hzAnalyzer: Detection, quantification, and visualization of contiguous homozygosity in human populations from high-density genotyping datasets using R and Java

Todd A. Johnson

RIKEN Center for Genomic Medicine

Tokyo Medical & Dental University

R User Conference - July 9, 2009

Homozygosity?

- Humans are diploid organisms, which means we each have two homologous chromosomes
- For a polymorphic locus that is bi-allelic, two alleles labeled A and a can be:
  - homozygous AA or aa
  - Heterozygous Aa
- We can recode:
  - AA and aa as 1
  - Aa as 0

  A contiguous homozygous segment then would be the red 1's in the following: 0111111111101011011

  Of course segments with 1, 2, or 3 homozygous loci is not so important, but other longer runs may be interesting…

International HapMap Project

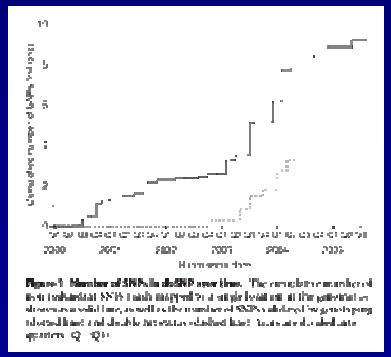
Vol 437|27 October 2005|doi:10.1038/nature04226 nature

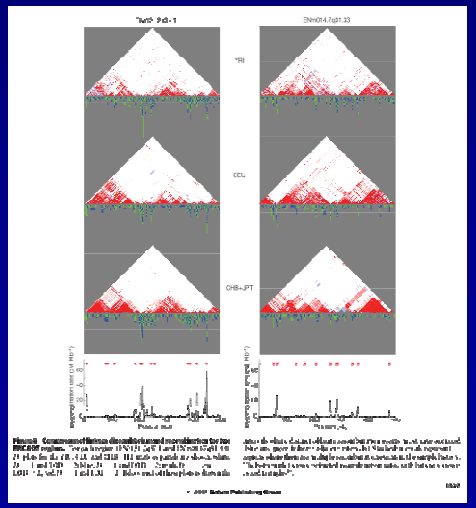
ARTICLES

A haplotype map of the human genome

The International HapMap Consortium*

Inherited genetic variation has a critical but as yet largely uncharacterized role in human disease. Here we report a public database of common variation in the human genome: more than one million single nucleotide polymorphisms (SNPs) for which accurate and complete genotypes have been obtained in 269 DNA samples from four populations, including ten 500-kilobase regions in which essentially all information about common DNA variation has been extracted. These data document the generality of recombination hotspots, a block-like structure of linkage disequilibrium and low haplotype diversity, leading to substantial correlations of SNPs with many of their neighbours. We show how the HapMap resource can guide the design and analysis of genetic association studies, shed light on structural variation and recombination, and identify loci that may have been subject to natural selection during human evolution.





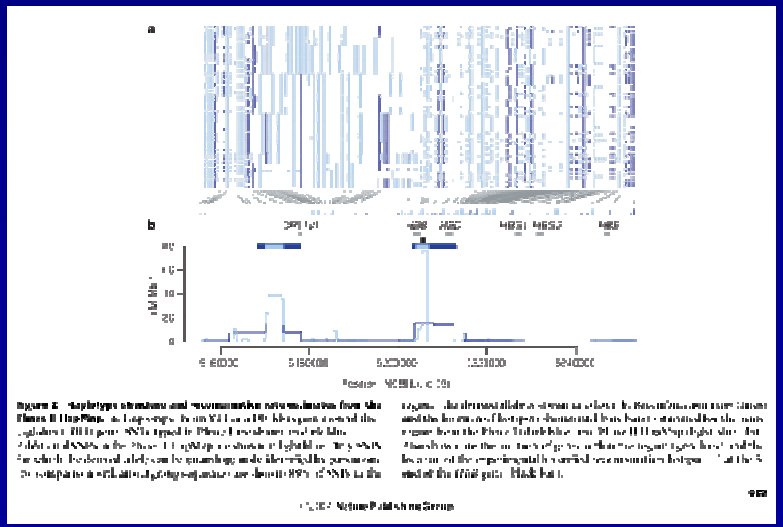
Vol 449|18 October 2007|doi:10.1038/nature06258 nature

ARTICLES

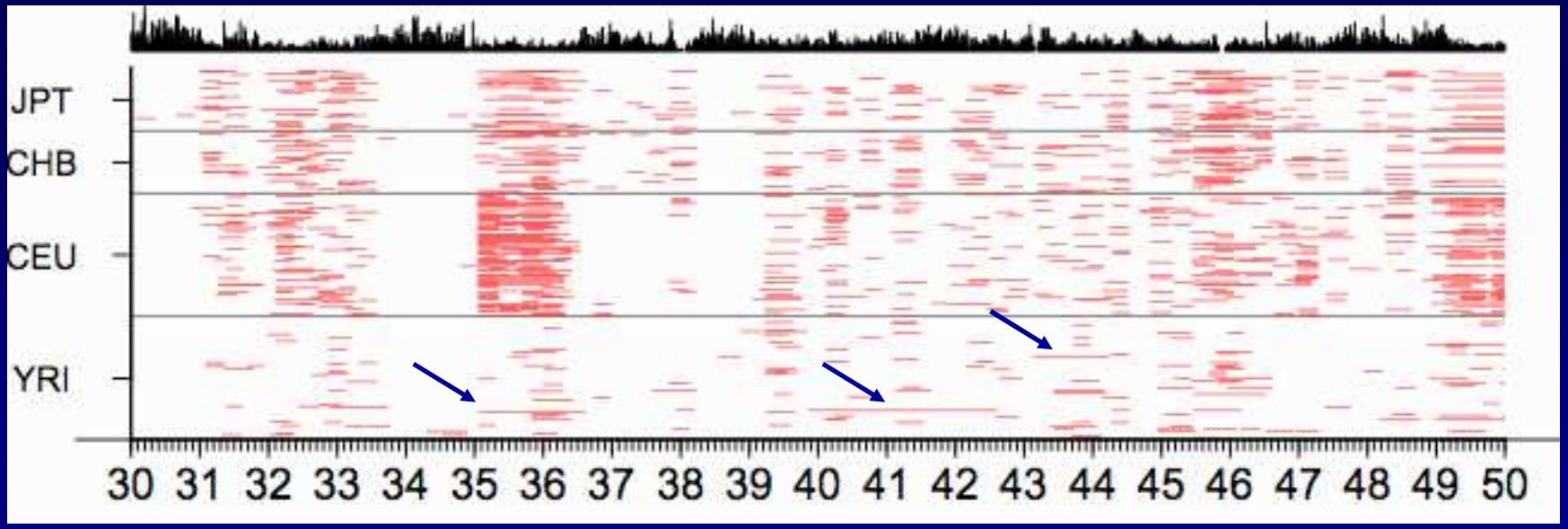
A second generation human haplotype map of over 3.1 million SNPs

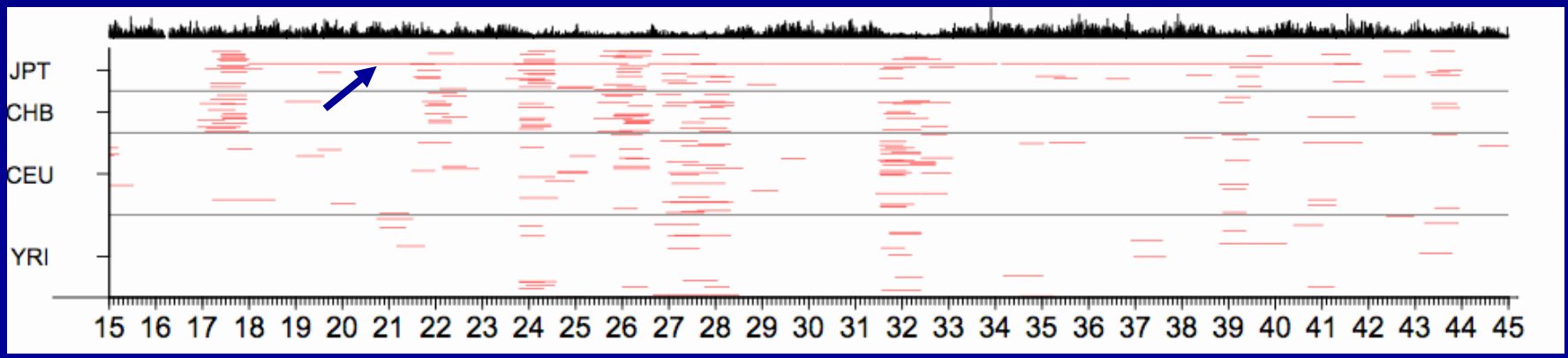
The International HapMap Consortium*

We describe the Phase II HapMap, which characterizes over 3.1 million human single nucleotide polymorphisms (SNPs) genotyped in 270 individuals from four geographically diverse populations and includes 25–35% of common SNP variation in the populations surveyed. The map is estimated to capture untyped common variation with an average maximum r^2 of between 0.9 and 0.96 depending on population. We demonstrate that the current generation of commercial genome-wide genotyping products captures common Phase II SNPs with an average maximum r^2 of up to 0.8 in African and up to 0.95 in non-African populations, and that potential gains in power in association studies can be obtained through imputation. These data also reveal novel aspects of the structure of linkage disequilibrium. We show that 10–30% of pairs of individuals within a population share at least one region of extended genetic identity arising from recent ancestry and that up to 1% of all common variants are untaggable, primarily because they lie within recombination hotspots. We show that recombination rates vary systematically around genes and between genes of different function. Finally, we demonstrate increased differentiation at non-synonymous, compared to synonymous, SNPs, resulting from systematic differences in the strength or efficacy of natural selection between populations.



Contiguous homozygous segments in two regions of HapMap sample data





Position (Mb)

Detection of homozygous segments

- hzAnalyzer incorporates a heuristic multi-step algorithm which was used to detect segments of contiguous homozygous loci within the 269 HapMap Phase 2 samples
  - 3,040,424 loci genome-wide SNPs
  - 2,956,629 autosomal loci
- Data processing
  - Minor allele frequency >0.01 in at least one population
  - Removed loci that intersected with copy-number variable regions, Ig $V_H/V_\kappa/V_\lambda$,segment duplications

Detection algorithm

- *snpMatrix*
  - Bioconductor package with excellent file input routines, compact binary data representation, and genotype/sample summary methods for storing and manipulating genotype data.
- Homozygous detection is run in a Java process that instantiates classes for:
  - Sample organization
    - Samplegroup
    - Individual with mother/father relationship info when appropriate
  - Data representation
    - Genotypes
    - Haplotypes
    - Segments of zygosity
  - Data processing
    - Instantiation of group, individual, genotype objects
    - Segment detection function

Detection algorithm

- Basic homozygous segment detection
  - Detect runs of homozygous loci allowing no-call genotypes but split at gaps>14kb

↓

Neighbor joining across regions of low SNP density
- Join segments A & B if:
  - A & B and combined segment A+B > 0.2 SNP/kb
  - A & B have length greater than 0.1*gap_size

  Or if A>0.1*gap_size but not B then scan past B and see if the addition of subsequent segments passes length and SNP density thresholds

↓

Modeling segments with low levels of heterozygosity
- Join segment HOM_A, HET_B, and HOM_C if:
  - $Freq_{HOMA+HETB}$<0.6% & $Freq_{HETB+HOMC}$<0.6%
  - Or if only $Freq_{HOMA+HETB}$<0.6% then scan past C and see if the addition of subsequent segments passes heterozygosity, length, and SNP density thresholds

Filtering terminology

- Homozygosity probability score (HPS)
  - Simple procedure
    - Measure the proportion of observed homozygous loci within a population for each SNP
      - $Freq_{HOMin}$ = frequency of homozygous genotypes within population
      - $Freq_{HOMex}$ = lowest frequency of homozygous genotypes across examine populations
      - HPS_{in} = Product of $Freq_{HOMin}$ for loci within a segment
      - HPS_{ex} = Product of $Freq_{HOMex}$ for loci within a segment
  - Goal is that each segment has some relative likelihood of being really homozygous based on the number of loci that are examined and each loci's heterozygosity.

Filtering terminology

- Minimum inclusive segment length (MISL)
  - Simple procedure
    - Find the maximum length segment (Max_L)in each individual
    - Find the minimum Max_L across the individuals
  - Depending upon sample populations or specific analysis, can choose subsets of groups or chromosomes
    - $MISL_{gw}$ = genome-wide
    - $MISL_{chr}$ = different value for each chromosome
    - $MISL_{chrn,n+1,...}$ = between a group of chromosomes

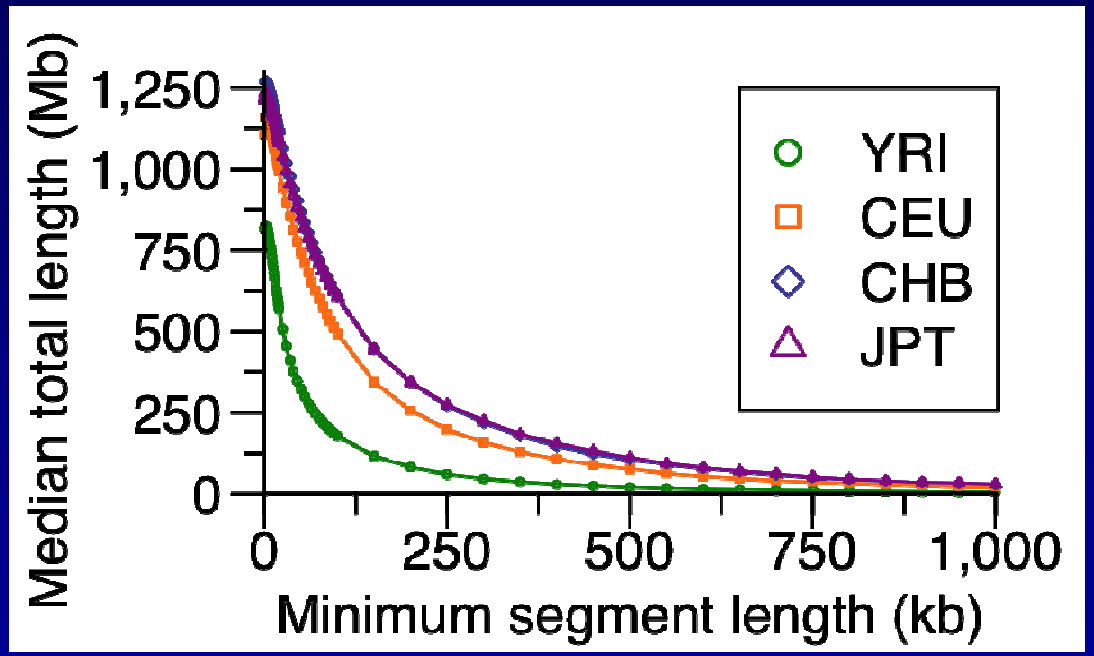
Chrom.	$MISL_{chr}$
1	391,555
2	385,789
3	400,822
4	355,550
5	264,726
6	309,973
7	308,518
8	315,796
9	228,061
10	229,520
11	293,727
12	311,633
13	248,643
14	268,112
15	242,482
16	239,646
17	270,268
18	179,120
19	270,633
20	168,531
21	131,431
22	155,041
X	457,502

Total length of homozygous segments in HapMap populations

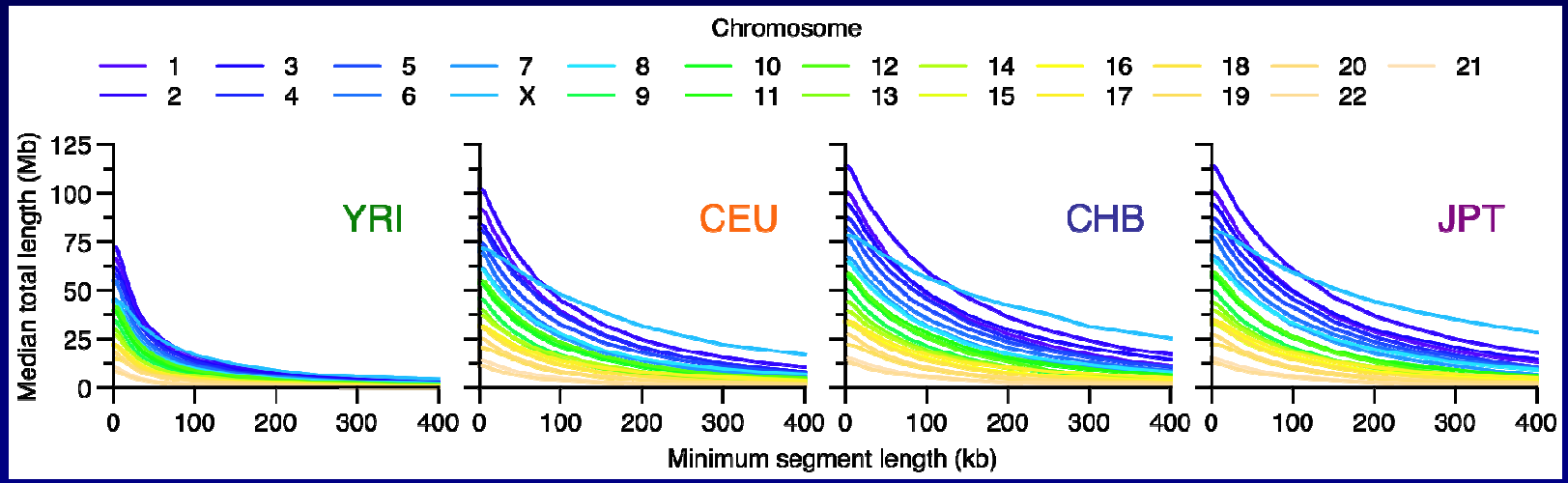
	Total length of homozygous segments (Total SNP count)		
Population	HPS_{in}<0.01	HPS_{ex}<0.01	HPS_{ex}<0.01, >=$MISL_{gw}$
YRI	0.67 x 10^9 (0.8 x10^6)	0.85 x 10^9 (1.0 x10^6)	0.15 x 10^9 (0.13 x10^6)
CEU	0.98 x 10^9 (1.1 x10^6)	1.15 x 10^9 (1.31 x10^6)	0.40 x 10^9 (0.37 x10^6)
CHB	1.06 x 10^9 (1.2 x10^6)	1.25 x 10^9 (1.42 x10^6)	0.50 x 10^9 (0.46 x10^6)
JPT	1.07 x 10^9 (1.2 x10^6)	1.27 x 10^9 (1.43 x10^6)	0.52 x 10^9 (0.48 x10^6)

Extended homozygosity on autosomes

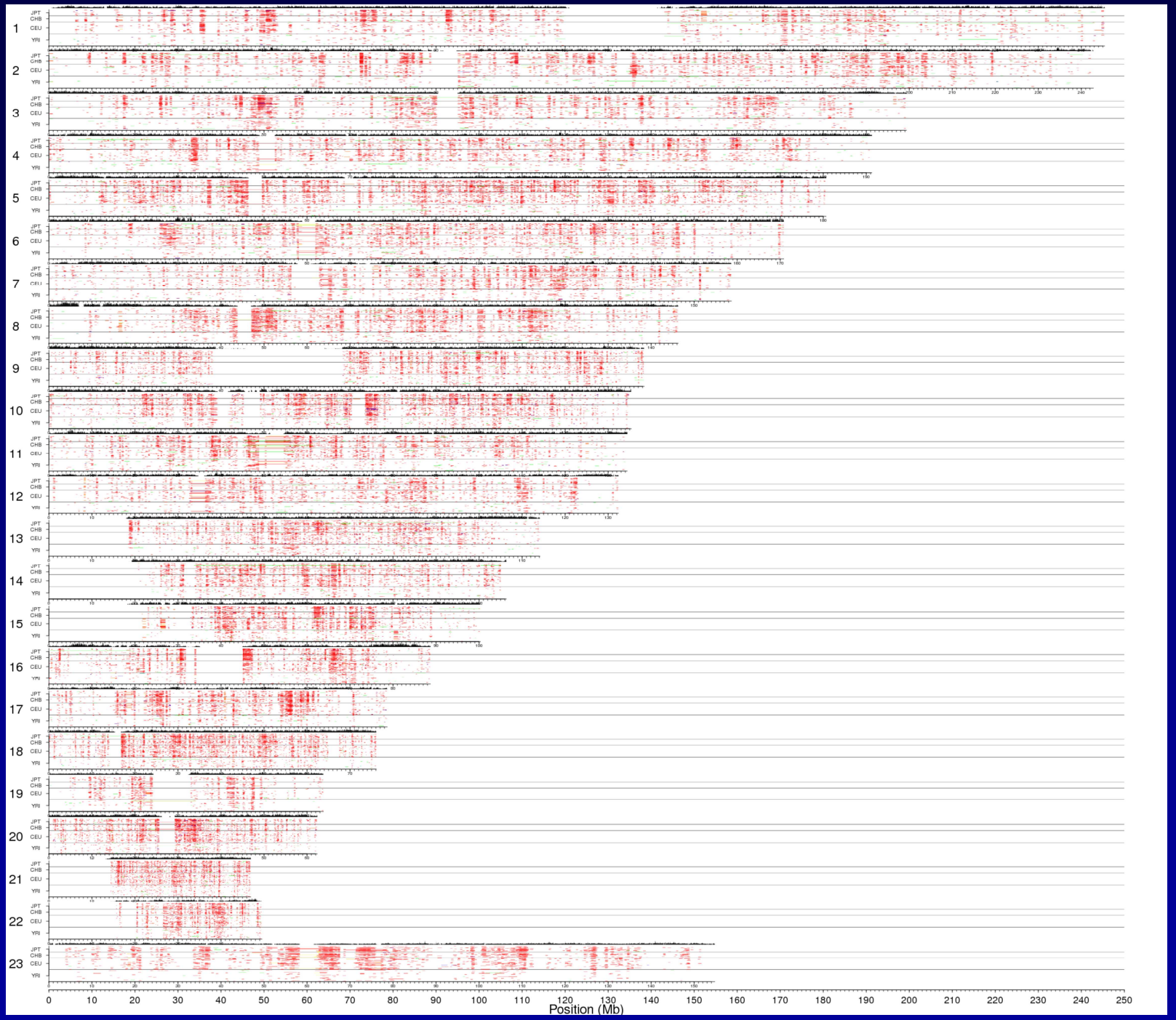
- YRI population shows much lower levels of contiguous homozygosity across all examined segment lengths as compared to the other three populations.



Distribution of homozygous segments on Chromosome X differs markedly from autosomes



	Median total length >=MISL$_{chr7,8,X}$			Chr.X median total length relative to:	
Population	Chr. 7	Chr. 8	Chr. X	Chr.7	Chr.8
YRI	2.4 x10^6	2.6 x10^6	5.7 x10^6	239%	220%
CEU	7.5 x10^6	9.3 x10^6	21.5 x10^6	286%	230%
CHB	10.1 x10^6	11.2 x10^6	31.0 x10^6	307%	277%
JPT	10.7 x10^6	12.6 x10^6	34.4 x10^6	322%	273%

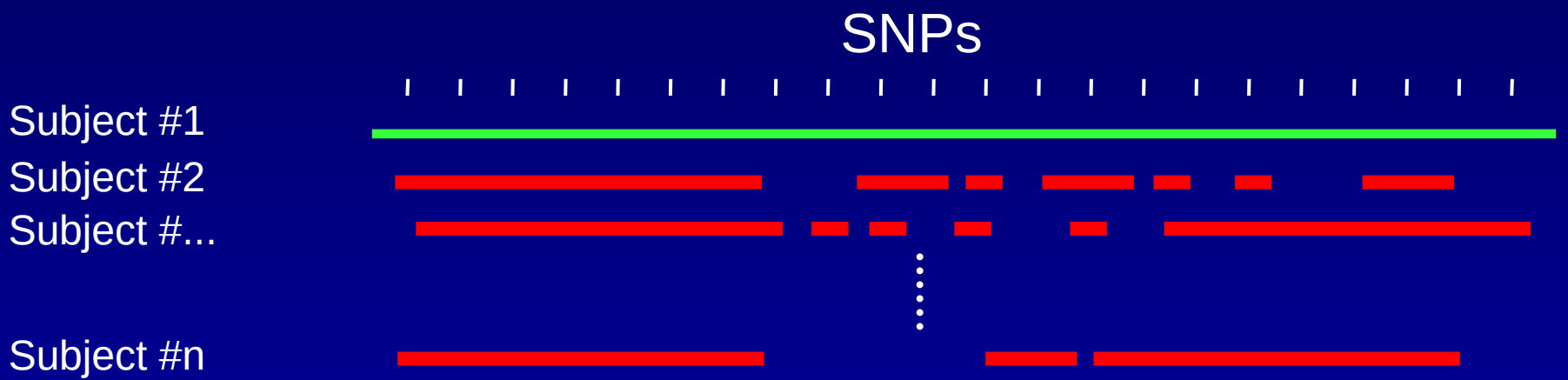

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
JPT
CHB
CEU
YRI
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240 250
Position (Mb)

How do we make sense out of all of those overlapping segments?

-> Develop a measure to quantify local variation of homozygous extent and relative population frequency.

Percentile-Extent matrix (PE_{mat}) derivation

- Tabulate for each locus the length of intersecting homozygous segments

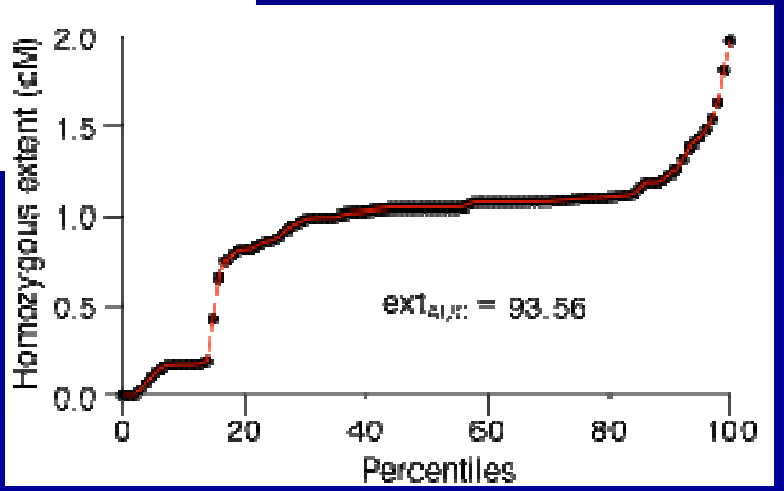


For each SNP, determine the percentile distribution of the lengths of any intersecting segments

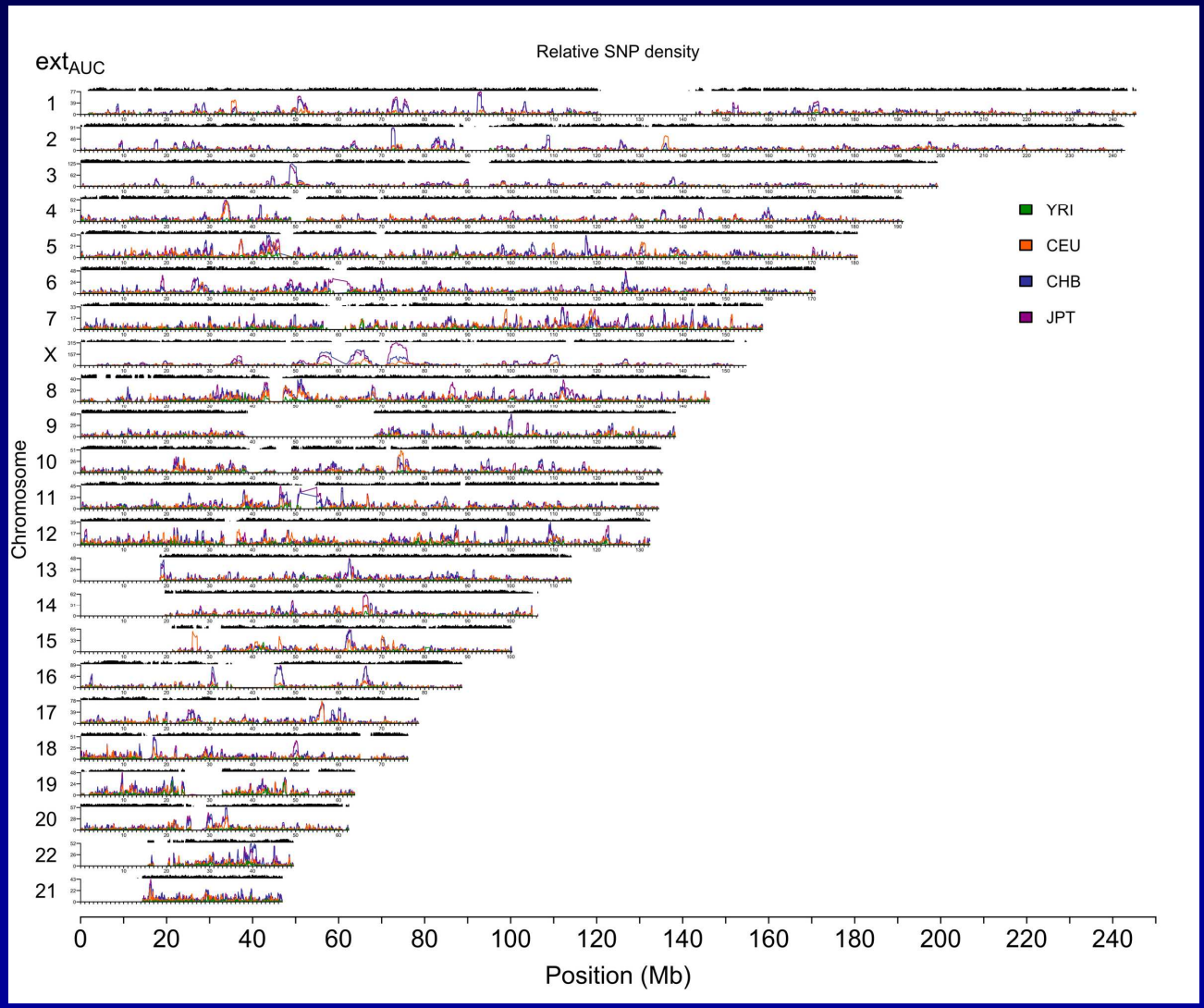
Deriving a locus-wise measure of homozygous extent from PE_{mat}

position	72,299,194	72,299,266	72,299,875	72,300,989	72,301,060	72,301,225	72,302,115	72,302,559	72,305,454	72,305,683	72,306,329	72,306,404	72,306,548
100	1.9778	1.9778	1.9778	1.9778	1.9778	1.9778	1.9778	1.9778	1.9778	1.9778	1.9778	1.9778	1.9778
99	1.8063	1.8063	1.8063	1.8063	1.8063	1.8063	1.8063	1.8063	1.8063	1.8063	1.8063	1.8063	1.8063
98	1.6349	1.6349	1.6349	1.6349	1.6349	1.6349	1.6349	1.6349	1.6349	1.6349	1.6349	1.6349	1.6349
97	1.5396	1.5396	1.5396	1.5396	1.5396	1.5396	1.5396	1.5396	1.5396	1.5396	1.5396	1.5396	1.5396
96	1.4811	1.4811	1.4811	1.4811	1.4811	1.4811	1.4811	1.4811	1.4811	1.4811	1.4811	1.4811	1.4811
95	1.4337	1.4337	1.4337	1.4337	1.4337	1.4337	1.4337	1.4337	1.4337	1.4337	1.4337	1.4337	1.4337
94	1.4071	1.4071	1.4071	1.4071	1.4071	1.4071	1.4071	1.4071	1.4071	1.4071	1.4071	1.4071	1.4071
93	1.3797	1.3797	1.3797	1.3797	1.3797	1.3797	1.3797	1.3797	1.3797	1.3797	1.3797	1.3797	1.3797
92	1.3213	1.3213	1.3213	1.3213	1.3213	1.3213	1.3213	1.3213	1.3213	1.3213	1.3213	1.3213	1.3213
91	1.2630	1.2630	1.2630	1.2630	1.2630	1.2630	1.2630	1.2630	1.2630	1.2630	1.2630	1.2630	1.2630
90	1.2271	1.2271	1.2271	1.2271	1.2271	1.2271	1.2271	1.2271	1.2271	1.2271	1.2271	1.2271	1.2271
89	1.2009	1.2009	1.2009	1.2009	1.2009	1.2009	1.2009	1.2009	1.2009	1.2009	1.2009	1.2009	1.2009
88	1.1842	1.1842	1.1842	1.1842	1.1842	1.1842	1.1842	1.1842	1.1842	1.1842	1.1842	1.1842	1.1842
87	1.1837	1.1837	1.1837	1.1837	1.1837	1.1837	1.1837	1.1837	1.1837	1.1837	1.1837	1.1837	1.1837
86	1.1817	1.1817	1.1817	1.1817	1.1817	1.1817	1.1817	1.1817	1.1817	1.1817	1.1817	1.1817	1.1817
85	1.1508	1.1508	1.1508	1.1508	1.1508	1.1508	1.1508	1.1508	1.1508	1.1508	1.1508	1.1508	1.1508
84	1.1200	1.1200	1.1200	1.1200	1.1200	1.1200	1.1200	1.1200	1.1200	1.1200	1.1200	1.1200	1.1200
83	1.1096	1.1096	1.1096	1.1096	1.1096	1.1096	1.1096	1.1096	1.1096	1.1096	1.1096	1.1096	1.1096
82	1.1073	1.1073	1.1073	1.1073	1.1073	1.1073	1.1073	1.1073	1.1073	1.1073	1.1073	1.1073	1.1073
81	1.1044	1.1044	1.1044	1.1044	1.1044	1.1044	1.1044	1.1044	1.1044	1.1044	1.1044	1.1044	1.1044
80	1.1007	1.1007	1.1007	1.1007	1.1007	1.1007	1.1007	1.1007	1.1007	1.1007	1.1007	1.1007	1.1007
79	1.0971	1.0971	1.0971	1.0971	1.0971	1.0971	1.0971	1.0971	1.0971	1.0971	1.0971	1.0971	1.0971
78	1.0959	1.0959	1.0959	1.0959	1.0959	1.0959	1.0959	1.0959	1.0959	1.0959	1.0959	1.0959	1.0959
77	1.0946	1.0946	1.0946	1.0946	1.0946	1.0946	1.0946	1.0946	1.0946	1.0946	1.0946	1.0946	1.0946
76	1.0926	1.0926	1.0926	1.0926	1.0926	1.0926	1.0926	1.0926	1.0926	1.0926	1.0926	1.0926	1.0926
75	1.0904	1.0904	1.0904	1.0904	1.0904	1.0904	1.0904	1.0904	1.0904	1.0904	1.0904	1.0904	1.0904
74	1.0880	1.0880	1.0880	1.0880	1.0880	1.0880	1.0880	1.0880	1.0880	1.0880	1.0880	1.0880	1.0880
73	1.0852	1.0852	1.0852	1.0852	1.0852	1.0852	1.0852	1.0852	1.0852	1.0852			
72	1.0823	1.0823	1.0823	1.0823	1.0823	1.0823	1.0823	1.0823	1.0823	1.0823			
71	1.0790	1.0790	1.0790	1.0790	1.0790	1.0790	1.0790	1.0790	1.0790	1.0790			
70	1.0756	1.0756	1.0756	1.0756	1.0756	1.0756	1.0756	1.0756	1.0756	1.0756			
69	1.0748	1.0748	1.0748	1.0748	1.0748	1.0748	1.0748	1.0748	1.0748	1.0748			
68	1.0748	1.0748	1.0748	1.0748	1.0748	1.0748	1.0748	1.0748	1.0748	1.0748			
67	1.0745	1.0745	1.0745	1.0745	1.0745	1.0745	1.0745	1.0745	1.0745	1.0745			
66	1.0739	1.0739	1.0739	1.0739	1.0739	1.0739	1.0739	1.0739	1.0739	1.0739			

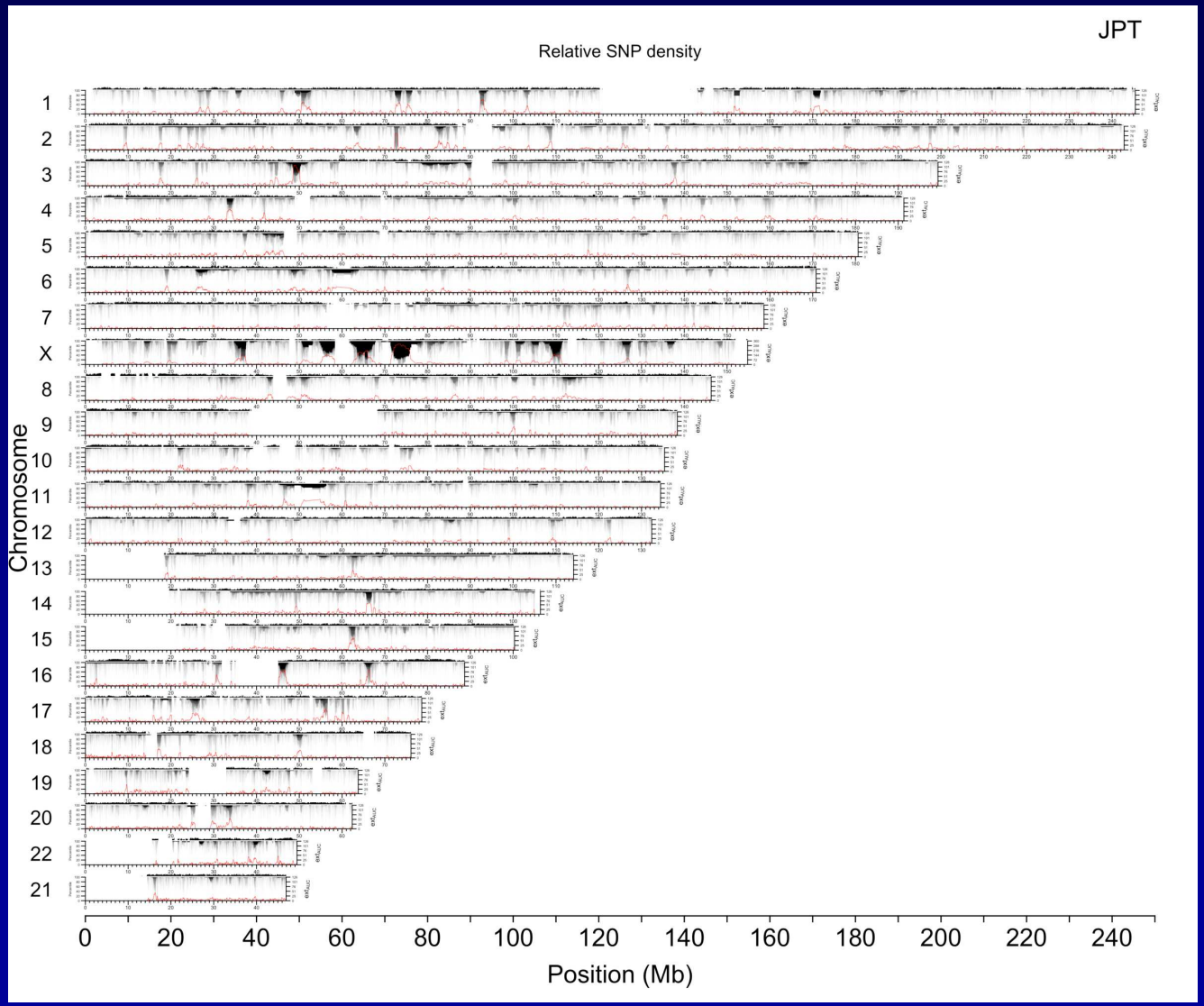
- ext_{AUC} = Extent (Area under the curve)
- Integrate the area under the curve for each locus in PE_{mat}



Smoothed ext_{AUC} values across the genome for all four populations

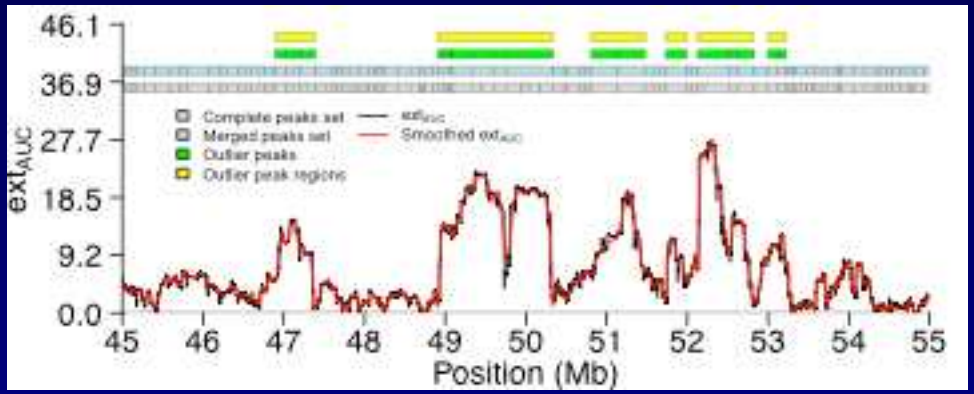


Plot of PE_{mat} and ext_{AUC} values as a means of visualizing haplotype diversity and structure across the genome



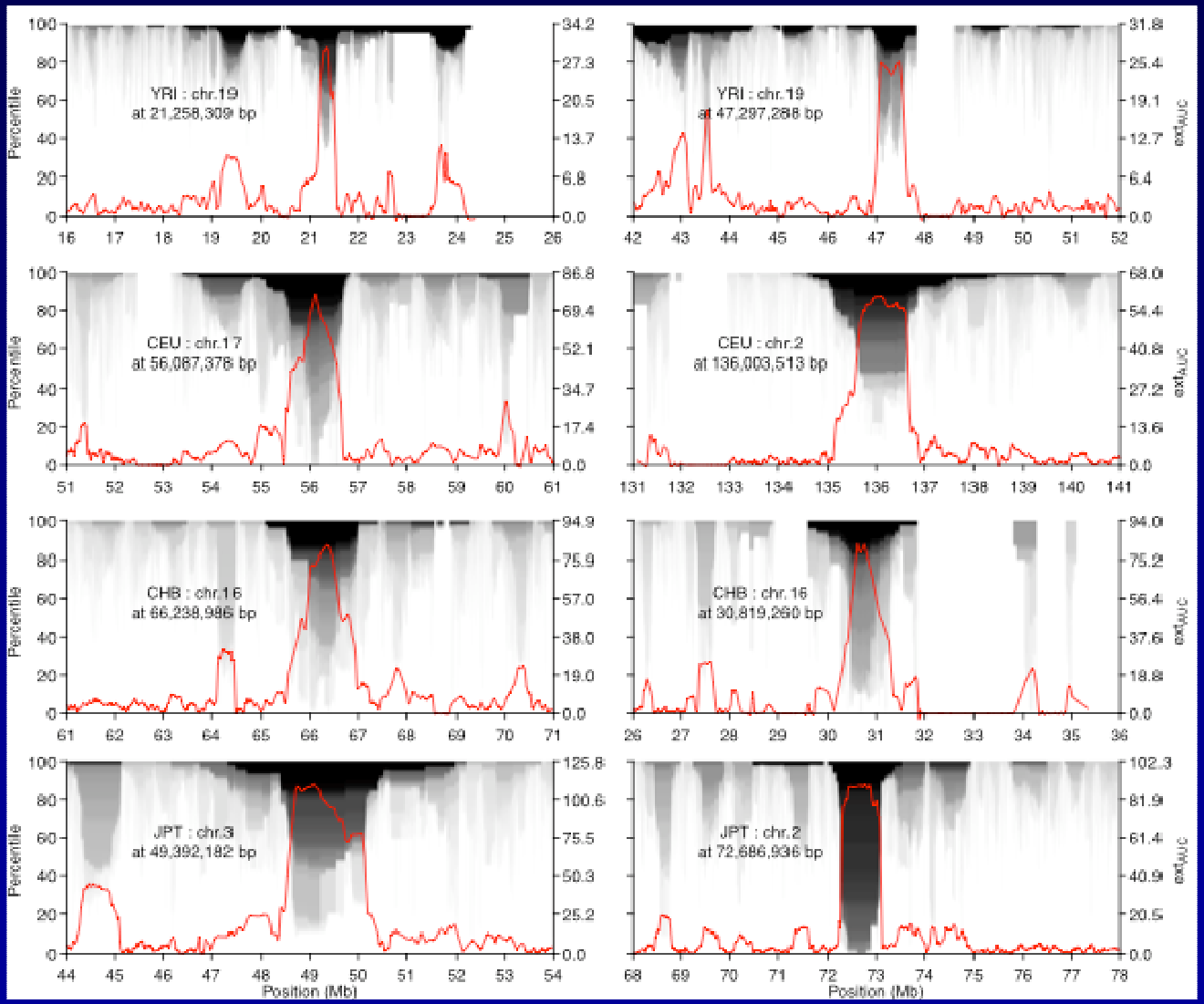
Peak detection and processing

- Peaks were detected using our own method based on predicting *dy/dx* and finding local maxima & minima.
- Similarly sized and separated peaks were then merged.
- Outlier peaks were extracted for each population and chromosome
- Contiguous outlier peaks were combined into outlier regions.

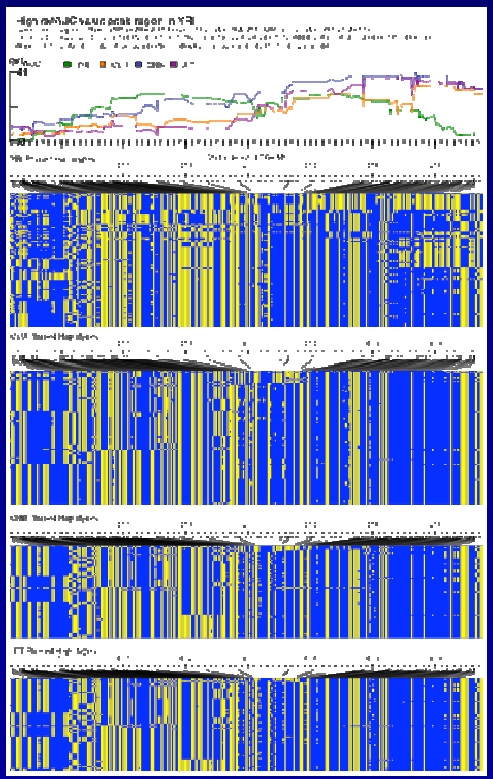


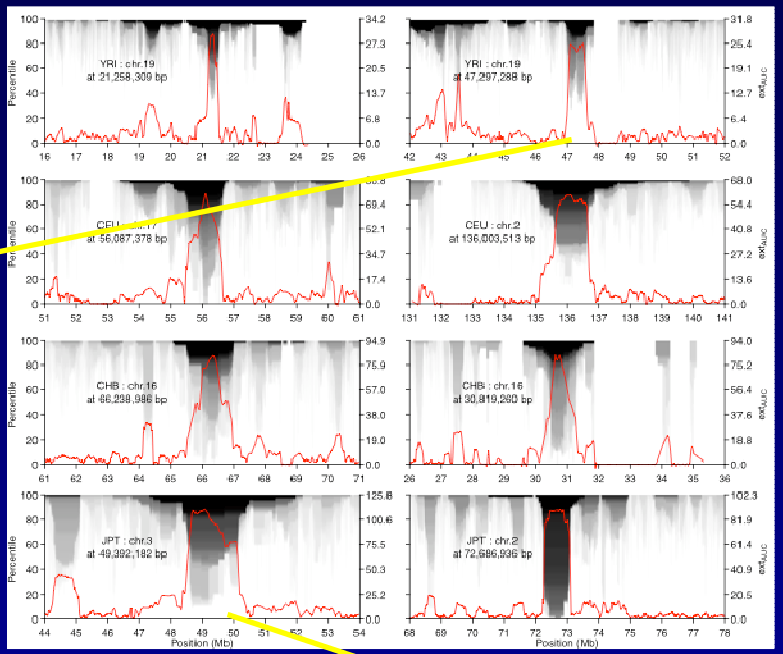
Merge type	YRI	CEU	CHB	JPT
Complete peak count	28,928	27,392	27,214	27,130
Merged peak count	23,284	22,653	22,660	22,615
Chromosome outlier peak count	1,575	1,492	1,606	1,567
Peak region count	902	656	605	579
Outlier regions	59	42	46	37
Peaks within outlier regions With height > 0.75*outlier height cutoff	124	120	136	115

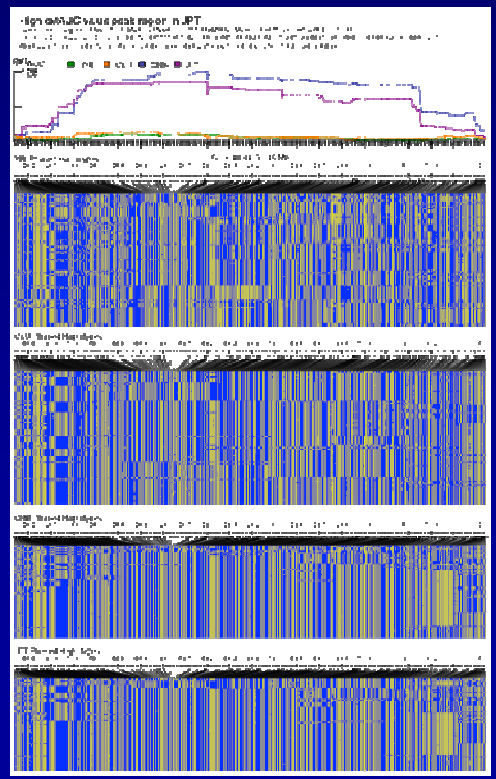
Top autosomal peak regions



Top autosomal peak regions compared to phased haplotype plots







Conclusions

- The distribution of contiguous homozygosity across the genome and populations mirrors patterns seen from plotting phased haplotypes.
- Although infrequent, YRI has genomic regions that have higher levels of homozygosity compared to the other three populations.
- Ongoing development suggests that we can utilize ext_{AUC} to search for regions that harbor multiple rare recessive disease variants in a population based case/control study.

Acknowledgments

- Advisors
  - Tatsuhiko Tsunoda (RIKEN CGM)
  - Yoshihito Niimura (TMDU)
- Computing systems
  - Takahisa Kawaguchi (prev. RIKEN CGM)
  - Muneyoshi Ohtsuka (RIKEN CGM)