Comparing and combining mutation callers

Terry Speed, WEHI

Joint work in the Department of Statistics at UC Berkeley, with

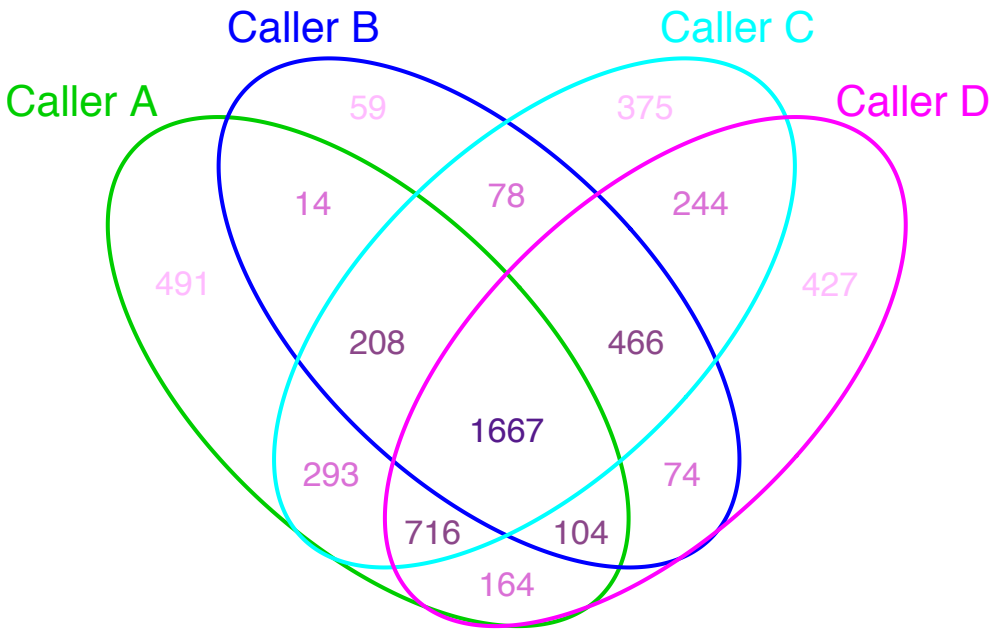
Su Yeon Kim (now at Veracyte) and Laurent Jacob (now at CNRS, Lyon)

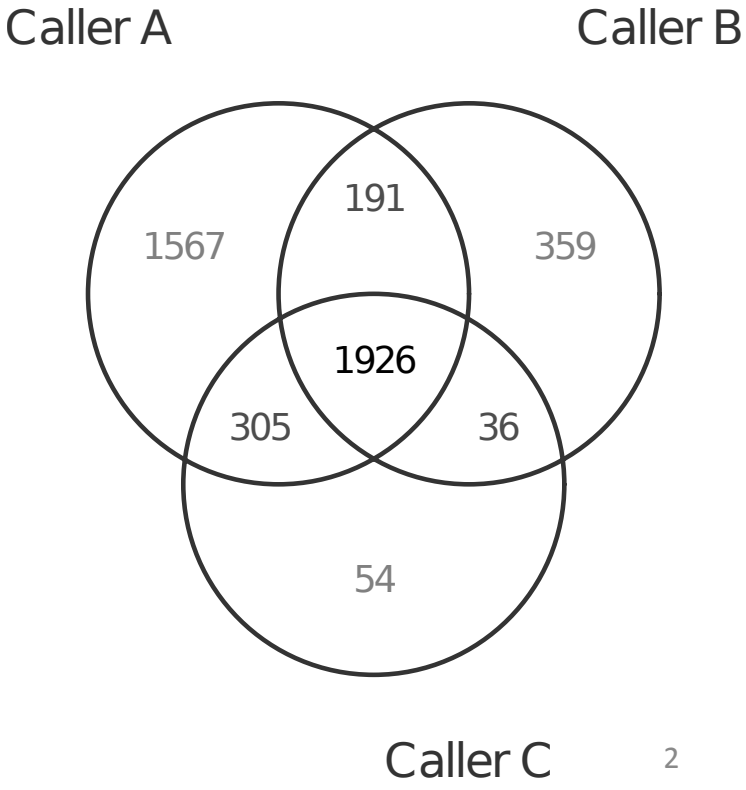
Motivation, funding, data and much assistance from

The Cancer Genome Atlas (TCGA)

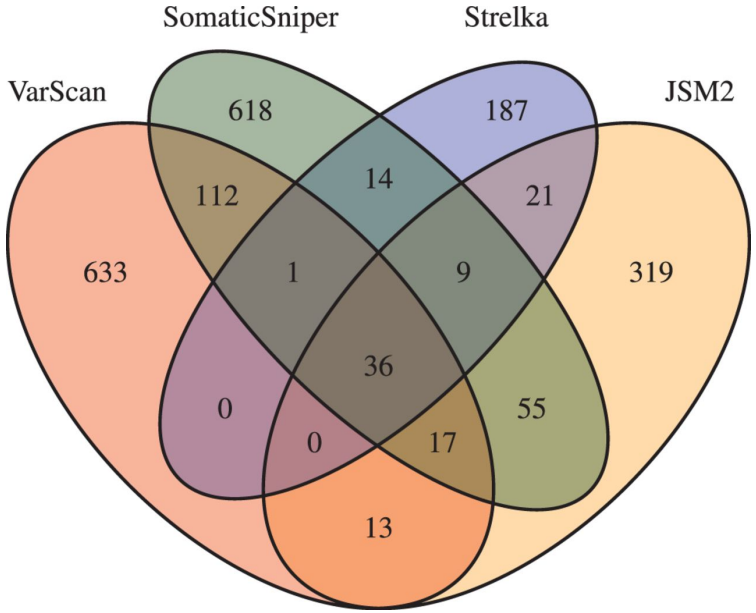
This talk is about going beyond Venn diagrams for comparing call-no call algorithms

Look at these diagrams. We see disturbing levels of disagreement, but what can we learn from them?

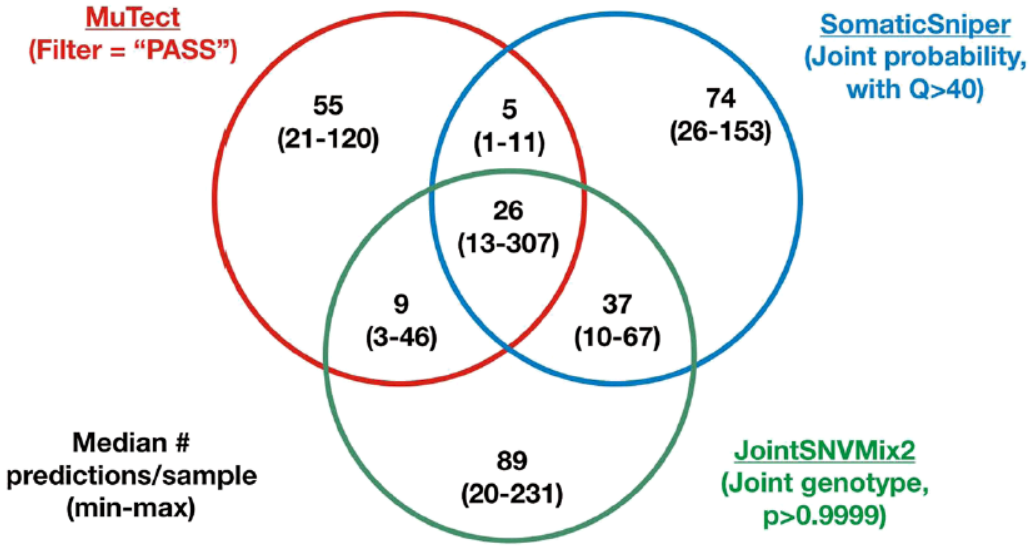




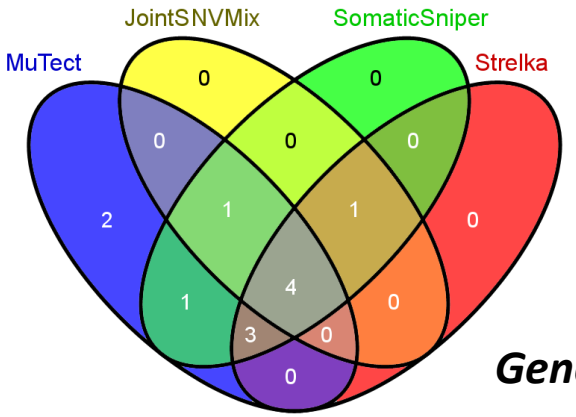
Published Venn diagrams on the same topic



Roberts *et al* ***Bioinformatics*** 2013



Goode *et al* ***Genomic Medicine*** 2013

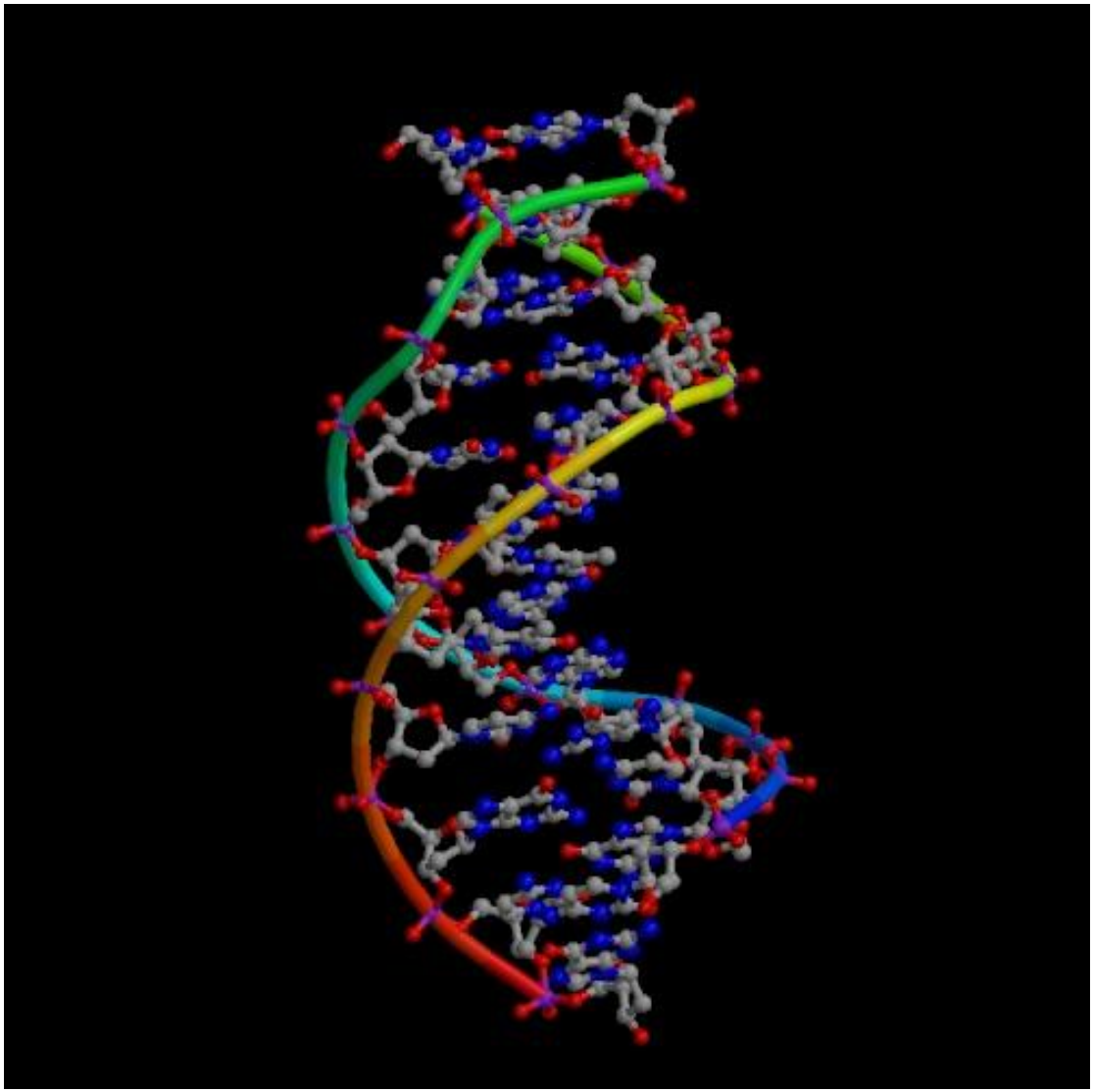


Wang *et al*
Genomic Medicine 2013

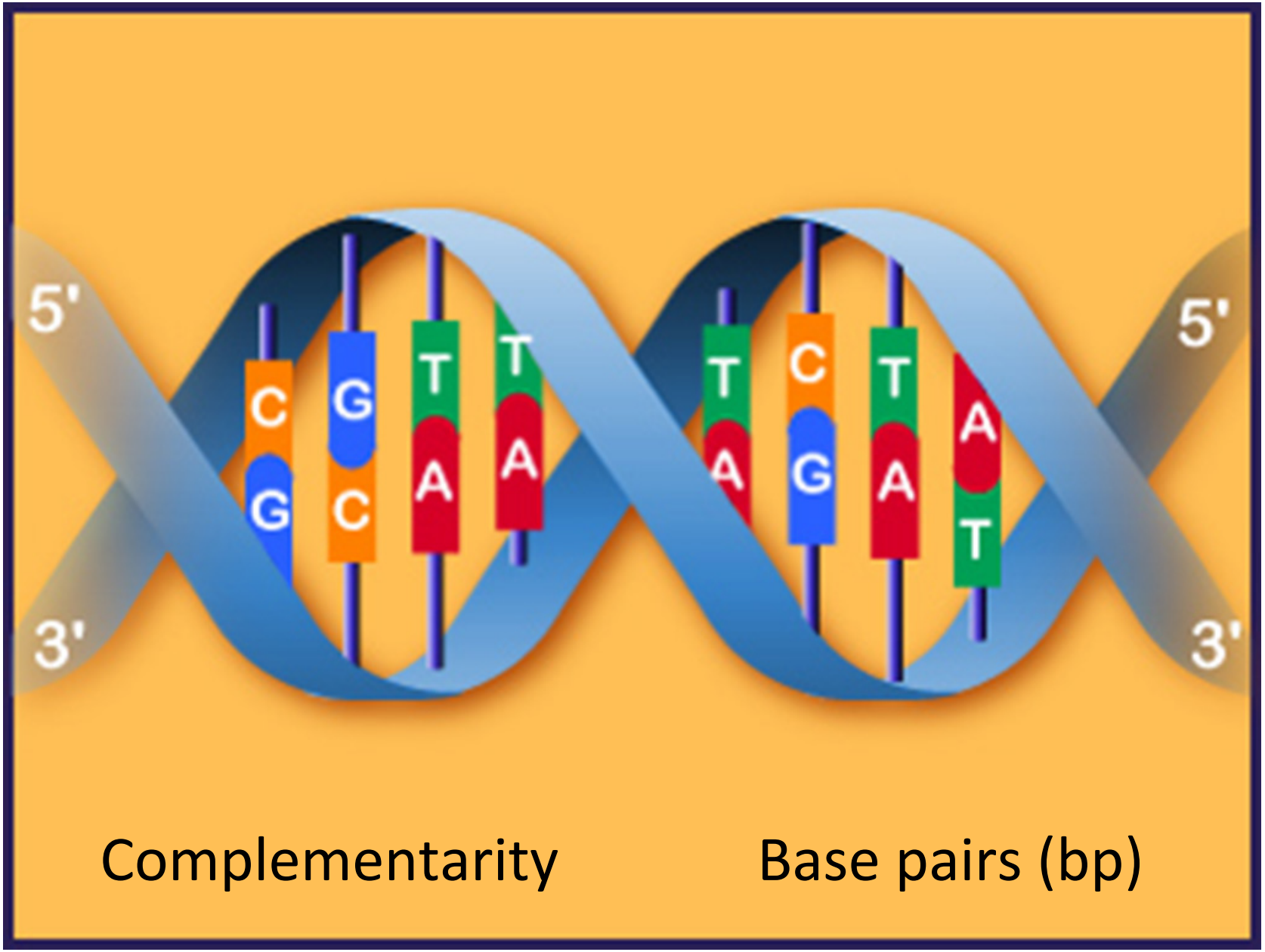
Aim and Synopsis

The purpose of this talk is to introduce you to some of the challenging statistical problems that arise in the analysis of cancer genome data, and **go beyond Venn diagrams**. The new material is published, and so people wanting more detail can go to the papers (see end). I'll spend much of my time setting the scene, and just sample the results.

- Some very basic molecular genetics
- Background to this talk
- The technology, data and algorithms
- Comparing mutation callers
- Combining mutation callers
- References

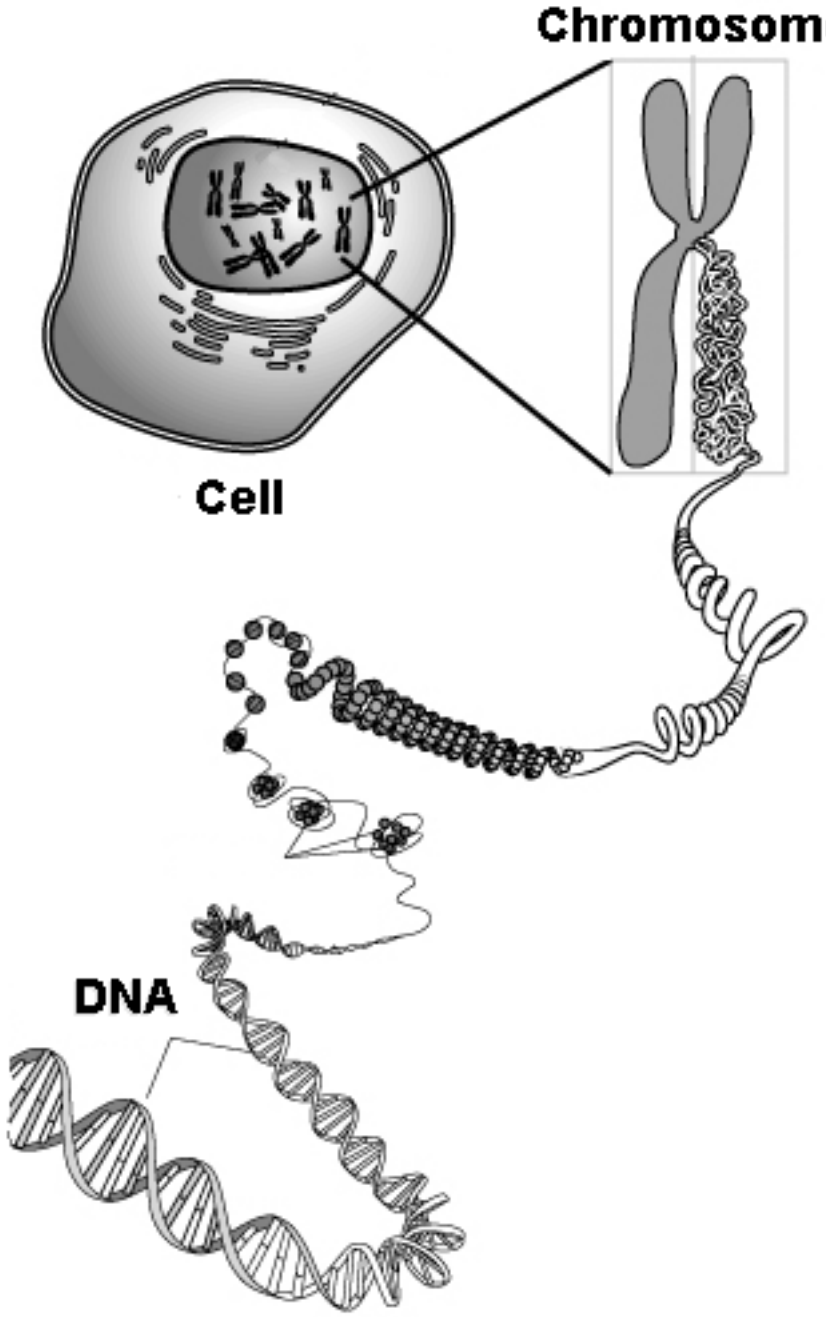


Nucleotide = unit of DNA = base + sugar + phosphate



Complementarity Base pairs (bp)

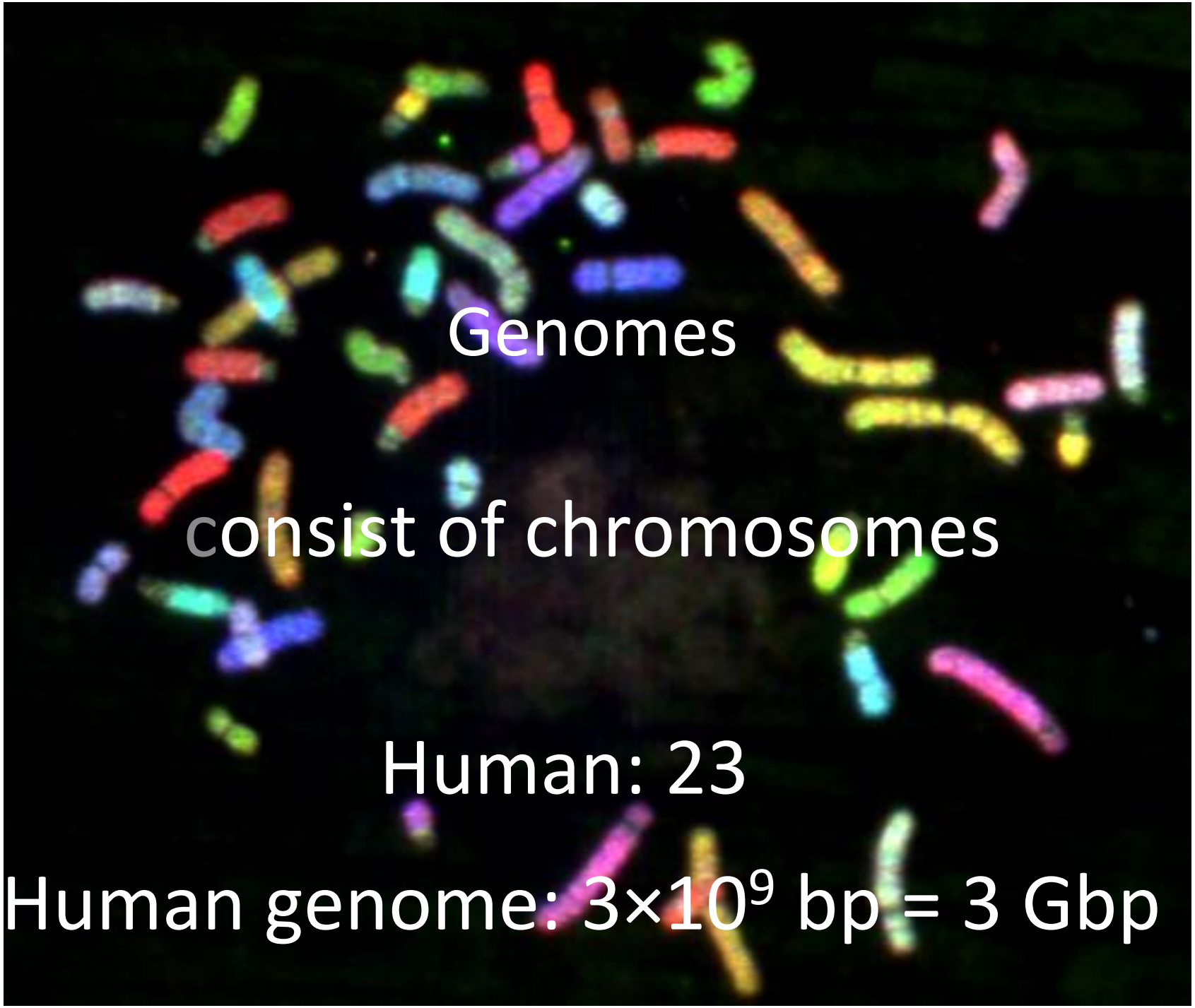
aaatatattc aatatggaga gaataaaaga actgagagat ctaatgtcgc
tcgcgagata ctcactaaga ccactgtgga ccatatggcc ataatcaaaa
aggaaggcaa gagaagaacc ccgcactcag aatgaagtgg atgatggcaa
aattacagca gacaagagaa taatggacat gattccagag aggaatgaac
cctctggagc aaaacaaacg atgctggatc agaccgagtg atggtatcac
aacatggtgg aataggaatg gcccaacaac aagtacagtt cattacccta
aacttatttc gaaaaggtcg aaaggttgaa acatggtacc ttcggccctg
aaatcaagtt aaaataagga ggagagttga tacaaaccct ggccatgcag
caaggaggca caggatgtga ttatggaagt tgttttccca aatgaagtgg
actgacatca gagtcacagc tggcaataac aaaagagaag aaagaagagc
taaaatt 1,000 bp = 1 kilobasepair = 1 kbp of DNA gaattggtcc
gtttctccca gtagccggcg gaacaggcag tgtttatatt gaagtgttgc
agggacgtgc tgggagcaga tgtacactcc aggaggagaa gtgagaaatg
ccaaagtttg attatcgctg ctagaaacat agtaagaaga gcagcagtgt
attagcatct ctcttggaaa tgtgccacag cacacagatt ggaggagtaa
catccttaga cagaatccaa ctgaggaaca agccgtagac atatgcaagg
gttgaggatt agctcatctt tcagttttgg tgggttcact ttcaaaagga
atcagtcaag aaagaagaag aagtgctaac gggcaacctc caaacactga
acatgaaggg tatgaagaat tcacaatggt tgggagaaga gcaacagcta
ggcaaccagg agattgatcc agttgatagt aagcgggaga gacgagcagt

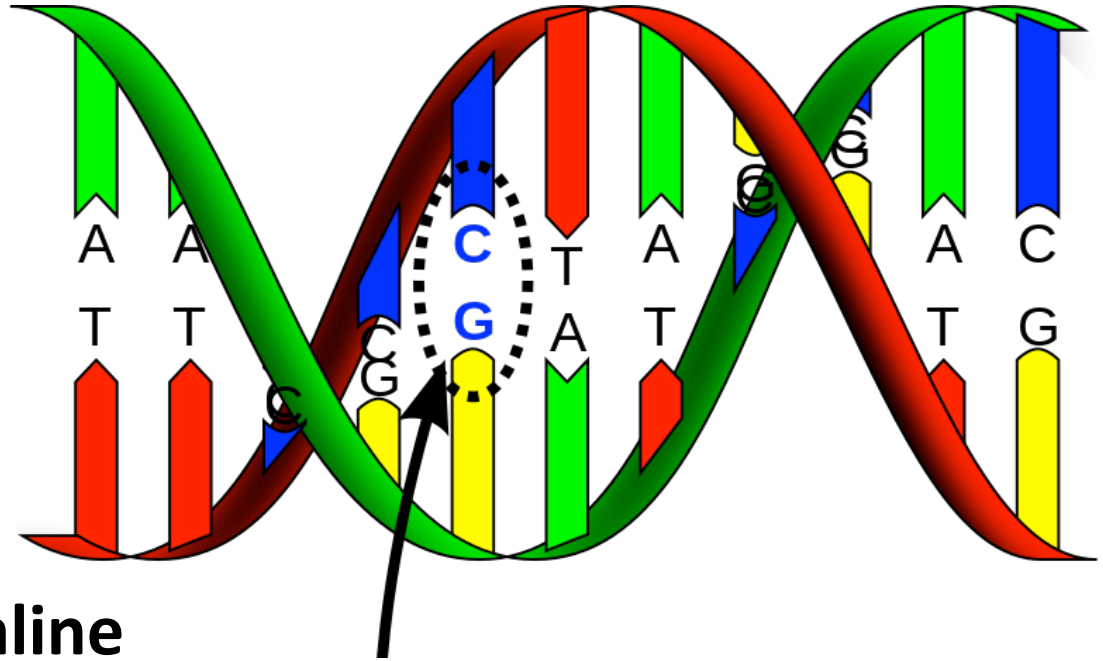


Chromosomes are long DNA molecules (there are exceptions)

Human chromosomes: tens to hundreds of millions of base pairs

av ~150 Mbp

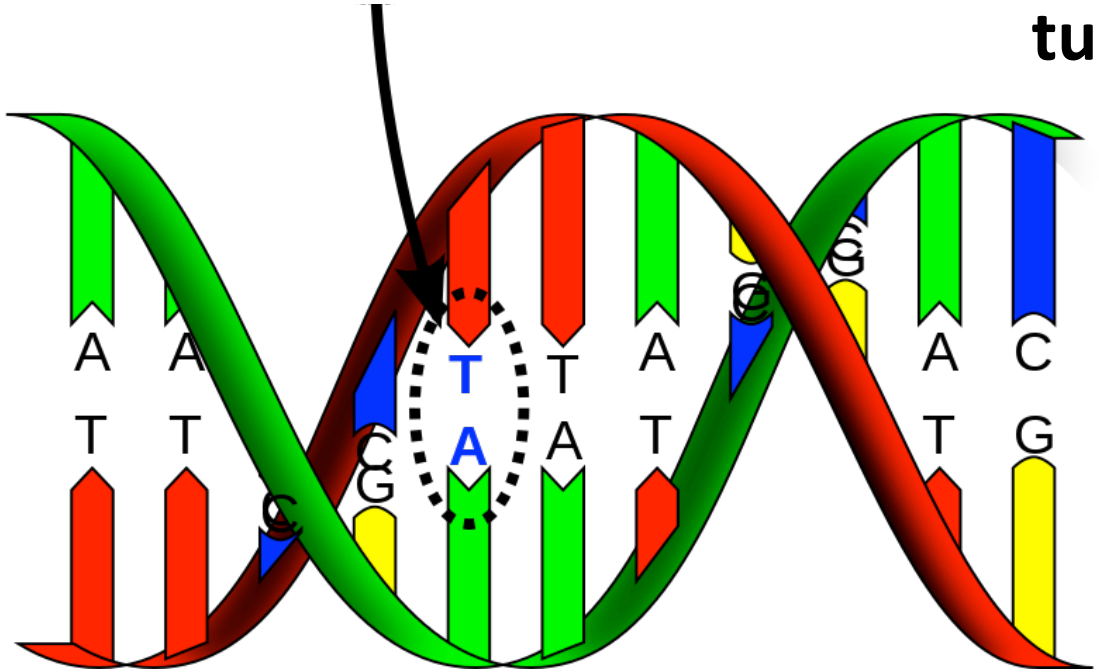

Genomes
consist of chromosomes
Human: 23
Human genome: 3×10^9 bp = 3 Gbp



Germline

- single nucleotide substitution -

tumor



Background to this talk

The problem

Single nucleotide substitutions relative to the germline genome are an important and common feature of tumor genomes. We'll call these **somatic mutations**.

germline = what you are born with

somatic = "of the body", develops later

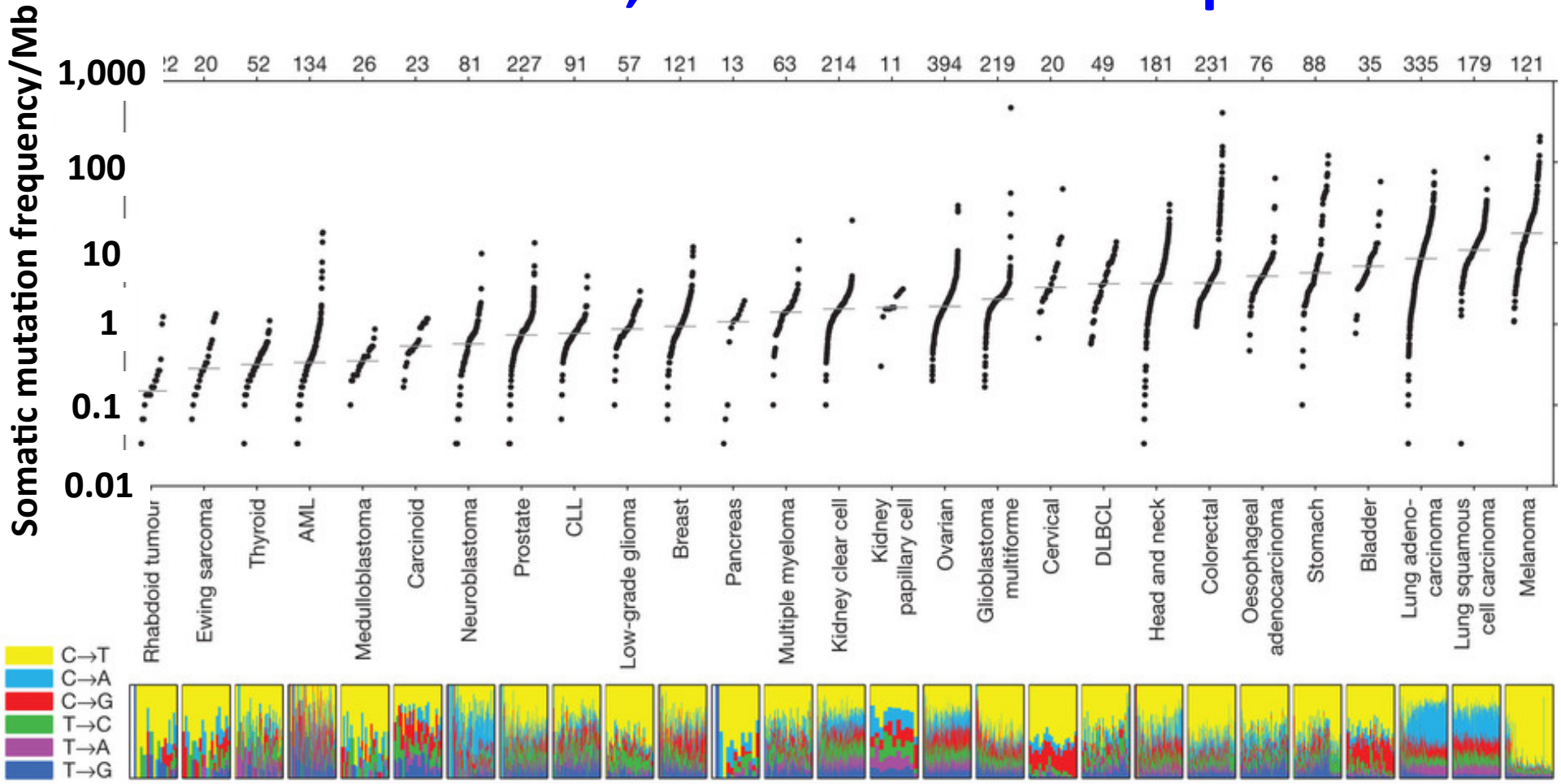
There are 3 billion possible locations for a somatic mutation, and people want to find them in tumors.

(Why? See later.)

Why is mutation-detection hard, I?

- Somatic mutations are rare, ~ 1 in a million.

Somatic mutation frequencies observed in exomes from 3,083 tumor–normal pairs.

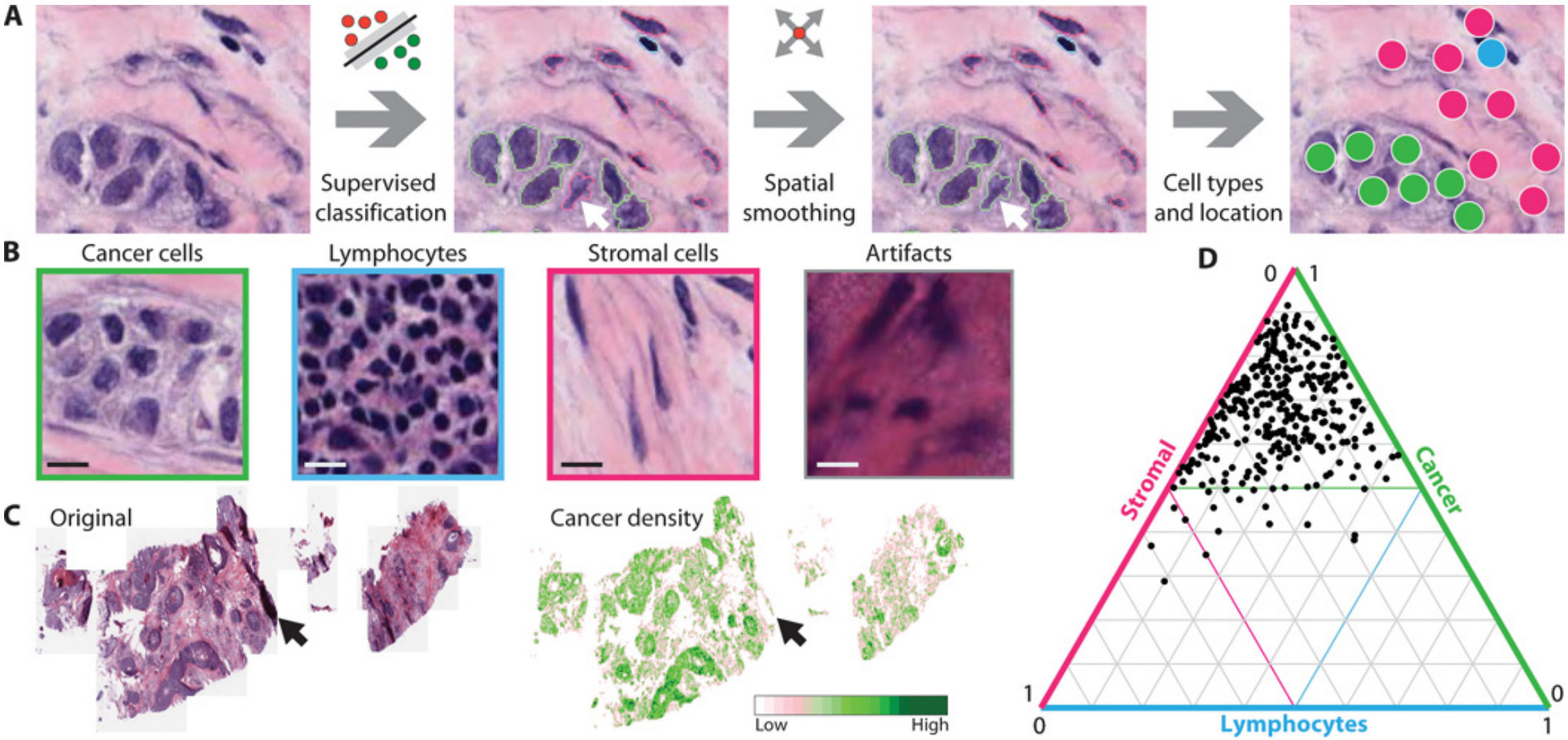


The human exome (see later) is ~30Mb, so multiply the numbers above by 30 to get the total per tumor exome. Lawrence *et al* ***Nature*** 2013

Why is mutation-detection hard, I?

- Somatic mutations are rare, ~ 1 in a million.
- The tumor tissue whose DNA we sequence is invariably contaminated with non-tumor cells, having germline (normal) DNA.

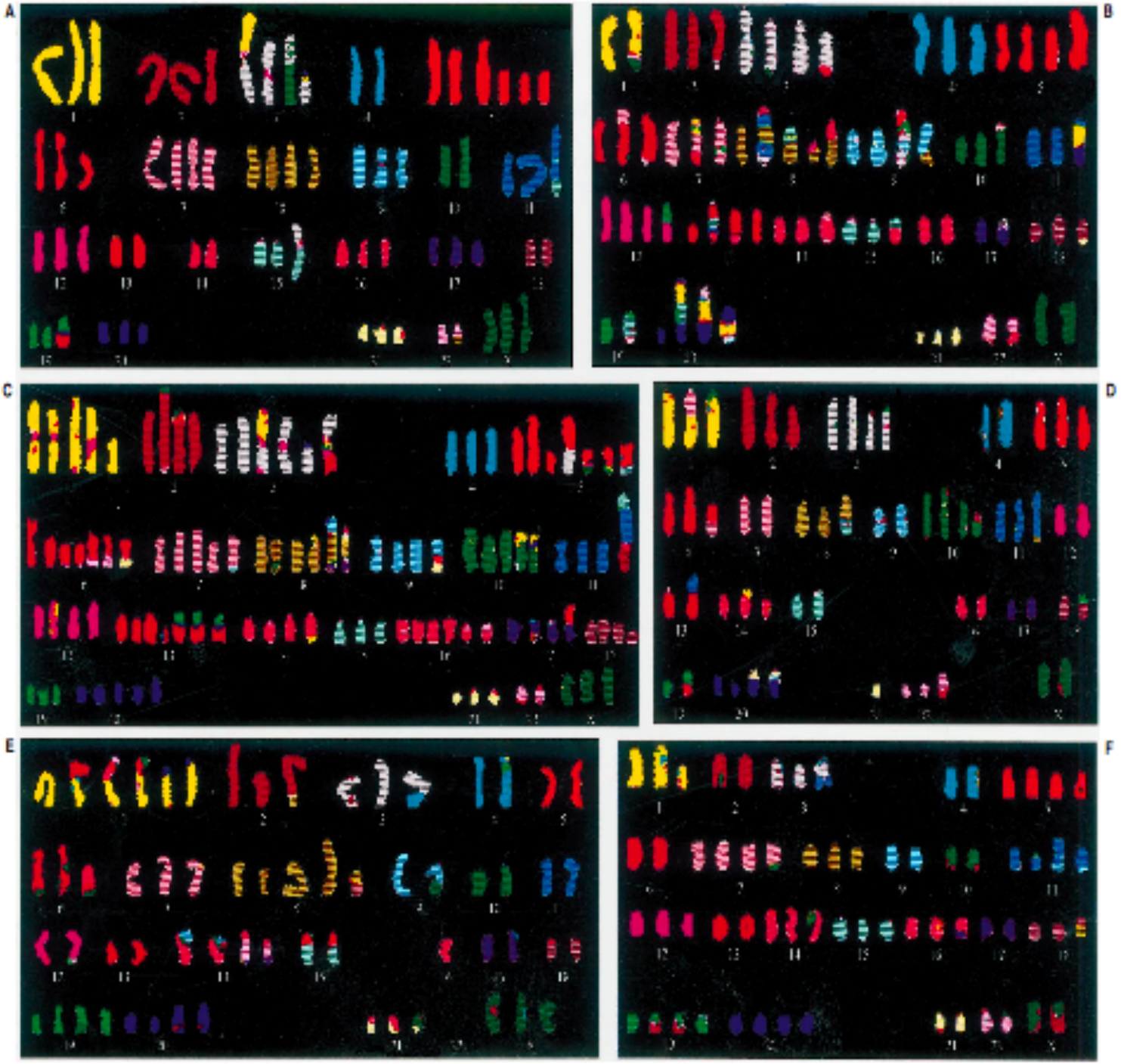
Normal cells "contaminating" a tumor



Yuan *et al,* ***Sci Transl Med*** (2012);

Why is mutation-detection hard, I?

- Somatic mutations are rare, ~ 1 in a million.
- The tumor tissue whose DNA we sequence is invariably contaminated with non-tumor cells, having germline (normal) DNA.
- Tumors often have local copy number aberrations, i.e. regions of the genome with one of both copies lost, and other regions with gains.



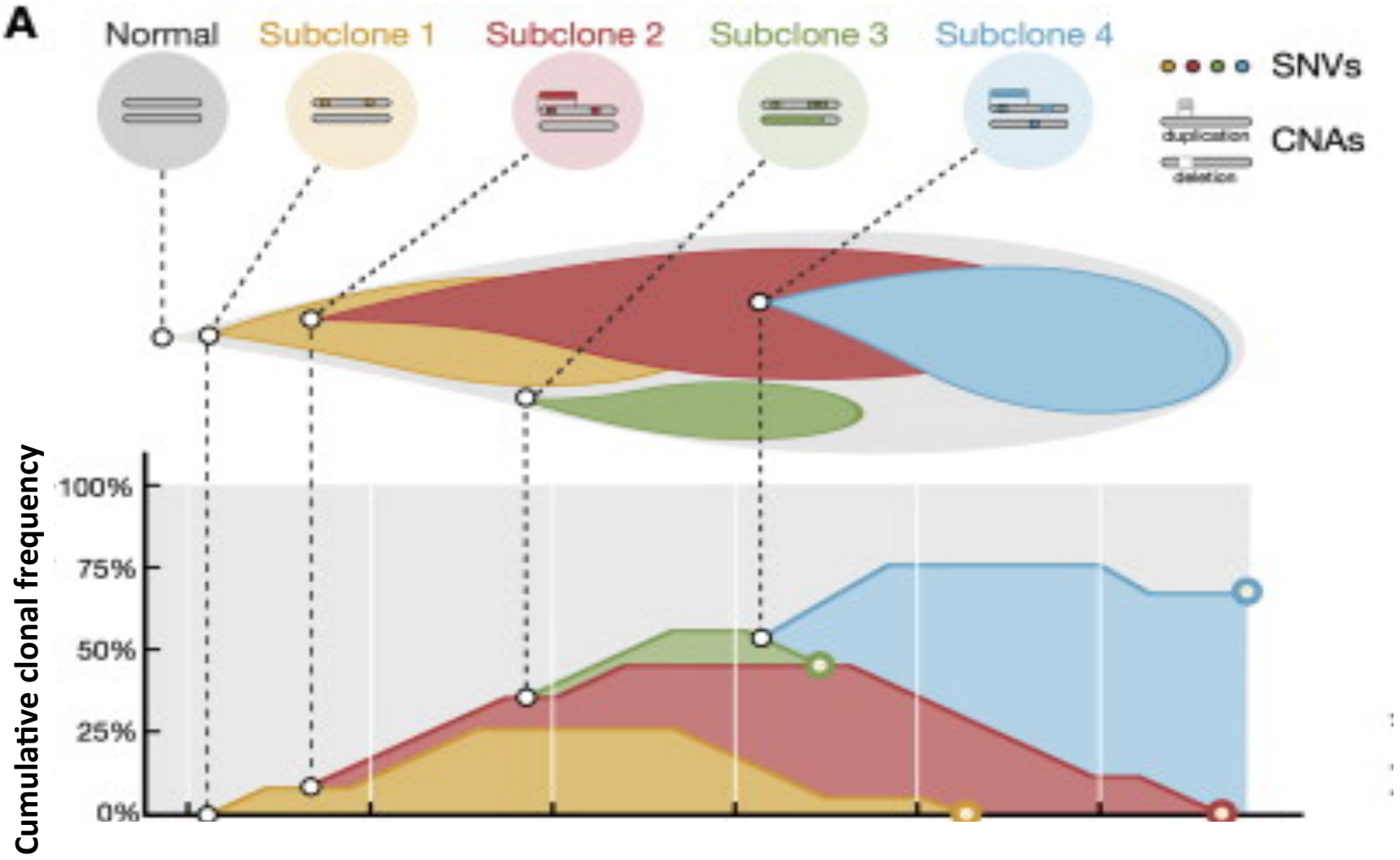
24 color karyotypes of 6 BrCa cell lines

Davidson *et al* BrJCa, 2000

Why is mutation-detection hard, I?

- Somatic mutations are rare, ~ 1 in a million.
- The tumor tissue whose DNA we sequence is invariably contaminated with non-tumor cells, having germline (normal) DNA.
- Tumors often have local copy number aberrations, i.e. regions of the genome with one of both copies lost, and other regions with gains.
- Tumors are frequently heterogeneous, that is, they harbor distinct subclones.

Tumor heterogeneity = subclonality

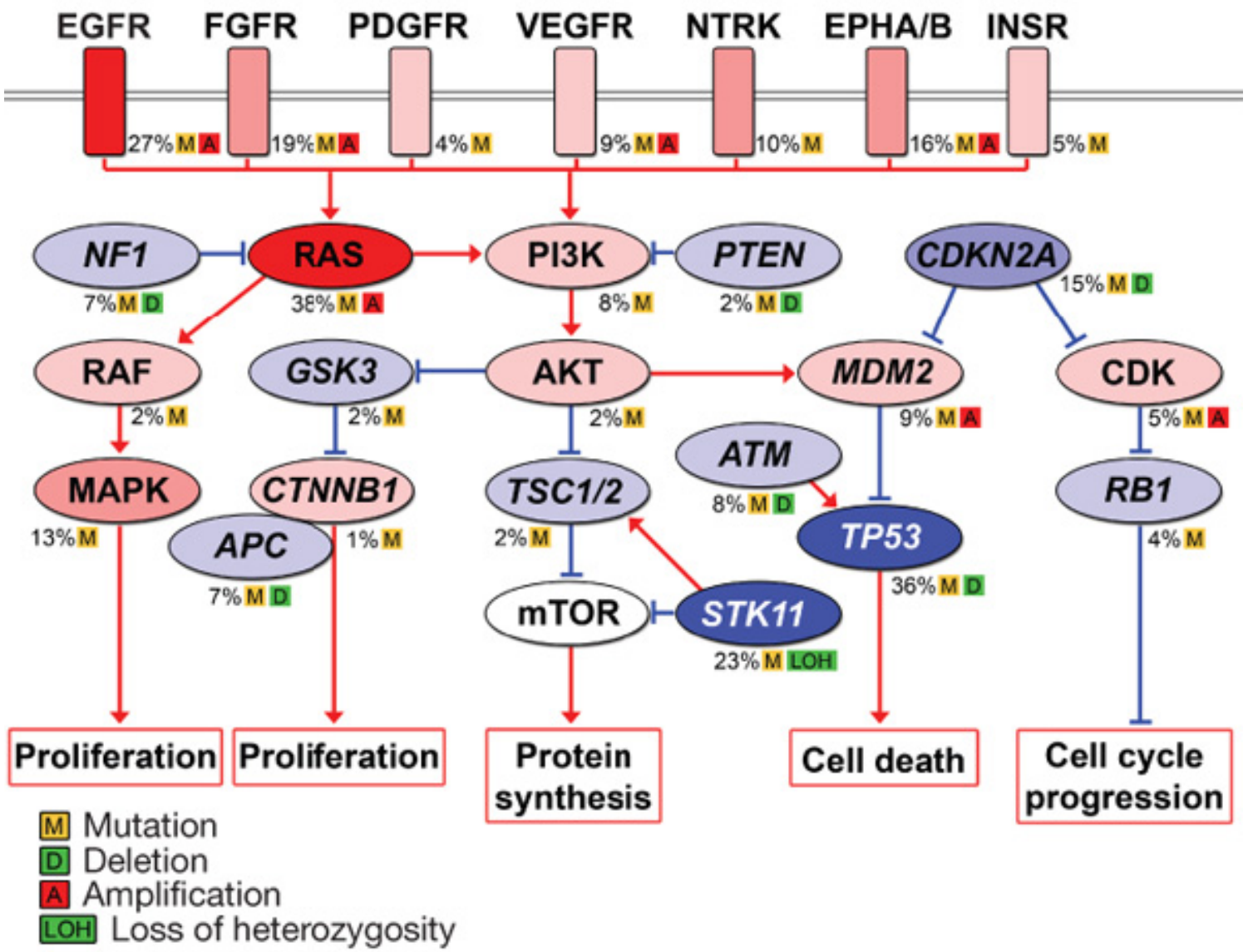


Adapted from Fischer *et al*, Cell Reports 2014

Why do we care?

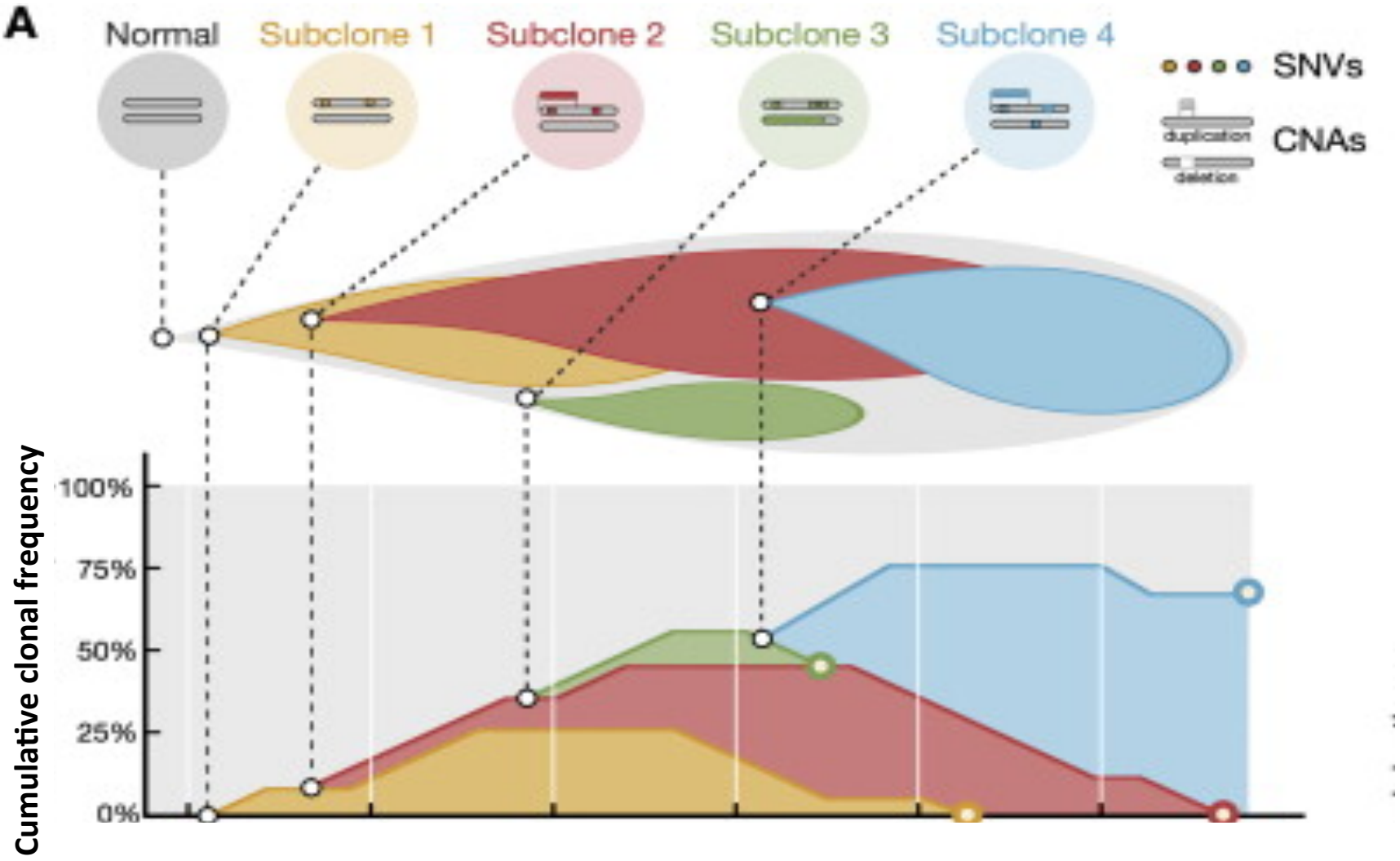
Many large-scale cancer projects are currently scanning for somatic mutations (and other aberrations) in tumors of various types, prior to conducting downstream analyses. These include detecting significantly mutated genes or pathways, inferring clonal history, and characterizing the landscape of the somatic mutations.

Significantly mutated pathways in lung adenocarcinomas



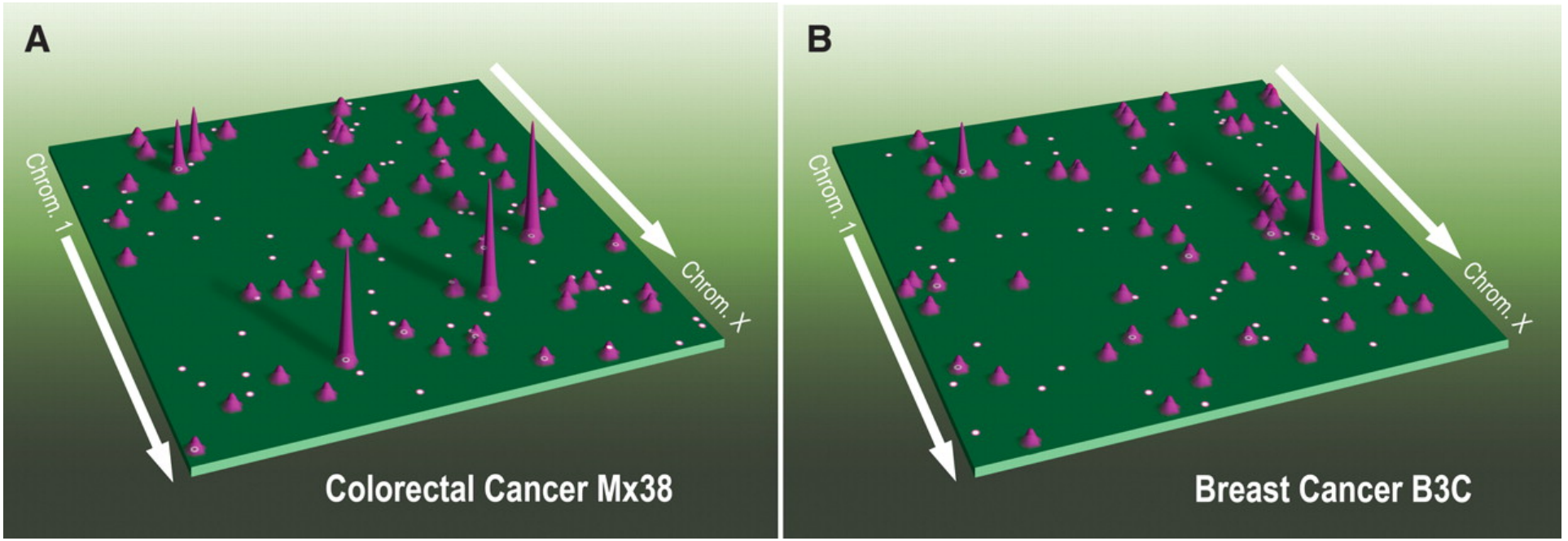
Ding *et al. Nature* **2008**

Inferring subclonality



Adapted from Fischer *et al*, Cell Reports 2014

Cancer genome landscapes



L D Wood *et al.* Science 2007

Why do we care?

Many large-scale cancer projects are currently scanning for somatic mutations (and other aberrations) in tumors of various types, prior to conducting downstream analyses. These include detecting significantly mutated genes or pathways, inferring clonal history, and characterizing the landscape of the somatic mutations.

Why do we care?

Many large-scale cancer projects are currently scanning for somatic mutations (and other aberrations) in tumors of various types, prior to conducting downstream analyses. These include detecting significantly mutated genes or pathways, inferring clonal history, and characterizing the landscape of the somatic mutations.

Some, but by no means all of these genomic aberrations will be responsible for the cancer phenotype, and for metastases. We want to find out which are, for treatment.

The technology, data and algorithms

Whole exome sequencing

The mutation calling process starts with what is known as whole exome sequence data on matched tumor-normal genomic DNA. Algorithms are used to "call" somatic mutations. They are in the tumor, and not the normal. Our starting point is the algorithms' output, called the VCF = variant call format file, see later.

An Illumina HiSeq 2000 similar to the one on which the data we discuss was generated.

Let's visualize some mapped reads, and variants, but first…

What is the exome?

- This term describes the 1-2% of the genome consisting of known protein-coding genes (plus a bit on the edges).
- It is an abbreviation of ***ex****pressed gen****ome***, but that is now outdated, as lots more of the genome gets expressed, not just the protein coding genes.
- Nevertheless, the name has stuck. Biologists now know that much more than protein-coding genes is relevant to cancer, but these insights are recent.
- Partly as a result of this, partly because of cost, most scanning for somatic mutations is restricted to the traditional exome.
- The genome is ~3Gbp, and so the exome is between 30Mbp and 60Mbp. The alternative is *deep*, **whole genome sequencing**, currently much more expensive.

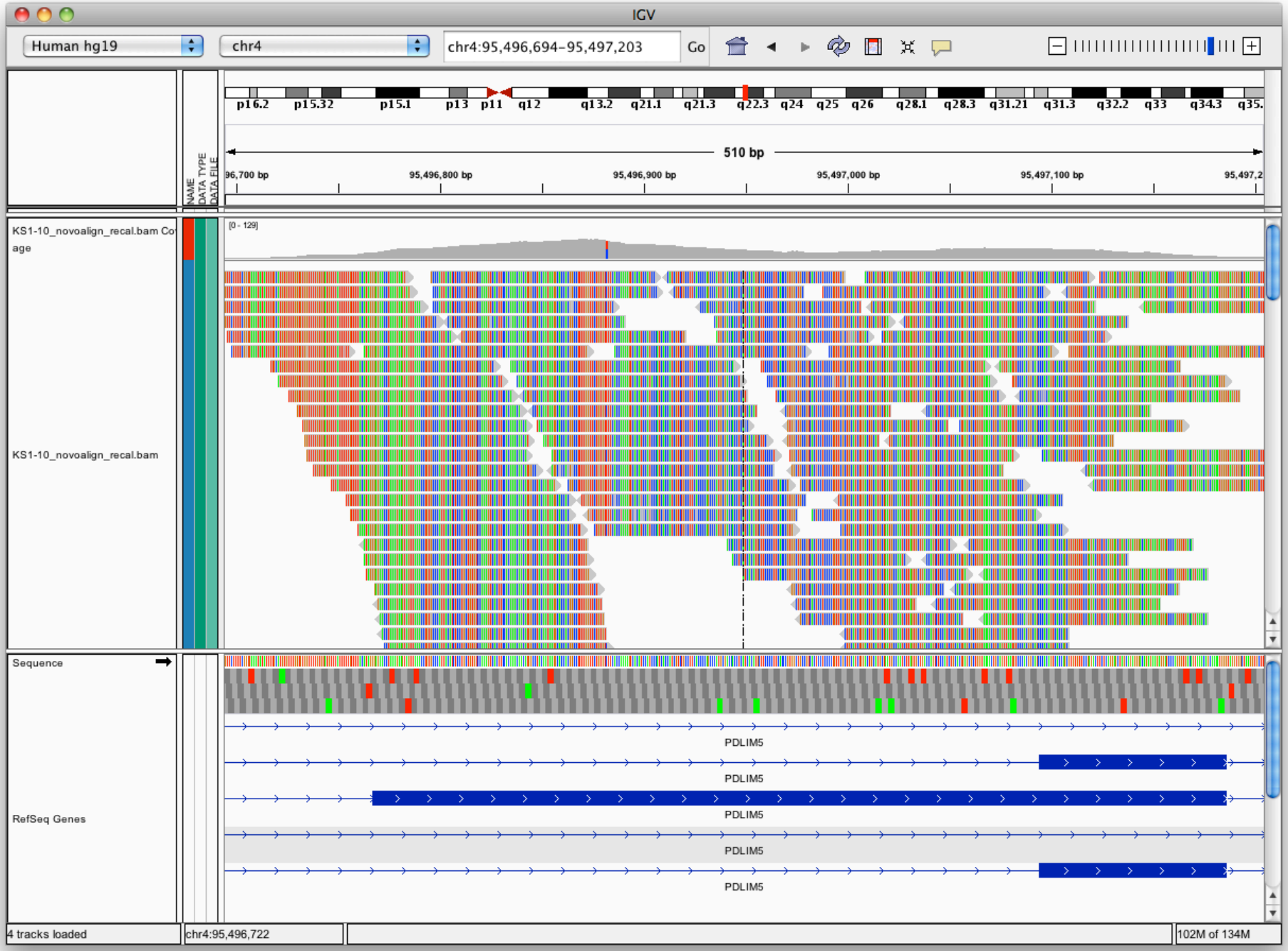


Image produced with Interactive Genomics Viewer http://www.broadinstitute.org/igv/ 22nd June 2012

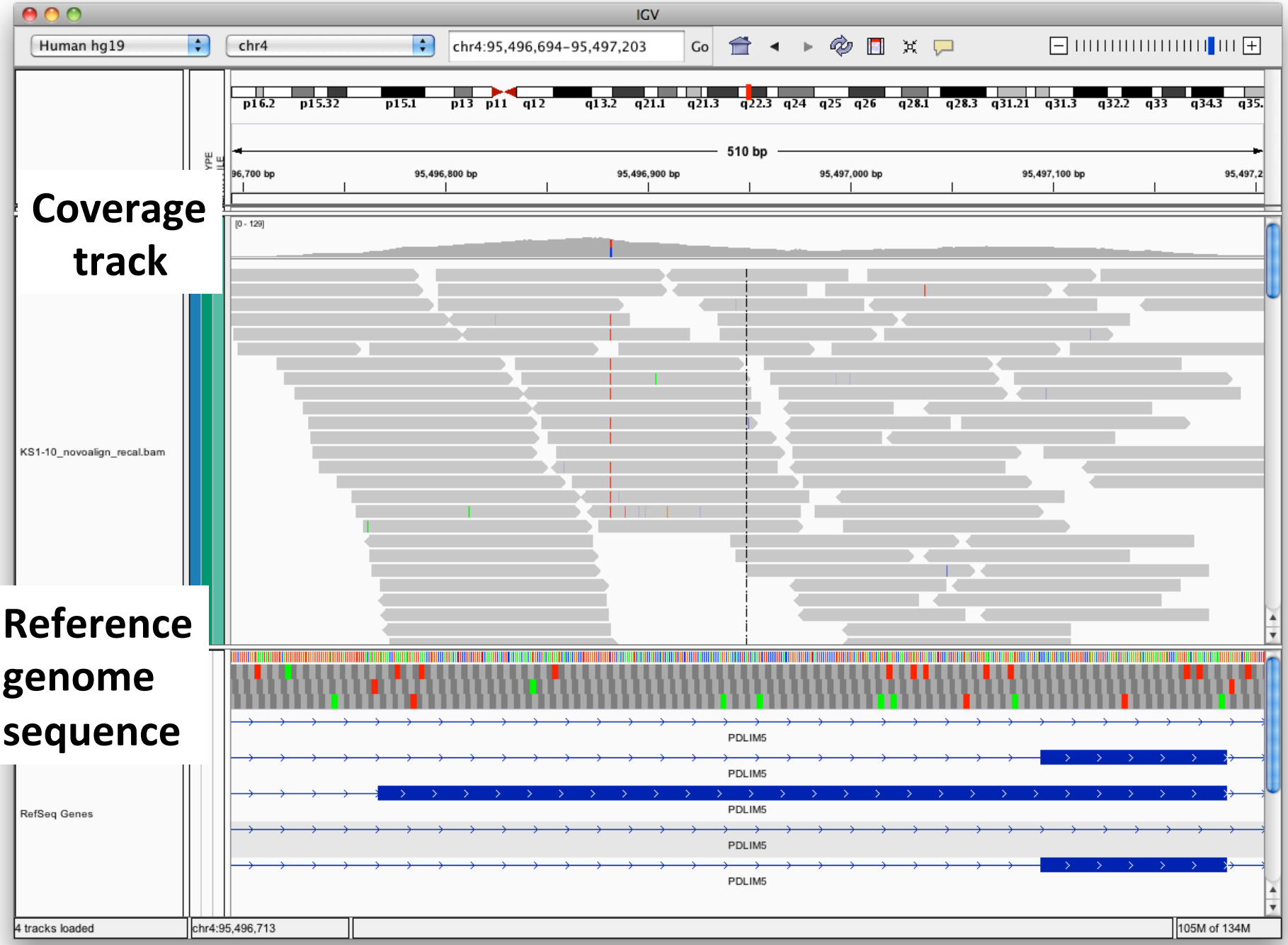


Image produced with Interactive Genomics Viewer http://www.broadinstitute.org/igv/ 22nd June 2012

Here we can see that about half the reads support the reference base **C** (not marked) and the other half support the alternate base **T** at position 95,496,882. Scattered calling errors can also be seen.

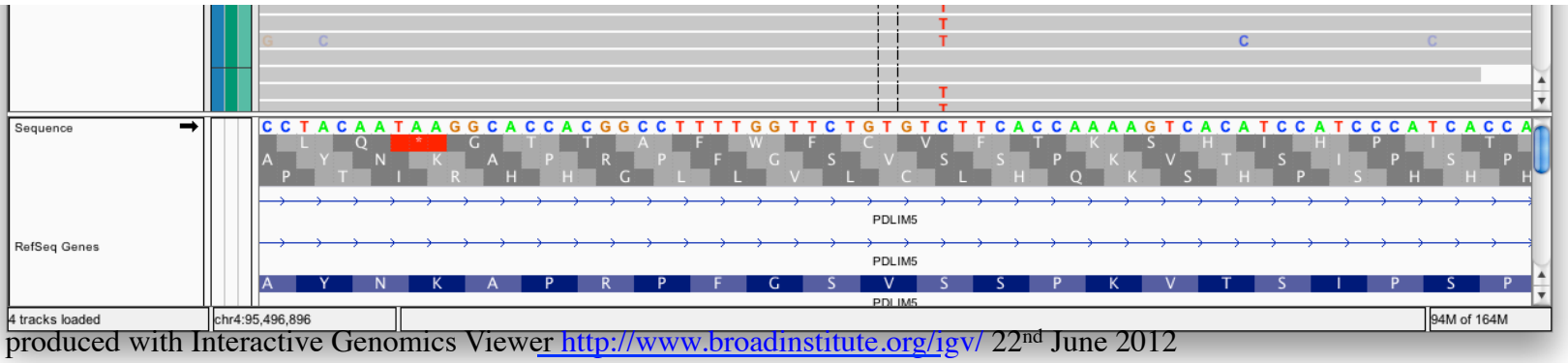


Image produced with Interactive Genomics Viewer http://www.broadinstitute.org/igv/ 22nd June 2012

Variant allele fraction (vaf)

Detecting a variant in an aligned sequence is looking for the existence of a *variant base* that is different from the *reference base*. In principle, the more reads carrying the variant allele, the stronger the evidence for it being a true variant. Thus, the fraction of reads carrying the variant allele (the variant allele fraction, *vaf*) is frequently used in variant calling analyses.

For somatic mutation-calling, the tumor and its matched normal sample are considered together. Therefore, a variant is determined by the joint status in tumor-normal sequence pairs:

- **somatic**: the variant is found in the tumor but not in the normal
- **germline**: variant found in both the tumor and the normal
- **wildtype**: no variant found in either the tumor or the normal

Why is mutation detection hard, II?

Finding mutations is challenging, even with high-throughput sequencing technology.

Coverage of the exome can be highly variable, even when we have high average coverage.

Artifacts can appear during targeted capture or PCR amplification, machine sequencing errors occur, as do incorrect local alignments of reads.

All in all, it's a hard problem, and so many methods have been proposed to solve it. How do they compare?

Mutation-calling algorithms

Several are published and more are sure to appear. Here are the names of some:

Strelka: Saunders *et al*, ***Bioinformatics*** 2012

VarScan2: Koboldt *et al*, ***Genome Research*** 2012

SomaticSniper: Larson *et al*, ***Bioinformatics*** 2012

JointSNVMix: Roth *et al*, ***Bioinformatics*** 2012

Mutect: Cibulskis *et al*, ***Nature Biotechnology*** 2013

EBCall: Shiraishi *et al*, ***Nucleic Acids Research*** 2013

There are also several unpublished in-house algorithms. In what follows, we'll use a mix of data from older versions of published callers, and unpublished ones. None will be named, as that's not necessary for what we are doing. Besides, the current versions of these algorithms will differ from those leading to the data we discuss, in part because of our results.

What I'm not and what I am discussing

I'm not discussing the inner workings of the mutation detection algorithms, although it will be helpful to know (where possible) which features of the data they use.

Our aims in this talk are simple:

- to try to go beyond Venn diagrams, when comparing callers for which no truth is known, and
- to combine the results of different callers into a better caller when some truth is known.

These are statistical analyses which can be carried out almost independently of the nature of the callers.

Why not just build a caller better than all of the existing ones?

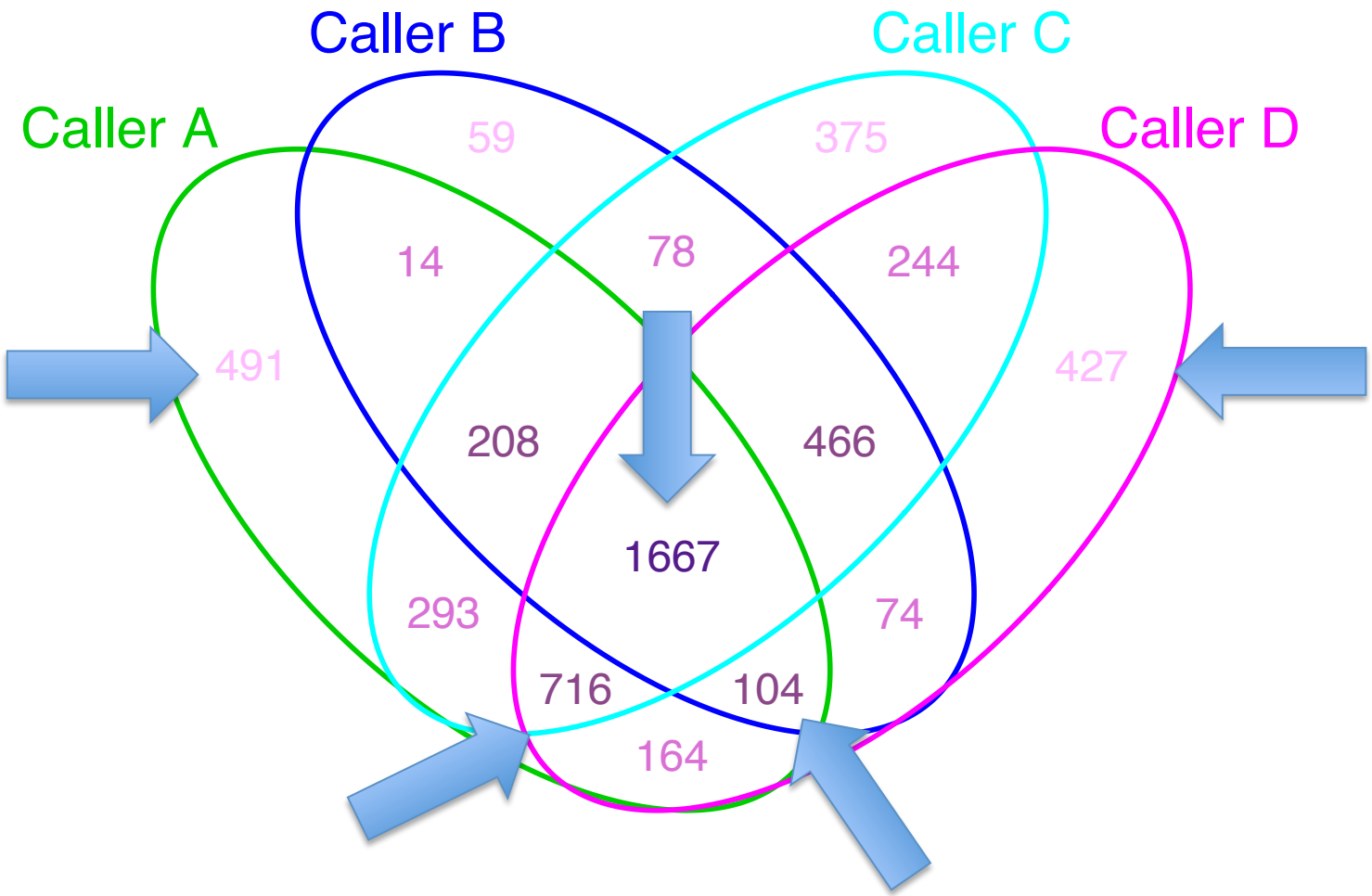
- Every caller will tackle the problem differently, and different callers are likely to deal more successfully with some issues and less well on others.
- As a consequence, finding a single best performing caller is not easy, and is perhaps not even feasible.
- With multiple callers on hand, anyone conducting a mutation analysis can apply all of the callers to his/her data with the aim of later constructing a list of final calls.
- In essence, combining calls from multiple callers amounts to developing a strategy to sort the calls to be included as final calls. This can be done effectively if one can systematically assign a confidence measure to being a somatic mutation across the full list.
- In general, pursuing this goal requires a validation dataset, at least to some extent.

Comparing mutation callers

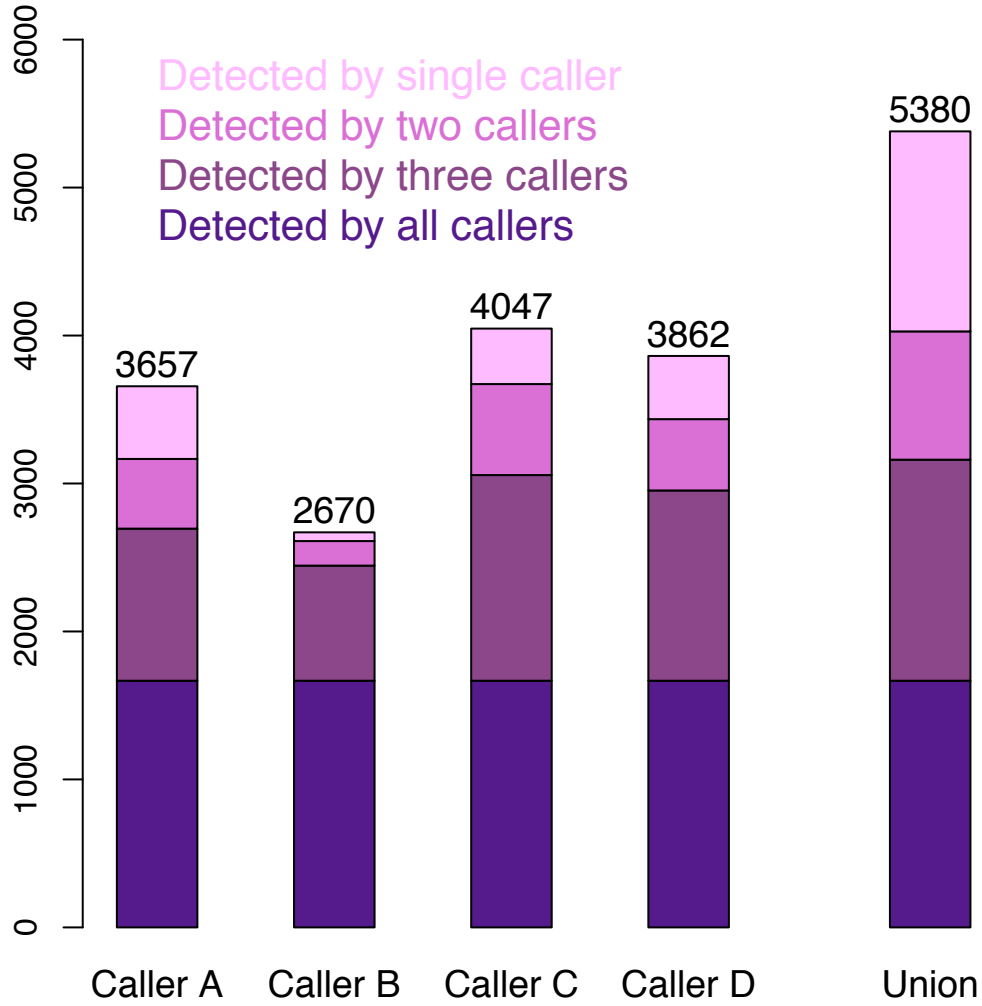
Lung squamous cell carcinoma (LUSC) dataset

- Mutation calling was done by four callers (named A, B, C and D) using Illumina exome-seq tumor-normal pairs from 16 LUSC patients.
- Some additional data exist for the same patients. One lot is high-coverage Illumina sequencing data available for tumor-normal pairs on a pre-selected set of 76 genes (540 Kb).
- It is ~3-fold higher coverage than the original exome-seq of ~80x, and called *deep-sequencing data* below.

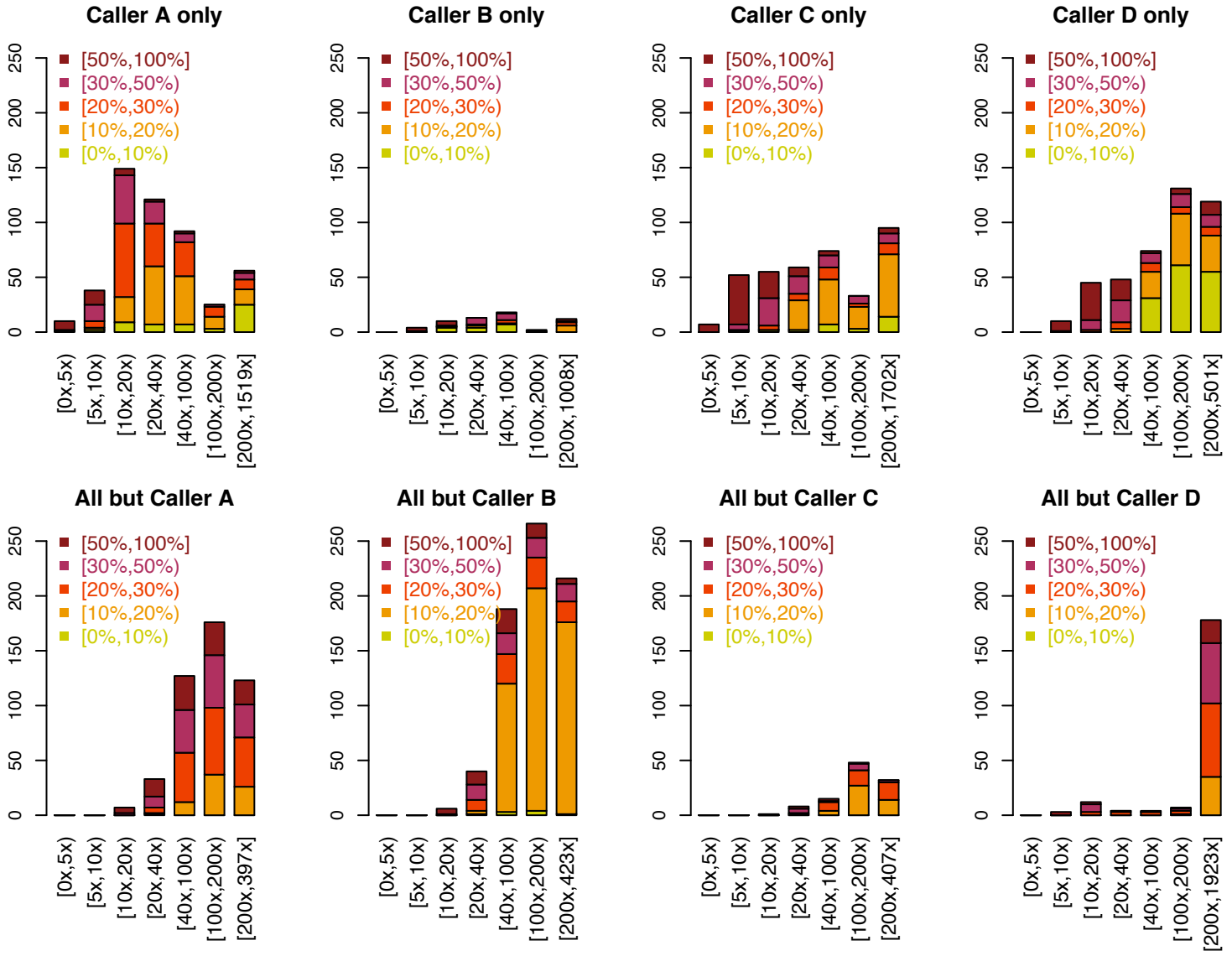
Counts of the mutations detected by four callers in the 16 LUSC tumor-normal exome-seq pairs



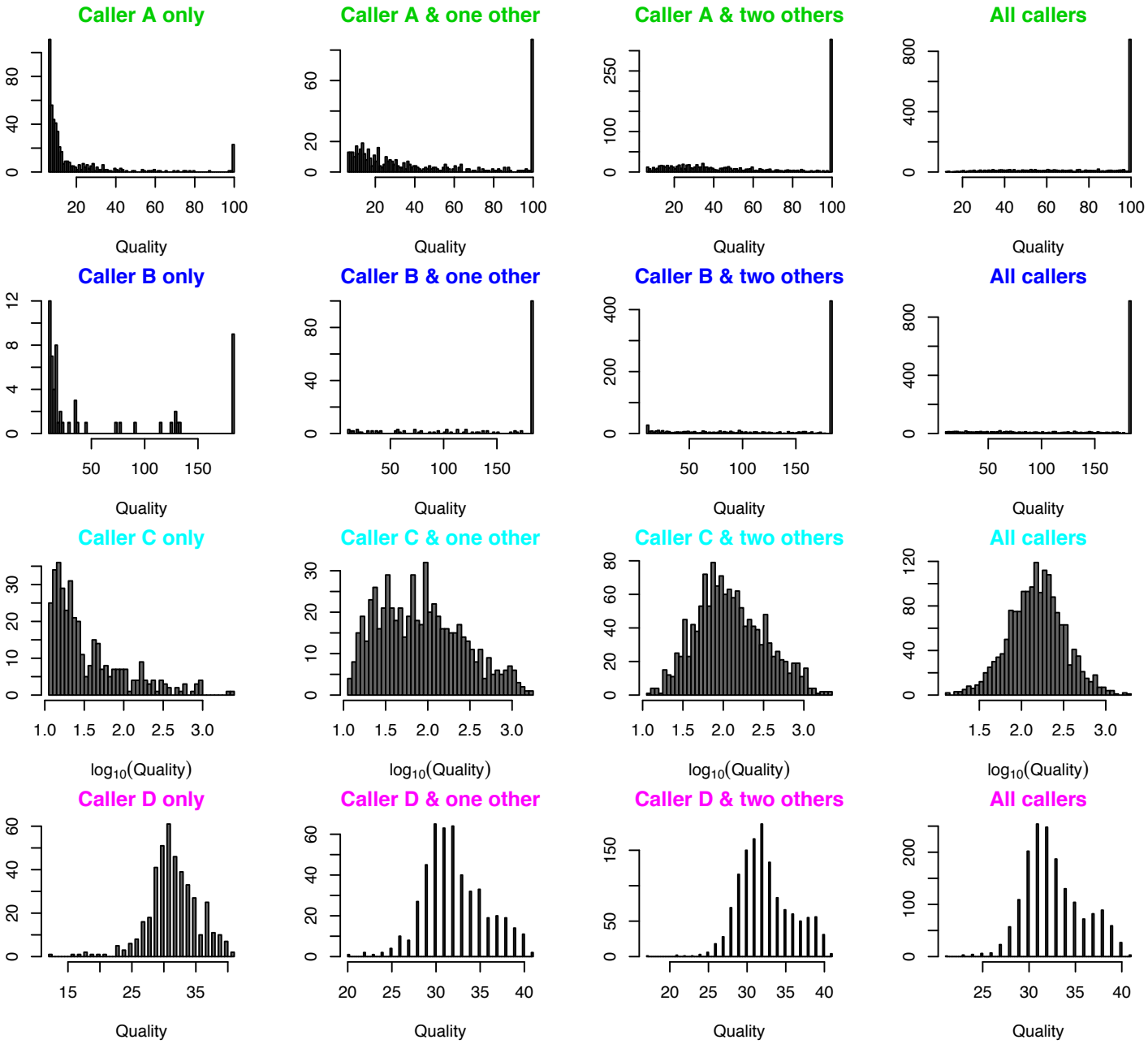
Mutations detected by each caller or by any caller ('Union'), classified based on the number of callers detecting the mutations.



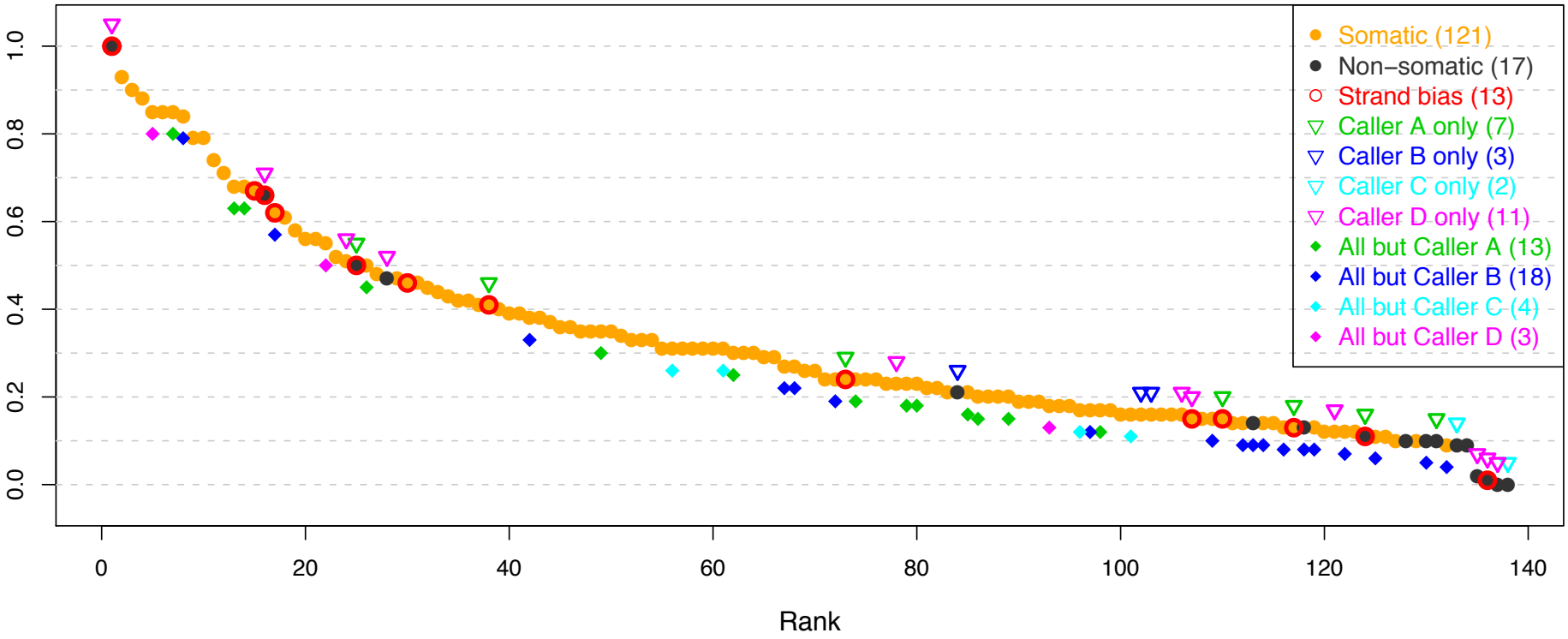
Distribution of the coverage (horizontal) and the variant allele fraction (vertical) in the tumor exome-seqs



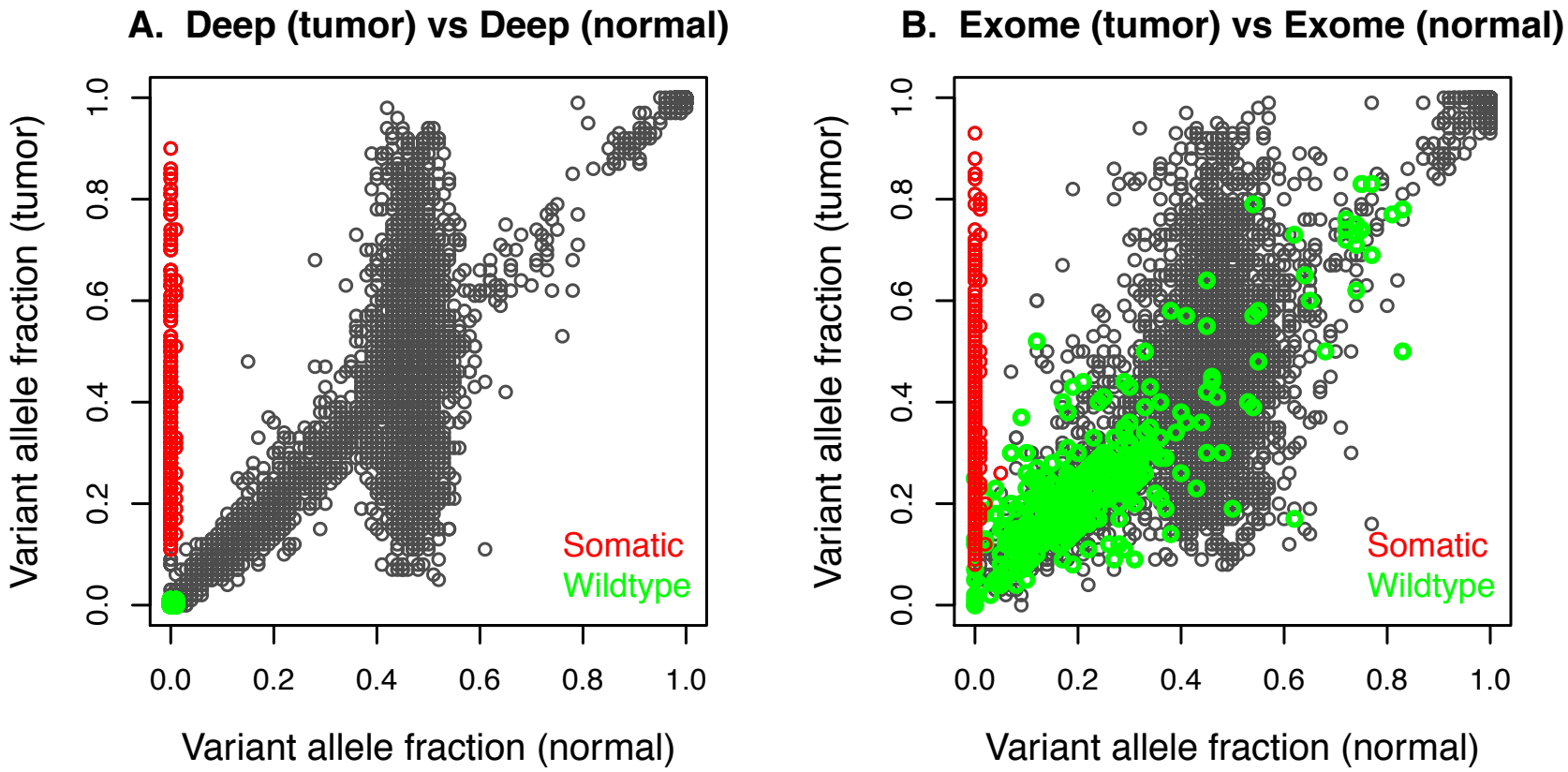
Distribution of mutation quality scores reported in VCF files



Validation status of individual mutations within the targeted regions of the deep-sequencing data (76 genes), among those detected from 16 LUSC whole exome-seq pairs using four callers.



Scatter plots of tumor vs normal variant allele fractions, using deep-seq pairs (A), exome-seq pairs (B) from 39 LUSC patients



Somatic: tumor vaf > 10%, normal vaf <2 %; wildtype: tumor vaf < 2%, normal vaf <2%.

Other results

In the paper, we described several other plots, and carried out a statistical modelling exercise, fitting a latent class model to the data, assuming that the callers gave results that were conditionally independent, given the true mutation status at a site.

We also made use of RNA-seq data on the expression levels of genes in these tumors. Such data has more variable, but frequently very high coverage of genes.

Combining mutation callers

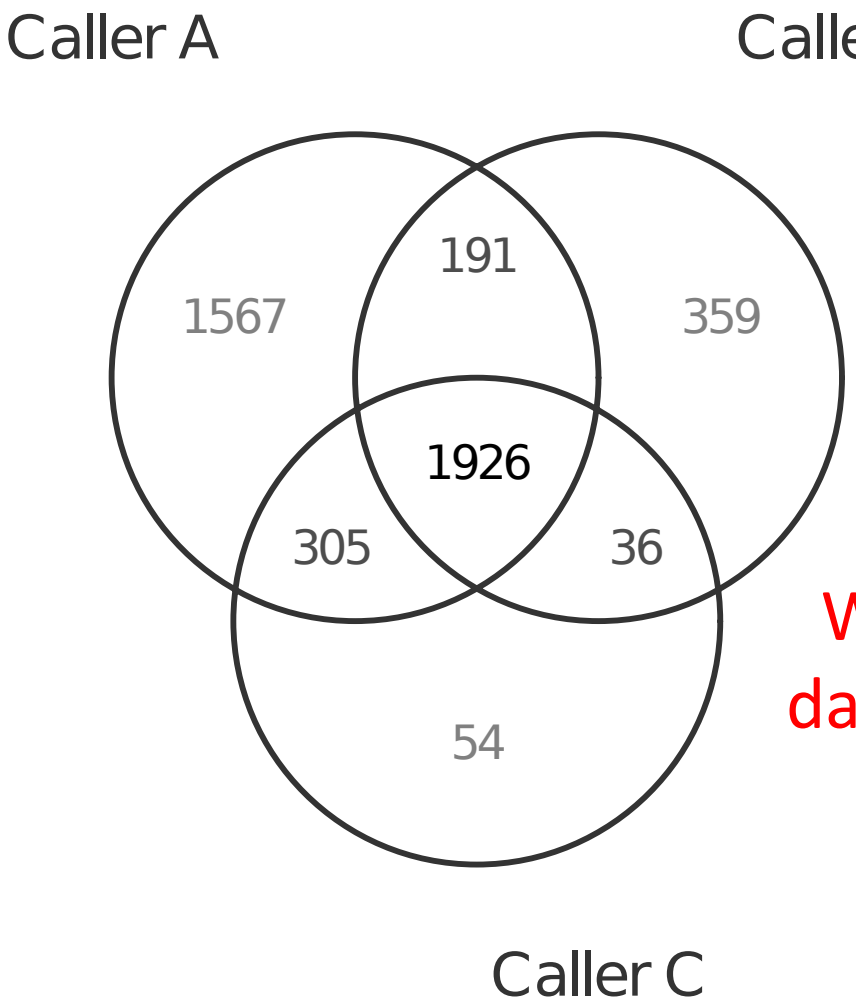
Coming up

- We build a combined caller using the mutation outputs generated by the 3 callers based on the same paired tumor-normal sequence data
- The most basic information available in each mutation output is the list of positions detected as point mutations. The output may also include additional features such as mutation substitution type, a mutation quality score, and perhaps details of filters applied to remove artifactual or low-quality variants.
- When the raw sequence data are available, genomic features can be computed for each mutation site such as sequencing depth and the variant allele fraction (the fraction of reads carrying the variant allele) for each tumor and normal sample.
- The more information that is available, the more powerful are the callers that can be constructed.

The TCGA endometrial cancer data

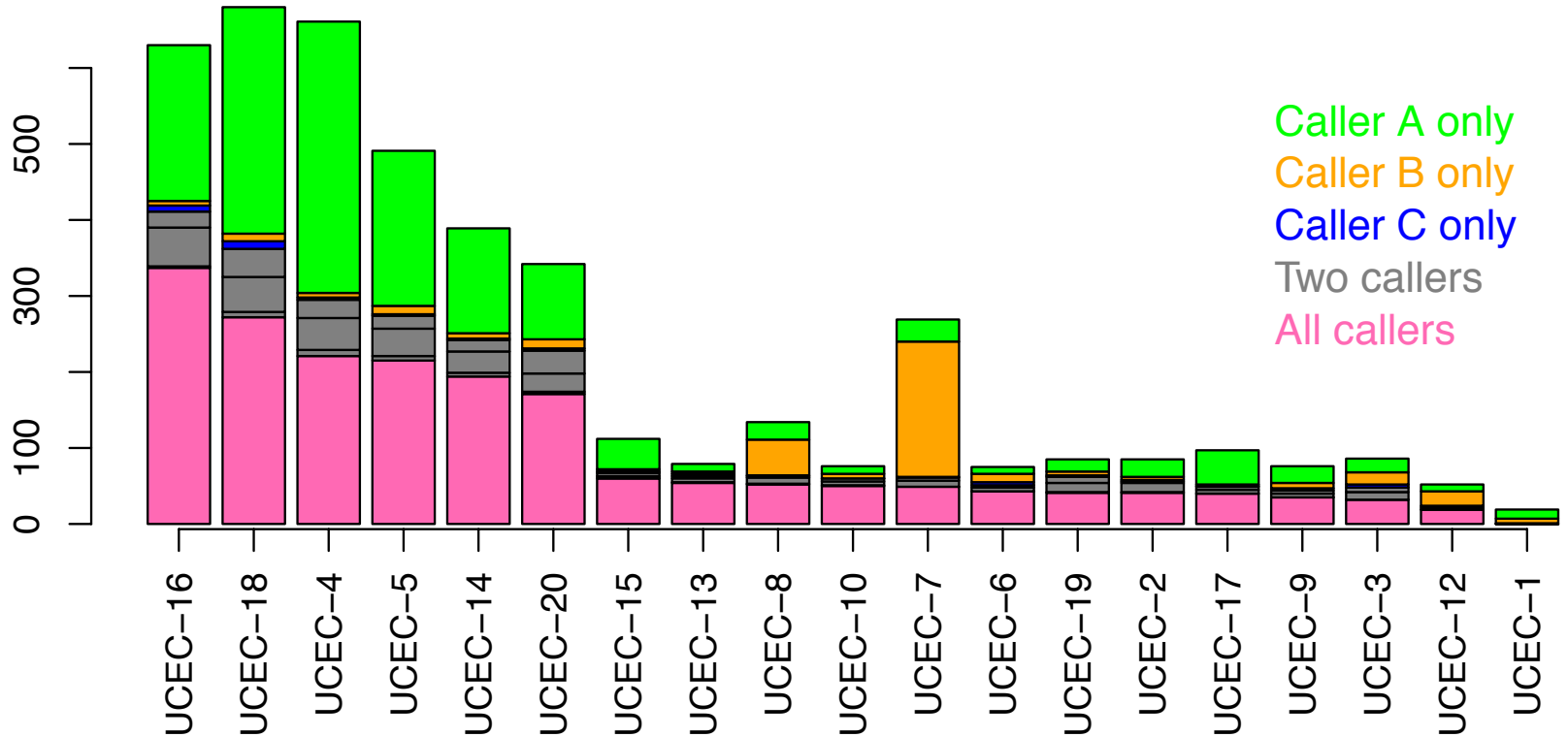
- For **194** tumor-normal Illumina exome-sequence pairs, somatic-mutation **calling was done by three centers**. In total, **51,648** single nucleotide variant mutations were detected.
- A large fraction of the mutations were targeted for custom capture validation. These sites were captured using the Nimblegen technology and then **re-sequenced independently** using an Illumina HighSeq 2000.
- In particular, impartial validation (i.e. validating all calls from all callers) was carried out for (1) **all** mutations in a randomly selected **20 patients** and (2) an **additional 243 genes** of interest from the remaining **174** patients.

Venn diagram of the mutations detected by 3 callers on 20/194 endometrial tumor-normal exome-seq pairs.



We have validation data on all these calls

Counts of mutations across 20 (19) selected patients classified based on the detection status of the callers



Taking intersections or unions

Cumulatively adding mutation sets based on combination call status

Validation results for the seven disjoint mutation sets (all genes in 20 patients) from the Venn diagram

Combination call status	Val. rate (%)	FP count	TP count	Cum FP rate	Cum TP rate
All callers	99.4	12	1,914	1.2	55.3
Caller A and C only	96.4	11	294	2.4	63.8
Caller A and B only	96.3	7	184	3.1	69.1
Caller B and C only	94.4	2	34	3.3	70.1
Caller C only	79.6	11	43	4.4	71.3
Caller A only	59.7	632	935	69.1	98.4
Caller B only	15.9	302	57	100	100

Stacking, including feature-weighted (logistic) linear stacking

Wolpert DH: Stacked generalization. ***Neural Networks*** 1992

Breiman L: Stacked regressions. ***Machine Learning*** 1996.

Sill J, Takács G, Mackey L, Lin D: Feature-weighted linear stacking. ***arXiv*** 2009.

Stacking builds a linear (logistic) function of the calls which predicts the true status of each site as accurately as possible. Each site is represented by its calls from the different callers, and a new classifier of mutation sites is learned in this feature space. Other features can be added when they are available.

Logistic models (fitted by maximum likelihood) for combining the callers:

Let $c_{ik} = 1$ if caller k **calls** site i as mutant, and $c_{ik} = 0$ otherwise. (Later we'll mention continuous scores.)

- (Logistic) linear predictor: $\Sigma_k \beta_k c_{ik}$ (similar to the above)

Suppose now that we have **genomic features** g_{ij} on each site. These can be used to enlarge the feature space. The linear predictor becomes (adding interactions with calls)

- (Logistic) linear predictor: $\Sigma_k \Sigma_j \alpha_{jk} g_{ij} c_{ik}$

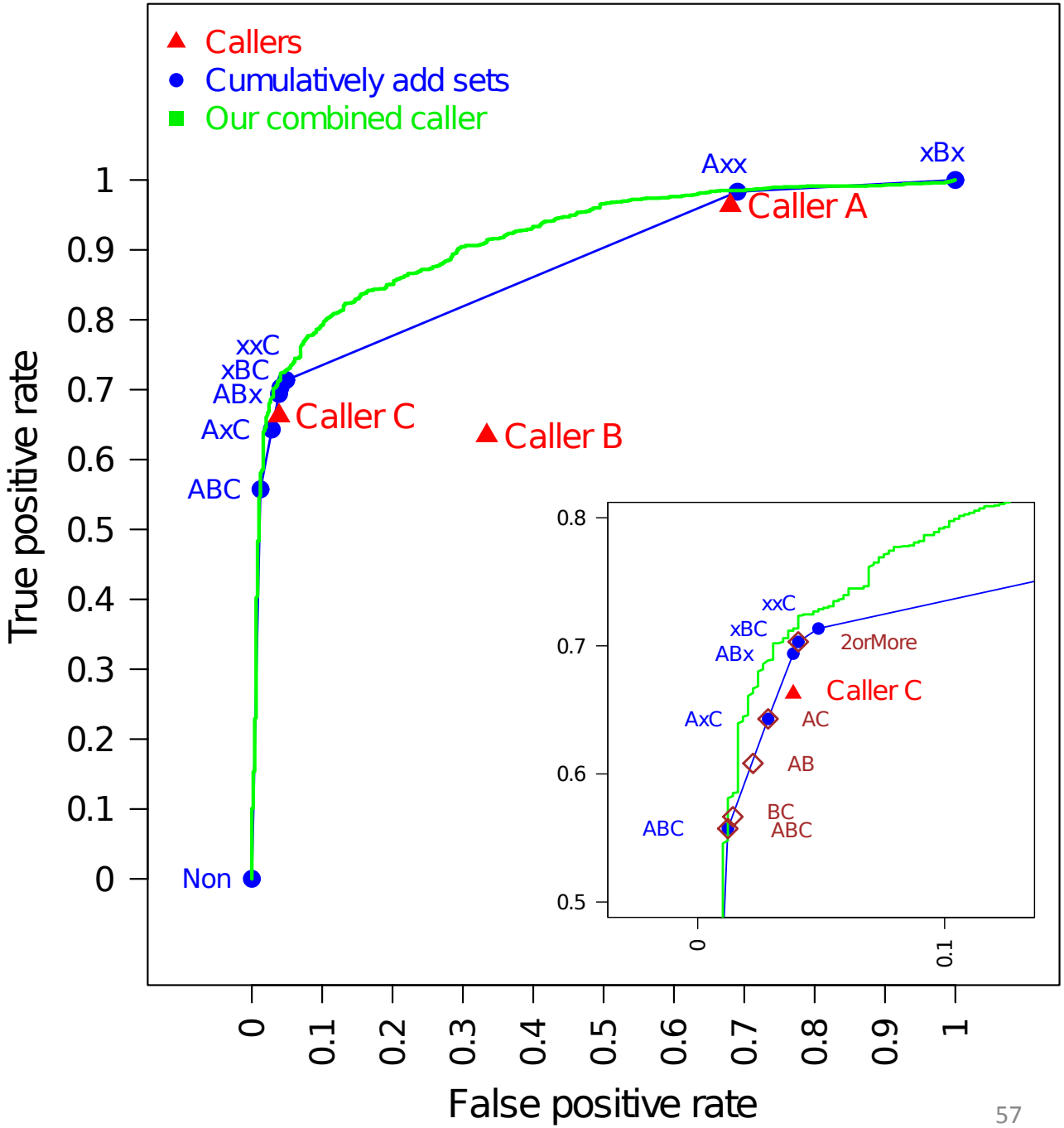
We fit this with a sparsity constraint on the αs using the R package `glmnet`. We then form an ROC curve by thresholding the fitted probability.

Genomic features used

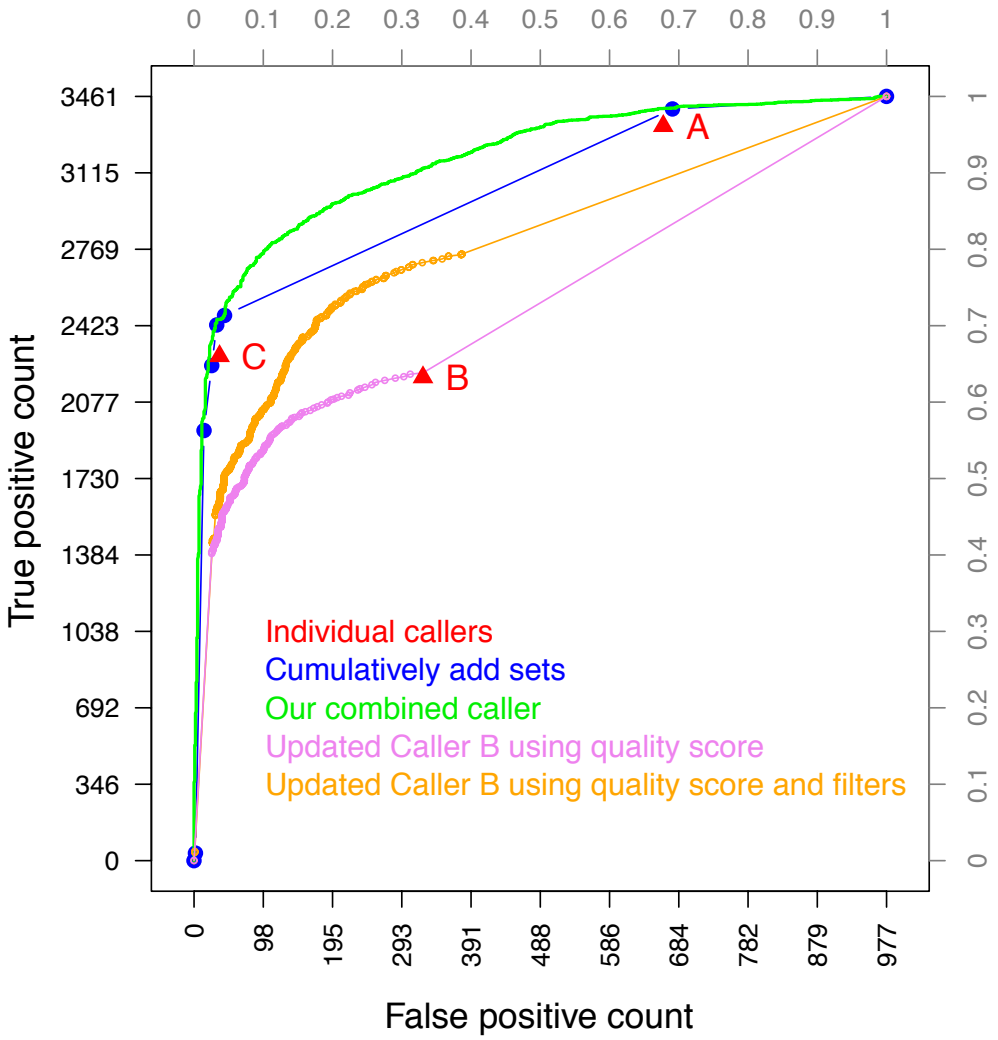
For each point mutation site in our final dataset, we know the validation status ('somatic' or 'non-somatic'), the call status (i.e., whether or not it was detected) by each of the three callers, the mutation substitution type (combination of the reference allele and the variant allele), and the sequencing depth and the variant allele fraction in each tumor and normal sample based on the exome sequence data that was used for mutation-calling.

Caller B provided more information besides the positions of the detected mutations. For a broader set of somatic variants (candidate mutations), it reported the mutation quality score as well as the pass/fail status of individual filters at each site. Although the detailed description of each filter was not available, the filter outcomes were available, which we were able to use for improving Caller B's performance.

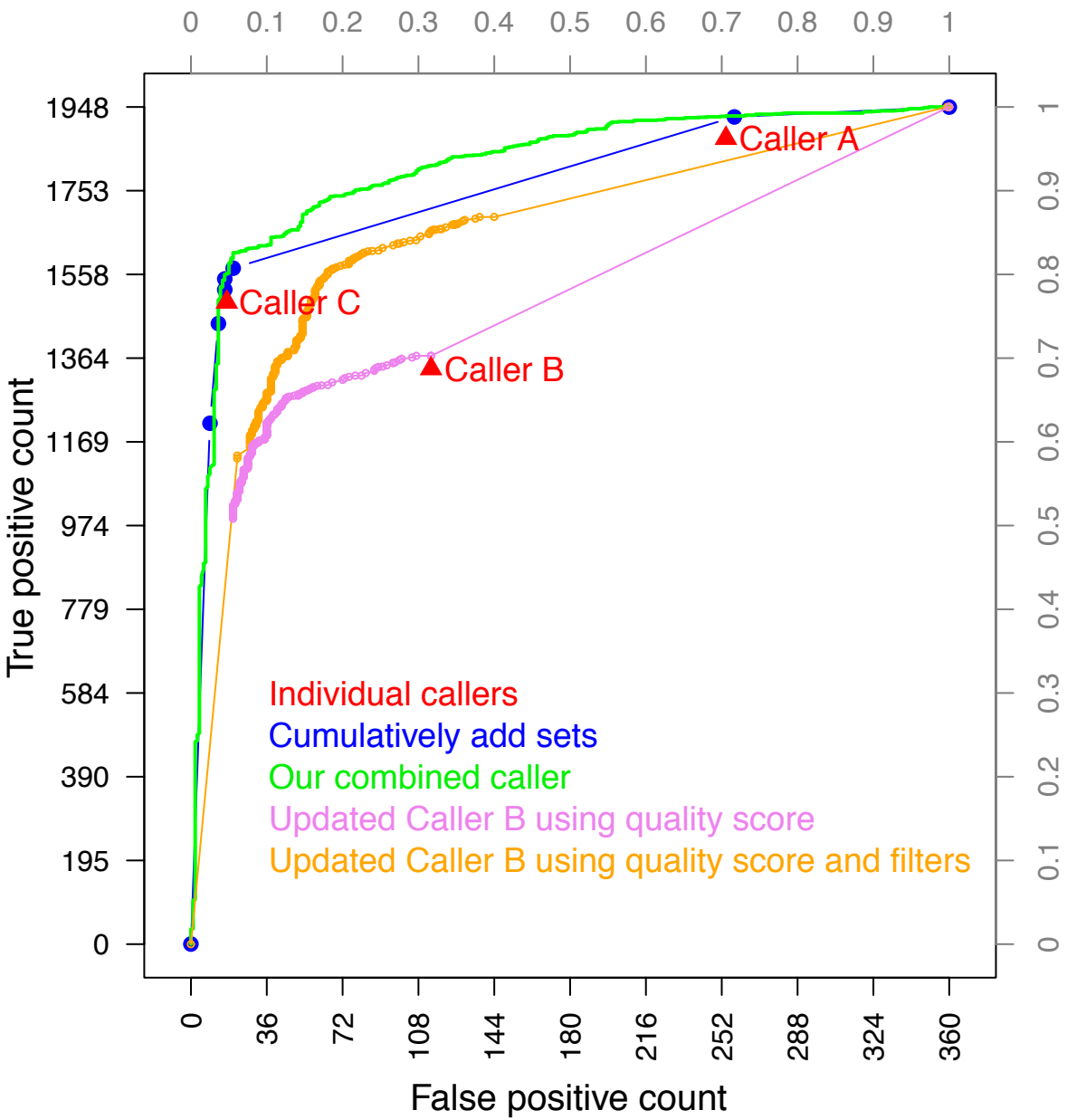
Model fitting was done using the point mutations in 243 genes of interest from 174 patients excluding the 20 patients, and evaluation was done on the point mutations in the 20 selected patients.



ROC curve of an improved Caller B built from a logistic model using its mutation quality score and filters



ROC curves with training and test sets reversed



References

Kim & Speed ***BMC Bioinformatics*** 2013, 14:189
http://www.biomedcentral.com/1471-2105/14/189

Kim *et al.* ***BMC Bioinformatics*** 2014, 15:154
http://www.biomedcentral.com/1471-2105/15/154

and many more in these papers.

Acknowledgements

My collaborators:

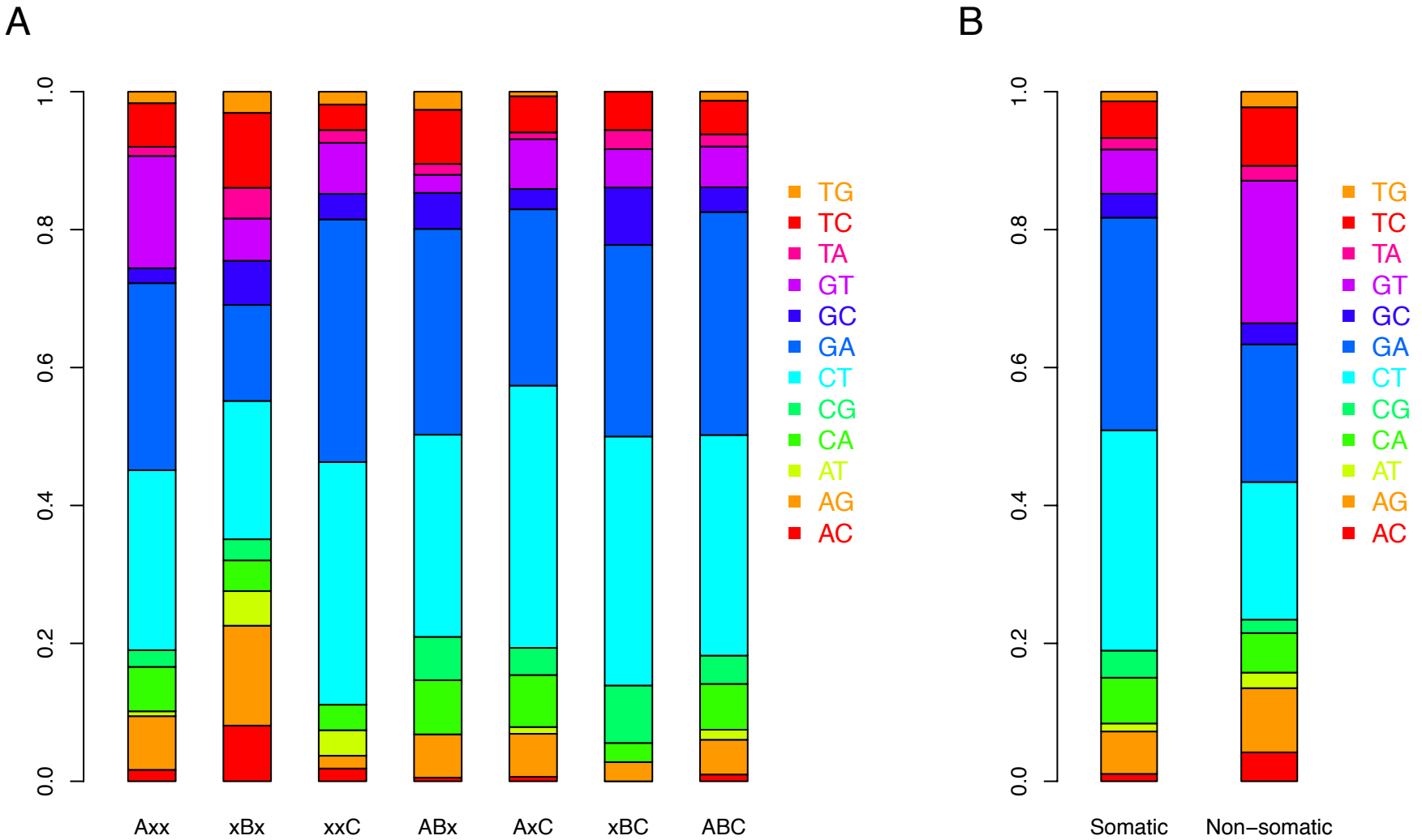
Su Yeon Kim, Veracyte, Inc.

Laurent Jacob, CNRS, Lyon

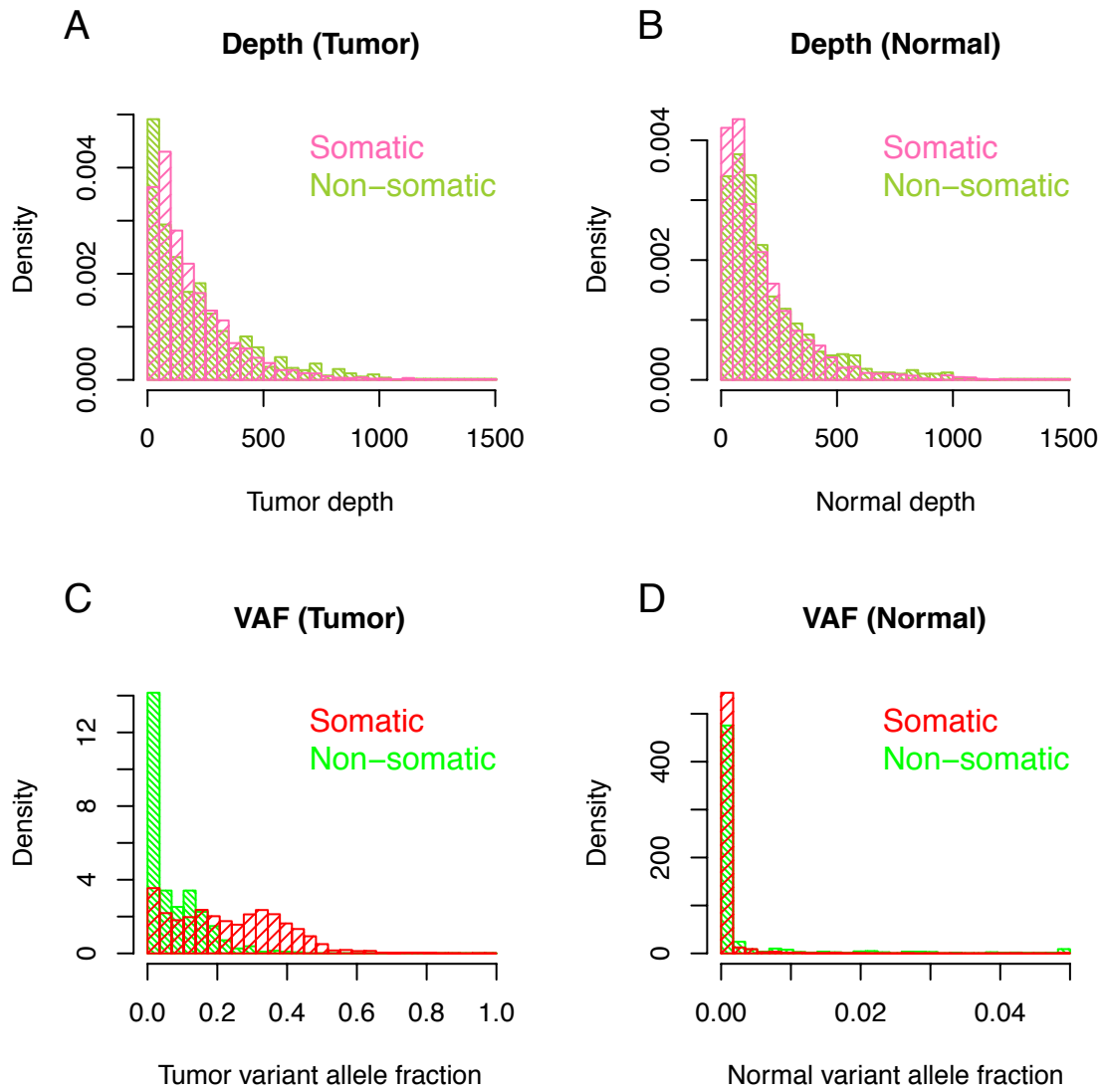
The TCGA mutation calling group:

led by David Haussler, Gad Getz, Li Ding and David Wheeler,

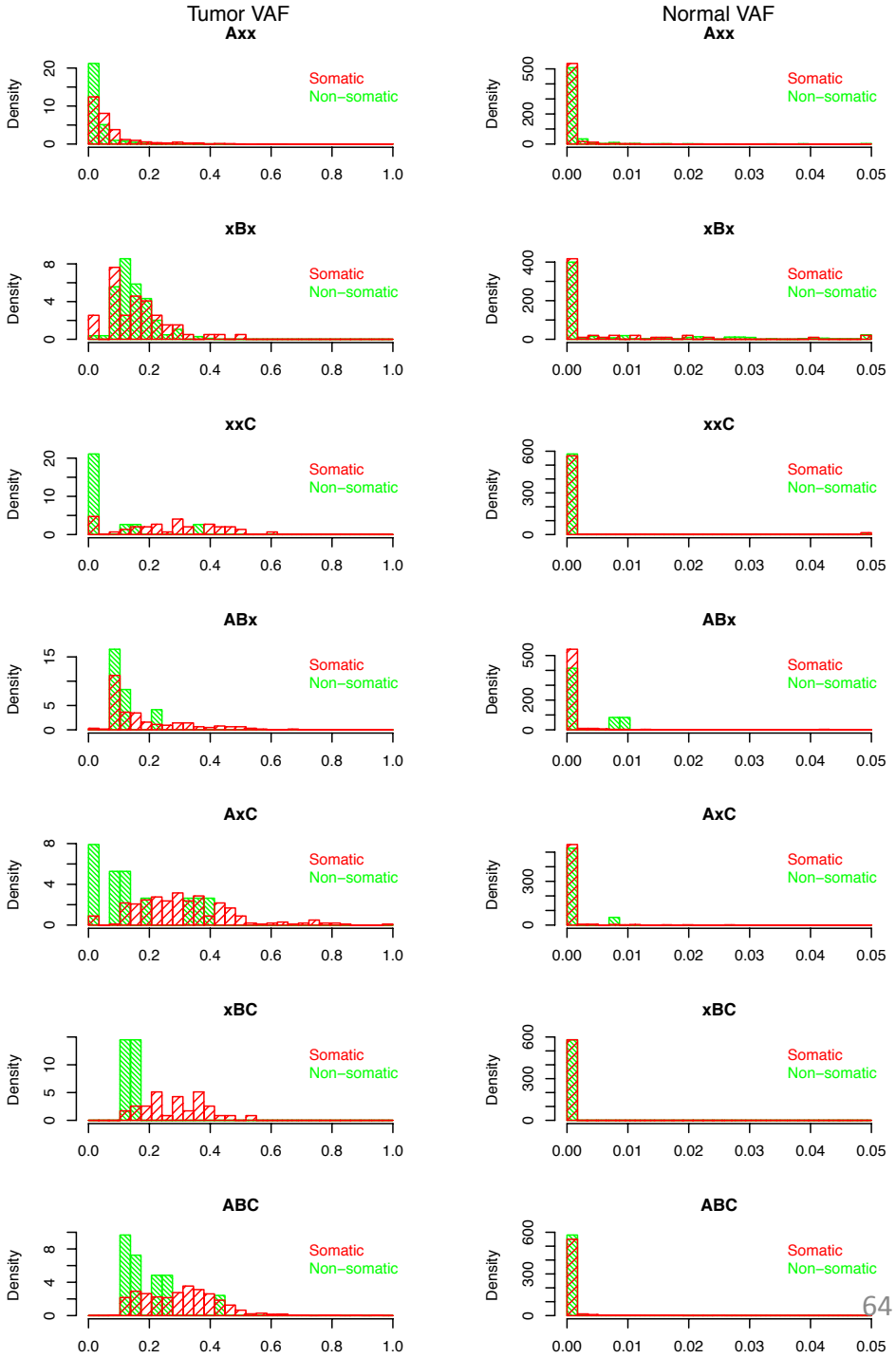
and to Singer Ma, Andrey Sivchenko, Cyriac Kandoth, Kyle Chang, Heidi Sofia, Kenna Shaw, Paul Spellman, Elizabeth Purdom, and many others in The Cancer Genome Atlas project.


A
B
0.0
0.2
0.4
0.6
0.8
1.0
Axx
xBx
xxC
ABx
AxC
xBC
ABC
Somatic
Non−somatic
TG
TC
TA
GT
GC
GA
CT
CG
CA
AT
AG
AC

Comparison of the distribution of four genomic features between *somatic* and *non-somatic* mutations.



The distribution of the variant allele fraction (VAF) in the tumor exome-seq (left) and the normal exome-seq (right column).



Variant Call Format (VCF)

- All four centers agreed on one annotation representing exome regions and generated calls only within those regions.
- Outputs were provided in a modified Variant Call Format (VCF), which reports the genomic position, somatic status, filter status, sequence information from each tumor and normal sample.
- The filter status indicates whether the variant (candidate mutation) passes all the filters implemented by each caller or not. The full details of all filters were not given in the VCF files though, partly because the modified VCF format was under active development.