Autism Exome Sequencing

March 10th 2011

Benjamin Neale

Boulder Workshop

MGH 1811 Massachusetts General Hospital

Harvard Medical School

BROAD INSTITUTE

Acknowledgements

Exome Analysis

Mark Daly
Elaine Lim
Andrew Kirby
Jared Maguire
Shaun Purcell
Manny Rivas
Ben Voight
Jason Flannick
Menachem Fromer
Heng Li
Todd Green

GATK

Eric Banks
Mark DePristo
Kiran Garimella
Chris Hartl

ARRA Autism sequencing

Mark Daly
Christine Stevens
Stacey Gabriel
Corin Boyko
Broad Sequencing Platform
Jen Baldwin, Jane Wilkinson, Liz Bevilacqua, Tim Fennell

Joe Buxbaum, Bernie Devlin, Richard Gibbs, Jerry Schellenberg, Jim Sutcliffe

NIMH, NHGRI

Many other Broad Collaborators and Colleagues

Overview

- Study Design
- Quality Control
  - Samples
  - Variants
  - Assessment
- Preliminary analysis
  - Single Locus
  - Regional

Autism

- Three core phenotypic characterizations
  - Repetitive behaviors
  - Impaired social interaction
  - Impaired communication skills
- Prevalence between 0.1-1%
- Onset starts ~2 years old
- Number of chromosomal abnormalities associated with disease
  - Maternally inherited 15q11-13 duplication (Angelman/ Praeder-Willi region)
  - 16p11.2 deletion/duplication

Study Design

- Match autism cases to NIMH Controls on PCA distance using common variation
  - Target 1,000 cases and 1,000 controls
  - Sequencing split between Broad and Baylor
- Pair samples through Next Generation Sequencing process
- Additionally sequencing trios
  - 41/100 completed, several hundred more proposed for Y2

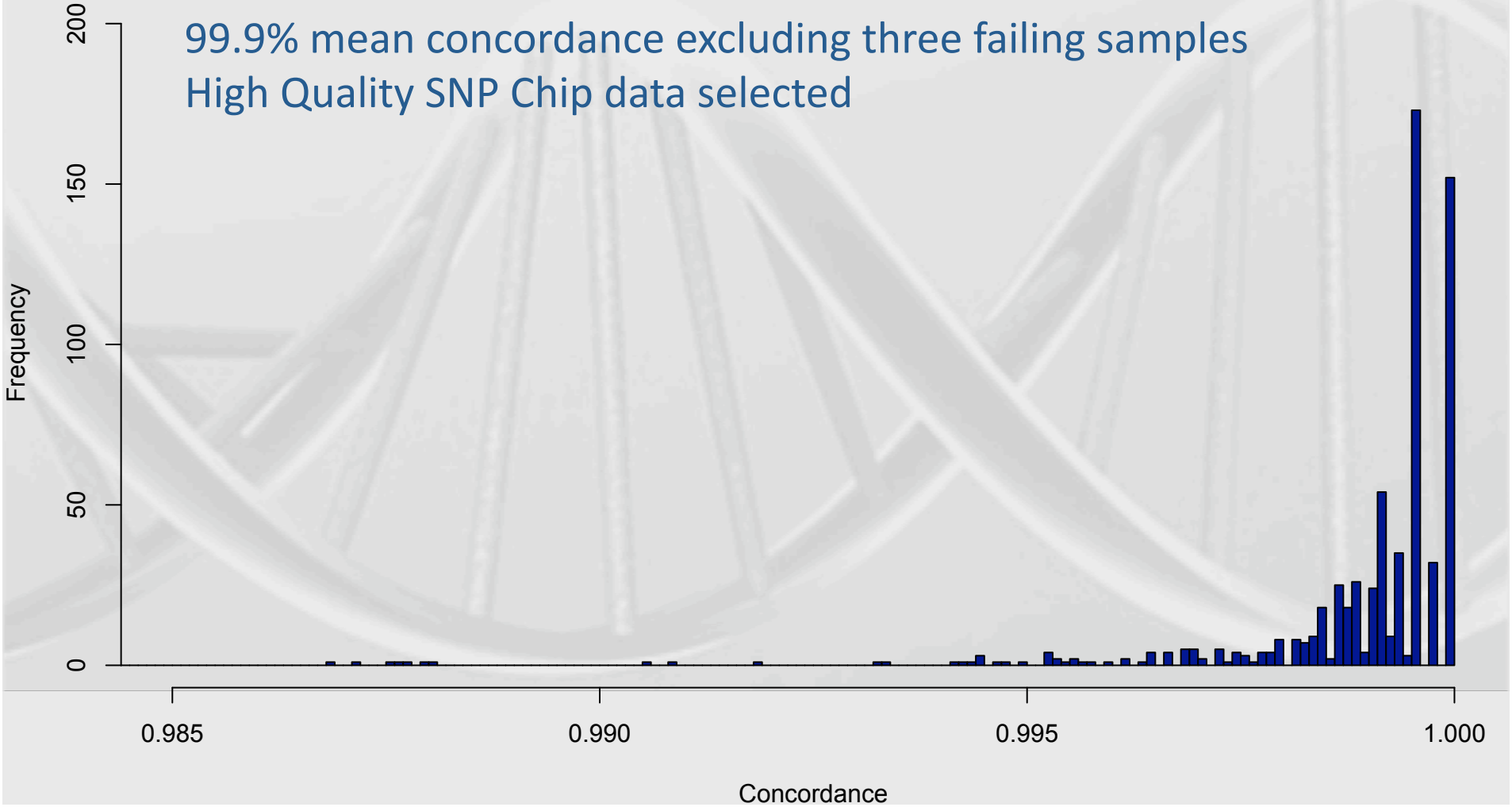
Data Generation

- Agilent Sure Select v2
- Coding regions for genes
- Splice sites (+/- 2bp)
- Known miRNA
- ~33 MB of target

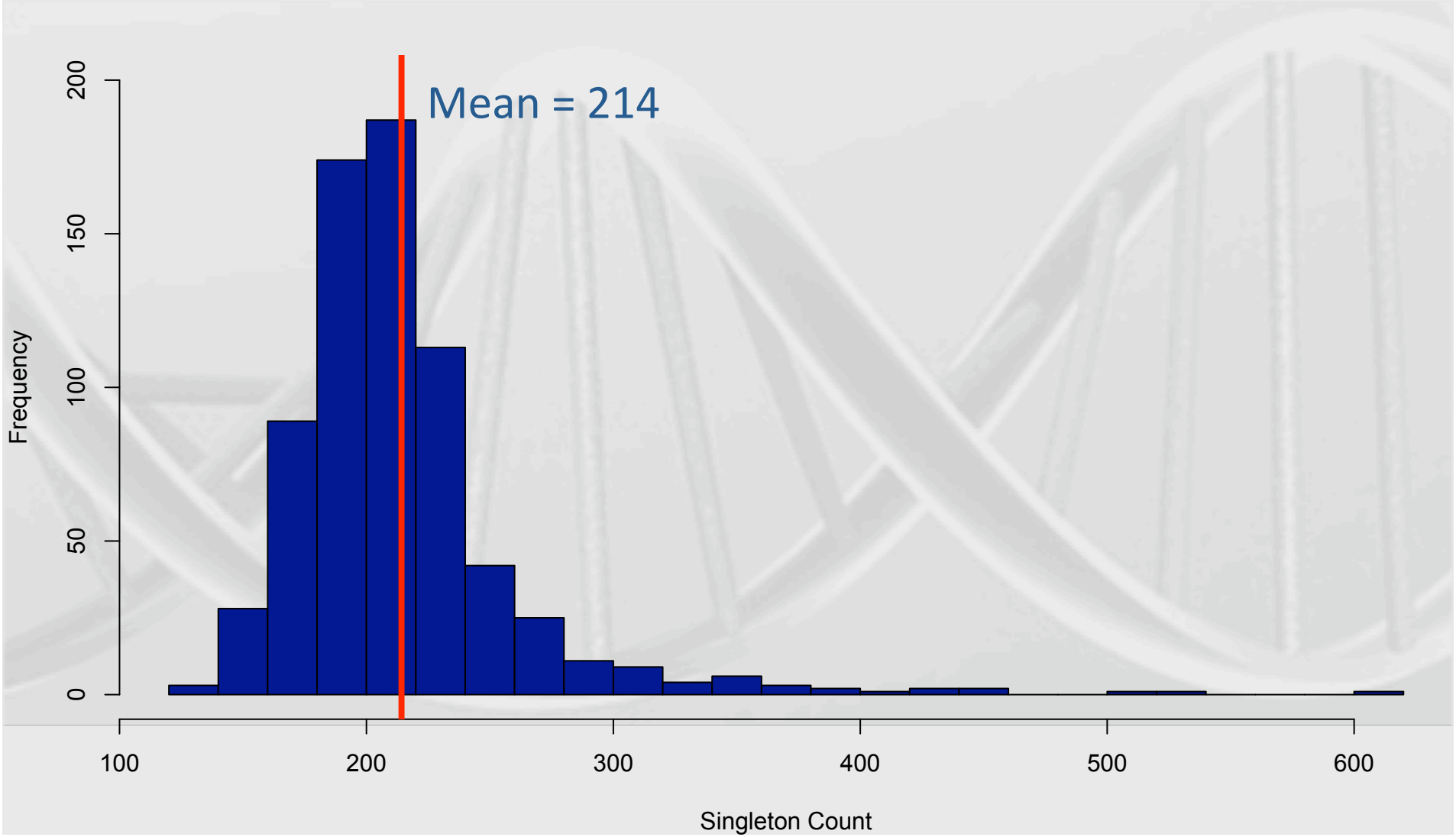
Individual QC

Concordance with SNP Chip

99.9% mean concordance excluding three failing samples
High Quality SNP Chip data selected

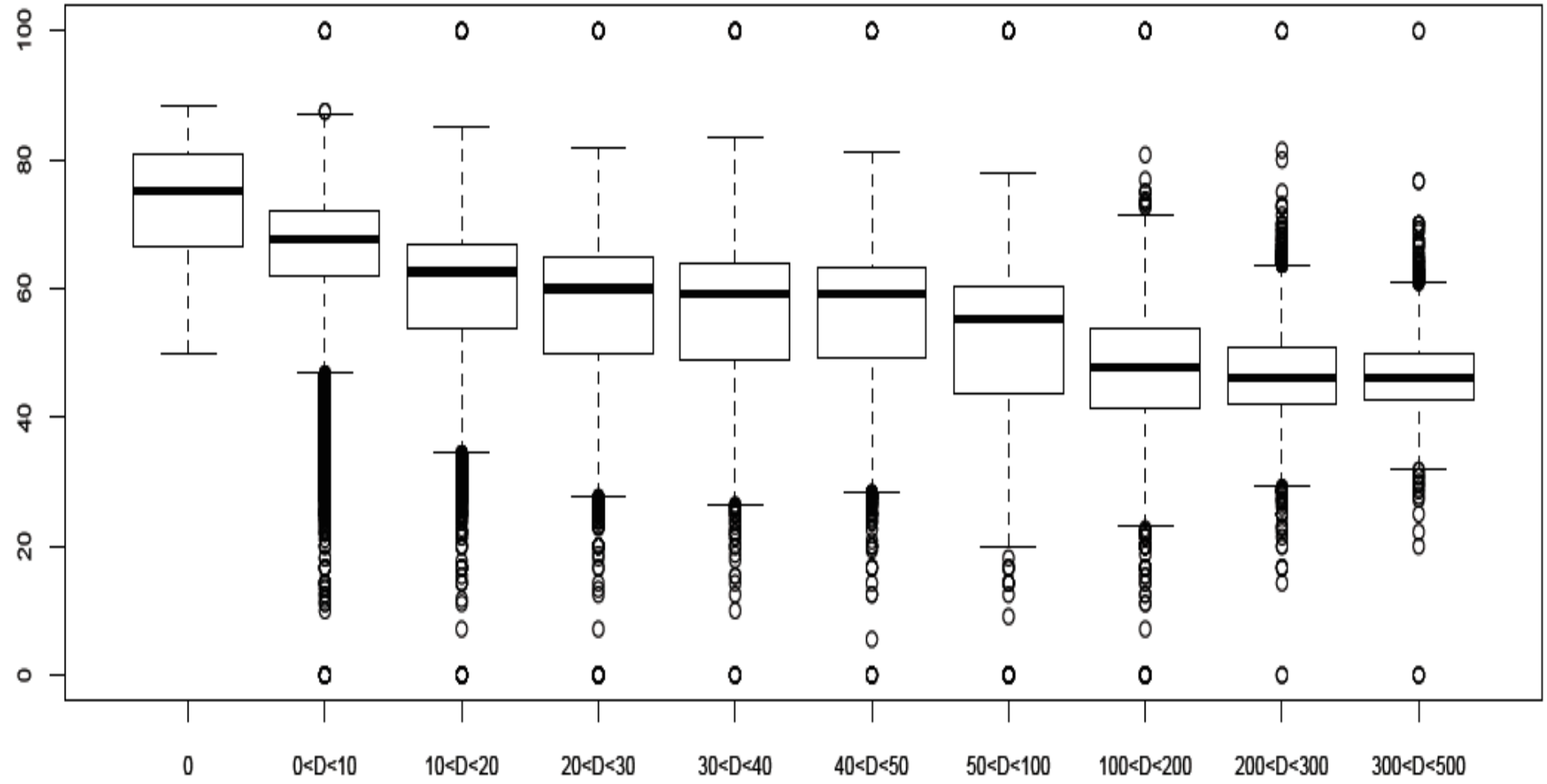


Singleton Counts

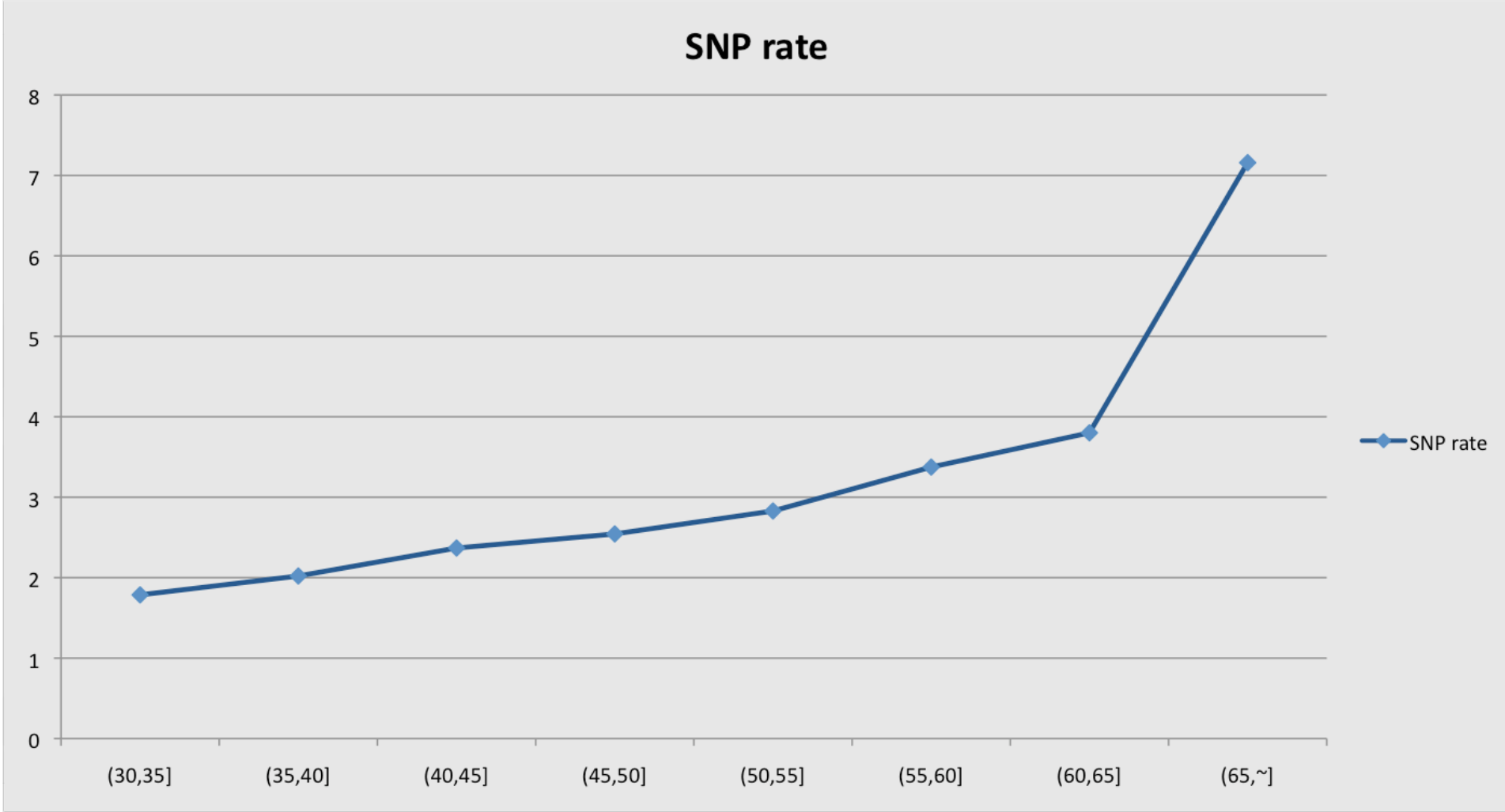


Target Considerations

Depth of coverage vs. GC



SNP rate as a function of GC Bin



GC content is strongly correlated with both lower coverage and higher polymorphism rate

(as a result we both miss more variants while simultaneously calling a higher rate of variants in high GC exons)

MGH 1811 Massachusetts General Hospital
Harvard Medical School
BROAD INSTITUTE

Call Comparison

Set	# Calls	Ti.Tv	%dbSNP
Intersection	227,657	3.24	27.0
Samtools Specific	26,860	2.37	27.2
Samtools Pass GATK Fail	13,505	1.57	17.9
GATK Specific	37,971	2.05	15.8

Assessing the call set

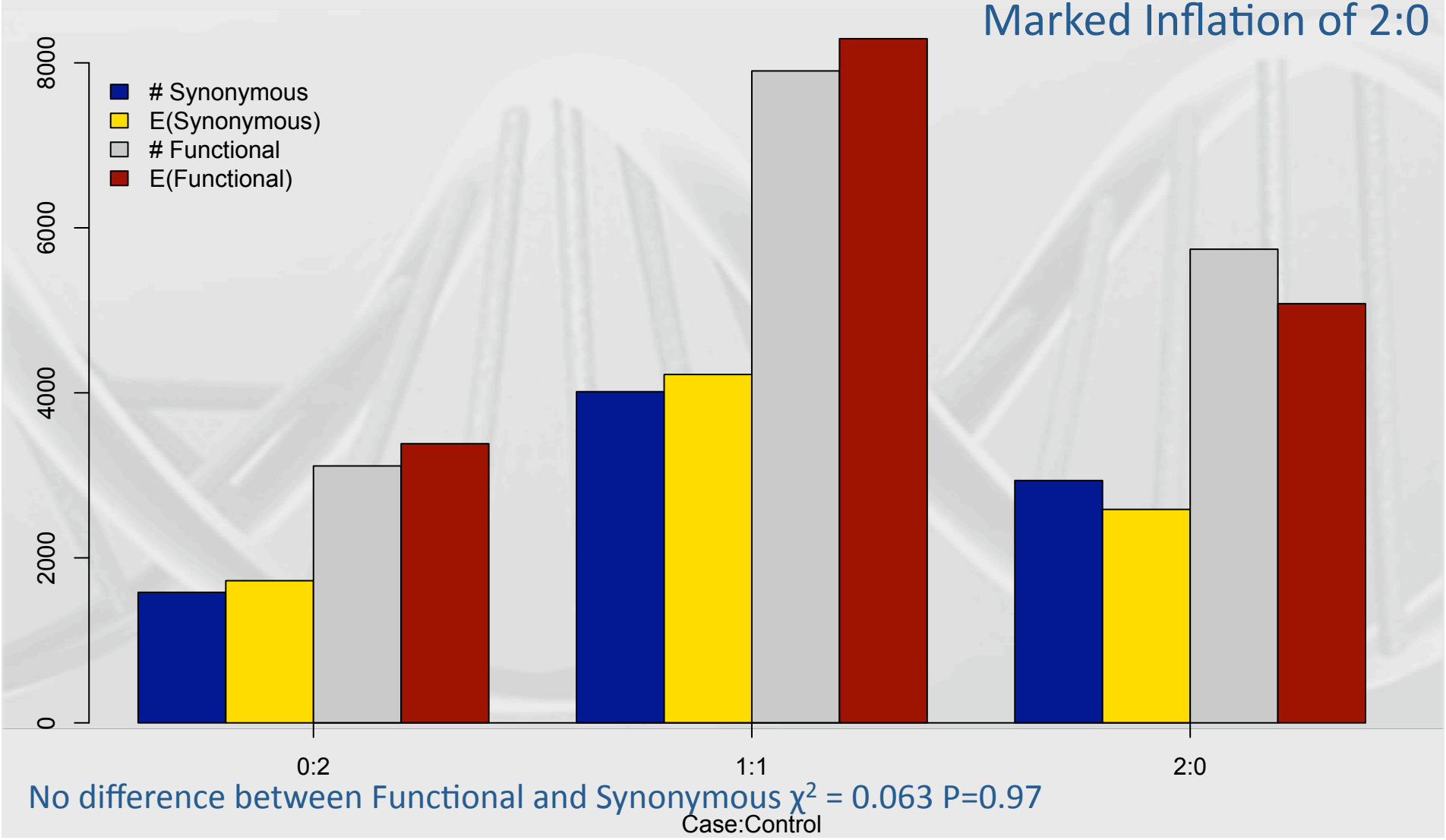
Inspect distributions of doubleton, tripleton etc.

Has the matching worked?

- Matched samples based on MDS distance from combined GWAS data
- Consider the set of doubletons (two copies in the dataset)
- Overall, we should see that there are comparable numbers of variants seen in 2 cases or 2 controls versus 1 case and 1 control; and we should see an excess of 1 case:1 control variants in matched pairs

Distribution of Doubletons [All]

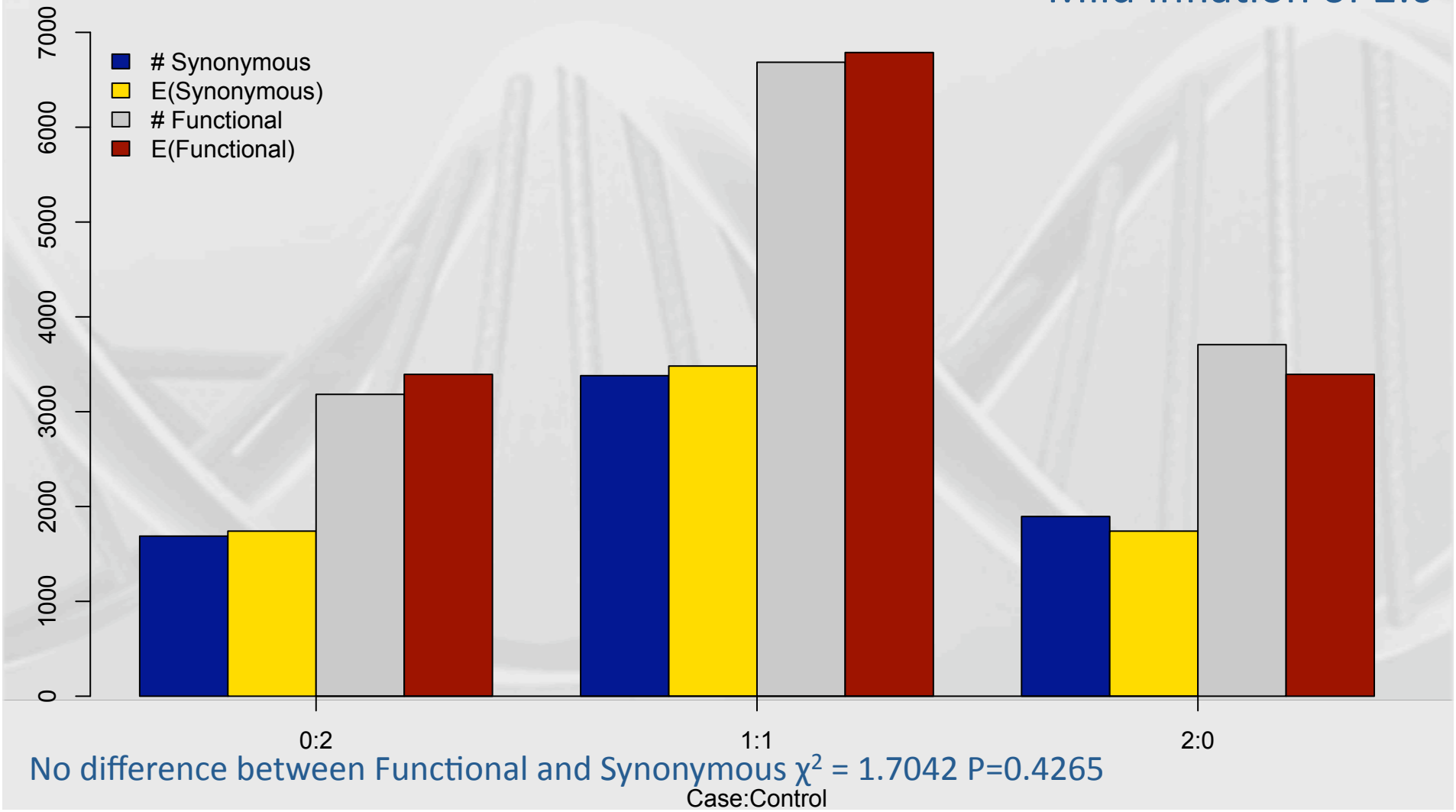
Marked Inflation of 2:0



No difference between Functional and Synonymous $\chi^2 = 0.063$ P=0.97

Distribution of Doubletons [Pairs]

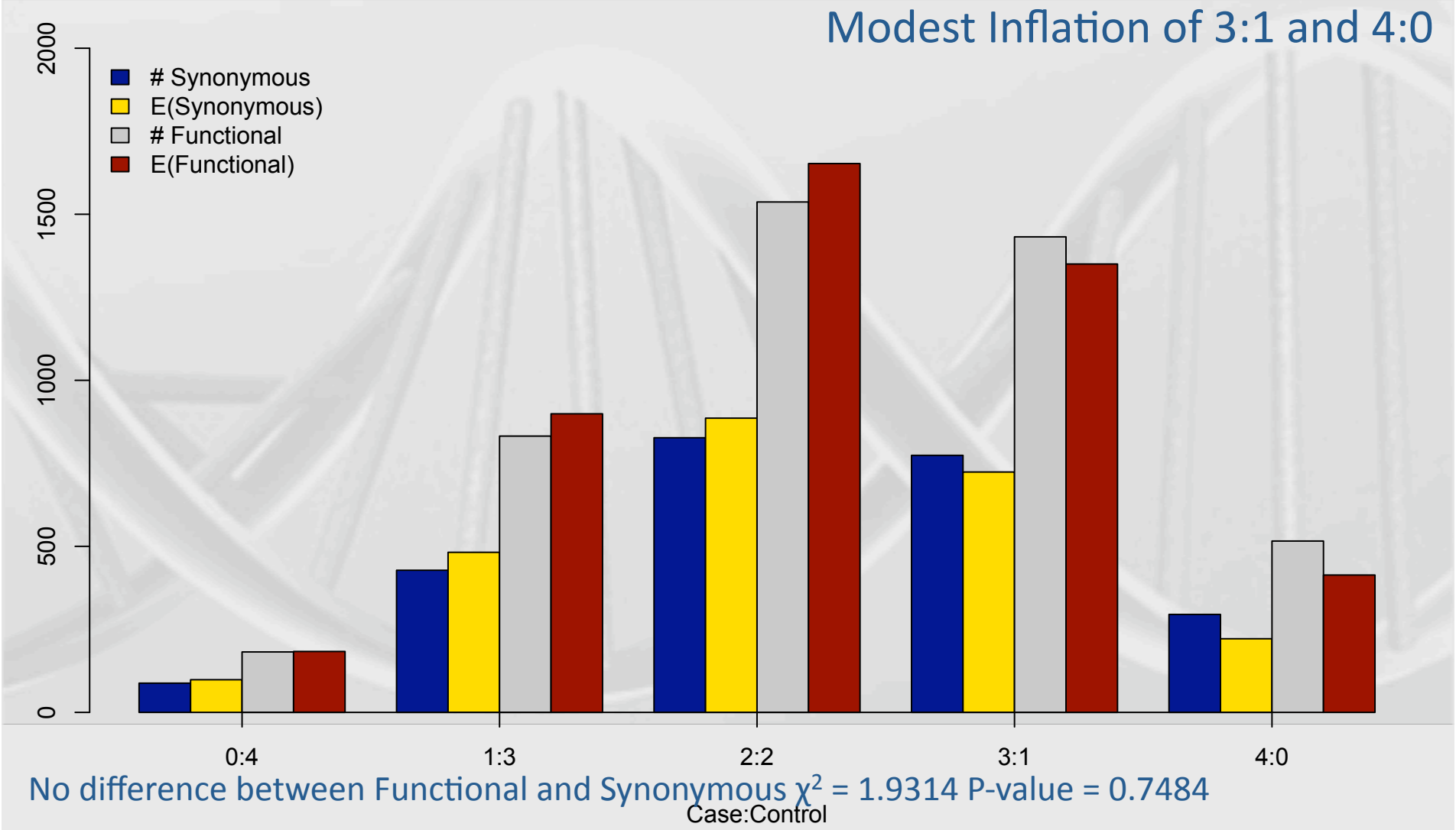
Mild Inflation of 2:0



No difference between Functional and Synonymous $\chi^2 = 1.7042$ P=0.4265

Distribution of 4-tons[All]

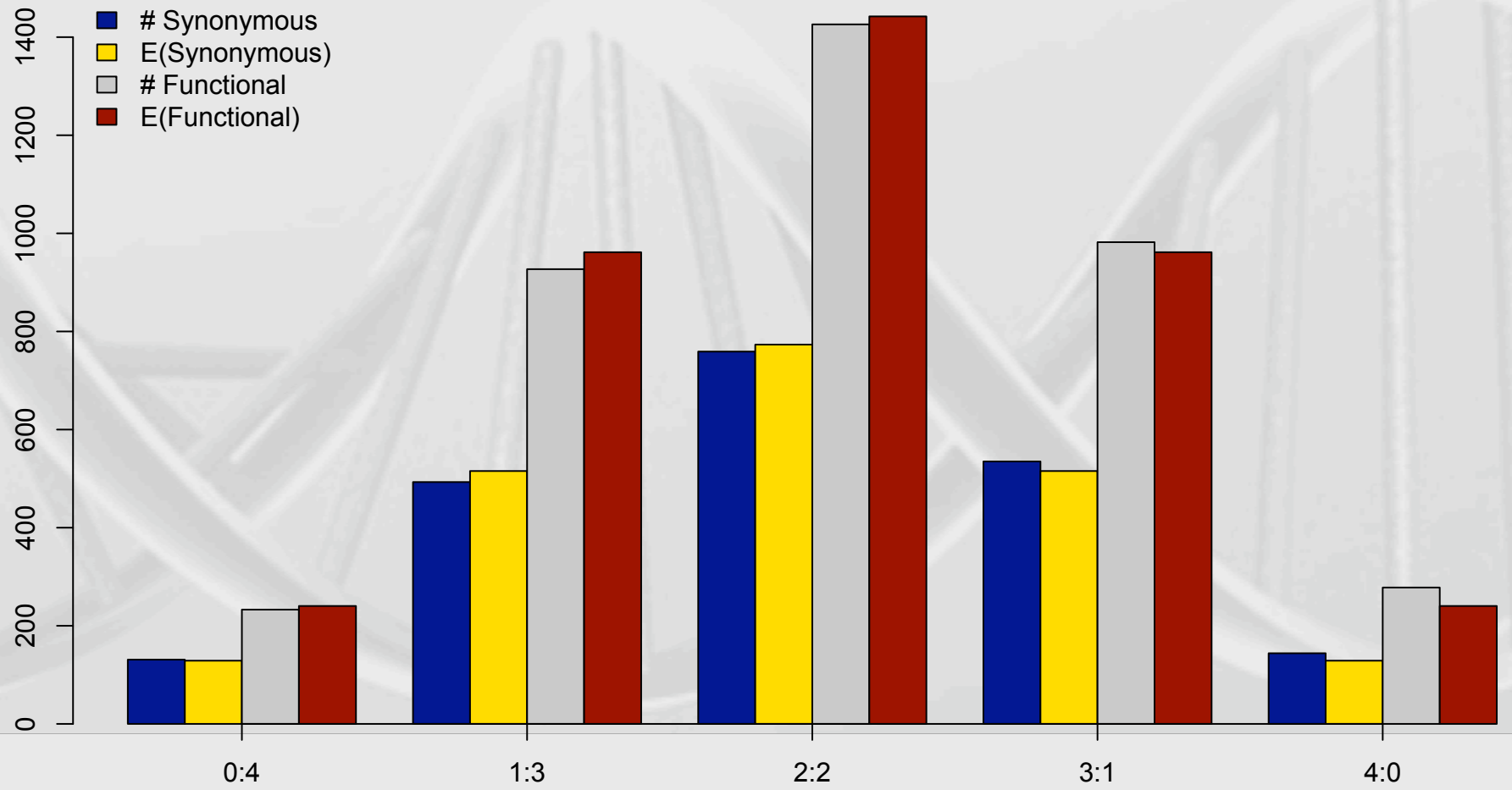
Modest Inflation of 3:1 and 4:0



No difference between Functional and Synonymous $\chi^2 = 1.9314$ P-value = 0.7484

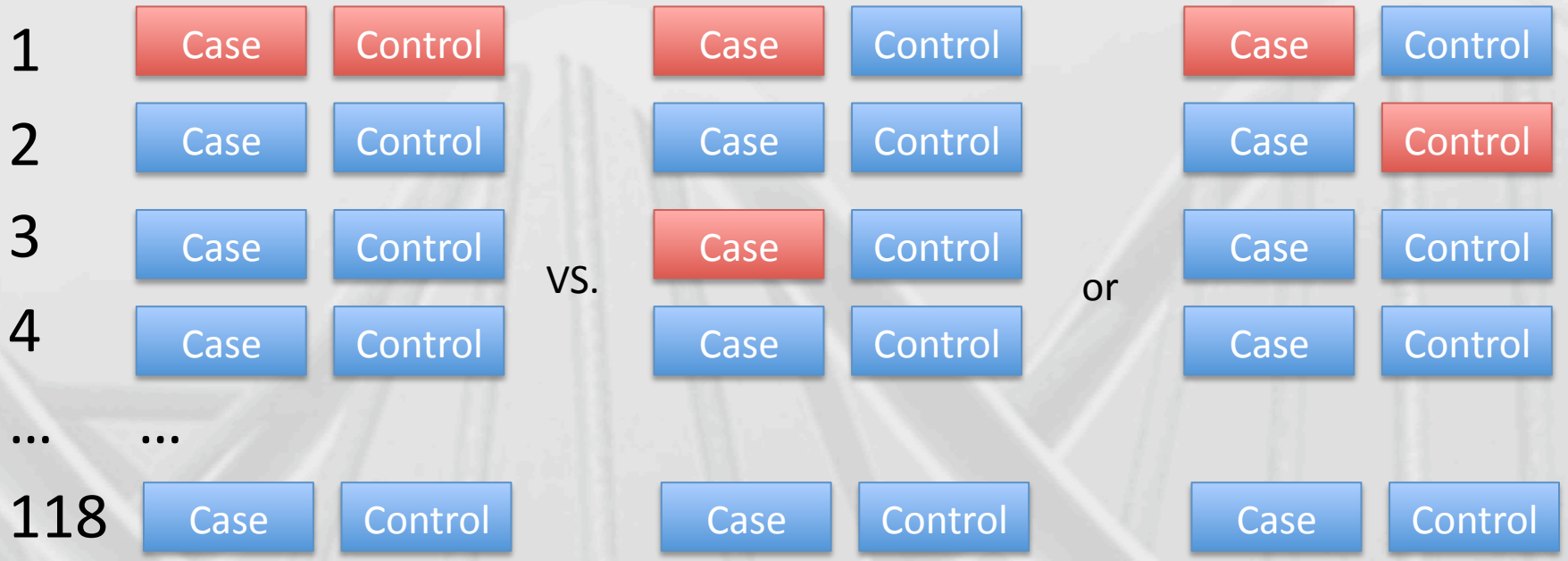
Distribution of 4-tons[Pairs]

Modest Inflation of 3:1 and 4:0



No difference between Functional and Synonymous χ^2=0.4382 P-value = 0.9792

Visual Representation



Data set contains 15,829 doubletons

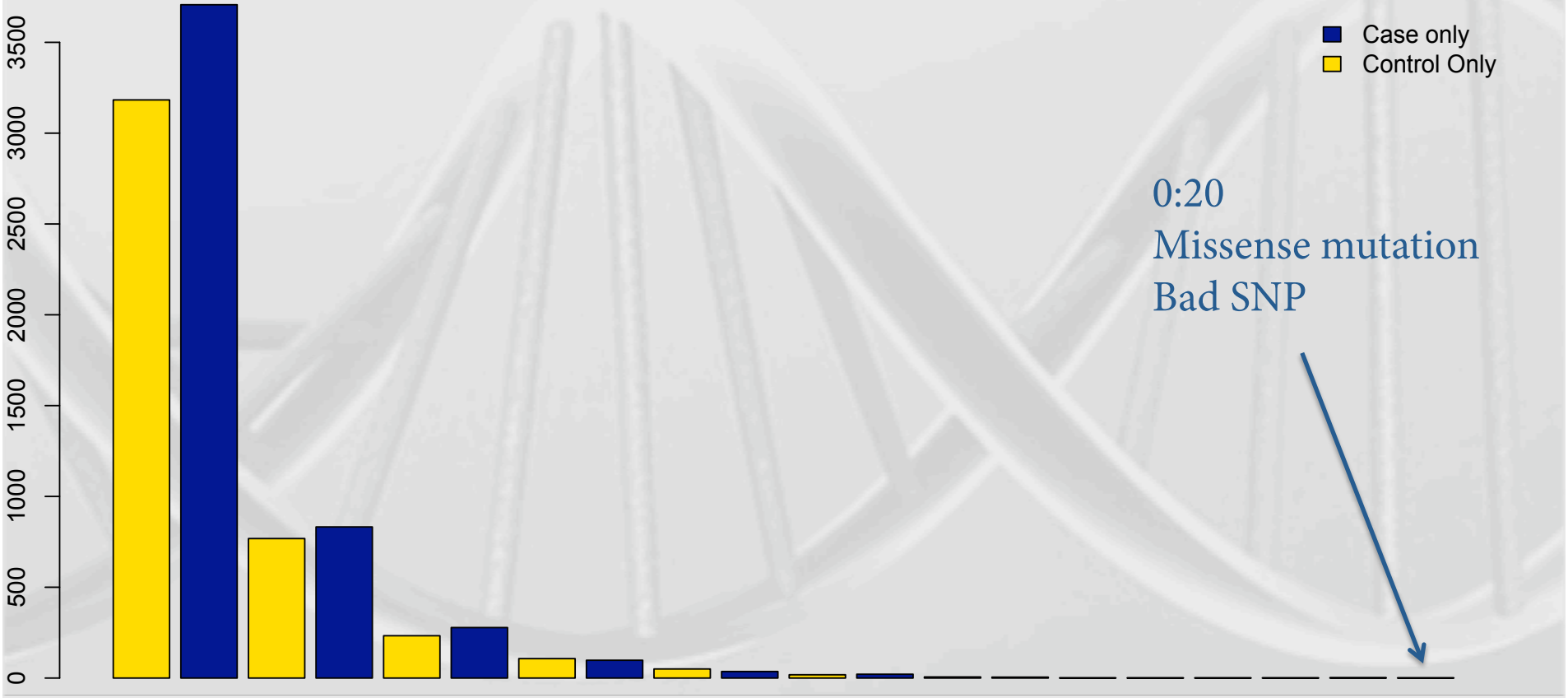
Random Expectation: 67.4 should be between matched pairs

Observation: 163 doubletons occur within matched pairs

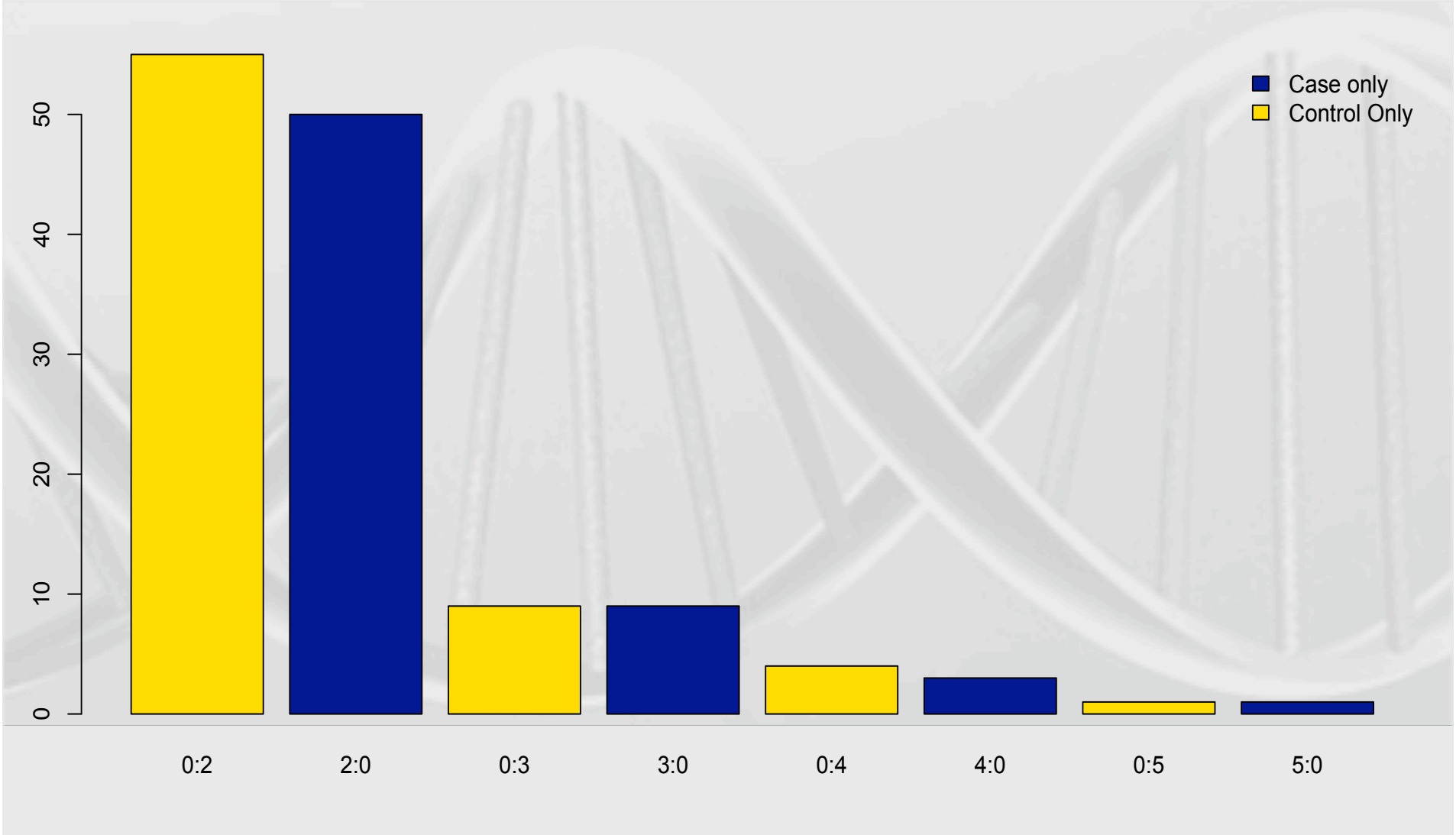
Single Variant distributions

Restricted to 260 matched pairs

Case or Control Only Functional Mutations

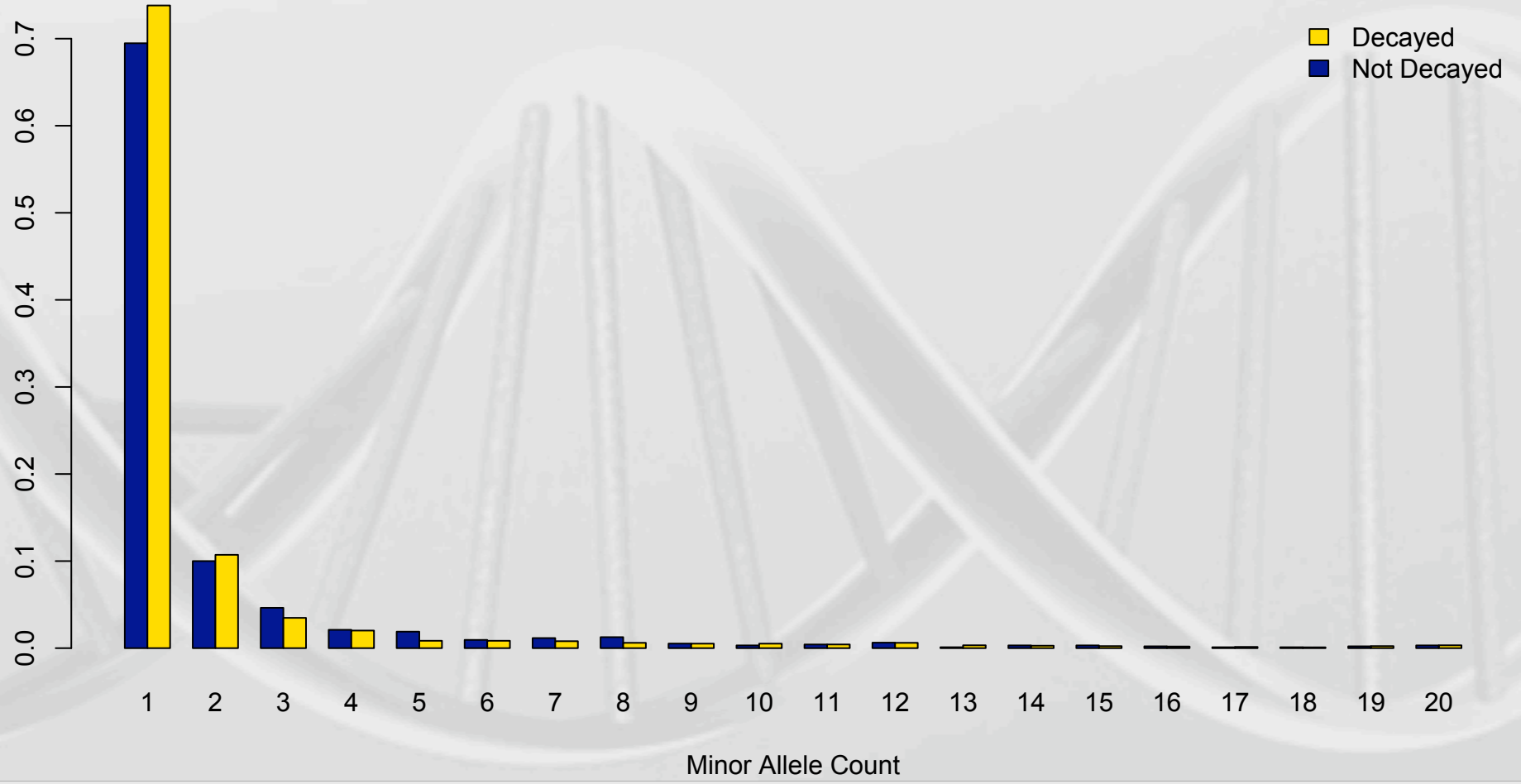


Nonsense Only Mutations



Nonsense Mediated Decay

Nonsense mediated decay



Nonsense Mediated Decay – If nonsense mutation occurs within 50 BP of terminal exon/intron boundary the transcript is made

Trio Analyses

GATK SNP stats

Samples processed: 125

Target size: 32950014 bp

NOVELTY	VARIANTS	TI/TV
all	95213	3.1948
known	53738	3.3393
novel	41475	3.0212

Representative of the very high quality and larger sets of variants coming through new and improved GATK processing pipeline (thanks Corin, Kiran and GSA team!)

Earlier project data now being reprocessed from scratch

Finding *de novos*

- Look for sites in which only a single child has the variant of interest called (*)
  - N=172
- Remove sites where parents have any copies of the minor allele and/or are not adequately covered (i.e., likely latent hets)
  - N=55

Ignores de novos at sites of common variation (very rare occurrences and sites that will be assessed in normal association). We also looked at 19 sites where >= 2 kids and no parents had a variant called and all were low coverage garbage

Distribution per sample

- 14 samples with 0
- 10 samples with 1
- 12 samples with 2
- 3 samples with 3
- 1 sample with 4

- 1 sample with 9 (eliminated pending further examination)

In total –
47 de novos called across 40 trios
Of those:

34 are missense
1 is nonsense
11 are synonymous
(1 noncoding)

So what did we expect?

Rough expectation

- Canonical point mutation rate of 1.2e-8 per bp/gen (Conrad et al., in press) would suggest about 0.7 per exome
- Comparing sequence context of coding exons versus the rest of the genome projects a 1.4x increase in mutation rate in exons (Jared Maguire et al.)
  - Revise expectation to 1.0 per exome
- Roughly 10% of exome not adequately covered for making the determination
  - Revise expectation to 0.9 per exome in our experiment
- Given mutational preferences and actual coding sequence of humans, we expect 29-30% to be synonymous, 67% missense and 3-4% nonsense.

Distribution per sample

- 14 samples with 0
- 10 samples with 1
- 12 samples with 2
- 3 samples with 3
- 1 sample with 4

In total –
47 de novos called across 40 trios
Of those:

34 are missense
1 is nonsense
11 are synonymous
(1 noncoding)

So what did we expect?

Our observations (46 events (36 expected), 76% 'functional' (70% expected) are slightly in excess of chance but not dramatically…

Conclusions

- Quality of data excellent
- No obvious ‘slam dunks’
- Excess of missense and nonsense X:0, but not reverse
  - Maybe IBD? Polygenic underpinnings in rare variation?