Module 6

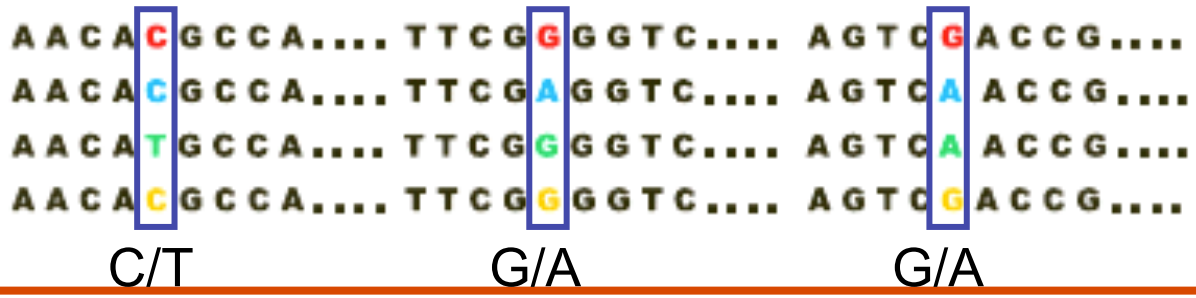
Genomic variation

Matthew D. Clark PhD
Group leader
Genomics, The Genome Analysis Centre
Norwich, UK

Any two copies of the human genome have 1 difference per every 1000 bases

Variation Types

- Cytological level:
  - Chromosome numbers
  - Segmental duplications, rearrangements, and deletions
- Molecular level:
  - Transposable Elements
  - Short Deletions/Insertions, Tandem Repeats
- Sequence level:
  - Single Nucleotide Polymorphisms (SNPs)
  - Small Nucleotide Insertions and Deletions (Indels)

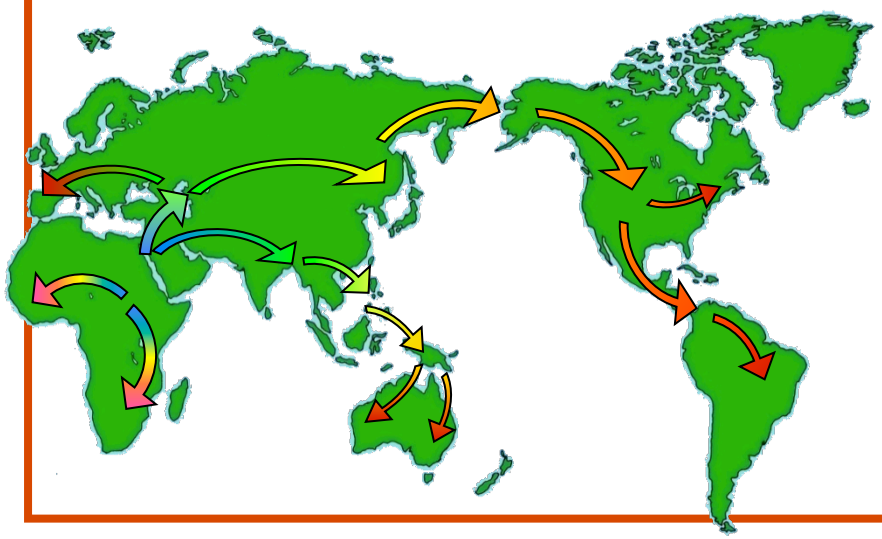


Variation is useful

- Determine disease risk
- Predict reactions to environmental triggers
- Predict responsiveness to drug treatments
- Forensics
- Evolution & migration

The Open Door Workshop

The Source of Variation – SNP's



CGCATTAAACTCATGAG

CGCATTAAGCTCATGAG

CGCATTAAACTCACGAG

CGCATTAAGCCCACGAG

CGCATTAAGCTCACGAG

CGCATTAAGCTCATGAG

AGTAATAGGCTCAAGTG

AGTAATAGGATCAAGAG

AGTAATAGGCTCAAGAG

AGCAATAGGCTCACGAG

AGTAATAGGCTCATGAG

AGCAATAGGCTCATGAG

AGCATTAAGCTCATGAG

AGCAATAAGCTCATGAG

Now

10,000

20,000

30,000

40,000

50,000

60,000

70,000

80,000

90,000

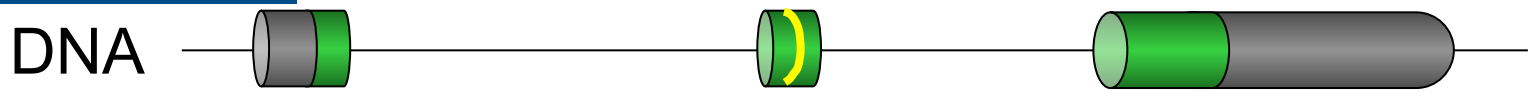
100,000

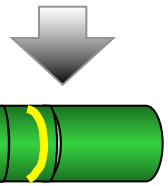
Years

Types of SNPs

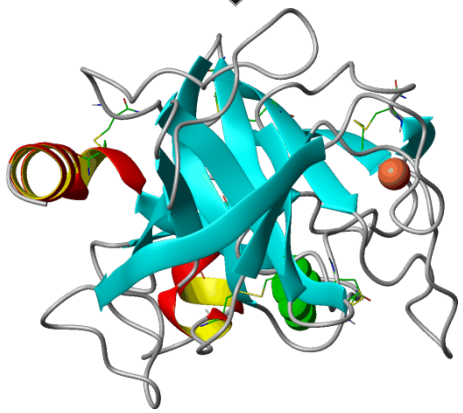
- Genic, coding SNPs
  - Frameshift
  - Splice site
  - Non-synonymous
  - Synonymous (splice enhancer/suppressor?)
- Genic, non-coding SNPs
  - Untranslated region
  - Regulatory SNPs
  - Intronic SNPs
- Intergenic

The Open Door Workshop





Drug metabolism:
The CYP2D6 gene



Coronary disease:
LDL receptor gene

. . . CAC TCC TGA CGC . . .

167 168

. . . His Ser Stop

. . . TTT TAC G T C ATG . . .

289 290 291 292

. . . Phe Tyr Ser Met . . .

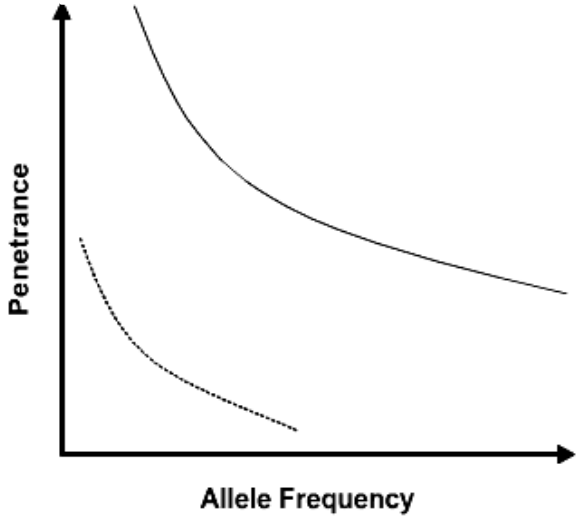
Deep-vein thrombosis:
The Factor V gene

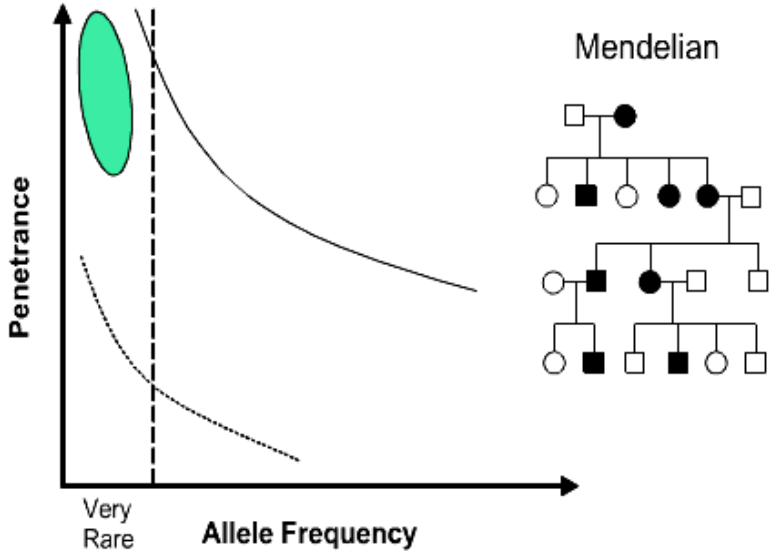
504 505 506 507

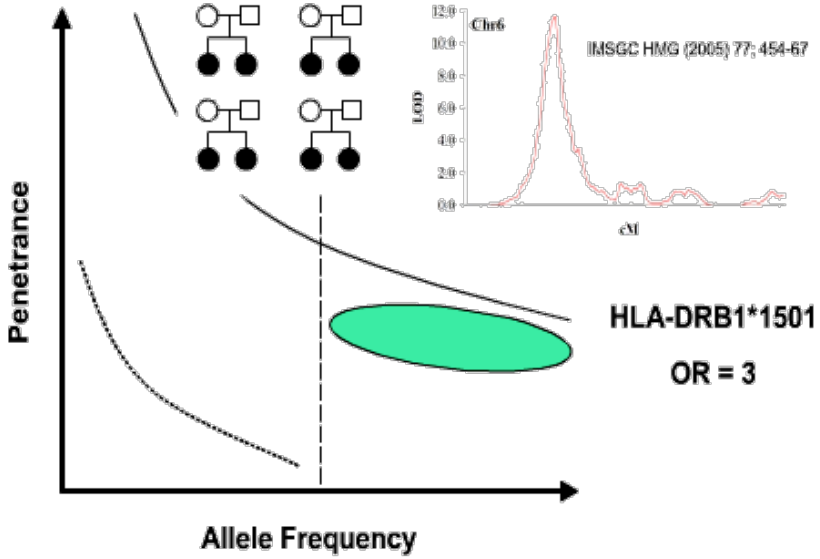
. . . Asp Arg Gln Gly . . .

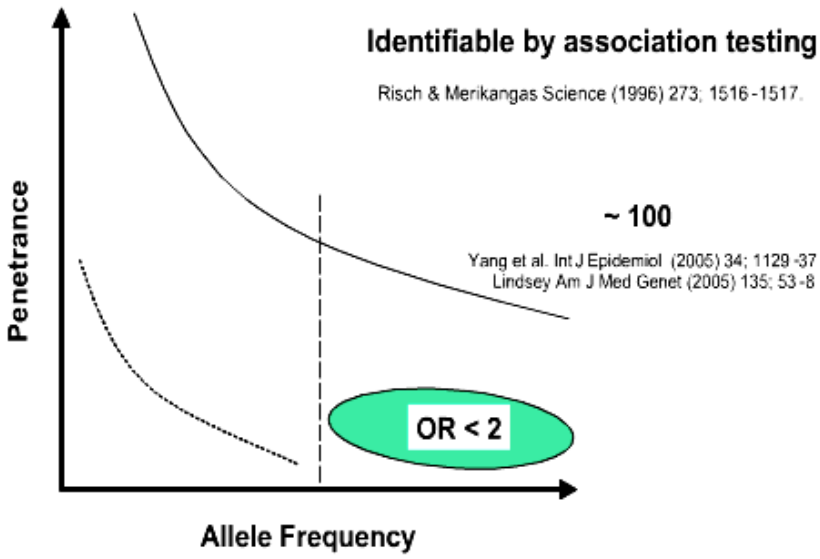
APC cleavage

The Genetic Architecture of Complex Disease





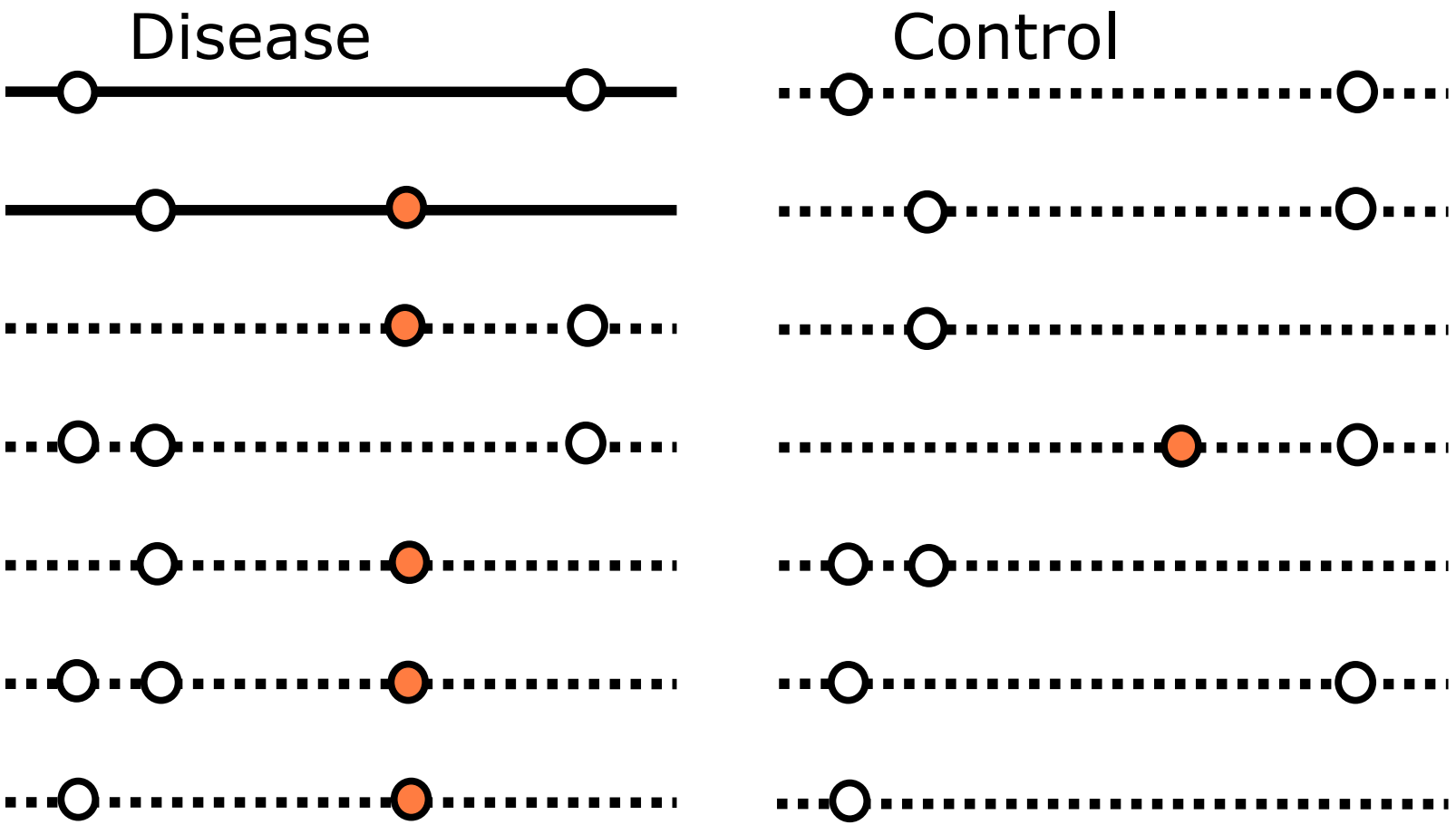




From Sawcer ACTRIMS 2008

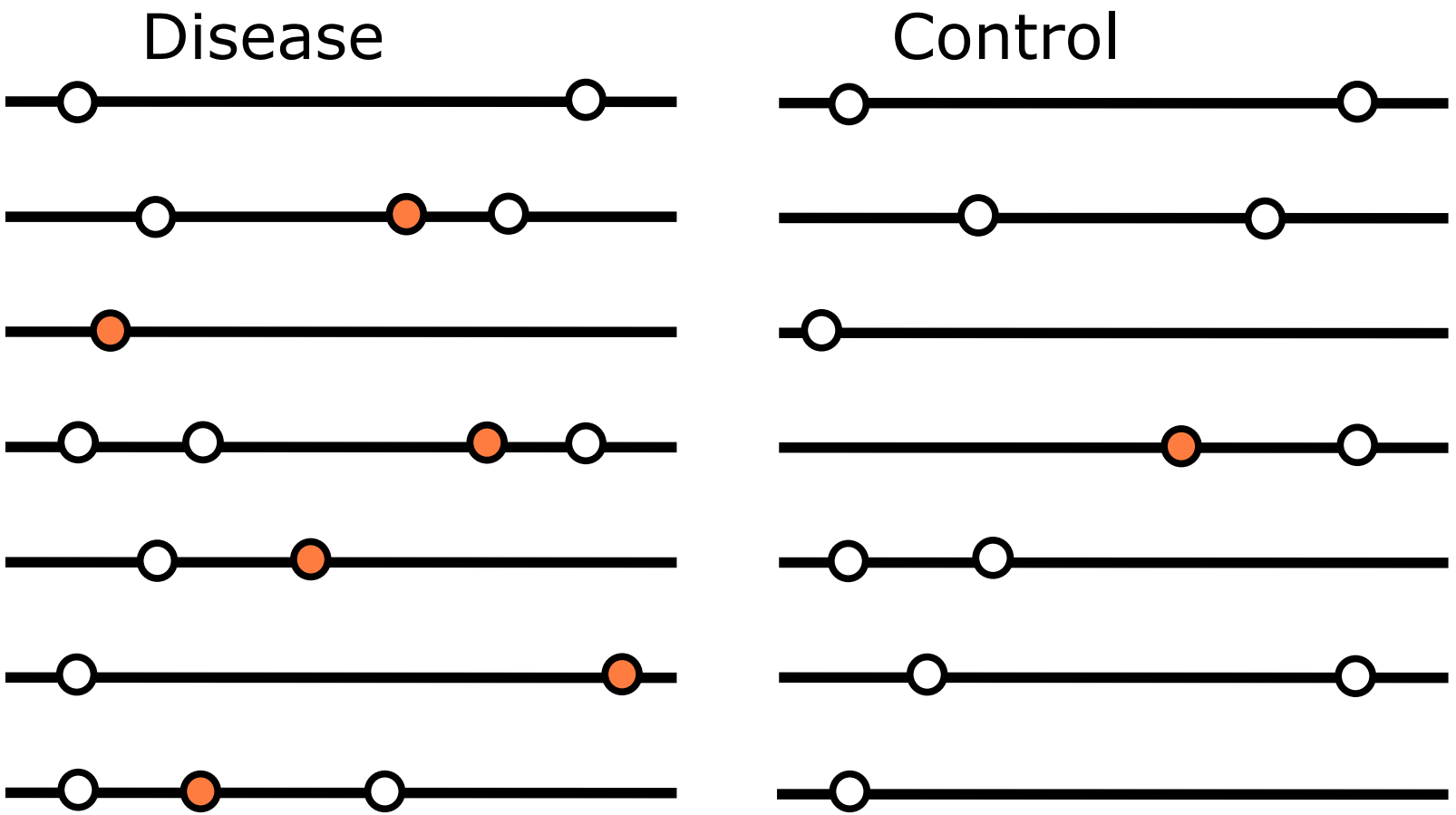
Classical association studies

Common allele



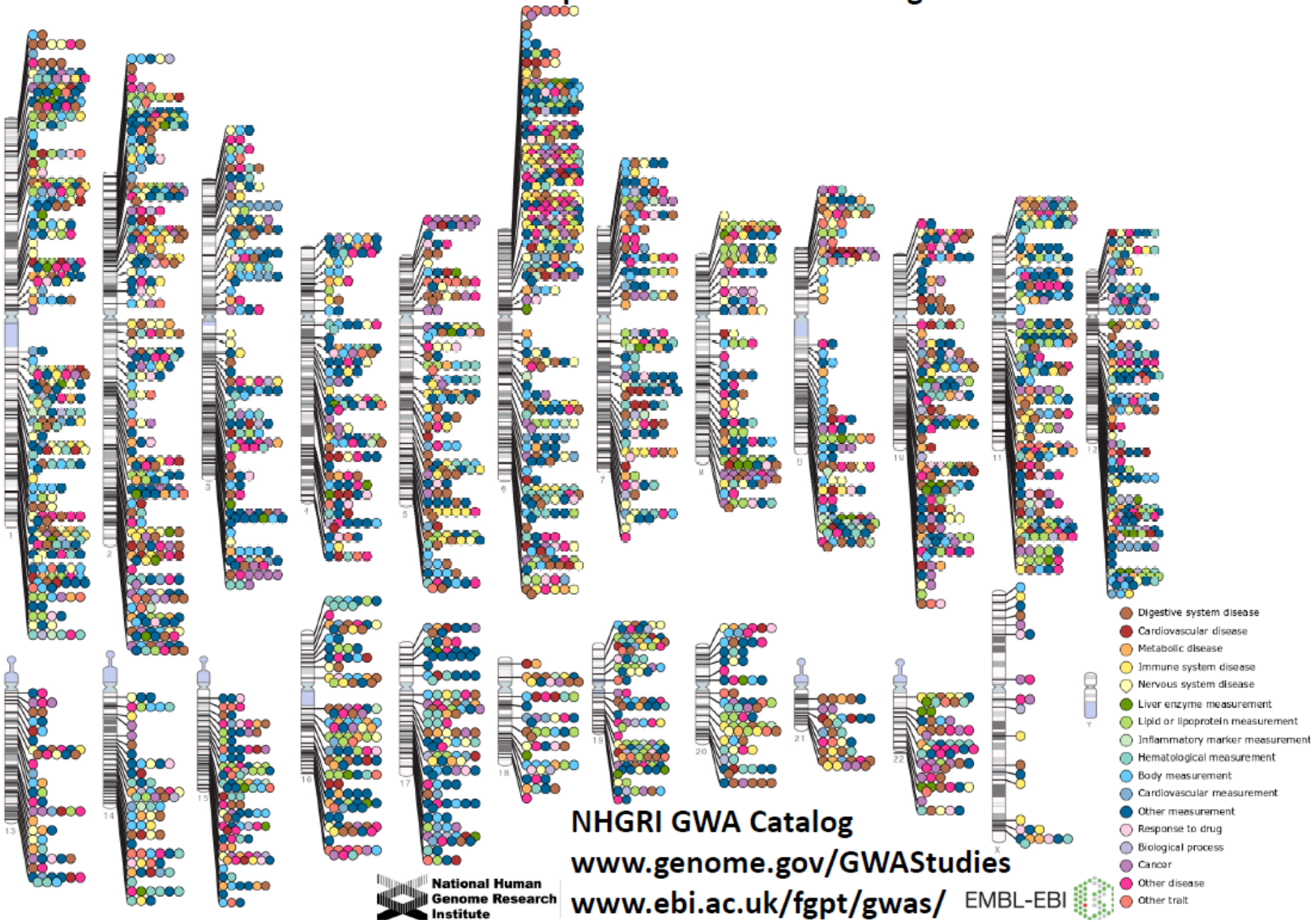
“Mutation enrichment” association studies

Rare alleles



Published Genome-Wide Associations through 12/2012

Published GWA at $p \leq 5X10^{-8}$ for 17 trait categories



NHGRI GWA Catalog

www.genome.gov/GWAStudies

www.ebi.ac.uk/fgpt/gwas/

National Human Genome Research Institute

EMBL-EBI

Table 2. Benefits, Misconceptions, and Limitations of the Genomewide Association Study.

Benefits
Does not require an initial hypothesis
Uses digital and additive data that can be mined and augmented without data degradation
Encourages the formation of collaborative consortia, which tend to continue their collaboration for subsequent analyses
Rules out specific genetic associations (e.g., by showing that no common alleles, other than *APOE*, are associated with Alzheimer's disease with a relative risk of more than 2)
Provides data on the ancestry of each subject, which assists in matching case subjects with control subjects
Provides data on both sequence and copy-number variations
Misconceptions
Thought to provide data on all genetic variability associated with disease, when in reality only common alleles with large effects are identified
Thought to screen out alleles with a small effect size, when in reality such findings may still be very useful in determining pathogenic biochemical pathways, even though low-risk alleles may be of little predictive value
Limitations
Requires samples from a large number of case subjects and control subjects and therefore can be challenging to organize
Finds loci, not genes, which can complicate the identification of pathogenic changes on an associated haplotype
Detects only alleles that are common (>5%) in a population
Requires replication in a similarly large number of samples

Hardy and Singleton NEJM 2009

Missing heritability question

- Twin studies reveal of extent of genetic inheritance
- Many quantitative traits e.g height are >50% heritable
- GWAS typically explains <<20% of phenotype

Possible answers?

- GWAS are too small
    - Meta-analysis does increase measured heritability
- Rare alleles (strong effect)
- Epistasis
    - Complex genetic interactions e.g. limiting pathway
- Epigenetics
    - Transgenerational environmental effects

Genome wide association studies

- Common (5%) variants, common disease
- http://www.genome.gov/26525384

Change in approach

- Importance of rare variants in common disease
- 1000 Genome project **(<0.05% variants)**
  - **www.1000genomes.org**

Personalised Genomes

SNPs + **indels, SV, CNV & unique content**

Individual	SNPs	Novel
J. C. Venter	3,074,574	160,370
J. Watson	2,060,544	98,926
Chinese	3,074,061	84,786
Korean	3,439,097	130,566
NA19240 (Yoruban)	3,586,490	**216,968**
D. Tutu (Bantu)	3,624,334	**412,754**
KB1 (Khiosan)	4,053,781	**743,714**
1000 genomes	17.3 million	9 million

Schuster et al. Nature 463, 943-947, and Via et al. *Genome Medicine* 2:3

Every Genome?

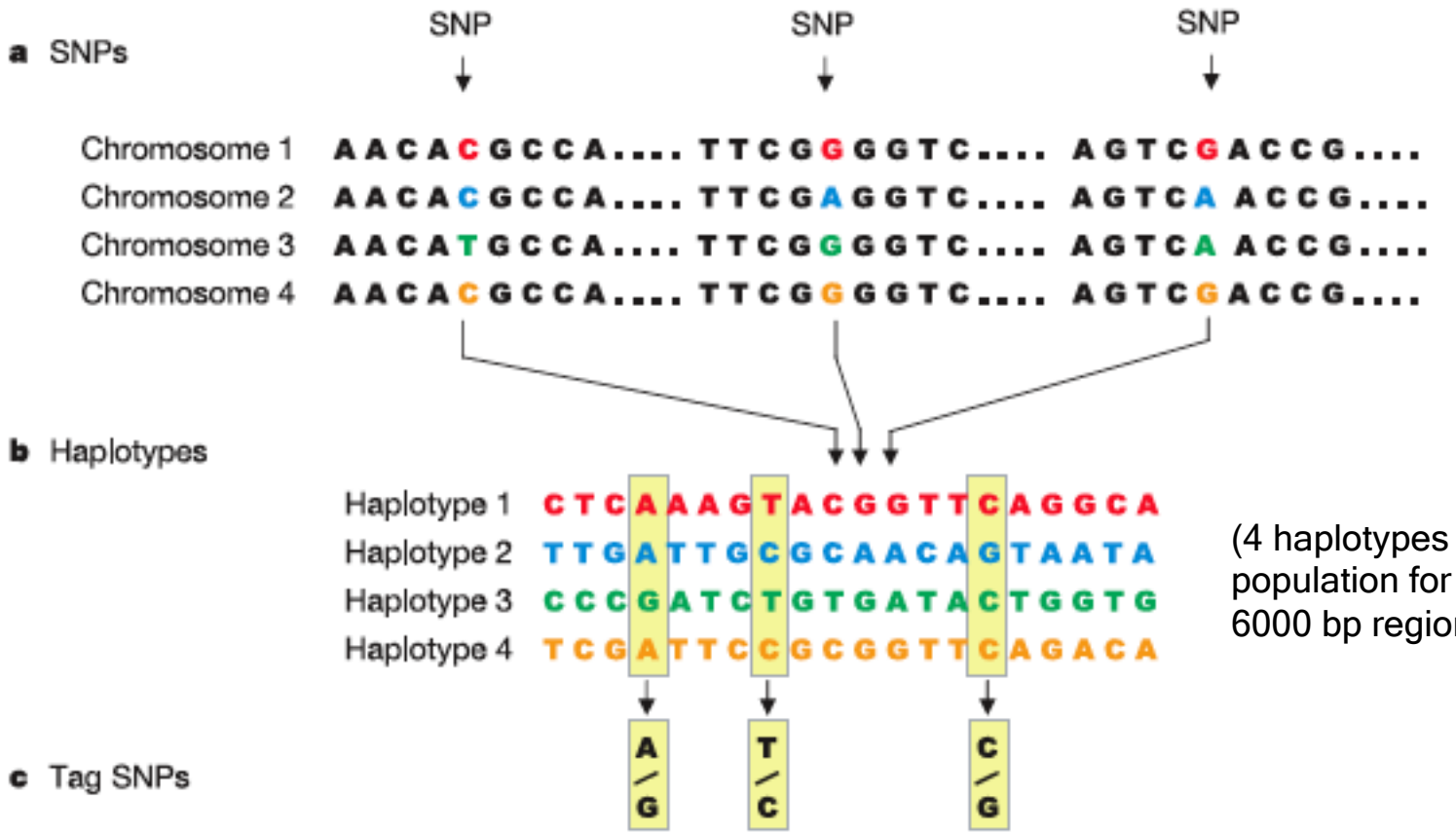
QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

Direct to consumer (DTC) marketing of genome sequencing

Worked example 1:

From microarray to polymorphisms

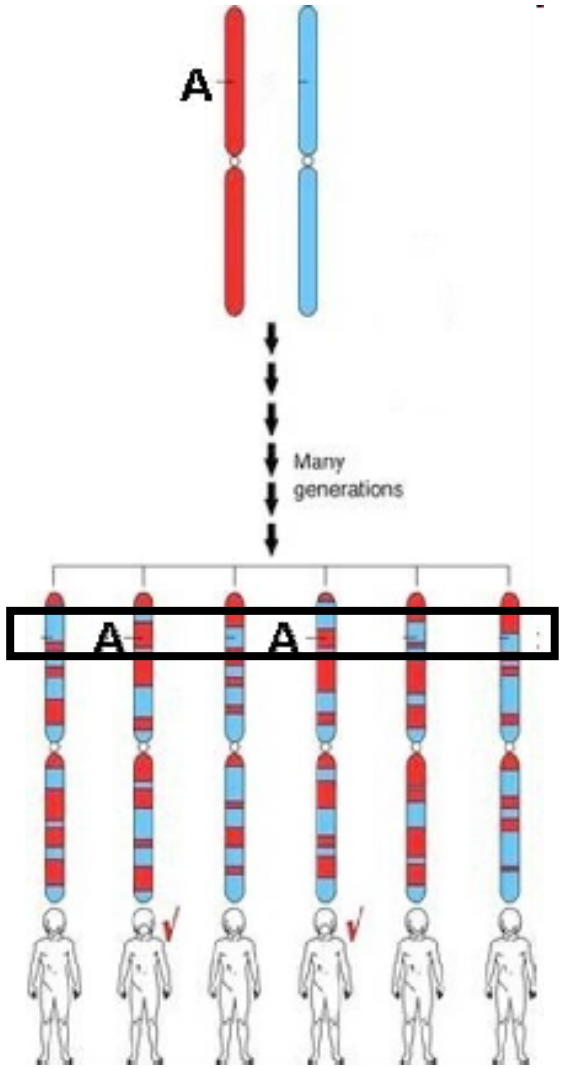
Haplotypes



(4 haplotypes in the population for a 6000 bp region)

The International HapMap Consortium, Nature 2003

Linkage and Haplotypes



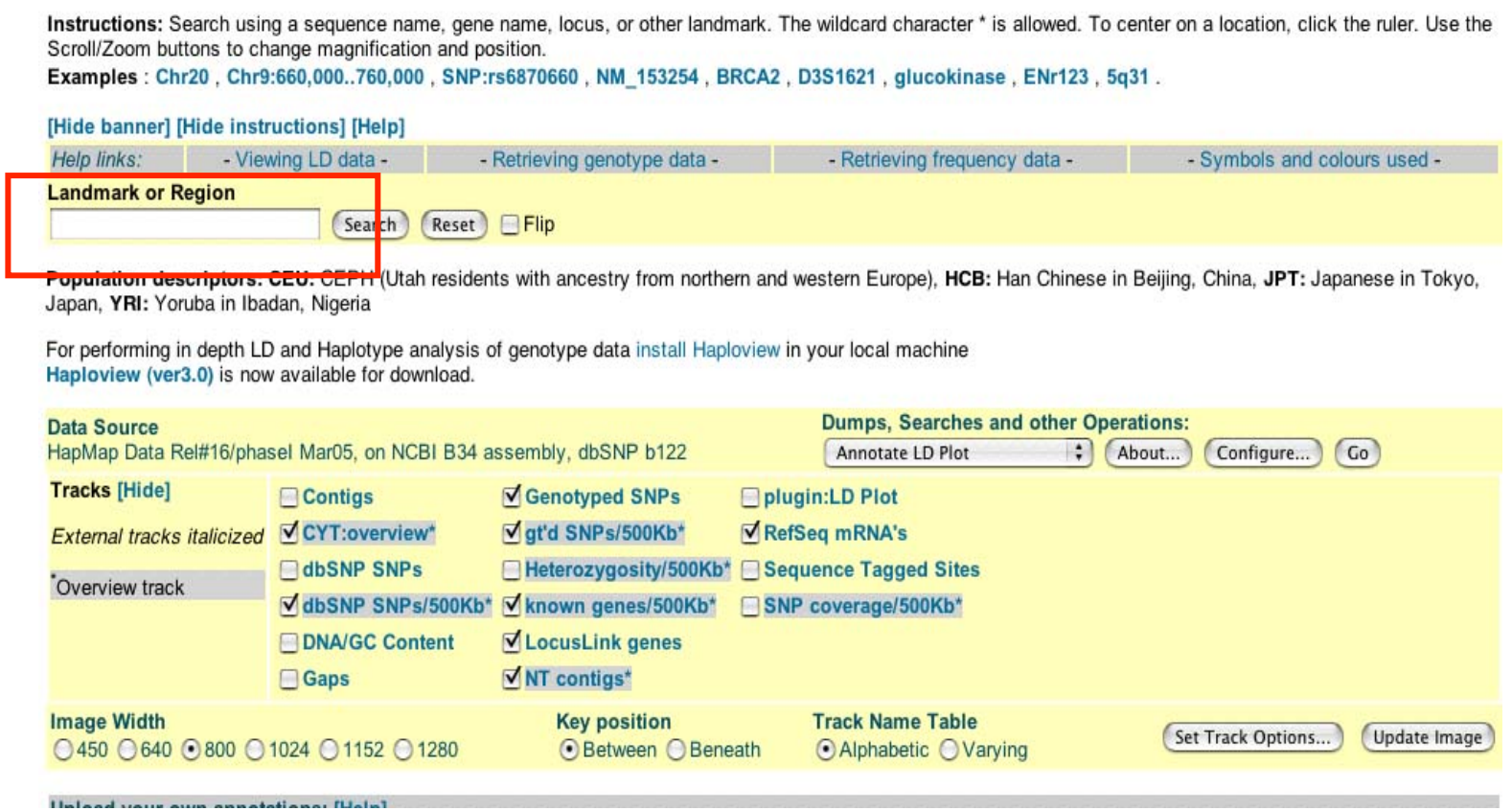
Recombination, mutation and selection

The International HapMap project

The International HapMap Project

- Phase I: 270 samples from four populations of Yoruban, Asian and European decent
  - 90 samples from US Utah population with European ancestry (30 CEPH trios)
  - 90 Yoruba samples (Nigeria, 30 trios)
  - 45 unrelated Japanese samples
  - 45 unrelated Chinese samples
- Phase II: Native American, Japanese, Kenyan, Mexican and Italian
  - Only sex and population membership known, no clinical phenotypes
  - Trios used to assess genotyping accuracy (Mendelian inheritance)
- Phase III: seven populations genotyped with >1.5 M SNPs


International HapMap Project
International HapMap Project
Home | About the Project | Data | Publications | Conference
中文 | English | Français | 日本語 | Yoruba
The International HapMap Project is a partnership of scientists and funding agencies from Canada, China, Japan, Nigeria, the United Kingdom and the United States to develop a public resource that will help researchers find genes associated with human disease and response to pharmaceuticals. See "About the International HapMap Project" for more information.
Project Information
About the Project
HapMap Publications
HapMap Conference
HapMap Mailing List
HapMap Project Participants
HapMap Mirror Site in Japan
Project Data
Browse Project Data
Bulk Data Download
Data Freezes for Publication
ENCODE Project
Guidelines For Data Use
Instructions: Search using a sequence name, gene name, locus, or other landmark. The wildcard character * is allowed. To center on a location, click the ruler. Use the Scroll/Zoom buttons to change magnification and position.
Examples : Chr20 , Chr9:660,000..760,000 , SNP:rs6870660 , NM_153254 , BRCA2 , D3S1621 , glucokinase , ENr123 , 5q31 .
[Hide banner] [Hide instructions] [Help]
Help links: - Viewing LD data - - Retrieving genotype data - - Retrieving frequency data - - Symbols and colours used -
Landmark or Region
Search
Reset
Flip
Population descriptors: CEU: CEPH (Utah residents with ancestry from northern and western Europe), HCB: Han Chinese in Beijing, China, JPT: Japanese in Tokyo, Japan, YRI: Yoruba in Ibadan, Nigeria
For performing in depth LD and Haplotype analysis of genotype data install Haploview in your local machine
Haploview (ver3.0) is now available for download.
Data Source
HapMap Data Rel#16/phasel Mar05, on NCBI B34 assembly, dbSNP b122
Dumps, Searches and other Operations:
Annotate LD Plot
Configure...
Tracks [Hide]
Contigs
Genotyped SNPs
plugin:LD Plot
External tracks italicized
CYT:overview*
gt'd SNPs/500Kb*
RefSeq mRNA's
Configure... LD Plot
Segment Size 500Kb
Genotype SNPs 200
Box Size Proportionate
LD Properties:
dprime
Color: Pairwise plot
red
Color:data not available
lightgrey
Populations:
CEU off on
HCB off on
JPT off on
YRI off on
Orientation:
normal
invert
normal
invert
Cancel
Configure
Home | About the Project | Data | Publications | Conference
Please send questions and comments on website to help@hapmap.org



Instructions: Search using a sequence name, gene name, locus, or other landmark. The wildcard character * is allowed. To center on a location, click the ruler. Use the Scroll/Zoom buttons to change magnification and position.
Examples : Chr20 , Chr9:660,000..760,000 , SNP:rs6870660 , NM_153254 , BRCA2 , D3S1621 , glucokinase , ENr123 , 5q31 .

[Hide banner] [Hide instructions] [Help]

Help links: - Viewing LD data - - Retrieving genotype data - - Retrieving frequency data - - Symbols and colours used -

Landmark or Region

Search Reset Flip

Population descriptors: CEU: CEPH (Utah residents with ancestry from northern and western Europe), **HCB:** Han Chinese in Beijing, China, **JPT:** Japanese in Tokyo, Japan, **YRI:** Yoruba in Ibadan, Nigeria

For performing in depth LD and Haplotype analysis of genotype data install Haploview in your local machine
Haploview (ver3.0) is now available for download.

Data Source
HapMap Data Rel#16/phaseI Mar05, on NCBI B34 assembly, dbSNP b122

Dumps, Searches and other Operations:
Annotate LD Plot About... Configure... Go

Tracks [Hide]

External tracks italicized

*Overview track

Contigs	Genotyped SNPs	plugin:LD Plot
CYT:overview*	gt'd SNPs/500Kb*	RefSeq mRNA's
dbSNP SNPs	Heterozygosity/500Kb*	Sequence Tagged Sites
dbSNP SNPs/500Kb*	known genes/500Kb*	SNP coverage/500Kb*
DNA/GC Content	LocusLink genes	
Gaps	NT contigs*	

Image Width
450 640 800 1024 1152 1280

Key position
Between Beneath

Track Name Table
Alphabetic Varying

Set Track Options... Update Image

Showing 399 bp from chr3, positions 69,947,470 to 69,947,868

Instructions: Search using a sequence name, gene name, locus, or other landmark. The wildcard character * is allowed. To center on a location, click the ruler. Use the Scroll/Zoom buttons to change magnification and position.
Examples : Chr20 , Chr9:660,000..760,000 , SNP:rs6870660 , NM_153254 , BRCA2 , 5q31 , ENm010 .

[Hide banner] [Hide instructions] [Bookmark this view] [Link to an image of this view] [Publication quality image] [Help]

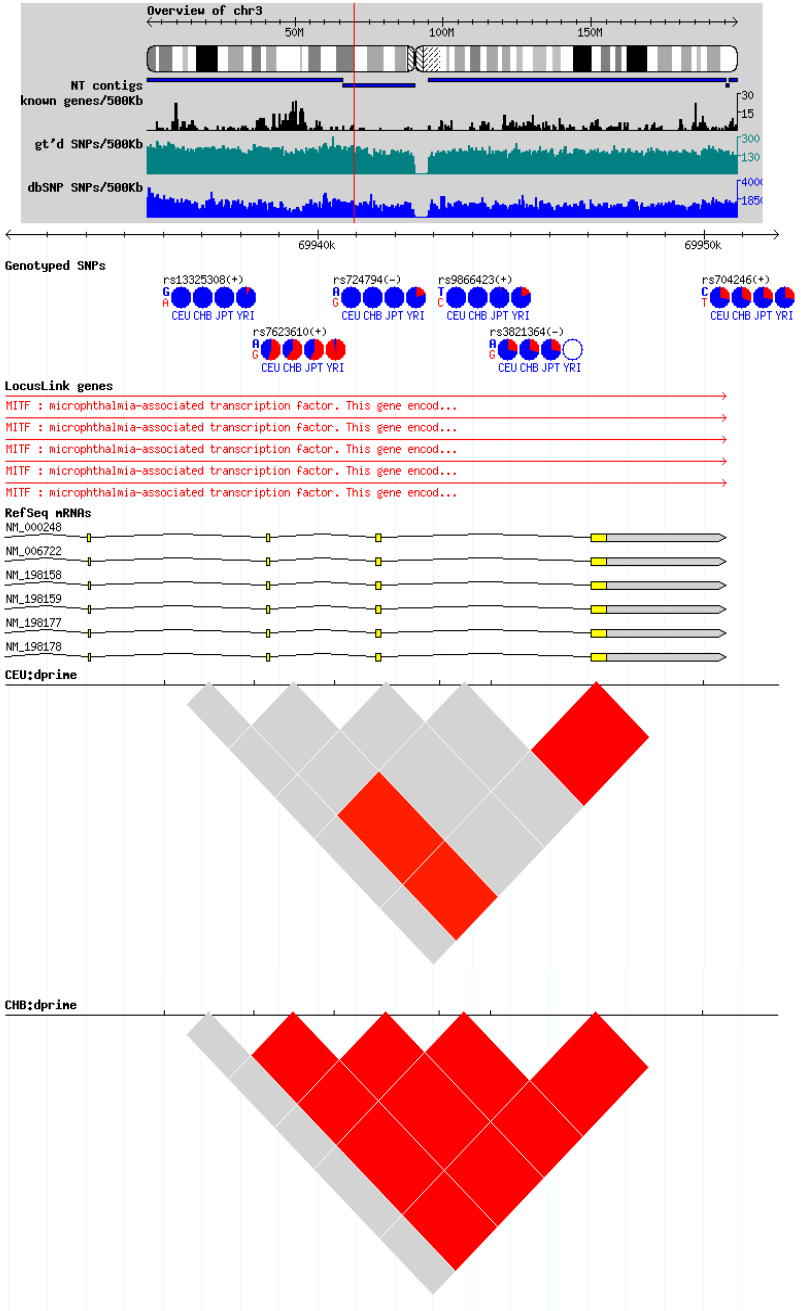
Help links: - Viewing LD data - - Retrieving genotype data - - Retrieving frequency data - - Symbols and colours used -

Landmark or Region

mitf Search Reset Flip

Scroll/Zoom:
Show 399 bp

Population descriptors:YRI: Yoruba in Ibadan, Nigeria, JPT: Japanese in Tokyo, Japan, CHB: Han Chinese in Beijing, China, CEU: CEPH (Utah residents with ancestry from northern and western Europe



refSNP rs7623610 with alleles A/G in dbSNP (dbSNP report | Ensembl SNPview)

Chr3:69938351..69938351, (+) strand relative to the human reference sequence

Population	Genotype frequencies: Ref-homozygote genotype	freq	count	Heterozygote genotype	freq	count	Other-homozygote genotype	freq	count	Total count	Allele frequencies: Ref-allele allele	freq	count	Other-allele allele	freq	count	Total count	
CEU	A/A	0.208	11	A/G	0.453	24	G/G	0.340	18	53	A	0.434	46	G	0.566	60	106	retrieve genotypes
CHB	A/A	0.111	5	A/G	0.533	24	G/G	0.356	16	45	A	0.378	34	G	0.622	56	90	retrieve genotypes
JPT	A/A	0.182	8	A/G	0.500	22	G/G	0.318	14	44	A	0.432	38	G	0.568	50	88	retrieve genotypes
YRI	A/A	0.018	1	A/G	0.018	1	G/G	0.965	55	57	A	0.026	3	G	0.974	111	114	retrieve genotypes

Note: the 'reference' allele is the base observed in the reference genome sequence at this location

Population descriptors:
YRI: Yoruba in Ibadan, Nigeria
JPT: Japanese in Tokyo, Japan
CHB: Han Chinese in Beijing, China
CEU: CEPH (Utah residents with ancestry from northern and western Europe)

Please see this page for more information about the populations, as well as a general discussion of the populations under study in the project.

Assay LSID	urn:lsid:chmo-h.hapmap.org:Assay:3Y25804:1
Protocol	urn:lsid:wicgr.hapmap.org:Protocol:assay_design_1:1 (Sequenom platform)
extension_probe	GCGAGGACATCCAACAATA
pcr_primer_forward2	ACGTTGGATGGGGTTAGGTTAGAATTTGGG
pcr_primer_reverse2	ACGTTGGATGTAGGGACTTGGCGAGGACAT
strand	reverse relative to dbSNP (minus relative to the human reference sequence)

Assay LSID	urn:lsid:chmo-h.hapmap.org:Assay:3q07817:1
Protocol	urn:lsid:wicgr.hapmap.org:Protocol:assay_design_1:1 (Sequenom platform)
extension_probe	ATTTGGGCTTTCACCAG
pcr_primer_forward2	ACGTTGGATGTTATTAGGGACTTGGCGAGG
pcr_primer_reverse2	ACGTTGGATGTGCTGACTTCTGCTCTAAGG
strand	forward relative to dbSNP (plus relative to the human reference sequence)

Assay LSID	urn:lsid:chmo-h.hapmap.org:Assay:3N25804:1
Protocol	urn:lsid:wicgr.hapmap.org:Protocol:assay_design_1:1 (Sequenom platform)
extension_probe	GCGAGGACATCCAACAATA
pcr_primer_forward2	ACGTTGGATGGGGTTAGGTTAGAATTTGGG
pcr_primer_reverse2	ACGTTGGATGTAGGGACTTGGCGAGGACAT
strand	reverse relative to dbSNP (minus relative to the human reference sequence)

Worked example 2: HapMap