



LOCAL VARIATION IN POLYMORPHISM RATES AND SIGNALS FOR RECENT SELECTION IN NONCODING DNA OF HOMINIDS

Belinda Giardine, Kuan-Bei Chen, Robert Harris, Aakrosh Ratan, Webb Miller, Stephan C. Schuster, Vanessa Hayes, Francesca Chiaromonte, Ross C. Hardison
Center for Comparative Genomics and Bioinformatics of the Huck Institutes of Life Sciences, The Pennsylvania State University, University Park, Pennsylvania

<http://main.genome-browser.bx.psu.edu>



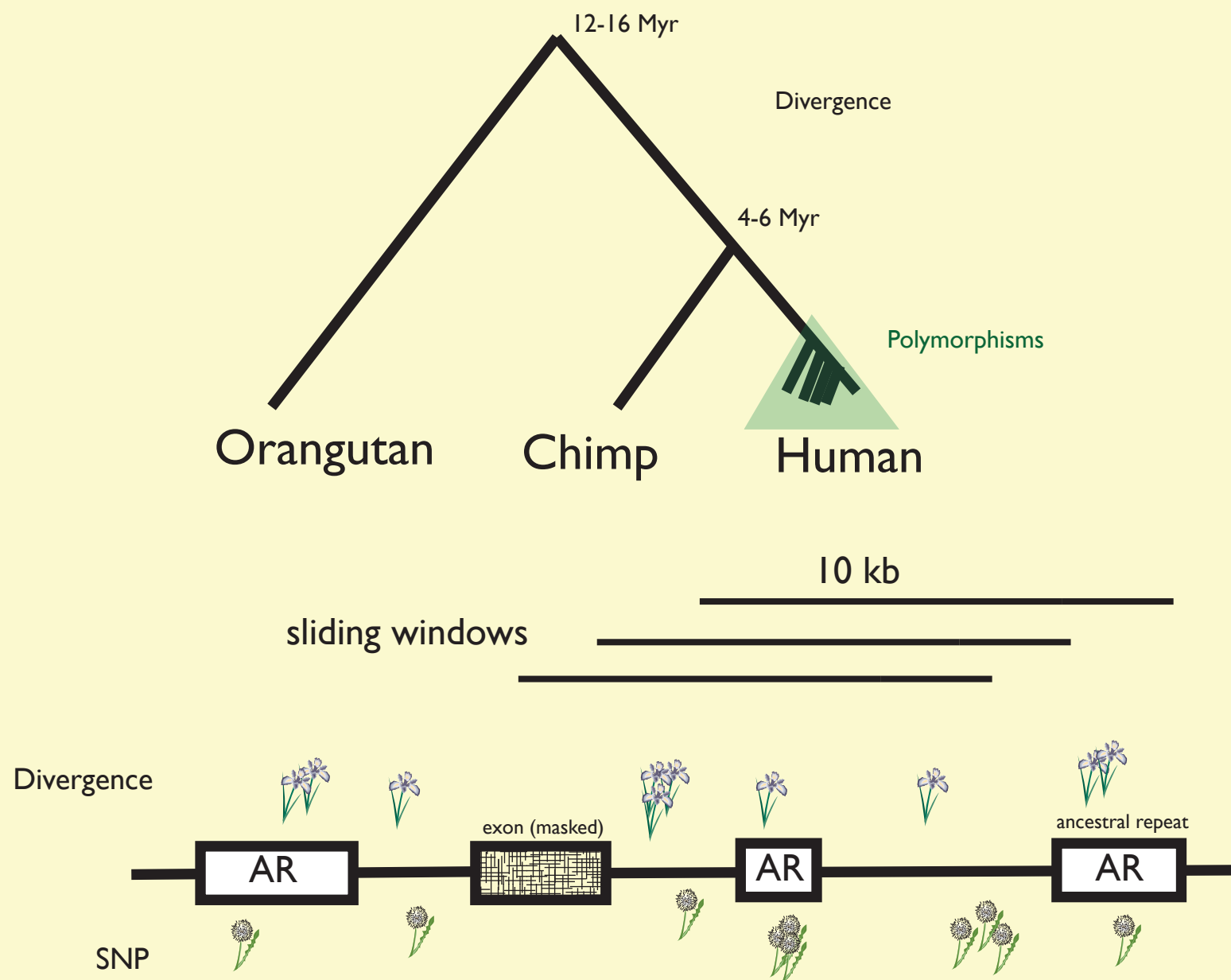
Comparison of interspecies divergence and within-species polymorphism can be used as a test for non-neutral evolution (McDonald-Kreitman test). As input data, we use alignments of the human reference sequence with chimp and orangutan, and SNPs from personal genomes (PGs). Ancestral repeats (AR) are the proxy for neutral DNA.

Divergence and SNP counts after masking coding exons and CpGs:

	Total	AR	not AR
Divergence from chimp	28,729,623	12,540,717	16,188,906
Divergence from orangutan	78,673,940	35,676,824	42,997,116
Polymorphic sites in PGs	8,725,848	3,895,705	4,830,143



McDonald-Kreitman test for non-neutral evolution in noncoding DNA, using Ancestral Repeats as a neutral reference (MKAR)



	SNP	divergence
AR (neutral model)	18	78
non-AR	15	276

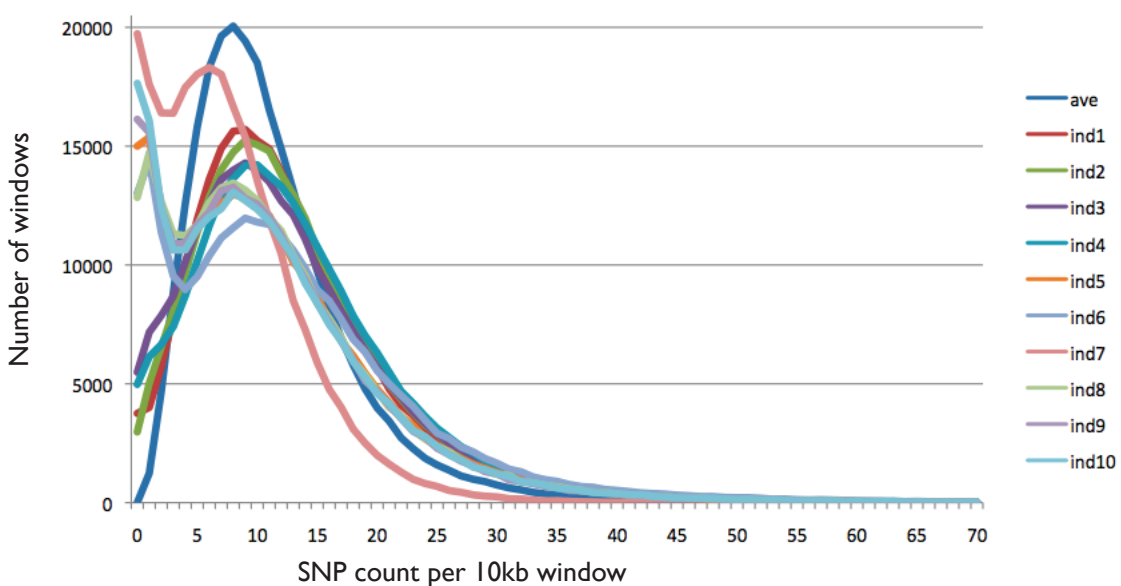
Numbers from chr2 136,303,000- 136,313,000

$p < 0.00001$

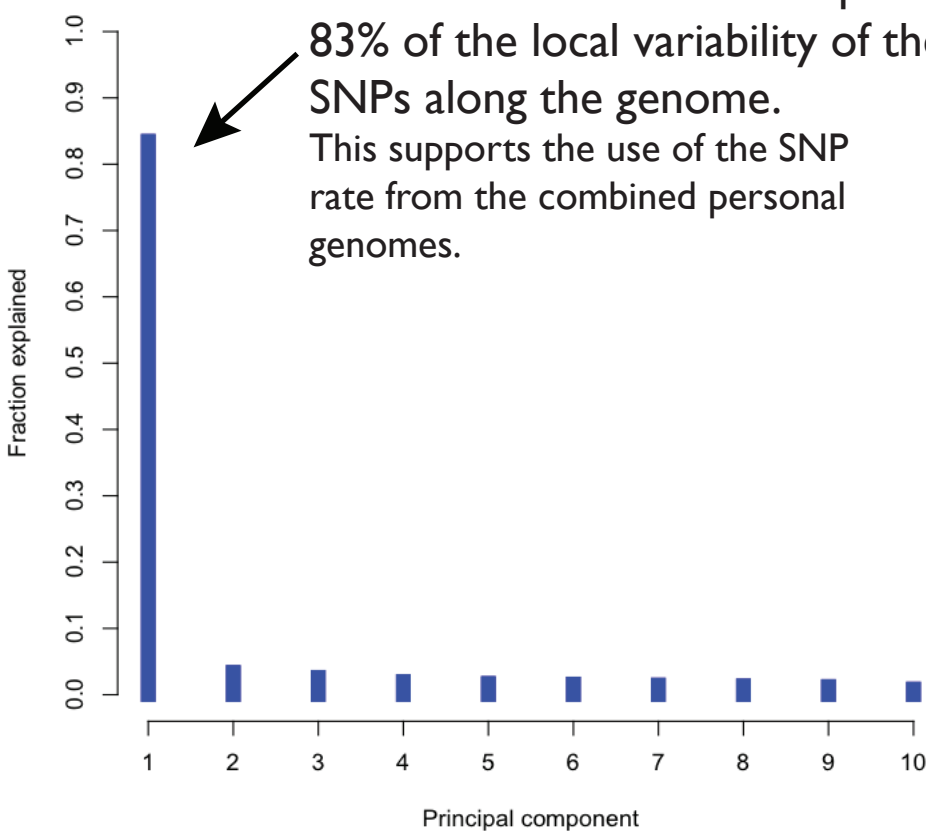
Hits are defined as significant after a FDR correction with a false positive rate of 0.10.

The SNP rate varies locally.

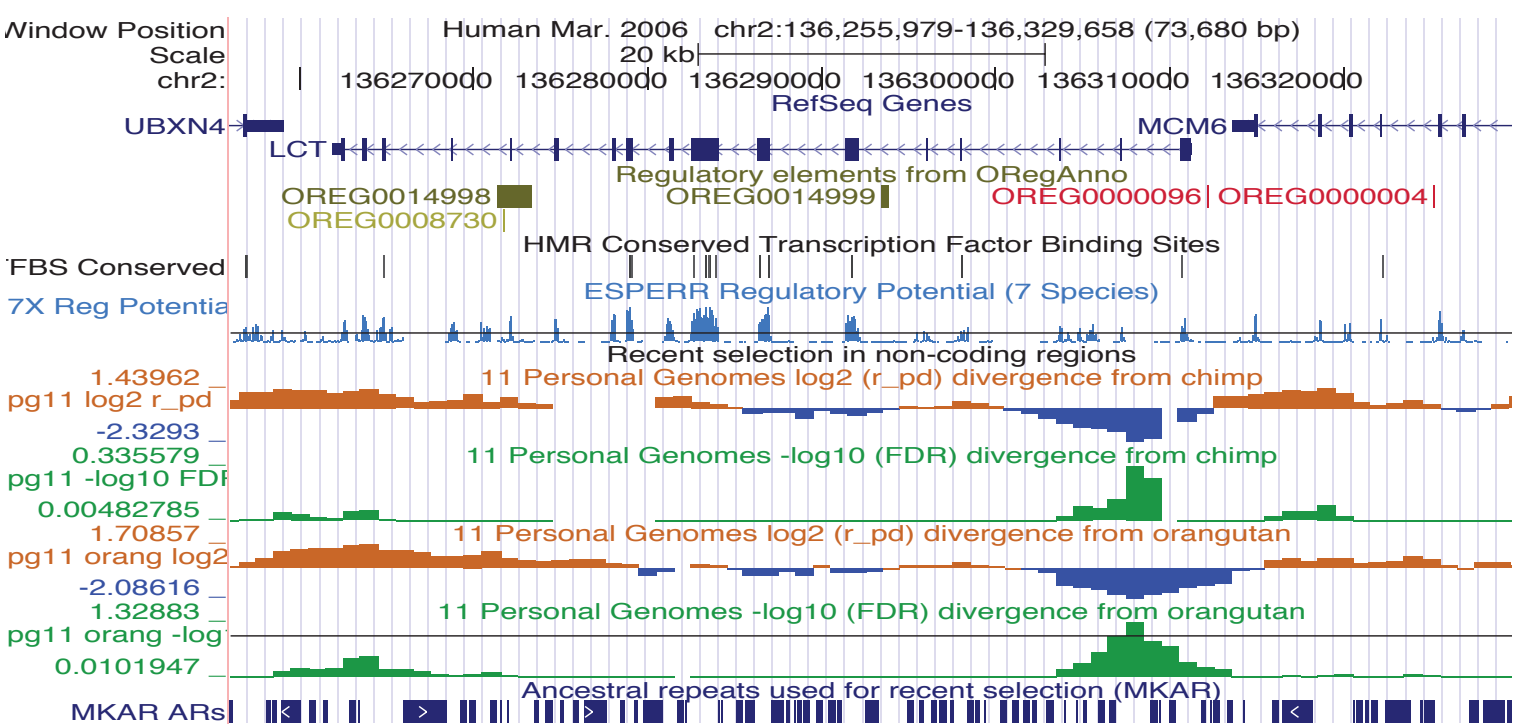
The graph is a histogram of SNP density in 10kb windows.



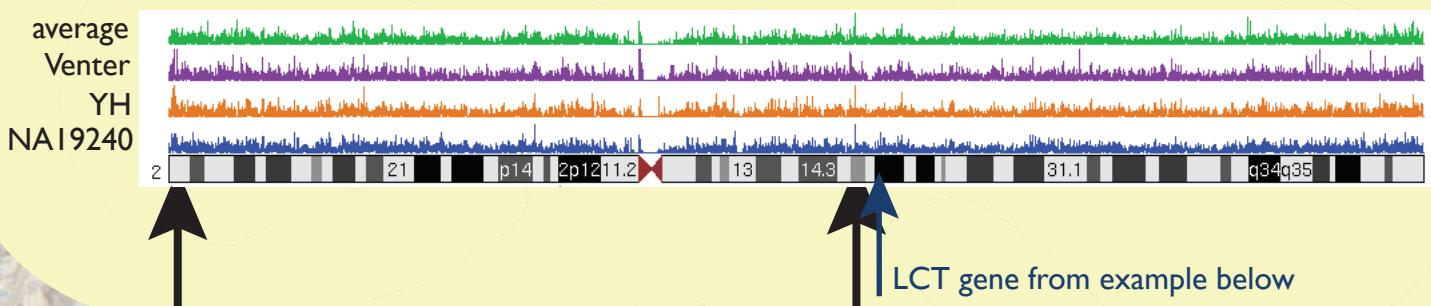
The combined SNP rate explains 83% of the local variability of the SNPs along the genome. This supports the use of the SNP rate from the combined personal genomes.



This example region has the lactase gene. There are 2 regulatory SNPs (in red) annotated in the ORegAnno track that affect lactase activity, one of which is found in a region we identify as being under positive selection.



Genome graphs showing SNP density along chr2. YH is a male Chinese individual. NAI9240 is a Yoruban daughter from the 1000 Genomes Project.



A peak found in all three individuals and another found in YH and NAI9240, but missing in Venter.

The MKAR positive selection hits (purple) are enriched for segmental duplications (blue) (p-value 1.34E-284).

