# Calling Variants from Sequence Data

Robert Gentleman

Genentech

Genentech
A Member of the Roche Group

# Outline

- The objective(s)
- Our experiment
- What we found out
- Next steps

- Caveats:
  - this is a work in progress, as you will see
  - Much of what I present is just based on Chr 1

# The objectives

- Identify a set of variants that are particular to an individual
  - Identify the genotype of an individual
  - Identify the mutations/variations that are specific to a tumor
- The first of these requires us to compare our data to a reference sequence
- The second requires that we compare the tumor genome to the germline (not quite) genome

# Landscape

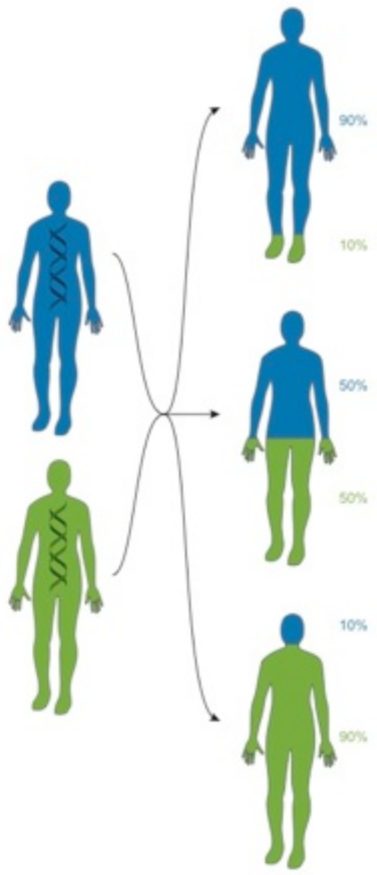
- There are many tools some for calling genotypes
  - SNVs in normal genomes (diploid for humans)
  - GATK, SOAP2, ….
  - Many that are not public, most labs have their own set of procedures
- Tools for calling variants
  - Atlas2 (seems to rely on GATK or similar)
- Tumor Normal Comparisons
  - Mutect
  - SomaticSniper
  - Strelka

# A way forward

- We do better at engineering than at discovery
  - By engineering I mean the process of iterative refinement of a solution
  - Iterative refinement requires a good and substantial *gold standard* data set containing substantial numbers of TPs and TNs
  - We want the TPs at varying frequencies (not just het and hom)
- Part of the reason there are so many competitors is the absence of good objective comparisons
  - A good *gold standard* data set could address this

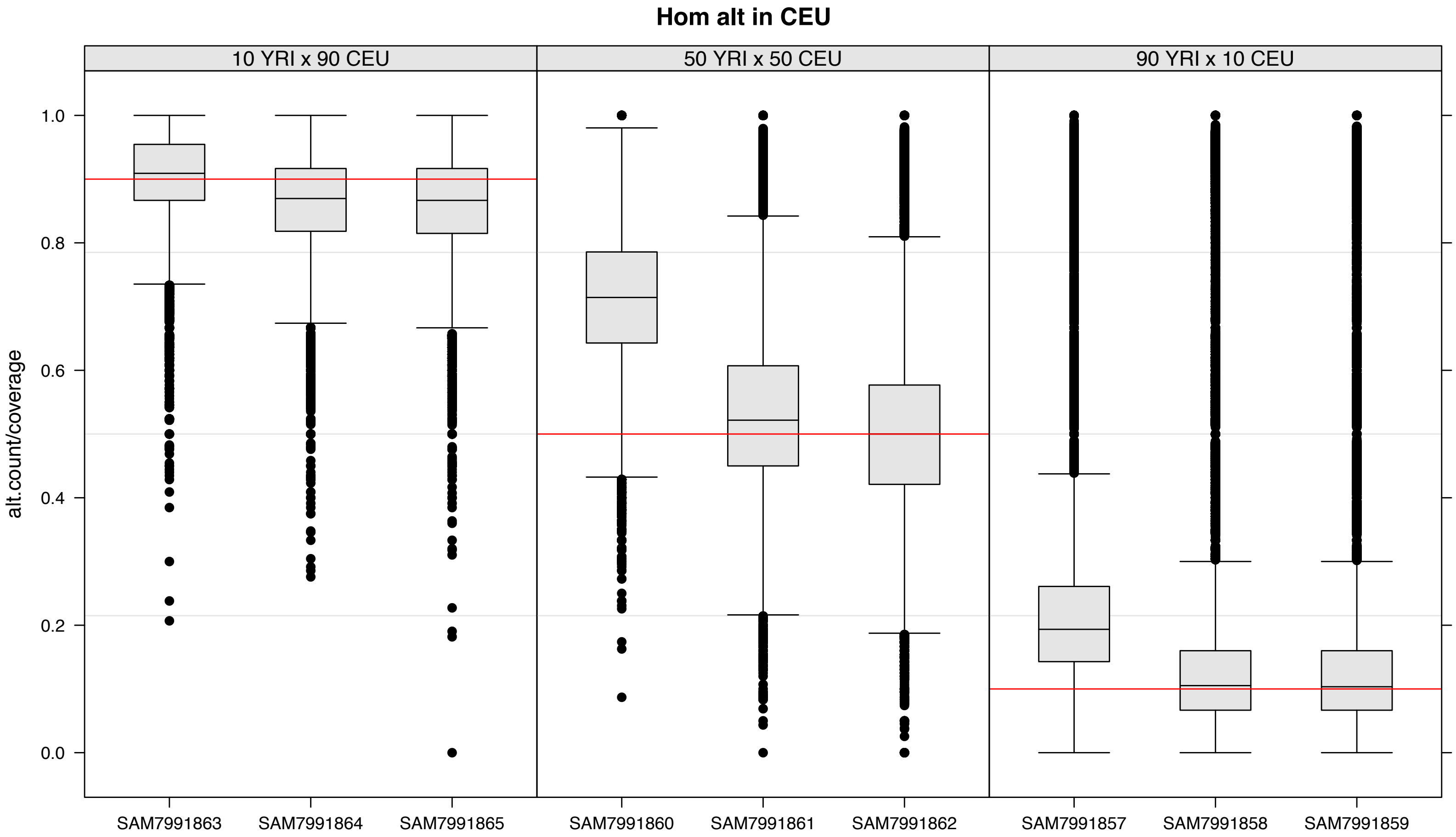
# The experiment

- Mix DNA from two well sequenced individuals and sequence the mixtures
  - NA12878, the daughter of a CEU trio
  - NA19240, the daughter of a YRI trio
  - Triplicate samples (biologic) at 10-90, 50-50 and 90-10
  - 20X coverage, 75nt paired end reads per sample



Michael Lawrence

# How did we do?



Genentech
A Member of the Roche Group

# How did we do?

- Not that bad – one obvious outlier
- But notice the lack of symmetry in the 90-10 and 10-90
  - For 90 YRI-10 CEU the dots go way up to around 1, suggesting that the YRI is actually non-ref at those loci, even though the 1000G genome says they are hom ref
  - We find substantial evidence that the YRI genome is less accurate than the CEU, and that will affect FP rates, as many of those may indeed be TPs

# Expected Frequencies of Alleles

- our samples contain mixed genotypes
- The expected frequency of an allele depends on whether it was het or hom in the original genome and on the mixture
- Example: 90-10 mixture (CEU/YRI)
  - Hom alt in both, EF=1.0
  - Hom alt in CEU, het in YRI, EF=0.95
  - Hom alt in CEU, WT in YRI, EF=0.9
  - Het alt in CEU, Hom alt in YRI, EF=0.55
  - Het alt in both, EF=0.5
  - Het alt in CEU, WT in YRI, EF=0.45
  - Hom alt in YRI, WT in CEU, EF=0.1
  - Het alt in YRI, WT in CEU, EF=0.05

# Experiment – Data

| %CEU | %YRI | Reads (analyzed) | Avg. Coverage |
|---|---|---|---|
| 90 | 10 | 461,449,560 | 22.3 |
| 90 | 10 | 475,567,437 | 23.0 |
| 90 | 10 | 460,196,498 | 22.3 |
| 50 | 50 | 489,166,262 | 23.7 |
| 50 | 50 | 442,737,941 | 21.4 |
| 50 | 50 | 430,779,023 | 20.8 |
| 10 | 90 | 496,958,600 | 24.0 |
| 10 | 90 | 494,245,570 | 23.9 |
| 10 | 90 | 534,458,340 | 25.8 |

- 6 sets of plates (3 of each), DNA extracted and mixed separately for each replicate
- Sample prep and sequencing was done separately
- We did not do either sample on its own

# Well estimated Genotypes

| Cell Line | Trio | Source | Reference | Coverage | Het/Hom |
|---|---|---|---|---|---|
| **NA12878** | CEU | Broad | Hg19 | 64x | **2402001/1423889** |
| **NA12878** | CEU | 1000G | Hg18 | 61x | **1678115/1047713** |
| **CEU UNION** | CEU | Both | Hg19 | | **2424095/1427209** |
| | CEU | Unique | | | **1643487/630909** |
| **NA19240** | YRI | 1000G | Hg18 | 66x | **2227251/1108784** |
| | YRI | Unique | | | **1416362/299673** |
| **UNION** | Both | ALL | Hg19 | | **3840201/1726882** |

- We mask regions of low complexity.
  - difficult to map to and not interesting
- We combine the two CEU genotypes using a
  - Union; Broad het calls are used in preference to the 1000G hom calls
- Notes:
  - Het/hom ratio is larger in YRI

# Some Definitions

- True Positive (TP): a variant that is present in the underlying mixture genome
- True Negative (TN): a locus where both CEU and YRI are WT
- False Positive (FP): a called variant where the CEU and YRI are WT
- False Negative (FN): a failure to call a *known* variant
- False Discovery Rate (FDR): the proportion of discoveries (calls) that are false
  - This is probably more meaningful than the FP rate
  - This is much easier to estimate
- These rates are affected by errors in the gold standard
  - FP might be TP
  - FN might be TN

# Statistical Challenges

- multiple testing
  - many millions of tests (discrete probability distribution)
- varying power
  - coverage determines power, coverage varies
- varying size
  - affected by coverage and frequency of the variant
- Bias
  - Many sources, most not known
  - Eg: we align to the reference genome (reference bias)

# Variant Calling

- where are there differences between the genome sequence data and the reference?
- our reference genome is homozygous at every locus
- $H_0$: the genome (G) and ref (R) are the same (G is homozygous identical to the reference)
- under $H_0$ all reads should be the reference allele
  - errors are due to sequencing errors
- every heterozygous locus is a variant (in this case), some homozygous loci are too

# Variant Calling

- usual algorithm: if X>1, and coverage > K, call a variant
  - K is artificial, the requirement should be based on evidence against $H_0$, not on coverage
  - Eg: coverage 5, but 4 non-ref alleles?
- Pr(2 or more non-ref reads (alleles)| $H_0$) is a Binomial calculation, $p_E=10^{-3}$, n=coverage
  - For n=10, the prob is $10^{-5}$
  - For n=50, the prob increases to $10^{-3}$
- So we will have lots of FPs if we are not careful

# Calling Variants

- We (and others) use a probability model
  - Can think of it as either a LRT or a Bayes Factor
- Look at the ratio of the likelihood under a model (initially Binomial) for
  - M1: the variant is a sequencing error (p=0.001)
  - M2: the variant is present at some frequency (p=0.2)

$$\frac{P(M1)}{P(M2)} = \frac{p_1^x(1-p_1)^{n-x}}{p_2^x(1-p_2)^{n-x}} = 1$$

Genentech
A Member of the Roche Group

# Calling Variants

$$\frac{p_1^x(1-p_1)^{n-x}}{p_2^x(1-p_2)^{n-x}} = 1$$

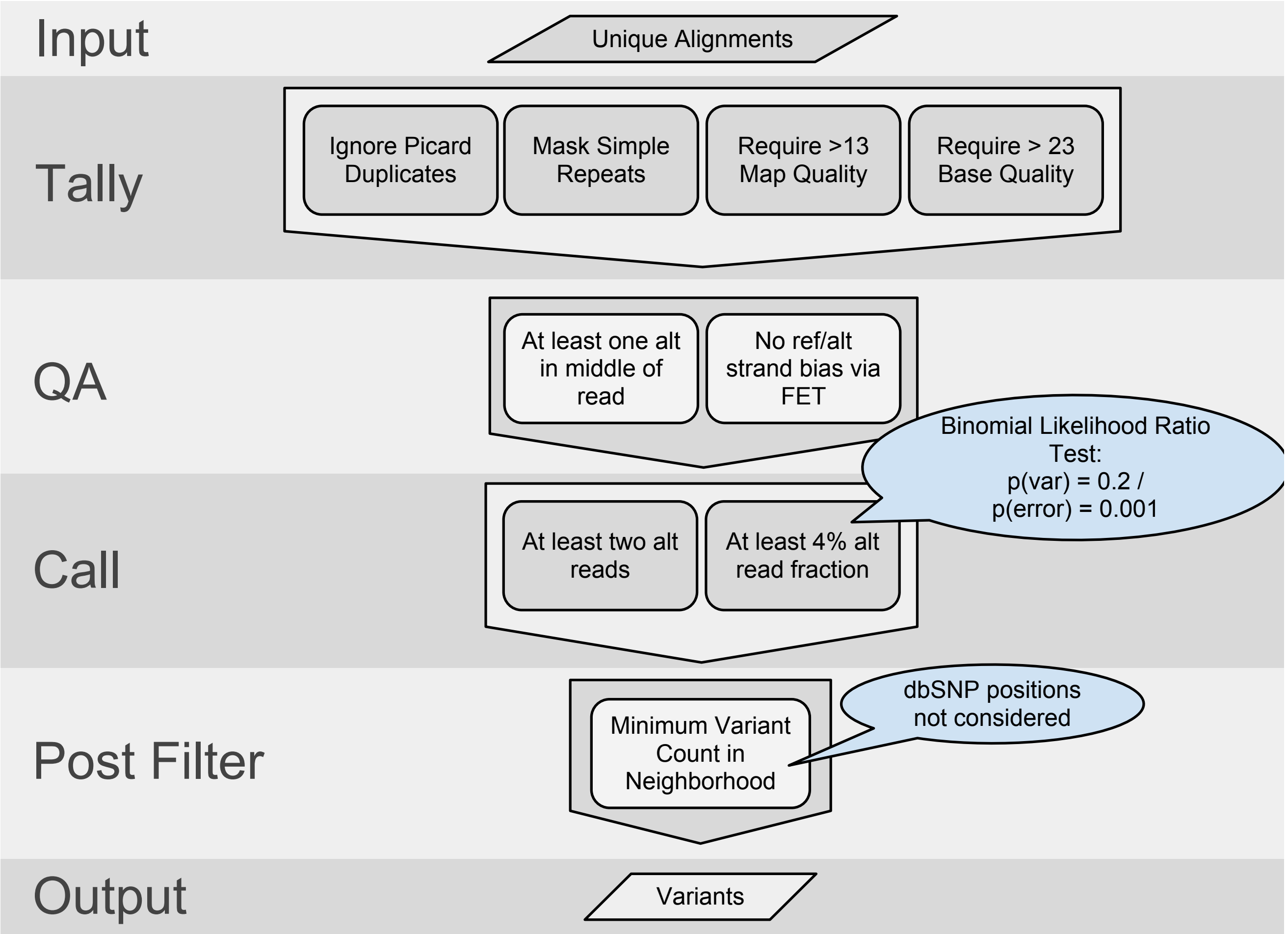
- When we solve this using $p_1=0.01$ and $p_2=0.2$
- We call a variant (M2) when $x/n>0.04$
- Issues:
  - More than one variant at the locus
  - Low coverage introduces discreteness

# Filtering the data

- The reads are aligned using gSNAP (T. Wu)
- And then a number of QA processes are used to filter out reads with anomalies that are more likely to be due to technical artifacts than real biology.
- Our test is a likelihood ratio (which can also be interpreted in a Bayesian fashion)

Michael Lawrence

# Workflow


Input
Unique Alignments
Tally
Ignore Picard Duplicates
Mask Simple Repeats
Require >13 Map Quality
Require > 23 Base Quality
QA
At least one alt in middle of read
No ref/alt strand bias via FET
Binomial Likelihood Ratio Test: p(var) = 0.2 / p(error) = 0.001
Call
At least two alt reads
At least 4% alt read fraction
Post Filter
Minimum Variant Count in Neighborhood
dbSNP positions not considered
Output
Variants

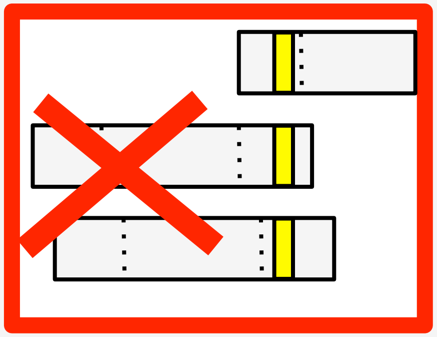
Genentech
A Member of the Roche Group

# QA Filters

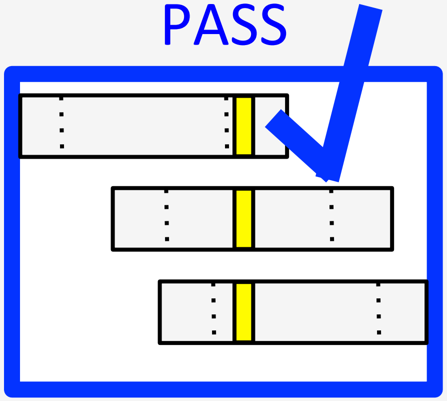
Michael Lawrence

Discard variants only seen at end of read.

FAIL



PASS



Discard variants with ref/alt strand bias.

FAIL

| | | |
|---|---|---|
| + | 5 | 30 |
| - | 23 | 25 |

PASS

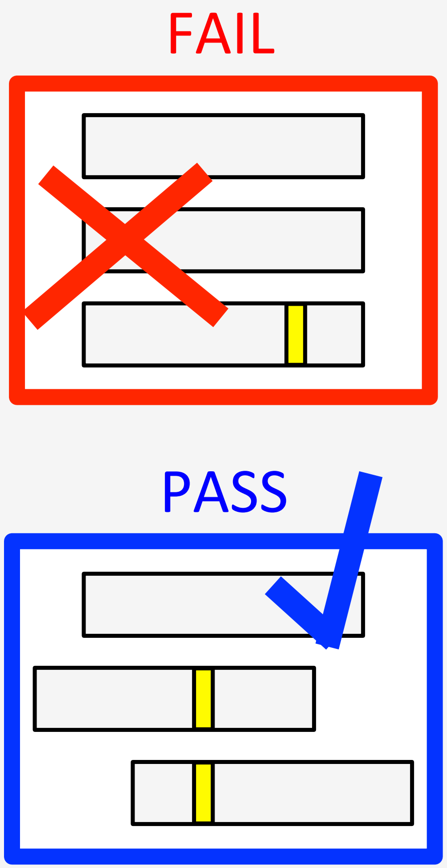
| | | |
|---|---|---|
| + | 5 | 8 |
| - | 23 | 25 |

Genentech
A Member of the Roche Group

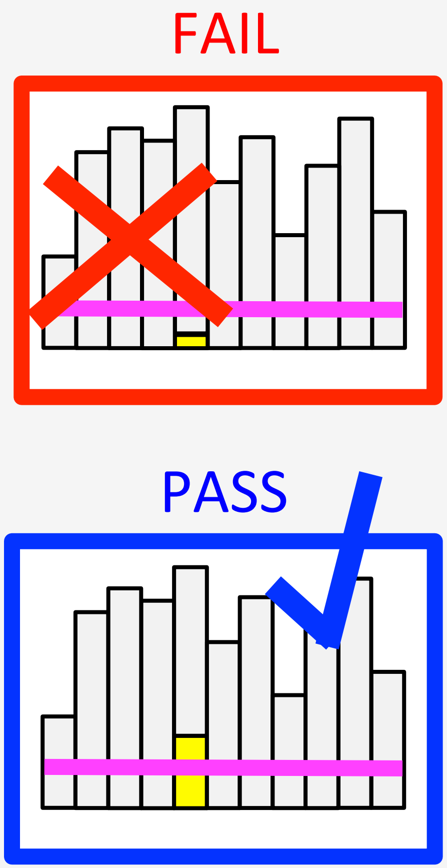
Michael Lawrence

# Calling Filters

Discard variants with only one alt read.



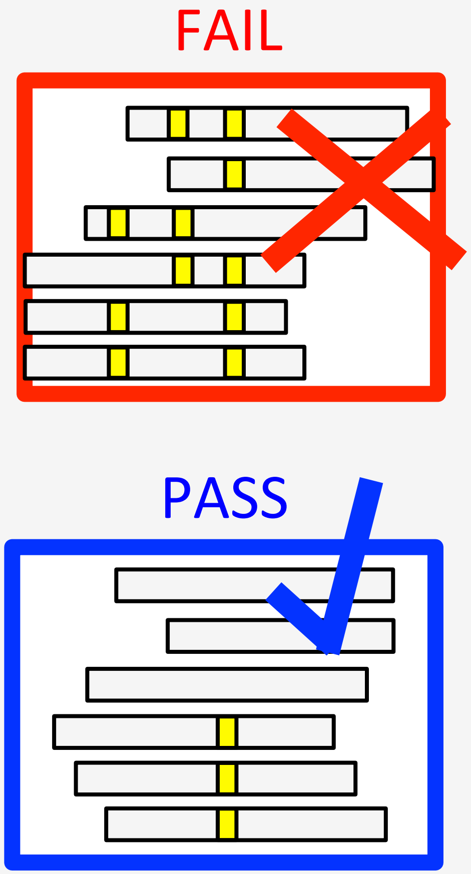
Discard variants with < 4% alt read fraction.



Michael Lawrence

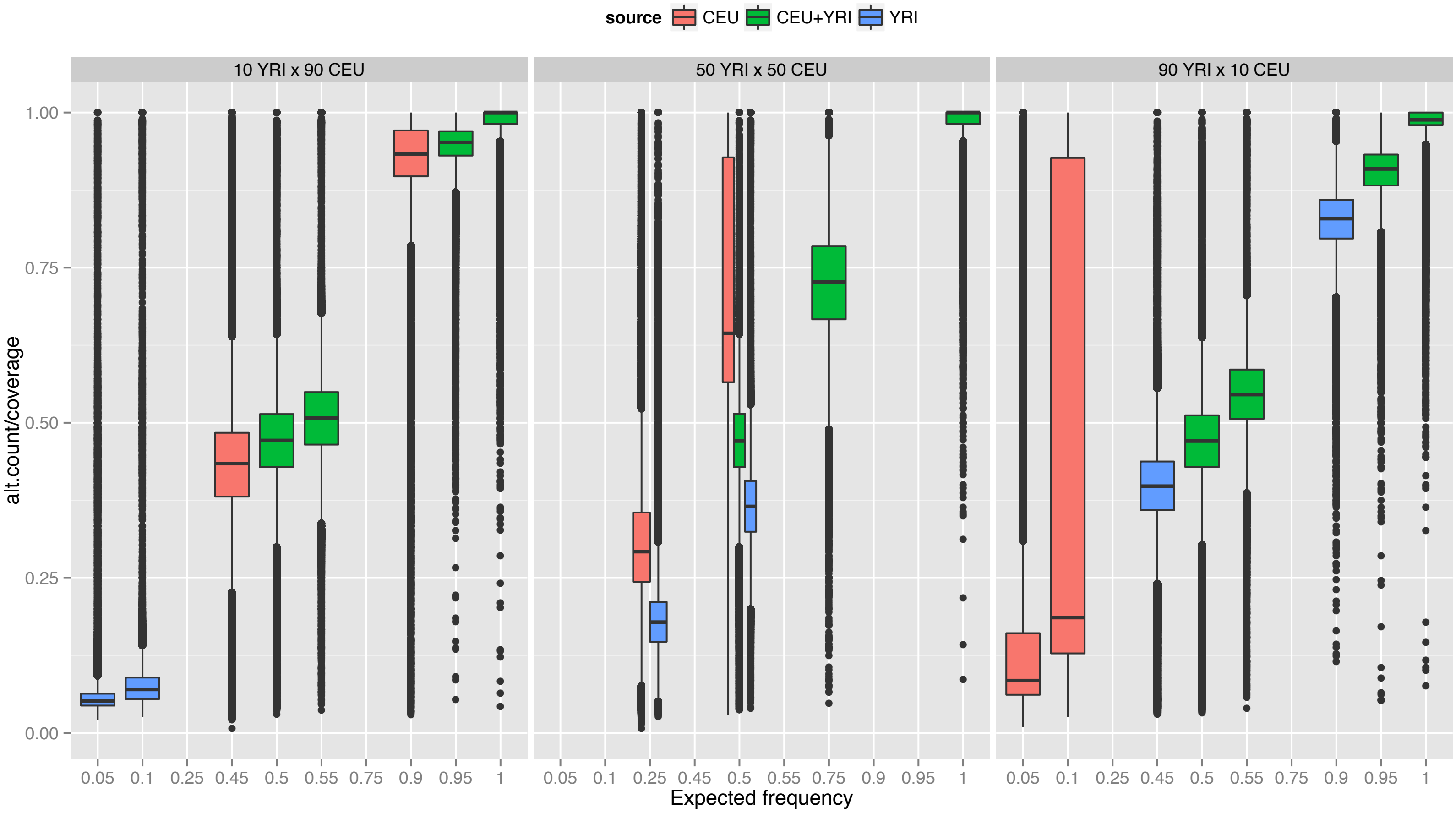
# Post Filter

Discard variants clumped on the chromosome.

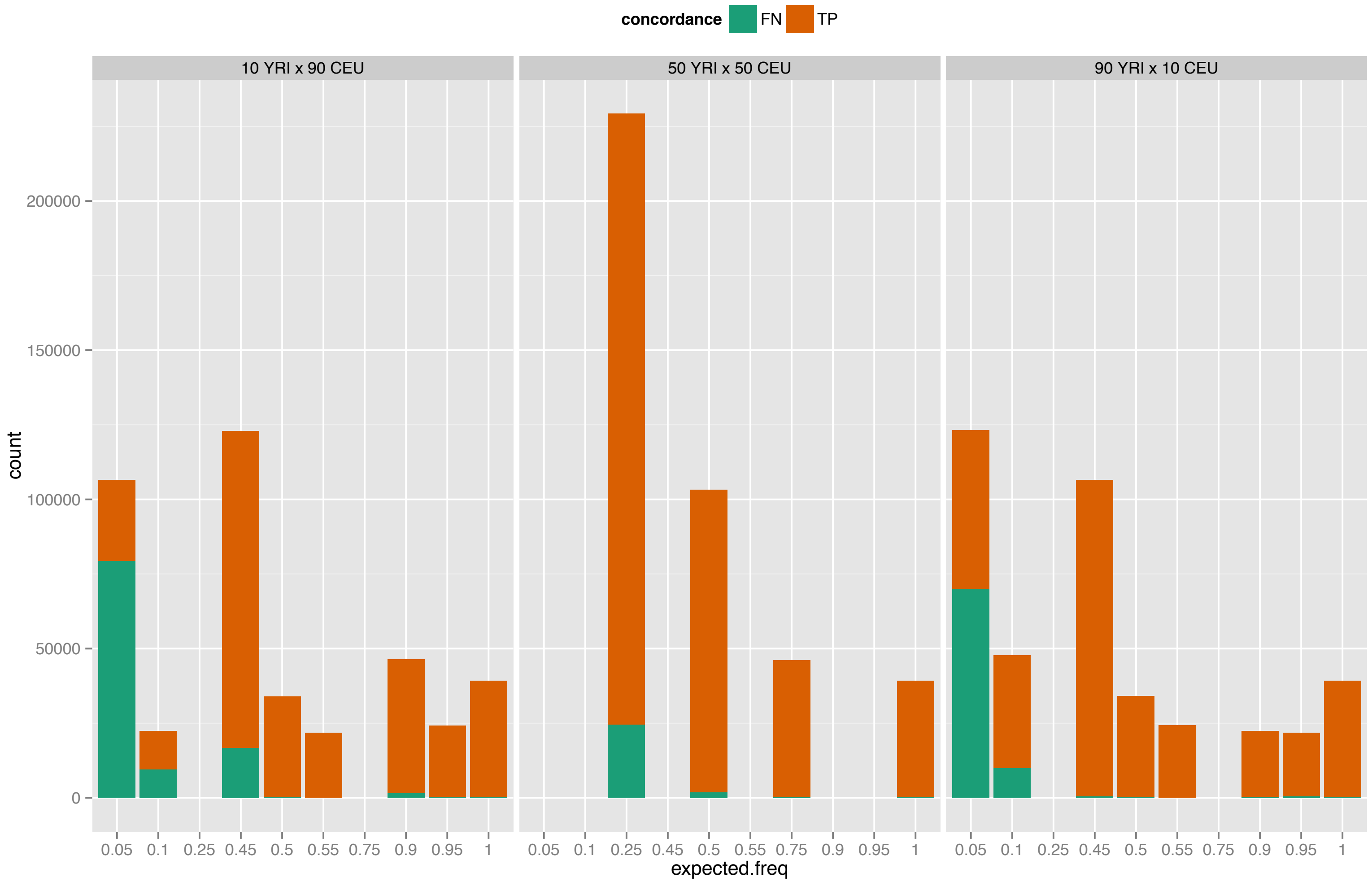


Genentech
A Member of the Roche Group

# Observed Variant Frequencies

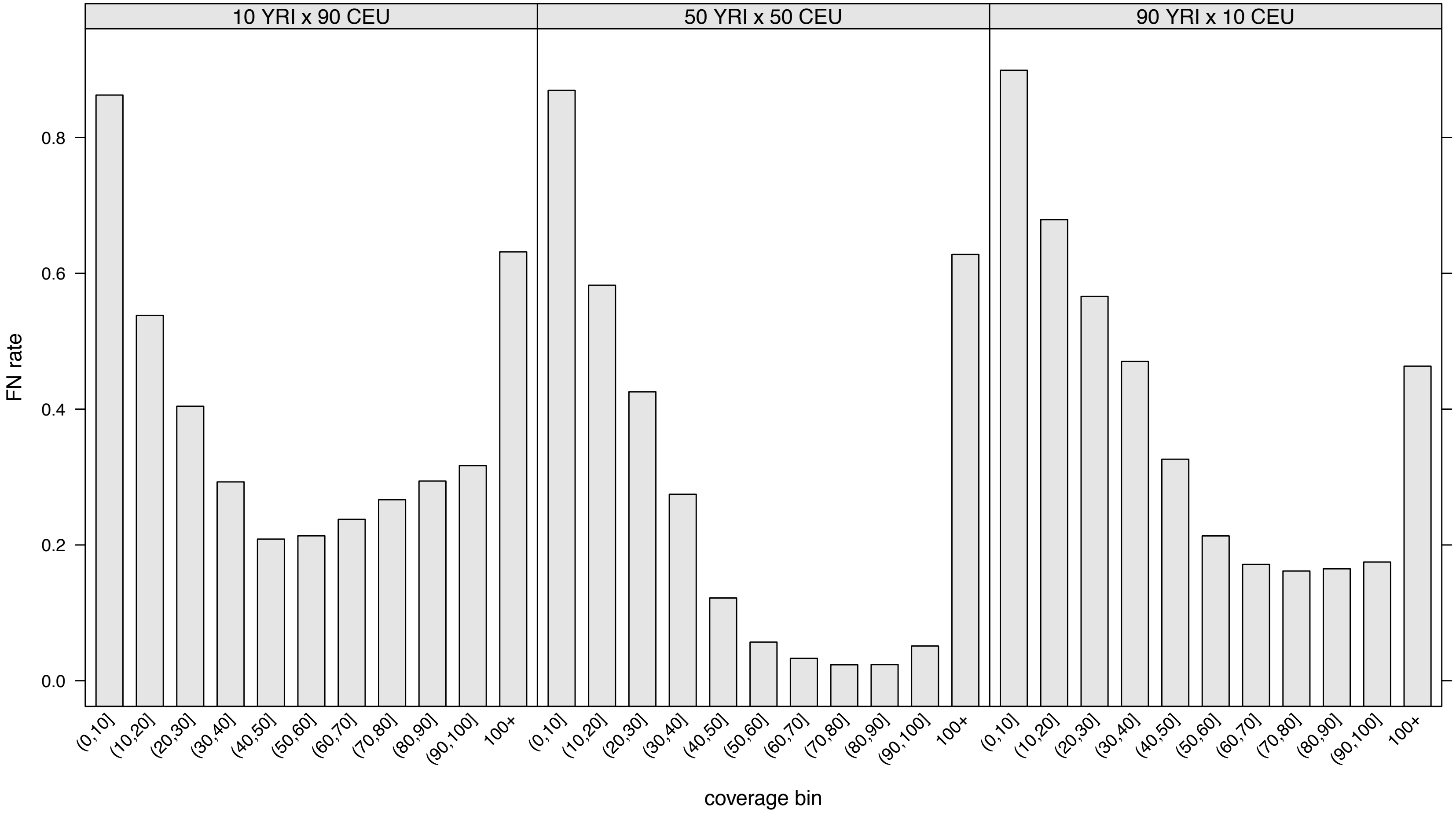


# FN by expected Frequency



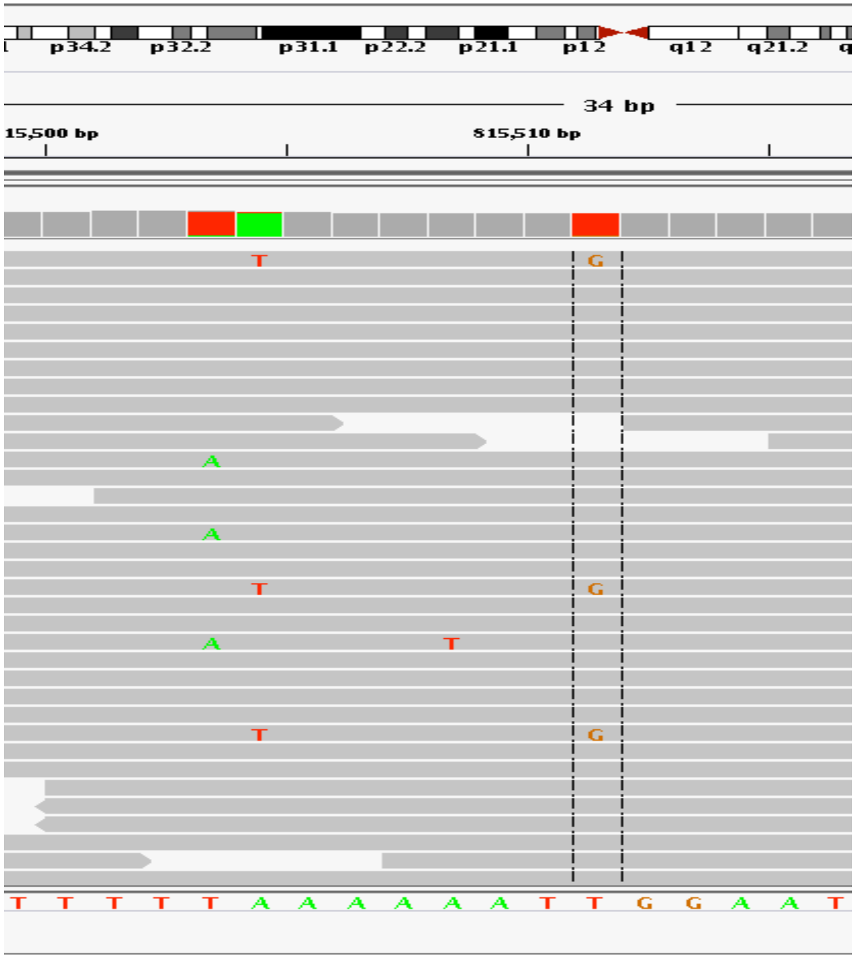
Michael Lawrence

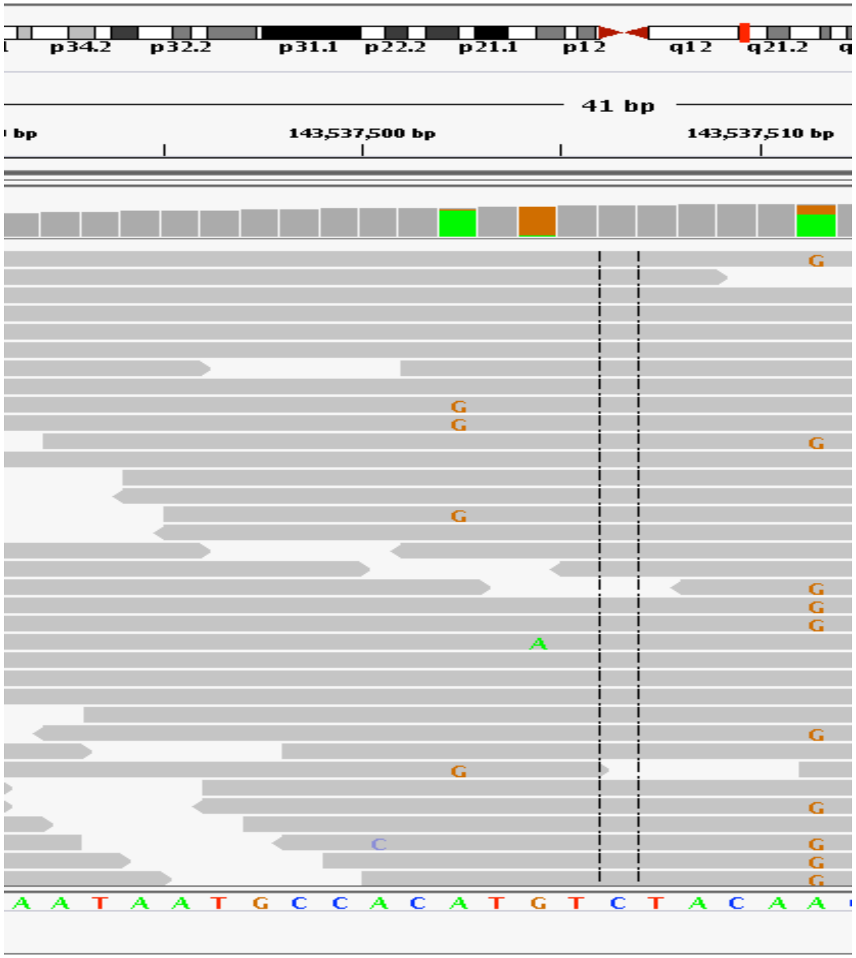
# FNR by Mixture and Coverage



Genentech
A Member of the Roche Group

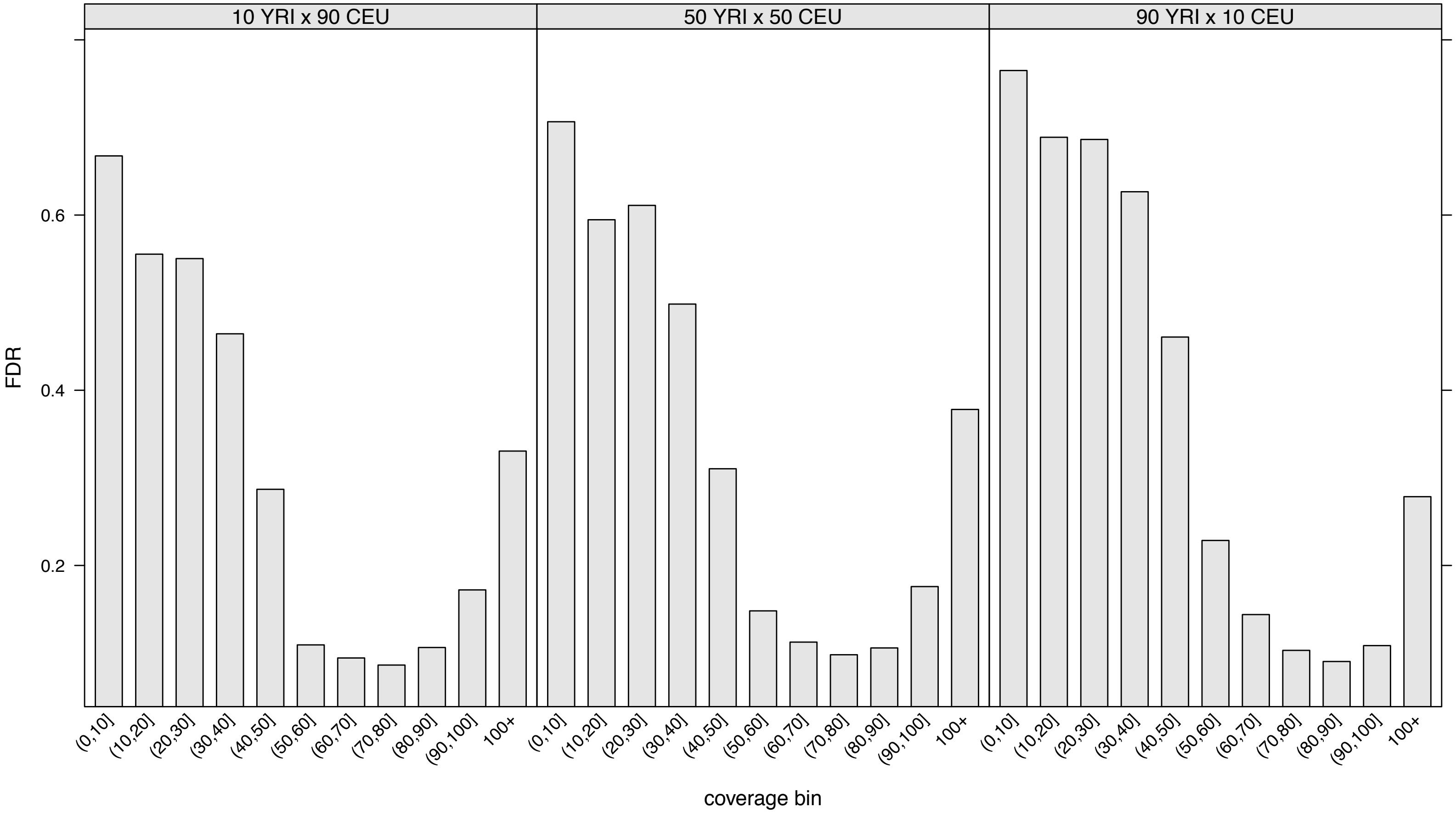
# What is going on in high coverage?





Michael Lawrence

# FDR rates by coverage



Genentech
A Member of the Roche Group

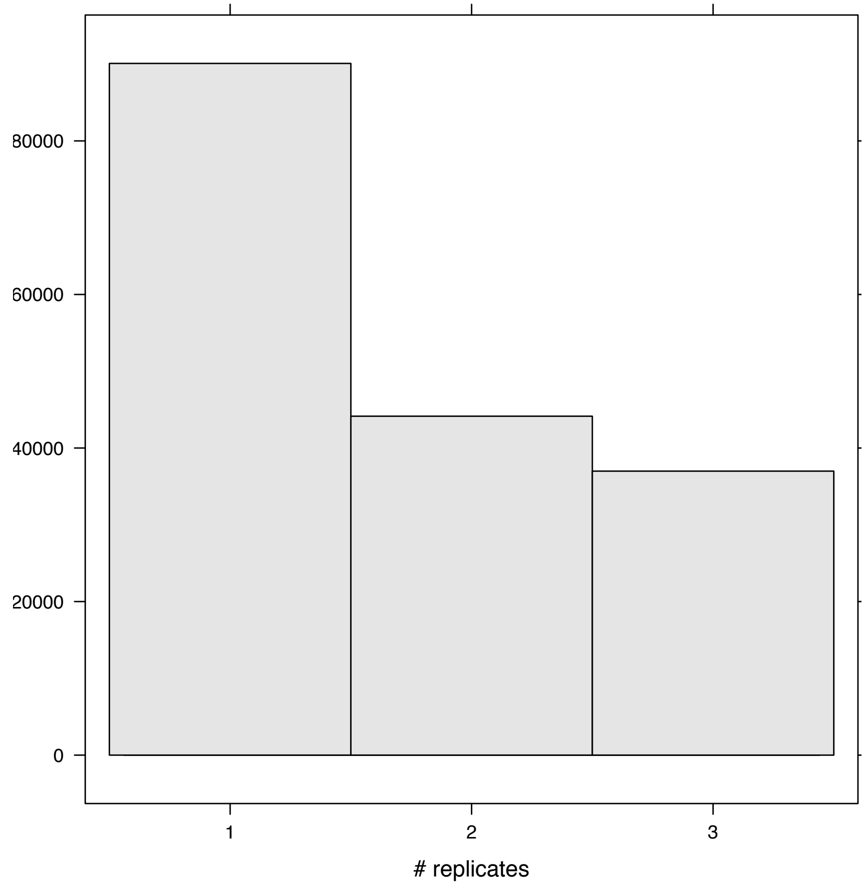
# How did we do

- Based on Chr1 we have 1-FNR = 0.91
- And FDR of about 19%
- But, we believe about 1/3 of the FPs are probably TPs
- We are still trying to determine how many of the FNs are TNs

# FP/FDR

- The data are pointing to the fact that the reference genomes (our gold standard) is not that accurate.
- Thus many presumed FPs are in fact TPs, but were missed for a variety of reasons in the original genotypes.
- We also see strong evidence that the YRI genome is less well determined than the CEU.

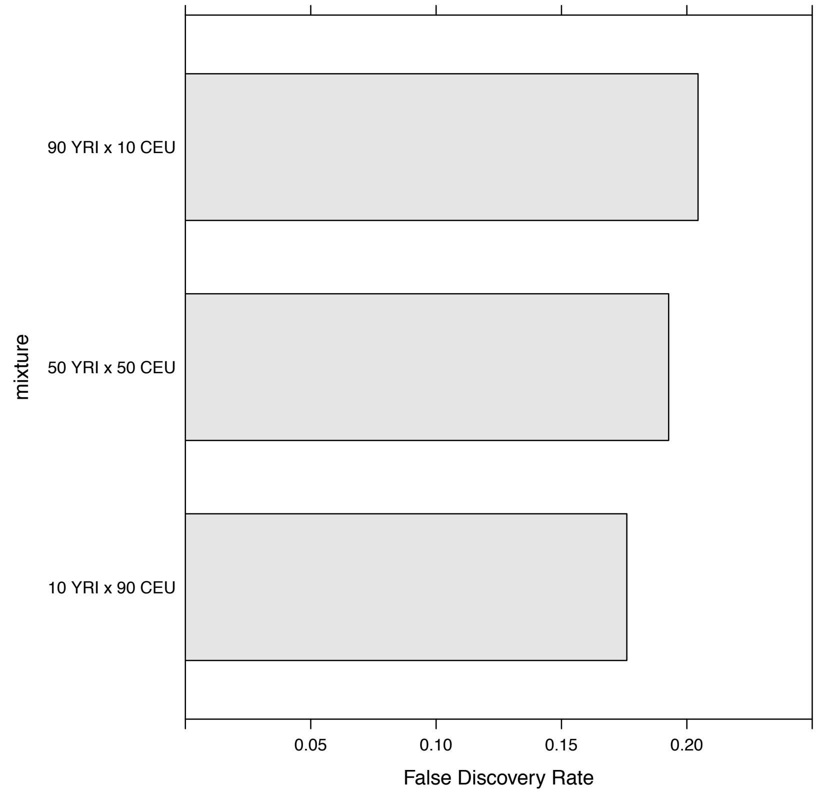
# Are our FPs really F?



- We see strong association between a variant being in dbSNP and whether or not it was an FP more than once.

| | dbSNP No | dbSNP Yes |
|---|---|---|
| **Rep 1** | 80003 | 10090 |
| **Rep 2/3** | 25052 | 56085 |

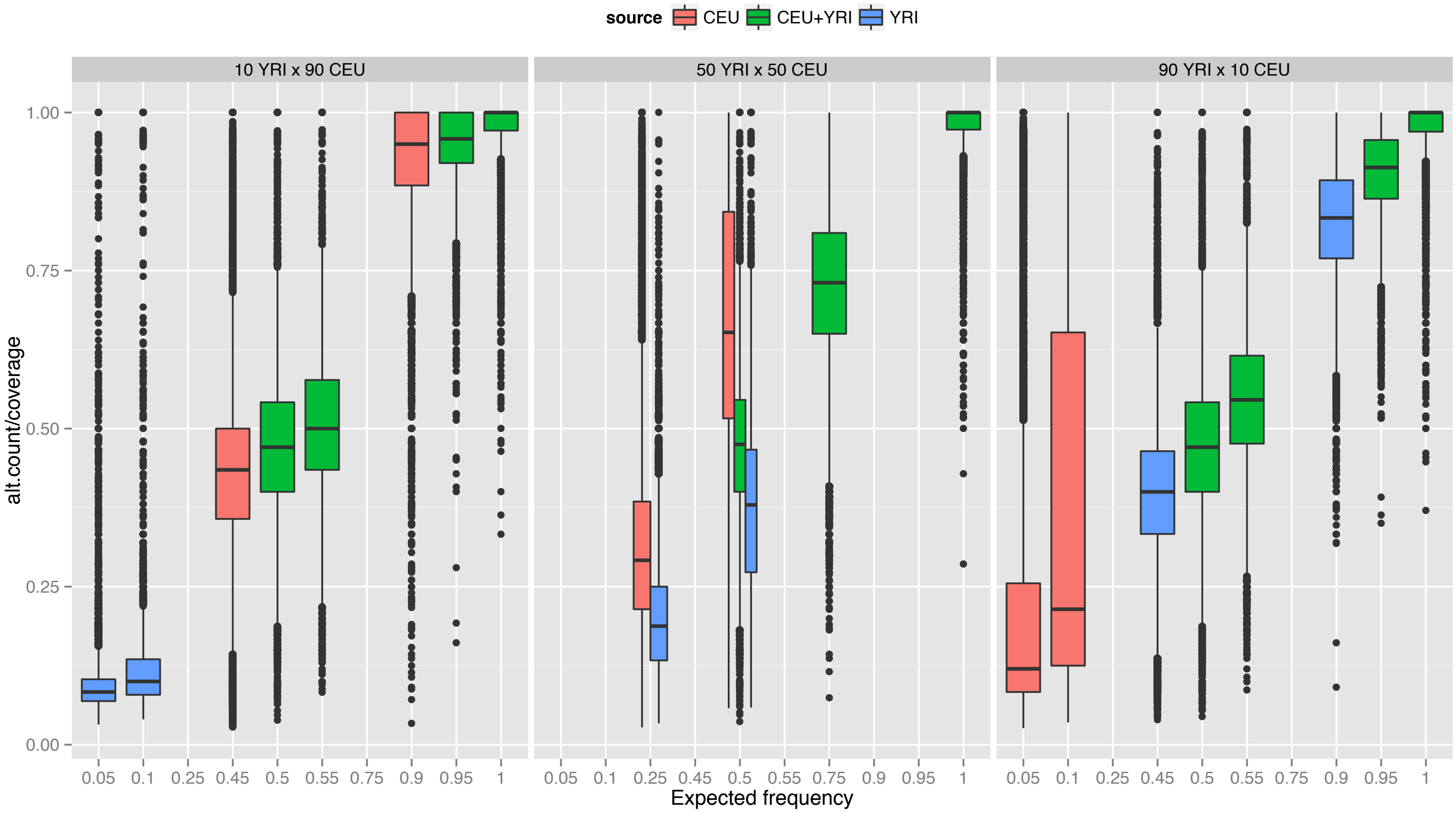
# How good is the YRI sample?



- We see that the FDR increases as the fraction of YRI increases.
- What else?

Michael Lawrence

# Observed Variant Frequencies



Genentech
A Member of the Roche Group

# What we cannot do

- APC: adenomatous polyposis coli,
  - A tumor suppressor, often mutated in cancer
  - Length 10740 nt
- WT calls: can we say the gene has no mutations/variants?
  - If we have power to detect a variant of 0.999
  - If each locus is independent then for the gene we have power of 0.999^10740=2.154485e-05
  - We need power around 0.99999 per variant (and much more for longer genes) to get power around 0.9
  - For a Binomial, p=0.1, we will need about 120 X coverage (minimum over the gene/genome depending on what you want to say)

# What we cannot do

- We currently do not phase (call haplotypes)
- Since the genomes are typically diploid (or greater for cancer) we cannot easily determine whether variants are in the same allele or in different alleles
  - Unless they are very close together
- For most variants we do not have good measures of their effect
  - Condel and similar can be used, but these are not the best tools
  - Finding the effect of a variant is challenging

# Discussion

- A large and comprehensive gold standard data set is an essential tool in improving variant calling
- With hundreds of thousands/millions of TP and TN we can study many aspects of the process
- We still need biochemical validation (being done now)

# Acknowledgements

- Michael Lawrence
- Melanie Huntley
- Yi Cao
- Jeremiah Degenhardt
- Sekar Sheshigiri
- Eric Stawiski
- Jens Reeder
- Gregoire Pau