Copy Number Variation

Biostatistics 666

Outline

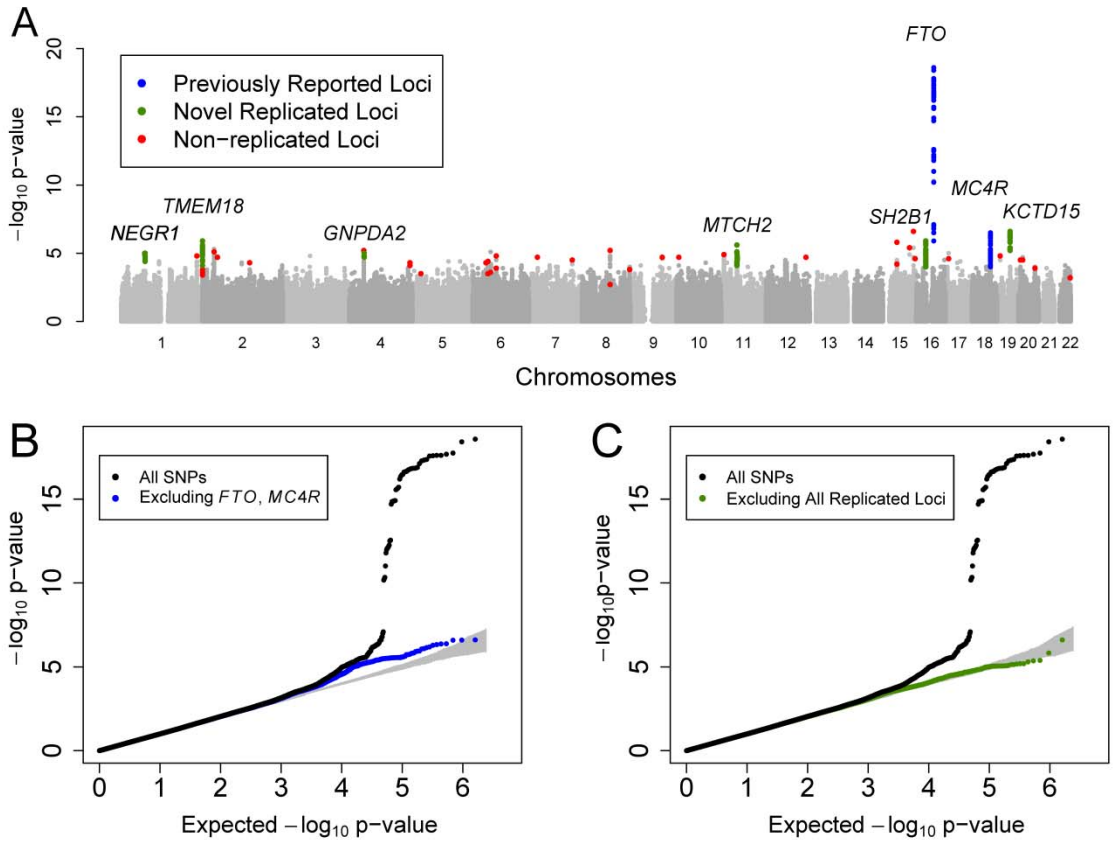
- Evidence that Copy Number Variation Contributes to Disease
- Signal of Copy Number Variation in Sequence Data
- A Method for Identifying Deletions Using Next-Generation Sequence Data

Evidence that Copy Number Variants Contribute to Disease

Some Background

- Copy number variation spans about 10% of the genome
- Copy number variation affects 5-10 Mb of sequence in an average individual
- Copy number variation is a major driver of disease in cancer
- A large fraction of copy number variants occur in and around regions of duplicated sequence
  - Due to “Non-Allelic Homologous Recombination” (NAHR)

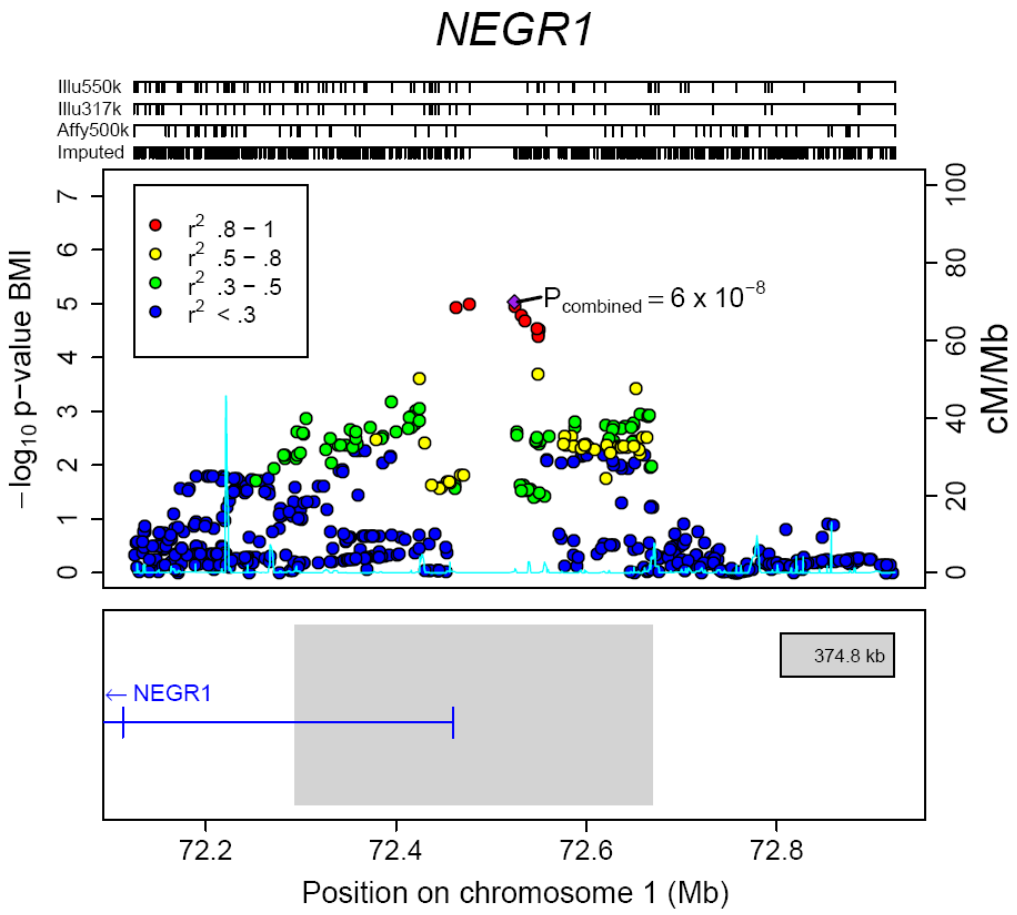
Evidence that Copy Number Variants Important Examples from Genetics of Obesity



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

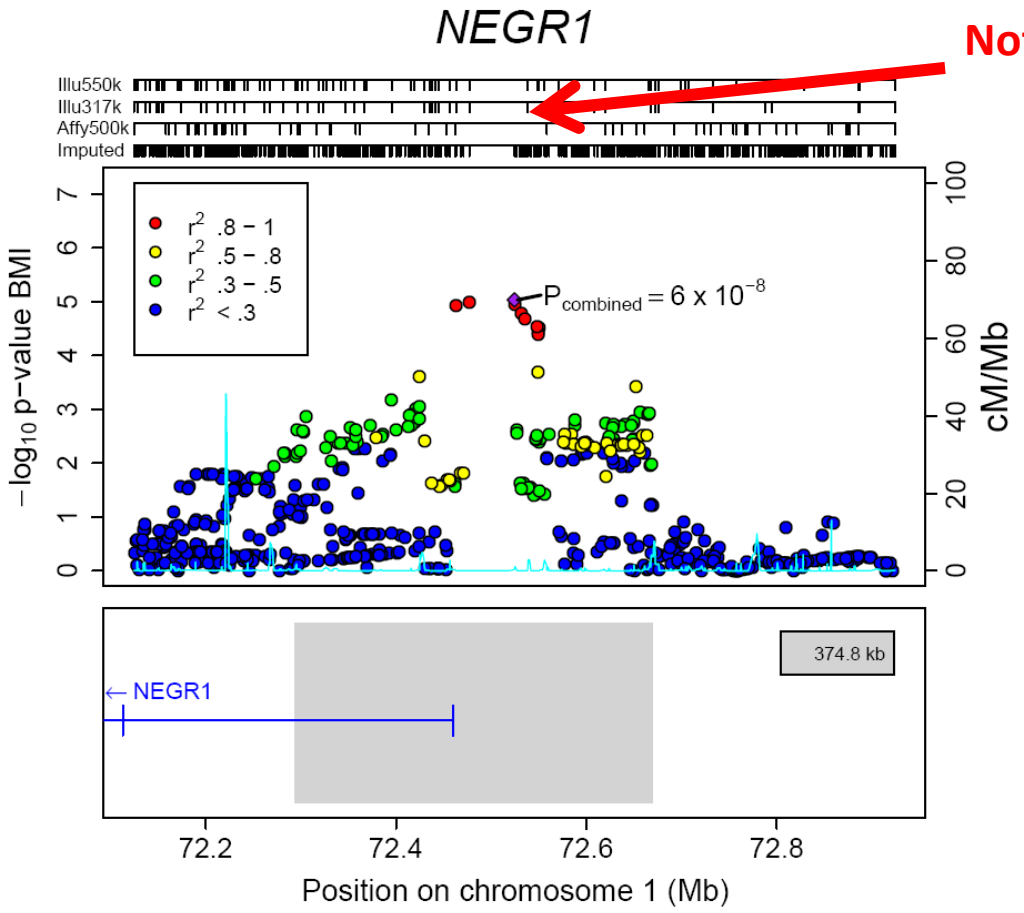
Willer et al, *Nature Genetics,* 2009

Some BMI Associated SNP Haplotypes …



Willer et al, *Nature Genetics,* 2009

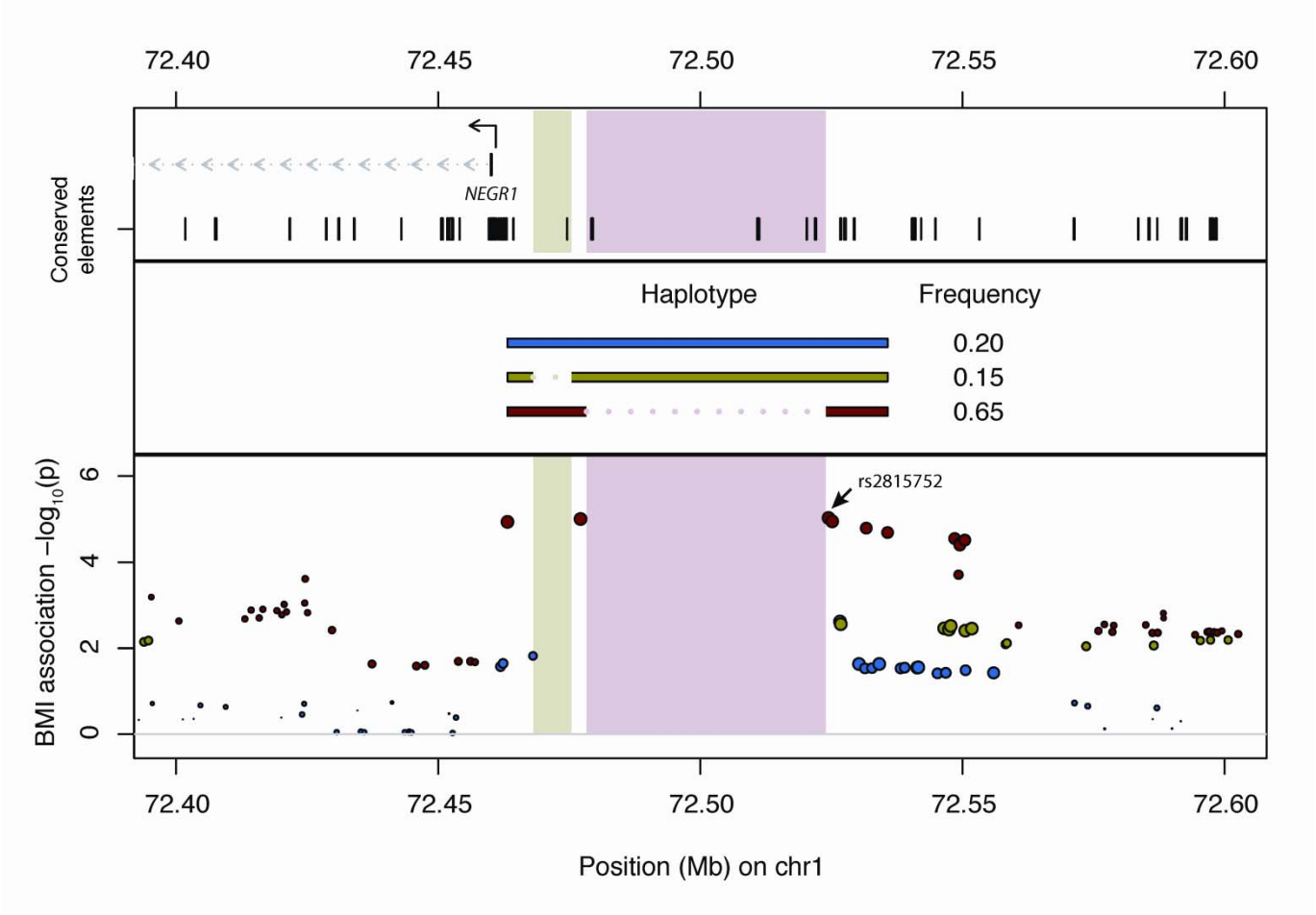
… Turn Out To Be Unusual …



Note hole in marker panels….

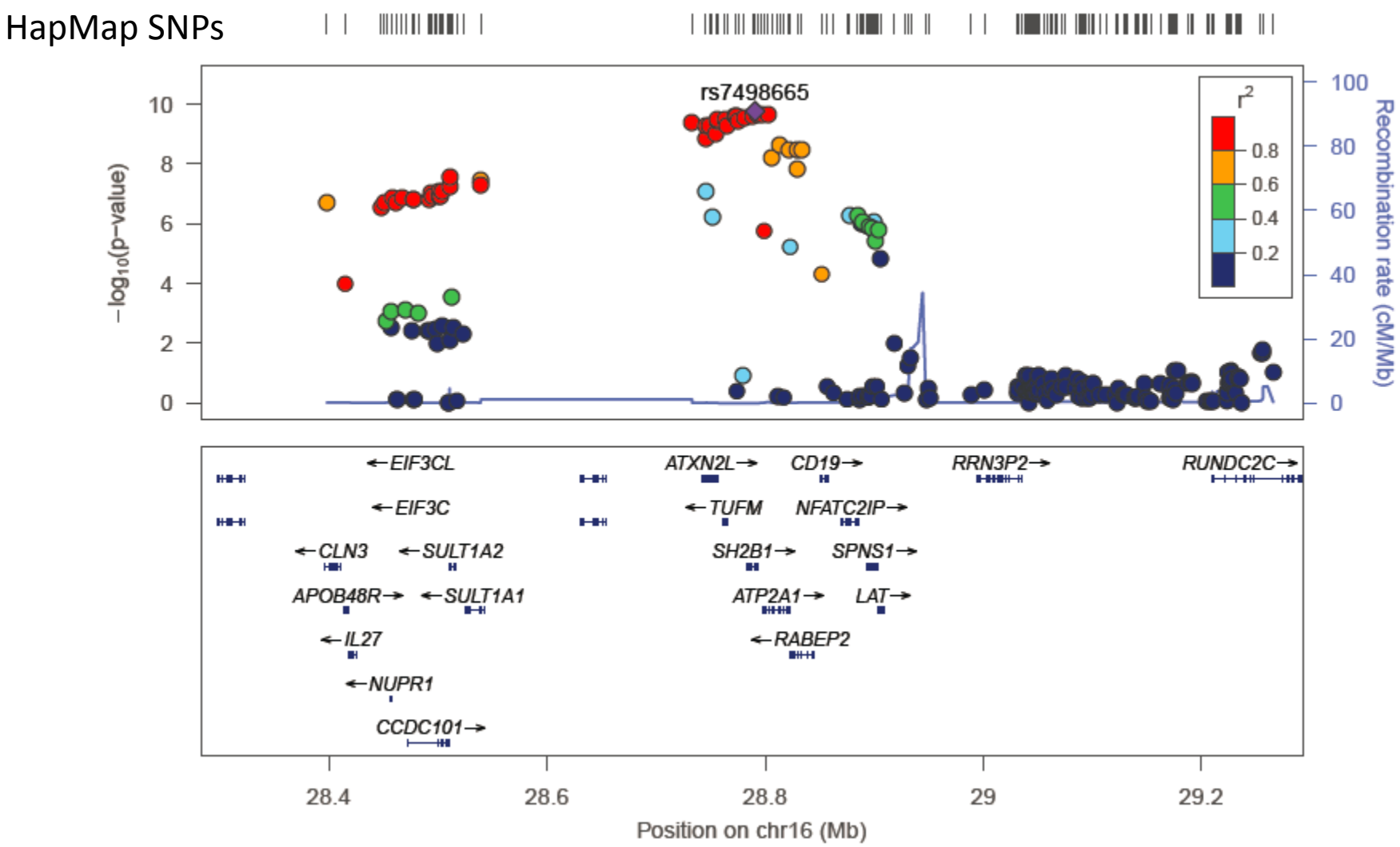
Willer et al, *Nature Genetics,* 2009

And Carry Unusual Deletion

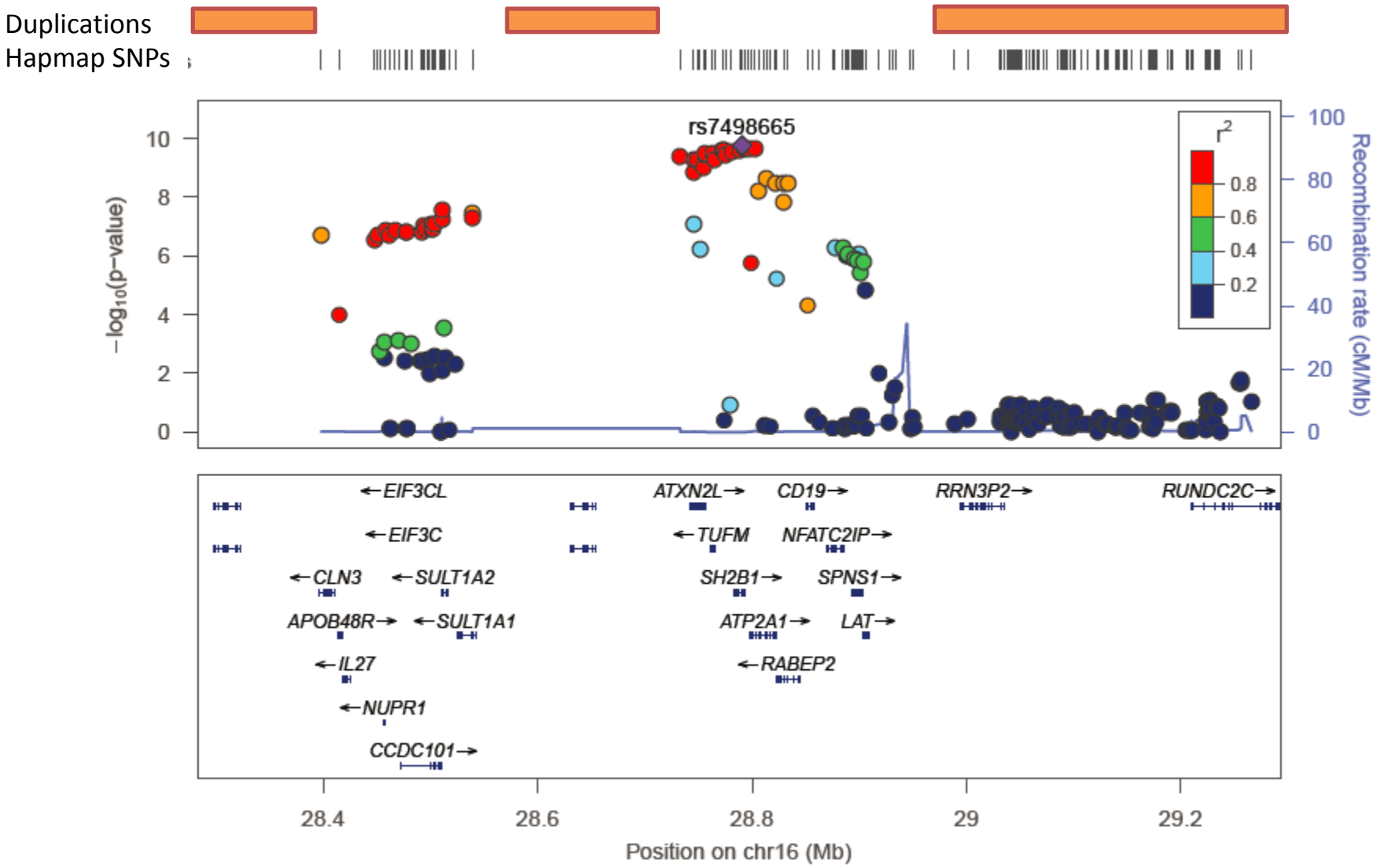


Willer et al, *Nature Genetics,* 2009

Other BMI Associated Haplotypes…



…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

Next Generation Sequencing And Copy Number Variation

Massive Throughput Sequencing

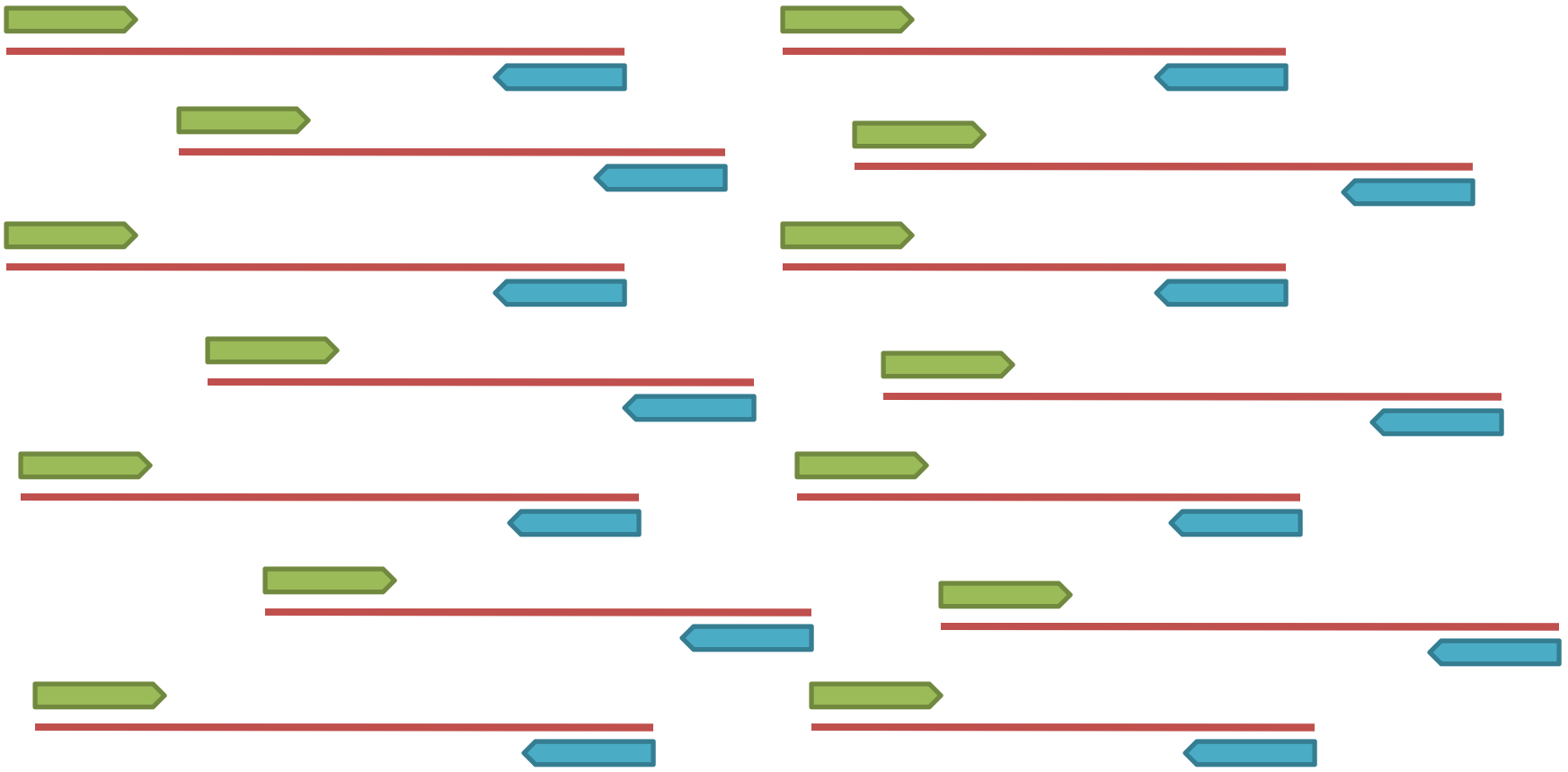
- Tools to generate sequence data evolving rapidly
- Commercial platforms produce gigabases of sequence rapidly and inexpensively
  - ABI SOLiD, Illumina Solexa, Roche 454, Complete Genomics, and others…
- Sequence data consist of thousands or millions of short sequence reads with moderate accuracy
  - 0.5 – 1.0% error rates per base may be typical

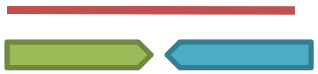
Shotgun Sequence Reads

- 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

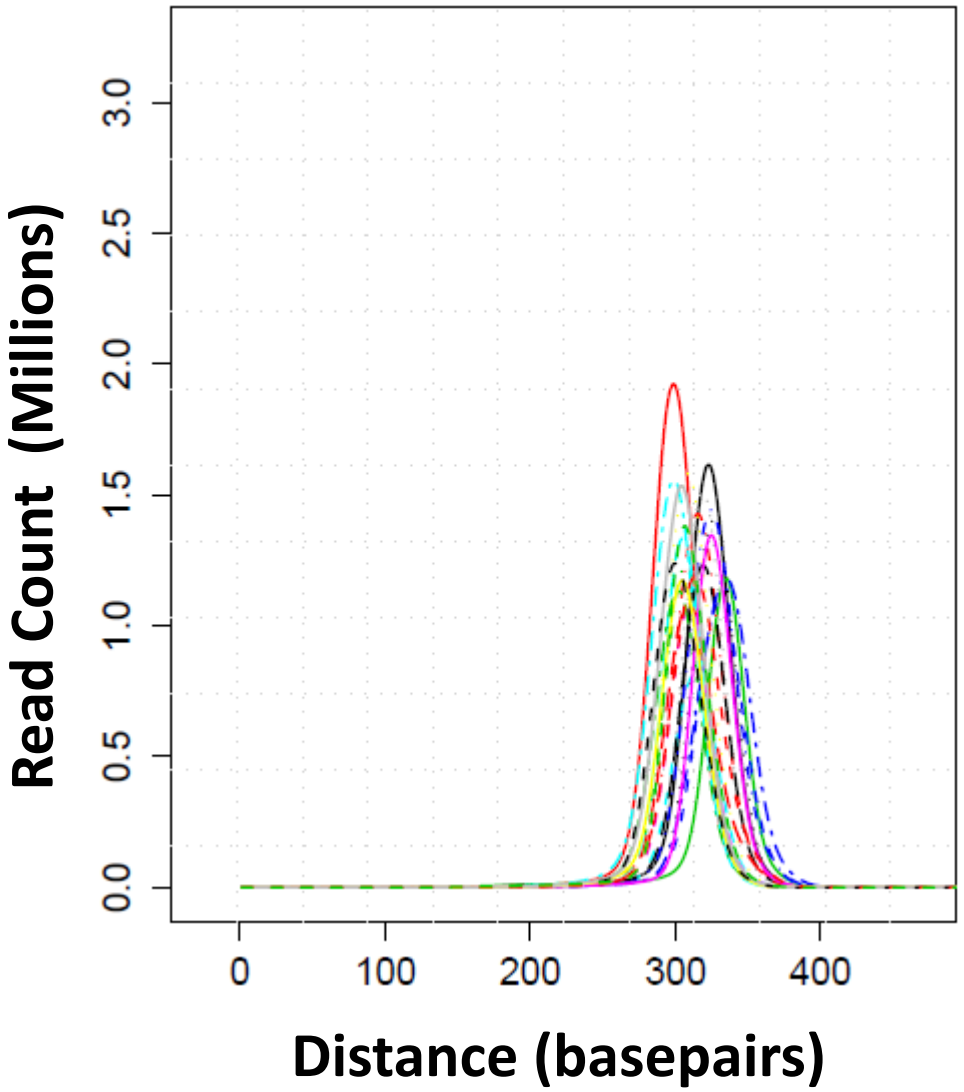
Paired End Sequencing

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

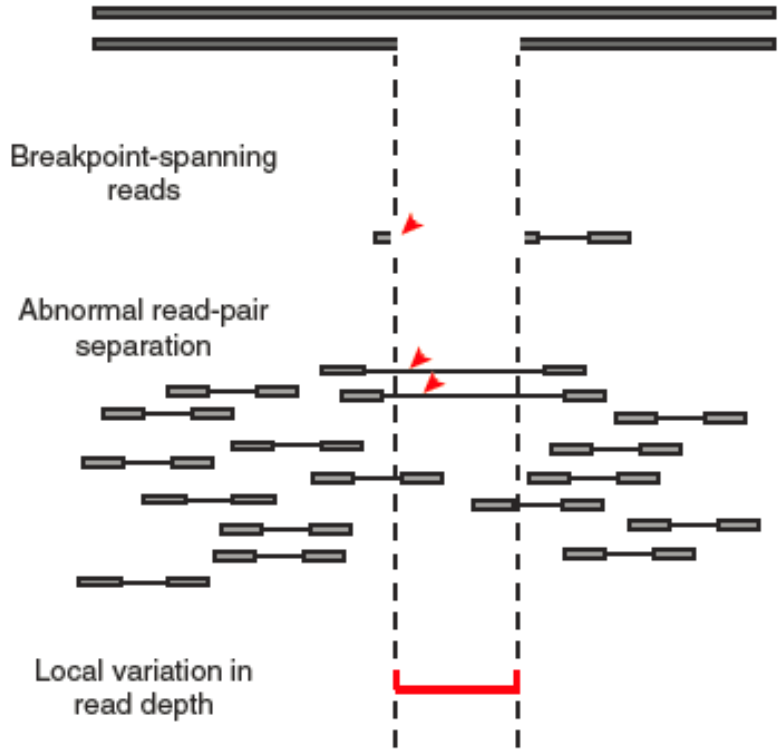


Figure from Handsaker et al (2011)

Detecting Copy Number Variation (Approach I)

- 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 (Approach I)

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

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*

Sara Rashkin

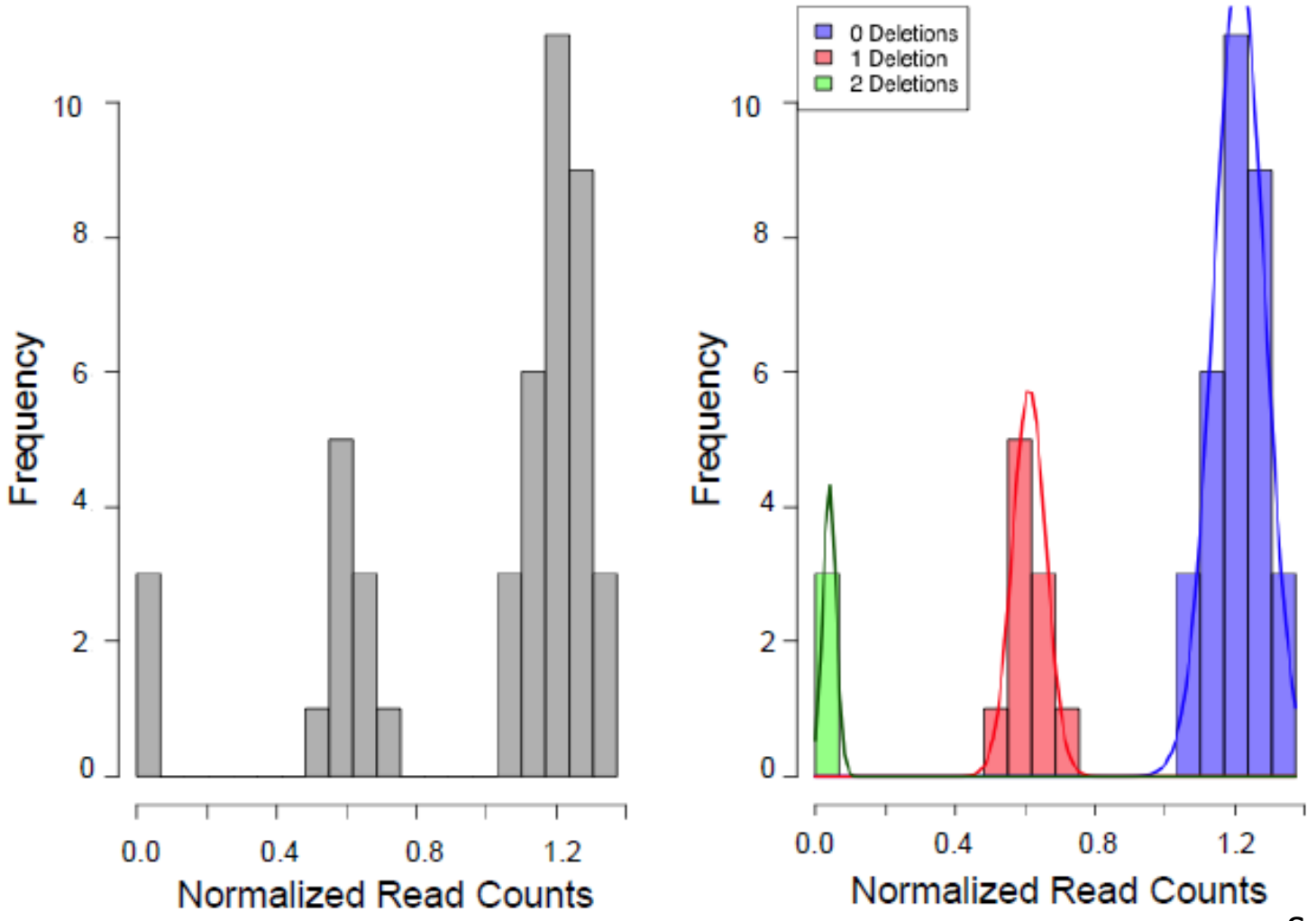
Detecting Copy Number Variation (Approach I)

- To estimate a deletion model, maximize

$$L(d_i) = \sum_j p_j (2\pi)^{\frac{1}{2}} \sigma_j^{-1} e^{-\frac{\frac{1}{2}(d_i - \mu_j)^2}{\sigma_j^2}}$$

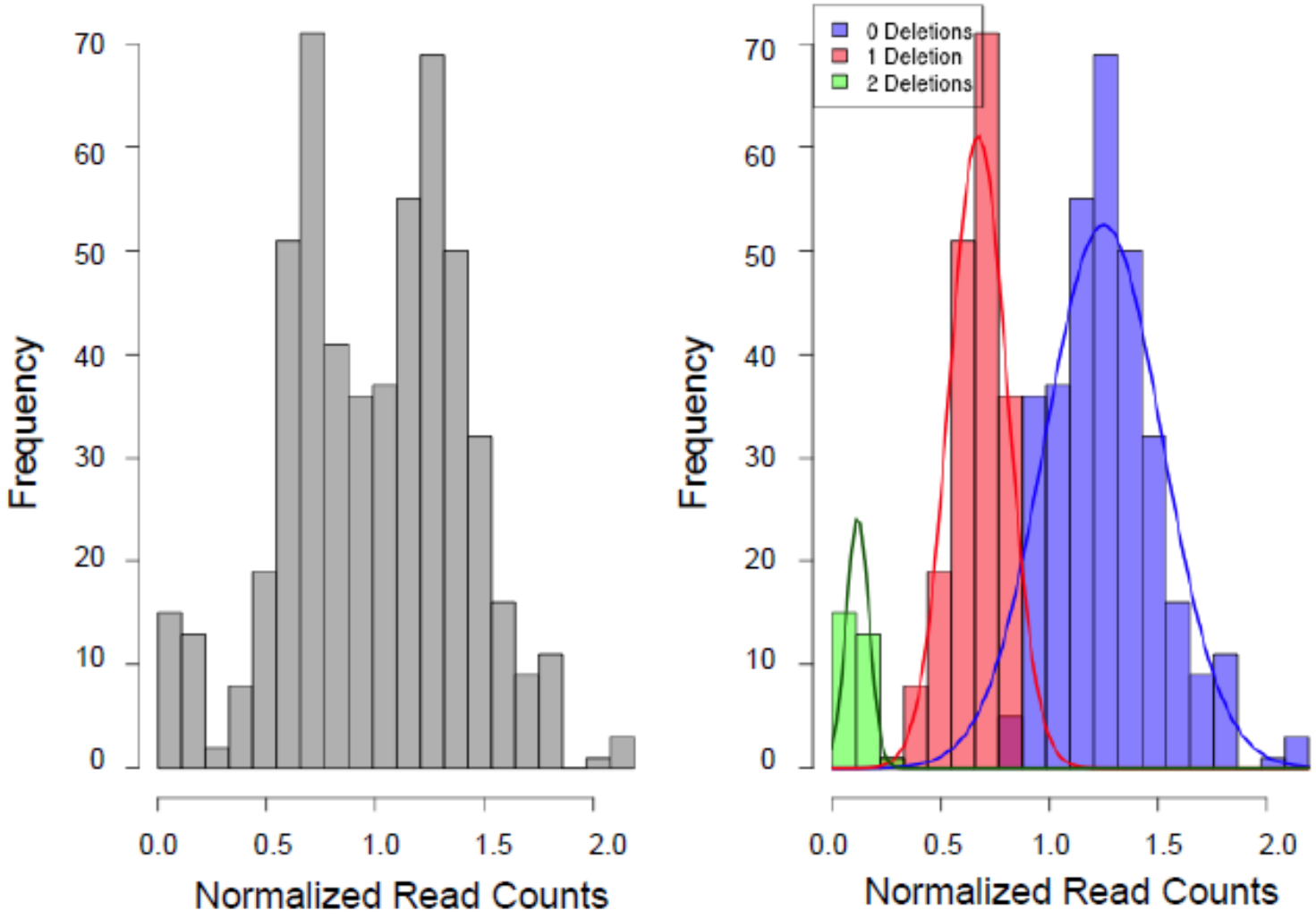
- To keep number of parameters modest, we use HWE for modeling p_j (one parameter for three frequencies) and can impose additional structure on means and variances

Well Separated Region



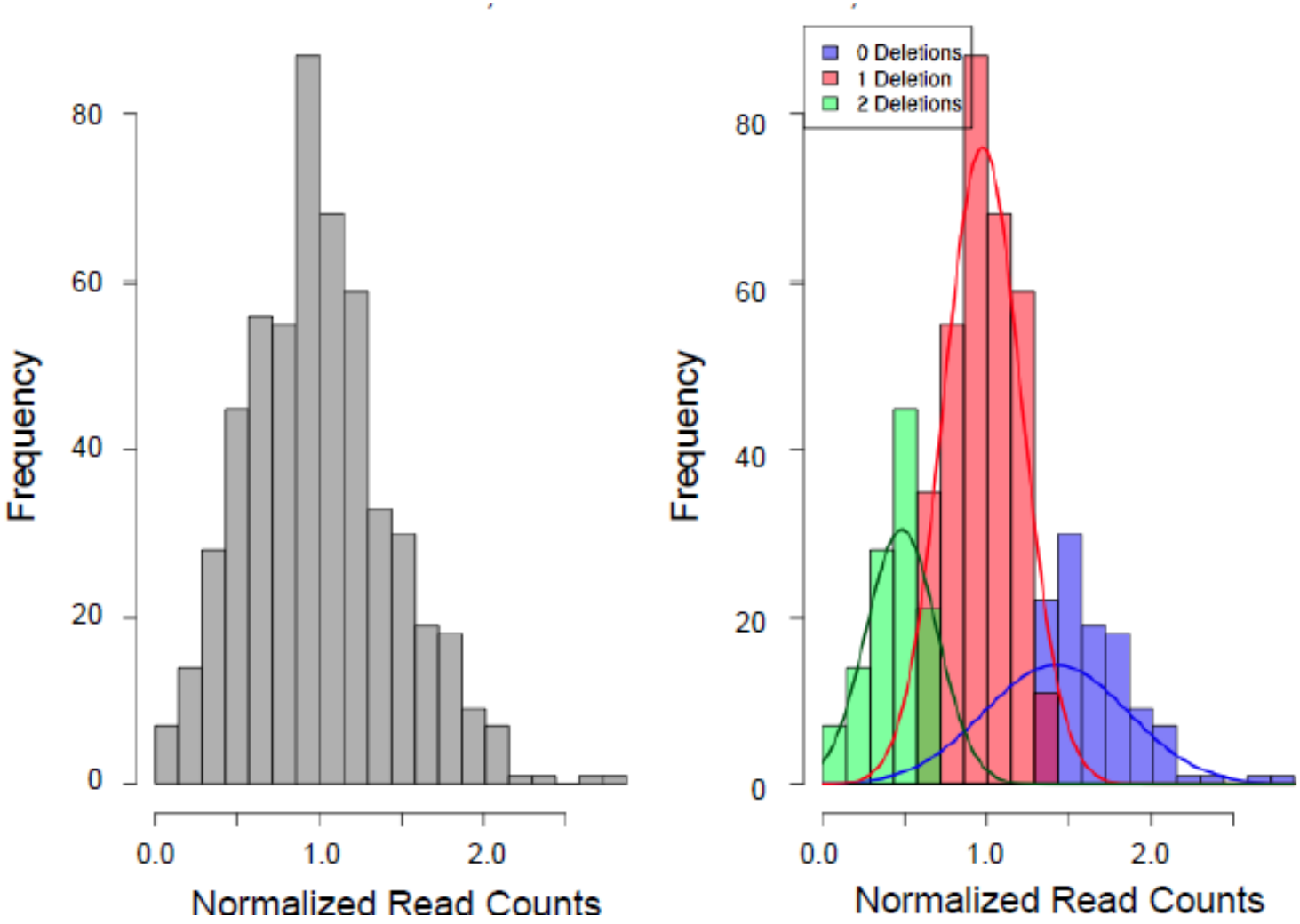
Sara Rashkin

Moderately Separated Region



Sara Rashkin

Hard to Call Region



Sara Rashkin

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)

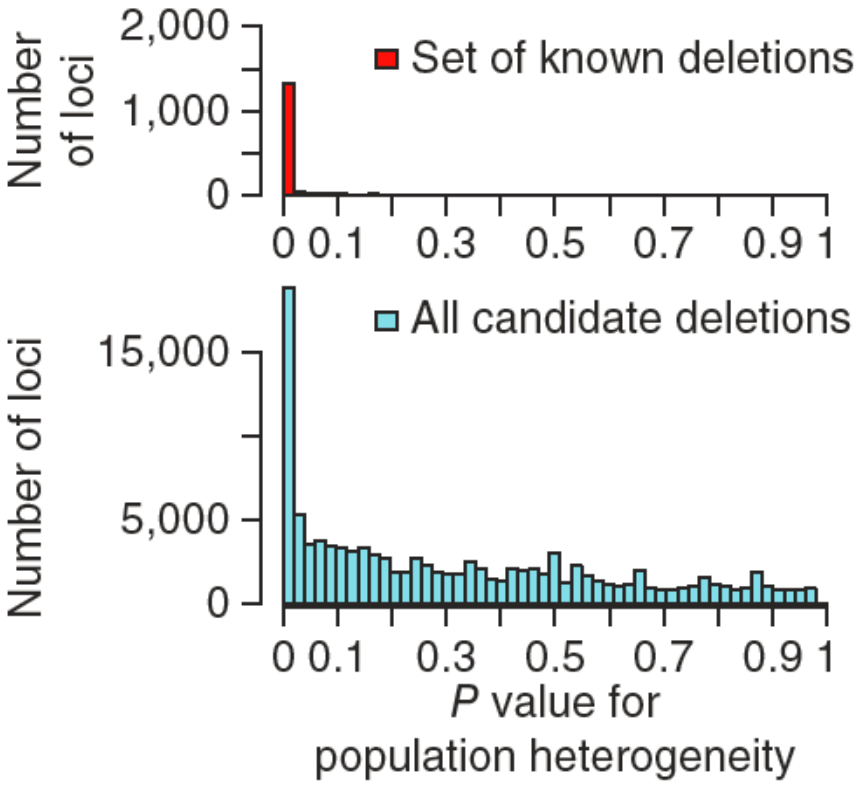
Search for Abnormal Read Pairs

- 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

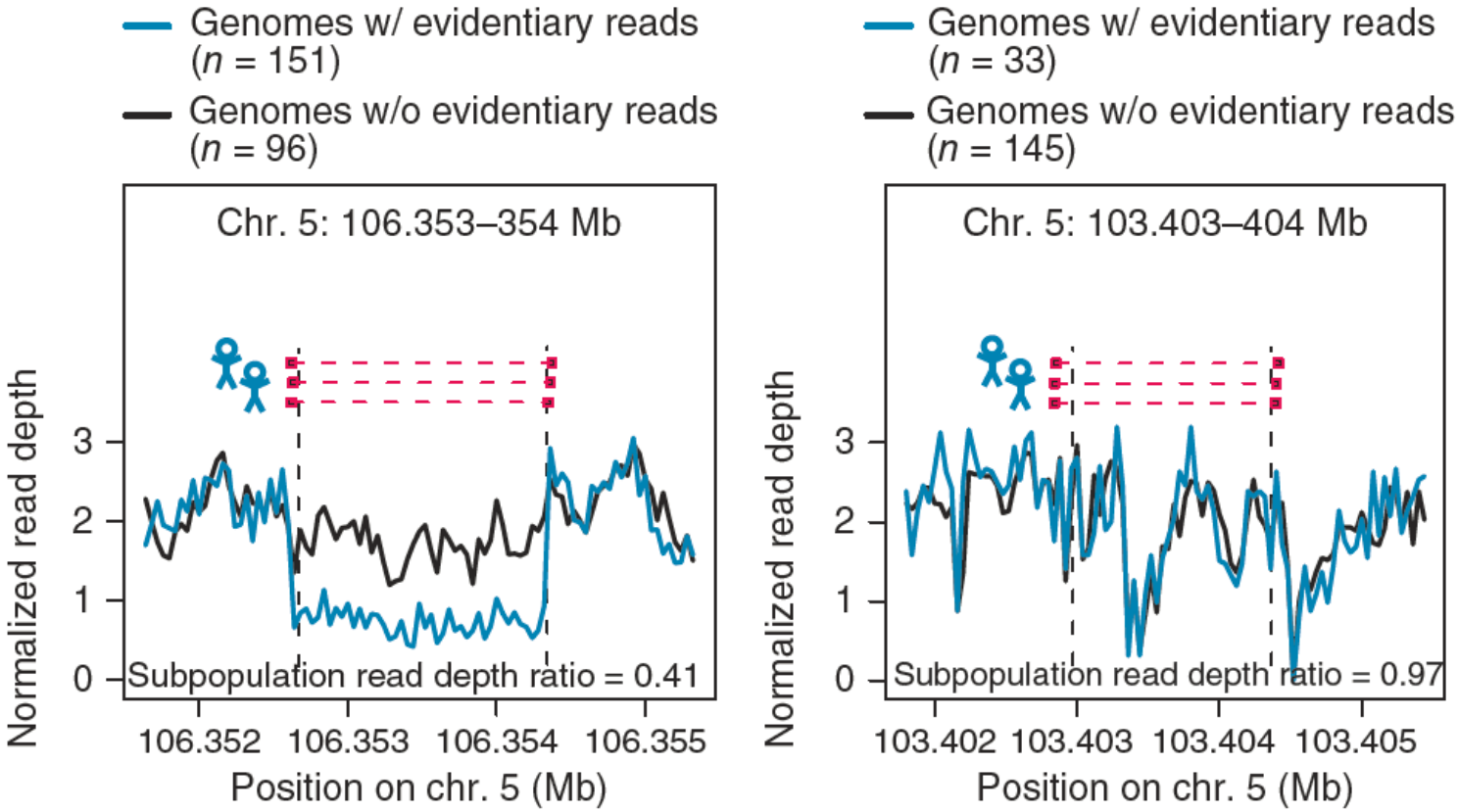
- The approach of Handsaker et al. requires that we calculate, for each individual, the expected number of widely separated read pairs

- 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*

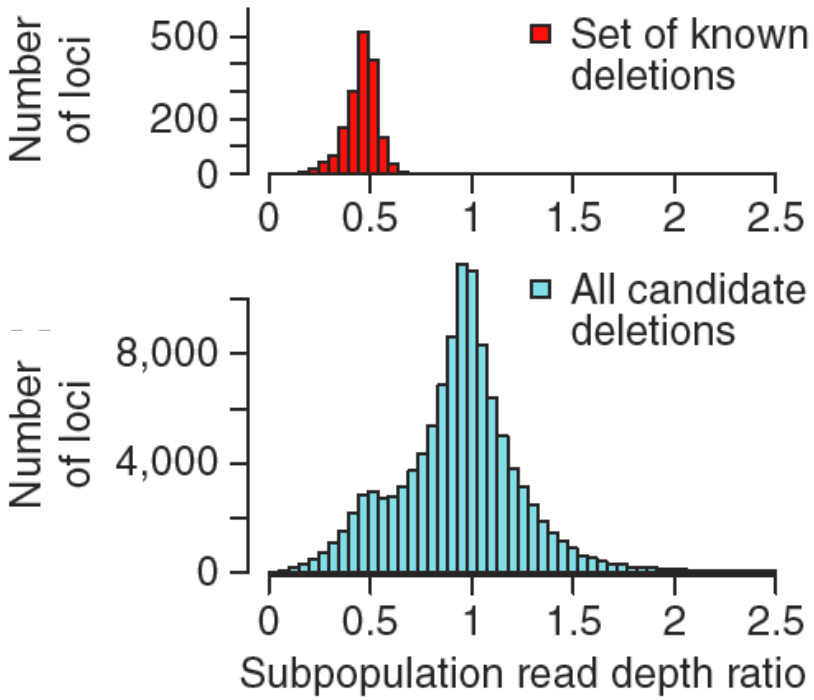
“Allelic Substitution”

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



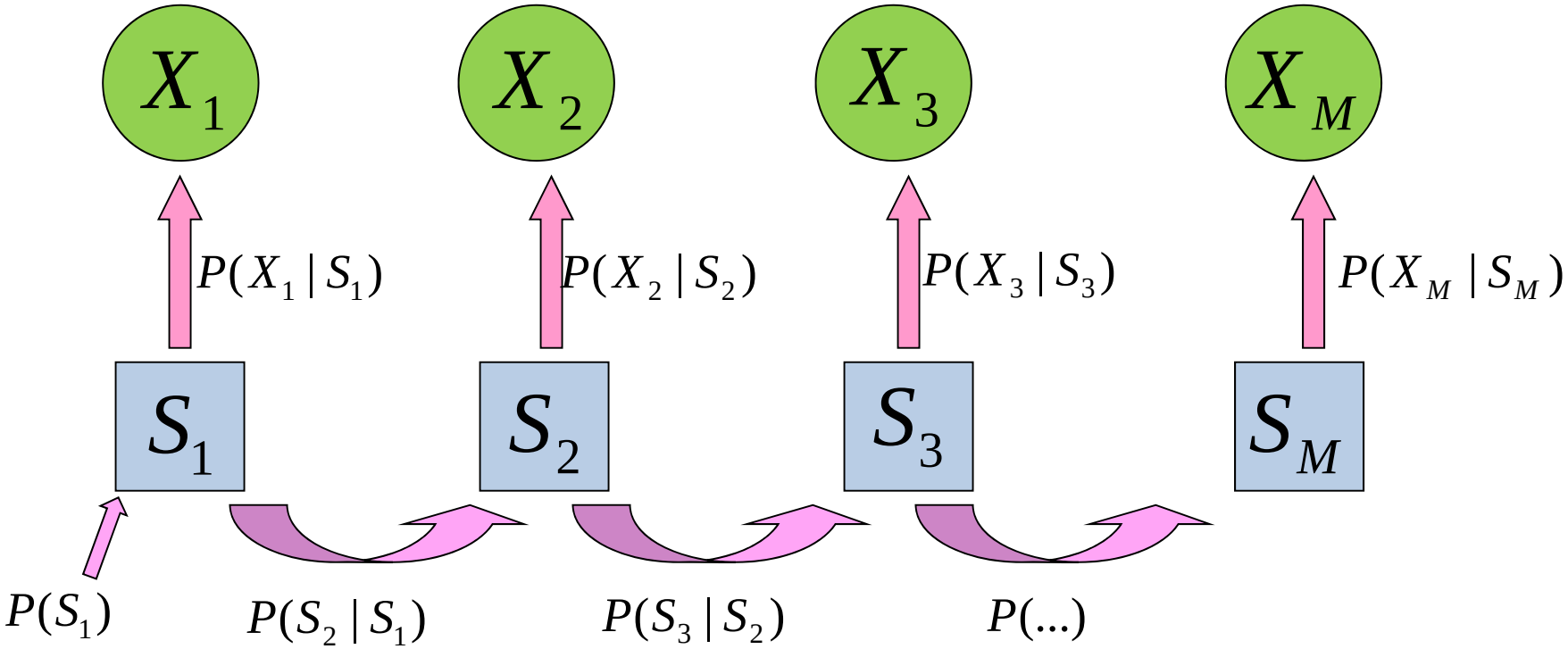
“Allelic Substitution”

- 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)

Integrate with Other Variant Types…

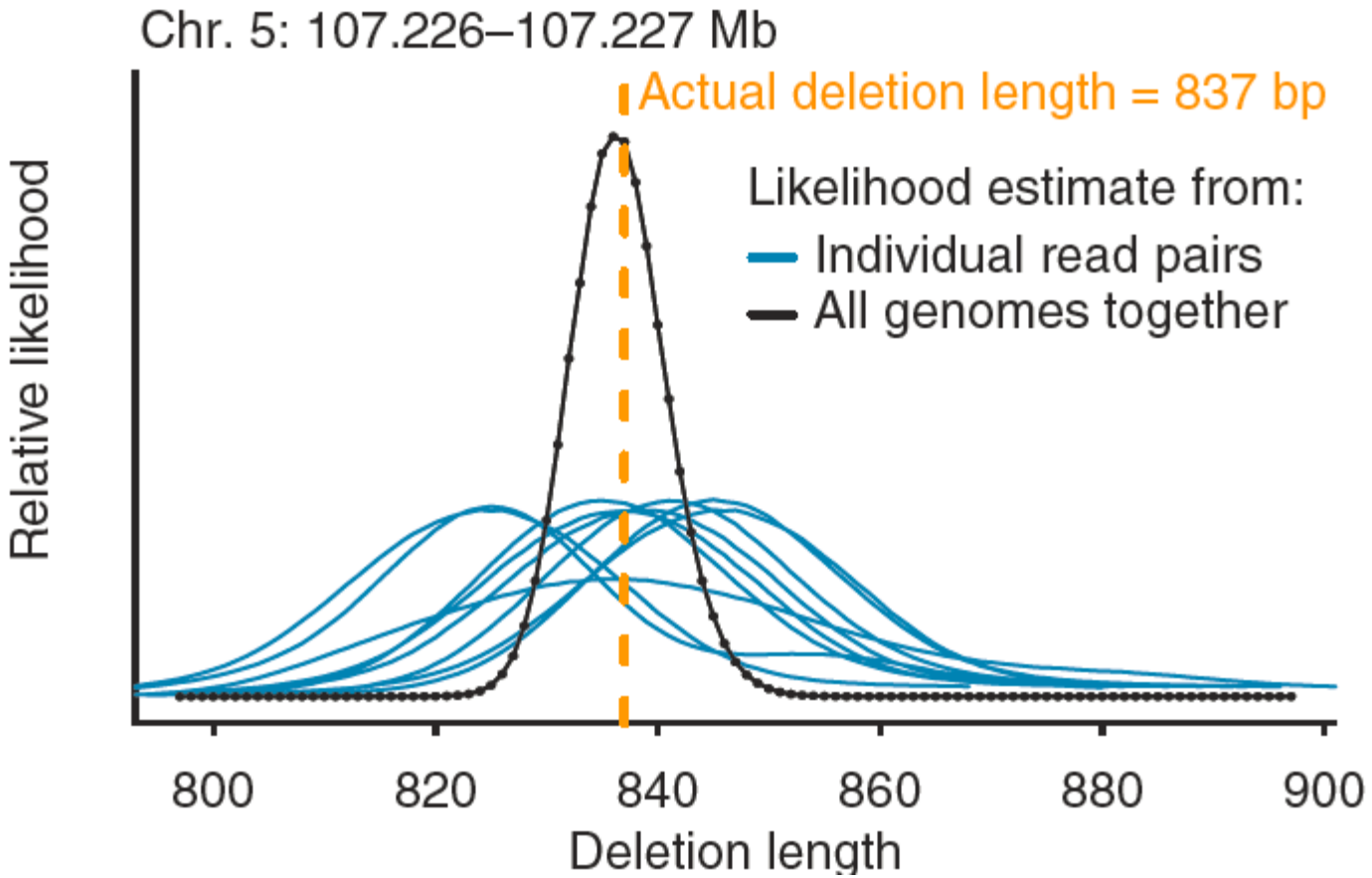


If we can calculated P(X|S) for a potential deletion, we can evaluate evidence for deletions jointly with other nearby variants

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