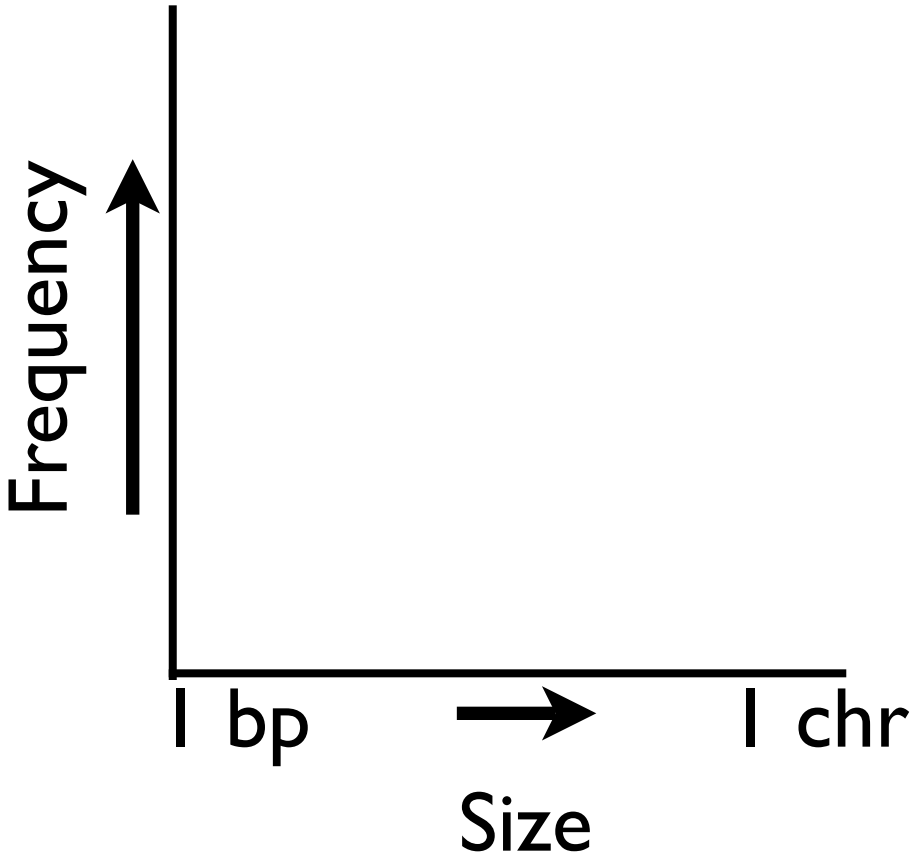
A Method for Rapid, Targeted CNV Genotyping Identifies Rare Variants Associated with Neurological Disease

Gregory Cooper, Ph.D.

Department of Genome Sciences, University of Washington

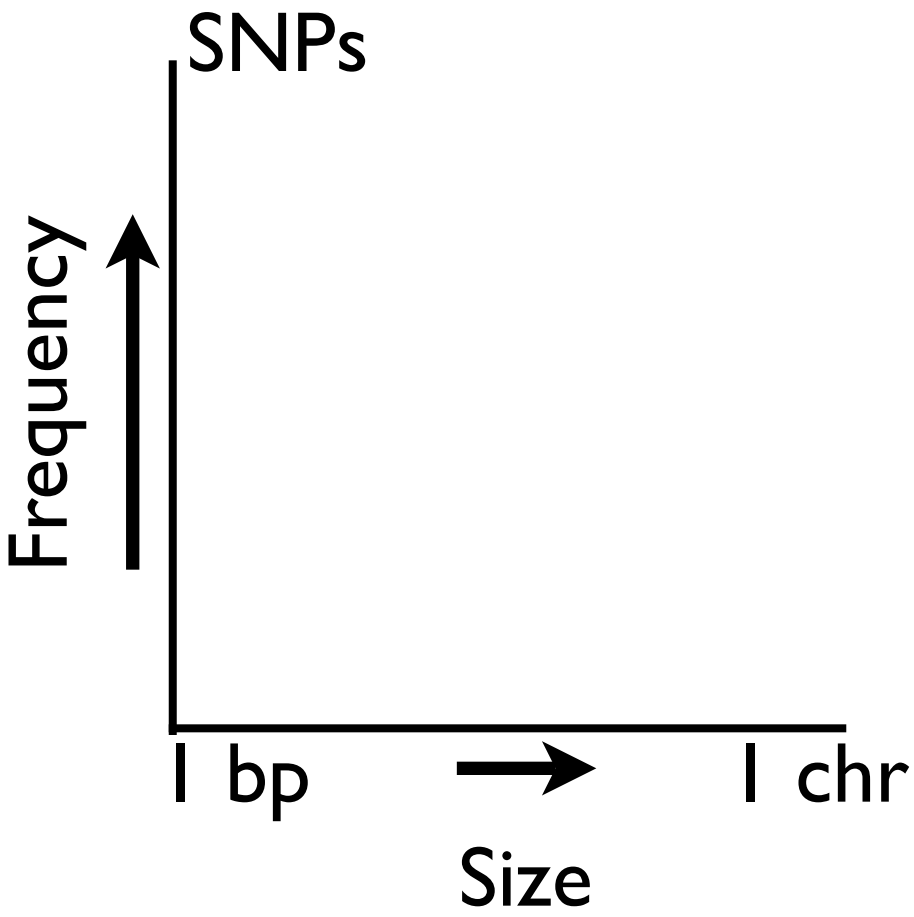
Genomic Structural Variation

Human Genetic Variation



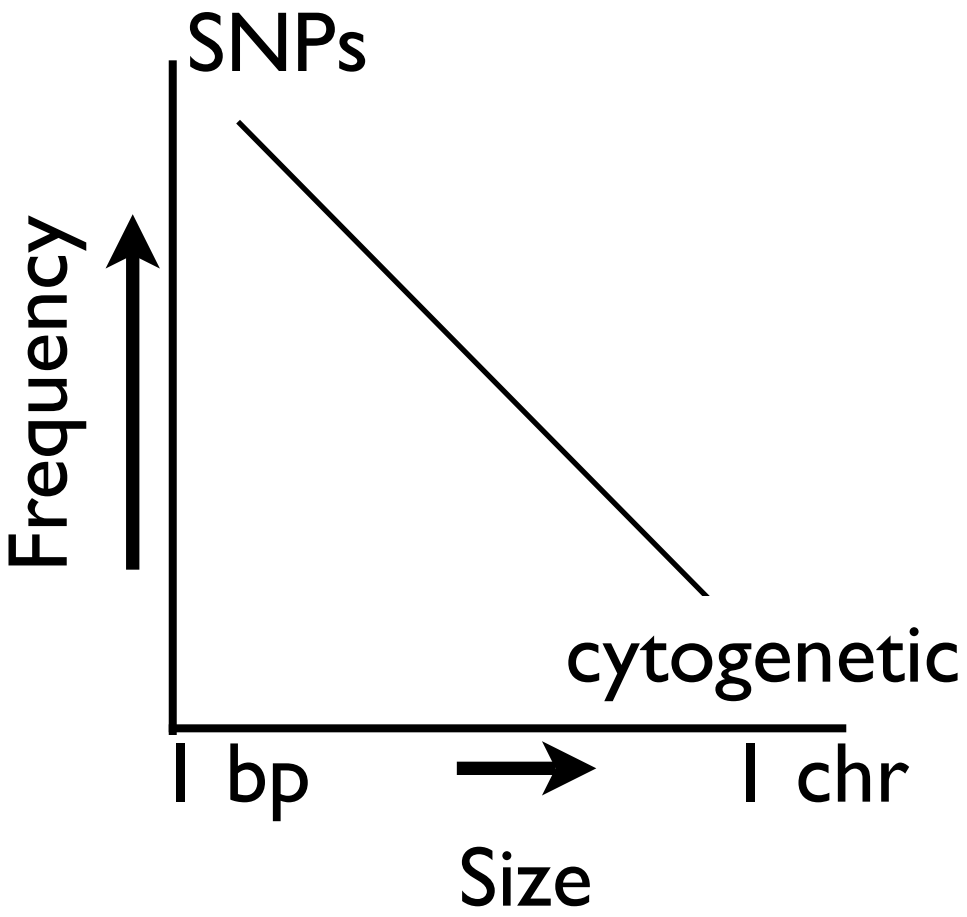
Genomic Structural Variation

Human Genetic Variation



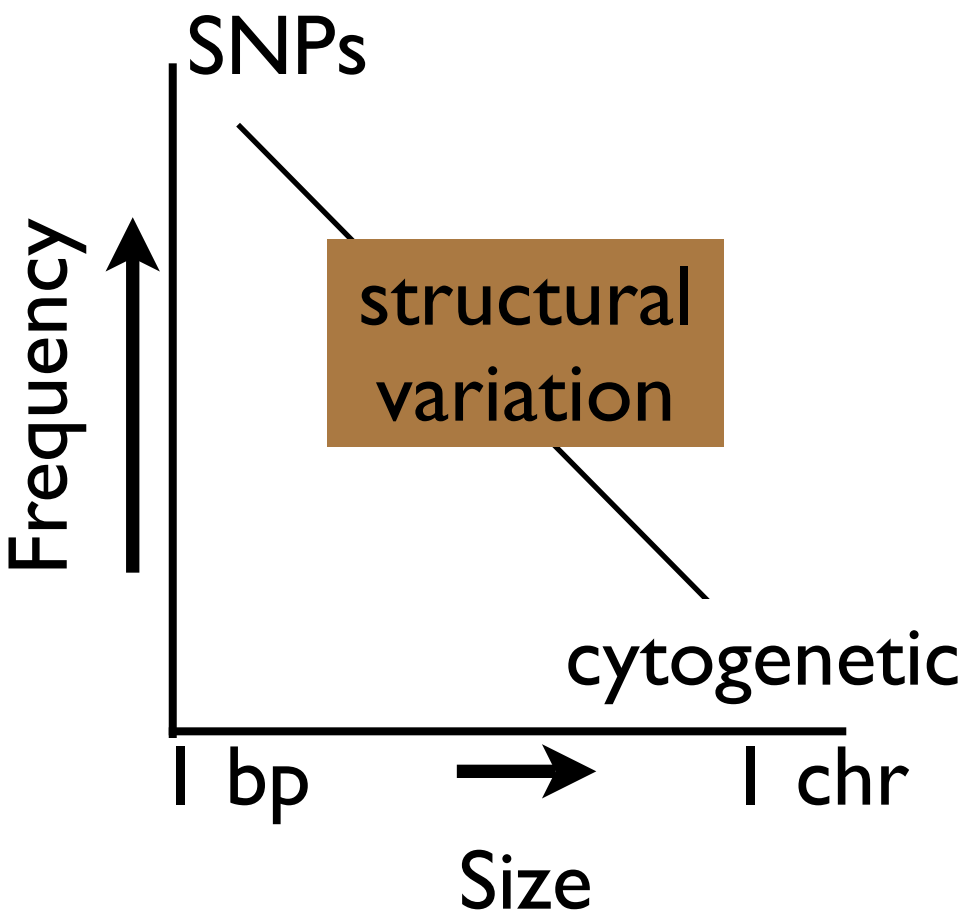
Genomic Structural Variation

Human Genetic Variation

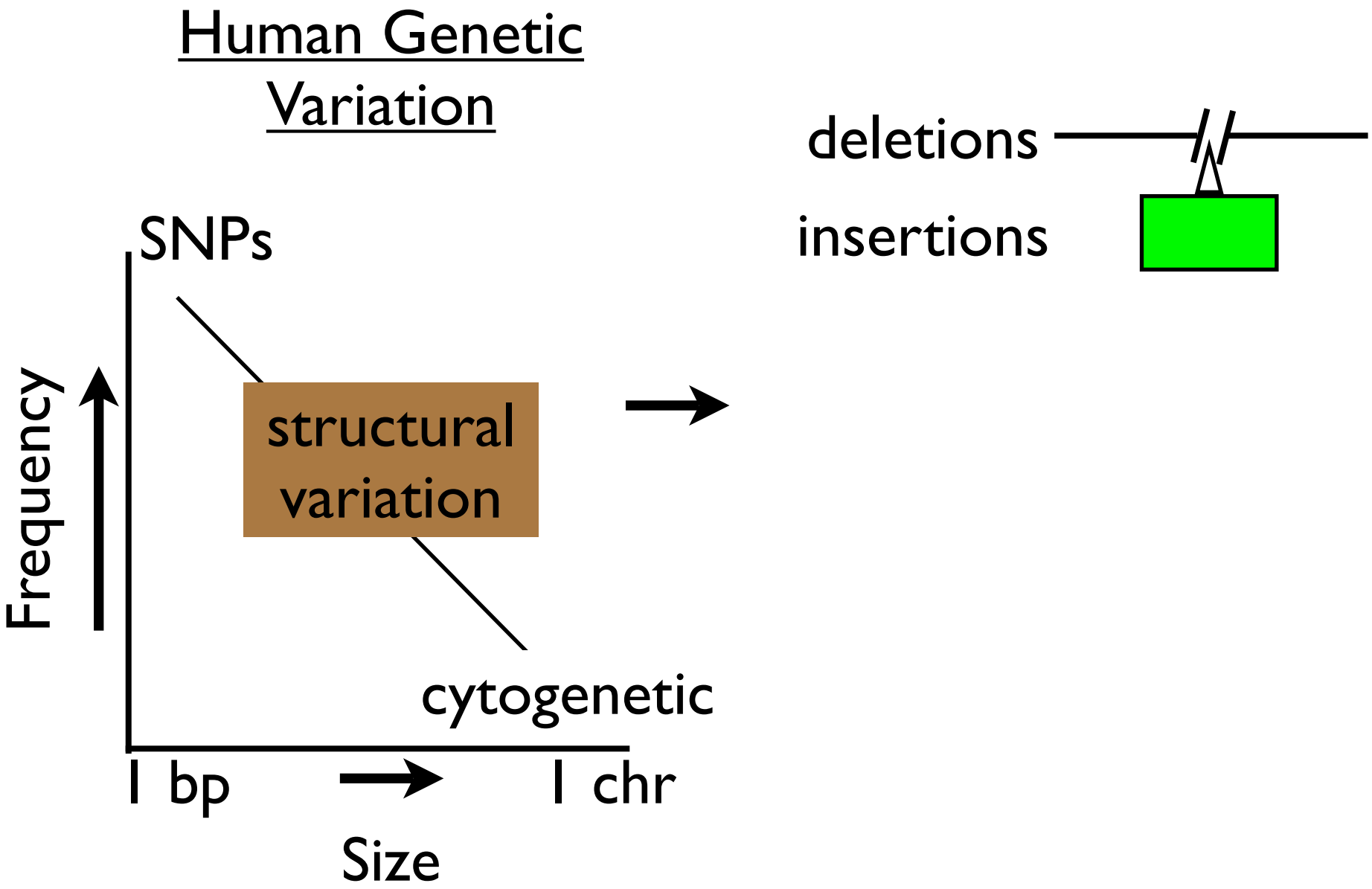


Genomic Structural Variation

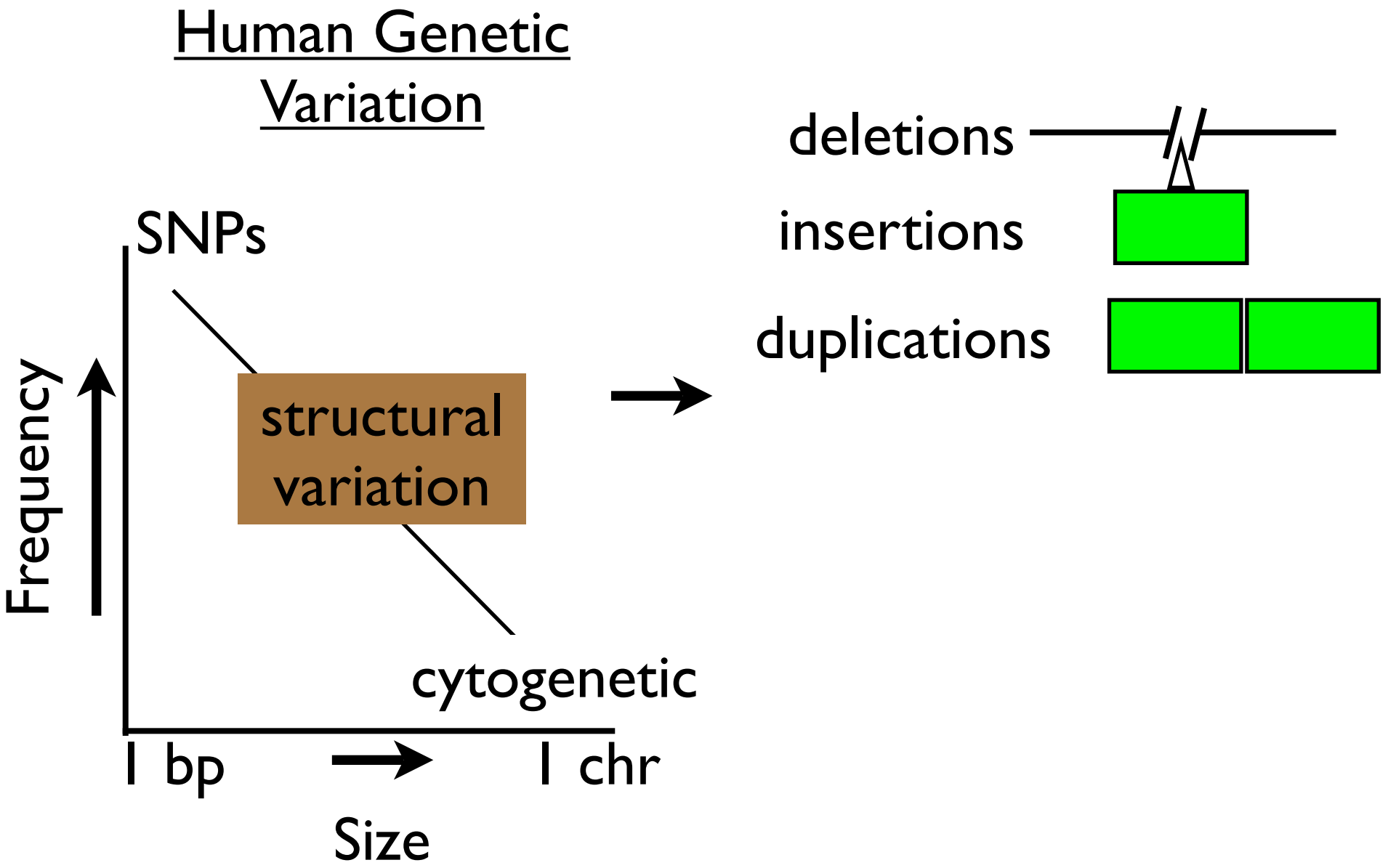
Human Genetic Variation



Genomic Structural Variation

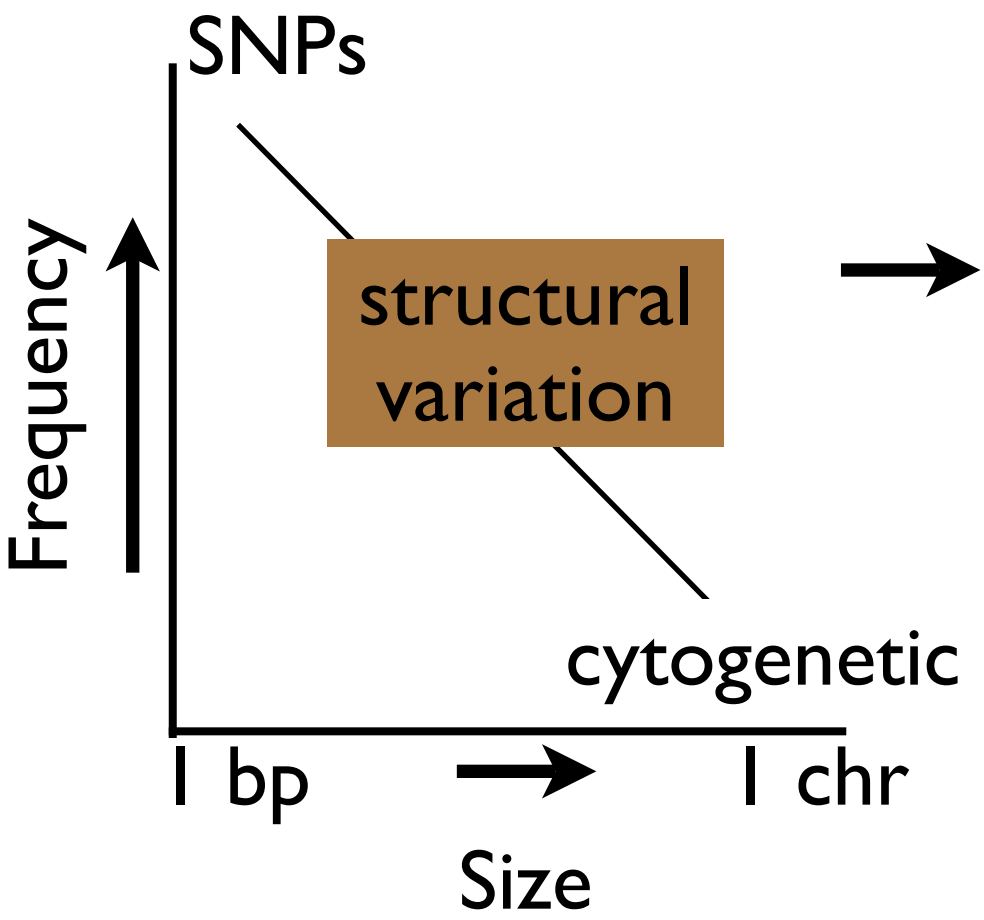


Genomic Structural Variation

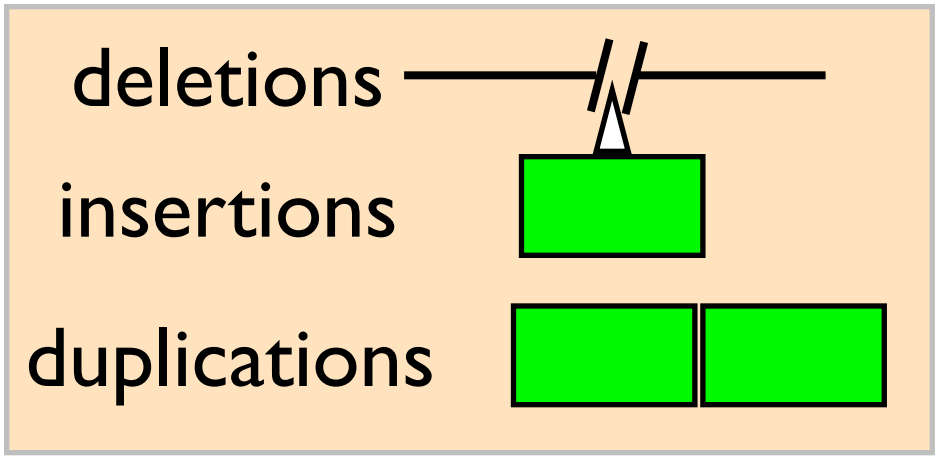


Genomic Structural Variation

Human Genetic Variation

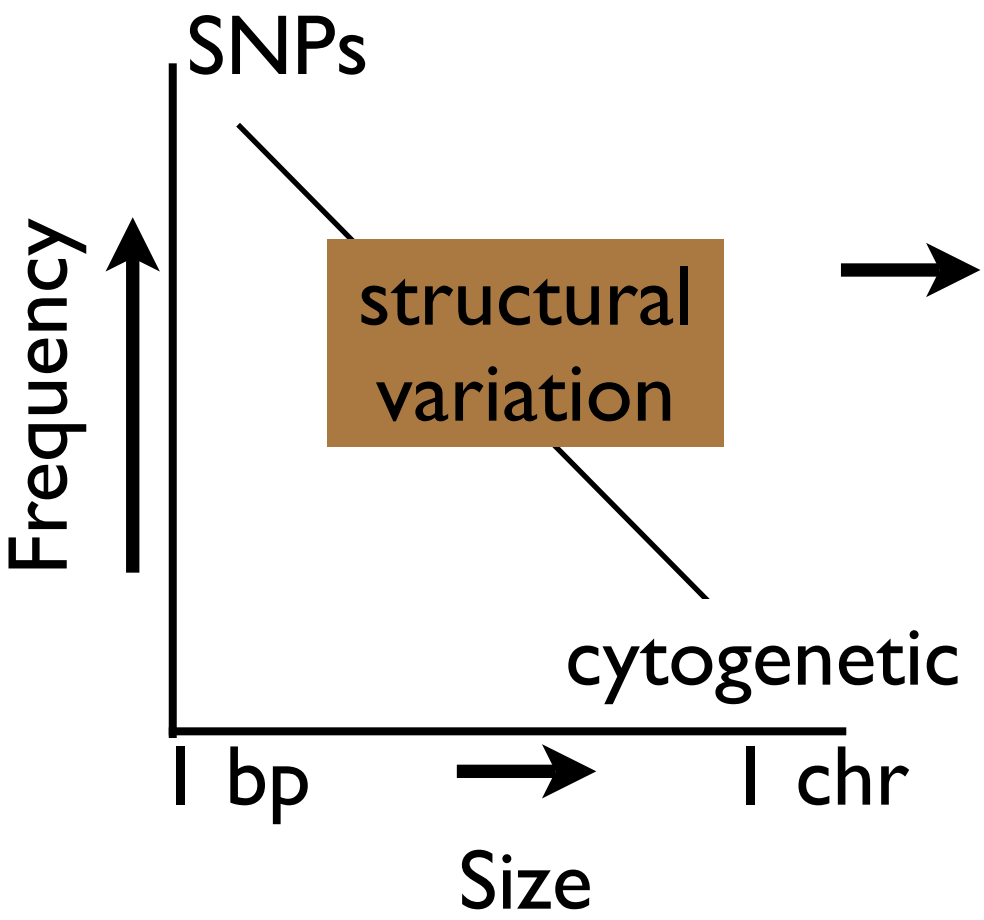


Copy-Number Variants

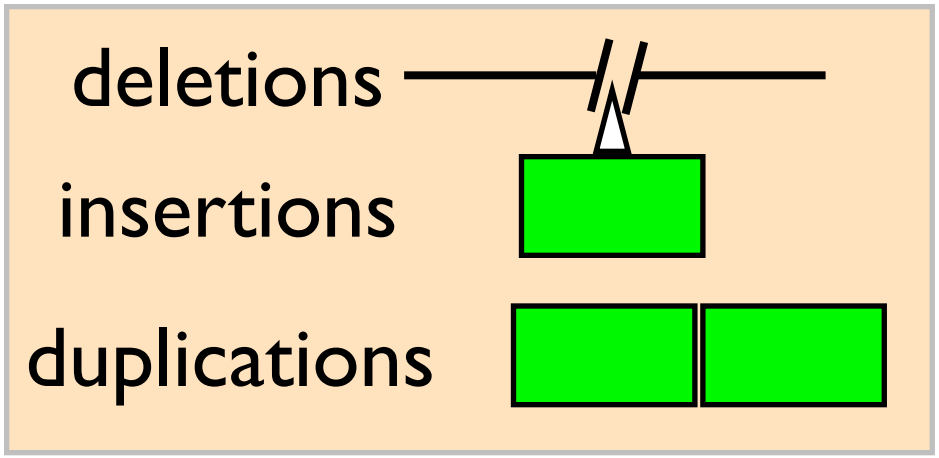


Genomic Structural Variation

Human Genetic Variation



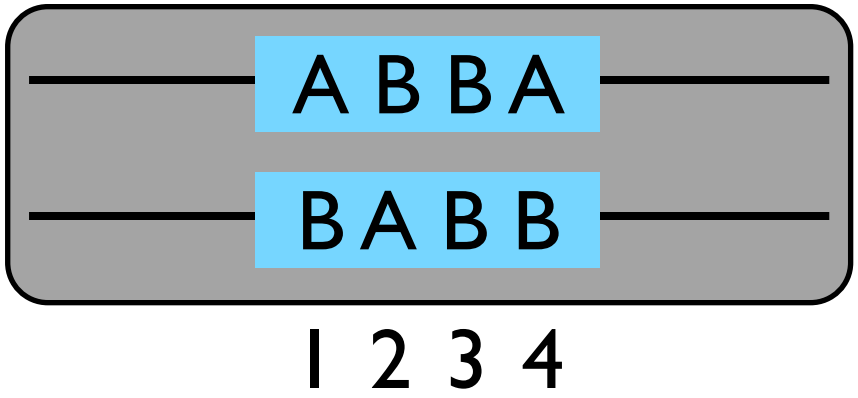
Copy-Number Variants



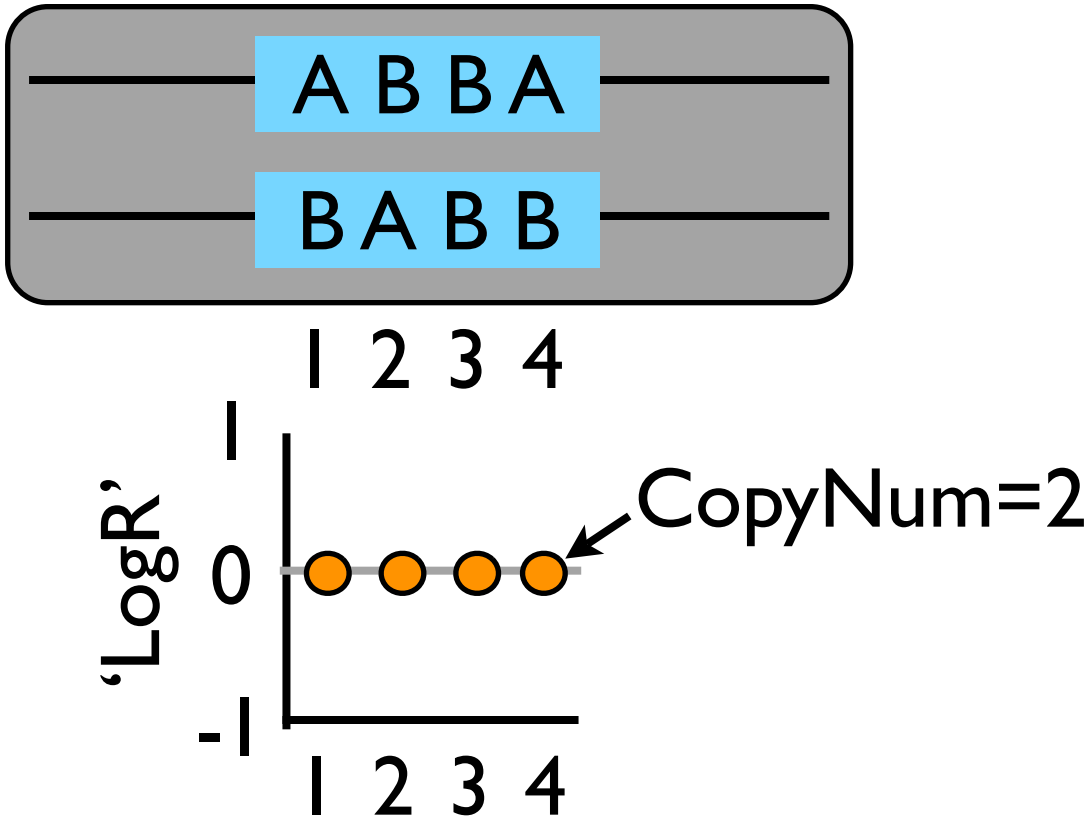
- Gene-rich, e.g. immune response, drug metabolism
- Abundant: majority of human heterozygosity
- Technological challenges have impeded large-scale analyses

SNP-based Deletion Discovery

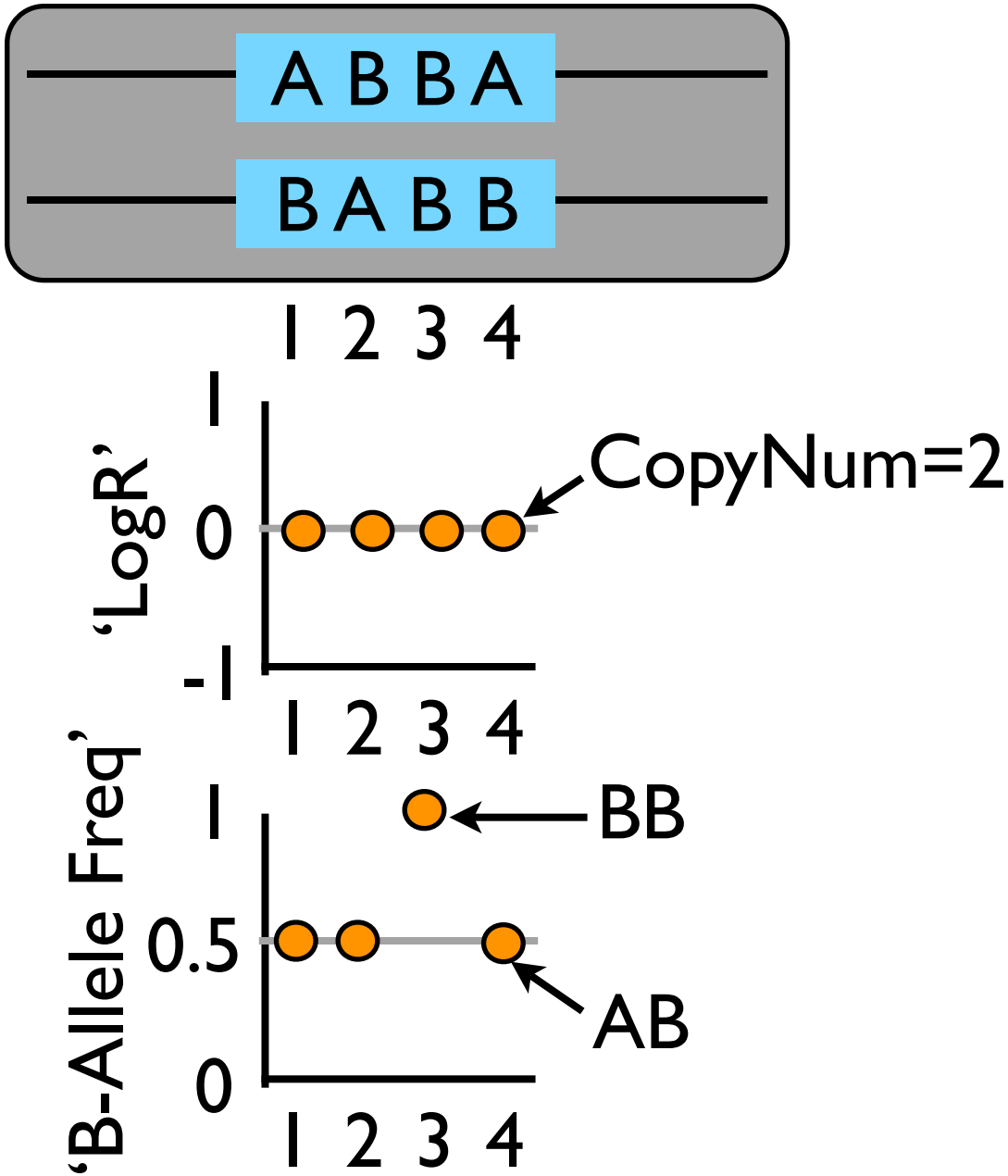
SNP-based Deletion Discovery



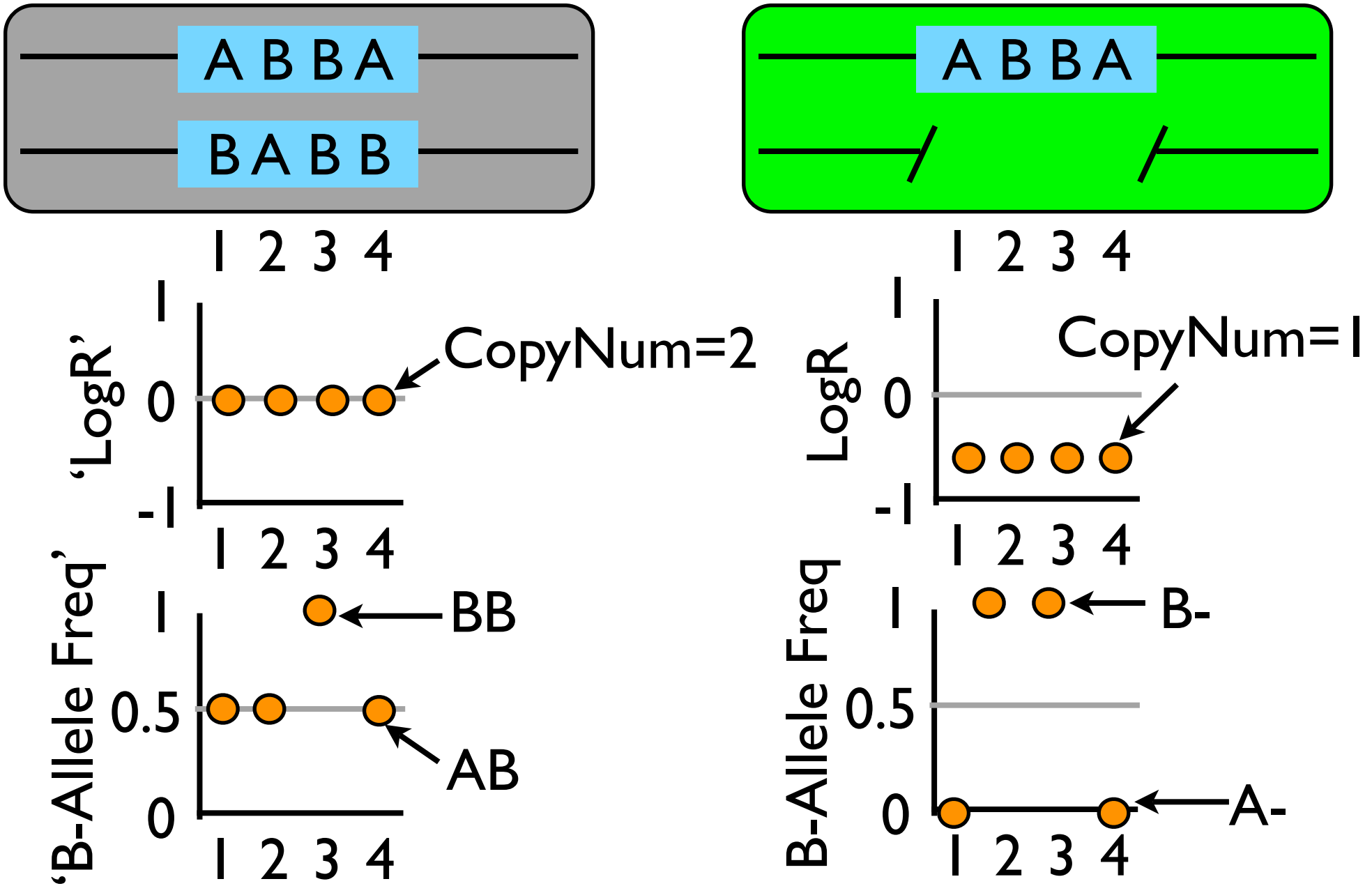
SNP-based Deletion Discovery



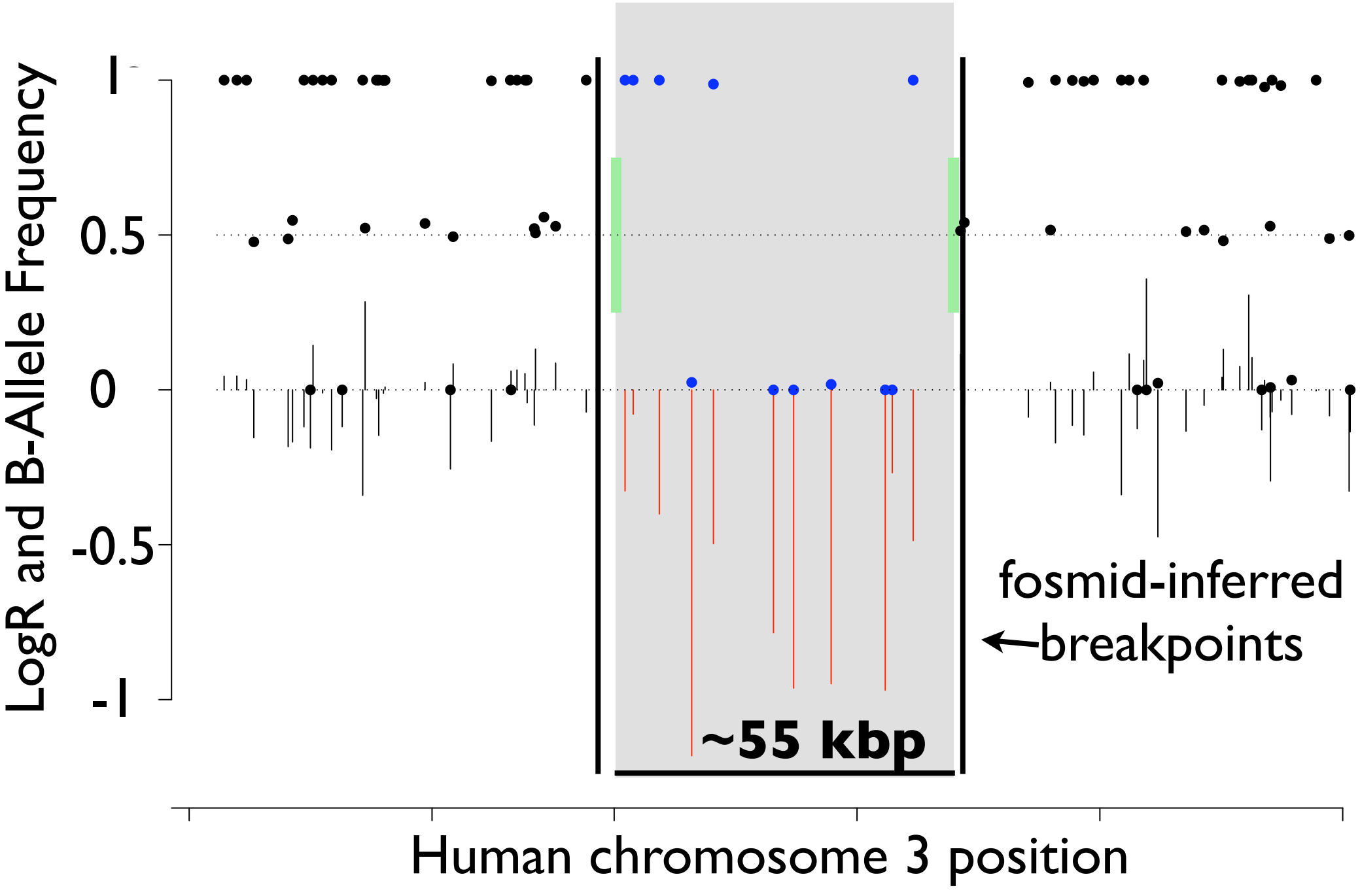
SNP-based Deletion Discovery



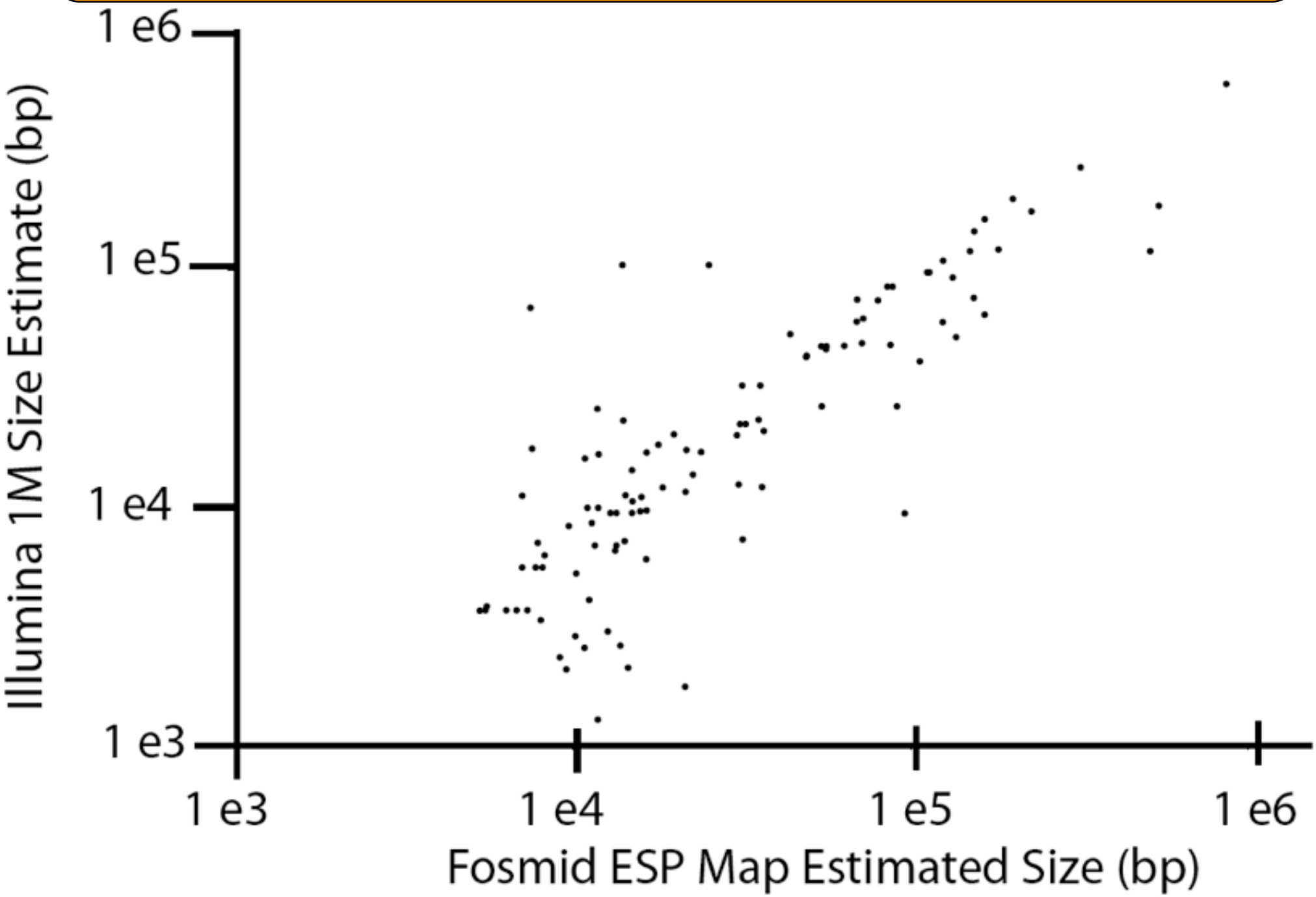
SNP-based Deletion Discovery



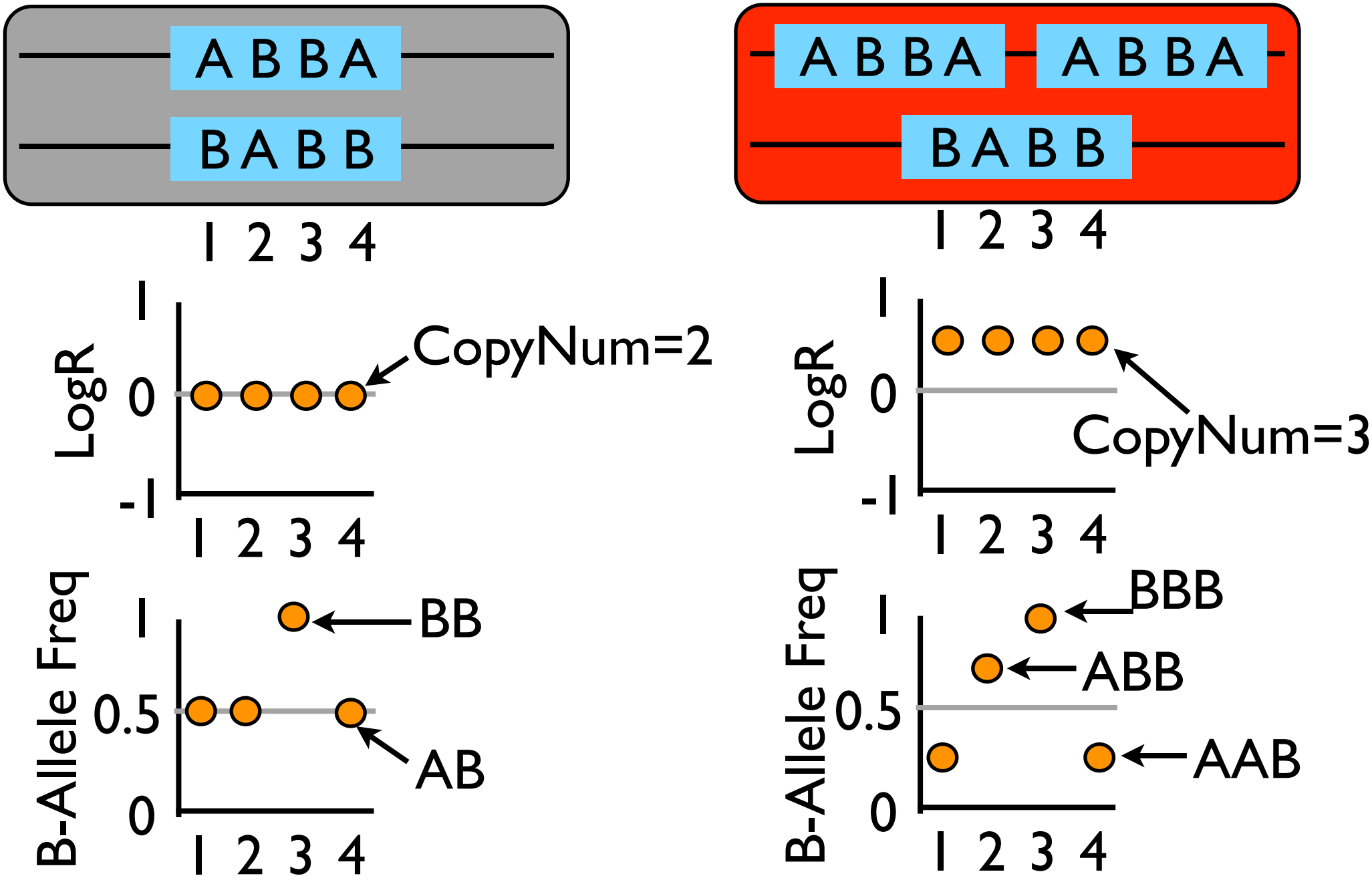
Illumina 1M Deletion Discovery



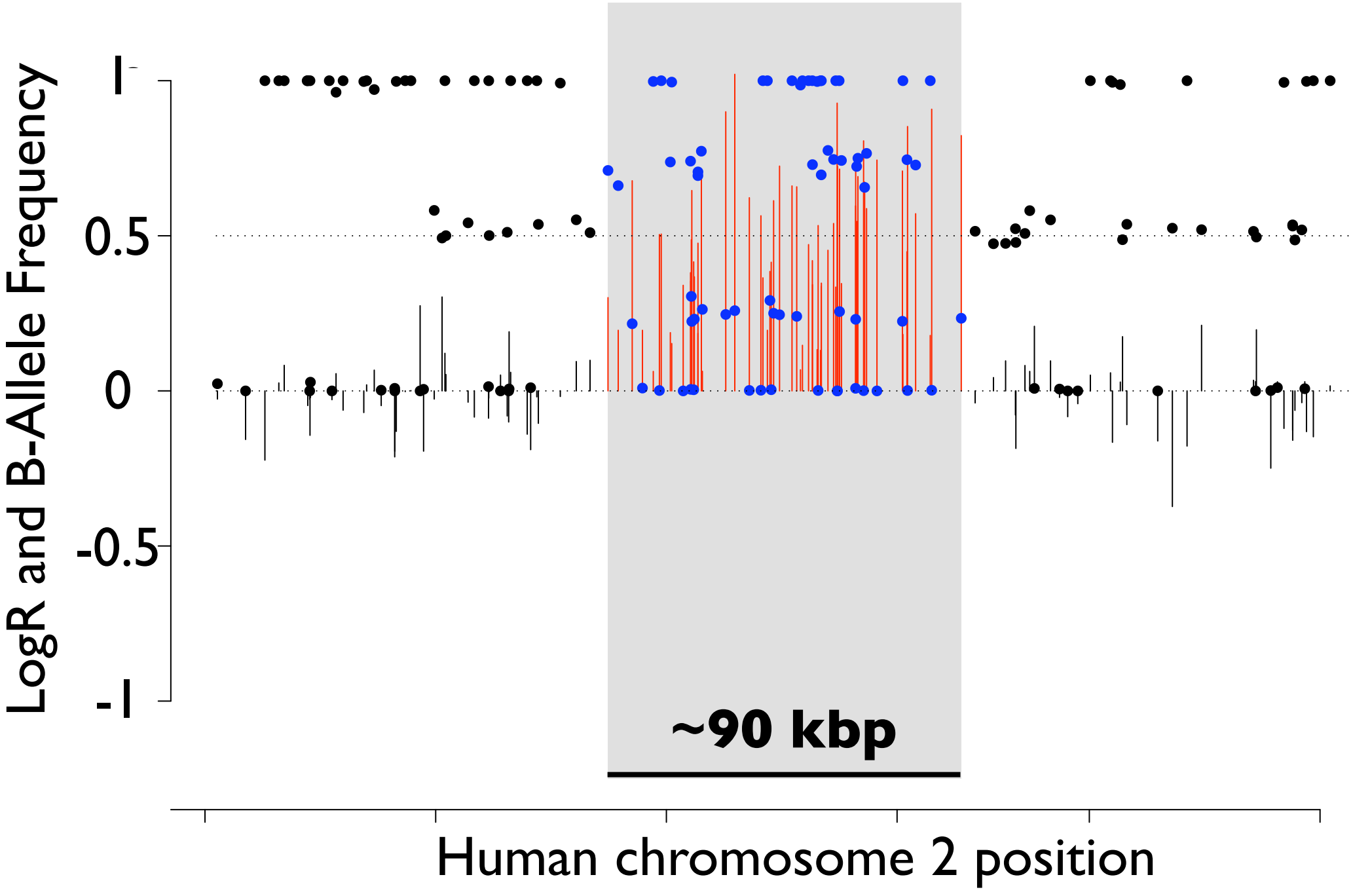
Illumina 1M Deletion Discovery



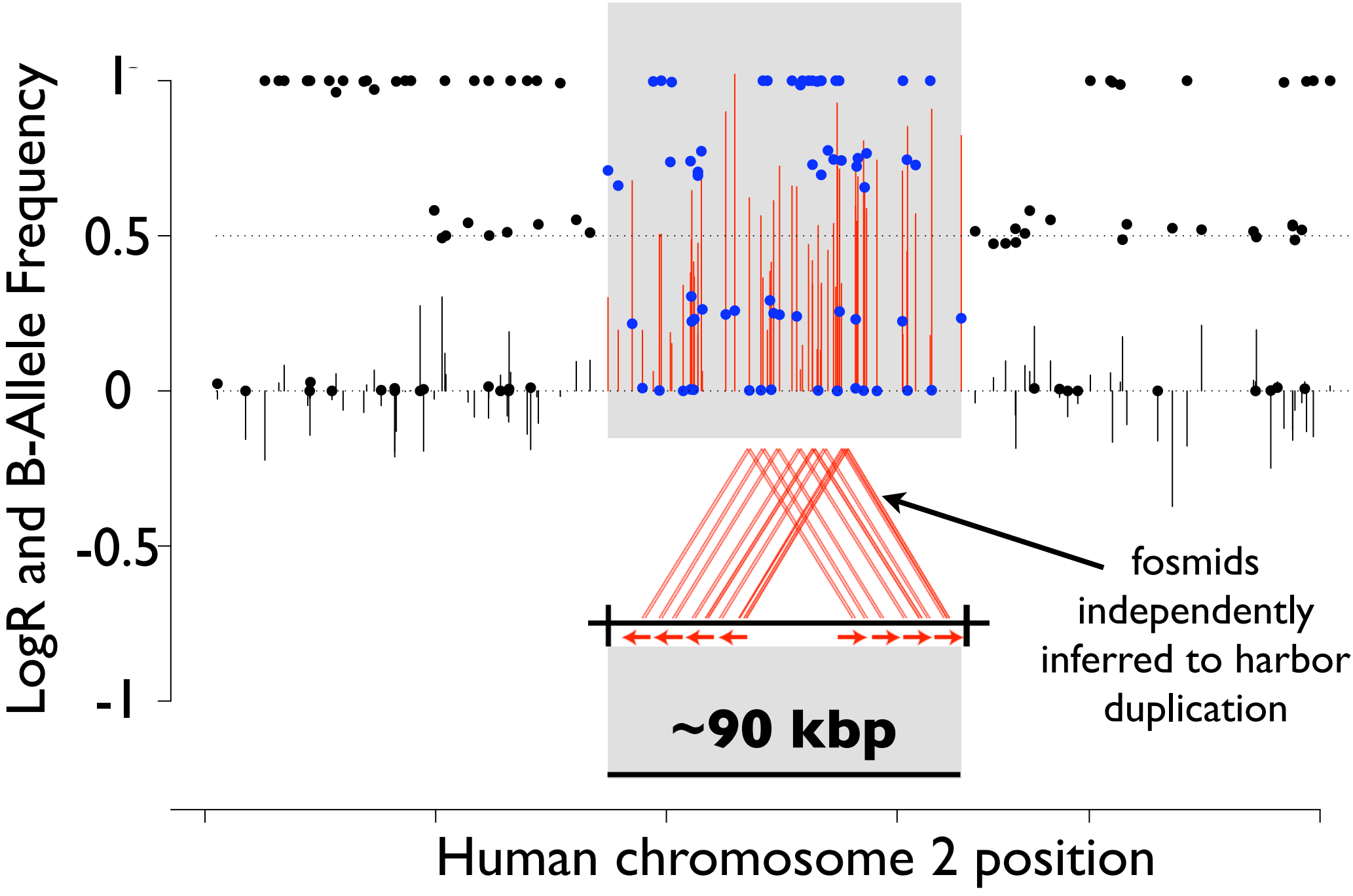
SNP-based Duplication Discovery



Illumina 1M Duplication Discovery



Illumina 1M Duplication Discovery



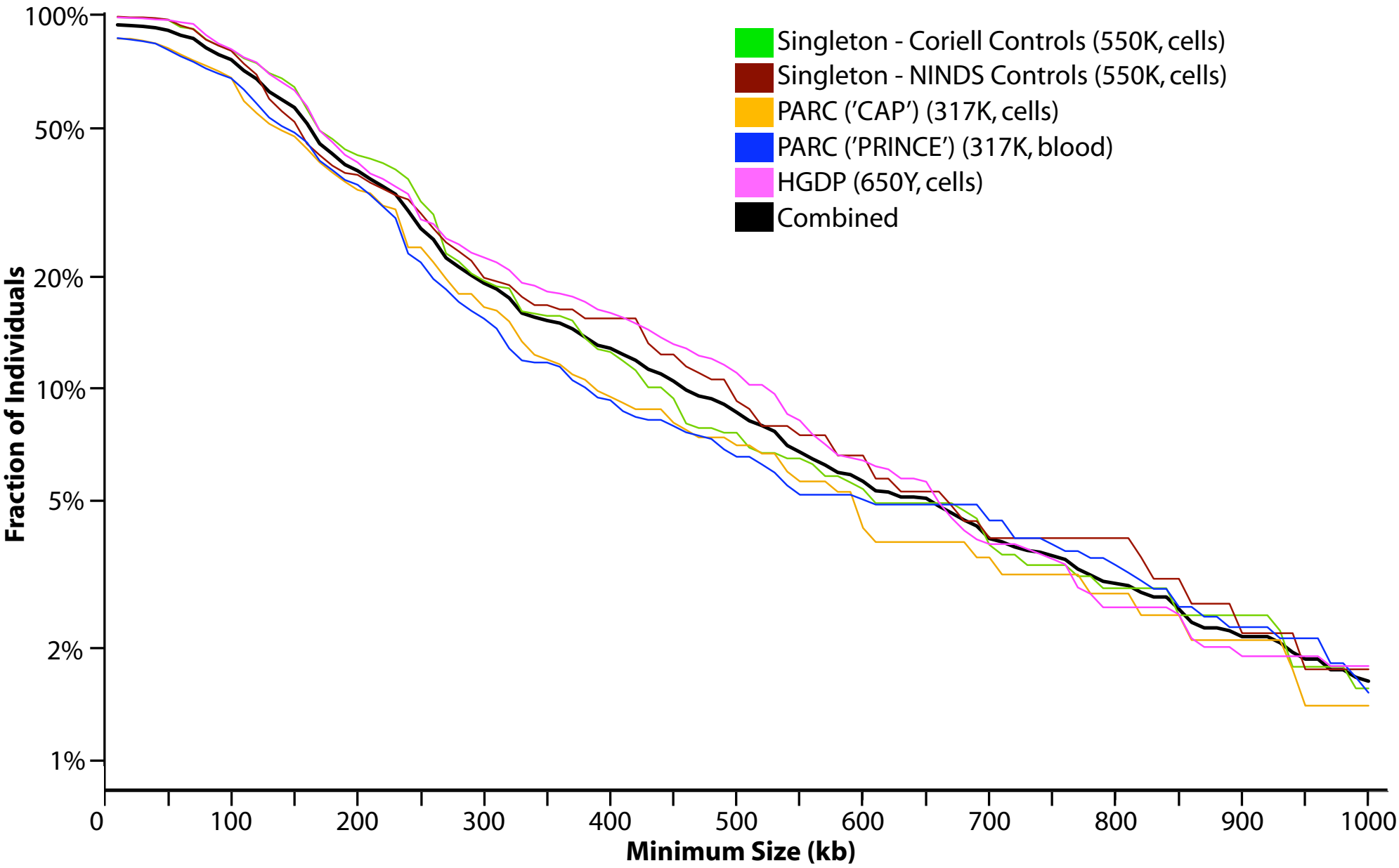
Discovering CNVs in Large Cohorts

- ~1,000 individuals from the PARC project (Illumina 317k SNP arrays)
  - Caucasian samples from a statin pharmacogenetics study
- ~1,000 samples from the Human Genome Diversity Panel (Illumina 650Y chips; Li, Absher, et al *Science* 2008):
  - samples collected in diverse regions of the world
- ~800 neurological disease controls (Illumina 550K chips; Andy Singleton; Walsh et al, *Science* 2008)
  - samples screened for symptoms of psychiatric disease
- n = 2,493 samples after QC (600 blood DNA)

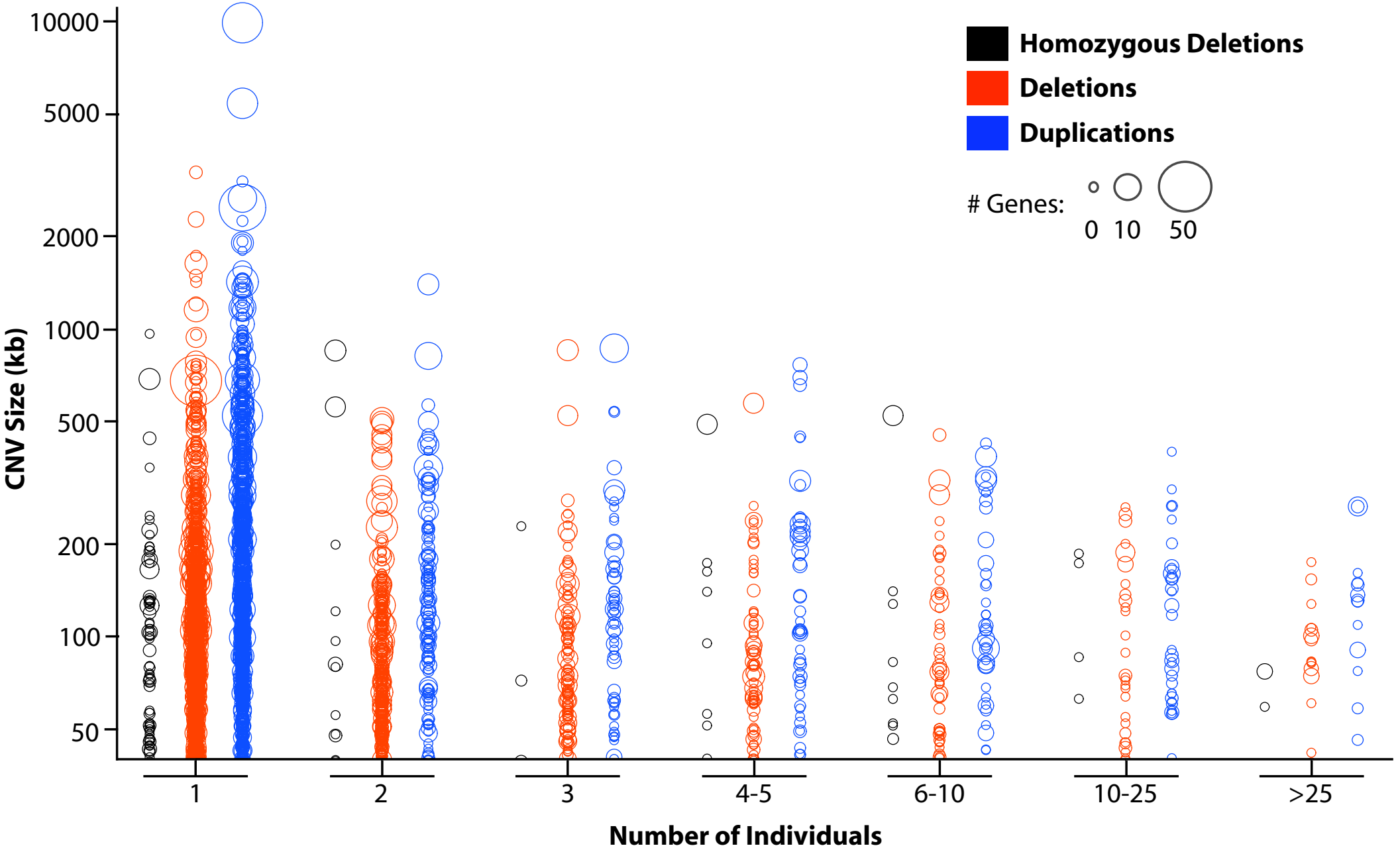
Large CNV Overview

Study	Platform	# Samples	CNVs	Calls/Sample
PARC	HH317K	936 (991)	2,664	2.85
Neurological Disease Controls	HH550K	671 (790)	4,641	6.92
HGDP	HH650Y	886 (941)	6,538	7.38
Total		**2,493 (2,722)**	**13,843**	**5.56**

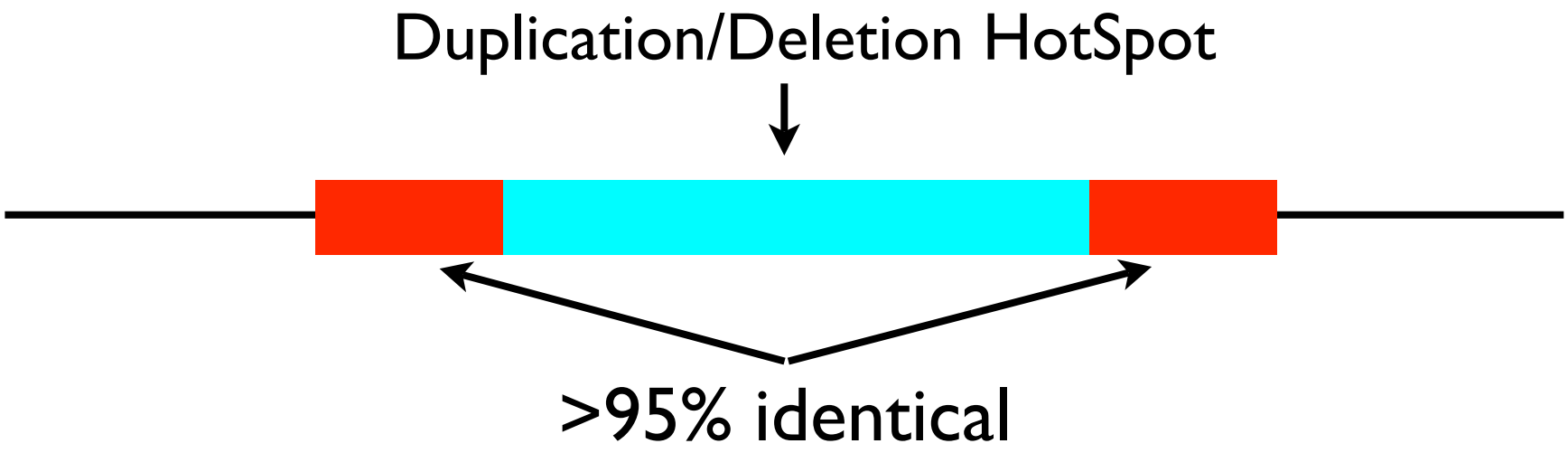
Many Individuals Carry Large Events



Collectively Frequent But Individually Rare

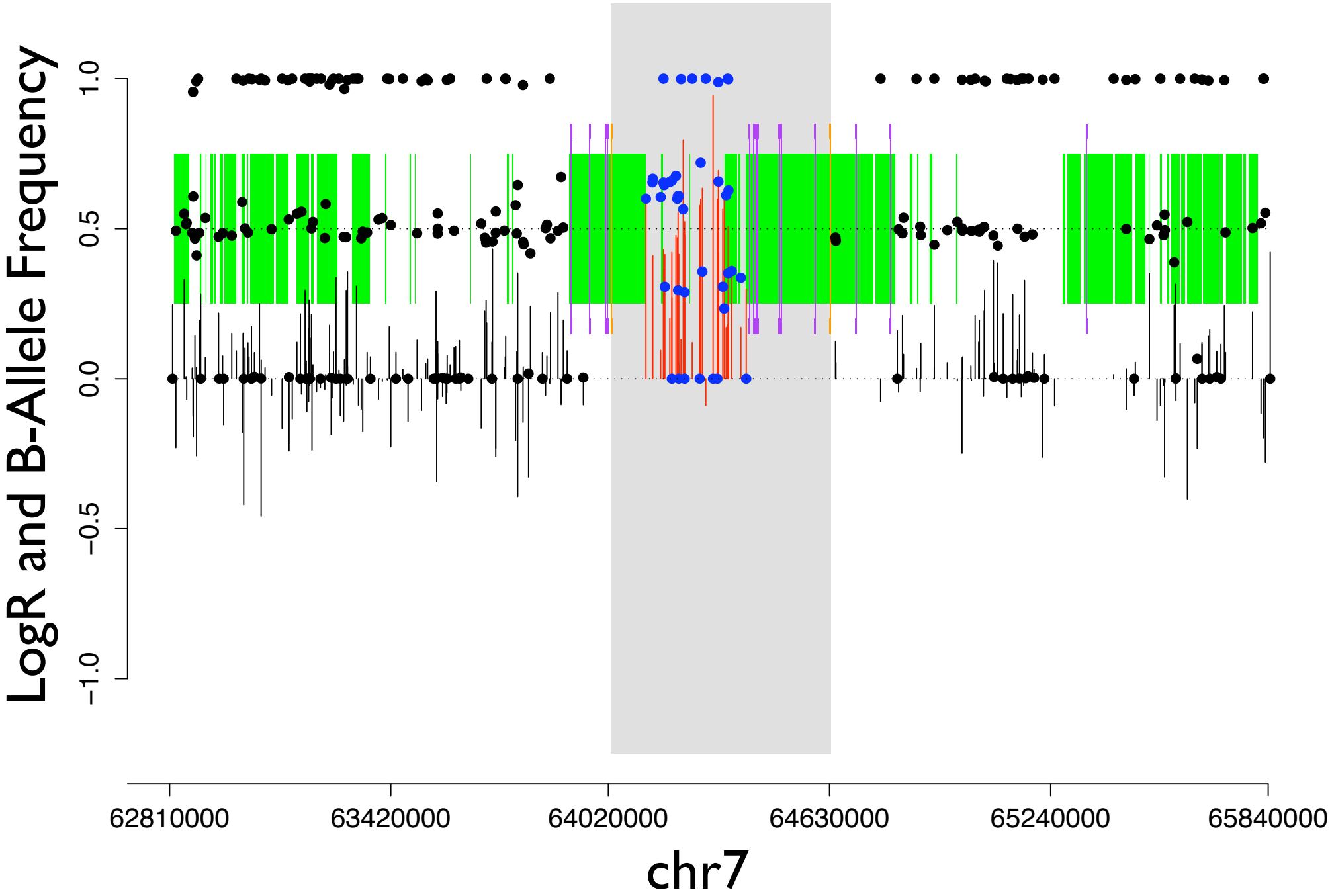


Genomic CNV 'HotSpots'

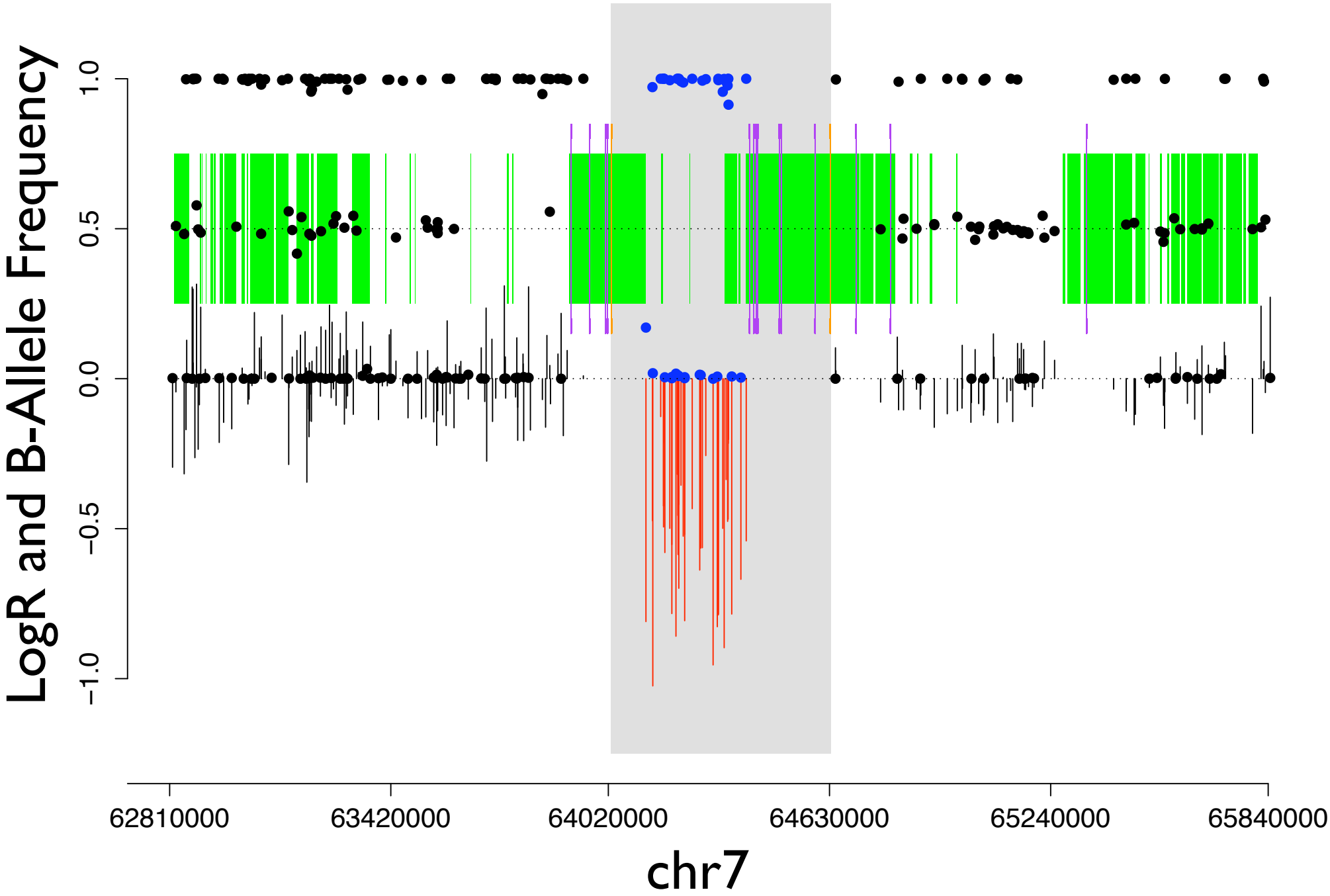


- Non-Allelic Homologous Recombination (NAHR) between duplicated sequences results in novel CNVs
- Thousands of potential hotspots in the reference assembly (1 kb to Mbps in size)
- *de novo* hotspot mutations have been implicated as causative for a number of diseases

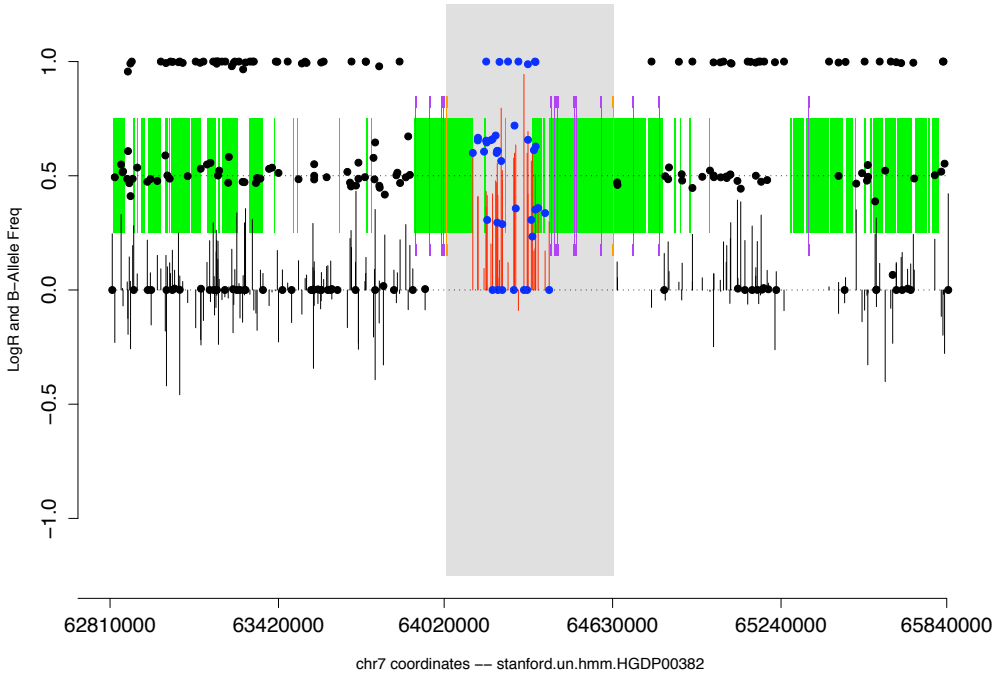
CNVs Enriched Near Segmental Duplications

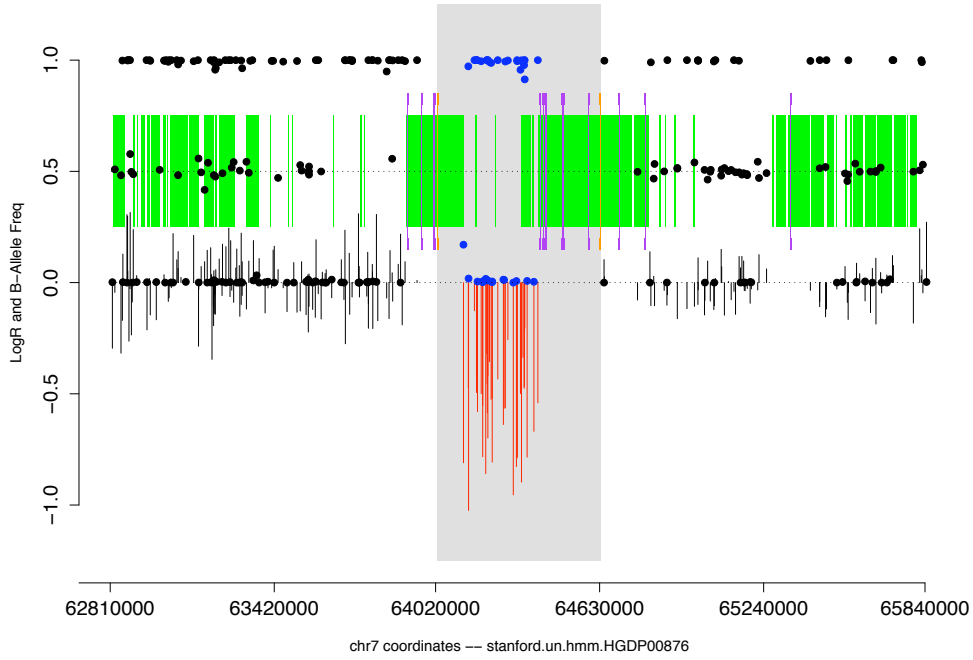


CNVs Enriched Near Segmental Duplications



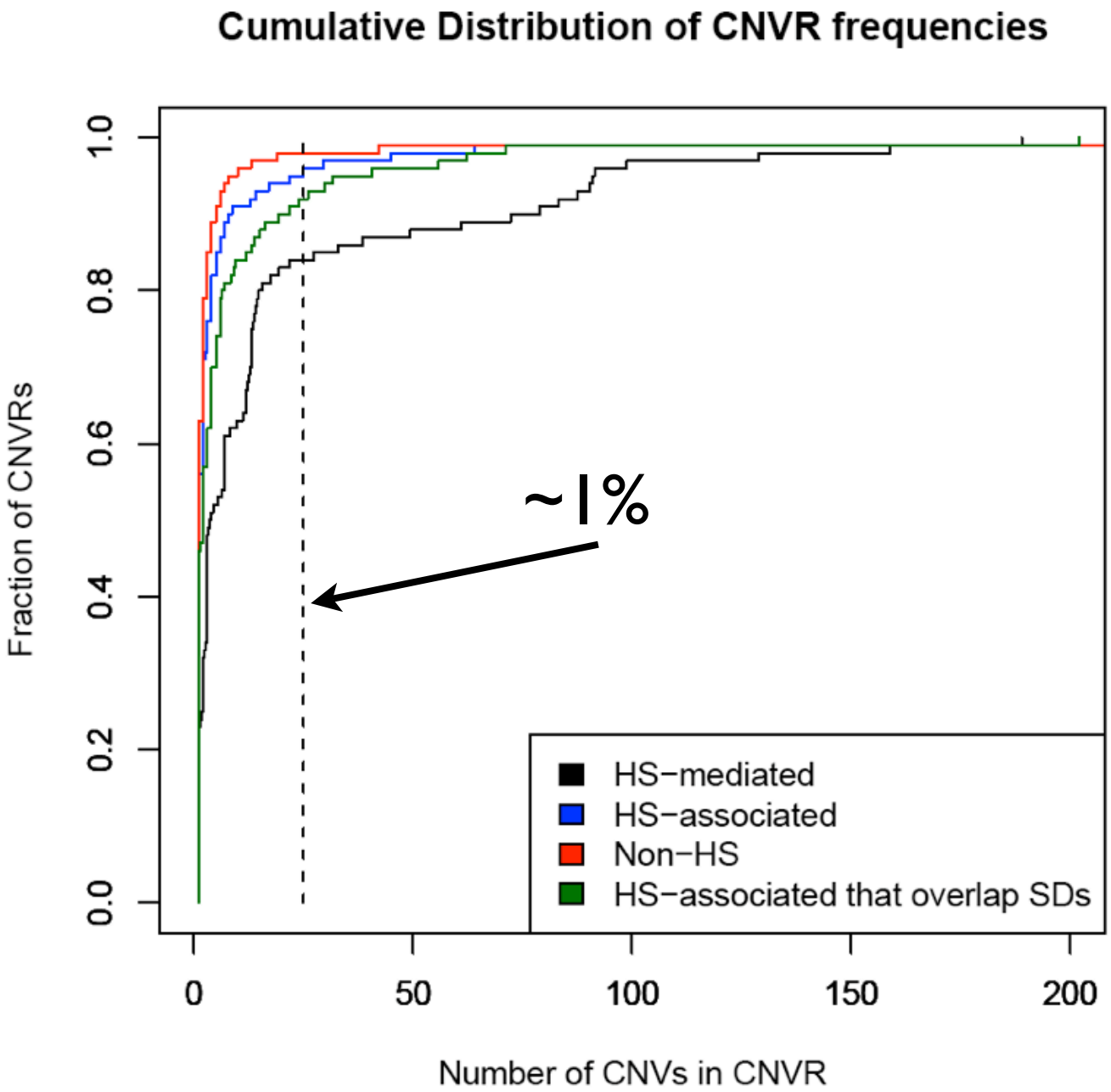
CNVs Enriched Near Segmental Duplications





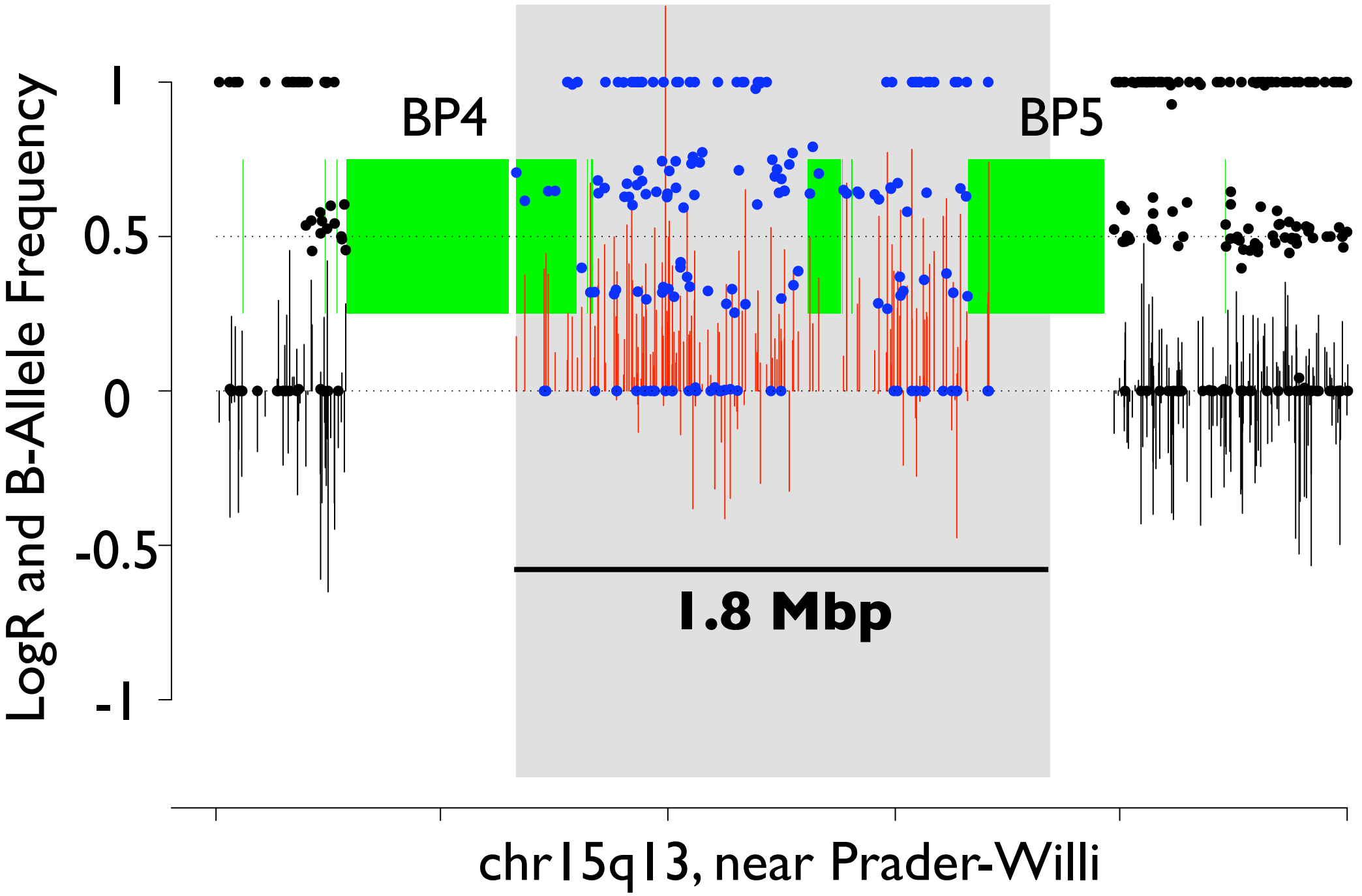
25X enrichment for CNVs between homologous duplications in the reference assembly

Hotspots Increase CNV Frequencies

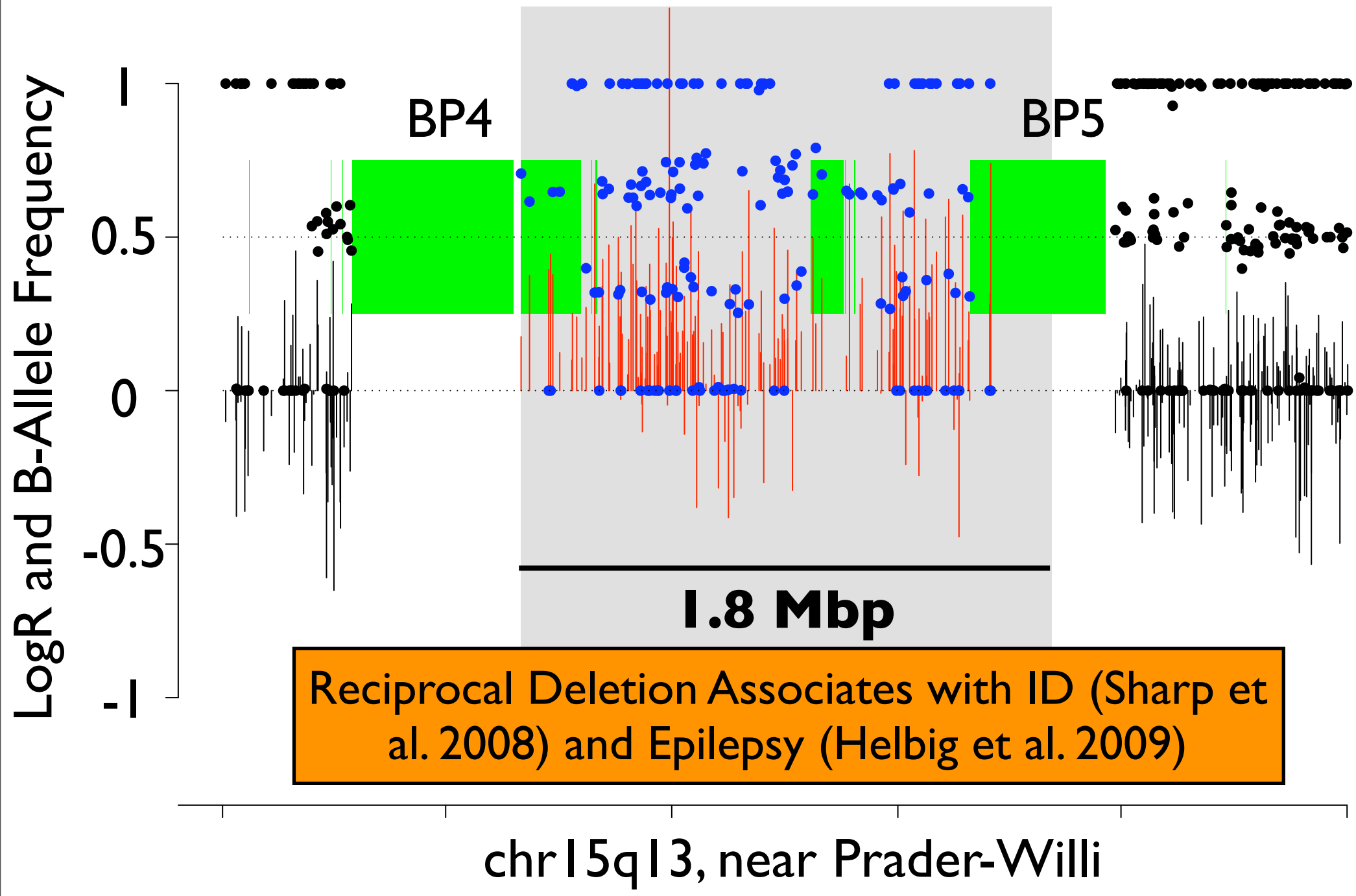


Duplications increase CNV allele frequency

Large ‘HotSpot’ Duplication



Large 'HotSpot' Duplication

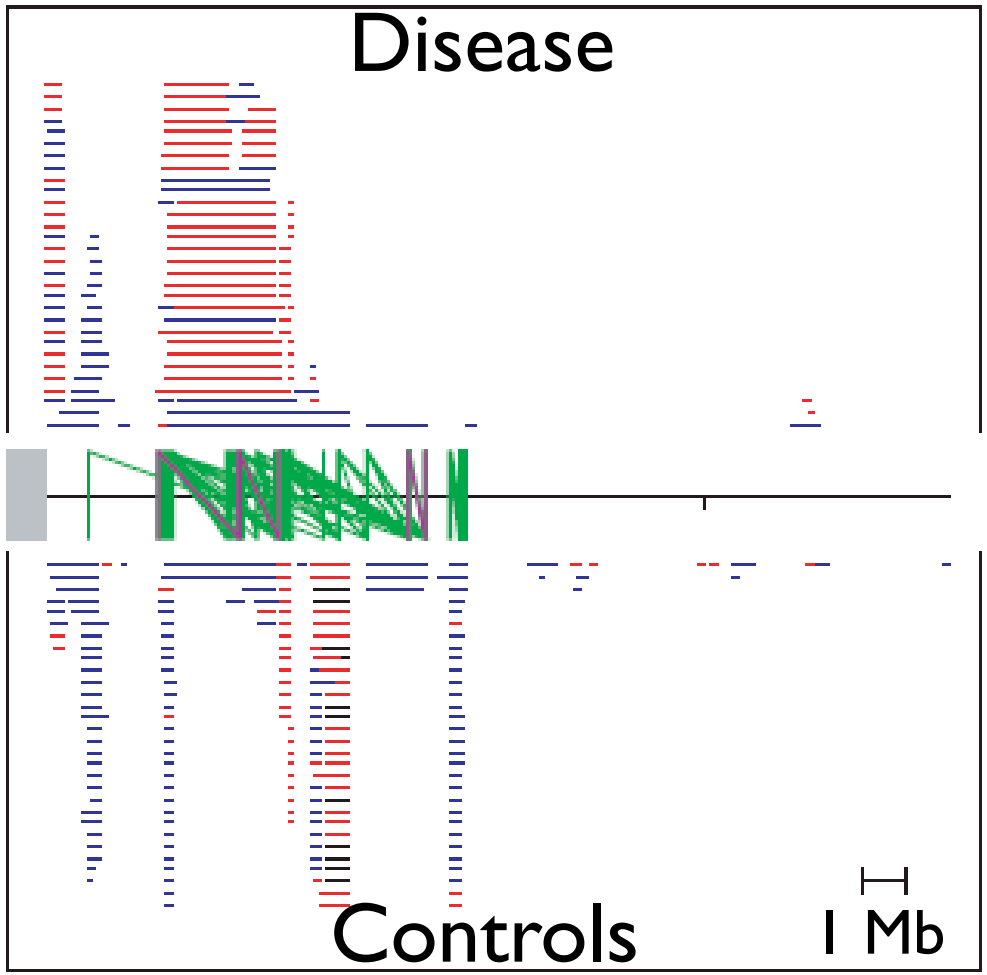


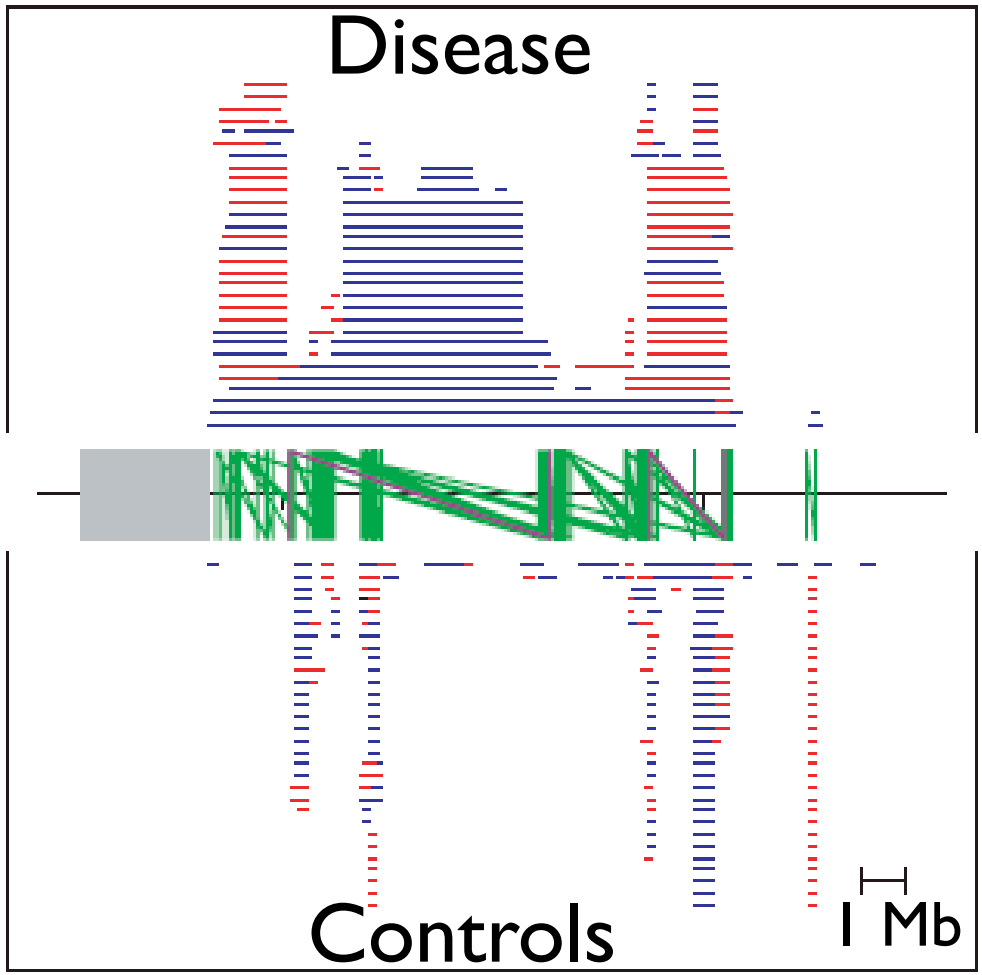
Reciprocal Deletion Associates with ID (Sharp et al. 2008) and Epilepsy (Helbig et al. 2009)

Neurological Disease Meta-Analysis

- Combined our ~2,500 samples with published CNV calls from ~3,000 controls analyzed using Affymetrix arrays (ISC *Nature* 2008)
- Combined genome-wide CNV annotations from 9 disease studies:
  - schizophrenia: ~3,500 individuals
  - autism: ~2,500 individuals
  - intellectual disability: ~500 individuals
  - mixture of Affy, Illumina, and CGH
  - only analyzed CNVs > 500 kb

Neurological Disease Meta-Analysis

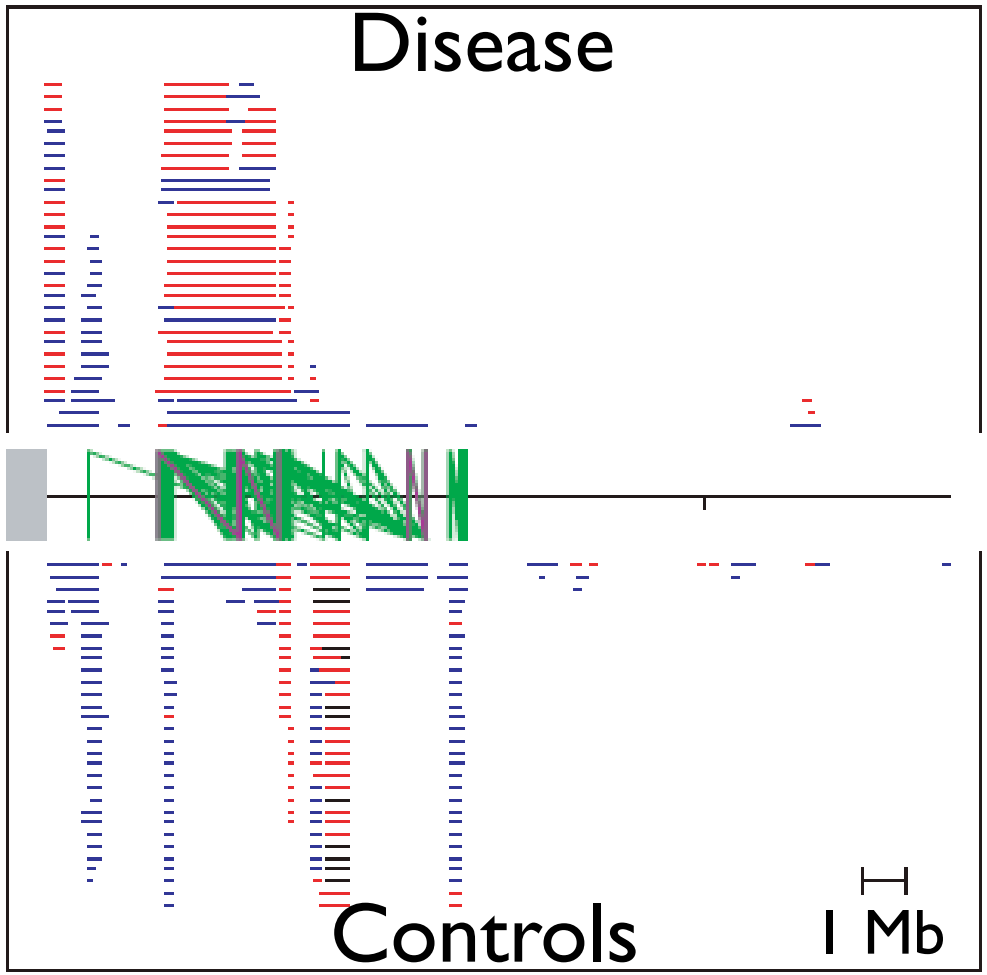




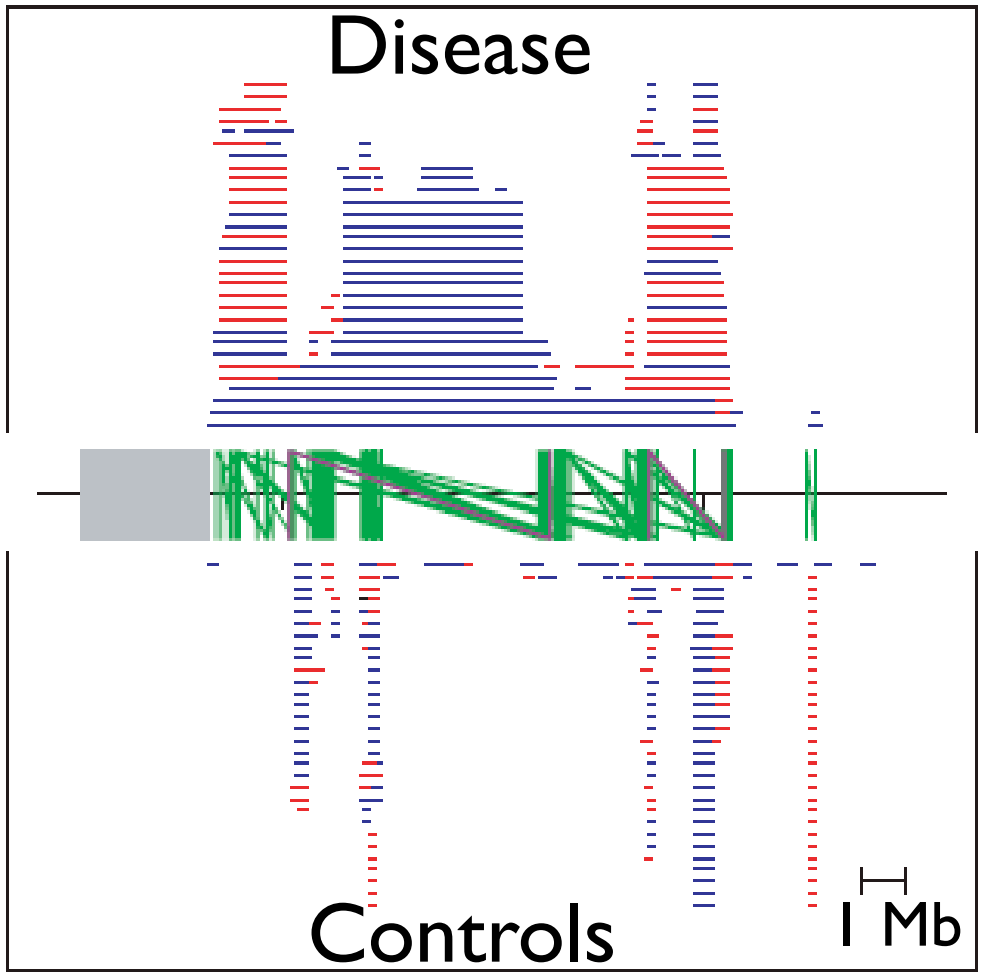
Deletions

Duplications

Neurological Disease Meta-Analysis



22q11-12 (VCFS)

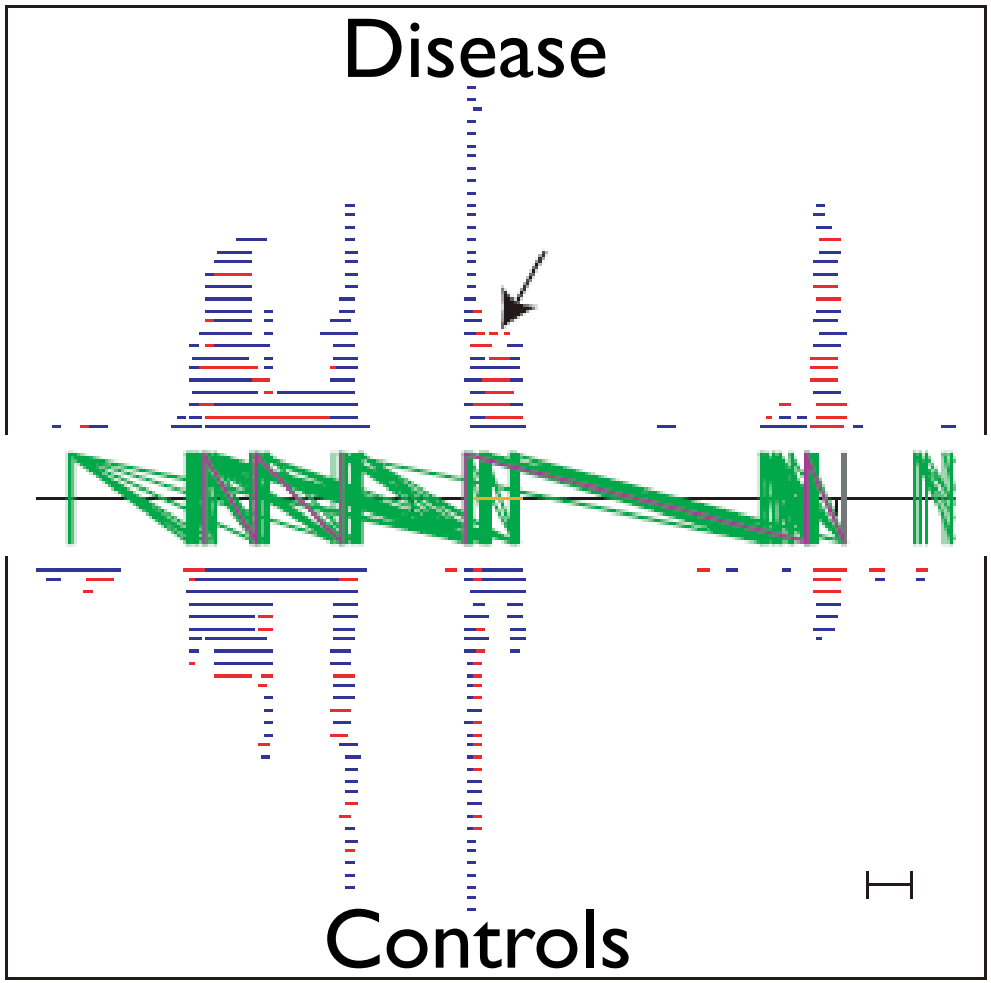


Prader Willi,15q13

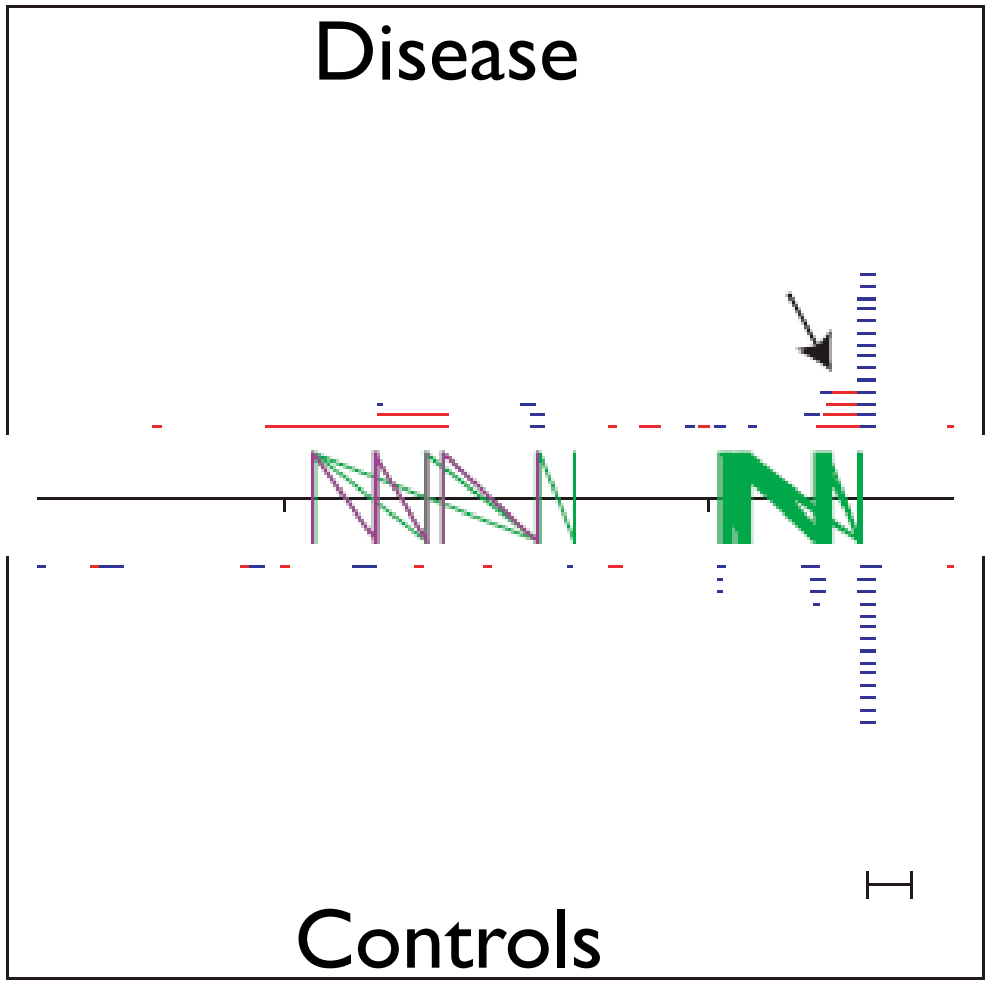
Deletions

Duplications

Neurological Disease Meta-Analysis



16p12 ?



15q25 ?

Deletions

Duplications

Neurological Disease CNVs

Chr	Start	Stop	Type	Note	NAHR?	Disease CNVs	Control CNVs	Diseases
chr22	17,014,900	19,993,127	loss	VCFS	yes	31	0	A,S,ID
chr15	27,015,263	30,650,000	loss	15q13	yes	19	0	S
chr1	142,540,000	146,059,433	loss	1q21.1	yes	24	3	A,S,C
chr15	18,376,200	30,756,771	gain	PW/15q13	yes	45	13	A,S,ID,C
chr1	142,800,580	146,009,436	gain	1q21.1	yes	12	3	A,S,ID,C
chr16	21,693,739	22,611,363	loss	16p12	yes	5	0	A,S
chr22	45,144,027	49,509,153	loss	Term 22	no	4	0	A
chr16	60,141,700	61,581,600	loss	16q21	no	4	0	A
chr15	82,573,421	83,631,697	loss	15q25	yes	4	0	A,S
chr16	80,737,839	82,208,451	gain	16q23.3	no	4	0	A,S
chr16	29,474,810	30,235,818	gain	16p11.2	yes	6	1	A,S,C
chr17	14,000,000	15,421,835	loss	HNPP	yes	6	1	A,S,C
chr22	47,572,875	48,323,417	gain	Term 22	no	5	1	A,S,C
chr11	78,120,000	85,610,000	loss	11q14	no	3	0	S,ID
chr2	184,270,000	186,892,000	gain	2q32	no	3	0	A
chr9	206456	1599250	gain	9p24	no	3	0	A,S
chr3	197,179,156	198,842,299	loss	3q29	yes	3	0	S
chr16	29,470,951	30,252,473	loss	16p11.2	yes	8	3	A,C
chr17	12,650,000	15,540,000	gain	CMT1A	yes	4	1	A,S,ID,C

Neurological Disease CNVs

Chr	Start	Stop	Type	Note	NAHR?	Disease CNVs	Control CNVs	Diseases
chr22	17,014,900	19,993,127	loss	VCFS	yes	31	0	A,S,ID
chr15	27,015,263	30,650,000	loss	15q13	yes	19	0	S
chr1	142,540,000	146,059,433	loss	1q21.1	yes	24	3	A,S,C
chr15	18,376,200	30,756,771	gain	PW/15q13	yes	45	13	A,S,ID,C
chr1	142,800,580	146,009,436	gain	1q21.1	yes	12	3	A,S,ID,C
chr16	21,693,739	22,611,363	loss	16p12	yes	5	0	A,S
chr22	45,144,027	49,509,153	loss	Term 22	no	4	0	A
chr16	60,141,700	61,581,600	loss	16q21	no	4	0	A
chr15	82,573,421	83,631,697	loss	15q25	yes	4	0	A,S
chr16	80,737,839	82,208,451	gain	16q23.3	no	4	0	A,S
chr16	29,474,810	30,235,818	gain	16p11.2	yes	6	1	A,S,C
chr17	14,000,000	15,421,835	loss	HNPP	yes	6	1	A,S,C
chr22	47,572,875	48,323,417	gain	Term 22	no	5	1	A,S,C
chr11	78,120,000	85,610,000	loss	11q14	no	3	0	S,ID
chr2	184,270,000	186,892,000	gain	2q32	no	3	0	A
chr9	206456	1599250	gain	9p24	no	3	0	A,S
chr3	197,179,156	198,842,299	loss	3q29	yes	3	0	S
chr16	29,470,951	30,252,473	loss	16p11.2	yes	8	3	A,C
chr17	12,650,000	15,540,000	gain	CMT1A	yes	4	1	A,S,ID,C

Neurological Disease CNVs

Chr	Start	Stop	Type	Note	NAHR?	Disease CNVs	Control CNVs	Diseases
chr22	17,014,900	19,993,127	loss	VCFS	yes	31	0	A,S,ID
chr15	27,015,263	30,650,000	loss	15q13	yes	19	0	S
chr1	142,540,000	146,059,433	loss	1q21.1	yes	24	3	A,S,C
chr15	18,376,200	30,756,771	gain	PW/15q13	yes	45	13	A,S,ID,C
chr1	142,800,580	146,009,436	gain	1q21.1	yes	12	3	A,S,ID,C
chr16	21,693,739	22,611,363	loss	16p12	yes	5	0	A,S
chr22	45,144,027	49,509,153	loss	Term 22	no	4	0	A
chr16	60,141,700	61,581,600	loss	16q21	no	4	0	A
chr15	82,573,421	83,631,697	loss	15q25	yes	4	0	A,S
chr16	80,737,839	82,208,451	gain	16q23.3	no	4	0	A,S
chr16	29,474,810	30,235,818	gain	16p11.2	yes	6	1	A,S,C
chr17	14,000,000	15,421,835	loss	HNPP	yes	6	1	A,S,C
chr22	47,572,875	48,323,417	gain	Term 22	no	5	1	A,S,C
chr11	78,120,000	85,610,000	loss	11q14	no	3	0	S,ID
chr2	184,270,000	186,892,000	gain	2q32	no	3	0	A
chr9	206456	1599250	gain	9p24	no	3	0	A,S
chr3	197,179,156	198,842,299	loss	3q29	yes	3	0	S
chr16	29,470,951	30,252,473	loss	16p11.2	yes	8	3	A,C
chr17	12,650,000	15,540,000	gain	CMT1A	yes	4	1	A,S,ID,C

Neurological Disease CNVs

Chr	Start	Stop	Type	Note	NAHR?	Disease CNVs	Control CNVs	Diseases
chr22	17,014,900	19,993,127	loss	VCFS	yes	31	0	A,S,ID
chr15	27,015,263	30,650,000	loss	15q13	yes	19	0	S
chr1	142,540,000	146,059,433	loss	1q21.1	yes	24	3	A,S,C
chr15	18,376,200	30,756,771	gain	PW/15q13	yes	45	13	A,S,ID,C
chr1	142,800,580	146,009,436	gain	1q21.1	yes	12	3	A,S,ID,C
chr16	21,693,739	22,611,363	loss	16p12	yes	5	0	A,S
chr3	197,179,156	198,842,299	loss	3q29	yes	3	0	S
chr16	29,470,951	30,252,473	loss	16p11.2	yes	8	3	A,C
chr17	12,650,000	15,540,000	gain	CMT1A	yes	4	1	A,S,ID,C

- 3q29 independently reported as an ID syndrome
- 16p12 deletion present in:
  - a schizophrenia/ID affected family (from Mary-Claire King)
  - 3 schizophrenic and 2 control samples (from Jonathan Sebat)
  - 12 out of ~10,000 children with various cognitive deficits (Lisa Shaffer and Signature Genomics)

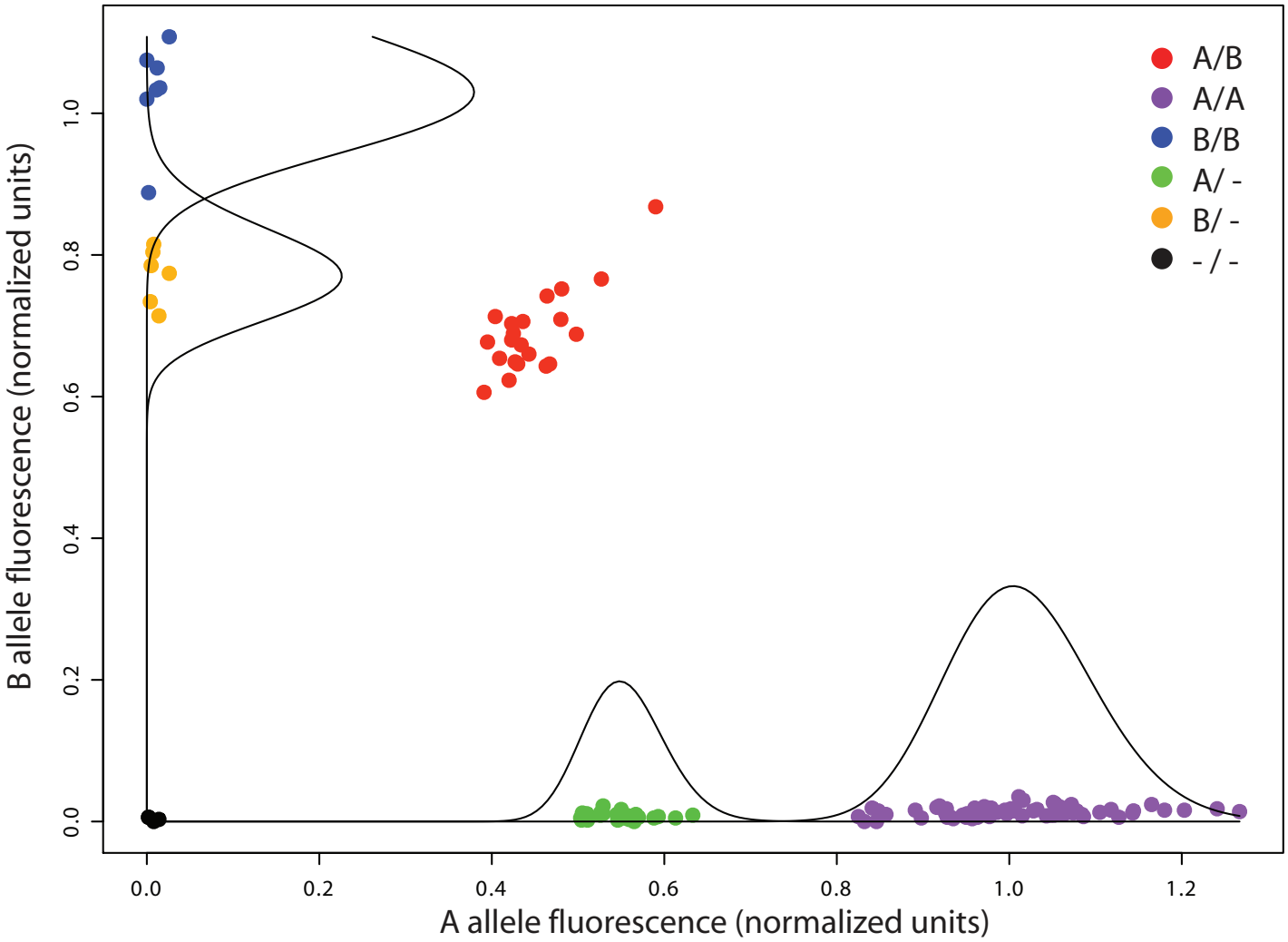
Large CNVs in Human Populations

- Large CNVs are collectively frequent (most individuals carry one or more large CNVs)
- Large CNVs are individually rare (usually <1% frequency)
- NAHR contributes to both common and rare variants:
  - many CNVs will not be 'taggable' via SNPs
  - ascertaining variation in/near duplications critical
- Highly penetrant CNVs, each explaining <1% of disease, may collectively make large disease contributions
  - Accurate genotyping of rare CNVs in large numbers will be required to identify these

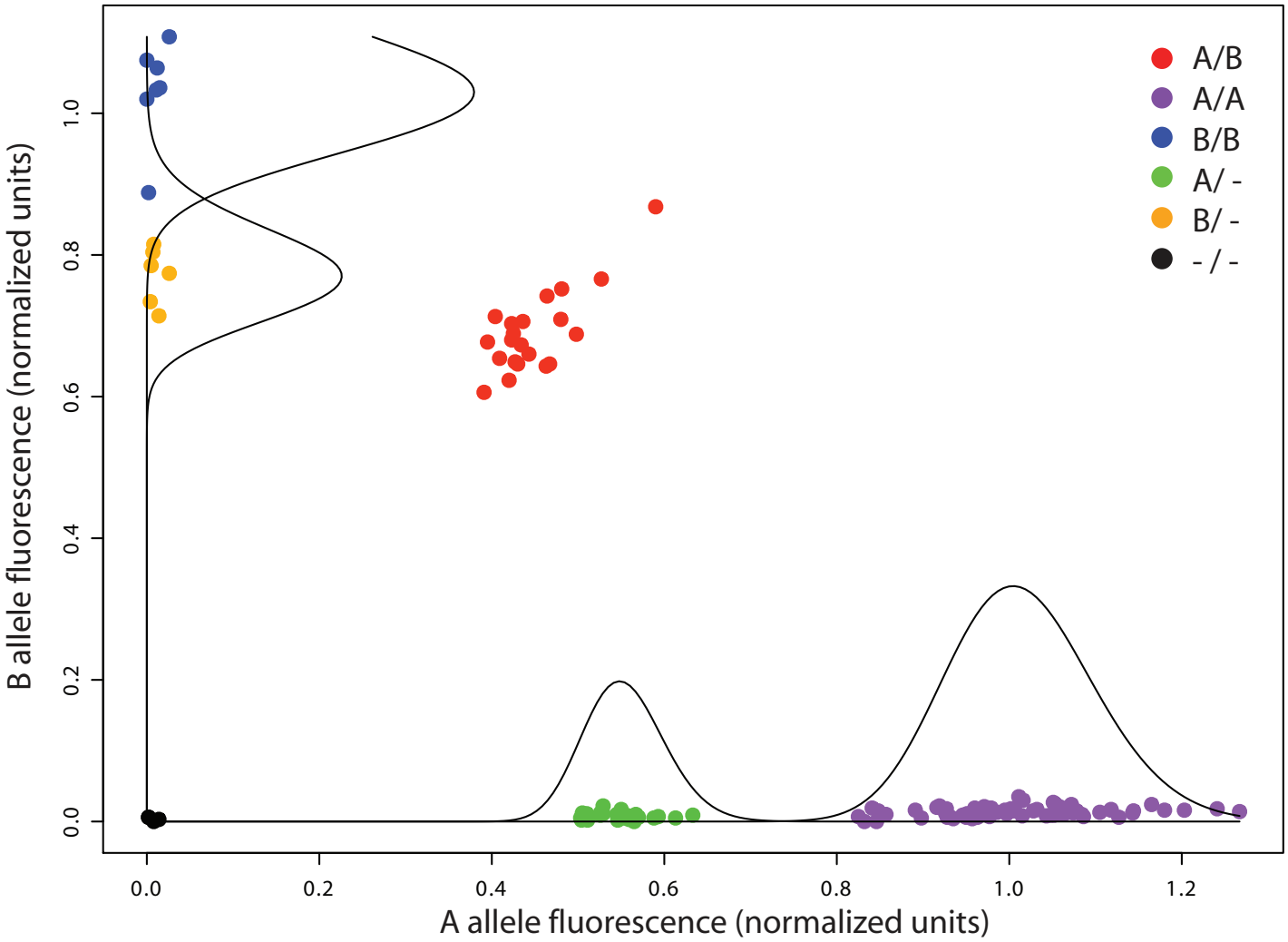
CNV Genotyping vs Discovery

- Discovery is done per-sample, genome-wide, and without assumptions about breakpoints
  - consequently, sensitivity is compromised to facilitate tolerable FDR
- Genotyping is targeted to known loci and applies to all samples simultaneously
  - good sensitivity **and** specificity are required
  - knowledge that a CNV is likely to exist and borrowing information across samples reduces the number of probes needed

SNP-based Common CNV Genotyping



SNP-based Common CNV Genotyping

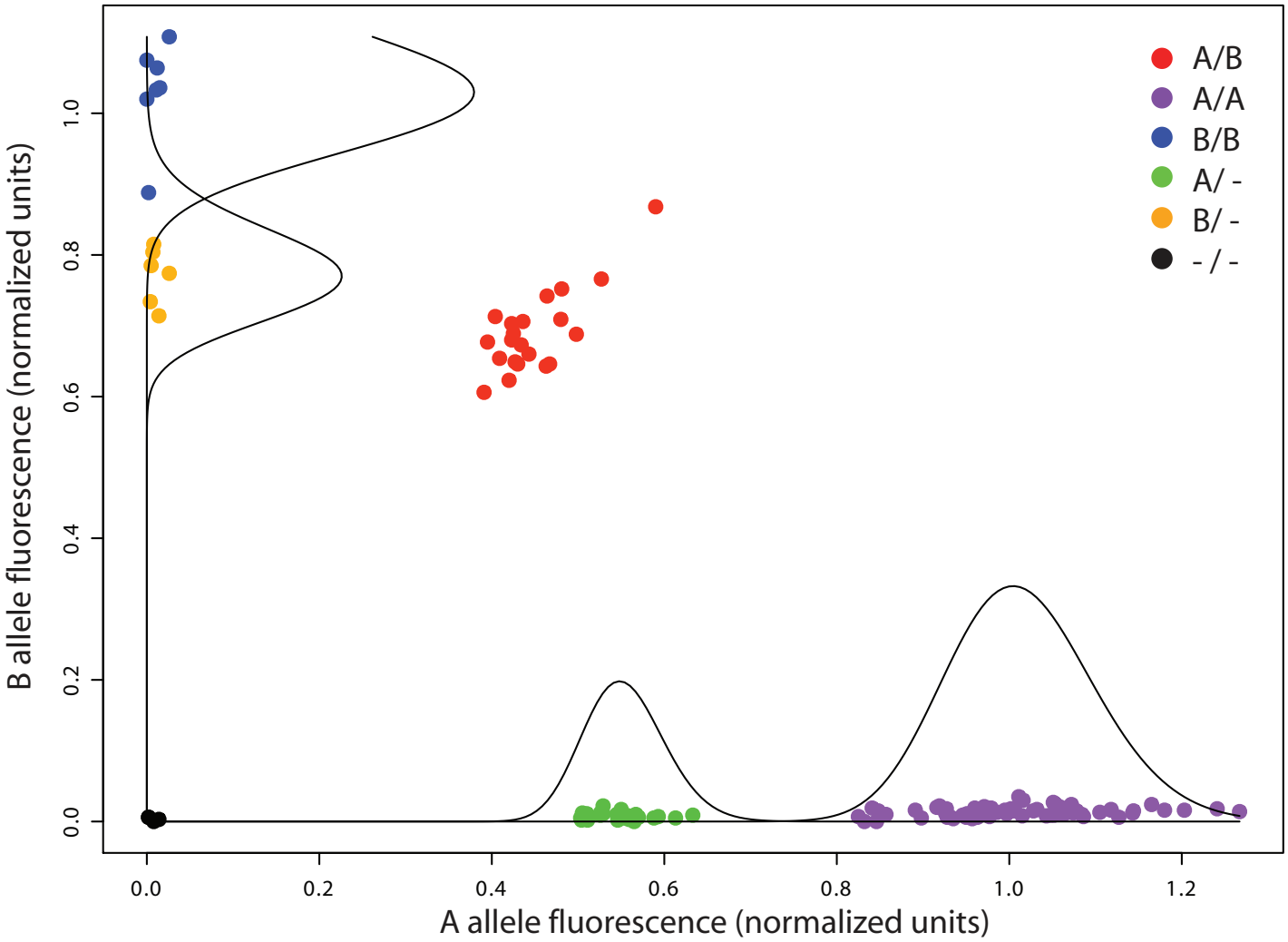


SNP-Conditional Mixture Modeling (SCIMM) for Deletion Genotyping

Input: SNP calls and quantitative data from multiple probes

Uses the EM algorithm to generate putative copy number calls (0, 1, 2) for each sample and a score for the probe set

SNP-based Common CNV Genotyping



SNP-Conditional Mixture Modeling (SCIMM) for Deletion Genotyping

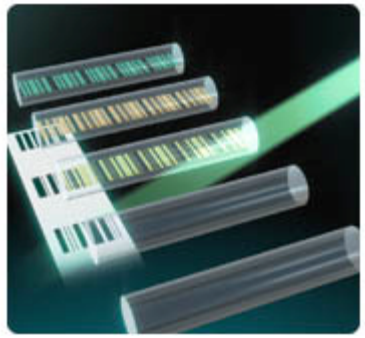
Input: SNP calls and quantitative data from multiple probes

Uses the EM algorithm to generate putative copy number calls (0, 1, 2) for each sample and a score for the probe set

SCIMM used to genotype hundreds of common deletions, replicate concordance and Mendelian consistency > 99%

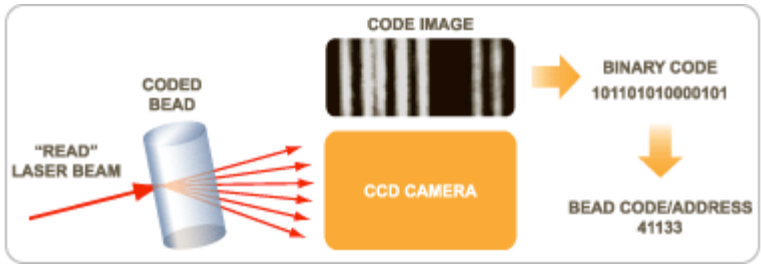
Scaling Up CNV Genotyping

- Identification of pathogenic variants seen in 0.1% - 1% of affected people requires sample sizes in excess of 10,000



- Genome-wide platforms are better for coverage, but expensive

- NAHR hotspots define breakpoints and enrich for copy-number variation



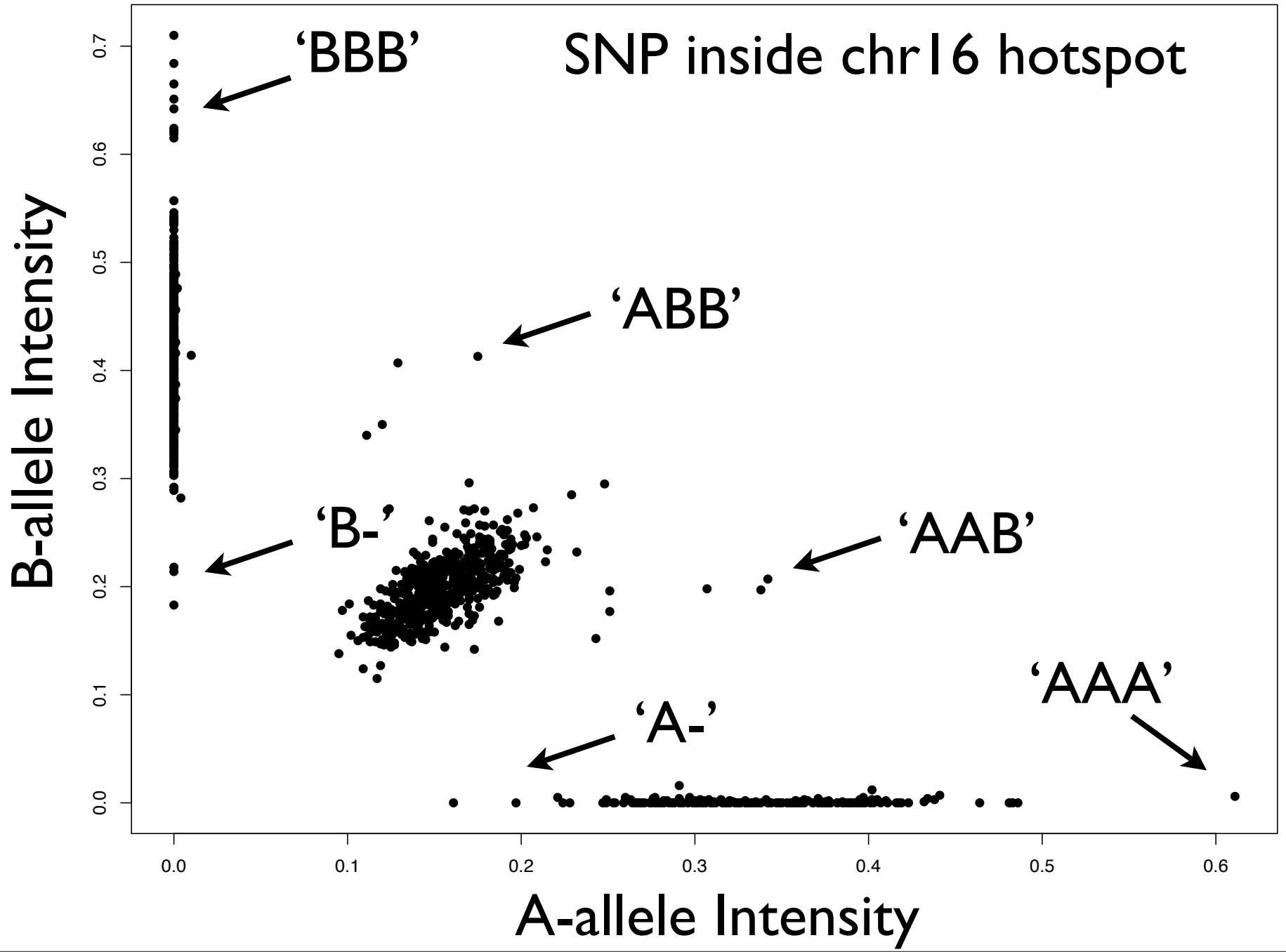
- We developed a custom 'BeadXpress' assay to genotype rare CNVs in children affected by idiopathic intellectual disability

BeadXpress CNV Detection

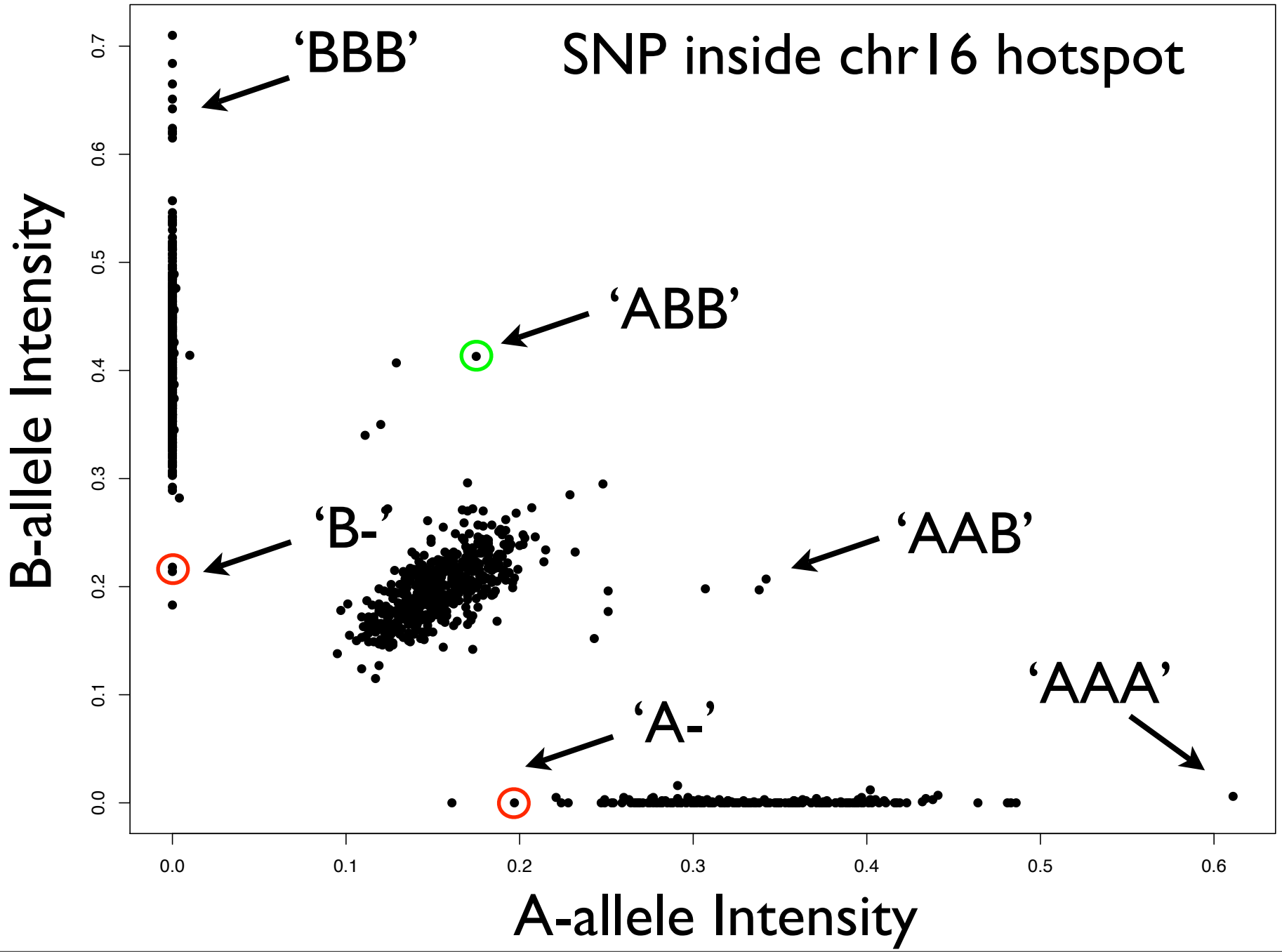
- 384-plex Illumina 'GoldenGate' SNP genotyping (PCR-based) assay performed in 96-well plates
- 69 known and putative disease hotspots targeted, ~5 probes each
- 2 common CNVs and 1 X-linked site
- Analyzed 1,105 affected children and 39 control samples
- Follow-up validation with targeted array-CGH

Snp-Conditional OUTlier (SCOUT) Detection

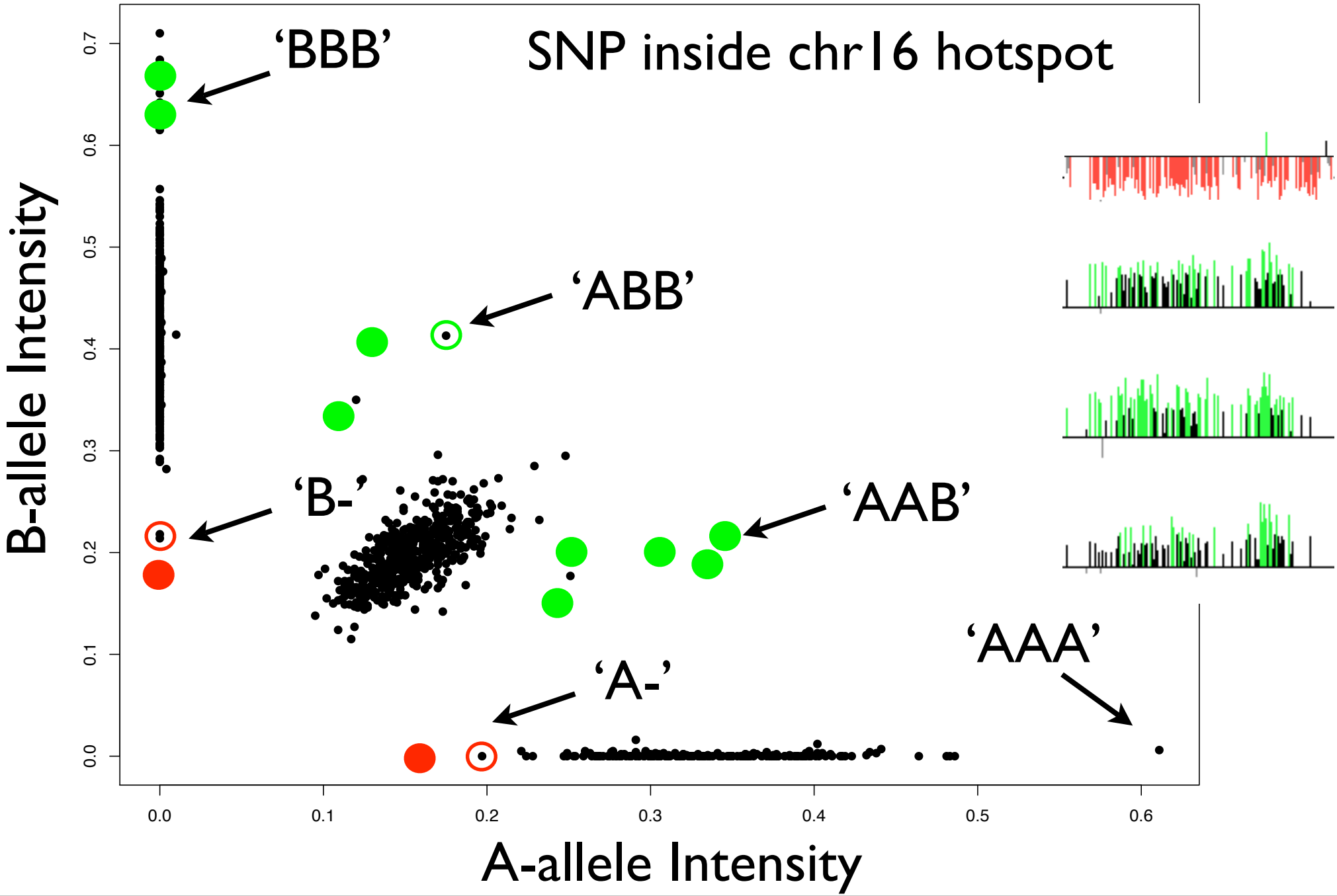
Snp-Conditional OUTlier (SCOUT) Detection



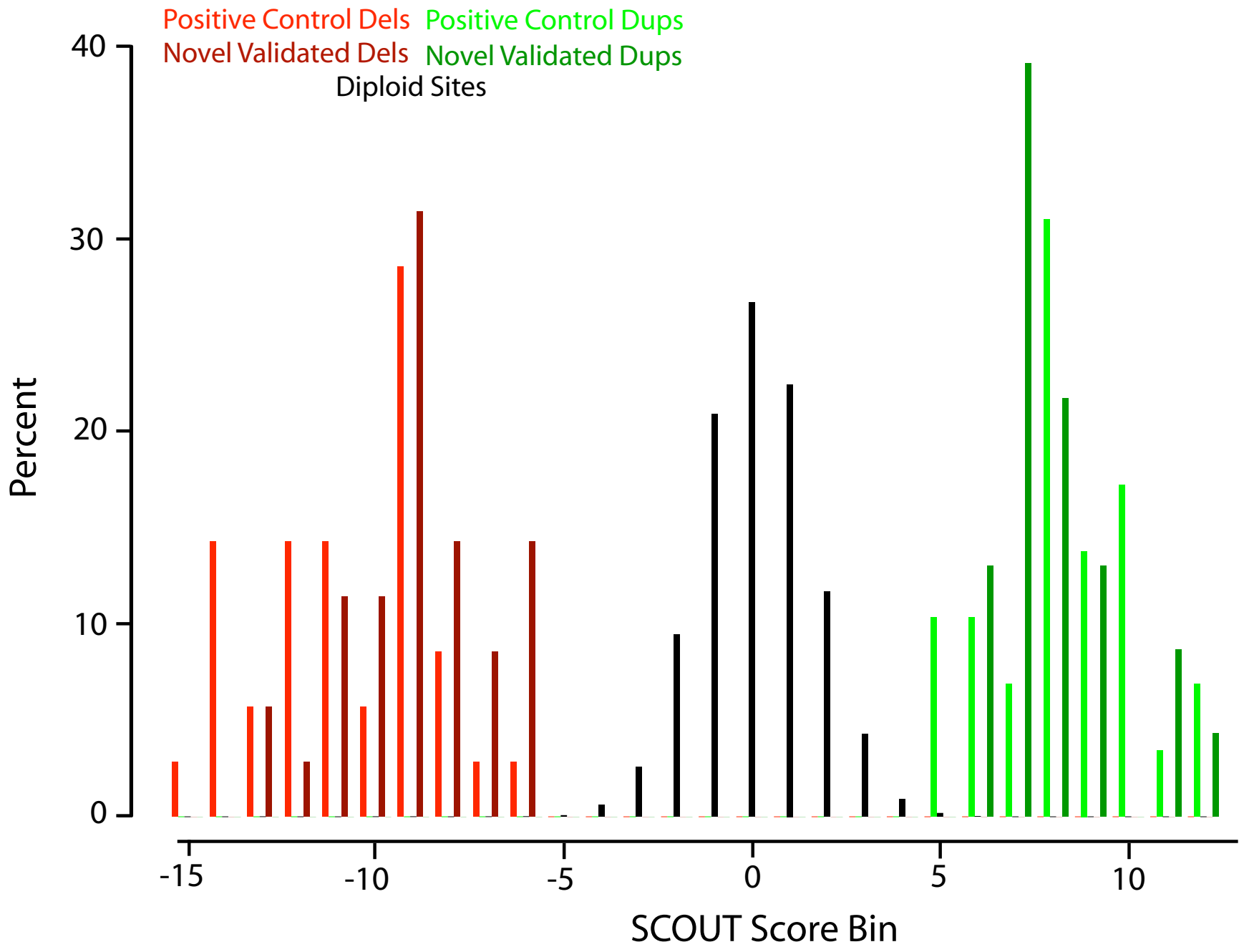
Snp-Conditional OUTlier (SCOUT) Detection



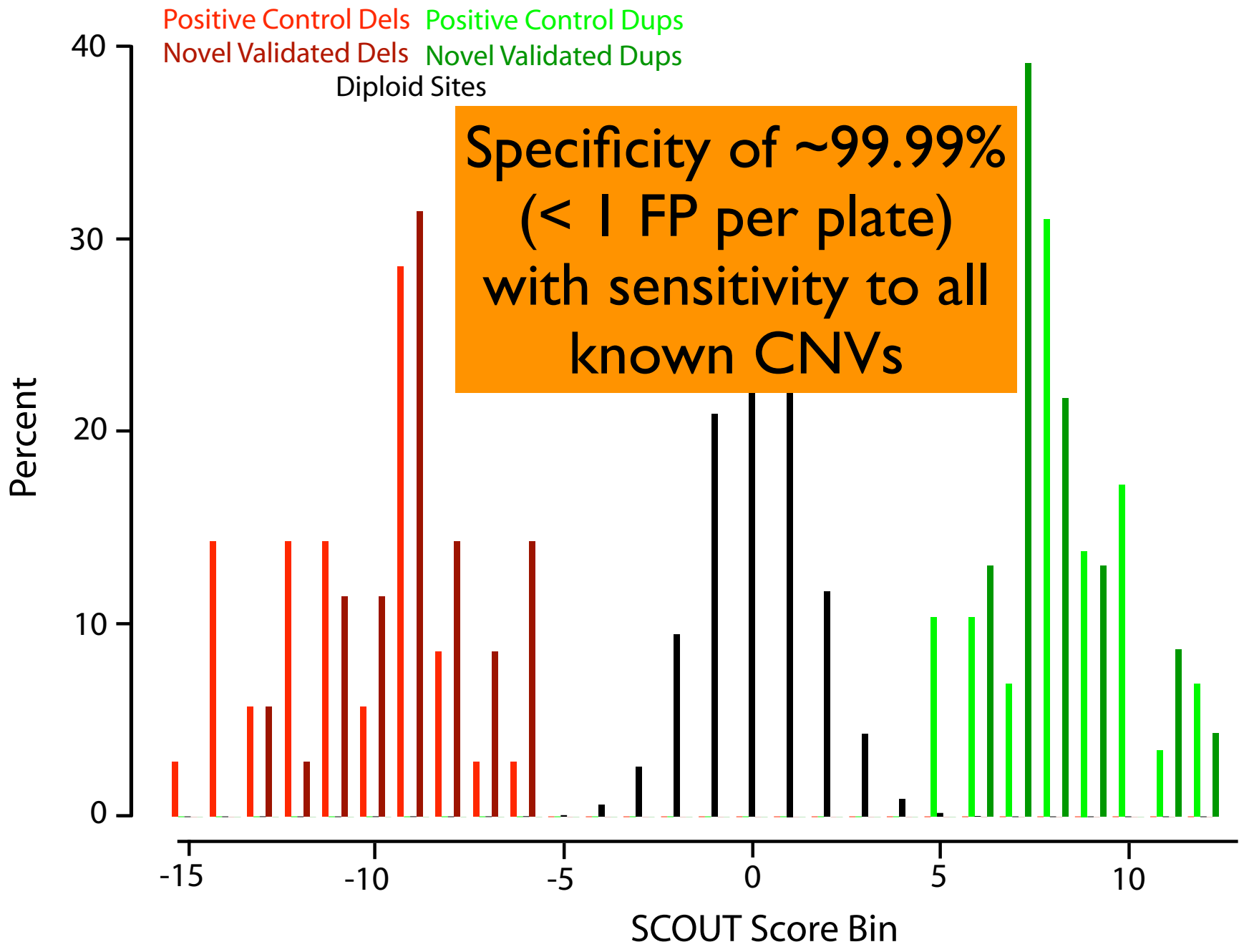
Snp-Conditional OUTlier (SCOUT) Detection



Accurate Detection of Rare Variants



Accurate Detection of Rare Variants



Disease-Relevant CNVs

Locus	Size (MB)	Confirmed by array-CGH	Predicted > |6|	Combined frequency (n = 1,010)
1q21.1 dup	1.4	1	2	0.30%
15q11 BP1-BP3 del	5.7	1	0	0.10%
15q11 BP2-BP3 del	5.3	1	0	0.10%
15q24 del	2.2	0	1	0.10%
16p11.2 del	0.9	6	1	0.69%
16p11.2 dup	0.9	1	0	0.10%
16p12.2-p11.2 del	8.1	0	1	0.10%
16p13 BP1-BP2 del	1.2	1	1	0.20%
16p13 BP1-BP3 del	3.2	1	0	0.10%
17p11.2 del (SMS)	3.5	2	0	0.20%
17p12 del (HNPP)	1.4	1	0	0.10%
17p12 dup (CMT1A)	1.4	0	1	0.10%
17q12 del	1.6	1	0	0.10%
22q11.21 3-Mb del (VCFS)	2.9	2	1	0.30%
22q11.21 3-Mb dup	2.9	2	0	0.20%
22q11.21-q11.22 distal del	1.4	2	1	0.30%
TOTAL		22	9	3.08%

Disease-Relevant CNVs

Locus	Size (MB)	Confirmed by array-CGH	Predicted > |6|	Combined frequency (n = 1,010)
1q21.1 dup	1.4	1	2	0.30%
15q11 BP1-BP3 del	5.7	1	0	0.10%
15q11 BP2-BP3 del	5.3	1	0	0.10%
15q24 del	2.2	0	1	0.10%
16p11.2 del	0.9	6	1	0.69%

16p11.2 previously associated with autism, but most of the individuals here have intellectual disability without autism

17p11.2 del (SMS)	3.5	2	0	0.20%
17p12 del (HNPP)	1.4	1	0	0.10%
17p12 dup (CMT1A)	1.4	0	1	0.10%
17q12 del	1.6	1	0	0.10%
22q11.21 3-Mb del (VCFS)	2.9	2	1	0.30%
22q11.21 3-Mb dup	2.9	2	0	0.20%
22q11.21-q11.22 distal del	1.4	2	1	0.30%
TOTAL		22	9	3.08%

CNVs of Uncertain Pathogenicity

- Duplications at 16p13 seen in 11 (1.1%) of samples
  - only seen in 2/2,493 controls ($p = 4.7 \times 10^{-5}$)
  - deletion is known to be pathogenic
  - duplication also enriched in schizophrenia
- Deletions at 15q11.2 seen in 8 (0.8%) of samples
  - only seen in 3/2,493 controls ($p = 0.003$)
  - near Prader-Willi critical region
  - also enriched in schizophrenia

BeadXpress Pilot Results

- High-confidence diagnosis of 31 (3.1%) children carrying pathogenic variants (e.g. 1q21.1, 16p11.2, VCFS, SMS, etc)
  - 16p11.2 associates with intellectual disability and is not an autism allele *per se*
- Evidence for disease relevance of 16p13 dups and 15q11.2 deletions collectively seen in 19 (1.9%) affected children
- 1,105 samples processed in < 1 month with costs 5-20X less than CGH or genome-wide SNP platforms
- New assay being designed for a larger batch of samples (5,000 - 10,000) and with probe selection optimizations

General Conclusions

- Copy-number variation is an influential source of genomic variation in human populations
- Duplication architecture is a major contributor to both common and rare variation via NAHR
- Rare variants are contributors to substantial amounts of disease heritability
- Tools are now available to dramatically scale the the size and accuracy of CNV-disease association studies

Acknowledgments

University of Washington

Evan Eichler and Debbie Nickerson

Andy Itsara, Troy Zerr, Heather Mefford, Jeff Kidd, Josh Smith, Mark Rieder

Eichler and Nickerson labs

Illumina Human 1M HapMap Data: Dan Peiffer (Illumina)

PARC Project: Ron Krauss (CHORI), Jerry Rotter (Cedar-Sinai)

Neurological Disease Controls: Andy Singleton (NINDS)

HGDP Samples: Devin Absher, Jun Li, and Rick Myers (HudsonAlpha)

Funding: NHGRI/NHLBI; Merck, Jane Coffin Childs Foundation