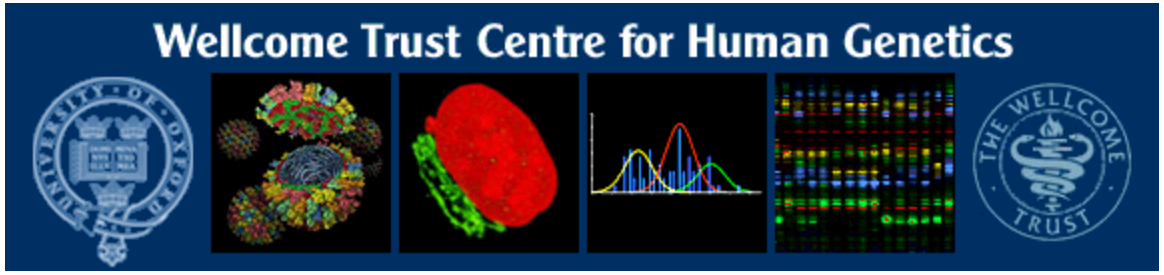
The HapMap Project and Haploview

David Evans

Ben Neale

University of Oxford

Wellcome Trust Centre for Human Genetics



Human Haplotype Map

- General Idea: C*haracterize the distribution of Linkage Disequilibrium across the genome.*
- Why?: Infeasible to type every polymorphism in the human genome => Because of LD, type a subset of variants that captures most of *common* variation in genome
- Output:
  - -- Raw genotype data freely available (monthly release)
  - -- www.hapmap.org
- Deliverables: Sets of haplotype tagging SNPs

Human Haplotype Map
- Funding -

- **Total US $120 million**
- Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT), Tokyo
- National Institutes of Health, US
- The Wellcome Trust, UK
- Genome Canada in Ottawa and Genome Quebec, Montreal
- Chinese Academy of Sciences, Chinese Ministry of Science and Technology, Natural Science Foundation of China, Beijing
- The SNP Consortium (TSC), US

Human Haplotype Map
- Participants -

- Genotyping
  - 25% RIKEN/Univ Tokyo (Nakamura)
  - 24% Sanger Institute (Bentley)
  - 16% Illumina (Chee)
  - 10% Genome Quebec (Hudson)
  - 10% Beijing/Shanghai/Hong Kong (Yang, Zeng, Huang, Tsui)
  - 9% Whitehead Institute (Altshuler)
  - 4% Baylor Coll Medicine, US (Gibbs)
  - 2% Univ Calif San Francisco (Kwok)
- Ethical, Legal, Social Issues
  - Japan (Matsuda)
  - China (Zhang, Zeng)
  - US (Leppert)
  - Nigeria (Rotimi)
- Samples
  - Nigeria (Yoruba; Ibadan)
    - 30 trios → 90 individuals
  - US (CEPH)
    - 30 trios → 90 individuals
  - China (Han)
    - 45 unrelateds
  - Japan (Tokyo)
    - 45 unrelateds
- Data Analysis
  - Whitehead (Altshuler, Daly)
  - Johns Hopkins Univ (Chakravarti, Cutler)
  - Oxford Statistics (Donnelly, McVean)
  - Oxford Genetics (Cardon, Weir, Abecasis)
- Data Coordination
  - Cold Spring Harbor (Stein)

Human Haplotype Map

Status March 2005

- "Phase I" complete
  - ~1 million SNPs typed in 270 individuals at an average spacing of 1 SNP per 5 KB
  - Study of data accuracy across centres (1,500 markers) revealed concordance, internal consistency > 99.8%
    - For several centres, accuracy > 99.9%
- "Phase II" underway
  - Type an additional 2.25 million SNPs in the same samples (~1 SNP per 1 KB)

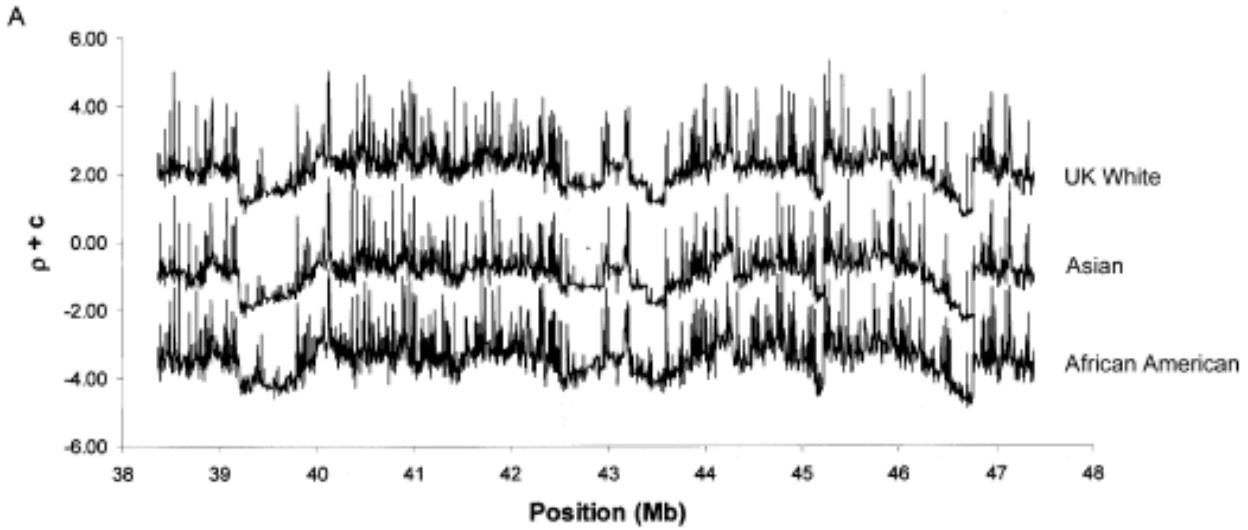
Encode Regions

- Resequence ten ~500KB regions in 16 CEPH, 16 Yoruba, 8 Japanese and 8 Chinese
- Genotype all dbSNPs and “new” SNPs in all 270 individuals

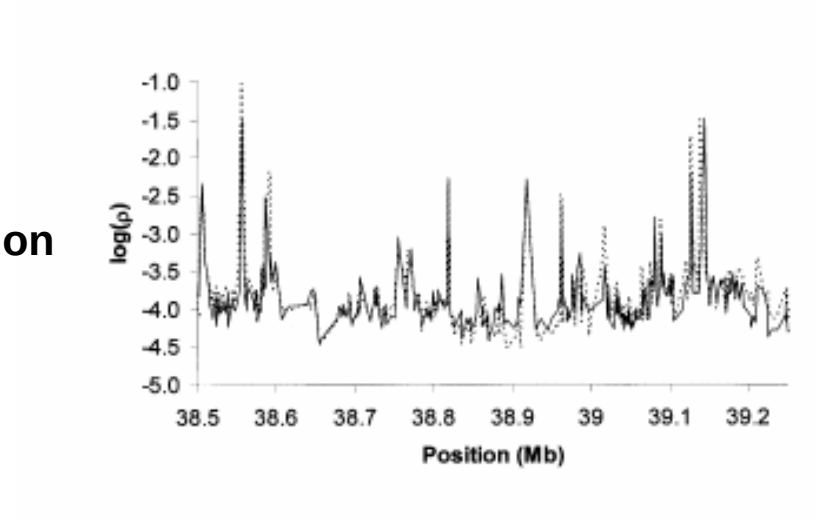
ENCODE Regions Genotype Information

Region name	Chromosome band	Genomic interval (NCBI)	Available SNPs		Genotyped SNPs								Genotyping group
			dbSNP	New SNPs	CEU		HCB		JPT		YRI		
					rs#	no rs#	rs#	no rs#	rs#	no rs#	rs#	no rs#	
ENr112	2p16.3	Chr2:51633239..52133238	1,624	1,720	1,064	937	867	900	868	900	879	922	McGill-GQIC, Perlegen
ENr131	2q37.1	Chr2:234778639..235278638	1,787	1,233	1,179	719	923	690	925	690	932	704	McGill-GQIC, Perlegen
ENr113	4q26	Chr4:118705475..119205474	1,516	1,819	1,017	1,614	878	1,589	878	1,589	879	1,597	Broad, Perlegen
ENm010	7p15.2	Chr7:26699793..27199792	1,274	1,857	757	459	291	500	291	500	284	456	UCSF-WU, Perlegen
ENm013	7q21.13	Chr7:89395718..89895717	1,545	1,713	927	1,382	740	1,393	740	1,393	748	1,391	Broad, Perlegen
ENm014	7q31.33	Chr7:126135436..126632577	1,354	1,562	963	1,428	794	1,417	794	1,417	800	1,419	Broad, Perlegen
ENr321	8q24.11	Chr8:118769628..119269627	1,468	1,682	936	905	726	907	726	907	713	903	Illumina, Perlegen
ENr232	9q34.11	Chr9:127061347..127561346	1,494	1,646	694	707	508	702	508	702	517	689	Illumina, Perlegen
ENr123	12q12	Chr12:38626477..39126476	1,904	1,551	859	0	80	0	78	0	74	0	BCM, Perlegen
ENr213	18q12.1	Chr18:23717221..24217220	1,391	1,465	809	820	643	816	643	817	644	819	Illumina, Perlegen
		Total	**15,357**	**16,248**	**9,205**	**8,971**	**6,450**	**8,914**	**6,451**	**8,915**	**6,470**	**8,900**	

Population Recombination Rate



Population Recombination Rate



US Caucasian vs UK Caucasian

- HapMap website:
  - www.hapmap.org
- Haploview website:
  - www.broad.mit.edu/mpg/haploview/index.php

Haploview

.ped file

```
1  1  0  0  1  2    1 2  0 0
1  2  0  0  2  2    1 2  3 3
1  3  0  0  1  2    1 1  1 1
1  4  1  2  2  2    1 2  0 0
1  5  3  4  2  2    1 1  1 1
1  6  3  4  1  2    2 2  1 3
```

.info file

```
rs1474567    38362947
rs2179083    38364233
```

Exercise

F:\davide\Boulder2005\hapmap

African dataset	af1.dat af1.ped
Caucasian dataset	cauc1.dat cauc1.ped

How many “blocks” are there in the Caucasian dataset?

Do the number and position of blocks vary according to whether the Gabriel et al or four gamete block definition is employed?

Choose a set of tagging SNPs for the Caucasian dataset to summarize the genotype data efficiently.

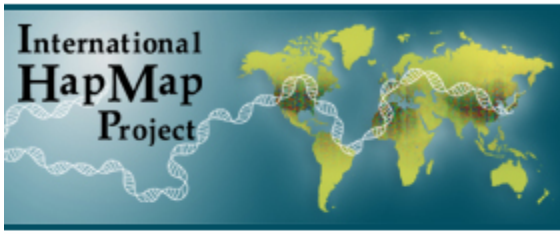
Do LD patterns vary between the Caucasian and African datasets? Why?

Edit View Go Bookmarks Tools Help

http://www.hapmap.org/

Go

Getting Started Latest Headlines Google



International HapMap Project

Home | About the Project | Data | Publications | Conference

ζ | English | Français | 日本語 | Yoruba

International HapMap Project is a partnership of scientists and funding agencies from Canada, China, Japan, Nigeria, the United Kingdom and the United States to elop a public resource that will help researchers find genes associated with human disease and response to pharmaceuticals. See "**About the International HapMap ject**" for more information.

oject Information

- out the Project
- oMap Publications
- oMap Conference
- oMap Mailing List
- oMap Project Participants
- oMap Mirror Site in Japan

oject Data

- neric Genome Browser
- k Data Download
- CODE Project
- idelines For Data Use

eful Links

- oMap Project Press Release
- GRI HapMap Page
- BI Variation Database (dbSNP)

News

- 2005-03-01: **HapMap public release #16.**
  ATTN: This the so-called Phase I data freeze which marks a major milestone of the project: a genotyped common SNP every 5Kb in all populations under study. Data available for **bulk download** and **graphical browsing**. Summary of genotyped SNPs:

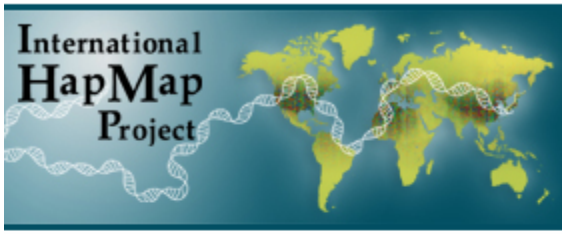
Populations	CEU	HCB	JPT	YRI
Genotyped SNPs	1,073,663	1,044,686	1,044,416	1,034,205

- 2005-02-08: **HapMap News Volume 1, 2004**
  This is the first in a series of newsletters to be published by the Coriell Institute for Medical Research to inform communities how their samples are being used. Each issue of the newsletter will be available in the primary languages of all the participating communities.
- 2005-02-07: **International HapMap Consortium Expands Mapping Effort**
  The International HapMap Consortium, boosted by an additional 3.3 million in public-private support, announces plans to create an even more powerful map of human genetic variation than originally envisioned. The map will accelerate the discovery of genes related to common diseases, such as asthma, cancer, diabetes and heart disease.
- ***Old News***

Edit View Go Bookmarks Tools Help

http://www.hapmap.org/downloads/index.html.en

Getting Started Latest Headlines Google



International HapMap Project

Home | About the Project | Data | Publications | Conference

ζ | English | Français | 日本語 | Yoruba

delines for the Responsible Use and Publication of HapMap data is available **here**.

wse data graphically

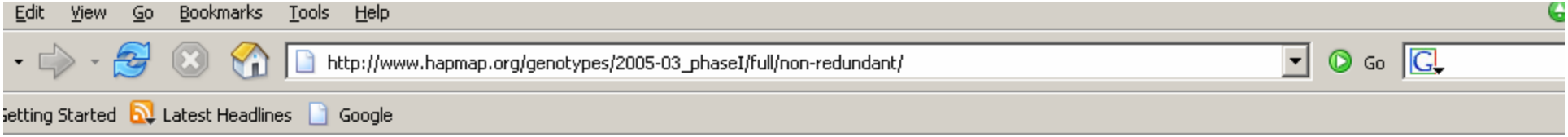
the **Generic Genome Browser** to view HapMap Project data in the context of other genomic features, as well as retrieve genotypes & frequencies for specific genomi
ons. Jump directly to chromosome —— Submit or see an **overview of all chromosomes**.

k data downloads

following directories contain project data that have been made publicly available. (See **HapMap Data Access Policy** for more information). More details about each set can be found in READMEs in the respective directories:

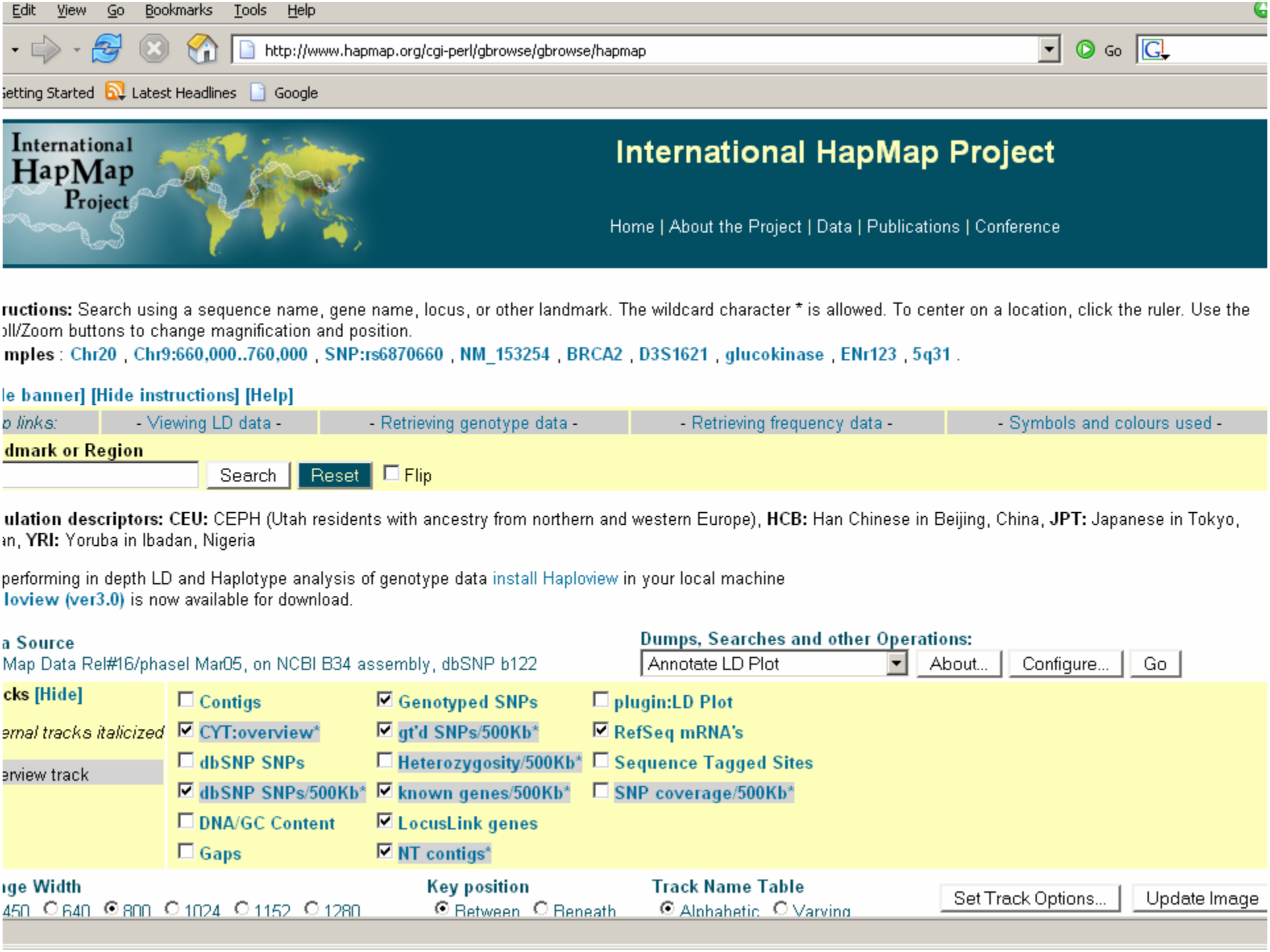
- **Genotypes**: Individual genotype data submitted to the DCC to date.
- **LD Data**: Linkage disequilibrium properties D', LOD , R^2 compiled from the genotype data to date
- **Allocated SNPs**: dbSNP reference SNP clusters that have been picked and prioritized for genotyping according to several criteria (see info on **how SNPs were selected**). The file 00README contains per-chromosome SNP counts and further details.
- **Frequencies**: Allele & genotype frequencies compiled from genotyping data submitted to the DCC to date. These have also been submitted to **dbSNP** and should be available in the next dbSNP build.
- **SNP assays**: Details about assays submitted to the DCC to date. PCR primers, extension probes etc., specific to each genotyping platform.
- **Protocols**: Information on assay design, genotyping and other protocols used in the project.
- **Samples/individuals**: Information on the samples used in the project and the individuals from which they were drawn. (See **About the project: Which Population are Being Sampled**).
- **XML docs**: Documentation on the XML format used in the project.

CODE regions



ndex of /genotypes/2005-03_phaseI/full/non-redundant

Name	Last modified	Size
Parent Directory	01-Mar-2005 16:40	-
00md5 checksums	01-Mar-2005 17:06	5k
genotypes_chr10_CEU.txt.gz	01-Mar-2005 16:41	2.0M
genotypes_chr10_HCB.txt.gz	01-Mar-2005 16:41	1.4M
genotypes_chr10_JPT.txt.gz	01-Mar-2005 16:41	1.4M
genotypes_chr10_YRI.txt.gz	01-Mar-2005 16:41	2.0M
genotypes_chr11_CEU.txt.gz	01-Mar-2005 16:41	1.8M
genotypes_chr11_HCB.txt.gz	01-Mar-2005 16:41	1.2M
genotypes_chr11_JPT.txt.gz	01-Mar-2005 16:41	1.2M
genotypes_chr11_YRI.txt.gz	01-Mar-2005 16:41	1.7M
genotypes_chr12_CEU.txt.gz	01-Mar-2005 16:41	2.0M
genotypes_chr12_HCB.txt.gz	01-Mar-2005 16:41	1.3M
genotypes_chr12_JPT.txt.gz	01-Mar-2005 16:41	1.3M
genotypes_chr12_YRI.txt.gz	01-Mar-2005 16:41	2.0M
genotypes_chr13_CEU.txt.gz	01-Mar-2005 16:41	1.4M
genotypes_chr13_HCB.txt.gz	01-Mar-2005 16:41	1.0M
genotypes_chr13_JPT.txt.gz	01-Mar-2005 16:41	1.0M
genotypes_chr13_YRI.txt.gz	01-Mar-2005 16:41	1.5M
genotypes_chr14_CEU.txt.gz	01-Mar-2005 16:41	1.2M
genotypes_chr14_HCB.txt.gz	01-Mar-2005 16:41	803k
genotypes_chr14_JPT.txt.gz	01-Mar-2005 16:41	790k
genotypes_chr14_YRI.txt.gz	01-Mar-2005 16:41	1.1M
genotypes_chr15_CEU.txt.gz	01-Mar-2005 16:41	1.1M
genotypes_chr15_HCB.txt.gz	01-Mar-2005 16:41	725k
genotypes_chr15_JPT.txt.gz	01-Mar-2005 16:41	714k
genotypes_chr15_YRI.txt.gz	01-Mar-2005 16:41	1.0M


Edit View Go Bookmarks Tools Help
http://www.hapmap.org/cgi-perl/gbrowse/gbrowse/hapmap
Go
Getting Started
Latest Headlines
Google
International HapMap Project
International HapMap Project
Home | About the Project | Data | Publications | Conference
ructions: 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 ll/Zoom buttons to change magnification and position.
mples : Chr20 , Chr9:660,000..760,000 , SNP:rs6870660 , NM_153254 , BRCA2 , D3S1621 , glucokinase , ENr123 , 5q31 .
le banner] [Hide instructions] [Help]
p links: - Viewing LD data - - Retrieving genotype data - - Retrieving frequency data - - Symbols and colours used -
dmark or Region
Search
Reset
Flip
ulation descriptors: CEU: CEPH (Utah residents with ancestry from northern and western Europe), HCB: Han Chinese in Beijing, China, JPT: Japanese in Tokyo, an, YRI: Yoruba in Ibadan, Nigeria
performing in depth LD and Haplotype analysis of genotype data install Haploview in your local machine
loview (ver3.0) is now available for download.
a Source
Map Data Rel#16/phaseI Mar05, on NCBI B34 assembly, dbSNP b122
Dumps, Searches and other Operations:
Annotate LD Plot
About...
Configure...
Go
cks [Hide]
ernal tracks italicized
erview 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*
ge Width
450 640 800 1024 1152 1280
Key position
Between Beneath
Track Name Table
Alphabetic Varying
Set Track Options...
Update Image

Edit View Go Bookmarks Tools Help

http://www.hapmap.org/cgi-perl/gbrowse/gbrowse/hapmap

Go

Getting Started Latest Headlines Google

owing 100 kbp from chr19, positions 200,000 to 300,000

ructions: 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 ll/Zoom buttons to change magnification and position.

mples : Chr20 , Chr9:660,000..760,000 , SNP:rs6870660 , NM_153254 , BRCA2 , D3S1621 , glucokinase , ENr123 , 5q31 .

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

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

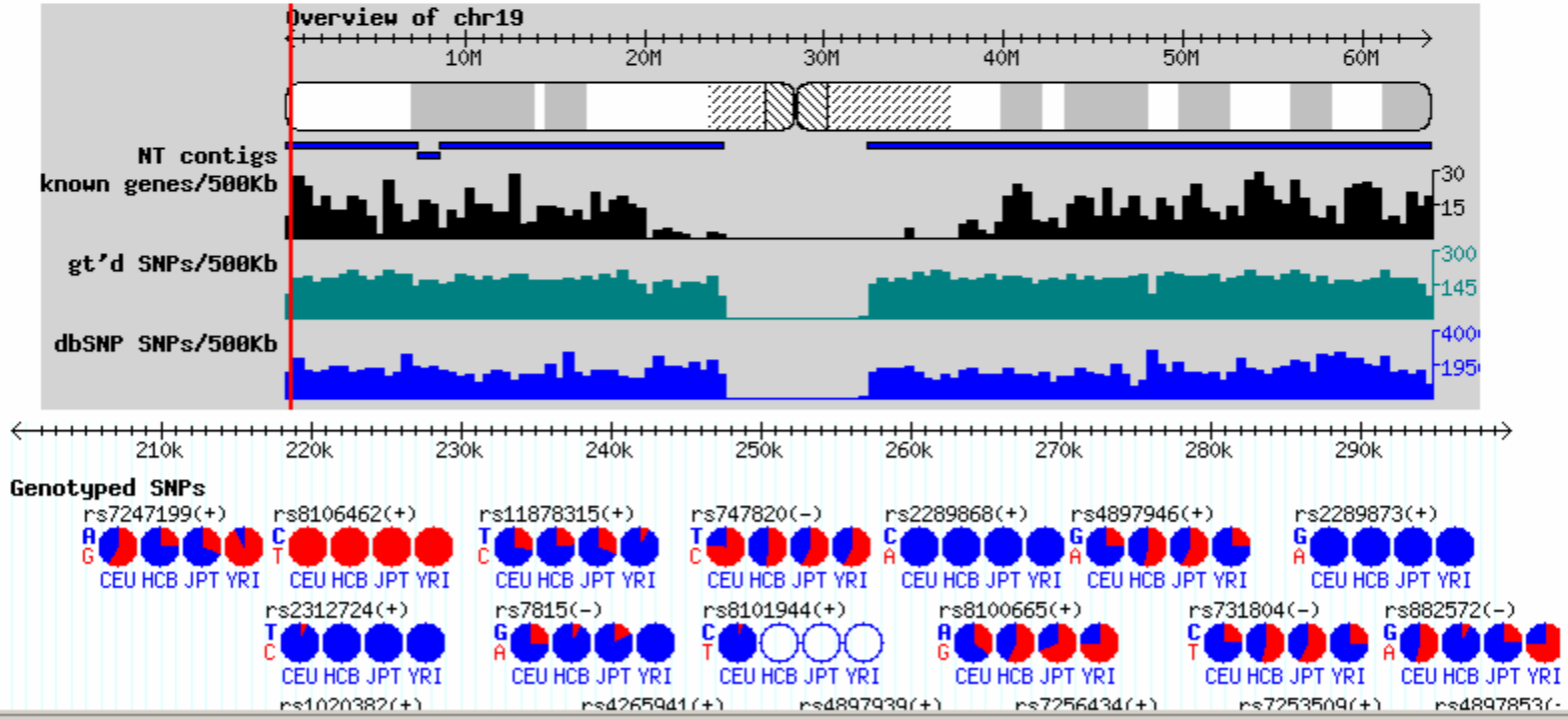
dmark or Region

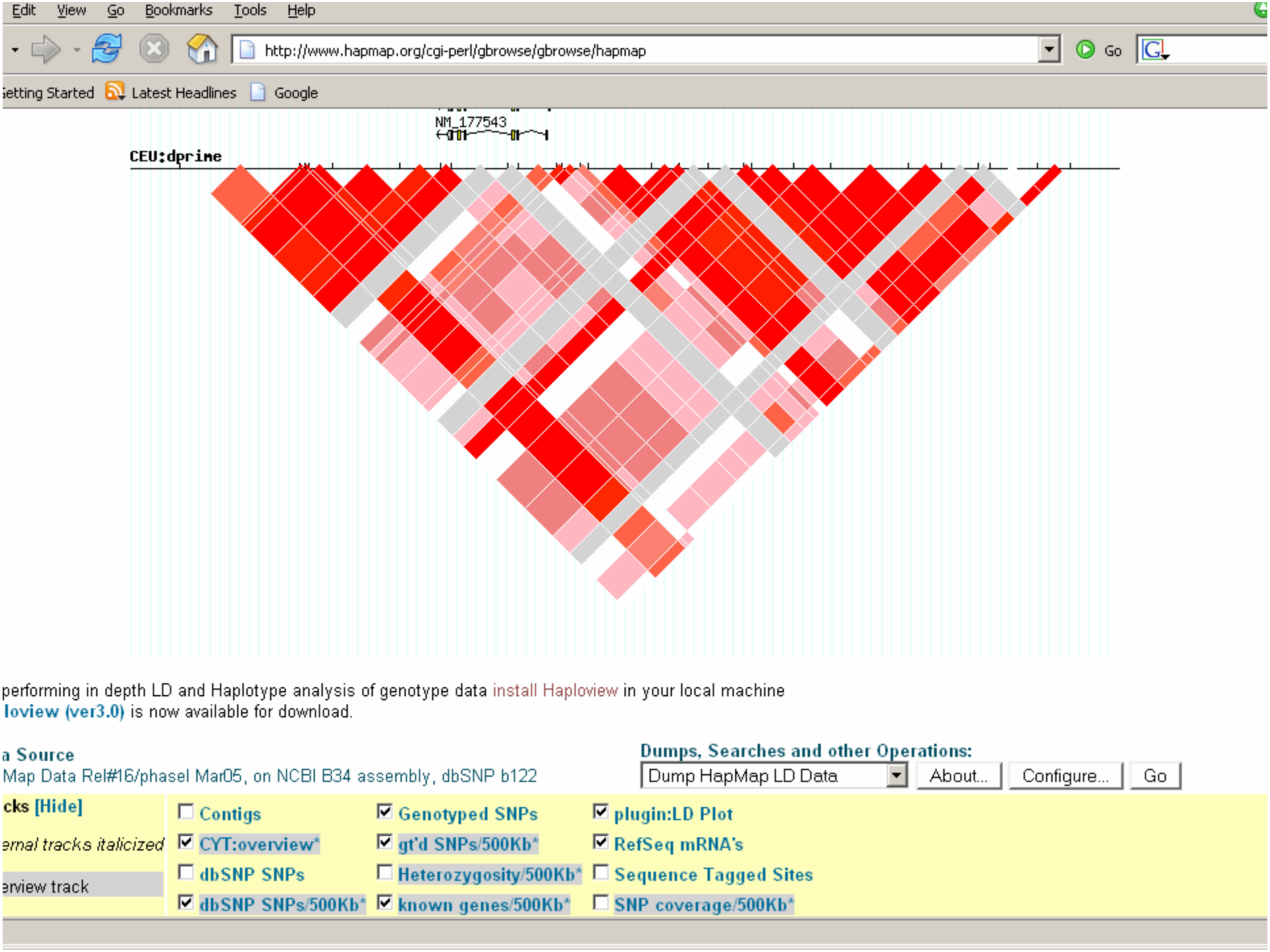
19:200000..300000 Search Reset Flip

Scroll/Zoom:

Show 100 kbp

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




Edit View Go Bookmarks Tools Help
http://www.hapmap.org/cgi-perl/gbrowse/gbrowse/hapmap
Go
Getting Started
Latest Headlines
Google
NM_177543
CEU:dprime
performing in depth LD and Haplotype analysis of genotype data install Haploview in your local machine
loview (ver3.0) is now available for download.
a Source
Map Data Rel#16/phase1 Mar05, on NCBI B34 assembly, dbSNP b122
Dumps, Searches and other Operations:
Dump HapMap LD Data
About...
Configure...
Go
cks [Hide]
ernal tracks italicized
erview 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*