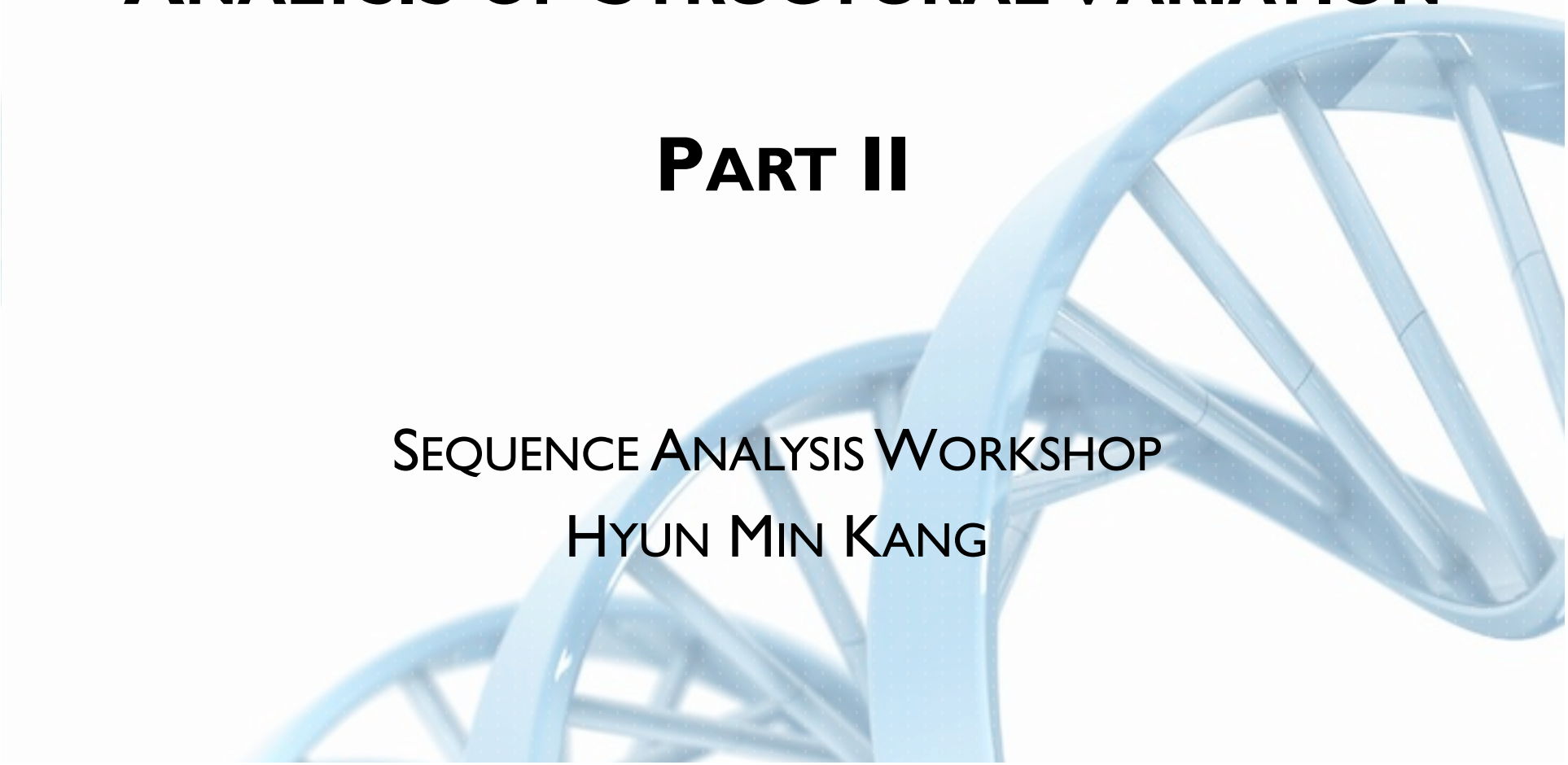
Analysis of Structural Variation

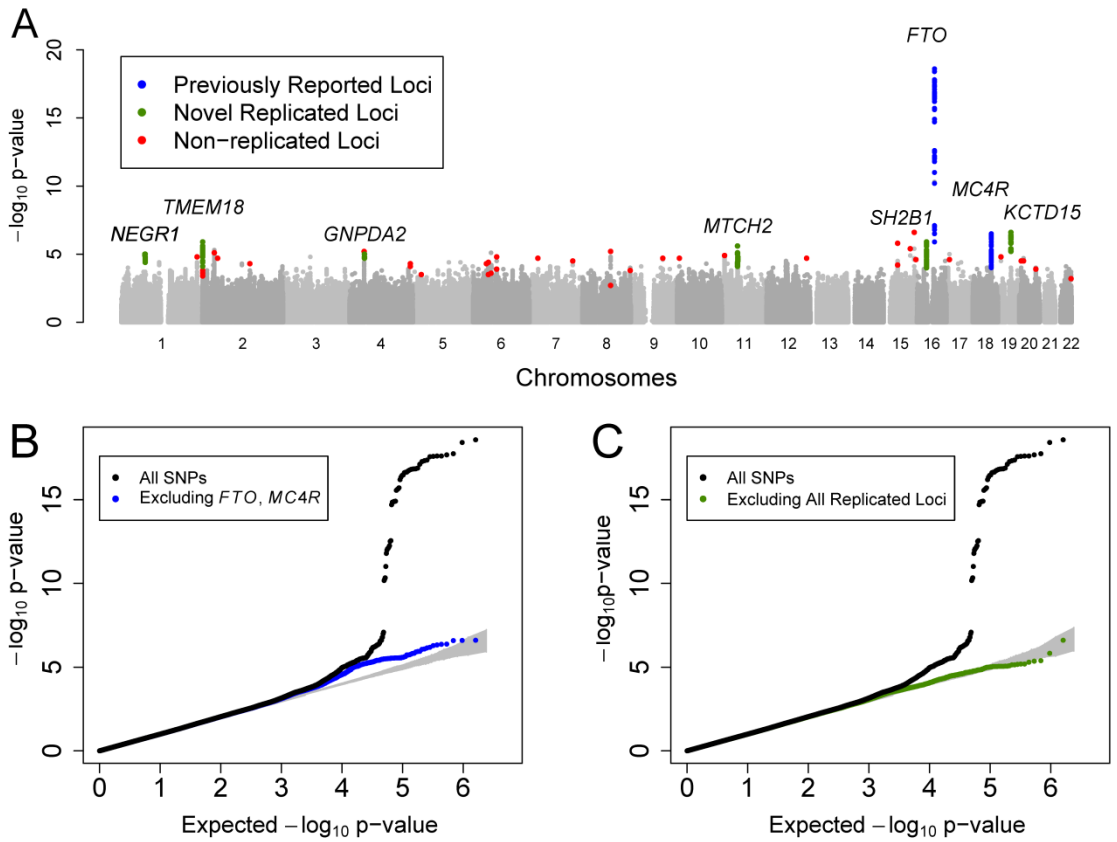
Part II



Sequence Analysis Workshop

Hyun Min Kang

Copy Number Variants Are Important Examples from Genetics of Obesity

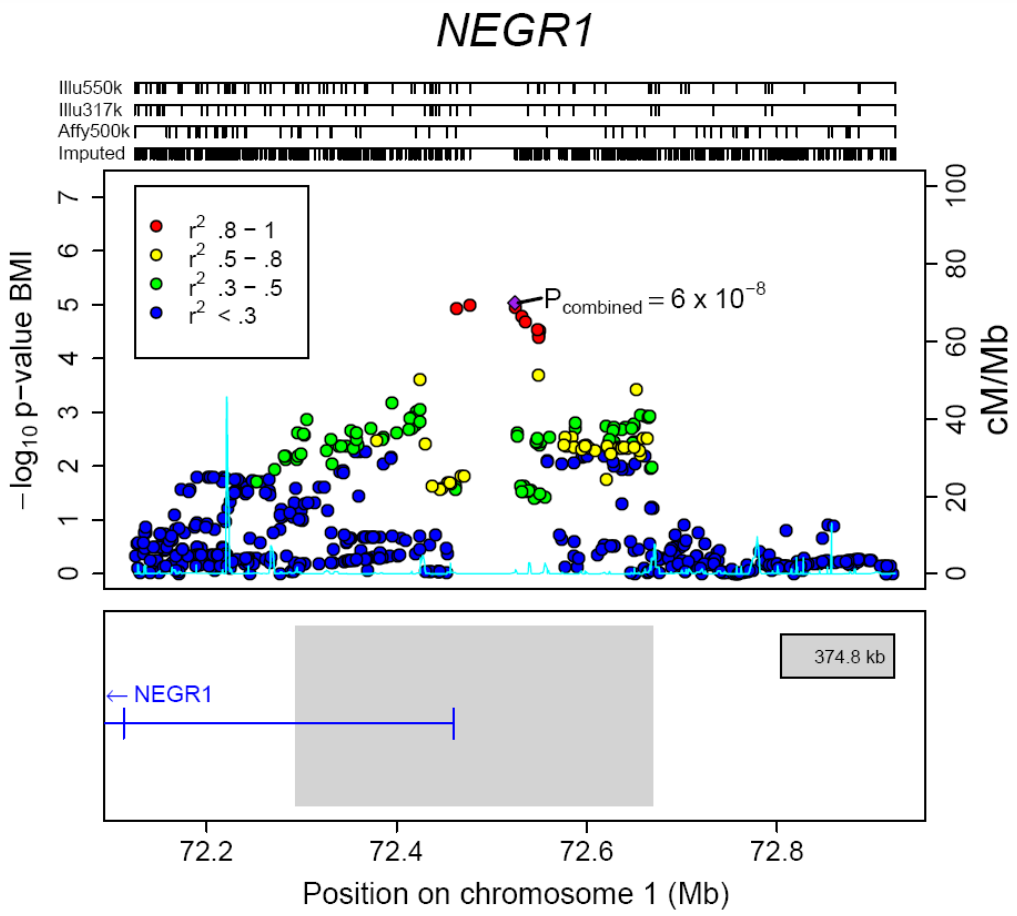


Seven of eight confirmed BMI loci show strongest expression in the brain…

Willer et al, *Nature Genetics,* 2009

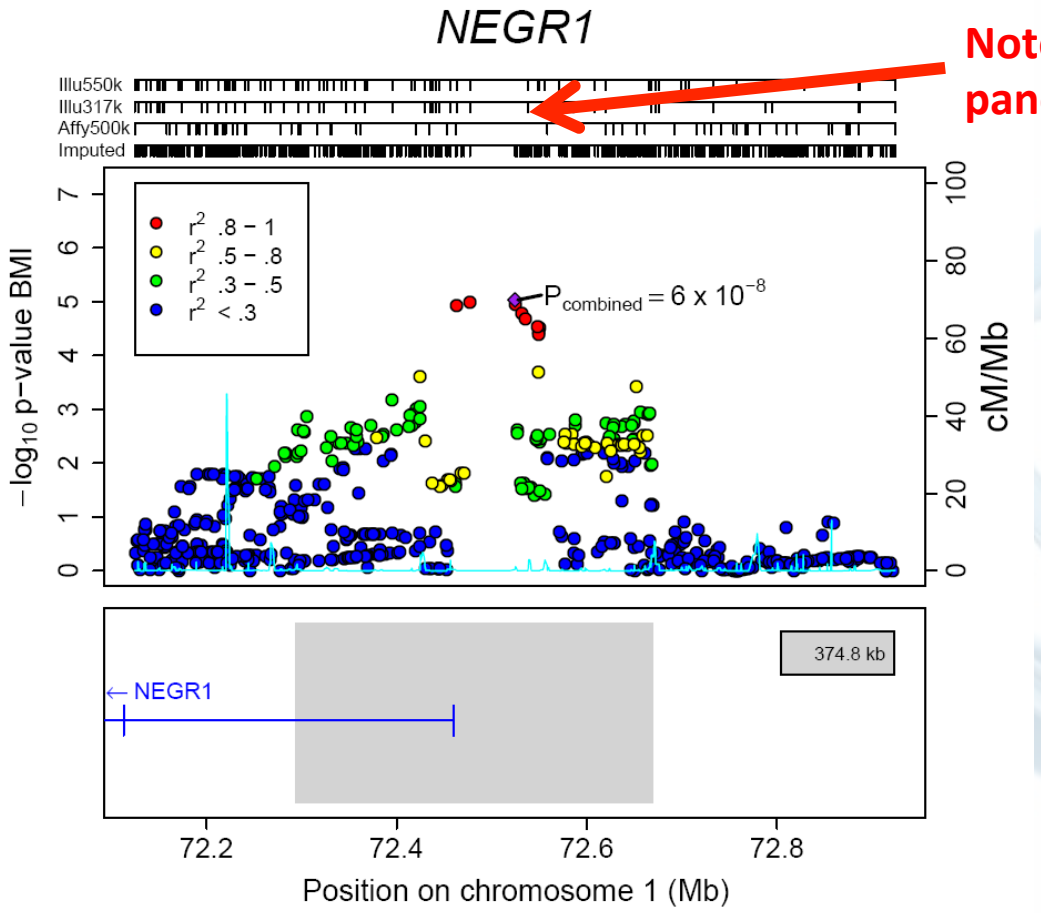
Some BMI Associated SNP Haplotypes

...



Willer et al, *Nature Genetics,* 2009

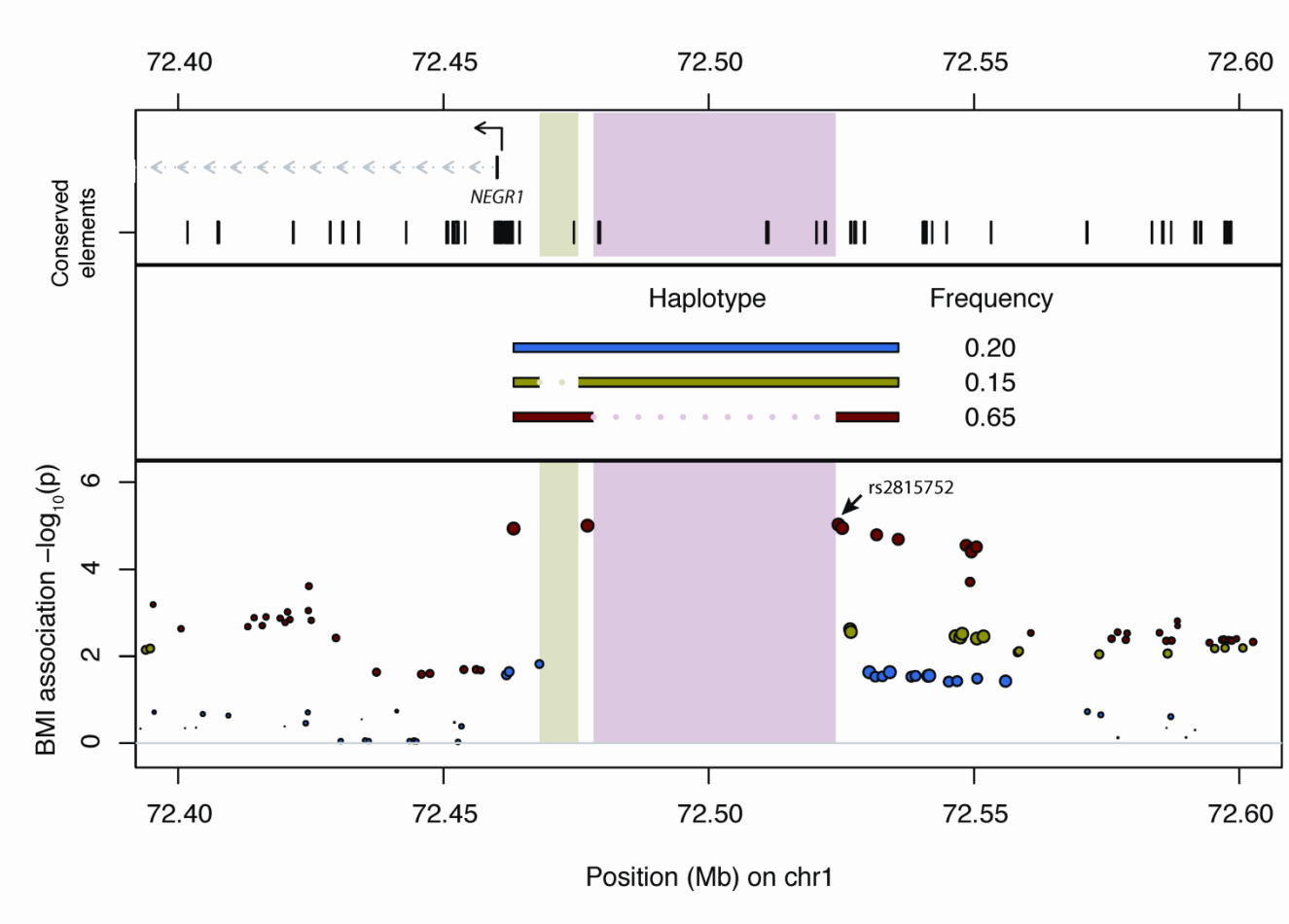
… Turn Out To Be Unusual …



Note hole in marker panels….

Willer et al, *Nature Genetics,* 2009

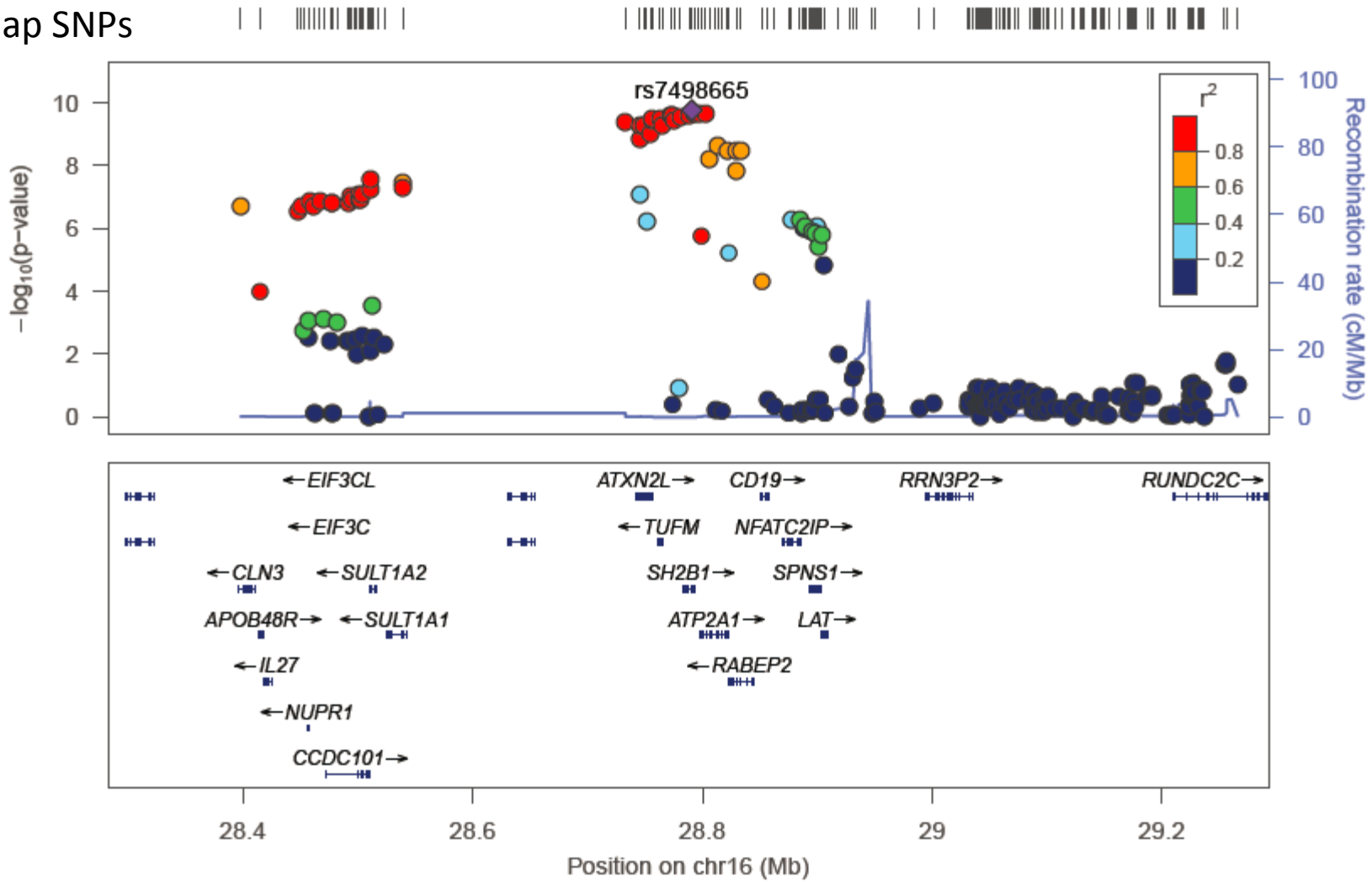
And Carry Unusual Deletion



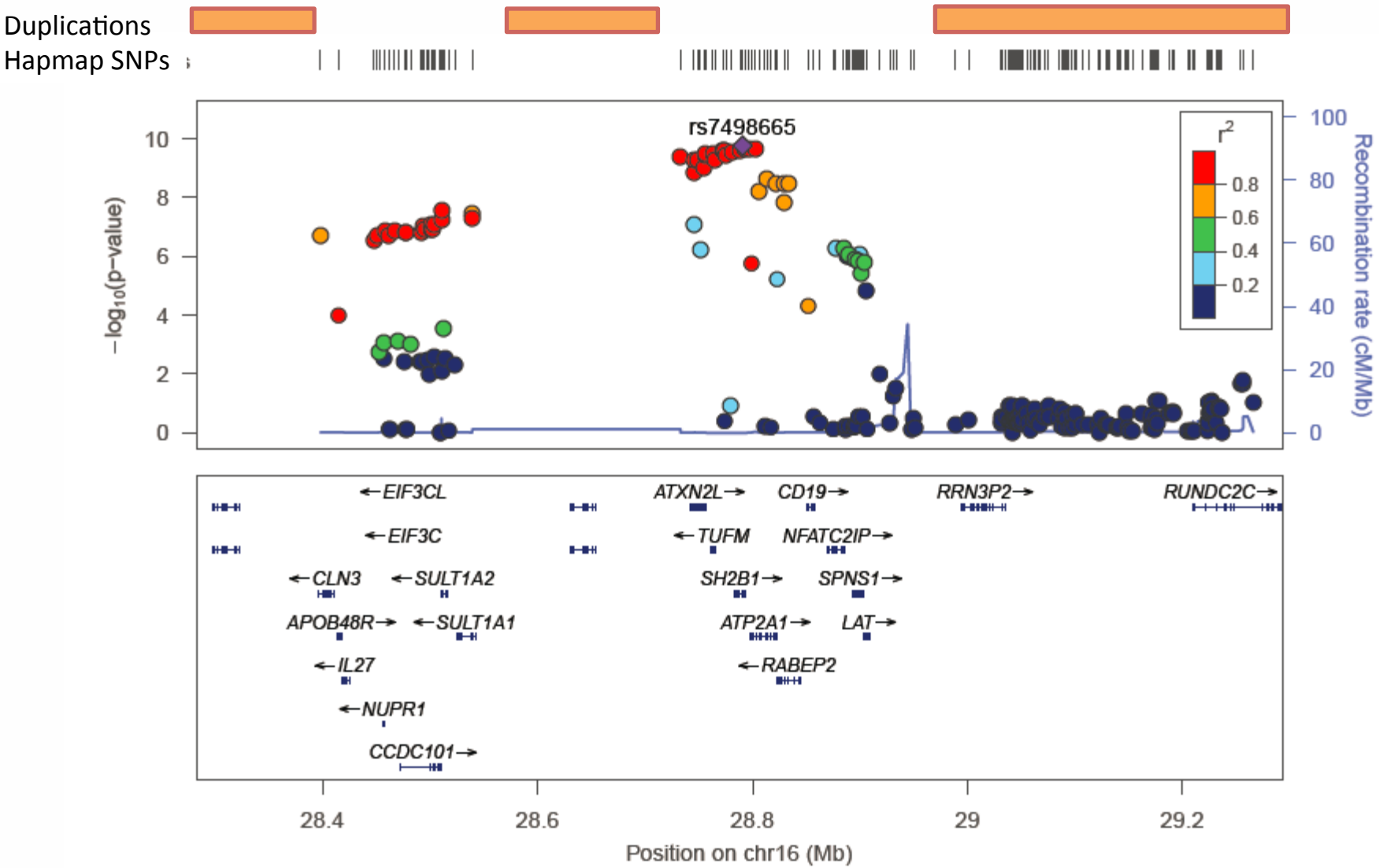
Willer et al, *Nature Genetics,* 2009

OTHER BMI ASSOCIATED HAPLOTYPES…

HapMap SNPs



…Point To Complex Genomic Regions…



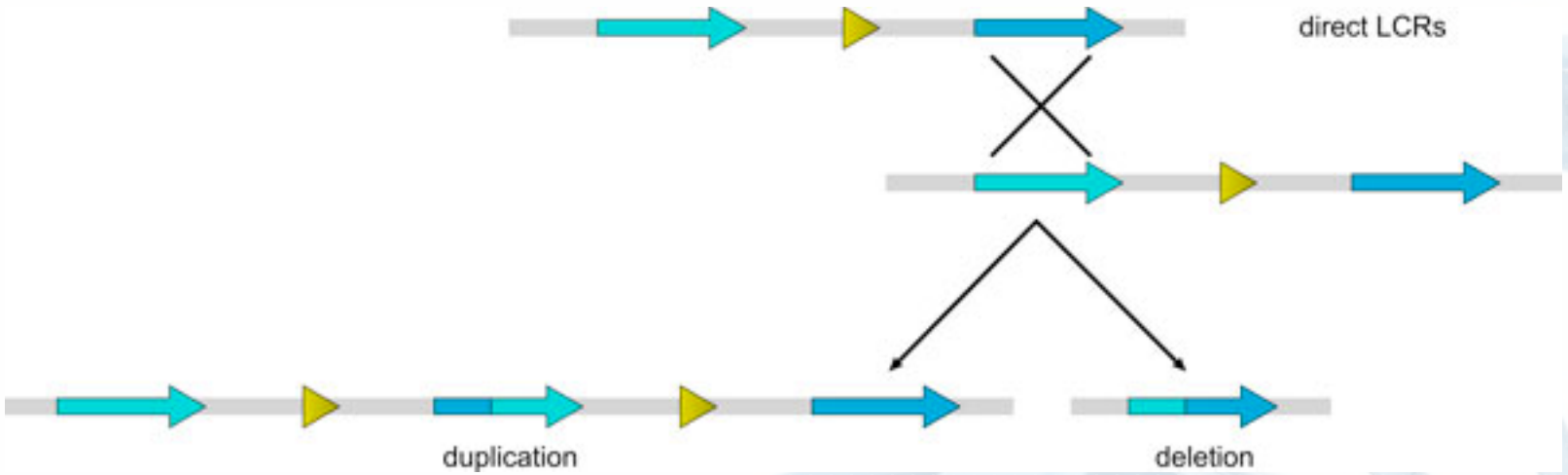
… With Large Obesity Associated Deletions

- Bochukova et al (Nature, 2009) showed that large deletions were present in 1.0 – 2.0% of children with severe obesity, 0.5% of controls

- The most common of these deletions was observed in 5 of 300 children with severe obesity but only 2 of 7,366 controls

- Similar patterns reported for other complex traits, including autism and schizophrenia

Many CNVs Are Contributed by Non-Allelic Homologous Recombination (NAHR)



SHOTGUN SEQUENCE READS

- Typical short read might be <25-100 bp long and not very informative on its own
- Reads must be arranged (*aligned*) relative to each other to reconstruct longer sequences

PAIRED END SEQUENCING

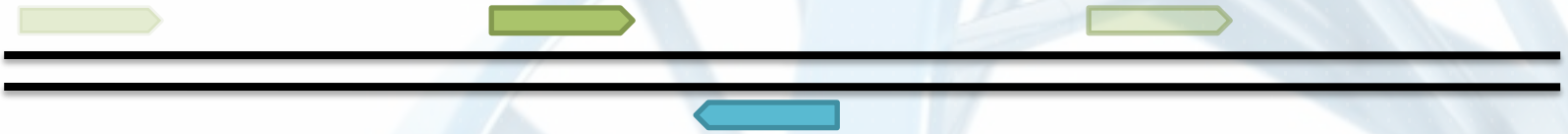
Population of DNA fragments of known size (mean + stdev)
Paired end sequences

PAIRED END SEQUENCING

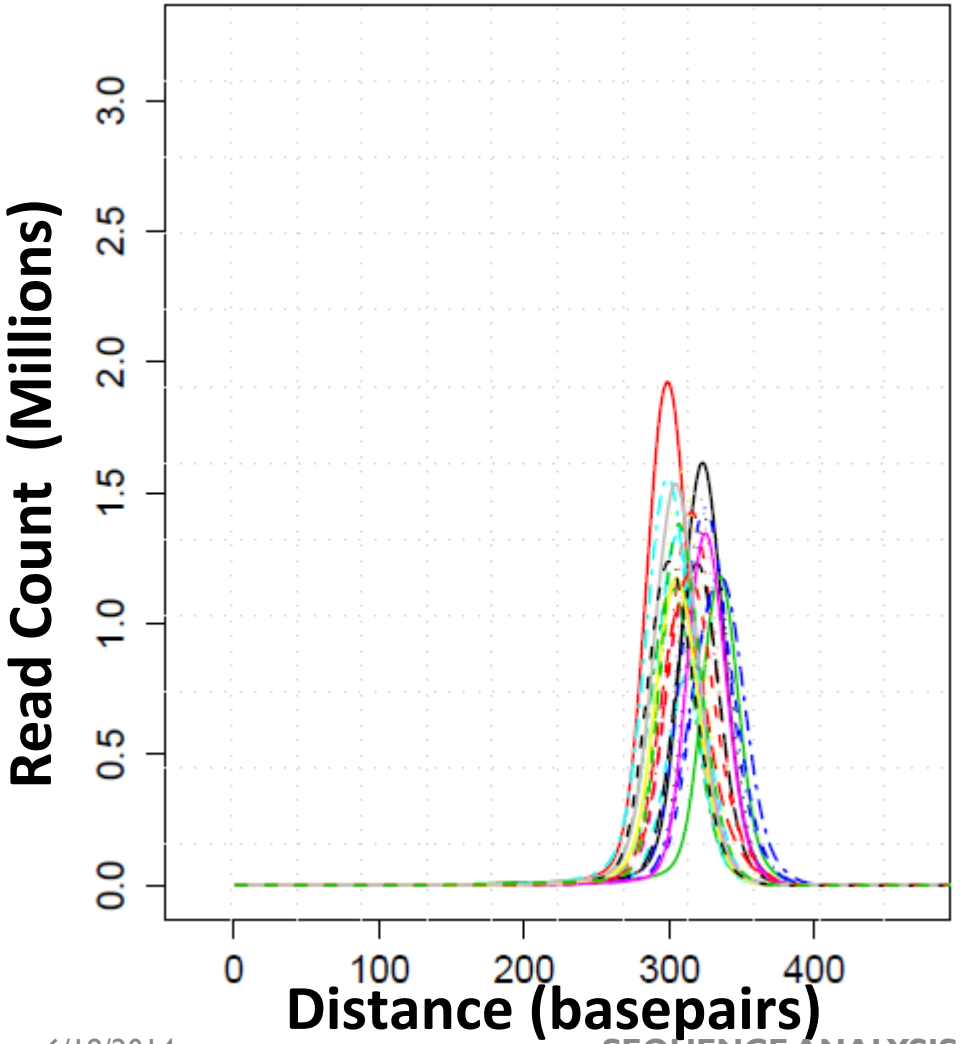
Paired Reads

Initial alignment to the reference genome

Paired end resolution



Distance Between Pair Ends



- The graph shows distance between paired end reads
- Data summarized across 24 samples
- Courtesy: Xiaowei Zhan, University of Michigan DNA Sequencing Core

Evidence For A Deletion Within A Single Individual

- Split Reads
- Read Pair Separation
- Read Depth

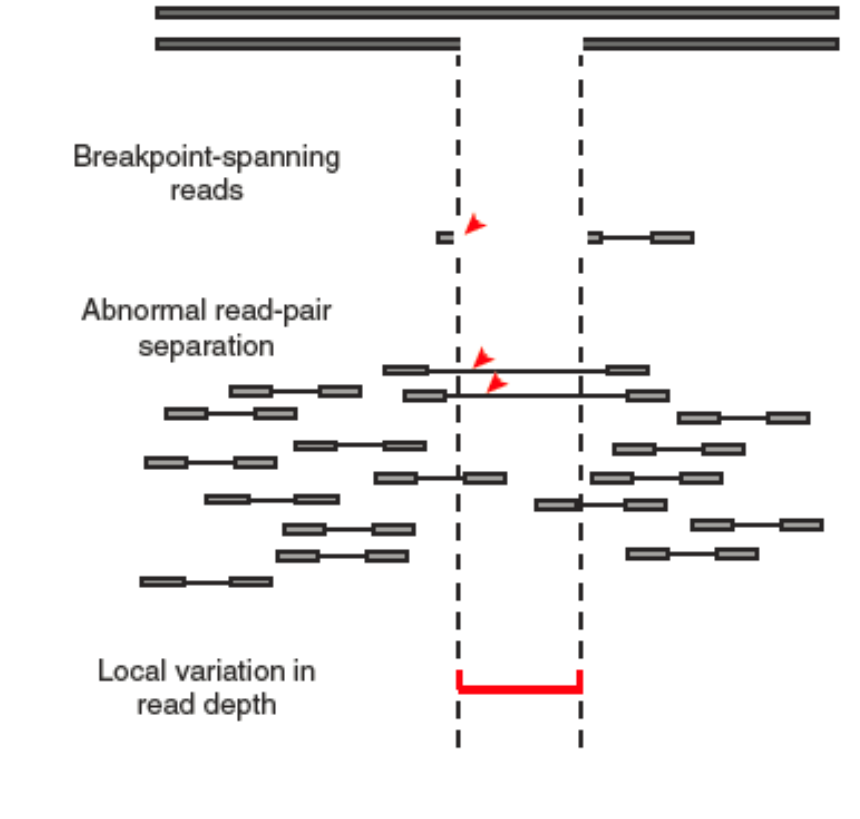
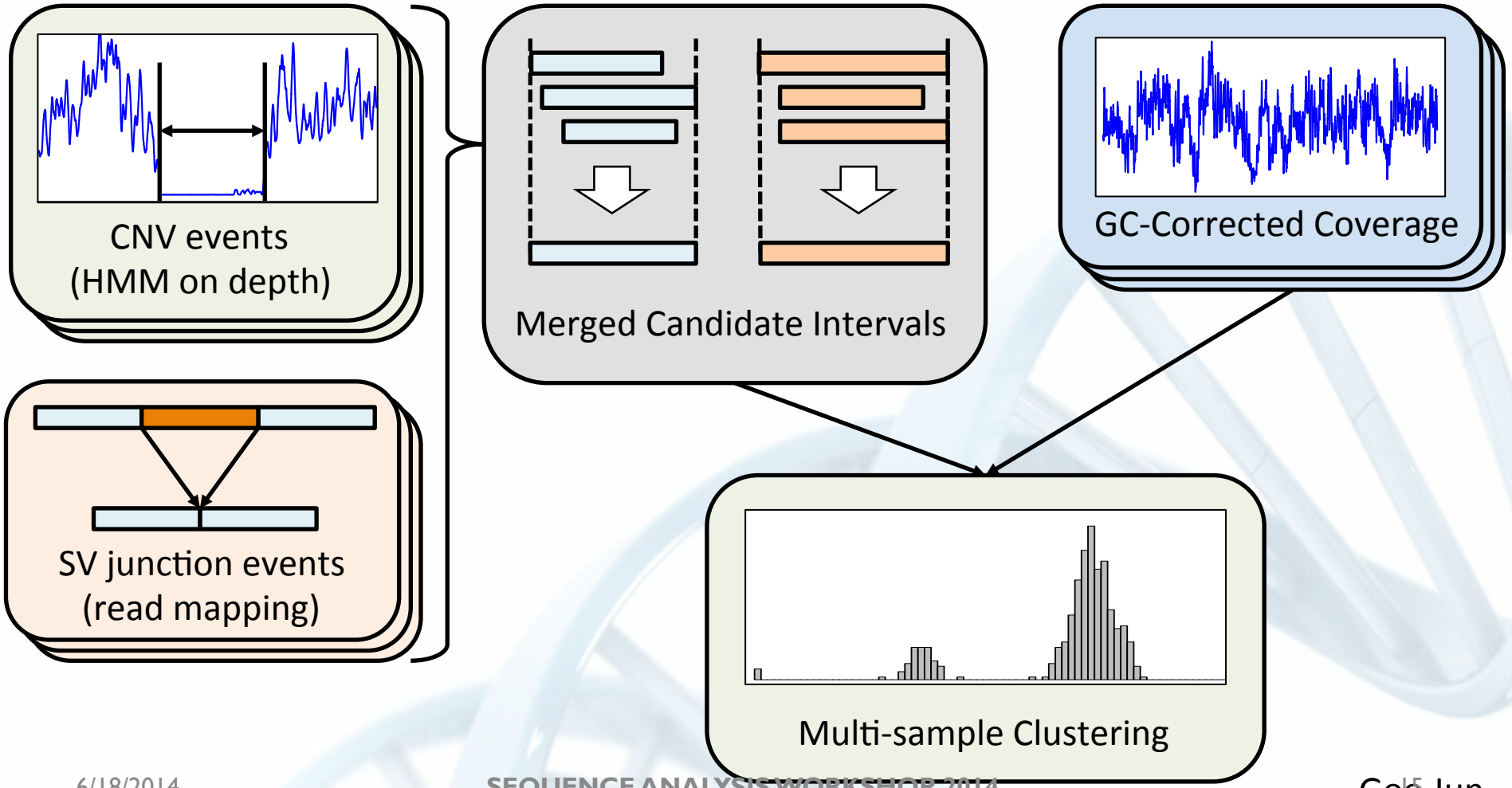


Figure from Handsaker et al (2011)

MULTI-SAMPLE CNV (DELETIONS) CALLING



Detecting Copy Number Variation Based on Read Depths

- Focus on a particular feature of the data
  - e.g., read depth
- Normalize depth for each individual
  - e.g., adjust for total read count
  - e.g., adjust for GC content specific read count
- Model data as a mixture of distributions, characterized using maximum likelihood

Sara Rashkin

Detecting Copy Number Variation Based on Read Depths

$$d_i \sim p_0\mathcal{N}(\mu_0, \sigma_0^2) + p_1\mathcal{N}(\mu_1, \sigma_1^2) + p_2\mathcal{N}(\mu_2, \sigma_2^2)$$

Where

d_i is the depth for individual i

p_j is the frequency of individuals with j deletions (assuming Hardy Weinberg Equilibrium)

μ_j and σ_j^2 are the mean and variance of adjusted read depth distribution for deletion count j

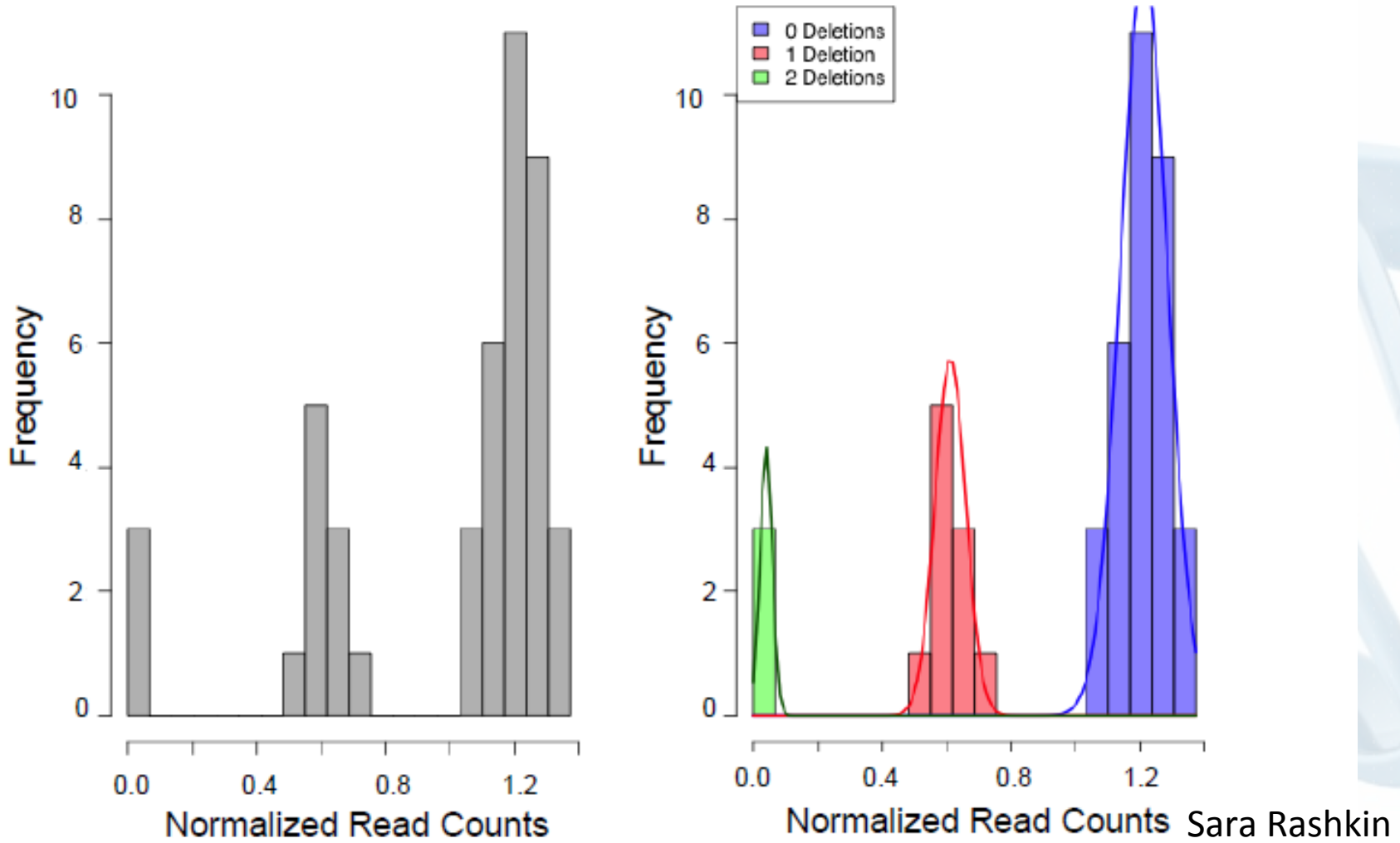
Detecting Copy Number Variation Based on Read Depths

- To estimate a deletion model, maximize

$$L(d_i) = \sum_j p_j (2\pi)^{\frac{1}{2}} \sigma_j^{-1} \exp\left[-\frac{(d_j - \mu_j)^2}{2\sigma_j^2} \right]$$

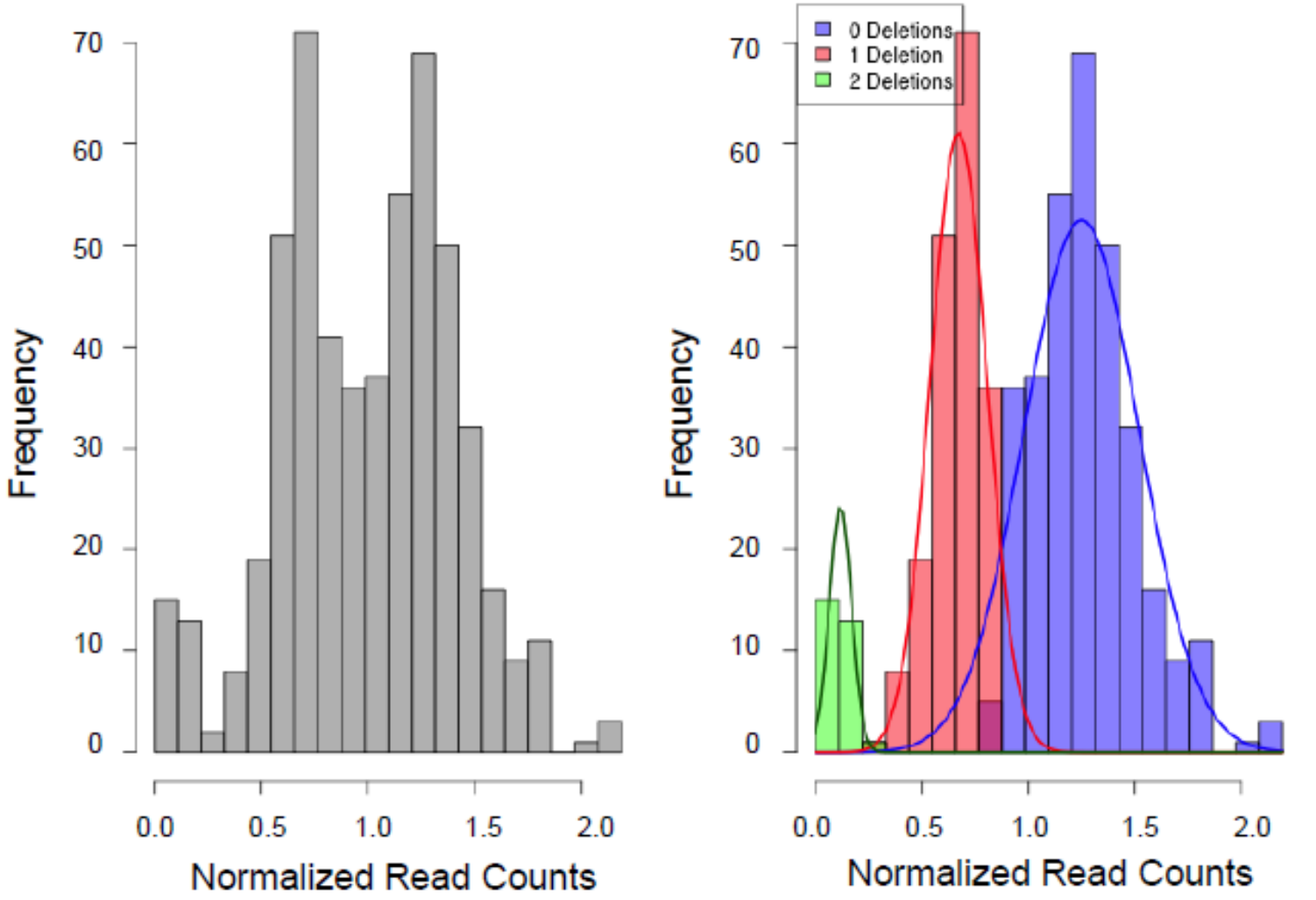
- To keep number of parameters modest, we use HWE for modeling (one parameter for three frequencies) and can impose additional structure on means and variances

Well Separated Region



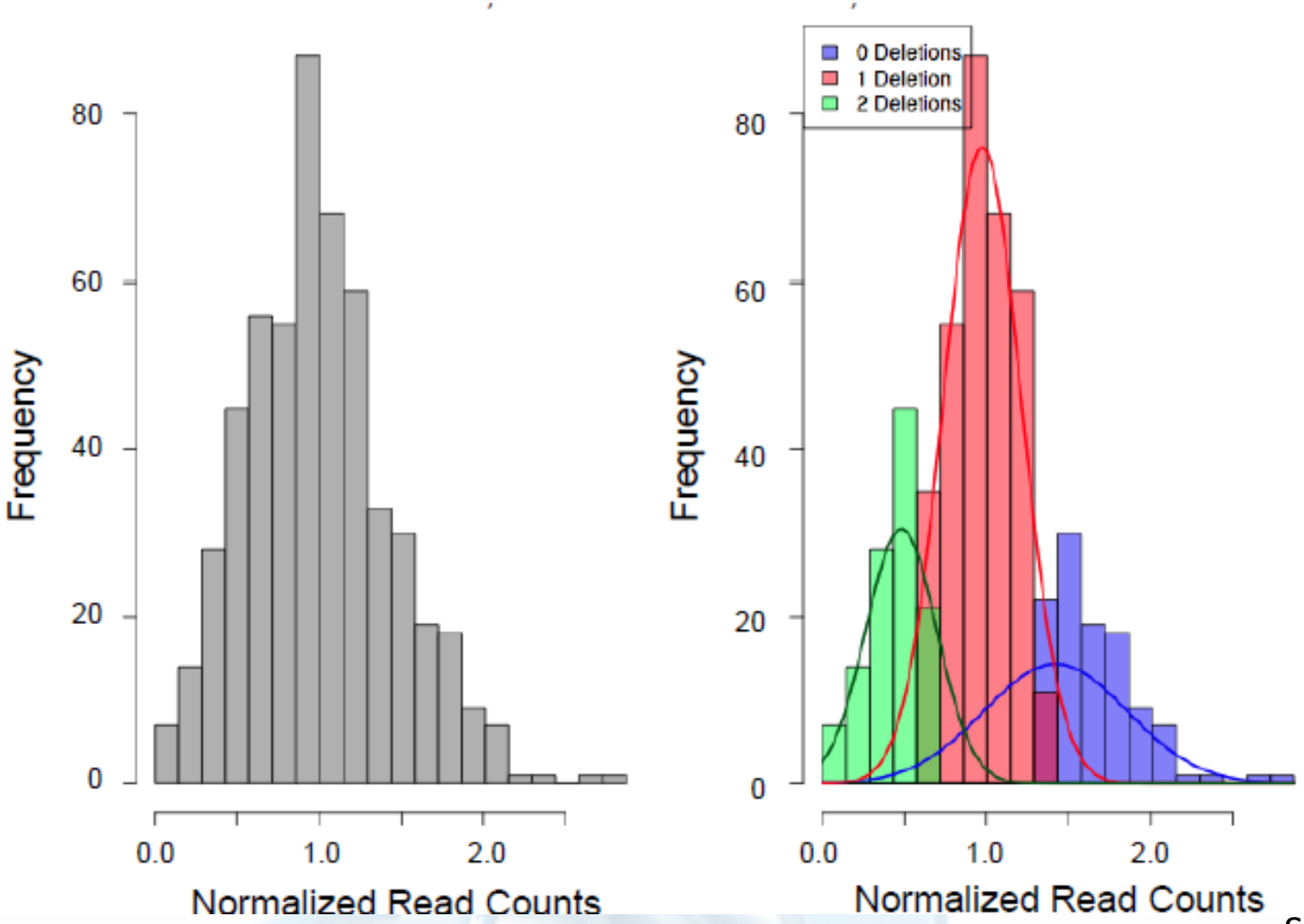
Sara Rashkin

MODERATELY SEPARATED REGION



Sara Rashkin

Hard to Call Region



Sara Rashkin

CLUSTER EVALUATION - # OF COMPONENTS

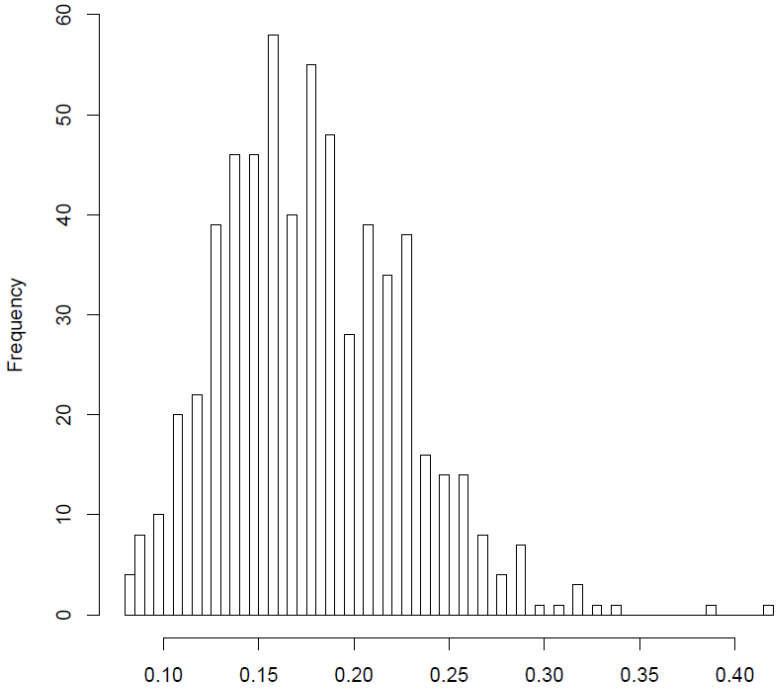
- Bayesian Information Criterion (BIC)

$$\mathrm{BIC} = -2\ \mathrm{LLK} + k \ln(n)$$

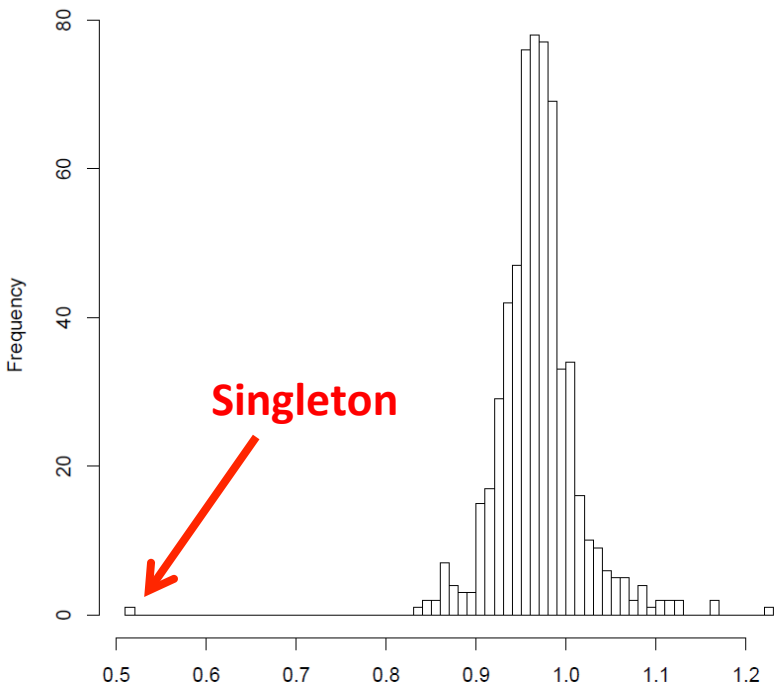
  - LLK : Log-likelihood
  - k : number of clusters
  - n : number of samples

- Fit Gaussian mixtures with 1, 2 and 3 components, and compare BIC of 1, 2, 3 component models to decide number of clusters.

BIC Example

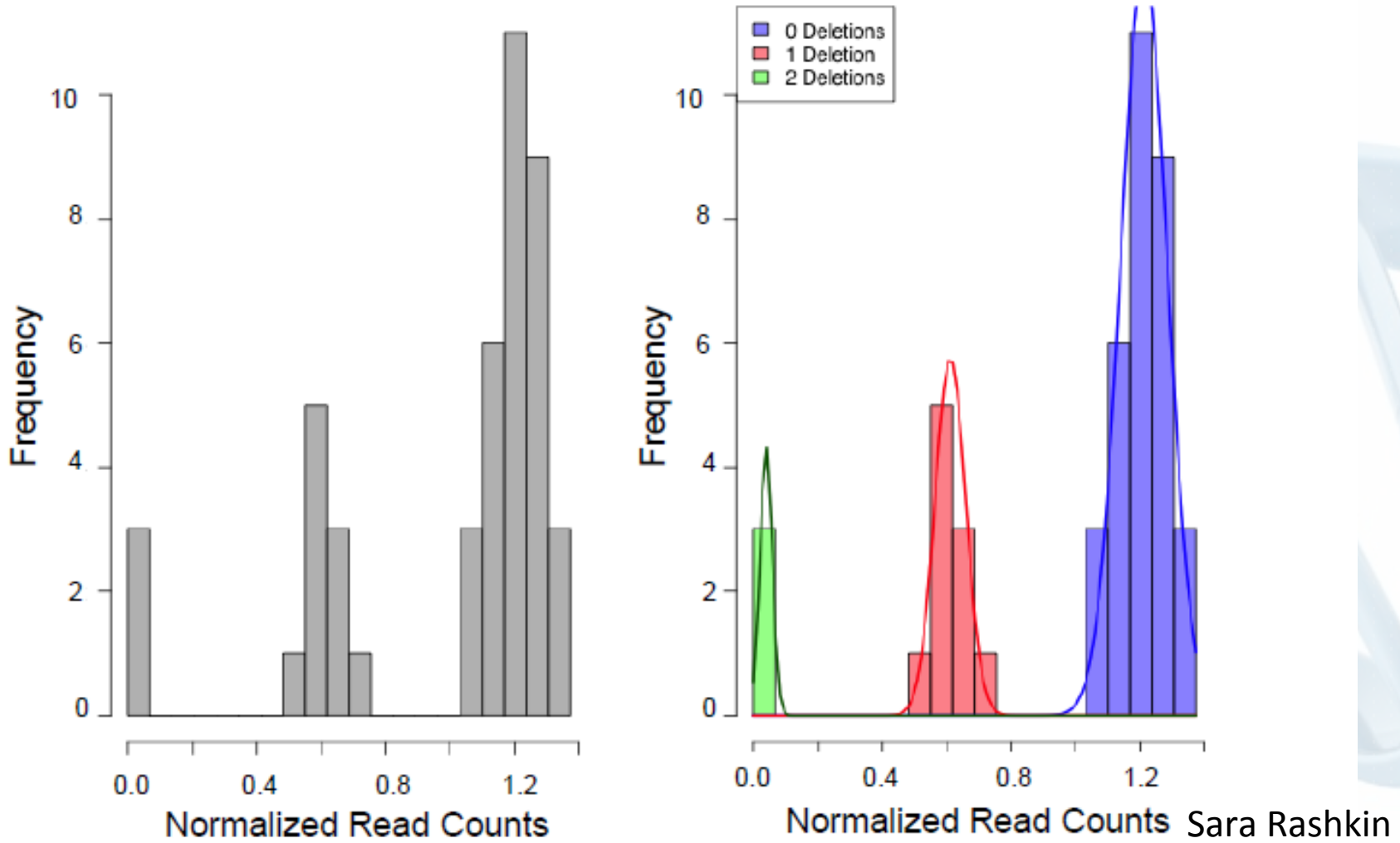


BIC chooses 1 component



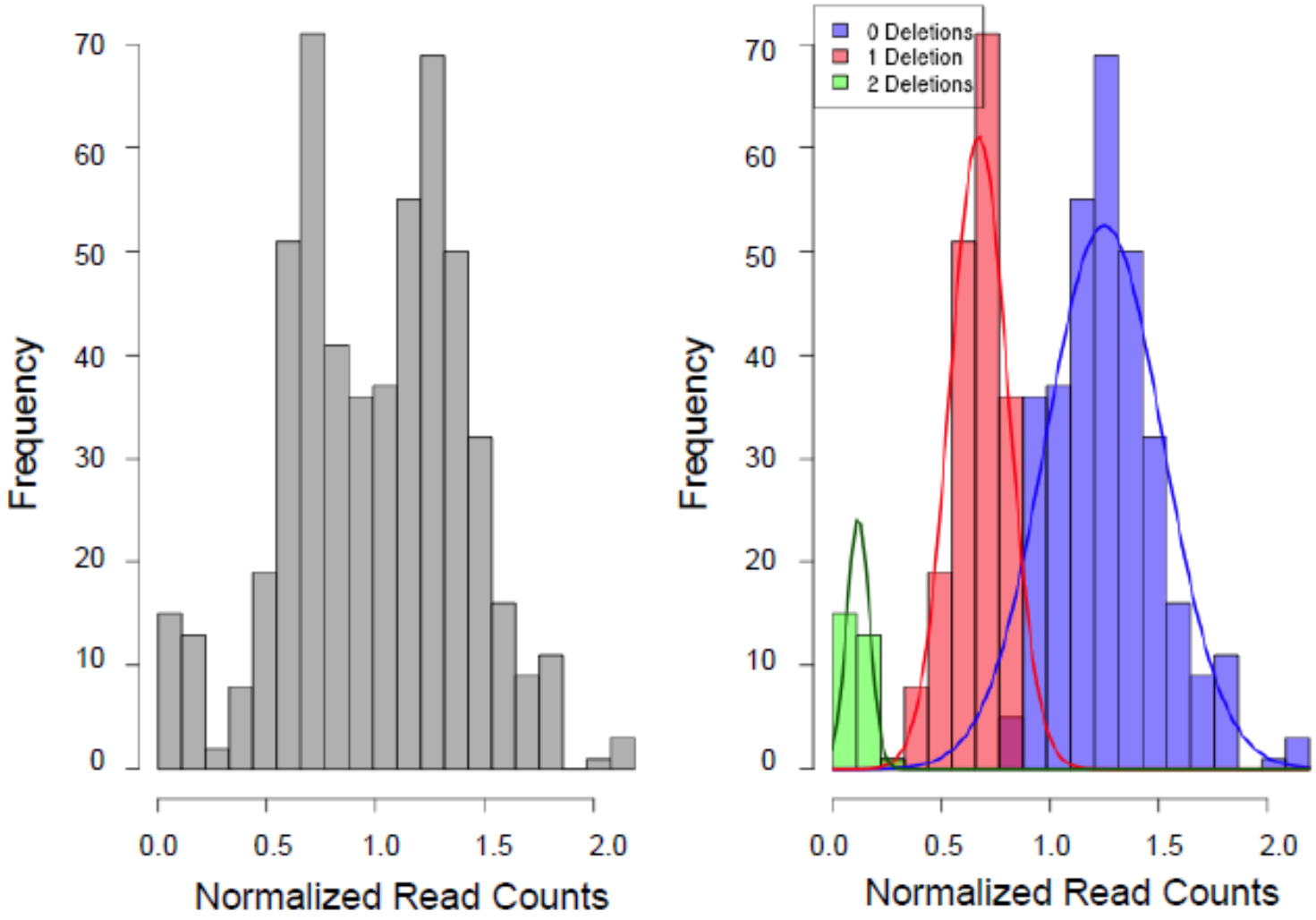
BIC chooses 2 components

Well Separated Region



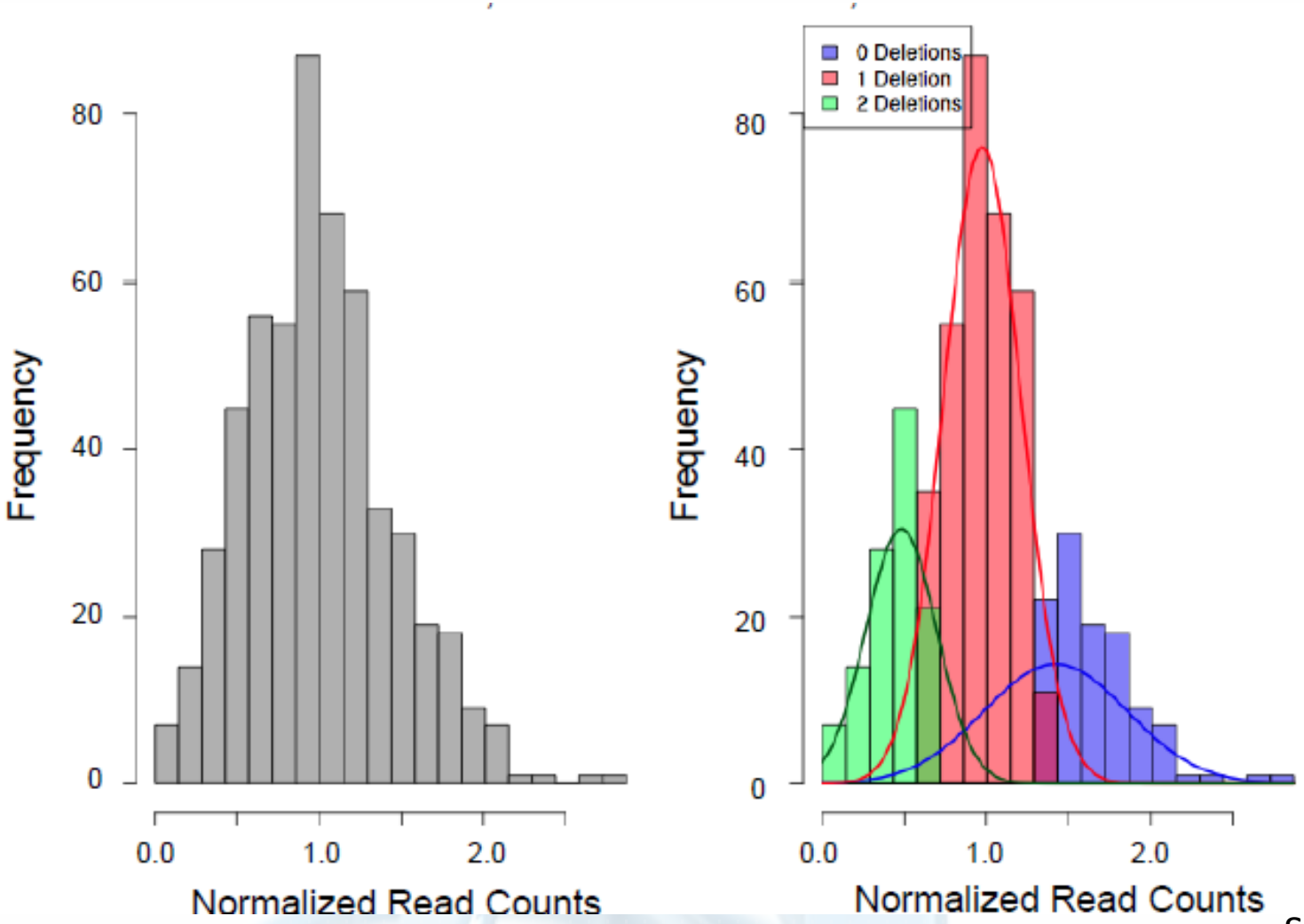
Sara Rashkin

Moderately Separated Region



Sara Rashkin

Hard to Call Region



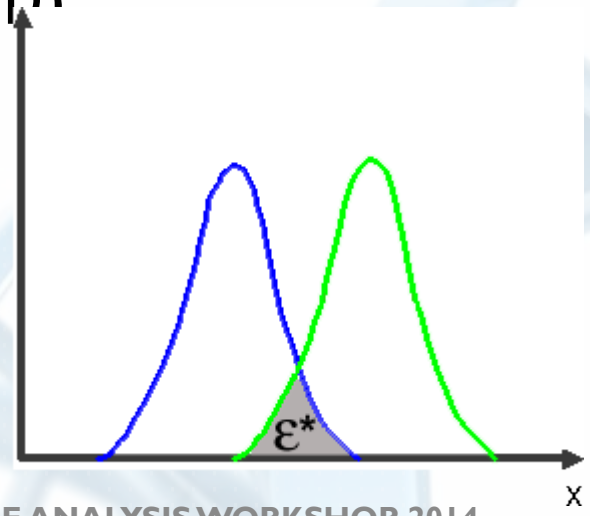
Sara Rashkin

CLUSTER EVALUATION – OVERLAP

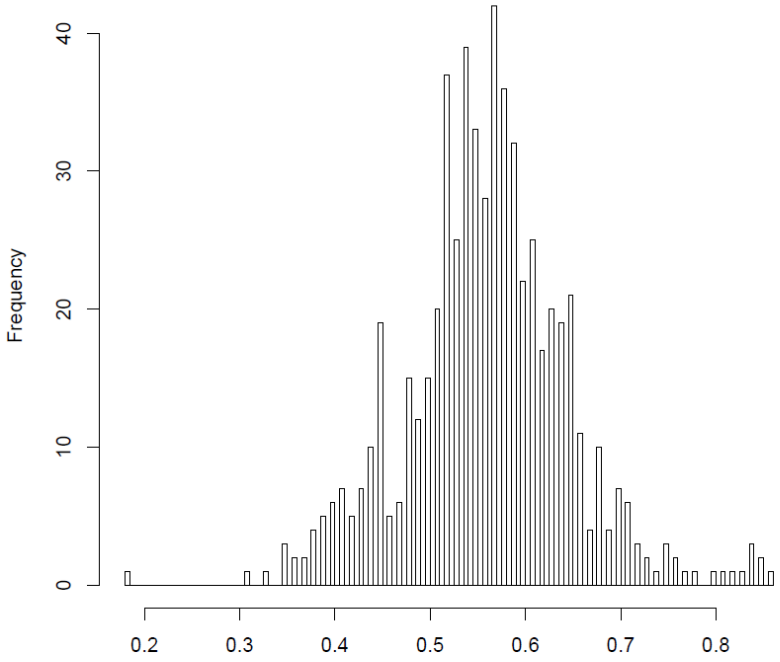
- Evaluation of cluster separation
  - Unavoidable error : overlap between two distributions
  - Bayes error rate ~ Bhattacharyya coefficient
  - For two Gaussian distributions,

    $D = (\mu_1 - \mu_2)^2 / (8\ \sigma_{avg}^2) + (1/2) \log [\sigma_{avg} / \mathrm{sqrt}(\sigma_a \sigma_a)]$

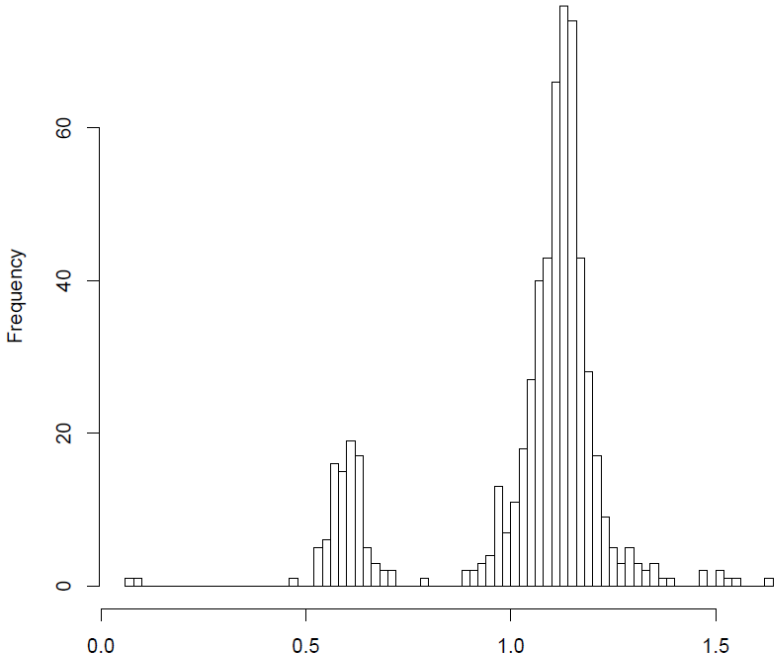
    $P(\text{Overlap}) = \exp(-D)$



P(OVERLAP) EXAMPLE

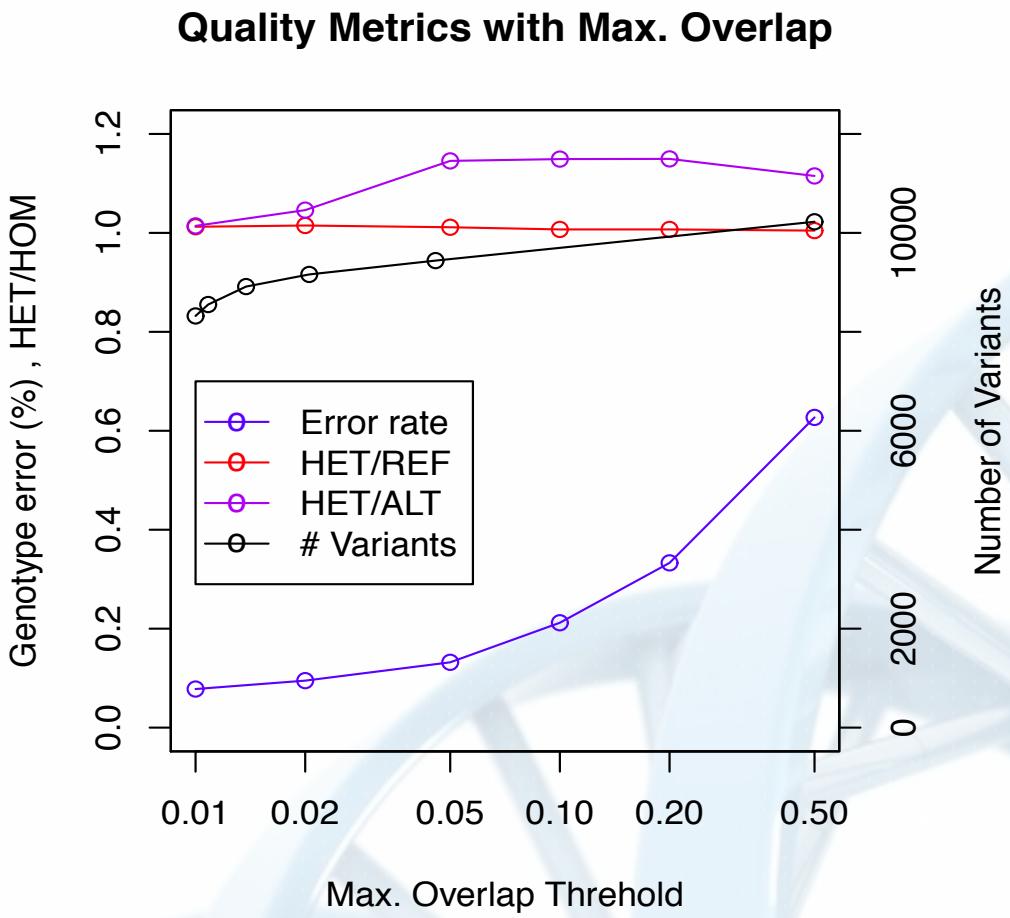


max exp(-D) = 0.933



max exp(-D) = 0.0032

SENSITIVITY/ERROR VS P(OVERLAP)



1000G CNV Default CG vs Multi-sample

	CG Default	Multi-sample	SNPs
# Large Deletions	2,374	**8,321**	-
Call Rate (%)	95.2	**99.9**	96.73
Merlin-estimated Error Rate (%)	>10	**0.078**	0.494
Trio HET/HomREF	0.750	**1.012**	0.962
Trio HET/HomALT	1.055	**1.014**	1.001

Challenges in Read Depth Based Calling

- Ideal if number of reads per region is large
- As technologies improve and reads get longer …
- … read depth based calling becomes harder
- Important to integrate different types of signal!

Evidence at the Population Level

- Allele Shared Between Multiple Individuals
  - Multiple individuals show cluster of reads with unusual separation in the same location
- Evidence for Deletion Recurs in the Same Individuals
  - Individuals with one unusually separated pair of reads, likely to show additional nearby read pairs with unusual separation
- Evidence for Reference Allele Decreases as Evidence for Deletion Increases
  - When the number of reads with unusual separation increases, the number of nearby reads with expected separation decreases
- Deletions Segregate on Specific Haplotypes

REFINED ALGORITHM

- Build list of candidate variants by finding read pairs with abnormal separation
- Focus on regions supported by multiple pairs
- Check whether highly separated pairs are evenly distributed across individuals (*why?)*
- Evaluate read depth distribution
- Search for split reads spanning breakpoint
- Combine with haplotype based hidden Markov model analysis

Handsaker et al (2011)

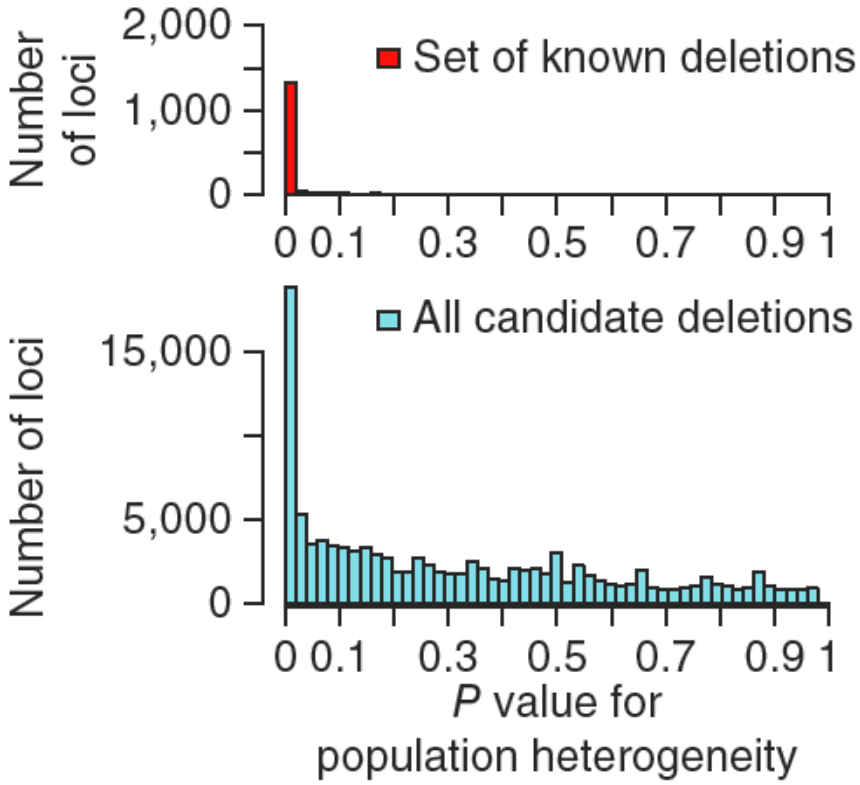
SEARCH FOR ABNORMAL READ PAIRS

- Search for read pairs where separation >10x the individual specific standard deviation
- Even if we require multiple supporting events, the number of potential copy number changes is ~10x larger than expected
- This is because of experimental limitation in preparing read pair libraries and of shortcomings in read mapping
- A major challenge is to reduce list of candidates

Handsaker et al (2011)

"HETEROGENEITY"

- Is rate at which widely separated read pairs occur constant among individuals?
- Calculated expected number of widely separated pairs using sequencing depth, average pair separation



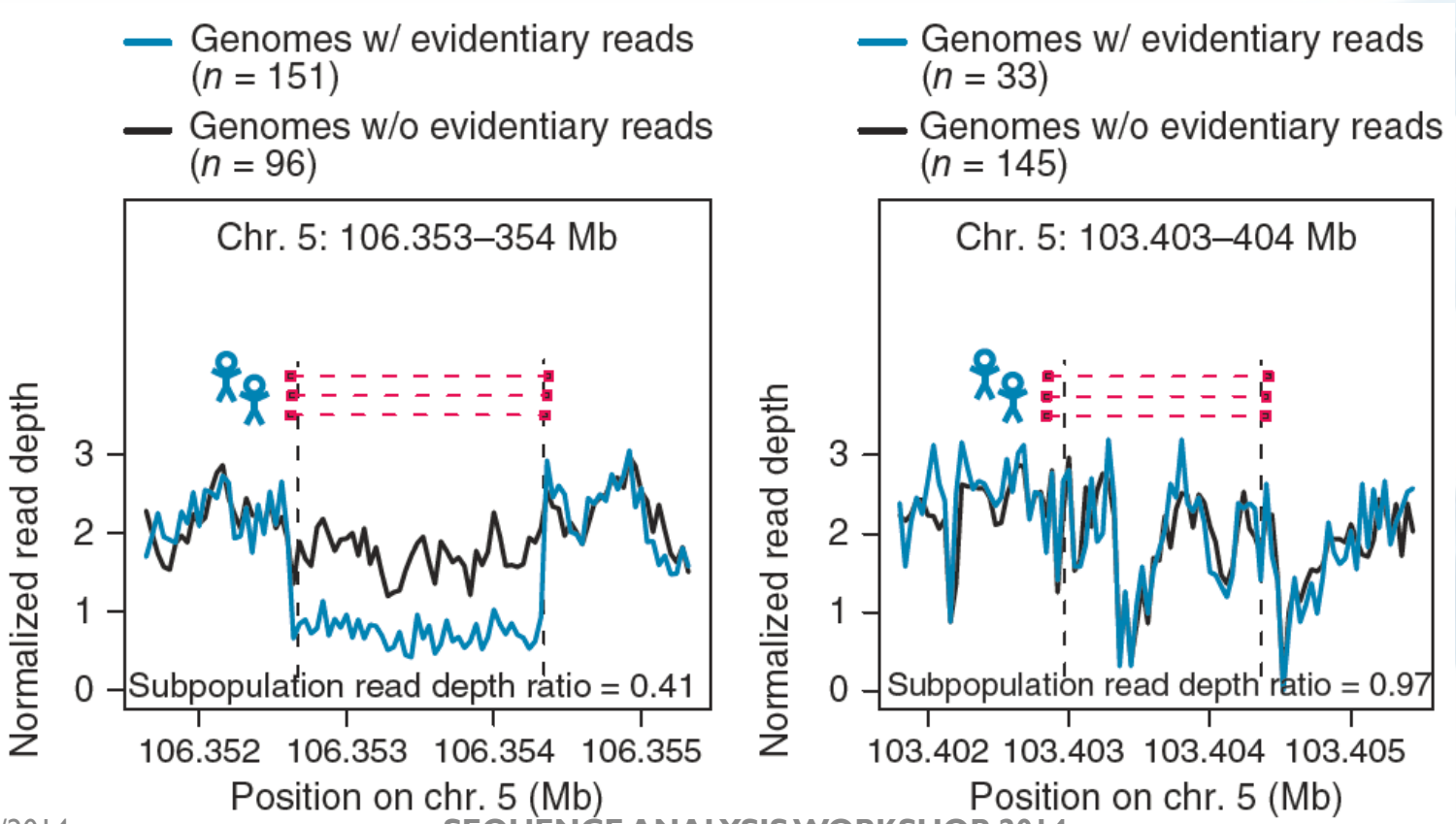
Handsaker et al (2011)

Expected Number of Widely Separated Read Pairs

- The approach of Handsaker et al. requires that we calculate, for each individual, the expected number of widely separated read pairs
- To do this, Handsaker et al (2011) calculate the distance between every mapped pair of reads
- They then assume that the number of read pairs separated by >*x* bp is proportional to the number of reads (across the genome) for which this distance exceeds *x*

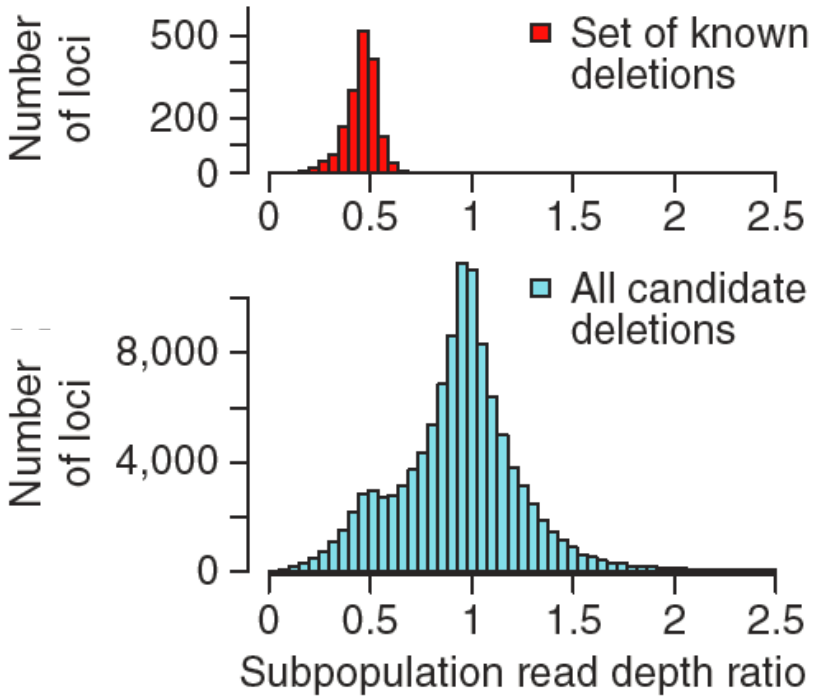
"Allelic Substitution"

- If we see evidence for deletion, based on read pair separation …
- Expect to see reduced evidence for reference based on read depth



"Allelic Substitution"

- If we see evidence for deletion, based on read pair separation …
- Expect to see reduced evidence for reference based on read depth

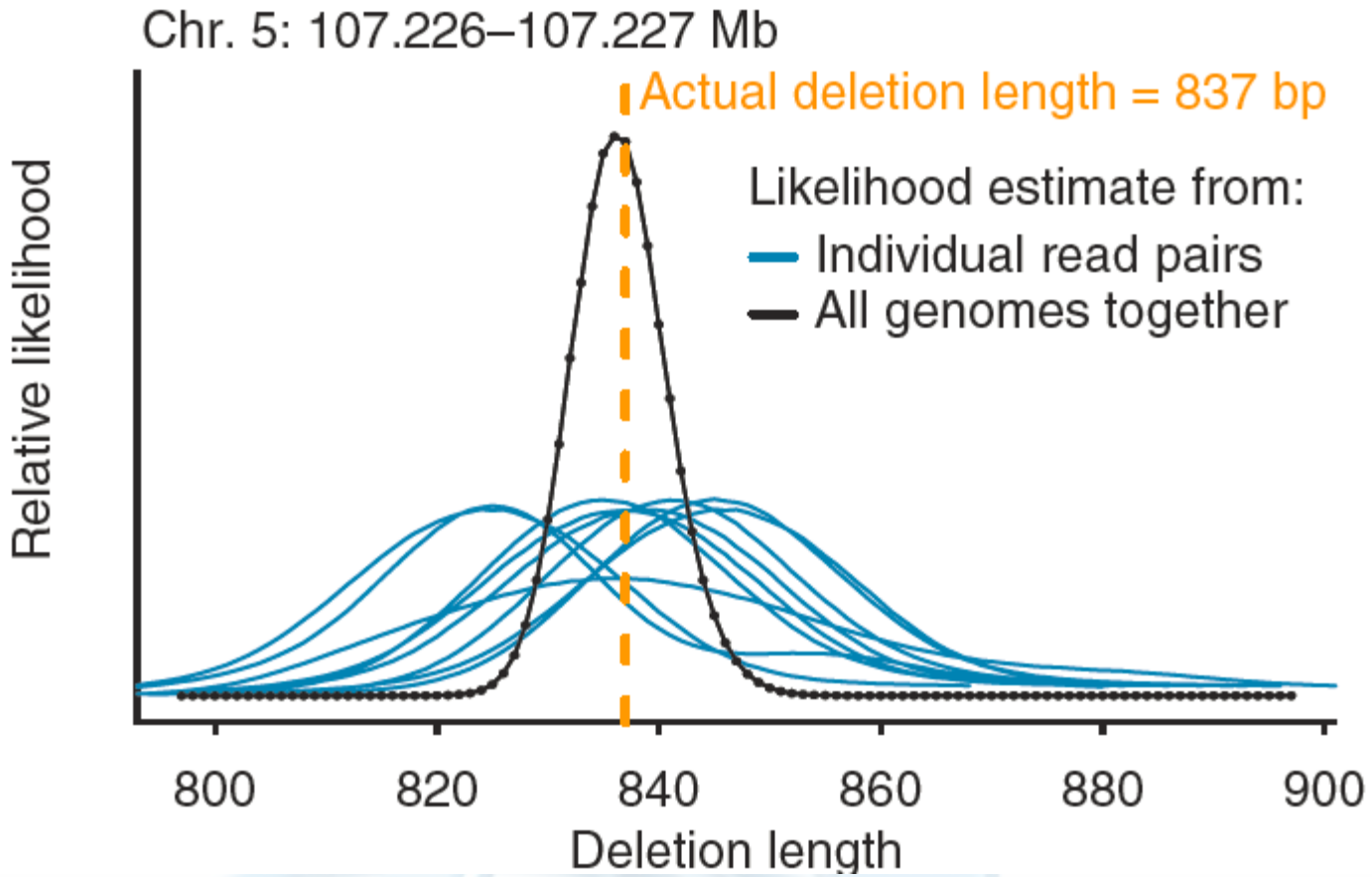


Handsaker et al (2011)

Sizing the Deletion

- If we know the distribution of read pair distances for one individual…
- Observing an abnormal read pair suggests a specific deletion size, but with low confidence
- Observing many abnormal read pairs gradually suggests more specific deletion sizes and locations

Combining Information Across Individuals is Key



Conclusions

- Combining information across individuals improves the power of deletion analyses
- Combining different sources of information within each individual also provides increased resolution
- Avoiding experimental artifacts is a major challenge in analysis of copy number

Recommended Reading

- Handsaker, Korn, Nemesh and McCarroll (2011) Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nature Genetics* **43**:269 - 276