or if there are ≥ 3 reads with deletions in the same 11-base-pair window.

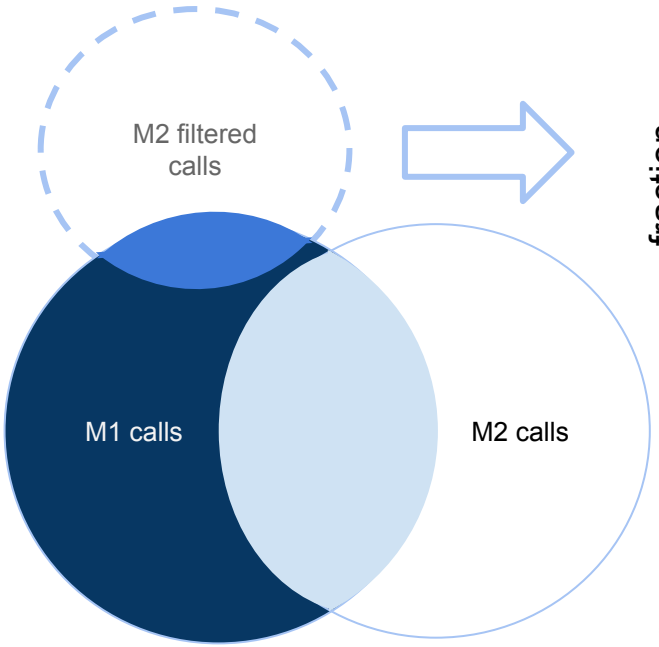
M2 implements an additional filter: “Clustered_events”: limit the number of mutations sharing an assembly region

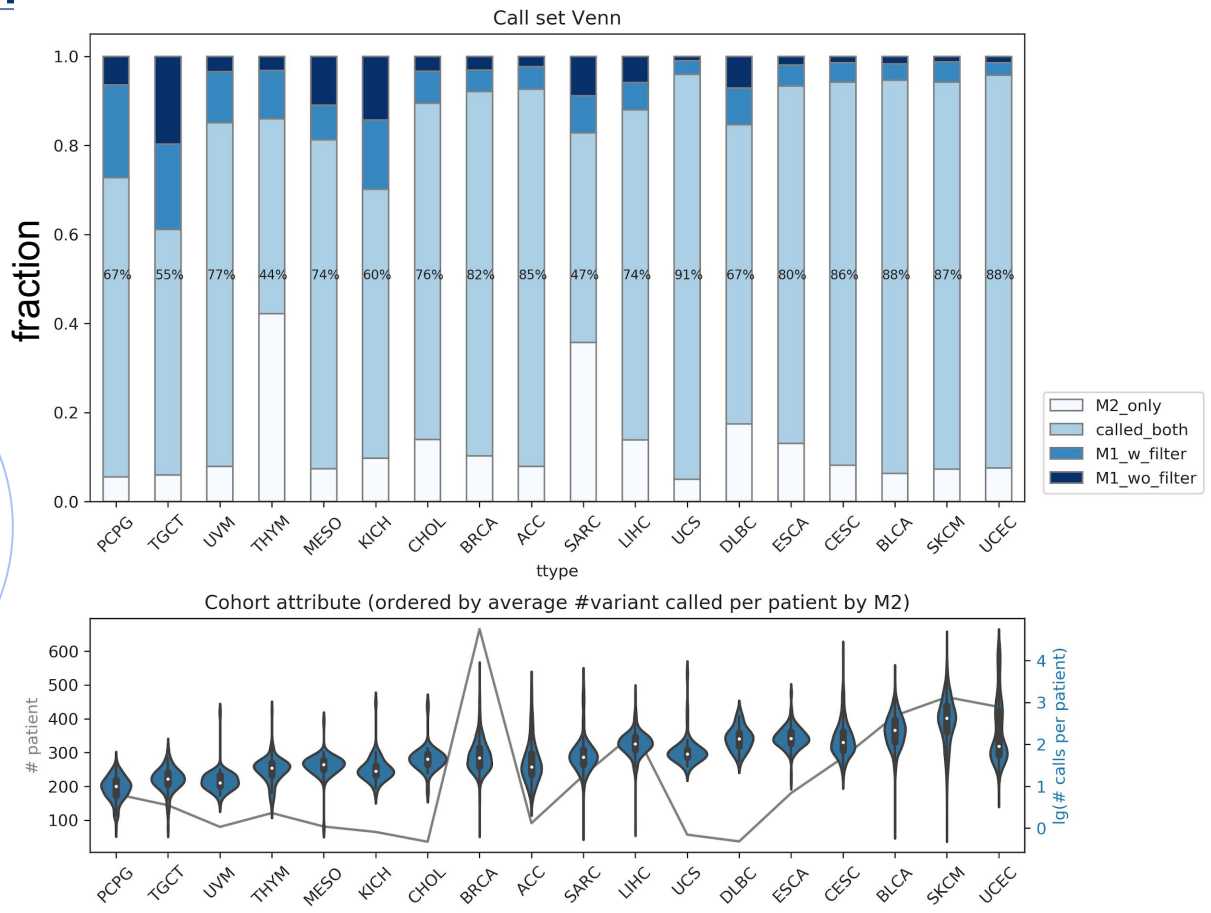
M2 has many probabilistic filter: <germline> <strand-bias> <orientiational bias> etc based on AF-clustering model

Result

Call set comparison

Call set comparison

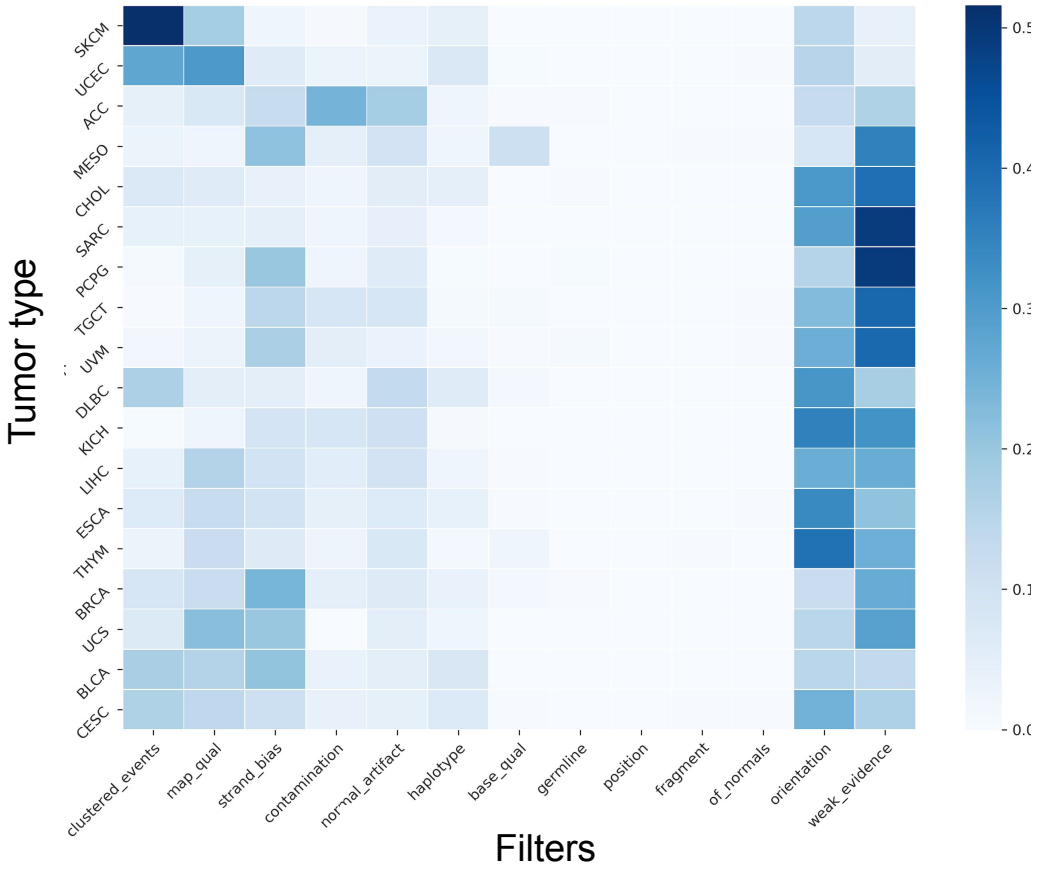




Call set overlaps most in UCS (91%), least in THYM(45%) and SARC(47%)
For most cohorts, more than 50% of M1-only calls have at least one M2 filtering reason, meaning that they have passed M2's somatic calling model (but not its postfiltering).

M2 filtering reason for M1-only calls

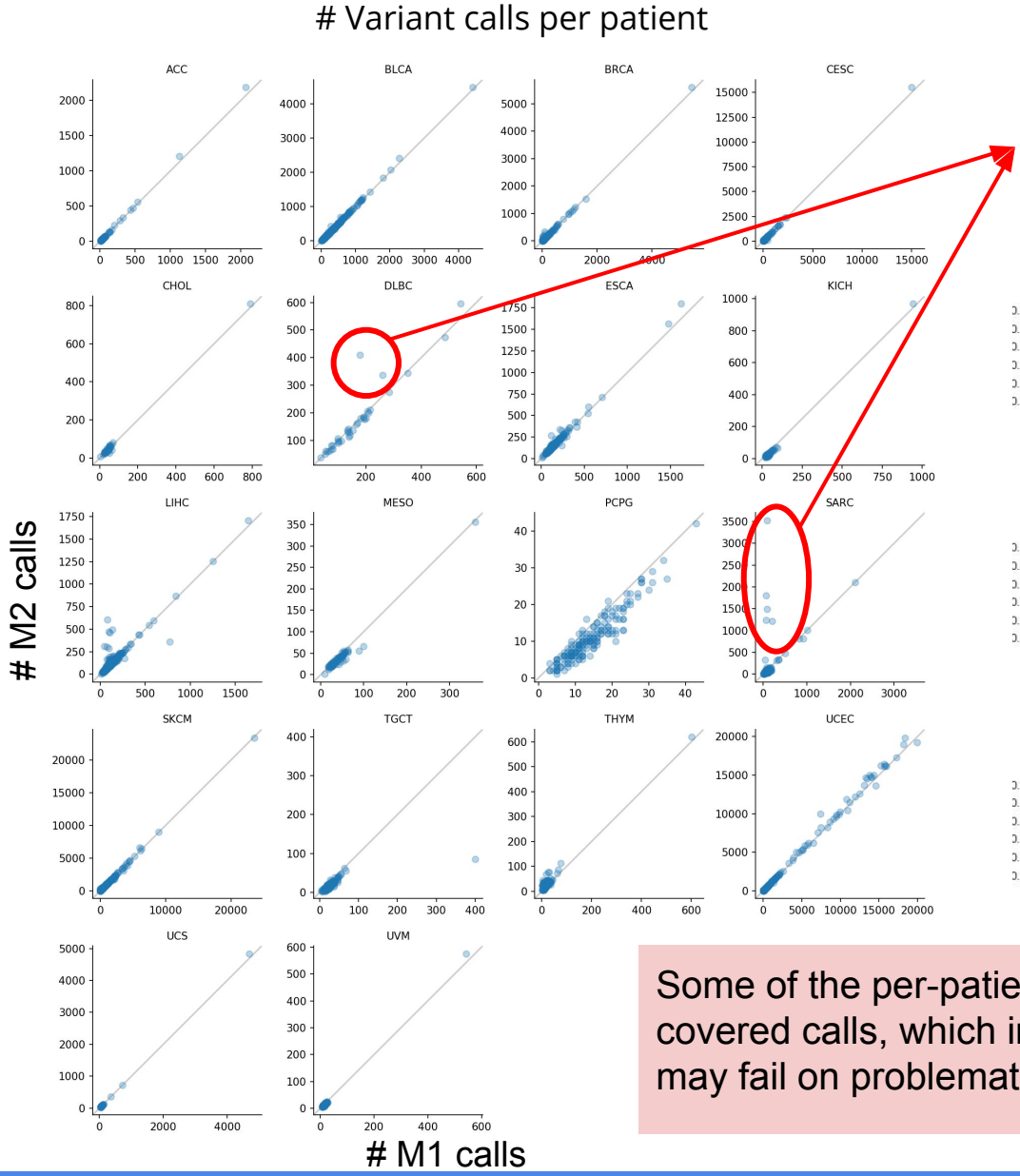
Number of variants removed by each filter, normalized by number of variant called per tumor type



Weak evidence and strand bias are popular filters across tumor types, they are less cohort-specific.

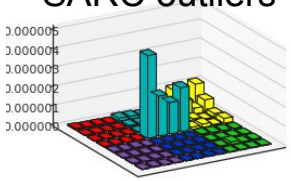
Clustered events filters most of the variants, and are very specific to UCEC and SKCM cohorts.

#Calls per patient aligns, despite some patients

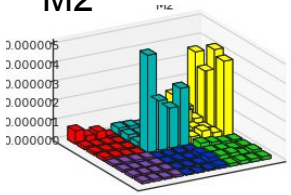


Patients with a lot more M2 calls are called at very low alt count - they are almost exclusively from alt_count < 5

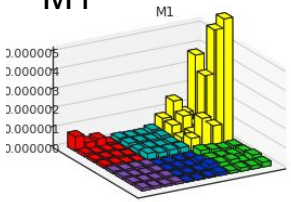
SARC outliers



M2



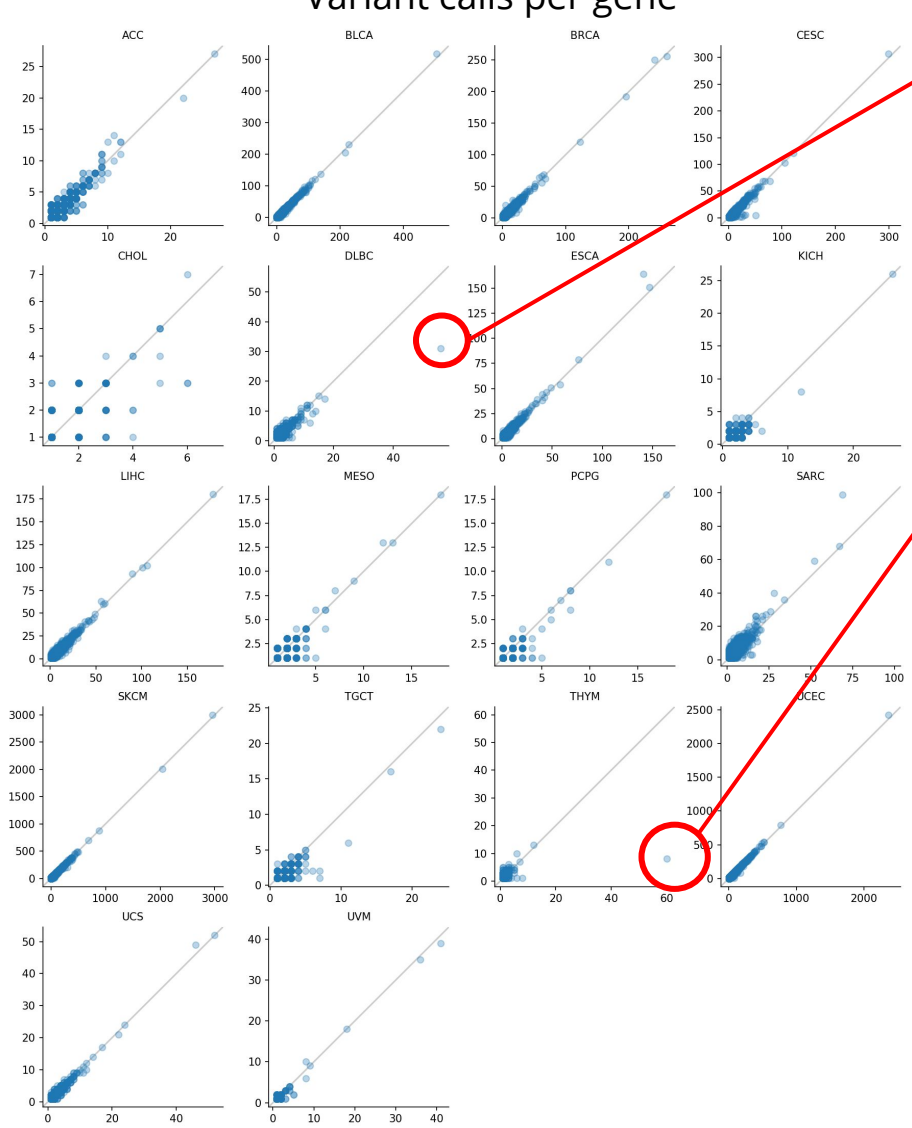
M1



SARC G(C->A)* mutation seems to be an artifact and are enriched in the SARC outlier patients.

Some of the per-patient discrepancies come from poorly covered calls, which indicates M2's data dependent modeling may fail on problematic samples.

#Calls per gene is largely consistent, but M2 misses some known ones



27 PIM1 calls are removed in DLBC by “clustered_events” filter

Open Access | Published: 21 June 2012

PIM kinases are progression markers and emerging therapeutic targets in diffuse large B-cell lymphoma

L Brault, T Menter, E C Obermann, S Knapp, S Thommen, J Schwaller & A Tzankov

In THYM cohort, 72/80 calls from M1 at chr7:74146970 are filtered by poor mapping quality.

Published: 29 June 2014

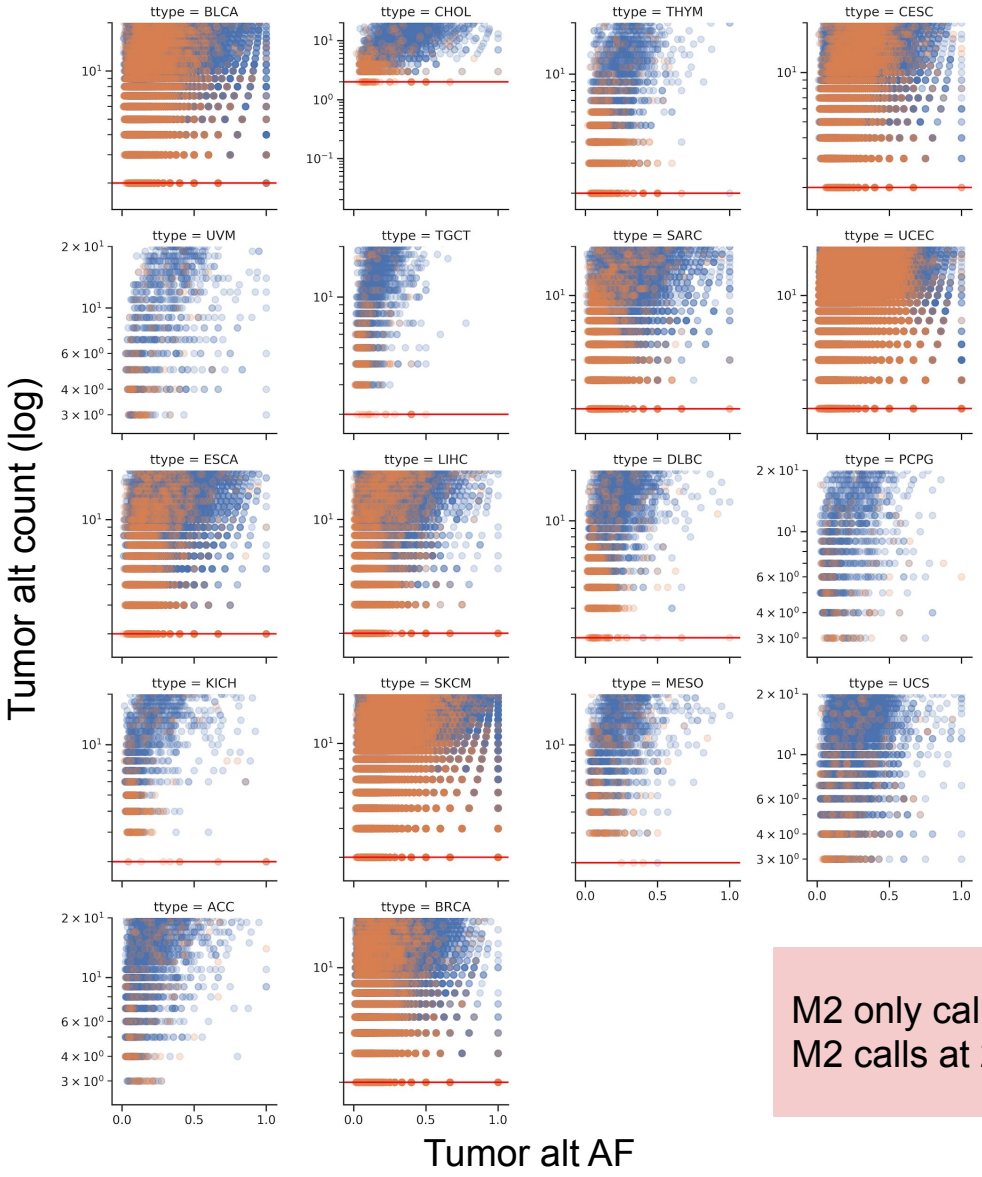
A specific missense mutation in GTF2I occurs at high frequency in thymic epithelial tumors

Iacopo Petrini, Paul S Meltzer, In-Kyu Kim, Marco Lucchi, Kang-Seo Park, Gabriella Fontanini, James Gao, Paolo A Zucali, Fiorella Calabrese, Adolfo Favaretto, Federico Rea, Jaime Rodriguez-Canales, Robert L Walker, Marbin Pineda, Yuelin J Zhu, Christopher Lau, Keith J Killian, Sven Bilke, Donna Voeller, Sivanesan Dakshanamurthy, Yisong Wang & Giuseppe Giaccone

Nature Genetics **46**, 844–849(2014) | Cite this article

M2 filters removes calls from some well-studied variants that have been verified

M2 calls at low alt coverage-1



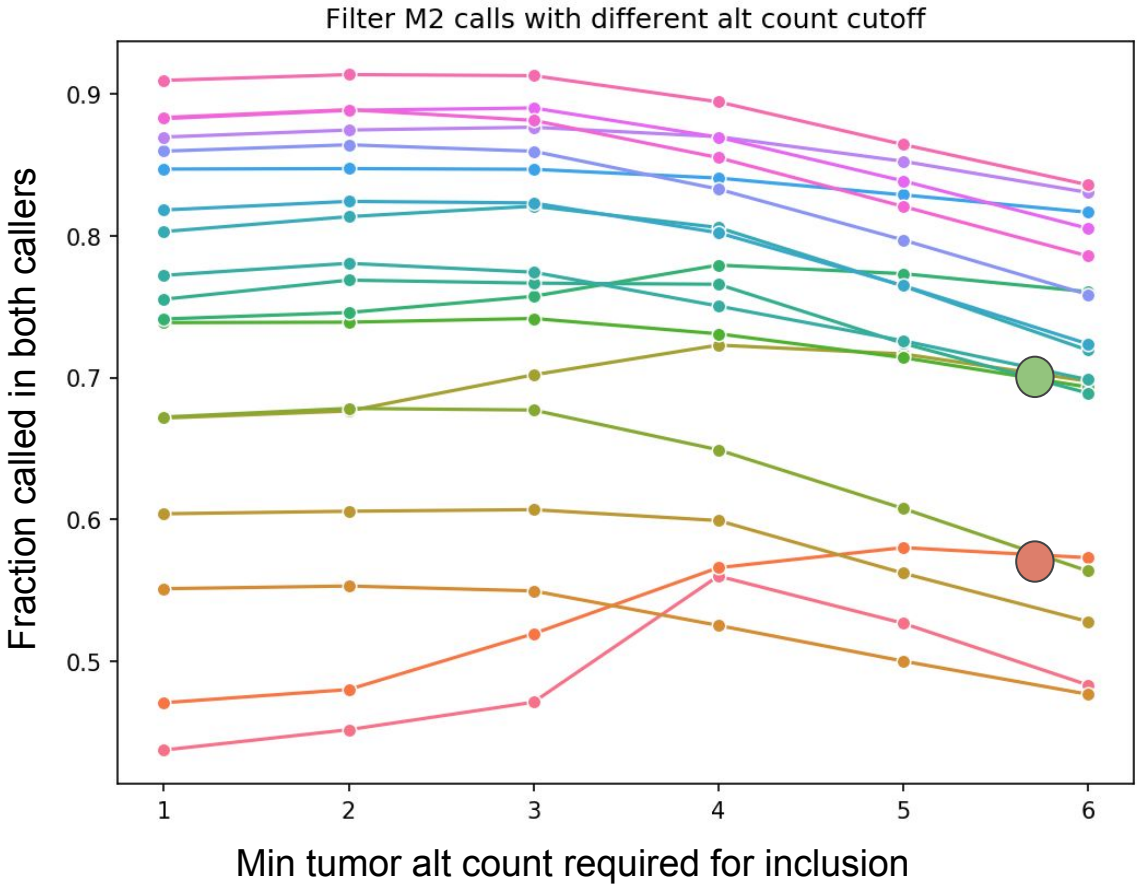
M2-only calls

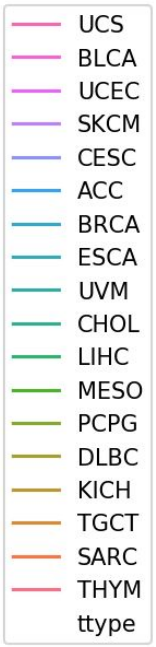
Overlapping calls

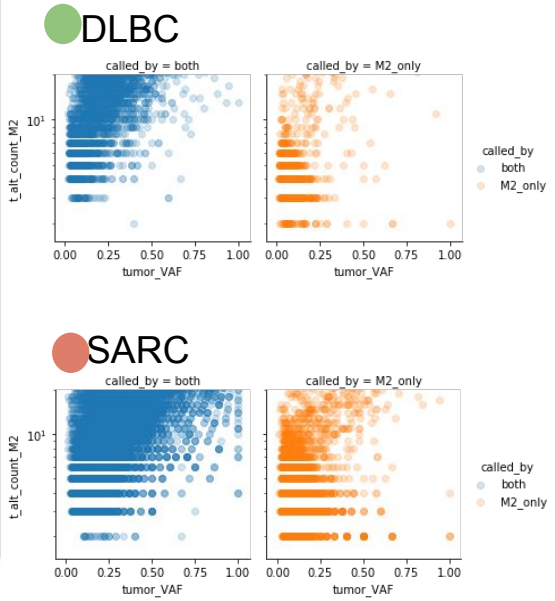
Alt_count =2

M2 only calls are enriched at low AF with low alt count.
M2 calls at 2 alt count in many cohorts!

M2 calls at low alt coverage-2



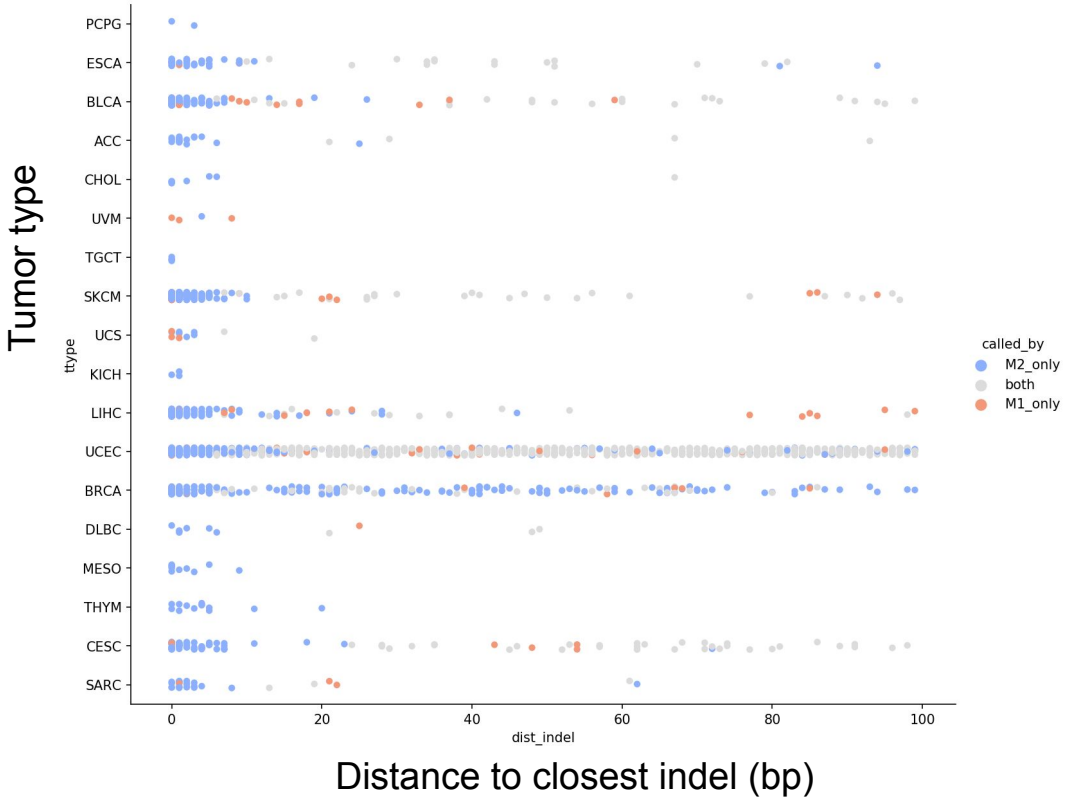




We can expect no more than 3% increase in %overlap between M1/2 call sets for most of cohort.

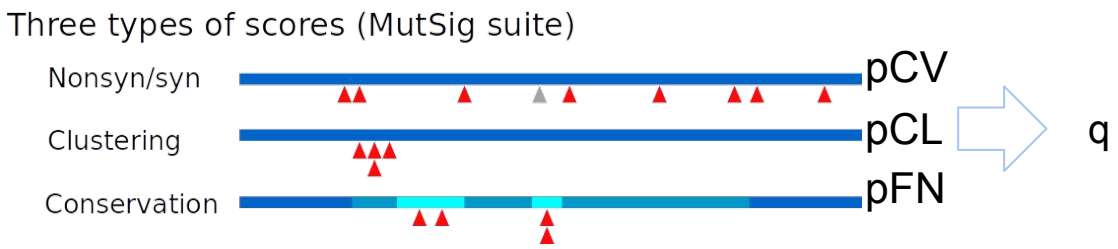
M2 calls more near (potential) indels

Variant called vs distance to nearest indel (called by M2)



M2 calls at regions with SV and indel, where those regions have been removed in M1 postfilter.

However, this only explains less than 2% of M2-only calls

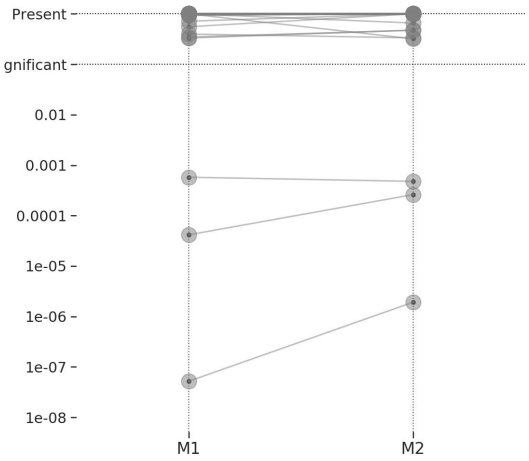


Result

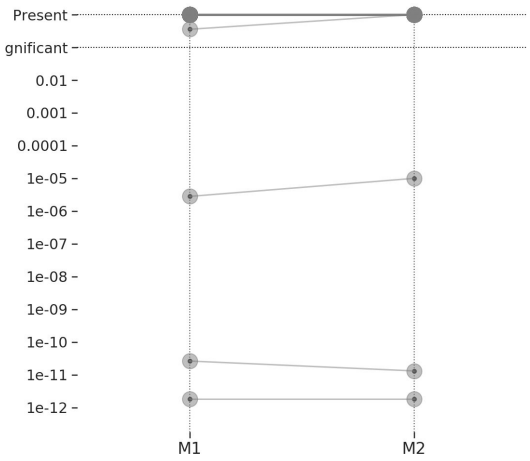
Lessons learned from driver gene discovery (MutSig)

MutSig: Consistency in several tumor types

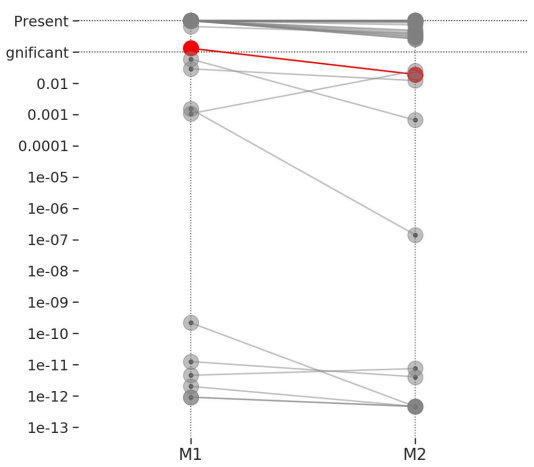
CHOL



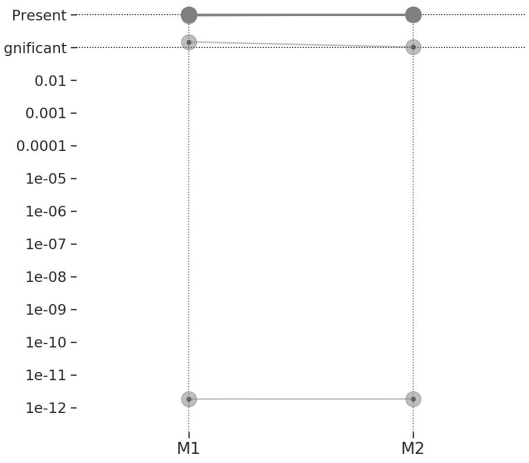
PCPG



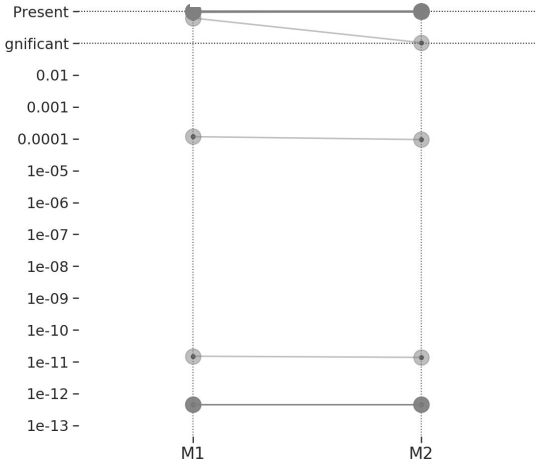
UCS



KICH

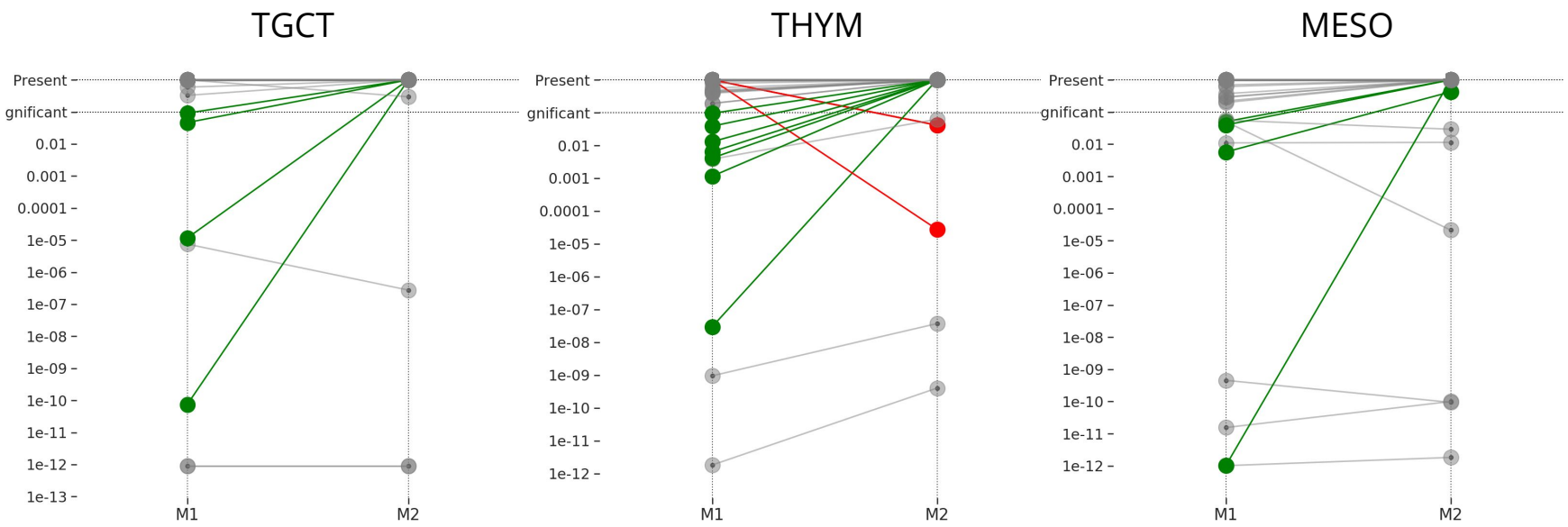


UVM



These cohorts are small; fewer patients limits MutSig's discovery power.

MutSig: Significance only from M1 calls



In TGCT/THYM/MESO, significant genes from M1 usually have mapping issues, e.g. zinc finger proteins, PAK2, PTMA. We hypothesize that there is a systemic problem with M1 handling **mapping artifact**.

Those cohorts are late TCGA cohorts, therefore not included in the M1 PoN.

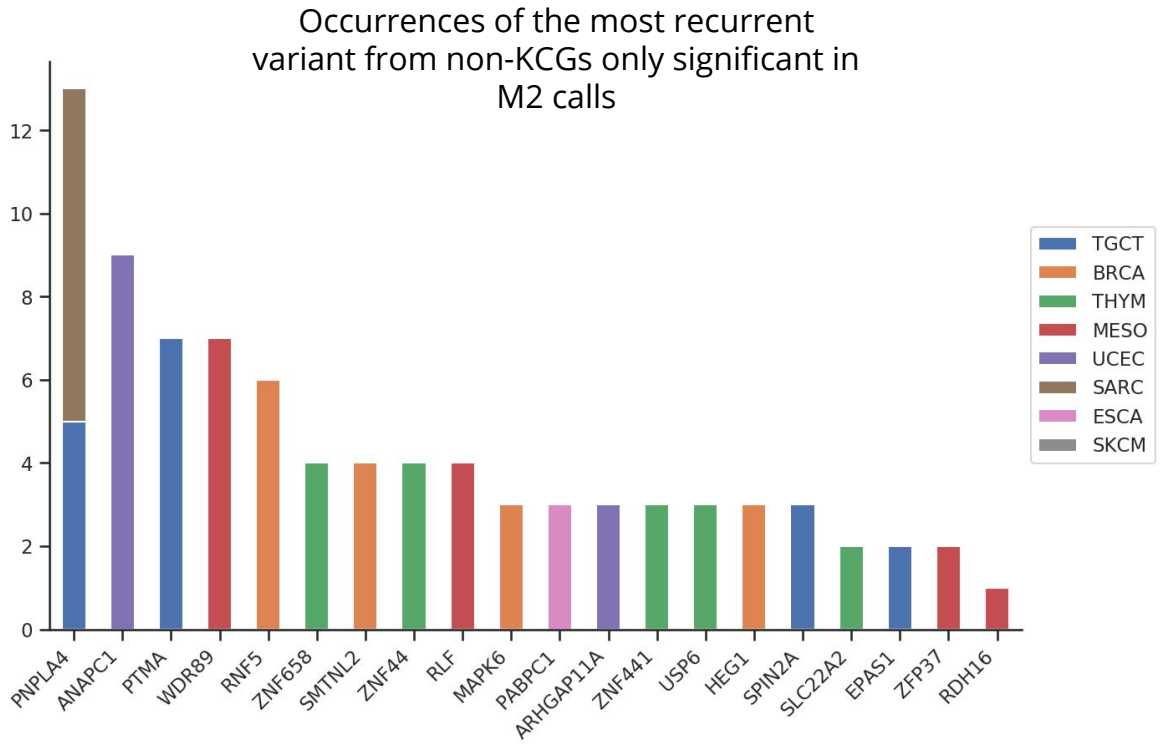
Genes only significant in M1

Known Cancer Gene (KCG): CTCF & FOXA1 from BRCA have lost significance in M2, however calls are not recurrent. FOXA1 have 3 calls filtered by “clustered events”.

Most of non-KCGs have mapping issues, can be categorized into

- Germline polymorphism: HLA-B
- Homopolymer bleedthrough: PNPLA4
- General mapping issue: AP2B1

Base quality issue: HEG1



FOXA1 in BRCA: three missense mutation in M1 are filtered by M2 “clustered_events”

BRCA: 14-38061249 FOXA1

BRCA: 17-12016571 MAP2K4

GATK Mutect2 removes variant coming from **a reassembled region** that had too many other variants, which is a proxy for mapping artifacts (similar idea applied in M1 prefilter: sum of quality scores of the mismatches <= 100)

chr	start	end	strand	score	match
chr17	12016559	12016659	+	1000	100
chrX	72745196	72745296	-	880	94
chr17	39329140	39329167	+	220	23

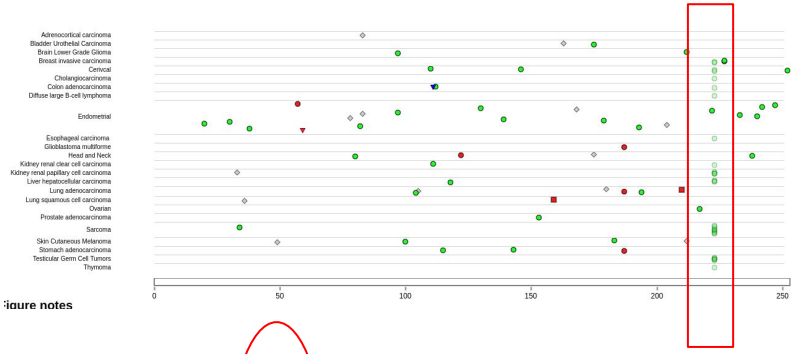
chr	start	end	strand	score	match
chr17	12016535	12016635	-	940	97
chrX	72745220	72745308	+	740	81
chr1	192506...	192506...	+	320	34
chr20	57410245	57410268	+	210	22

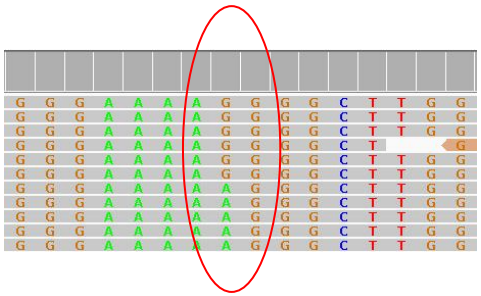
The three variants on *FOXA1* were filtered together by “clustered events”, however they do not come from the same haplotype, suggesting that this filter could be too aggressive.

nonKCG in M1: Recurrent mapping artifacts

PNPLA4, chrX 7868821

PNPLA4 is determined to be significant in SARC and TGCT. This variant (7868821, mis) is called in 8 SARC samples, 5 TGCT samples, 3 LIHC samples, 2 BCRA and 2 CESC samples. M2 was able to completely omit this variant





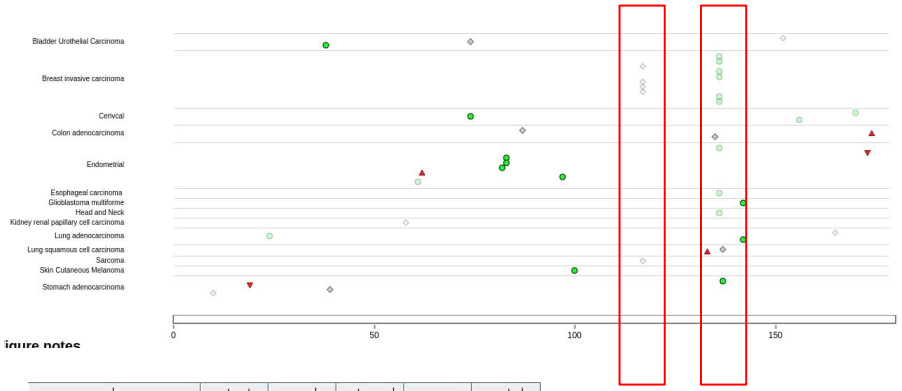
Neighboring Homopolymer could also be the source of mapping artifact

chr	start	end	strand	score	match	mis-match	rep. mat...	N's	Q gap c...	Q gap b...	T gap c...	T gap b...
chrX	7868786	7868886	-	980	99	1	0	0	0	0	0	0
chrY	16213527	162136...	-	730	87	12	0	0	1	1	1	3

RNF5, chr6 32147809/32147865

RNF5 is significant in BRCA cohort, 4 samples are called at 32147809 and 4 are called at 32147865 in M1.

32147809 32147865



chr	start	end	strand	score	match
chr6_ssto_hap7	3495511	3495611	-	1000	100
chr6_qbl_hap6	3408803	3408903	-	1000	100
chr6_mcf_hap5	3527610	3527710	-	1000	100
chr6_mann_hap4	3490607	3490707	-	1000	100
chr6_cox_hap2	3618464	3618564	-	1000	100
chr6_apd_hap1	3462524	3462624	-	1000	100
chr6	32147...	32147...	-	1000	100
chr8	38458...	38458...	+	970	98
chr16	87514...	87514...	+	220	23
chr4	16689...	16689...	-	210	21
chr12	15762...	15762...	+	210	21
chr5	13975...	13975...	+	200	20

Reads also align to contigs from chr6 MHC haplotypes. Similar alignment observed in *HLA-B.*

M2 base quality filter removes artifact

HEG1, chr3 124739694

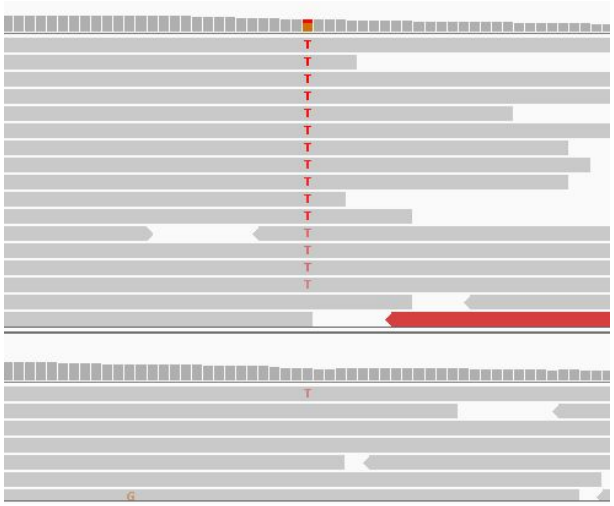
Significant in BRCA, 4 recurrent base quality artifact are not called.

No consensus sequence detected, but the sequencing quality varies - only half of the reads have base quality > 10.

SMTNL2, chr17: 4496467

Significant in BRCA by M1 calls.
4 recurrent variants detected by M1 in chr17: 4496467 are entirely omitted from M2. 3 of 4 are removed by "base quality filter".

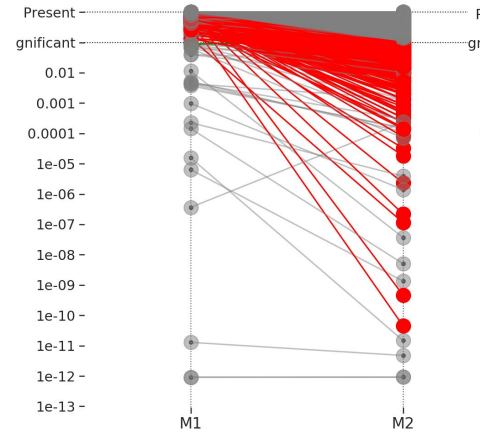
Base-quality filter is controlled by "min-median-base-quality" which is preset to 20.



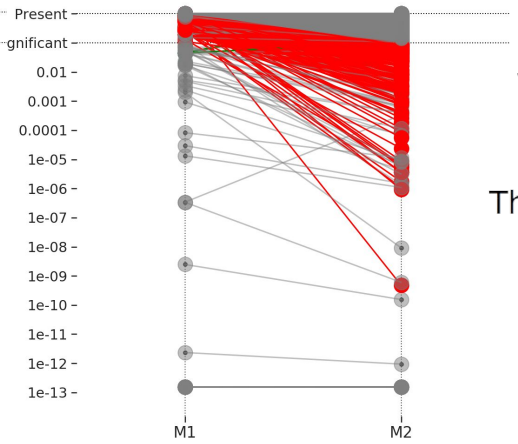
M2 does more than filter, it also keeps track of base quality of the reads that have been filtered.

MutSig: Significance only from M2 calls

LIHC

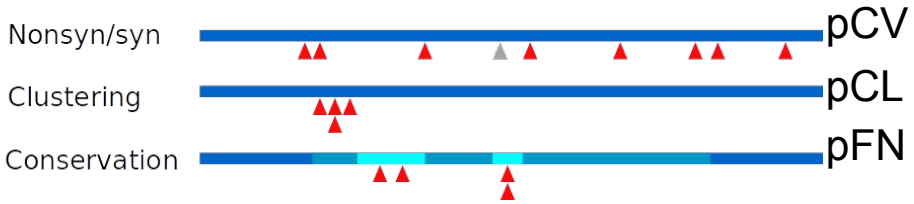


UCEC

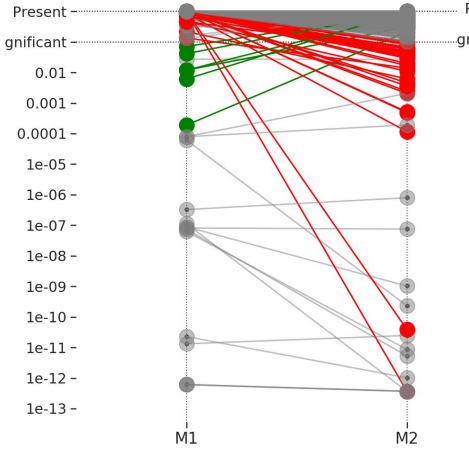


M2 null model seems off.
Which p value accounts for the big difference?

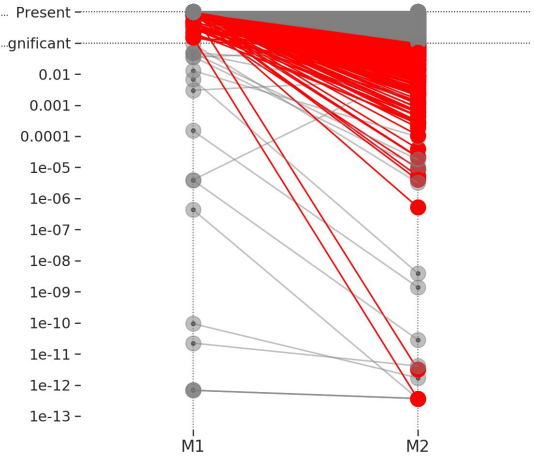
Three types of scores (MutSig suite)



BRCA

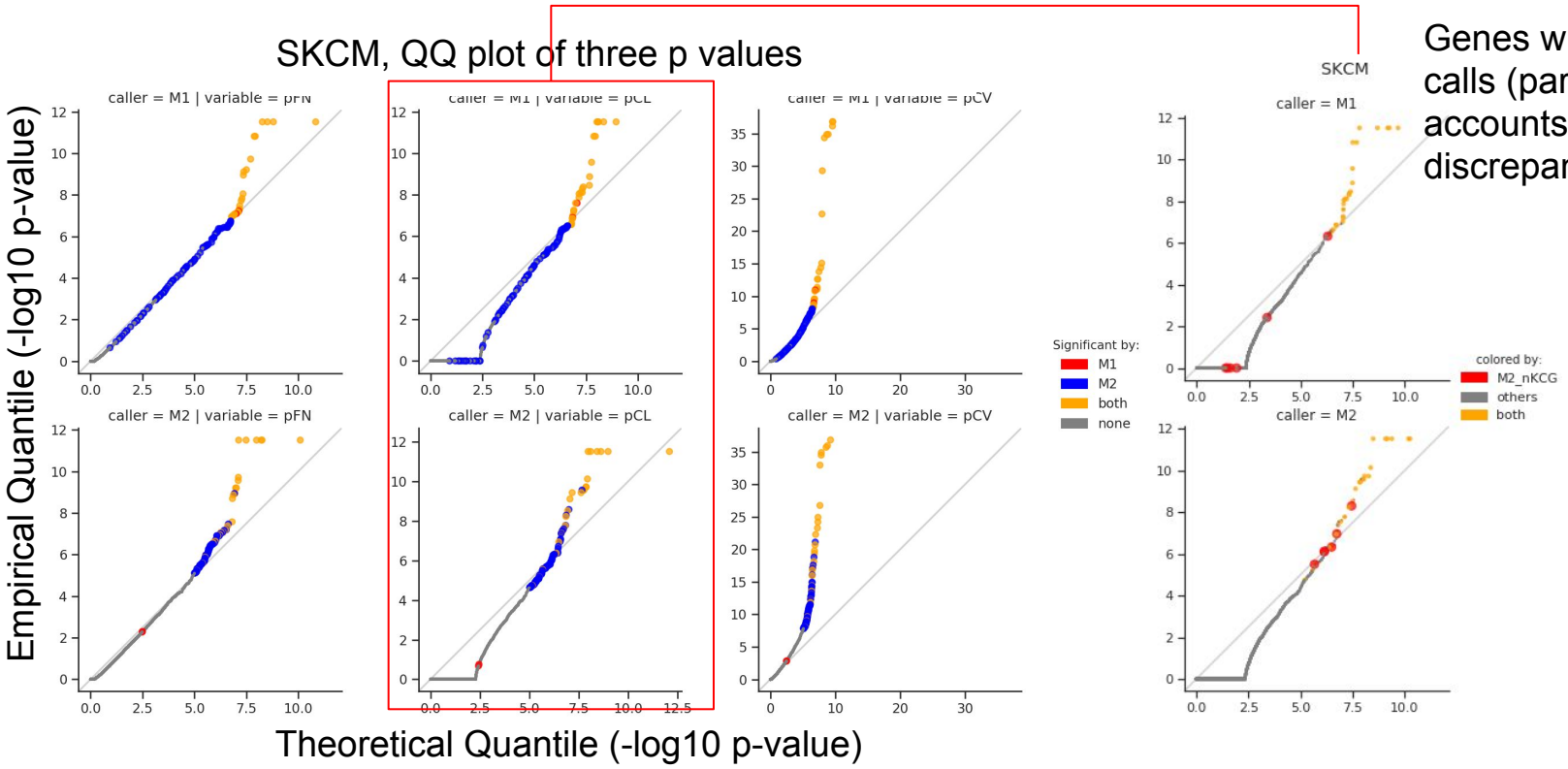


CESC



Inflated pCL suggests reviewing recurrent calls

SKCM, QQ plot of three p values



Genes with recurrent calls (partially) accounts for the pCL discrepancies

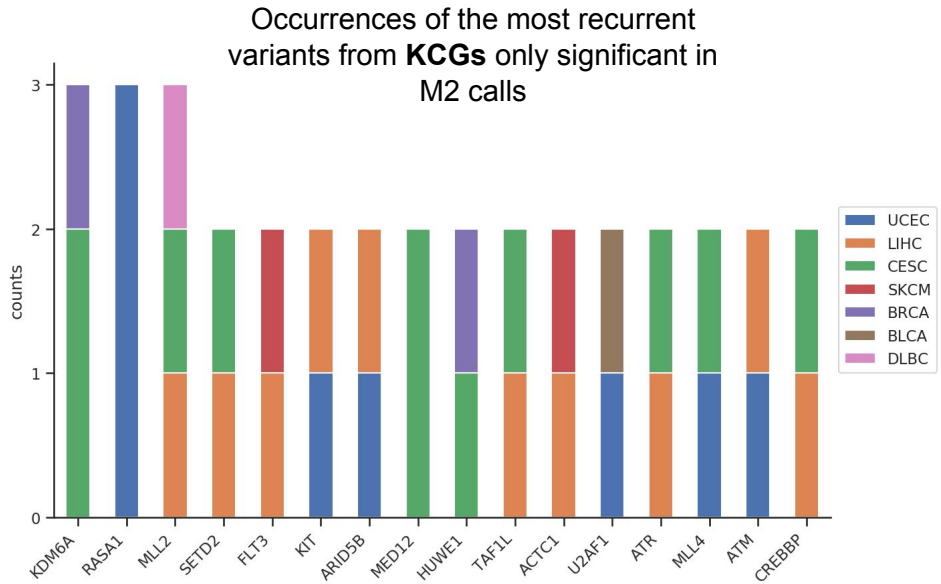
pCL shows the most inflation for cohorts where M2 identifies significantly more driver genes than M1.

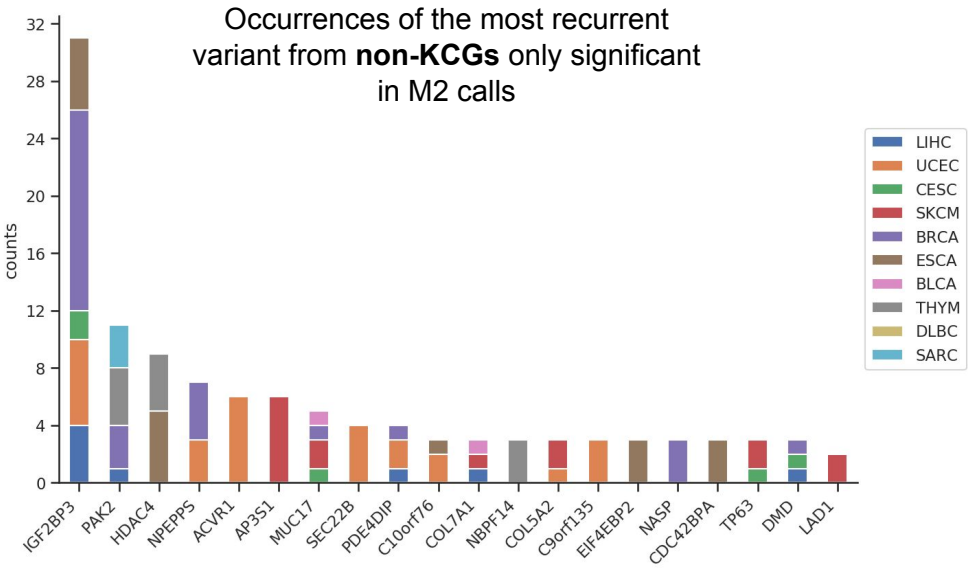
This has shed light on prioritization of manual review: Looking at the **recurrent calls** on **non-KCGs**.

Genes only significant in M2

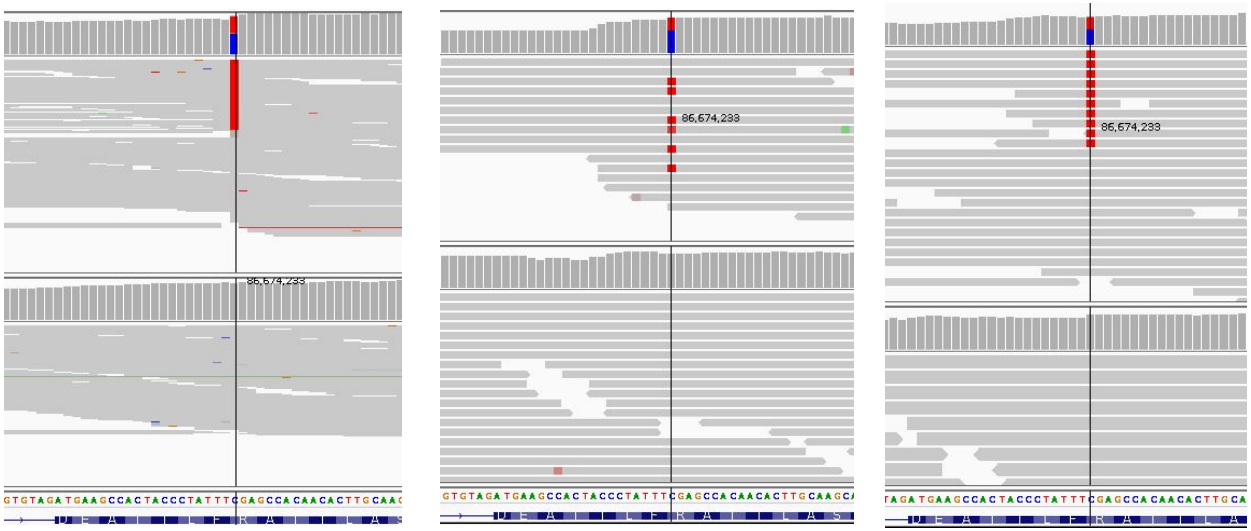
Genes only significant based on the M2 call set is a very long list, and there are some very long genes like SYNE. Some signal from M2 overcalling (compared with M1) seems to scale with gene length.

In manual review I would focus on the recurrent variants (called only by M2 in >=2 samples)





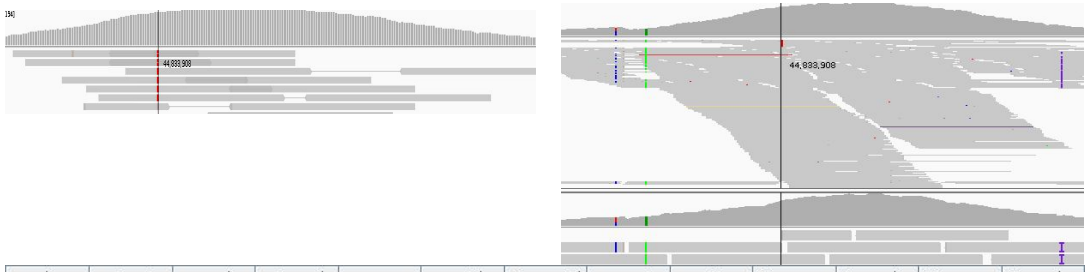
KCGs only significant in M2: reason for absence in M1 is not clear



RASA1,
5: 86674233

Significant in UCEC. There are three occurrences in M2. The reason for M1 exclusion is not clear - the region is noisy in PoN, but should not have been filtered given high AF.

chr	start	end	strand	score	match	mis-match	rep. mat...	N's	Q gap c...	Q gap b...	T gap c...	T gap b...
chr5	86674166	86674266	+	980	99	1	0	0	0	0	0	0
chr5	13866991	13867022	-	280	30	0	0	0	1	1	1	1
chr1	239869...	239869...	-	240	25	0	0	0	1	5	0	0
chr4	169756...	169756...	+	230	25	1	0	0	0	0	1	1
chr14	21455387	21455408	-	210	21	0	0	0	0	0	0	0

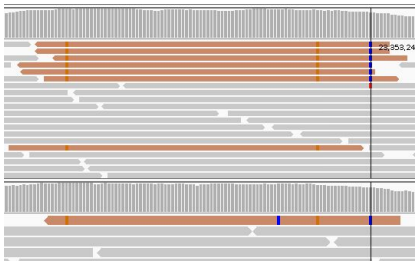


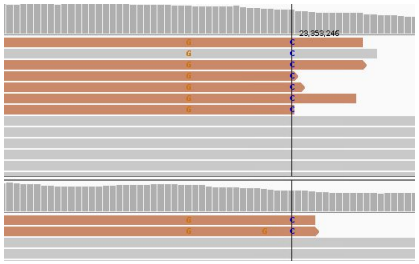
KDM6A
23: 44833908

chr	start	end	strand	score	match	mis-match	rep. mat...	N's	Q gap c...	Q gap b...	T gap c...	T gap b...
chrX	44833871	448339...	+	980	99	1	0	0	0	0	0	0
chrY	15560512	155609...	-	670	75	5	0	0	0	0	3	376

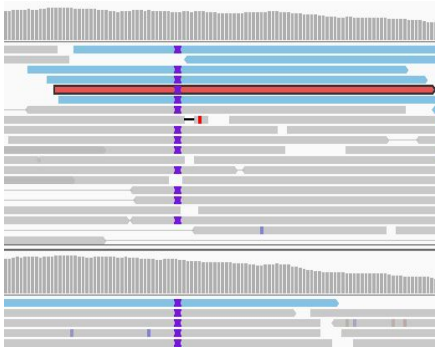
Non-KCG significant by M2: *IGF2BP3*, realignment

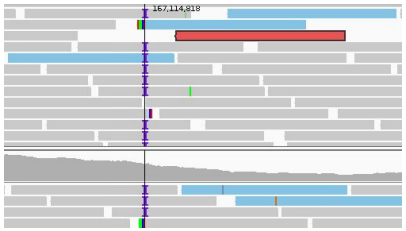
IGF2BP3 7:23353246 was called 31 times only from M2 (14 patients from BRCA, 2 from CESC, 5 from ESCA, 4 from LIHC, 6 from UCEC, significant in all these tumor types)





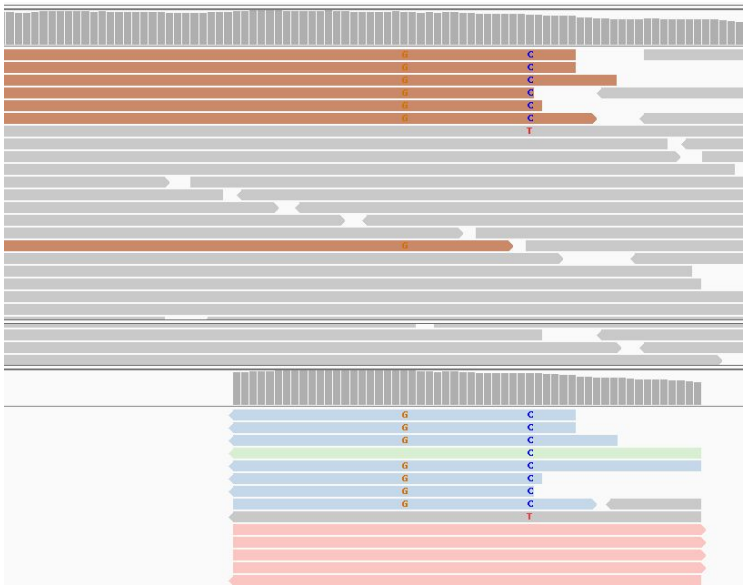
Chimera mate:



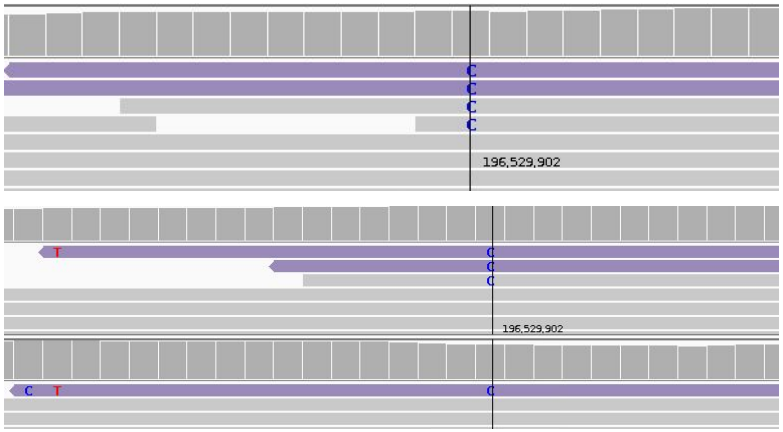


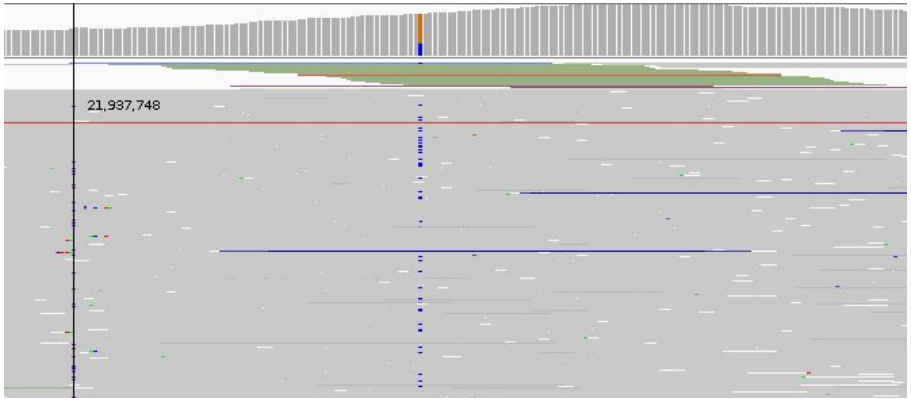
chr	start	end	strand	score	match
chr6	167114...	167114...	-	970	98
chr7	23353151	233532...	-	940	97
chr7	56665721	566658...	+	760	85
chr3	56952465	569525...	+	290	33

However (almost) all alt variants are on **chimeric** reads that should have been excluded by the prefiltering step. The mystery can be resolved if we look at the -bamout from M2.

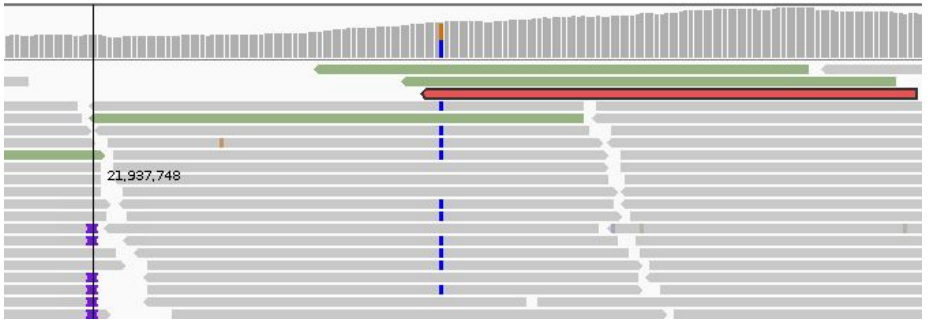


non-KCG significant by M2: *PAK2*, realignment of mate to the same contig





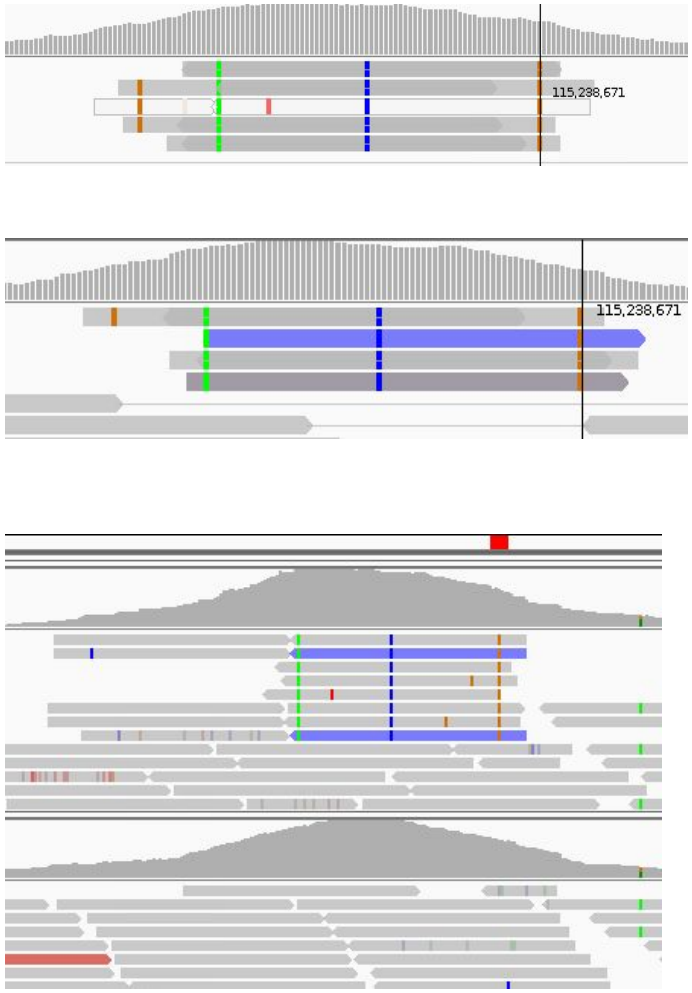
PAK2 3-196529902 is called only by M2 in 4 THYM, 3 SARC, 1 LIHC and 3 BRCA patients. Similarly, some reads supporting the alternate base are **chimeric**.



chr	start	end	strand	score	match	mis-match	rep. mat...	N's	Q gap c...	Q gap b...	T gap c...	T gap b...
chr3	196529...	196530...	-	920	120	5	0	0	0	0	0	0
chr15	21937720	21937836	+	920	116	0	0	0	1	9	0	0

non-KCG significant by M2: *AP3S1*, realignment

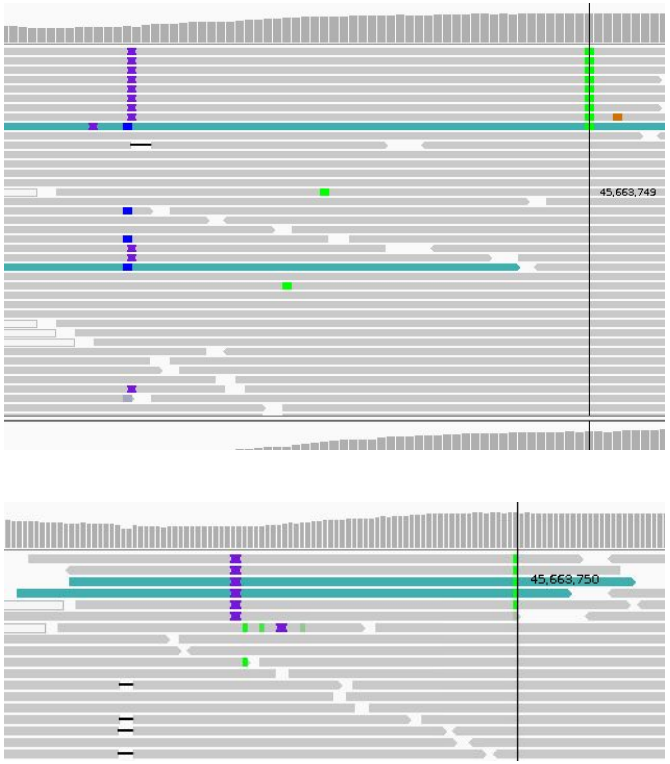
AP3S1 5: 115238670, M2 calls 6 samples from SKCM. Among all 6 samples, we can observe reads that harbors three substitutions. There is one alternative alignment (with better score) at chr1 bearing one insertion.

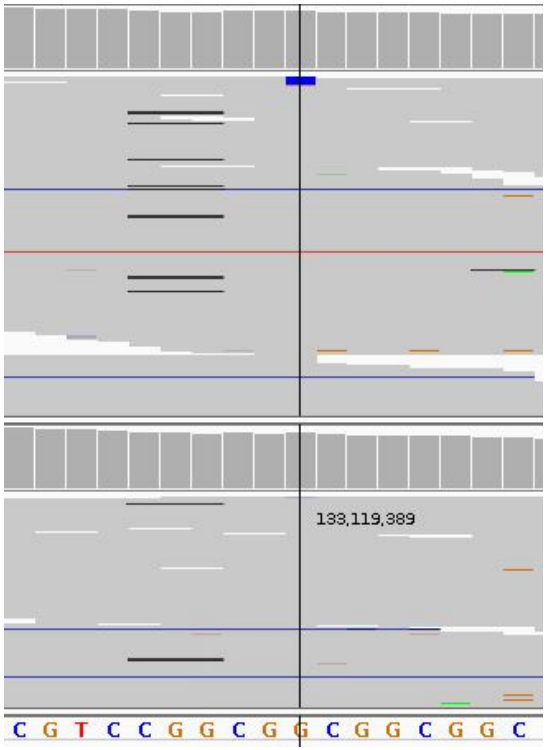


non-KCG significant by M2: *NPEPPS*, confounded by indels

NPEPPS 45663749 called in 5 BRCA patients, 3 UCEC patients and 2 SARC patient. A very consensus region can be found within the same contig.

In these two samples, the variant overlays with an upstream insertion (length varying from 2-4bp). It would also be possible that those variants are filtered by ”proximal gap” filter in M1.





Conclusion & Discussion

Conclusion

From call set comparison,

- Number of calls per gene / per patients correlates well, however M2 filters some well-studied variants, and produces artifactual calls at some outlier patients in SARC
- M2 calls at low AF, low alt supporting regions, which might capture false positives
- M2 calls in regions with lots of genomic alterations.

Conclusion

From significance analysis and manual review,

- pCL, which characterises clustering of variants, may point to recurrent artifacts that causes p value inflation in M2 significance analysis.
- Mapping artifacts are the major source of discrepancies
  - M2 filters many M1-only calls from mapping artifact at base-level.
  - However, if the mapping issue triggers realignment, M2 might produce calls at those challenging regions that M1 will not call.
- M2 filters
  - M1 handles base quality filtering in prefilter, it filters those bad reads and forgets their existence. M2 keep track of median_base_qual as a hard filter. (Those reads that are filtered need to be remembered)
  - "Normal_artifact" may filter a variant based on one support from normal, "Strand_bias" often filters variant with <5 alt support
  - "Clustered_events" might have been too sensitive in removing variants that are not on the same haplotype.

Perspective

Choice of caller

- If user is interested to call SNV near SV or indel, consider M2 to fully explore the assembly & alignment space.
- M2 makes lots of assumptions to fit M1's empirical parameters to the input data, might be stuck in local optima

Further direction

- There are other work we can do to evaluate call set, e.g. tumor subtyping, mutation signatures, and phylogeny inference.
- M2 has several good features that we can incorporate into a future version of MuTect (perhaps M3?).

Acknowledgement

Getz Lab

- **Julian Hess**
- **Chip Stewart**
- Chet Birger
- Aaron Graubert
- Jialin Ma
- Ziao Lin
- **Gad Getz**

Broad

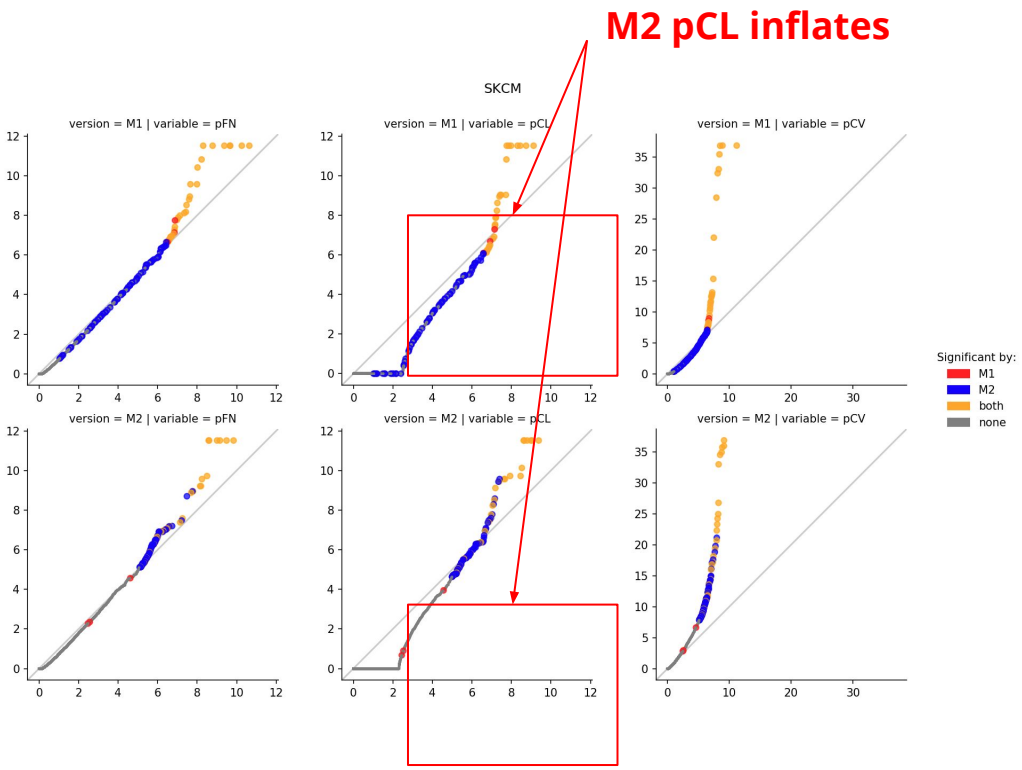
- Abhishek Niroula
- Cara Mason
- Bhanu Gandham
- Soo Hee Lee
- David Benjamin

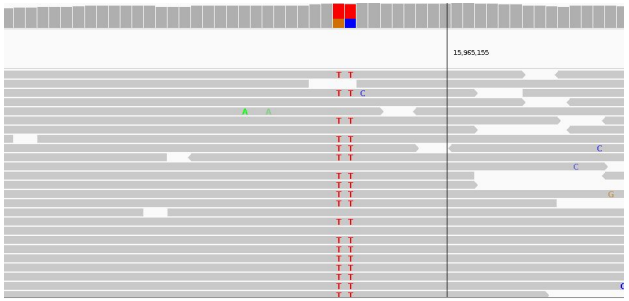
Prefilters: Getting good reads

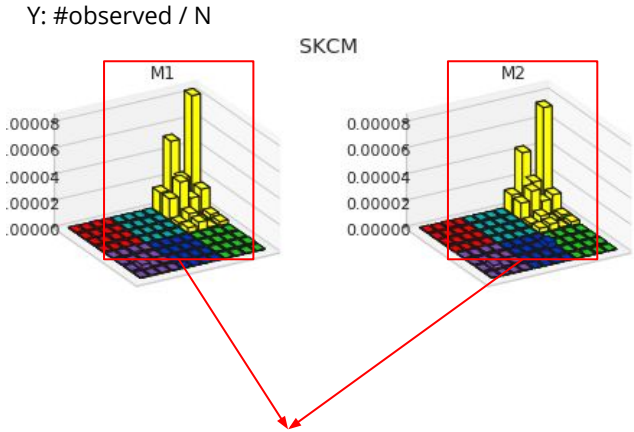
Not all reads are fed into callers - we need to select for true signals from sequencer / aligners.

Category		M1	M2
Mapping quality	Minimum map quality for reads to be included	> 0	> 0, also remove secondary alignments
	Mate rescue	Disabled	Disabled, **even removed read pairs that aligns to different contig!**
	Clipping	Remove reads if >30% bases are soft-clipped	Soft-clipped bases are included, but can be reverted by "--dont-use-soft-clipped-bases true" **[Hard-clip] Remove bases from end of reads until base quality > 10**
Base quality	minimum base quality to be considered as a variant	> 5	> 10
Overlapping read pair		If both agree, keep the one with better quality, otherwise both discarded	Will adjust model downstream
Other filters		**Sum of quality scores of mismatches <= 100**	Remove reads marked as duplicates(0X400), various formatting / consistency check **Tumor/Normal prefiltering crtieria identical**

M2 clustered events filter: bad case in SKCM





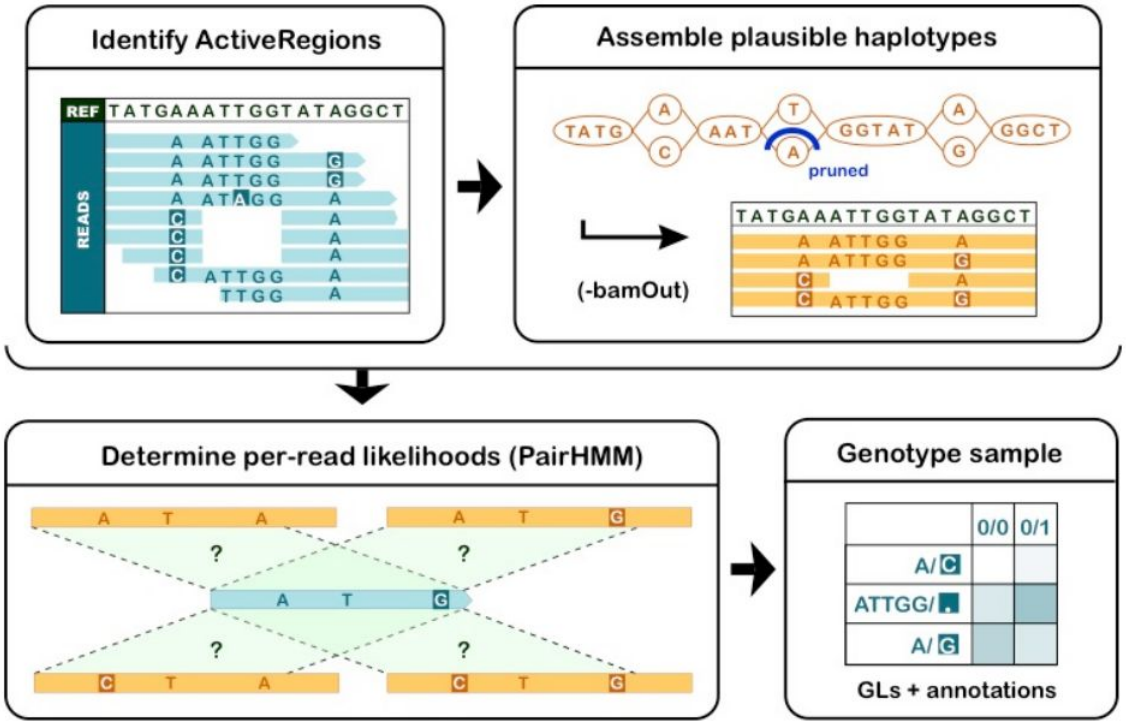


Despite some discrepancies, the intersect of sig genes from two callers are rather stable in three tests.
M2 could have missed quite some DNV!

C>T mutation rate is lowered

Prefilters: M2 can downsample / reassemble / realign

Run somatic likelihood model for one iteration; combine all alt allele into meta ALT. Downsamples 1000 reads per region.



Use K=10/25 to build DeBruijn graph, keep track of support of each path.

minPruning=2 will remove reads supported by only one read, number of haplotype is capped at 128

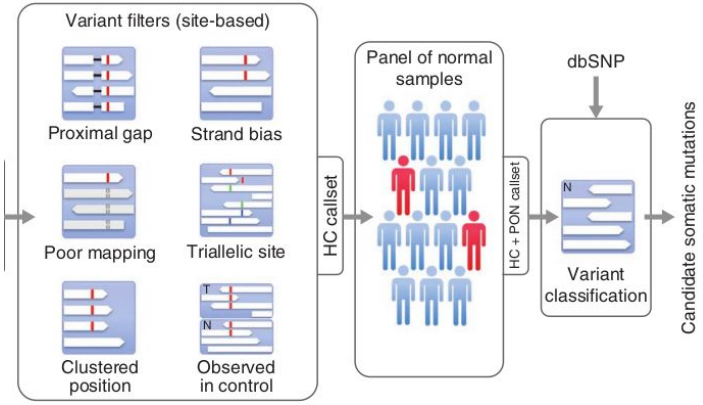
PairHMM performs alignment and calculates per-read likelihood for each allele, marginalized to get per-read likelihood for each allele.

M2 will skip sites in PoN, and sites with high support in normal
Phasing comes for free

From GATK tutorial

Post-filters: M2 hard filtering

Hard filter: directly assign P(error)=1



Trialleic site: M2 by default will give two alt allele.
Strand bias filter used to dependent on log odds, has become probabilistic.
Proximal gap is removed since M2 is a indel and SNV caller.
Duplicate_evidence, base_quality, fragment_length are new filters from M2.

Filter	Threshold	Explanation
clustered_events	max-events-in-region	mutations sharing an assembly region
duplicate_evidence	unique-alt-read-count	unique insert start/end pairs of alt reads
multiallelic	max-alt-alleles-count	passing alt alleles at a site
base_quality	min-median-base-quality	median base quality of alt reads
mapping_quality	min-median-mapping-quality	median mapping quality of alt reads
fragment_length	max-median-fragment-length-difference	difference of alt and ref reads' median fragment lengths
read_position	min-median-read-position	median distance of alt mutations from end of read
panel_of_normals	panel-of-normals	presence in panel of normals

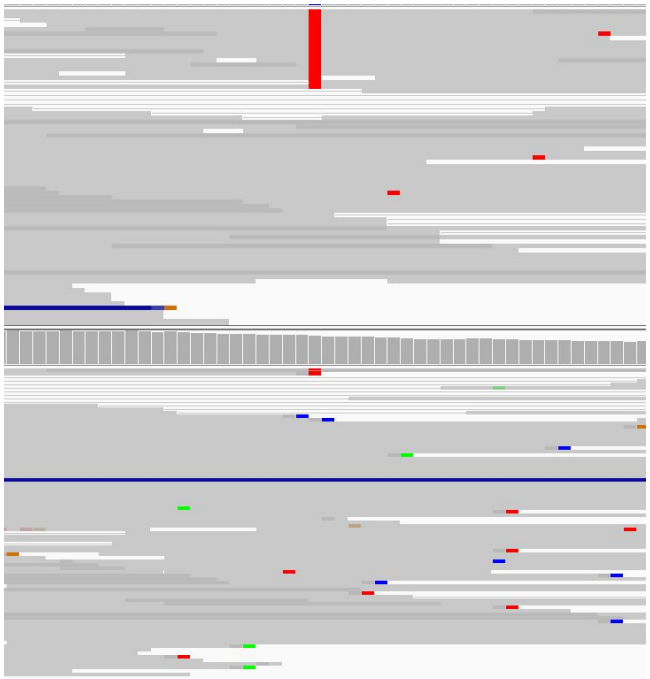
Post-filters: M2 probabilistic filtering

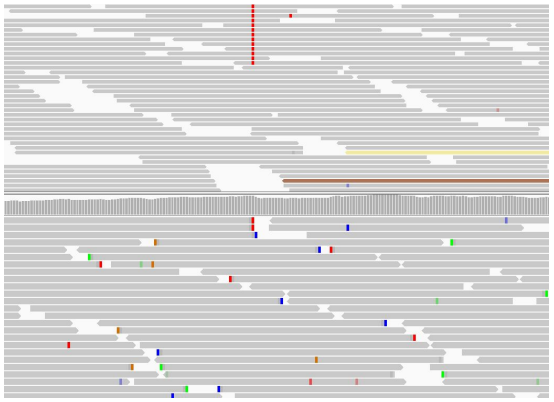
Probabilistic filtering, compared to hard filters, gives an error probability from different artifact modes. Allele fraction clustering model aims to capture the sharply defined heterogeneity among subclones. This model will output likelihood of a alt reads out of d total reads in case of real somatic variant.

- **Weak evidence by** cluster weights and allele-fraction-adjusted likelihoods
- **Strand artifact by PGM**
- Germline filter
- Contamination filter
- **Read Orientation Artifact Filter**
- Polymerase slippage: indel in STR
- Normal artifact: apply somatic likelihood model to normal
- Bad haplotypes: M2 does phasing for each assembly region. We assign a “bad haplotype” probability equal to the greatest technical artifact probability of any in-phase call within a certain distance, by default 100 bases.

I will have an another talk on M2 filters when I get M1 Mutect full directory.

M2 “normal artifact” filter





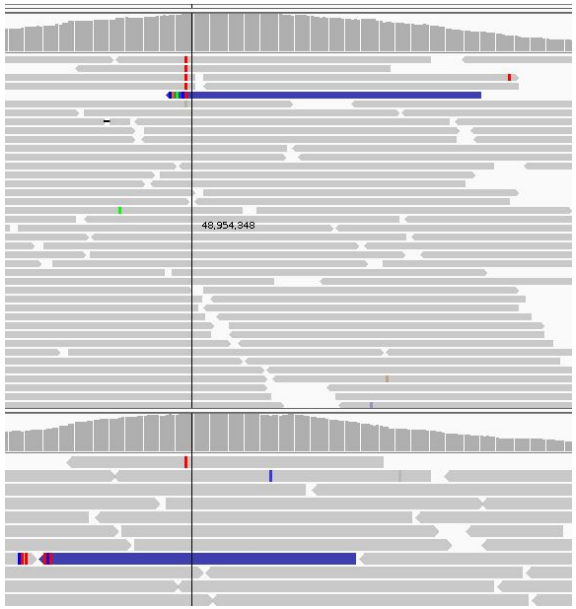
The presence of A in normal is due to improper trimming of adapters (?), should not be regarded as evidence.

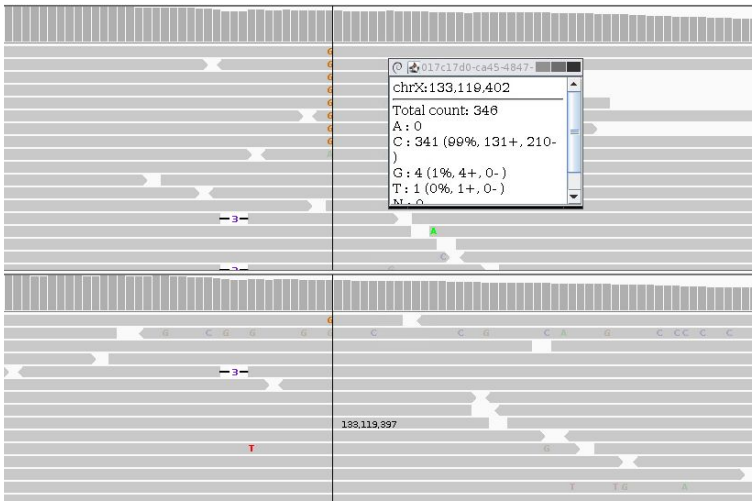
CESC: 12-49431002 MLL2
effect_M1=silence; effect_M2=nan;
q_M1=1.0, q_M2=3.206816e-12
position to view: chr12:49,431,002-49,431,003

The two example comes from one sample, suggesting there could be problems during handling..

CESC: 3-176750883 TBL1XR1
effect_M1=missense; effect_M2=nan;
q_M1=1.0, q_M2=0.03714929
position to view: chr3:176,750,883-176,750,884

“Normal artifact” removes by one occurence in normal





CESC: 13-48954346 RB1
effect_M1=silence; effect_M2=nan;
q_M1=1.0, q_M2=5.284868e-07
position to view: chr13:48,954,346-48,954,347

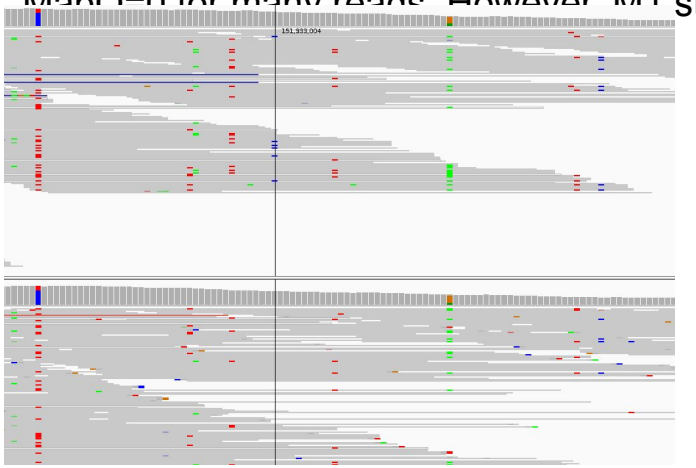
There is only **one** ALT observed in normal.

LIHC: 23-133119396 GPC3
effect_M1=silence; effect_M2=nan;
q_M1=1.0, q_M2=0.06934836
position to view: chr23:133,119,396-133,119,397

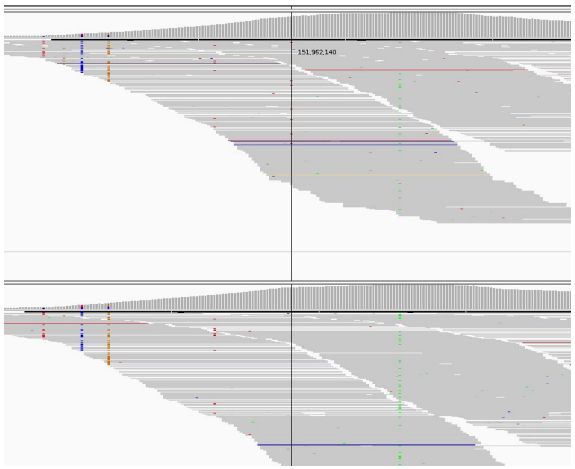
Note that this region is highly covered.

Map Quality: MLL3

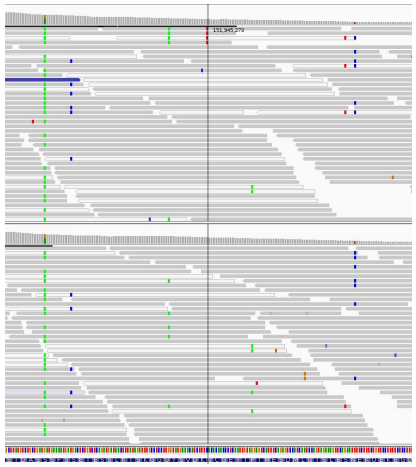
Terrible mapping, there are lots of variation in normal,
MapQ=0 for many reads. However, M1 still calls some variants here.



Silent mutation in M1, CESC



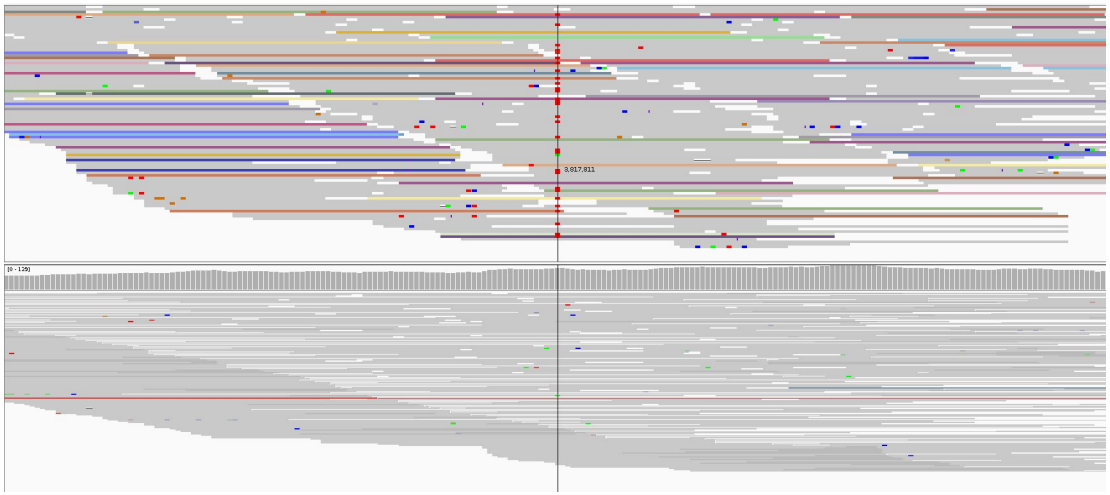
Missense mutation in M1, CESC



Silence mutation in M1, CESC
Filtered by map_qual;weak_evidence

chr	start	end	strand	score	match	mis-match	rep. mat...	N's	Q gap c...	Q gap b...	T gap co...	T gap b...
chr7	151932...	151933...	-	1000	100	0	0	0	0	0	0	0
chr2	91899840	91899940	-	940	97	3	0	0	0	0	0	0
chr1	148876...	148876...	+	940	97	3	0	0	0	0	0	0
chr21	11026791	11026891	-	920	96	4	0	0	0	0	0	0
chrUn_g...	163518	163618	+	920	96	4	0	0	0	0	0	0

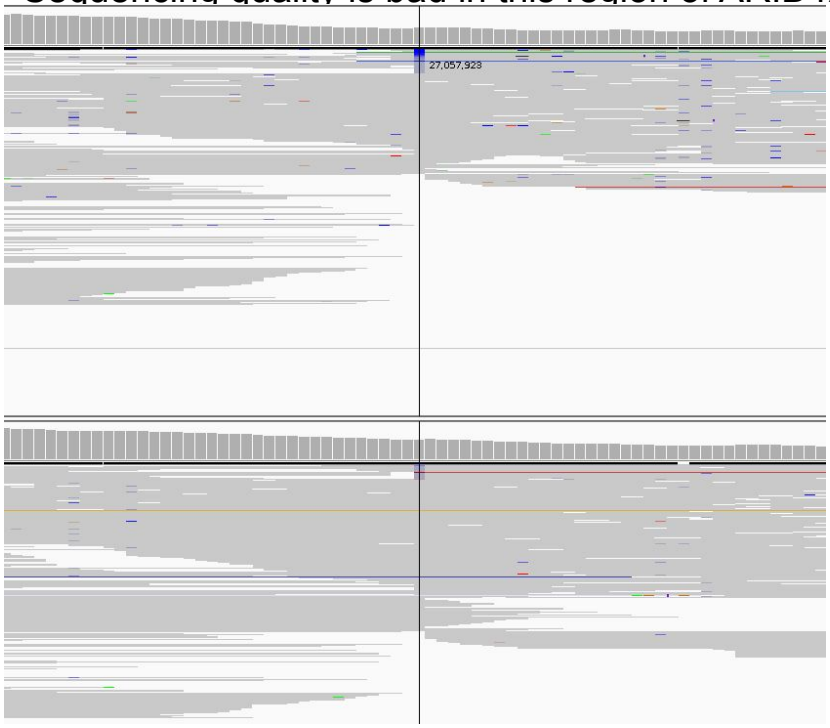
Prefilter makes a difference

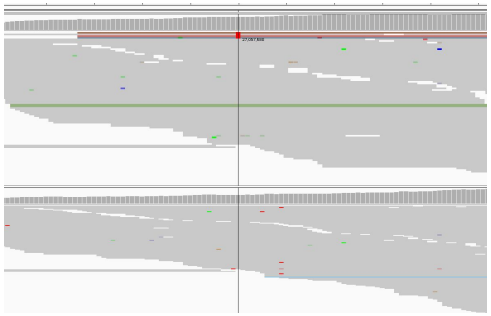


M2 calls a lot from 5a2e564d-6de3-4809-bd8c-0a591196ef4, and notably there are lots of reads aligned to different contig in tumor sample compared to normal - which would result in big difference from prefilter. Is this a hypermutation patient? Will a hypermutated patient disrupts our background model in MutSig?

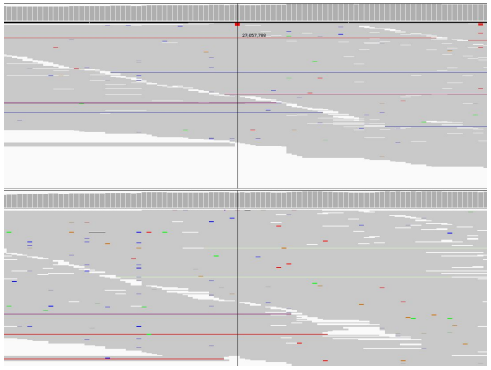
BaseQ: ARID1A sequencing error?

Sequencing quality is bad in this region of ARID1A, but M1 still calls the variant.





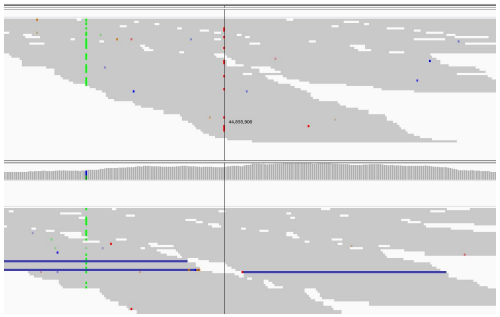
Missense called by M1, BRCA



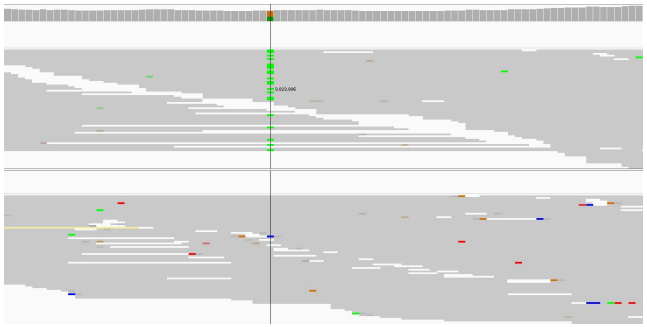
Missense called by M2, BRCA

Missense called by M1, BRCA

M1 reasons for removing variants is not very clear



M1 missed KDM6A(CESC)



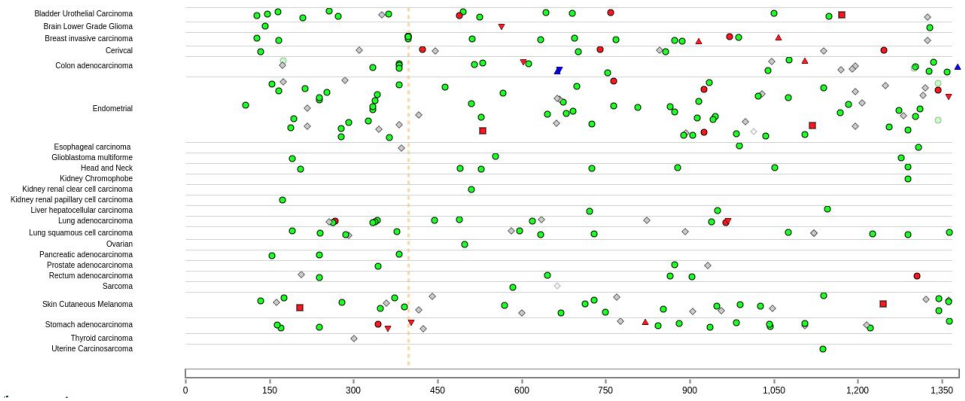
M1 missed CREBBP(CESC)

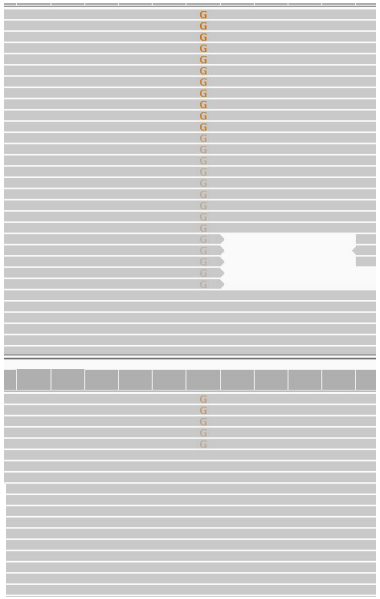
There are more examples (have not snapshot)
Should I go for the PoN?

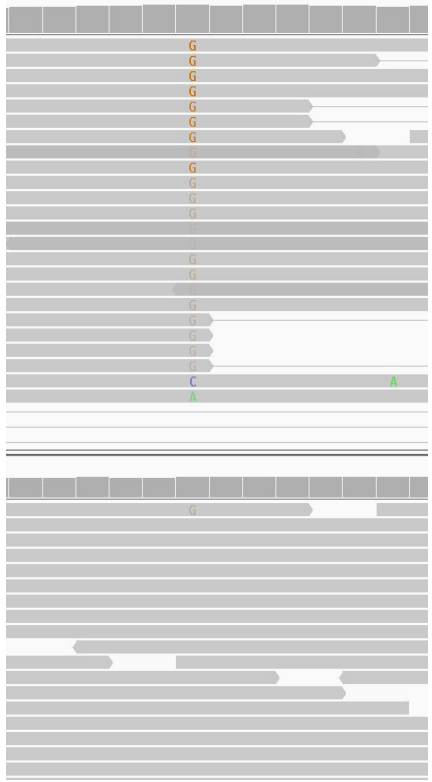
613, filters=nan

chr	start	end	strand	score	match
chr14	38061271	38061335	+	640	64
chr19	46375758	46375821	-	510	57
chr1	47882360	47882398	-	270	30

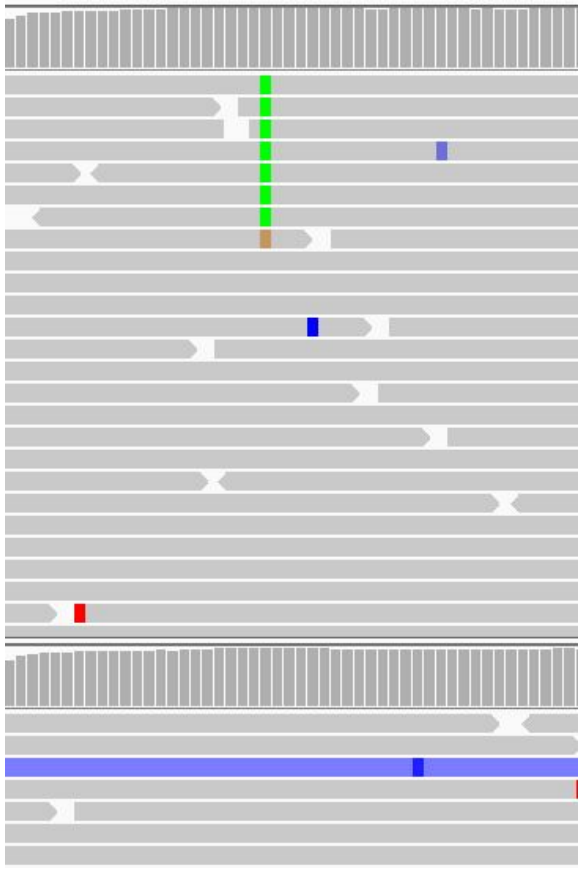
BRCA: 3-124739694 HEG1
effect_M1=mis; effect_M2=nan;
q_M1=0.07067863758599266, q_M2=1.0, filters=nan
position to view: chr3:124,739,694-124,739,695



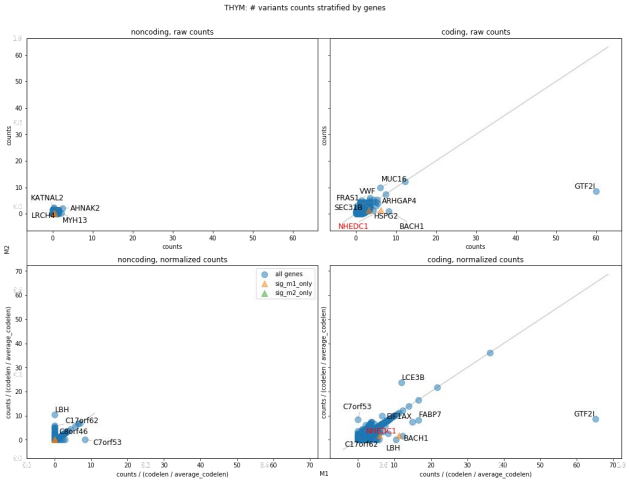




BRCA: 3-124748153 HEG1
effect_M1=mis; effect_M2=nan;
q_M1=0.07067863758599266,
q_M2=1.0, filters=nan
position to view:
chr3:124,748,153-124,748,154

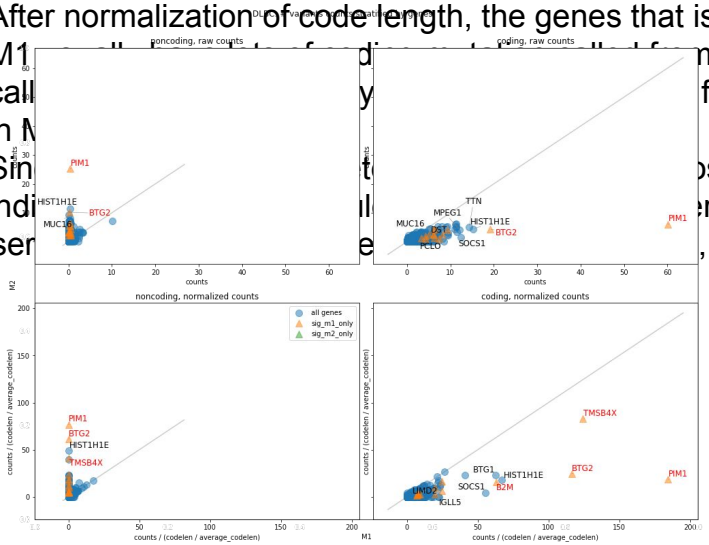


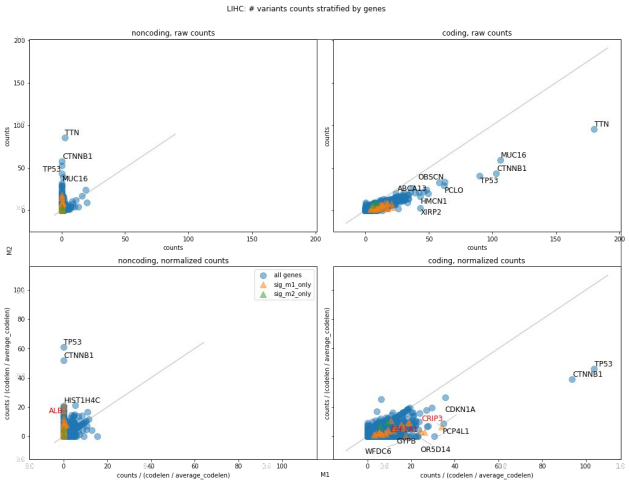
Same gene, different mutational effect



- Overcalling of noncoding mutations in M2 and overcalling of coding mutation in M1 scales with gene length.
- After normalization of code length, the genes that is only significant in M1 [illegible] from M1 while little call[illegible] for them are found in M[illegible]
- Sin[illegible]os, newbase}, this indi[illegible]erence.

The obser[illegible], LIHC





Visualize reads in IGV

A potential drawback of IGV visualization is that we are not viewing what the callers are seeing. Both have different rules for prefilter; M2 identifies active region and performs realignment.

We started from known drivers that is only considered significant from one callset, since those calls have better prior to be true events.

Take-away

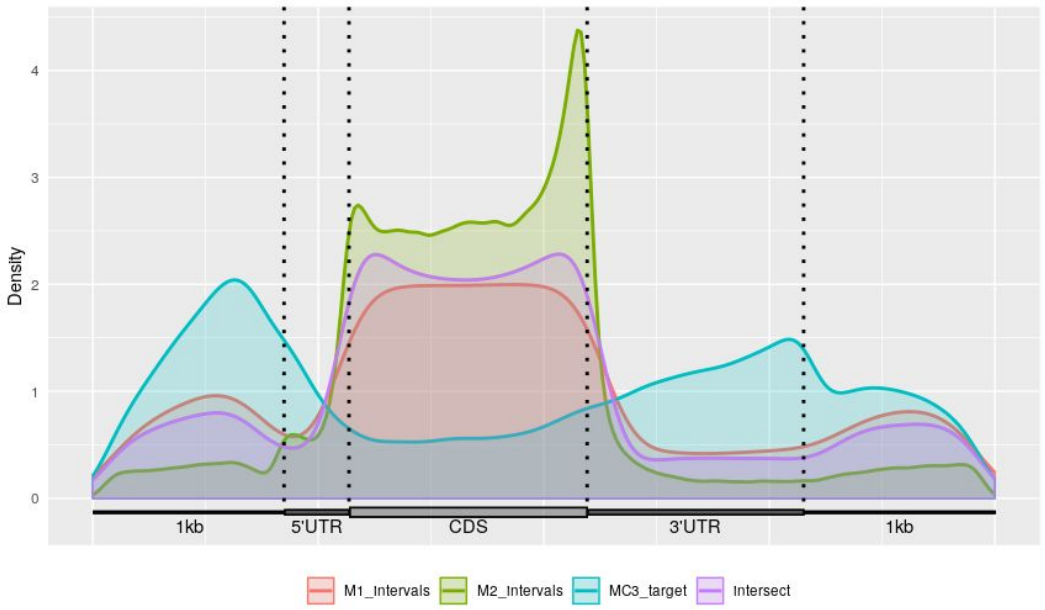
1. Every caller has three steps:
   a. preprocess of bam
   b. somatic calling model
   c. postfilers.

   Our goal is to characterize the differences and find the cause.
2. **M2 is not that bad**. It generally aligns well when stratified by patients, but there are some systemic difference when stratified by mutation effect. Significant disagreement arises when M1 calls more missense events, while M2 calls more IGR events - which is profound in **{DLBC, LIHC, CESC} cohorts.**
3. **A closer look at {DLBC, LIHC, CESC}** reveals that M2 calls lots of noncoding mutation and ignores many coding mutations called from M1, which fails MutSig for driver discovery. Scatter plot of coding vs noncoding mutations indicate some discrepancies are due to realignment.
4. Still need to explain why MutSig identifies so much sig genes from M2.
5. Other cohorts still keep the linear trend of mutation counts stratified by mutation effect, although the slope might be slightly off from 1.
6. Filtering reasons are only responsible for 30-40% M1-only calls, the rest are due to realignment and somatic model itself.

Characterization of interval list

GATK best practice uses bait list, while M1 calls were using a target list.
M2 intervals seems to have an interesting enrichment toward 3‘ UTR and has more transcript-specific signals.

Variant comparison is based on intersection of M1/2 intervals and MC3 target.



M1 Interval	whole_exome_agilent_1.1_refseq_plus_3_boosters_plus_10bp_padding_minus_mito.Homo_sapiens_assembly19.**targets**.interval_list
M2 interval	whole_exome_agilent_1.1_refseq_plus_3_boosters.Homo_sapiens_assembly19.**baits**.interval_list

Cui X, “Guitar: an R/Bioconductor package for gene annotation guided transcriptomic analysis of RNA related genomic features.”

Somatic likelihood model: M2 (2)

$$P(\mathbb{R}, \mathbf{z}, \mathbf{f}|\mathbb{A}) = P(\mathbf{f})P(\mathbf{z}|\mathbf{f})P(\mathbb{R}|\mathbf{z}, \mathbb{A}) = \mathrm{Dir}(\mathbf{f}|\boldsymbol{\alpha}) \prod_a \prod_r (f_a \ell_{ra})^{z_{ra}} .$$

Z_{ra} =1 iff **fragment r** came from allele a∈A

Subset of alleles

P(fragment r | allele a)

Here z (fragment-allele identity, if fragment really comes from allele a) and f (allele fraction) are the latent variable to marginalize to obtain evidence P(R|A). Mean-field approximation is applied (similar as the assumption we made for P(m,f)=P(m)*P(f) in M1) to make it tractable - the mean field for q(z) is a categorical and q(f) is a Dirichlet.
The following two equations are iterated until convergence

$$q(\mathbf{f}) \propto \exp E_{q(\mathbf{z})} [\ln P(\mathbb{R}, \mathbf{z}, \mathbf{f}|\mathbb{A})] \propto \mathrm{Dir}(\mathbf{f}|\boldsymbol{\alpha} + \sum_r \bar{\mathbf{z}}_r) \equiv \mathrm{Dir}(\mathbf{f}|\boldsymbol{\beta}), \quad \boldsymbol{\beta} = \boldsymbol{\alpha} + \sum_r \bar{\mathbf{z}}_r$$

$$q(\mathbf{z}_r) \propto \exp E_{q(\mathbf{f})} [\ln P(\mathbb{R}, \mathbf{z}, \mathbf{f}|\mathbb{A})] \propto \prod_a \left(\tilde{f}_a \ell_{ra}\right)^{z_{ra}},$$

Supp1: Mutect params

```
type_converter = {'Missense_Mutation':'Missense',
                  'Intron':'IGR',
                  'Silent':'Synonymous',
                  'Nonsense_Mutation':'Nonsense',
                  'Splice_Site':'Splice_site',
                  'RNA':'IGR',
                  "3'UTR": 'IGR',
                  "5'UTR": 'IGR',
                  "5'Flank": 'IGR',
                  'DE_NOVO_START_IN_FRAME':'IGR',
                  'DE_NOVO_START_OUT_FRAME':'IGR',
                  'Nonstop_Mutation':'Nonsense',
                  'START_CODON_SNP':'Nonsense'
                 }
```

	M1	M2
Interval list	whole_exome_agilent_1.1_refseq_plus_3_boosters_plus_10bp_padding_minus_mito.Homo_sapiens_assembly19.**targets**.interval_list	whole_exome_agilent_1.1_refseq_plus_3_boosters.Homo_sapiens_assembly19.**baits**.interval_list
Panel-of-Normal	controlled_access_token_pon_from_tcga8000.final_summed_tokens.hist.bin	