National Cancer Institute

Technologic Issues in GWAS and Follow-up Studies

Stephen Chanock, M.D.
Senior Investigator, POB,CCR &
Director, Core Genotyping Facility, DCEG NCI

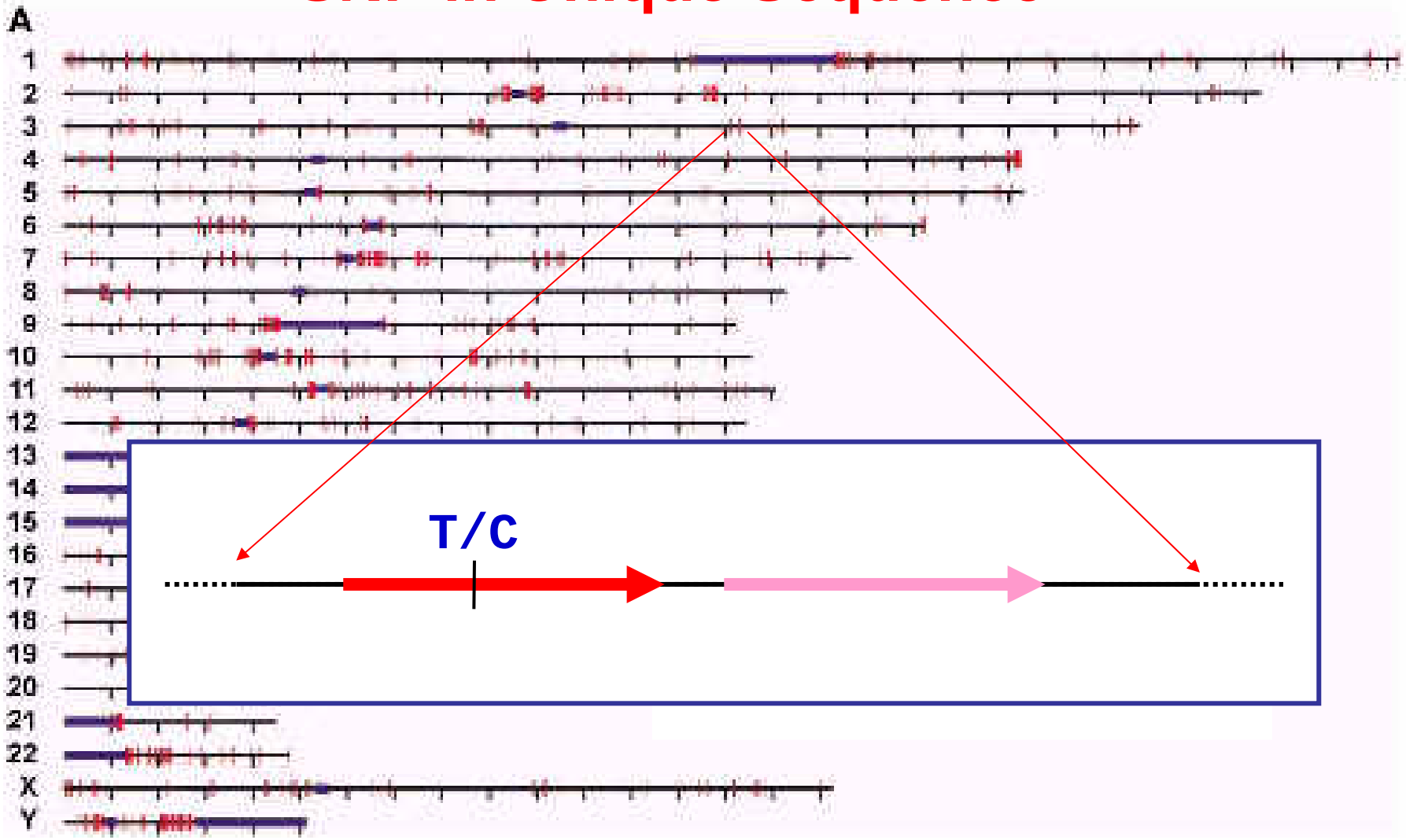
May 22, 2007

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
National Institutes of Health

Types of Polymorphisms

- Single nucleotide polymorphisms (SNPs)
  - Common SNPs are defined as >1% in at least one population
  - Rare SNPs are hard to identify and validate
    - *But*, it is estimated that there are a large number per individual
  - MAF= minor allele frequency

SNP in Unique Sequence



SNPs & Function: We know so little..

- Majority are “silent”
  - No known functional change
- Alter gene expression/regulation
  - Promoter/enhancer/silencer
  - mRNA stability
  - Small RNAs
- Alter function of gene product
  - Change sequence of protein

SNPs in Genes: Take one

Coding SNPs

Synonymous- no change in amino acid

previously termed “silent” but…..

Can alter mRNA stability

DRD2 (Duan et al 2002)

Nonsynonymous- changes amino acid

conservative and radical

Nonsense- insertion of stop codon

Indel- Disrupts codon sequence

Rare but disruptive

SNPs Outside Genes: Take many….

- Majority distributed throughout genome are “silent” (excellent as markers)
- Alter transcription
  - Promoter, enhancer, silencer
- Regulate expression
  - Locus control region, mRNA stability
- Most are assumed to be ‘silent hitchhikers’
  - No function by predictive models or analysis

Linkage disequilibrium (LD)

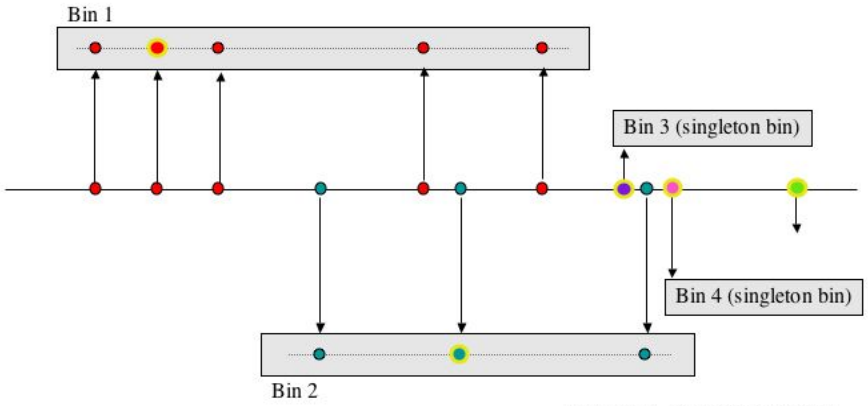
- The non-random association of alleles in the population
- Alleles at neighboring loci tend to cosegregate
- Linkage disequilibrium implies population allelic association

Strategy for SNP Selection

CGEMS
Cancer Genetic Markers of Susceptibility

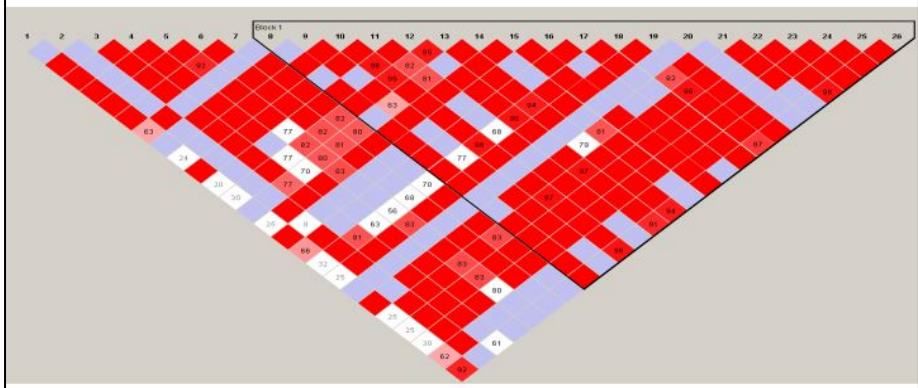
To test all SNPs is presently too costly
Utilize a strategy that capitalizes on linkage disequilibrium between SNPs

Grouping of SNPs into bins based on pairwise r^2.



Carlson et al. AJHG 74:106 (2004)

Haplotype blocks defined by Gabriel et al
Based on D' values for linkage disequilibrium

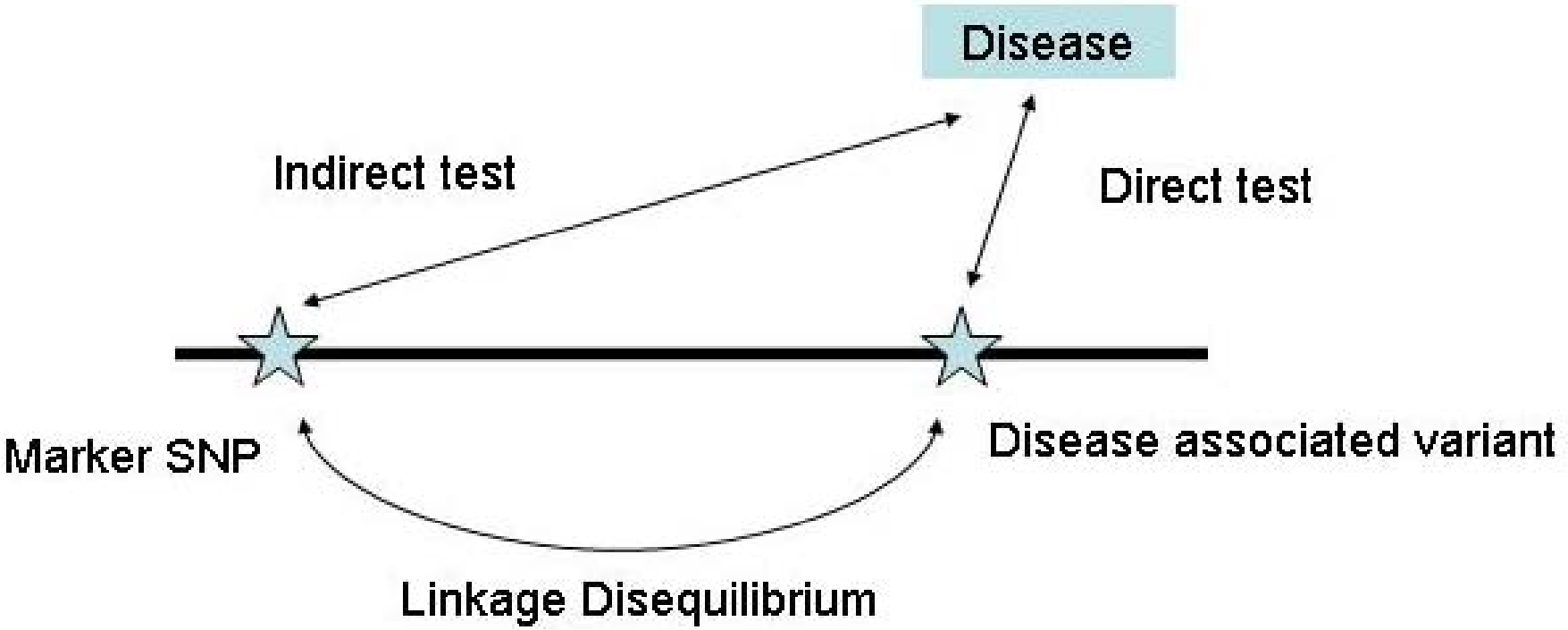


What can LD do for me?

- Knowledge of patterns of linkage disequilibrium can be quite useful in the design and analysis of genetic data
- Design:
  - Estimation of theoretical power to detect associations
  - Evaluation of degree completeness of sampling of genetic variants
  - Choice of most informative genetic variants to genotype

Genetic Association Testing: Finding Markers

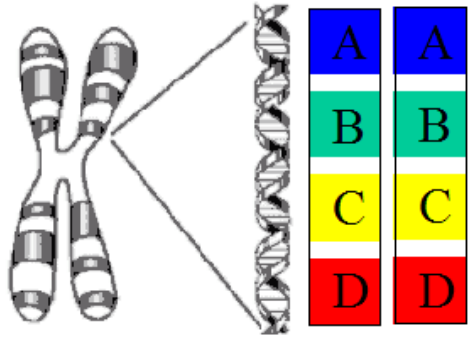
CGEMS
Cancer Genetic Markers of Susceptibility



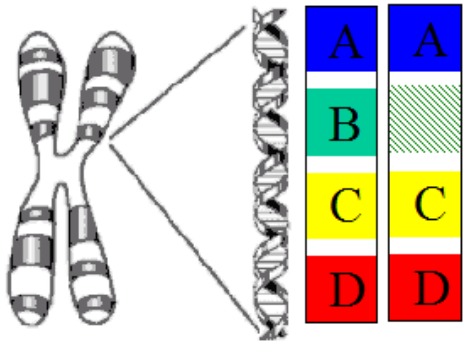
Large and Small Scale Polymorphisms

- **Copy Number of Polymorphisms**
  - Regional “repeat” of sequence
    - 10s to 100s kb of sequence
  - Estimate of >10% of human genome
  - Multi-copy in many individuals
- **Duplicons**
  - 90-100% similarity for >1 kb
  - 5-10% of genome (5% exons elsewhere)
  - Multi-copy (high N) in all individuals

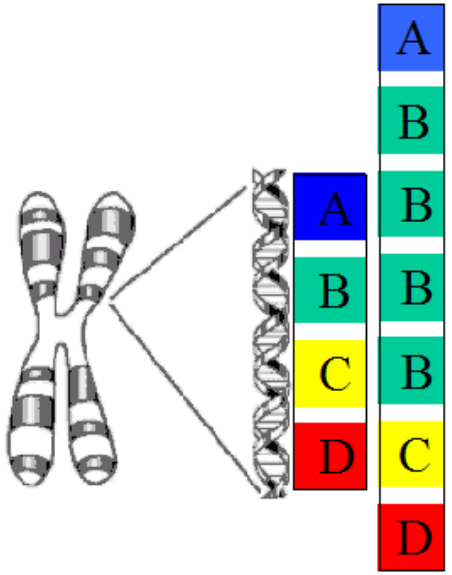
Germ-line DNA Copy Number Variation(CNV)



Diploid (n=2)

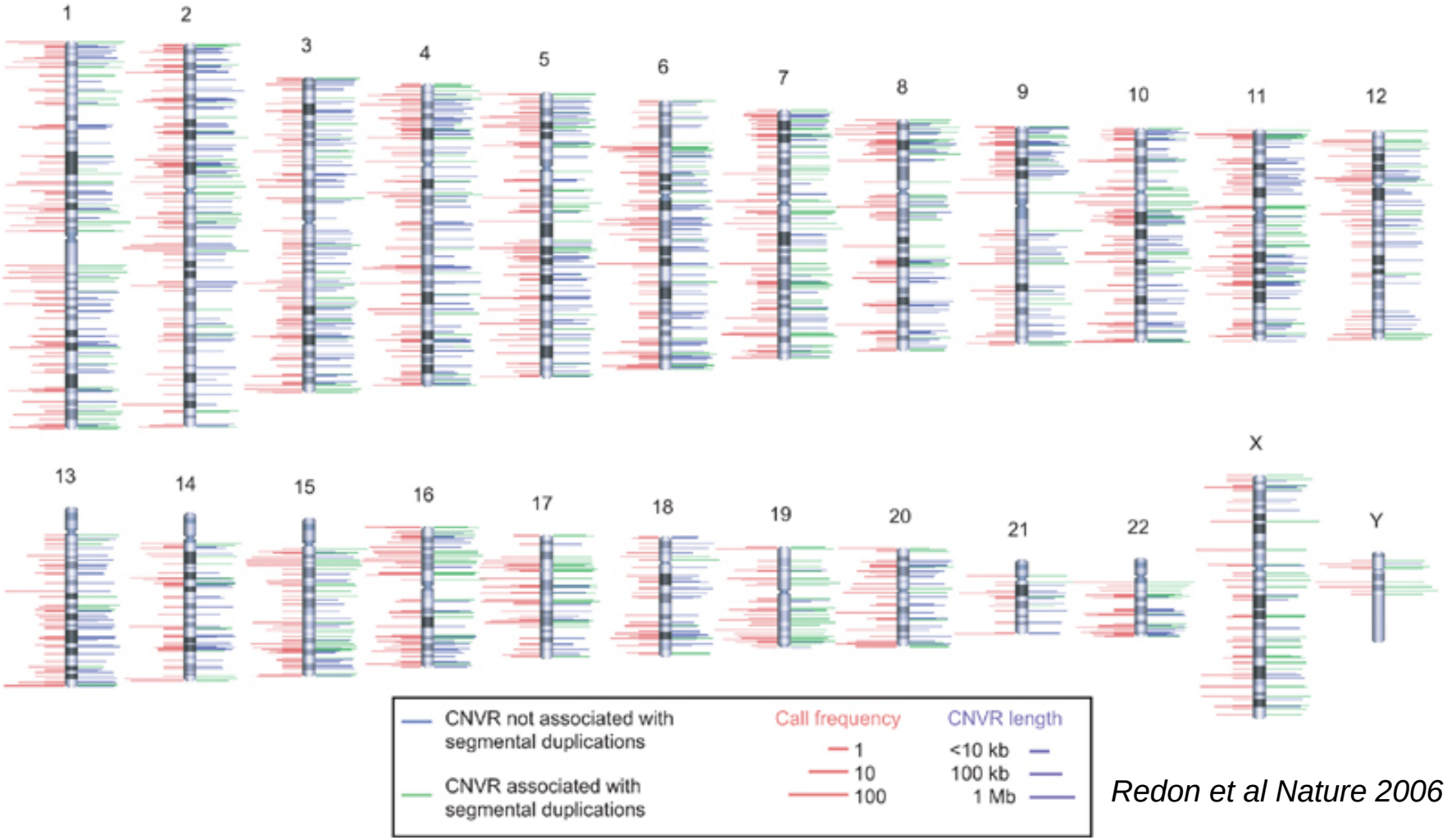


Deletion (n=1)



Gain (n=5)

Copy Number Variation Across the Genome



Redon et al Nature 2006

Copy Number Variation

Across the Genome

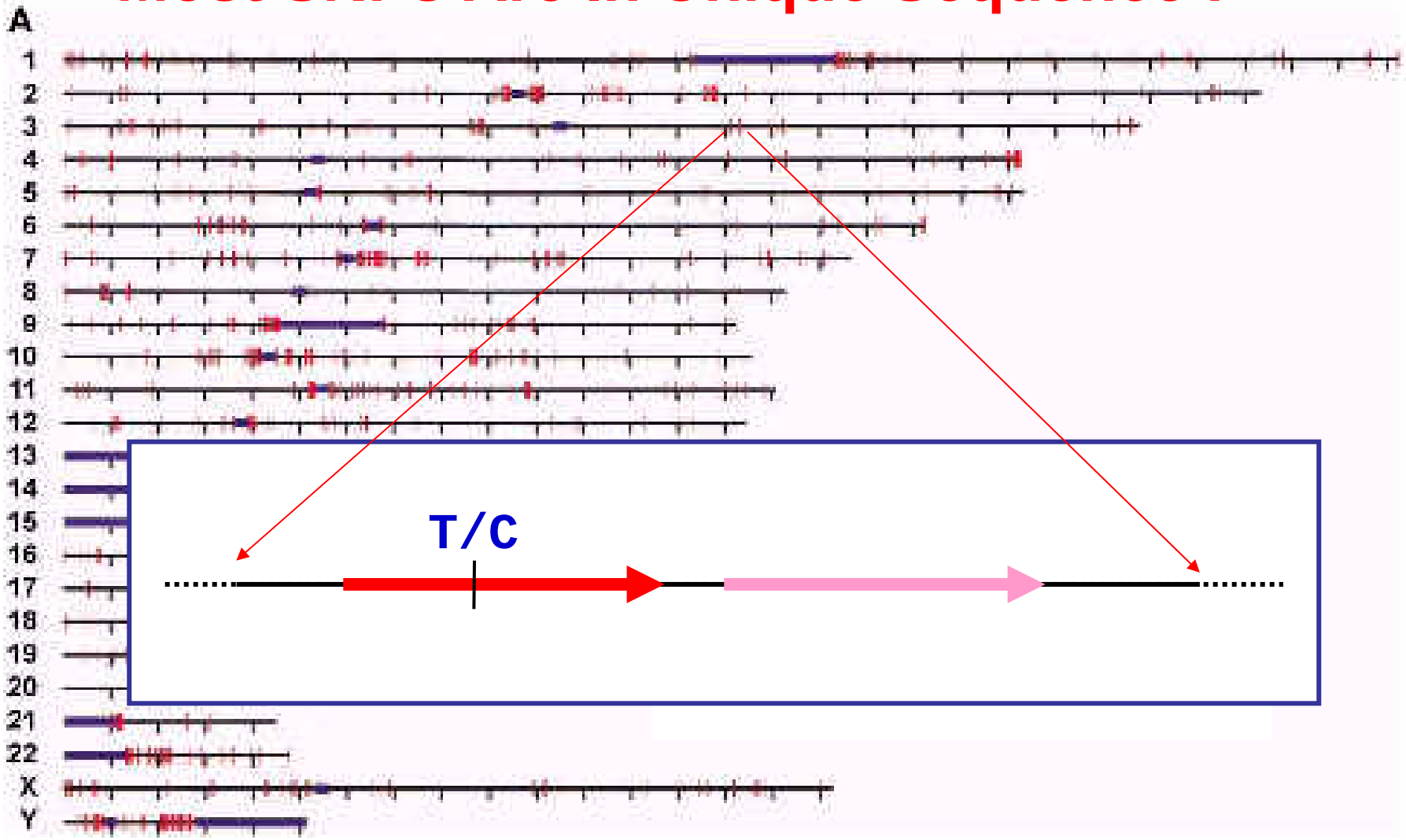
Frequency of CNVs

- Most are uncommon (<5%)
- Familial vs Unrelated Studies

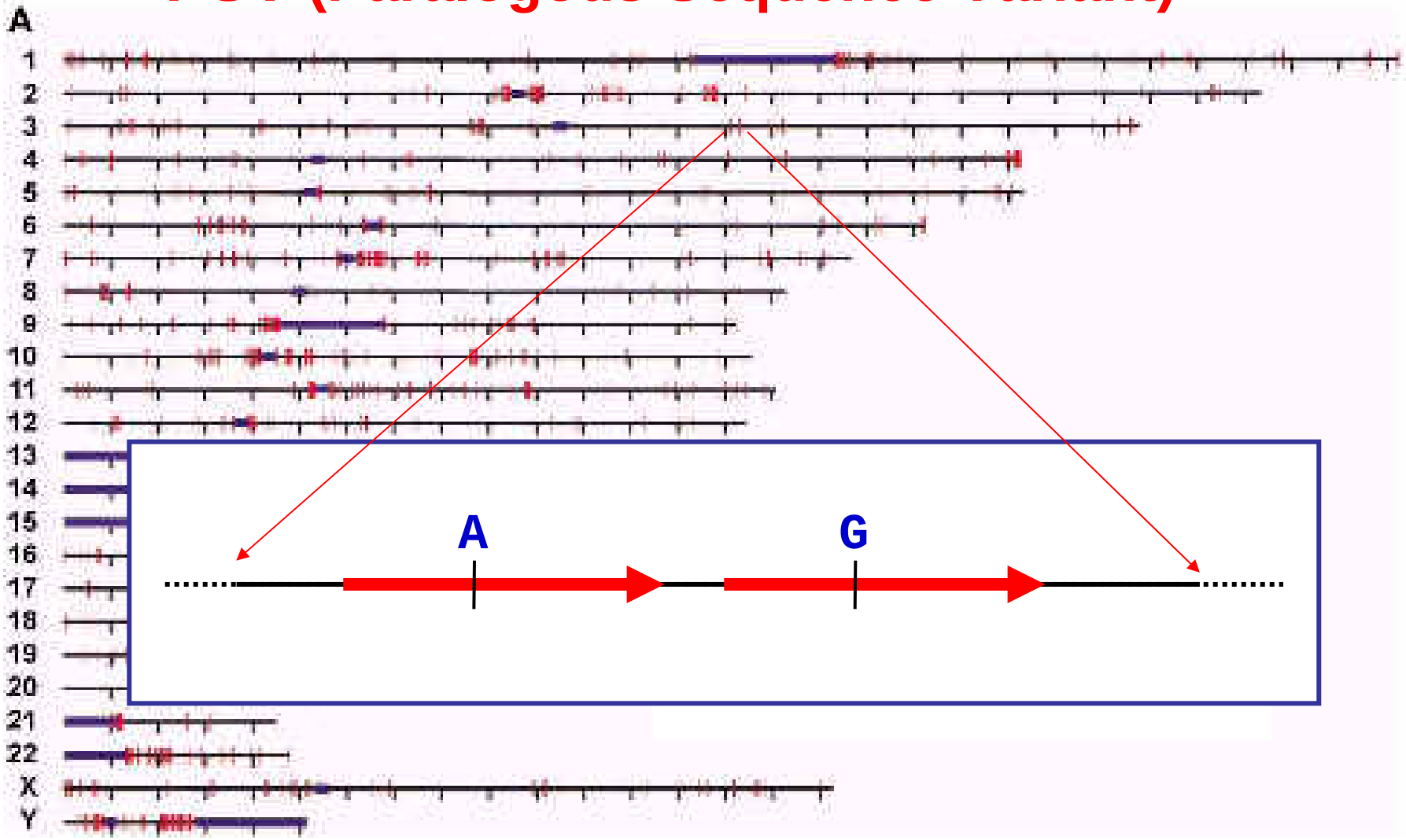
Associated with Disease

OPN1LW	Red/green colorblind
CCL3L1	Reduced HIV Infection
CYP2A6	Altered nicotine metabolism
VKORC1	Warfarin metabolism

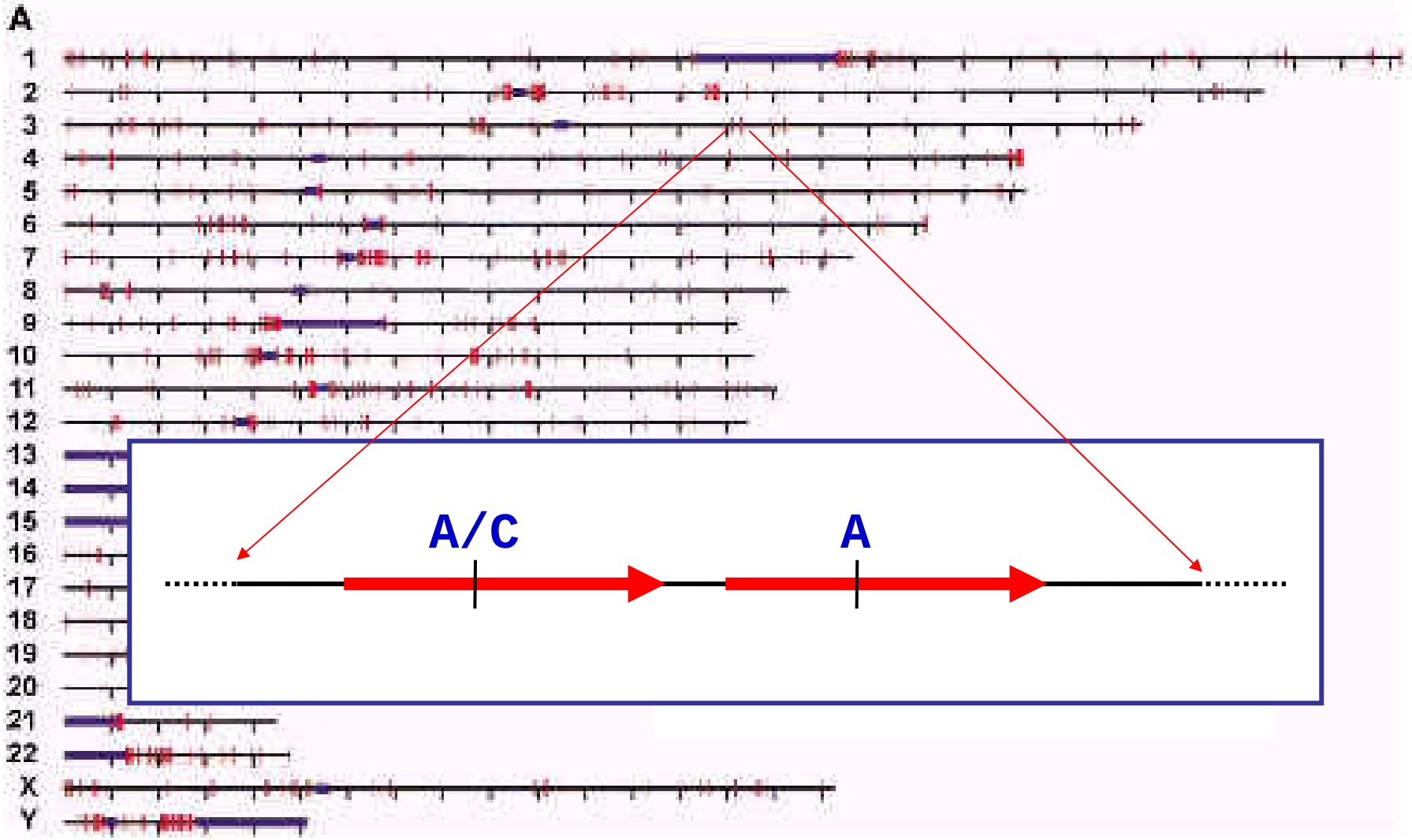
Most SNPs Are In Unique Sequence !



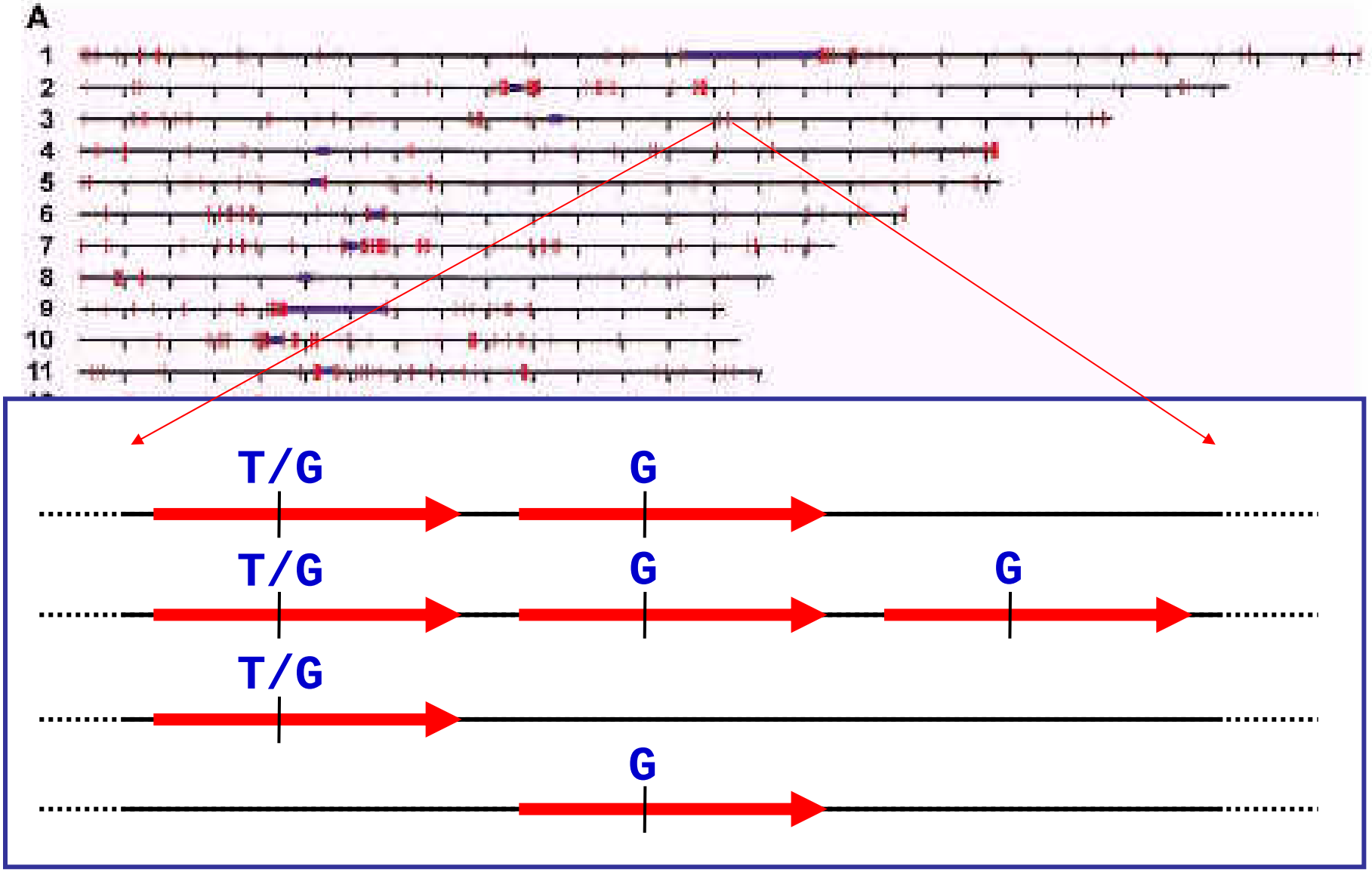
PSV (Paralogous Sequence Variant)



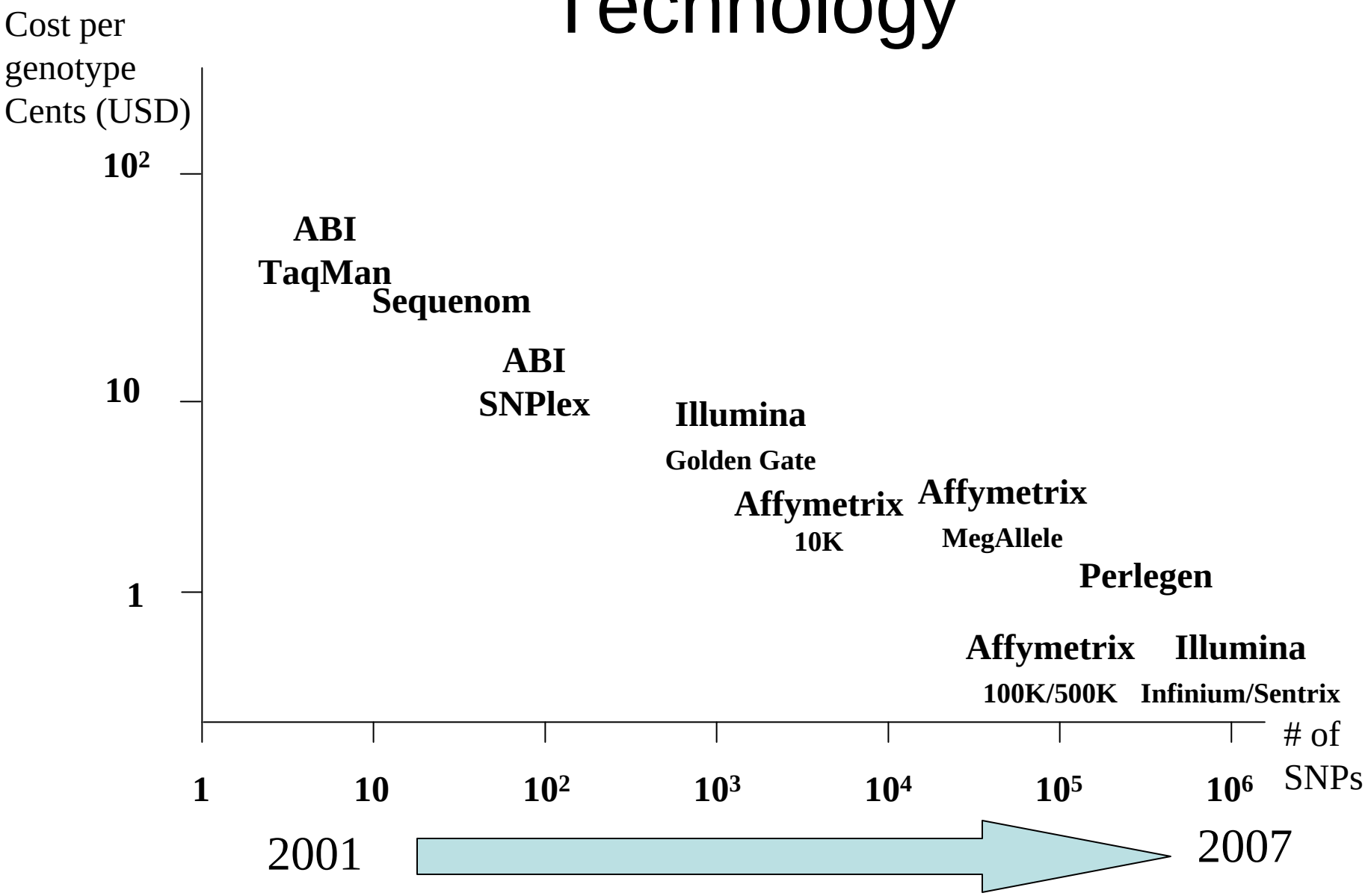
SNP in Duplicon Sequence



MSV: Multi-Site Variants



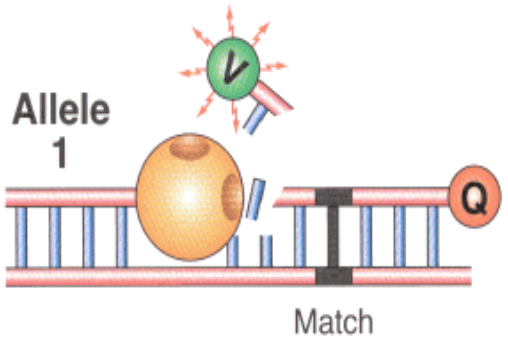
Progress in Genotyping Technology

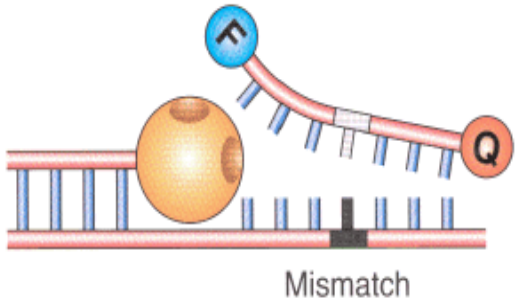


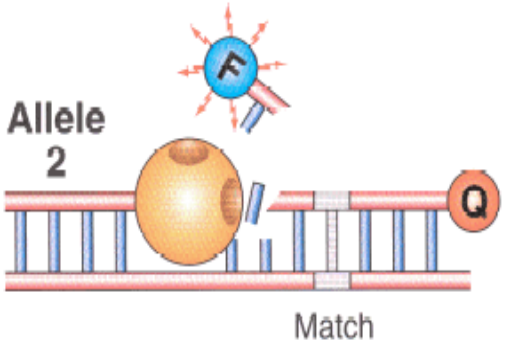
Genotype Technologies

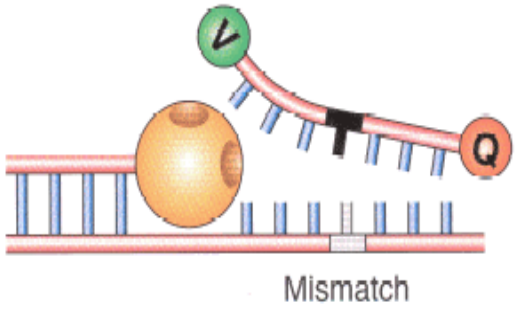
- Dropping costs
- Smaller amounts of DNA
  < 1ug for > 1 Million SNPs
- Economy of scale
  Frequent Flier Paradigm
- Increased density but with fixed products
- Custom products bear high development cost
- Challenge of mid-range (50 to 500 SNPs)

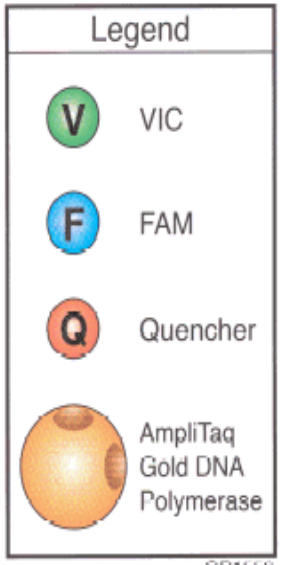
TaqMan™ (5’ Exonuclease)





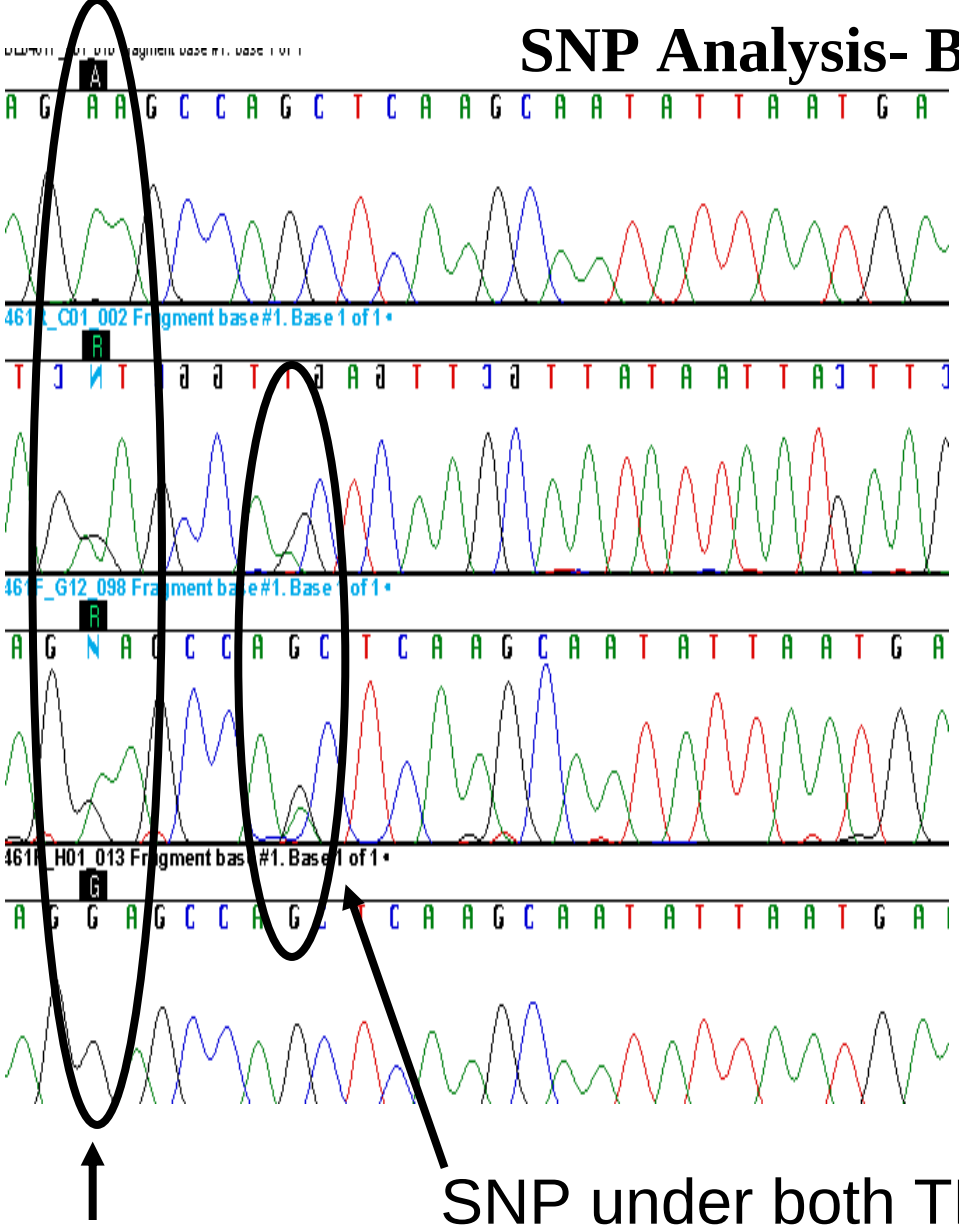






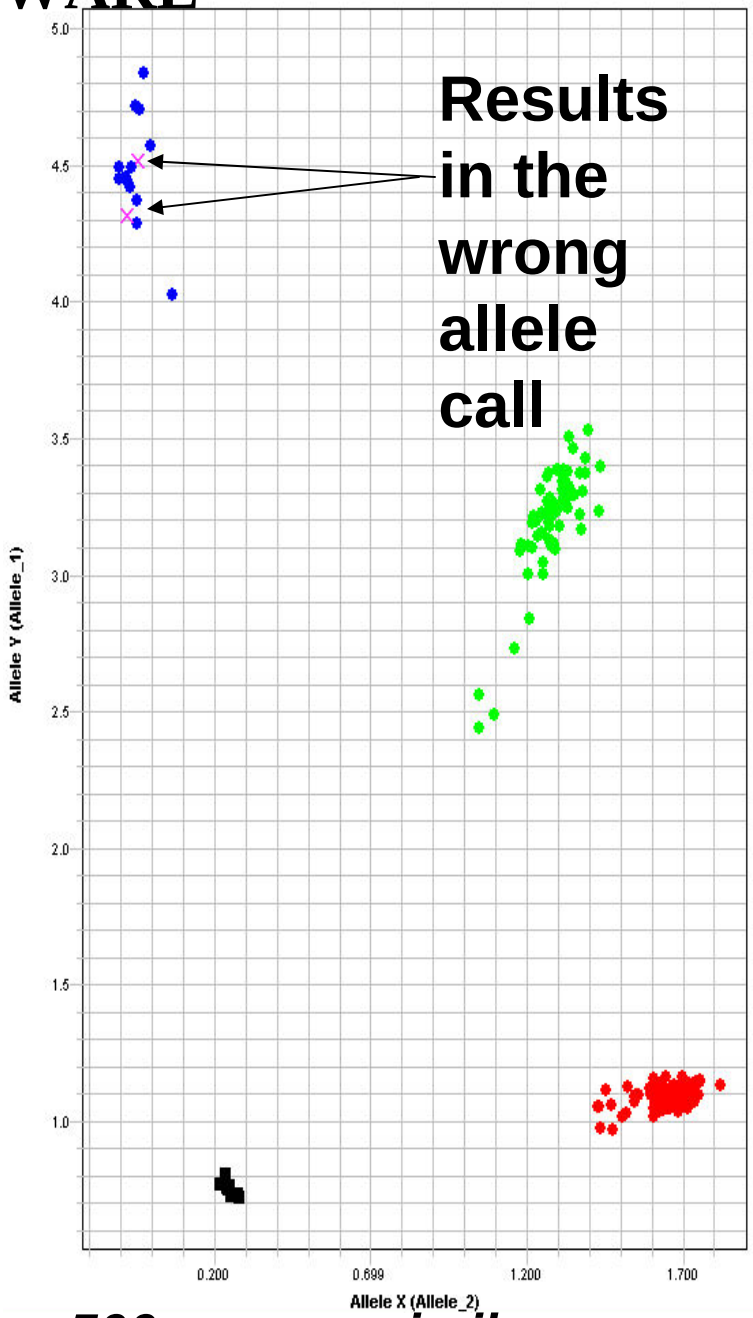
A substantial increase in...	Indicates...
VIC™ fluorescence only	homozygosity for Allele 1
FAM™ fluorescence only	homozygosity for Allele 2
Both fluorescences	heterozygosity

SNP Analysis- BEWARE



BRCA1-03

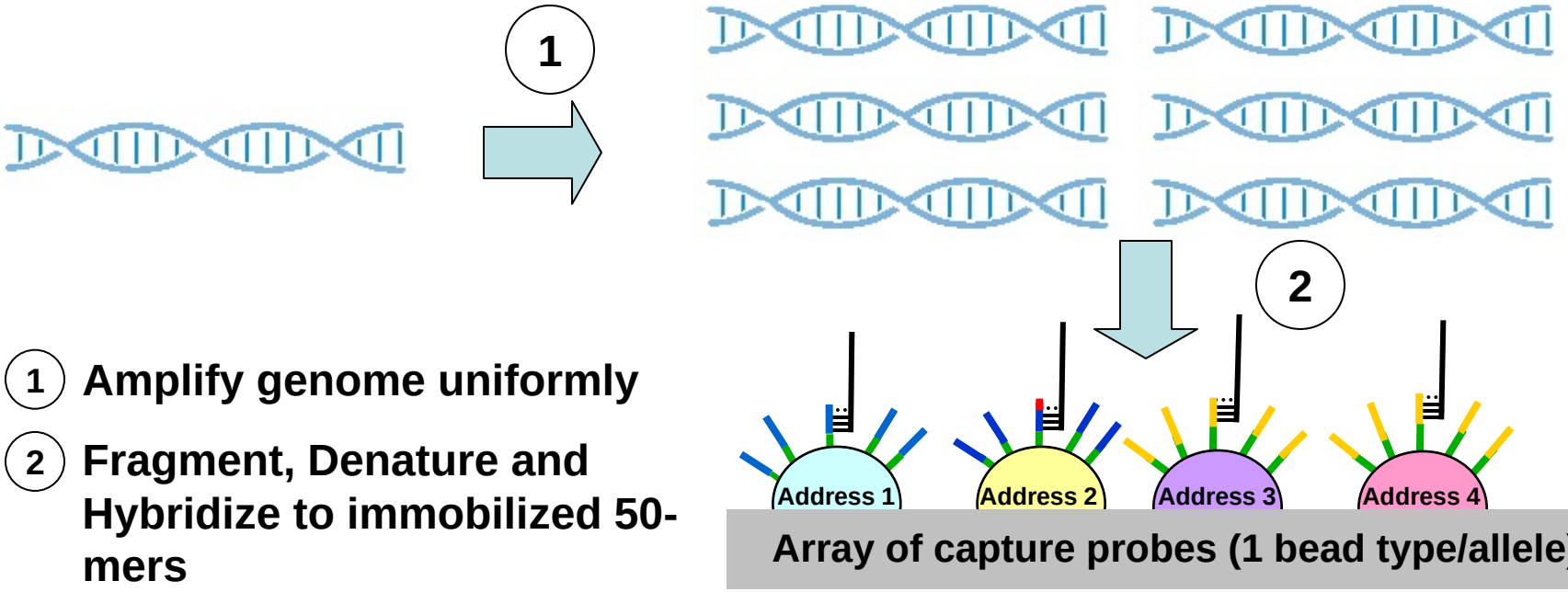
SNP under both TM probes



rs16941

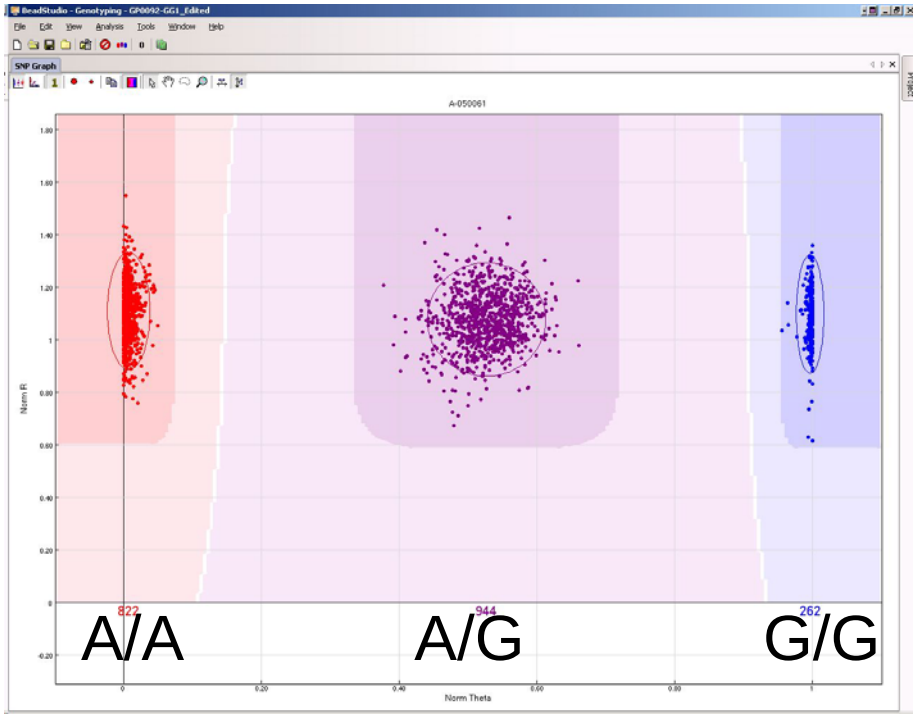
http://snp500cancer.nci.nih.gov

Infinium™ Assay

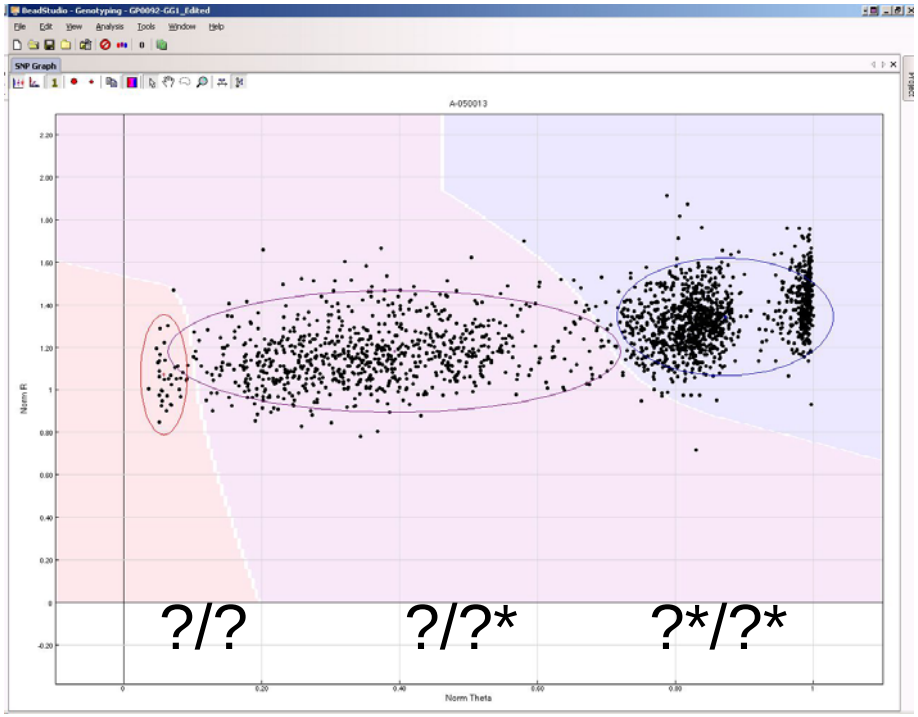


1. **Amplify genome uniformly**
2. **Fragment, Denature and Hybridize to immobilized 50-mers**
3. **Discriminate SNPs (enzyme-mediated)**
4. **Amplify signal and readout**

Illumina HumanHap500



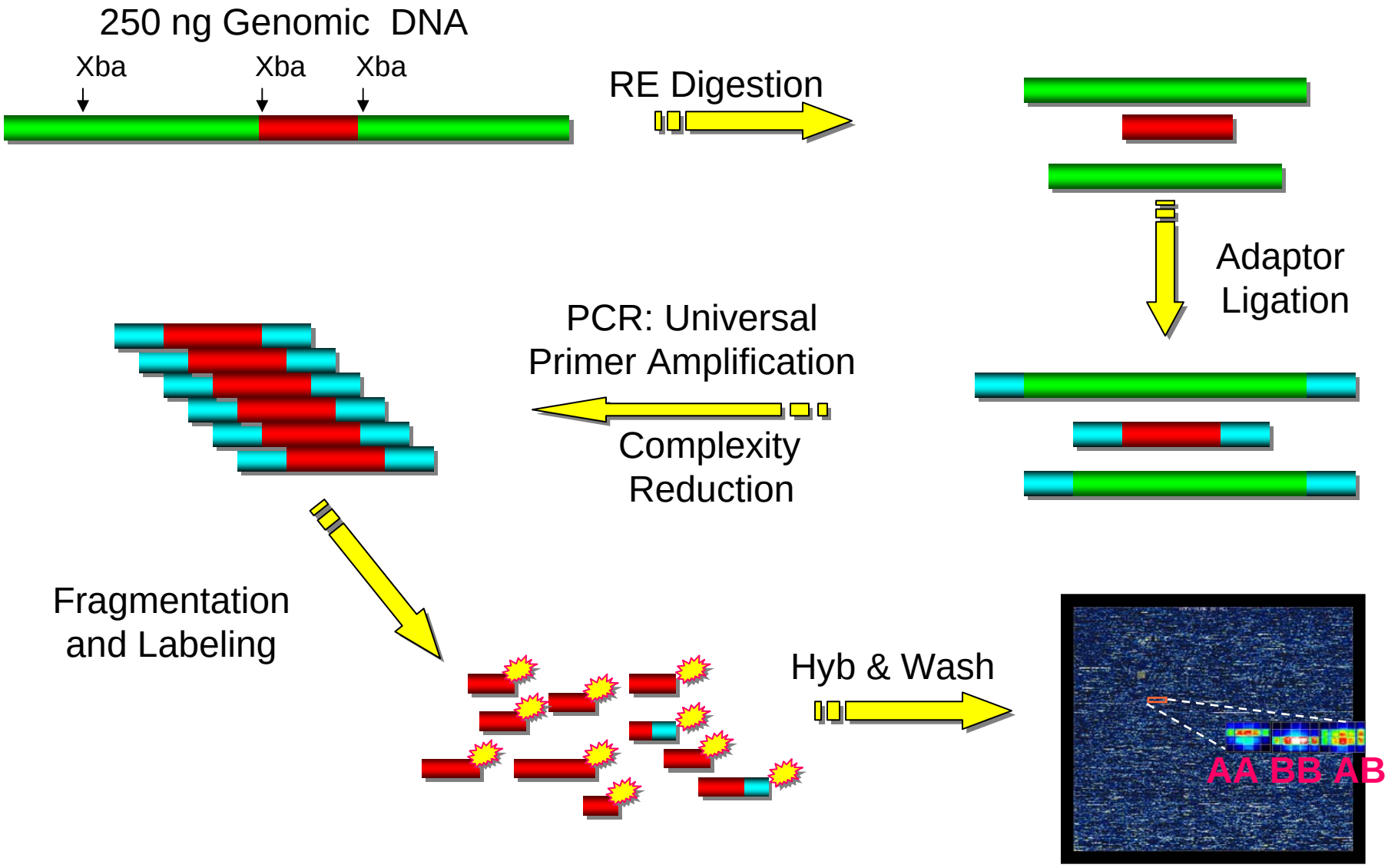
Good Cluster



Poor Cluster

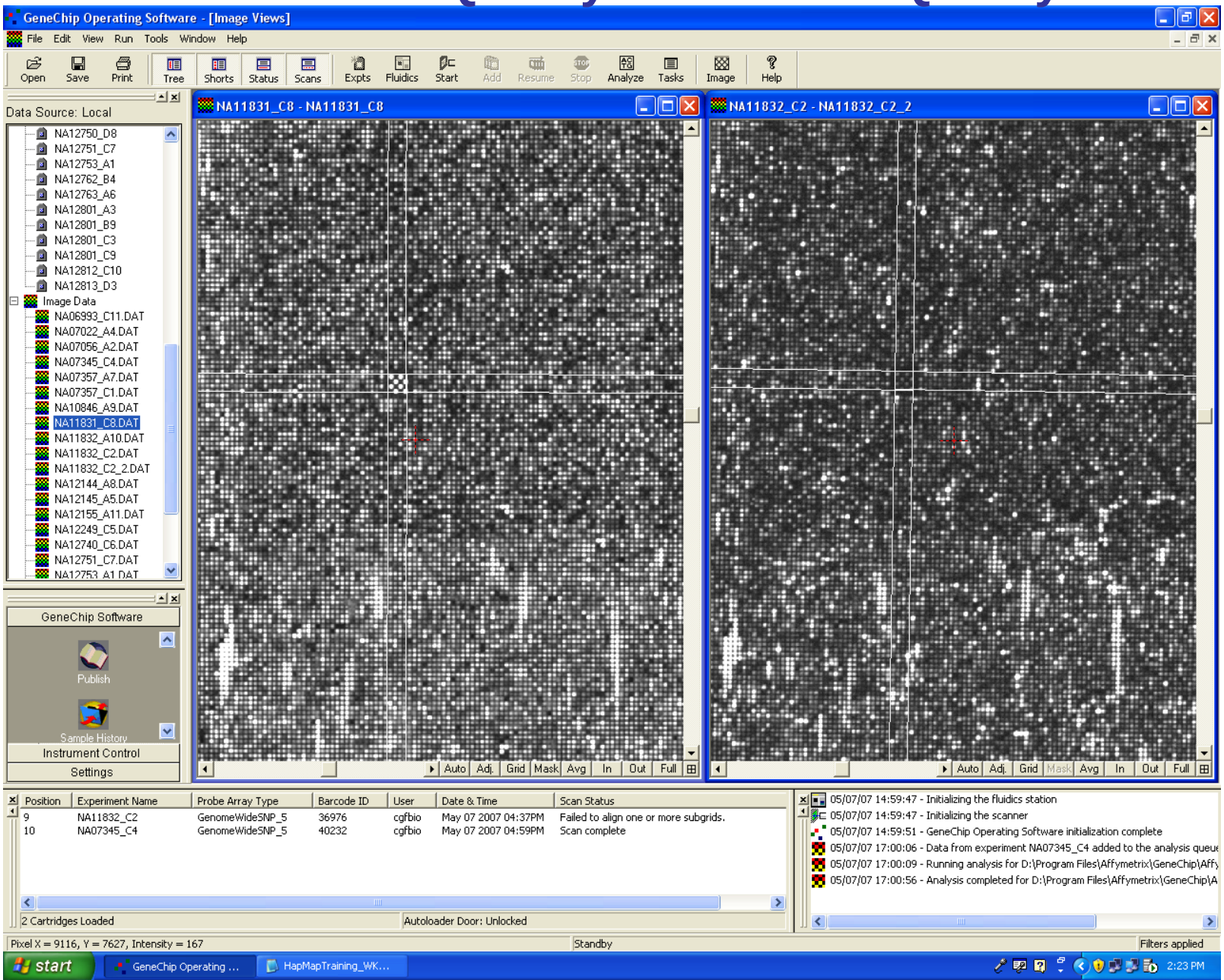
Read in BeadStudio™

Affymetrix GeneChip® Mapping Assay



Affymetrix 500K Chips

Poor Quality Good Quality



Important Points

- Too many data points to review individually
- Iterative algorithm for analysis
  - Still undergoing improvements
- Validation of notable SNPs with second technology
  - “Neighboring SNP-land mines”
- Do not do this at home- Only for highly trained personnel

Choice of Dense SNP Platforms

Affymetrix

Basic Points

- Based on ‘Spacing’
- 100k, 500k, 1 Million
- CNV Analysis
- WGA compatible

Issues

- Lower price
- 2 Enzyme Problem
- Calling Algorithms
- Redundancy (useful)

Illumina

Basic Points

- Based on ‘tagging’
- 317k, 550k, 1 Million
- CNV Analysis
- WGA not yet rec’d

Issues

- Higher price
- HapMap II Based

2007 What is Available for Whole Genome Scans

- **Coverage analysis based on HapMap II Data**
- **Build 20 MAF ≥5%, r^2 ≥ 0.8 (pair-wise)**

		CEU	YRI	JPT/CHB
Illumina	**HumanHap300**	**80%**	**35%**	**40%**
Illumina	**HumanHap500**	**91%**	**58%**	**88%**
Affymetrix*	**500k Mapping**	**63*%**	**41%**	**63%**

***77%** **(with 50k MegA)**

Quality control of genotype calls
&
DNA handling

Quality Control for Called Genotypes

PURPOSE:

Identification of unreliable SNPs and DNAs to be *entirely* removed from the analysis .

Evaluation of completion rate (DNAs)

Evaluation of call rate (SNPs)

Evaluation of discordance rate (error rate)

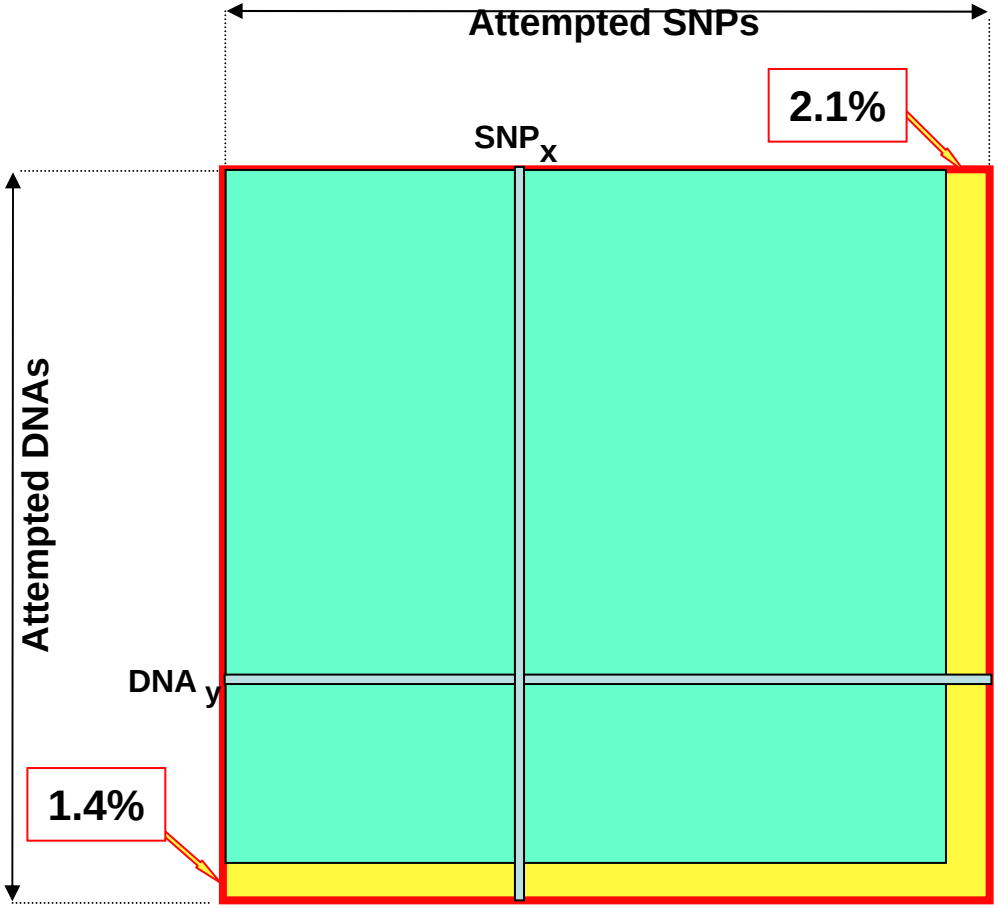
CGEMS SNP DNA Success Rates

Criterium
SNP call rate > 90%

	Number of individuals attempted	failed	success rate
PLCO	561,494	1,490	0.973
NHS	555,352	8,706	0.984

Threshold
DNA completion rate > 94%

Number of individuals attempted	failed	Success rate
4696	66	0.986



Discordance rate for CGEMS: HumanHap500 (Illumina)

Participants
142 duplicate pairs

Mean discordance rate
Prostate 2.0 x 10^{-4}
Breast 1.5 x 10^{-4}

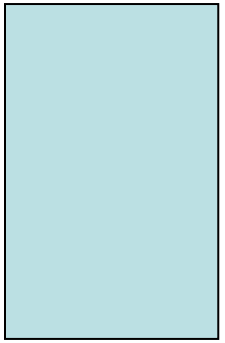
CEPH-CGEMS
74 duplicate pairs

Mean discordance rate
2 x 10^{-4}

28 individuals
(with 24 duplicates)

Mean discordance rate
1.4 x 10^{-3}

CEPH-HapMap



Concordance rates >99.5%
Subtle Differences in Quality of DNA

Quality Control for Recruitment DNA handling

Checking for:

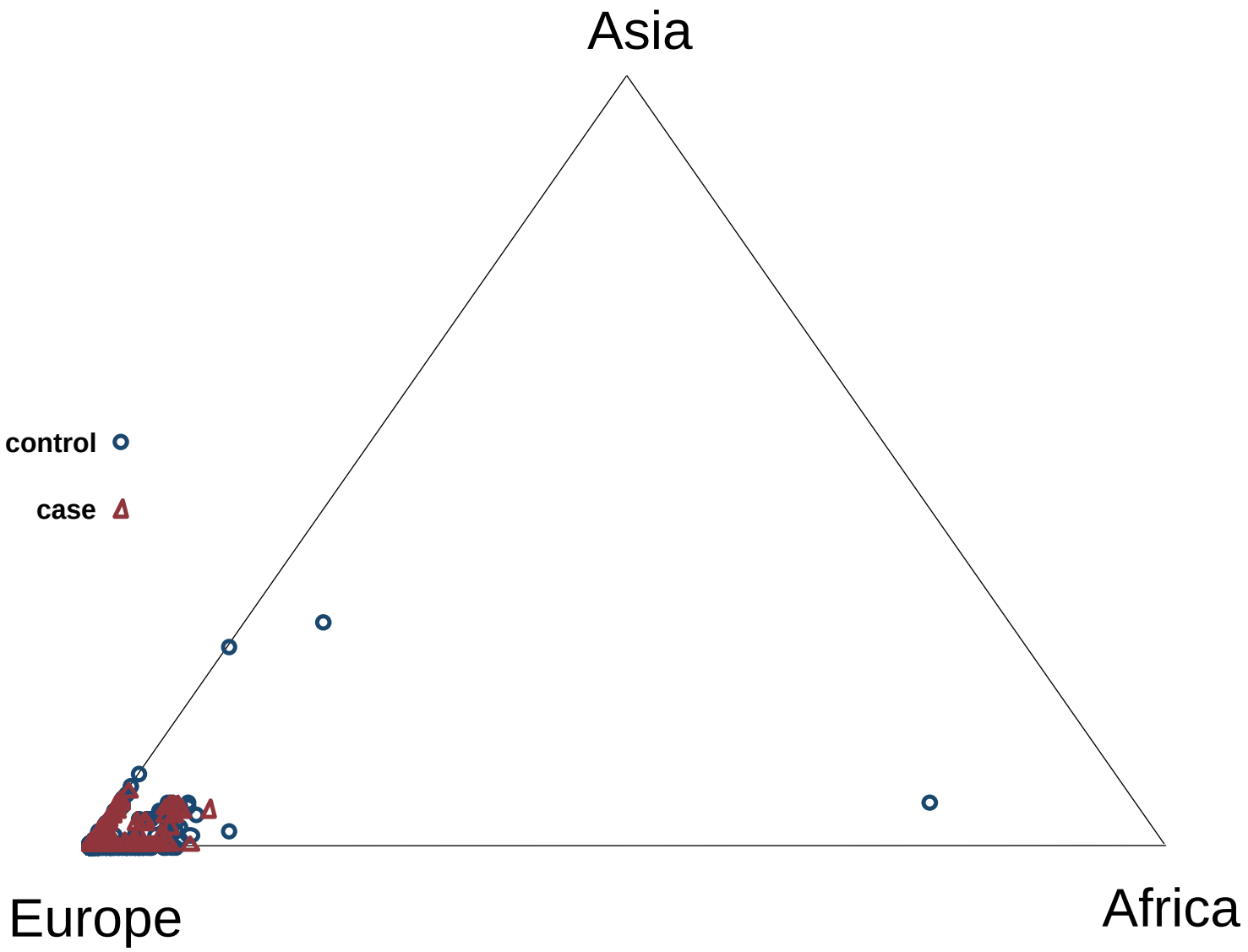
- Chromosome X Ploidy
- Identification of Familial Relationships
- Evaluation of Continental Admixture
  - Population Stratification
- Principal Component Analysis
- (Hardy Weinberg Statistics)

Analysis of CGEMS Data Sets

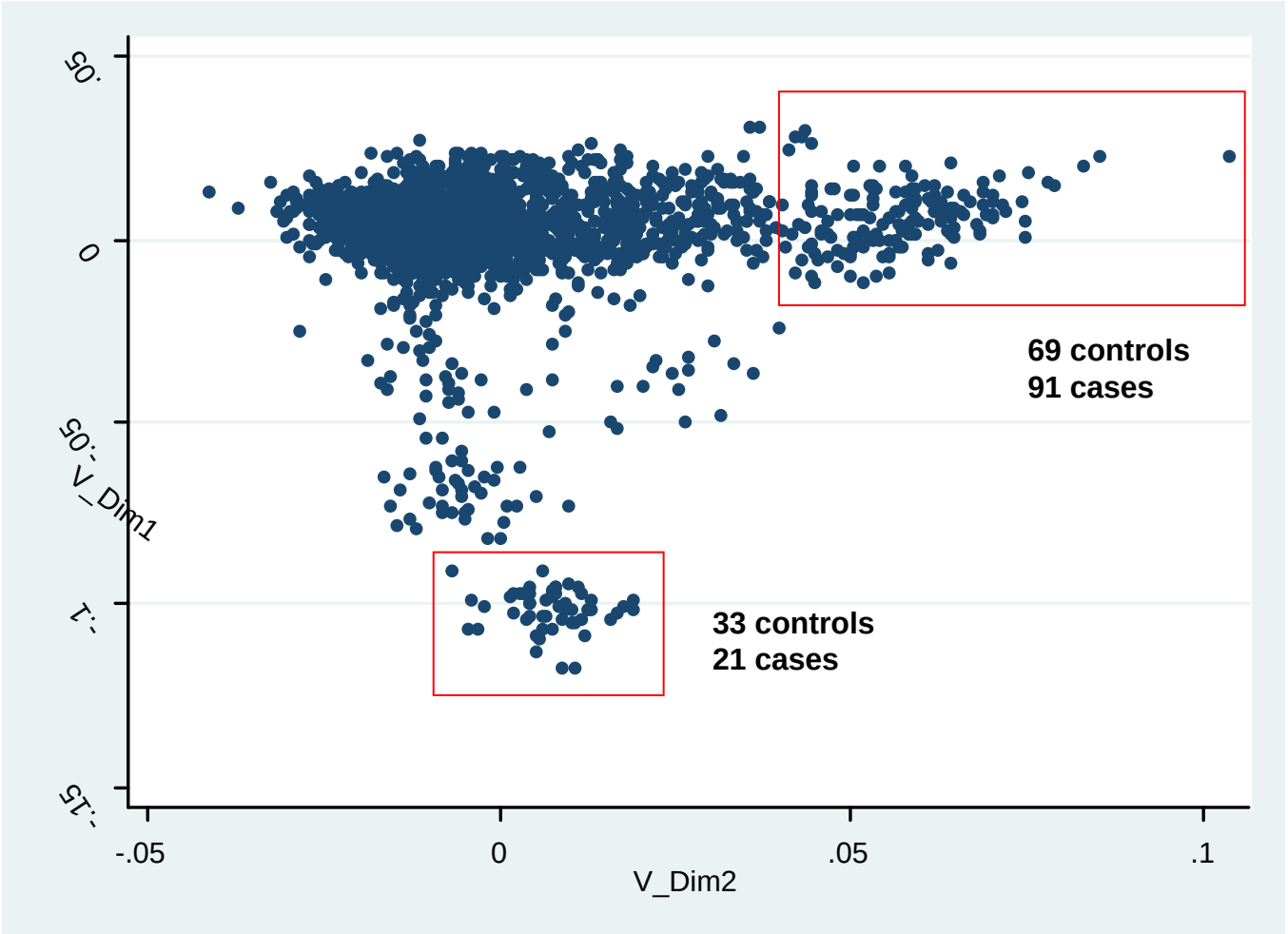
	Chromosome X		Unexpected duplicates	1st & 2nd degree relatives	Other 3rd to 5th degree relatives
	1 copy	2 copies			
Prostate	2279	3	3 pairs	5	20*
Breast	0	2299	3 pairs	1	to be done

*It was noted subsequently that both members of each pair had been recruited in the same center.

Admixture coefficient in PLCO samples



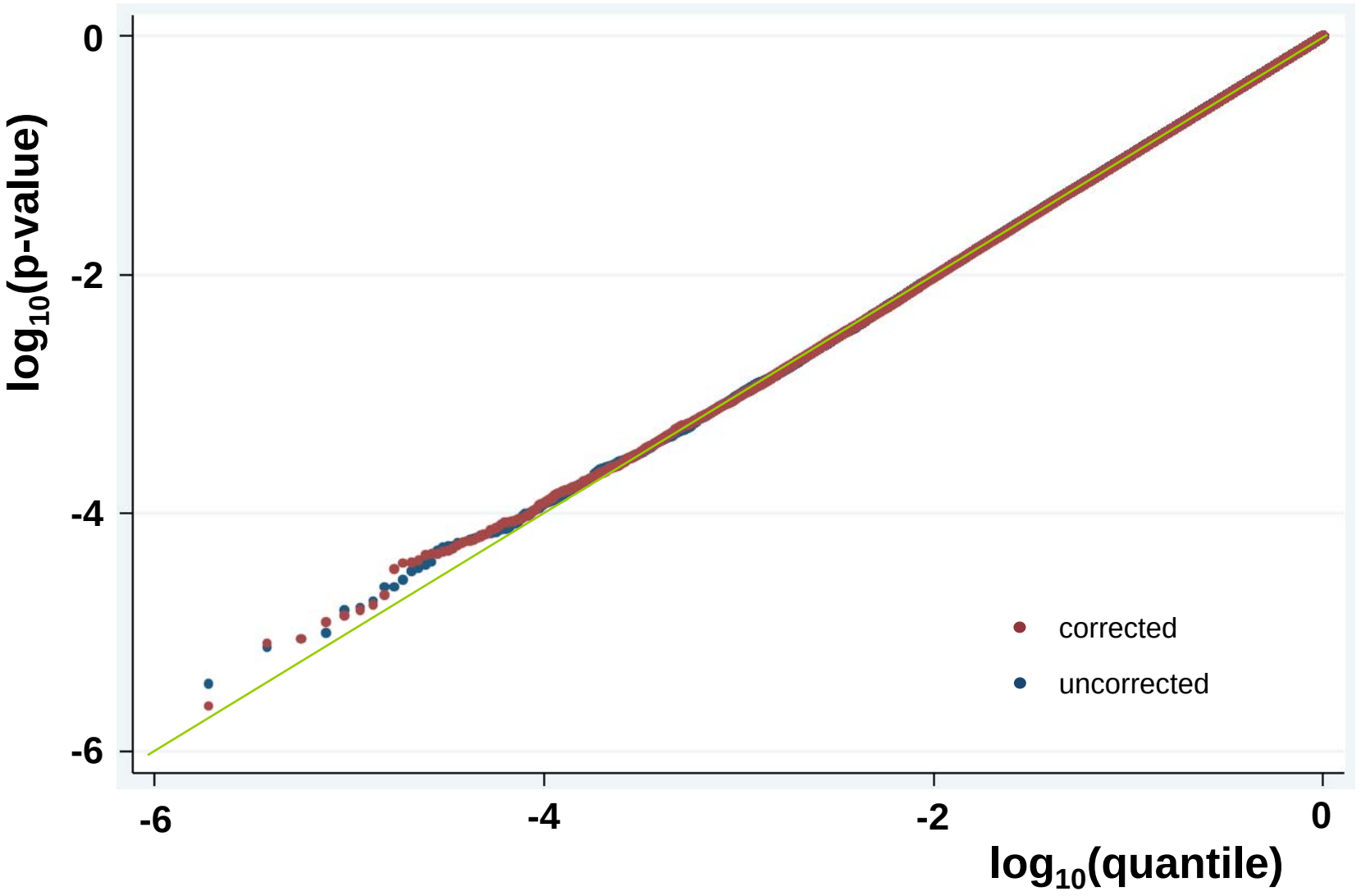
CGEMS Prostate Cases & Controls
Principal Component Analysis



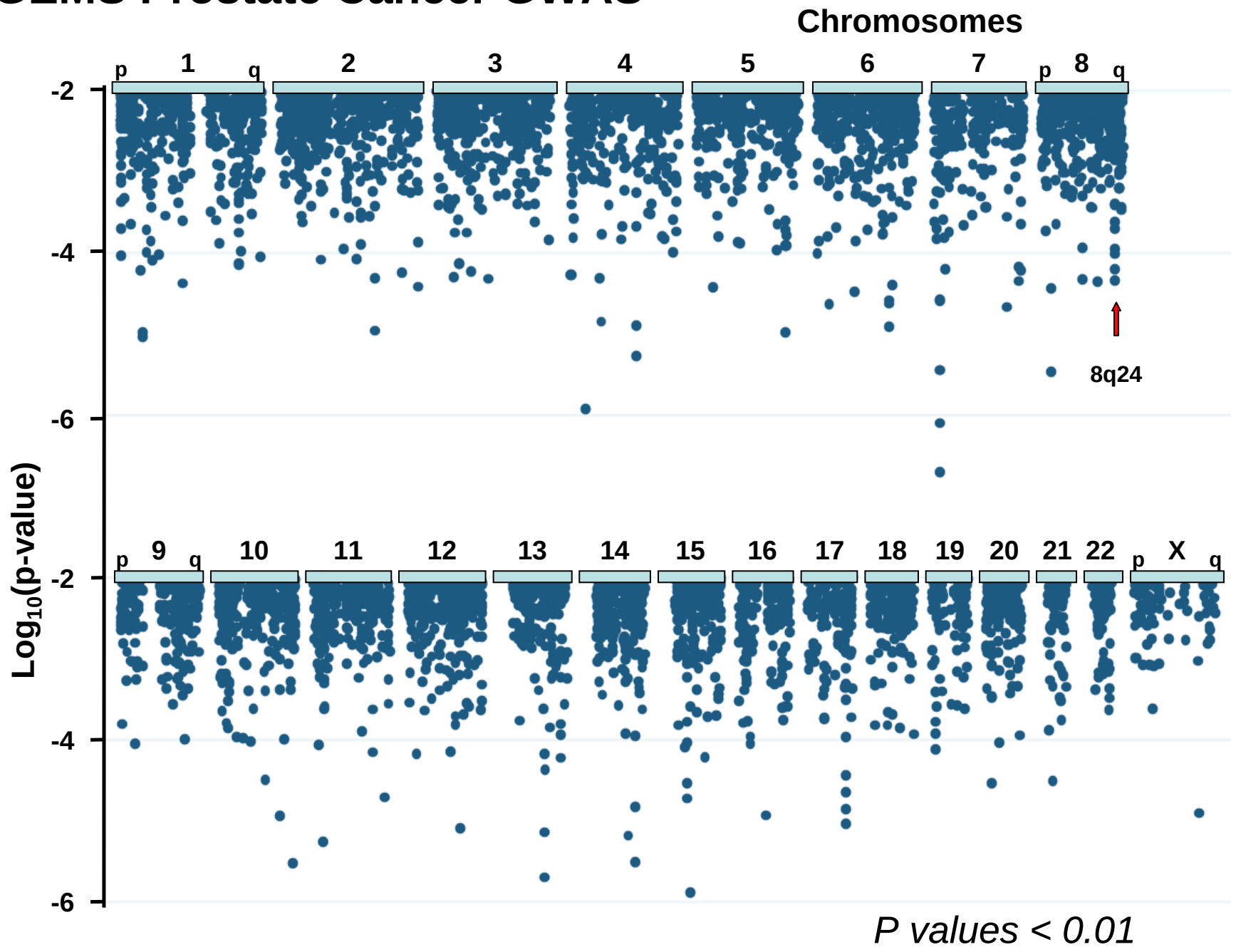
Based on Price et al Nat Genet 2006

CGEMS Breast Cancer Scan

log quantile plot of p-values for the Entire Set of Markers



CGEMS Prostate Cancer GWAS



incidence density sampling

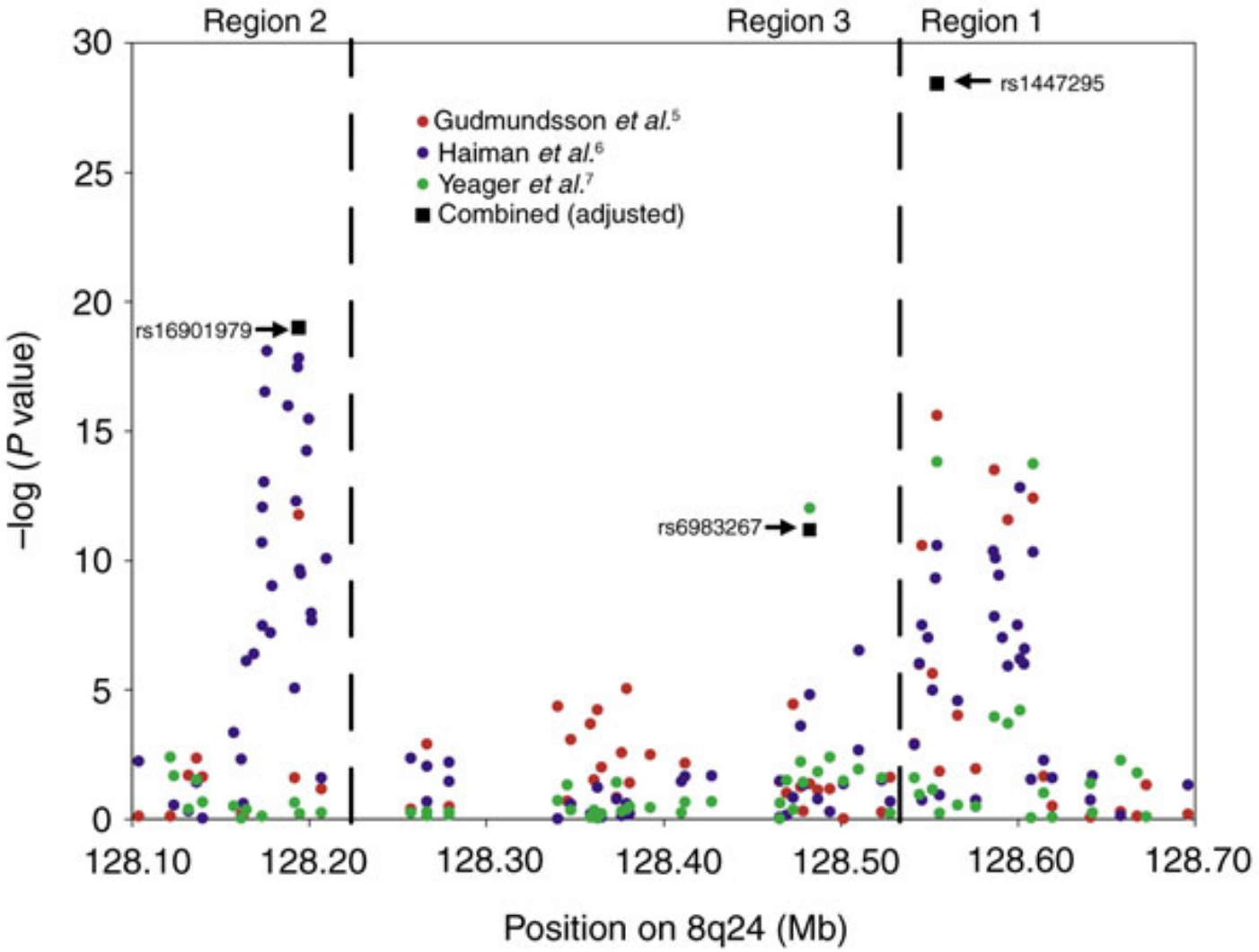
Replication Studies in CGEMS Prostate Cancer GWAS

	Subjects		rs6983267 Predisposing allele frequency		rs6983267 P-value	rs1447295 Predisposing allele frequency		rs1447295 P-value
			Cases	Cont.		Cases	Cont.	
PLCO	1157	1172	0.55	0.49	2.4×10^{-05}	0.14	0.10	9.8×10^{-05}
ACS	1151	1150	0.55	0.50	3.2×10^{-03}	0.12	0.08	2.7×10^{-05}
ATBC	896	894	0.57	0.51	1.9×10^{-03}	0.21	0.17	2.9×10^{-02}
FPCC	459	455	0.56	0.51	1.2×10^{-01}	0.12	0.07	4.4×10^{-03}
HPFS	636	625	0.57	0.51	1.0×10^{-02}	0.13	0.09	2.7×10^{-03}
ALL	4299	4296	0.56	0.50	9.4×10^{-13}	0.15	0.11	1.5×10^{-14}

Estimated Odds Ratios Overall

Heterozygotes	1.26	1.43
Homozygotes	1.58	2.23

Meta-Analysis of 8q24 papers in *Nature Genetics*: J Witte



J Witte Nature Genetics 2007