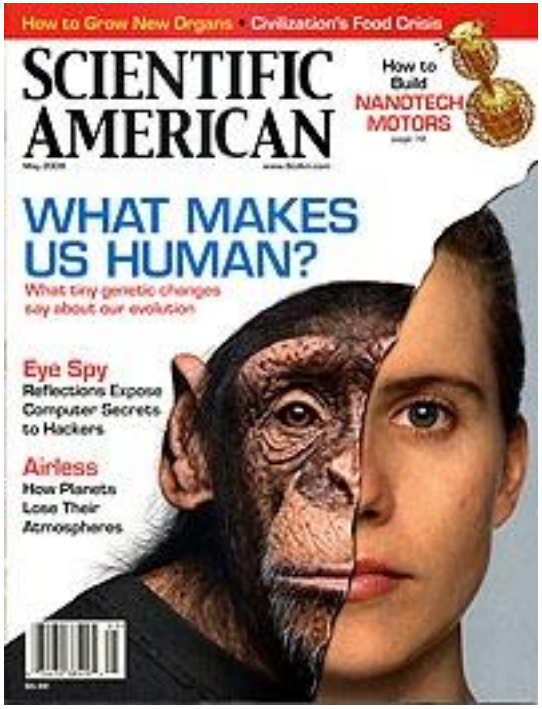
Ongoing GC-biased evolution is widespread in the human genome and enriched near recombination hotspots



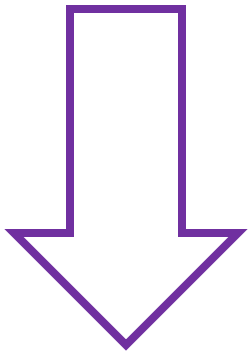
Sol Katzman, August 24, 2011 genecats

co-authors:
Tony Capra
David Haussler
Katie Pollard

The future of the human genome

An inconvenient truth?

GCCAACTAGTTCCGACTGGGTTAACCGTAGCT



GC-biased evolution

GCCGACCGGCGCTGACCGGATCGGCCGCAGCC

GC-biased Evolution: nomenclature

- Weak-to-Strong (W2S)
  - ancestral allele is A or T
  - derived allele is G or C
- GC-bias: W2S is more likely than S2W
- Historical GC-bias:
  - genomic change **along a species lineage**
  - determined from fixed substitutions between species
  - quantify with BDS (Bias in Divergent Sequences)

Capra and Pollard *Gen.Biol.Evol.* (2011)

- Ongoing GC-bias:
  - variation **within a species population**
  - determined from Derived Allele Frequency spectra
  - quantify with W2S DAF skew

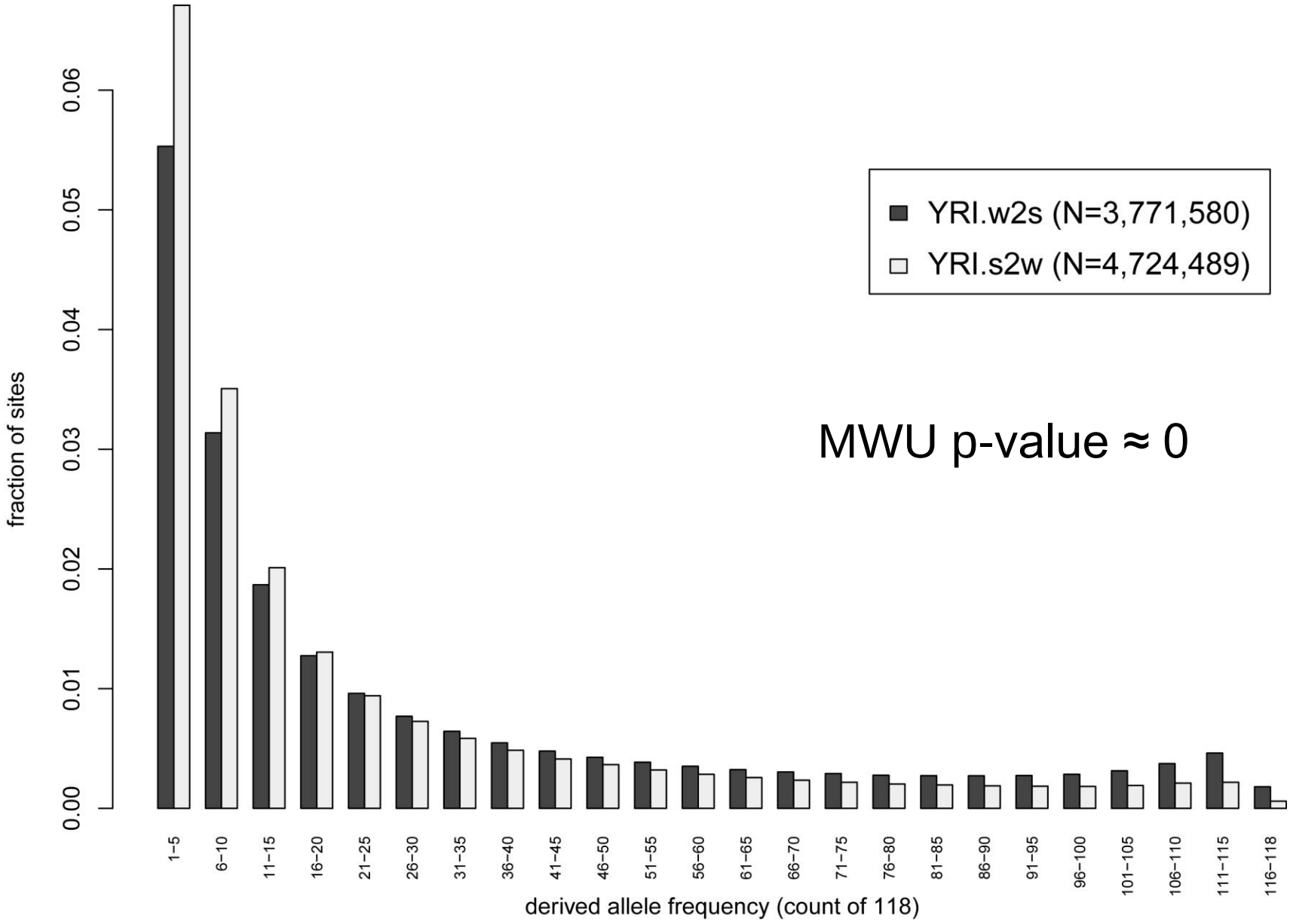
Got SNPs?

- 1000 Genomes low-coverage pilot

  1000Genomes.org *Nature* (2010)

  - 59 YRI individuals (also CEU and CHB+JPT)
  - High confidence SNP calls
  - Ancestral/Derived allele polarized using several primate genomes (EPO pipeline)
- Whole genome or specific regions. For each region:
  - Frequency spectrum calculated separately for W2S and S2W SNPs
  - Compare 2 spectra using MWU (rank sum) test to get p-value and direction of bias

YRI whole genome DAF skew



DAF Skew across the genome

- Non-overlapping windows
- P < 0.05, 2-sided MWU test
  - 2.5% expected W2S
  - 2.5% expected S2W

Window Size	Number of windows	SNPs per window	W2S p < 0.05 fraction	S2W p < 0.05 fraction
4Mb	699	14,430	98.7%	0.0%

DAF Skew across the genome

- Non-overlapping windows
- P < 0.05, 2-sided MWU test
  - 2.5% expected W2S
  - 2.5% expected S2W

Window Size	Number of windows	SNPs per window	W2S p < 0.05 fraction	S2W p < 0.05 fraction
4Mb	699	14,430	98.7%	0.0%
1Mb	2,715	3,715	96.3%	0.0%

DAF Skew across the genome

- Non-overlapping windows
- P < 0.05, 2-sided MWU test
  - 2.5% expected W2S
  - 2.5% expected S2W

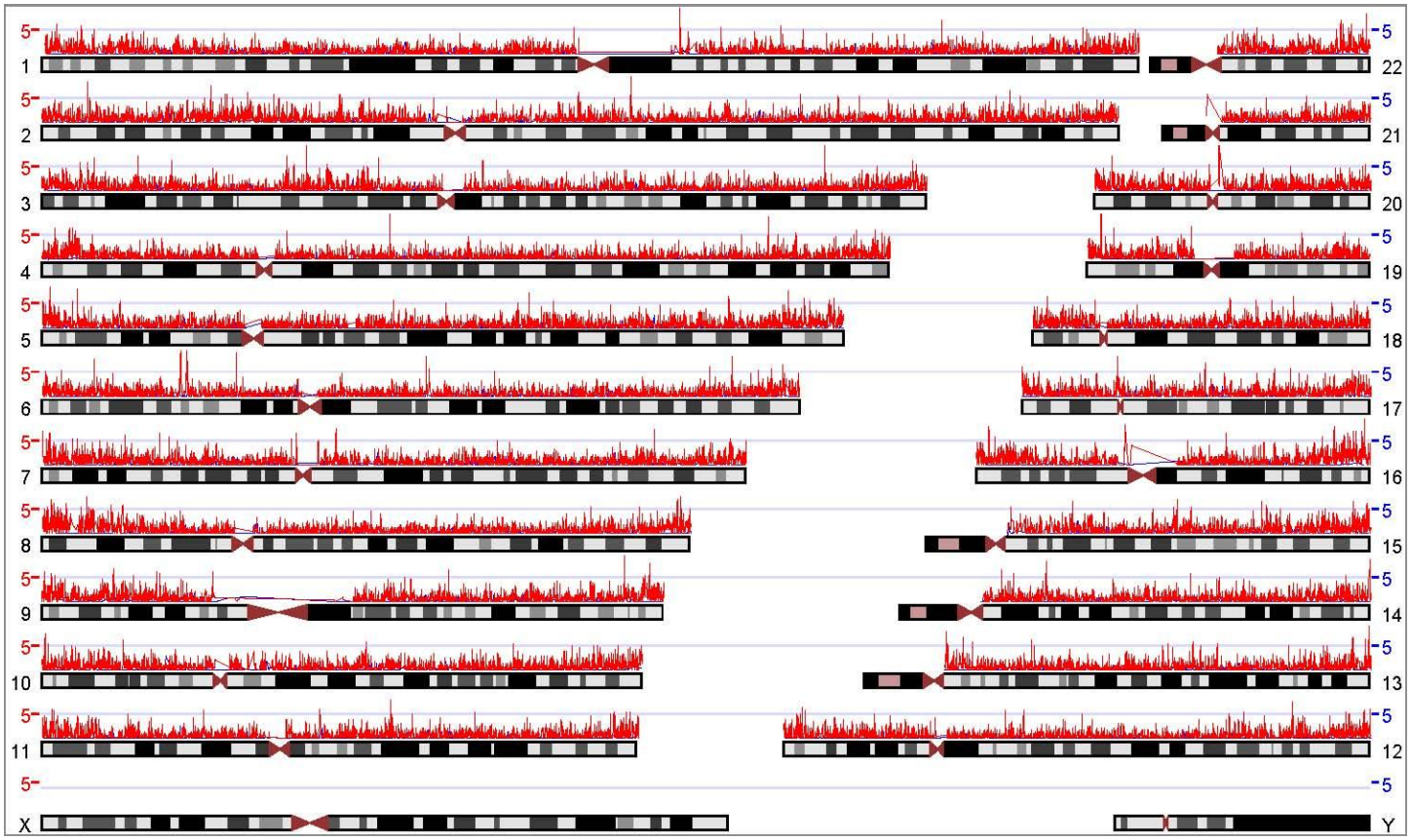
Window Size	Number of windows	SNPs per window	W2S p < 0.05 fraction	S2W p < 0.05 fraction
4Mb	699	14,430	98.7%	0.0%
1Mb	2,715	3,715	96.3%	0.0%
…				
40kb	65,510	154	22.4%	0.3%

-log10 MWU p-value across genome

- S2W-biased cases only: few significant 40kb regions

-log10 MWU p-value across genome

- S2W-biased and W2S-biased shown
- Overwhelming majority of 40kb regions are W2S-biased
- But no clear pattern – bias is not particularly localized



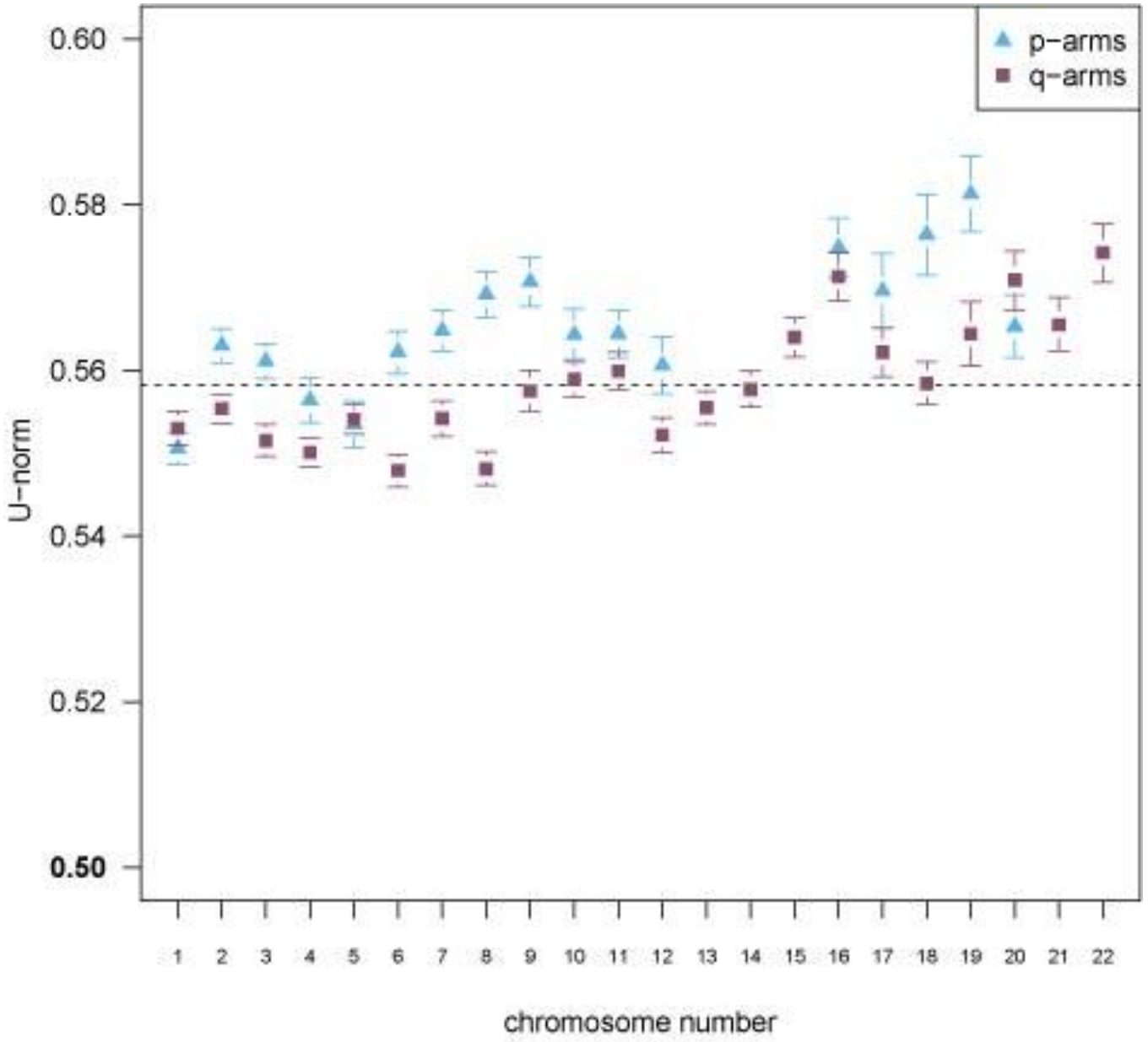
U-norm

- Quantify DAF skew to compare regions with different numbers of SNPs
- Normalized Mann-Whitney U
  - U/Umax (Umax = #W2S x #S2W)
  - Range 0 to 1
  - Interpretation: P(W2S DAF > S2W DAF)
    + 0.5 P(W2S DAF = S2W DAF)
  - 0.50 in absence of bias
- For YRI genome-wide: ??

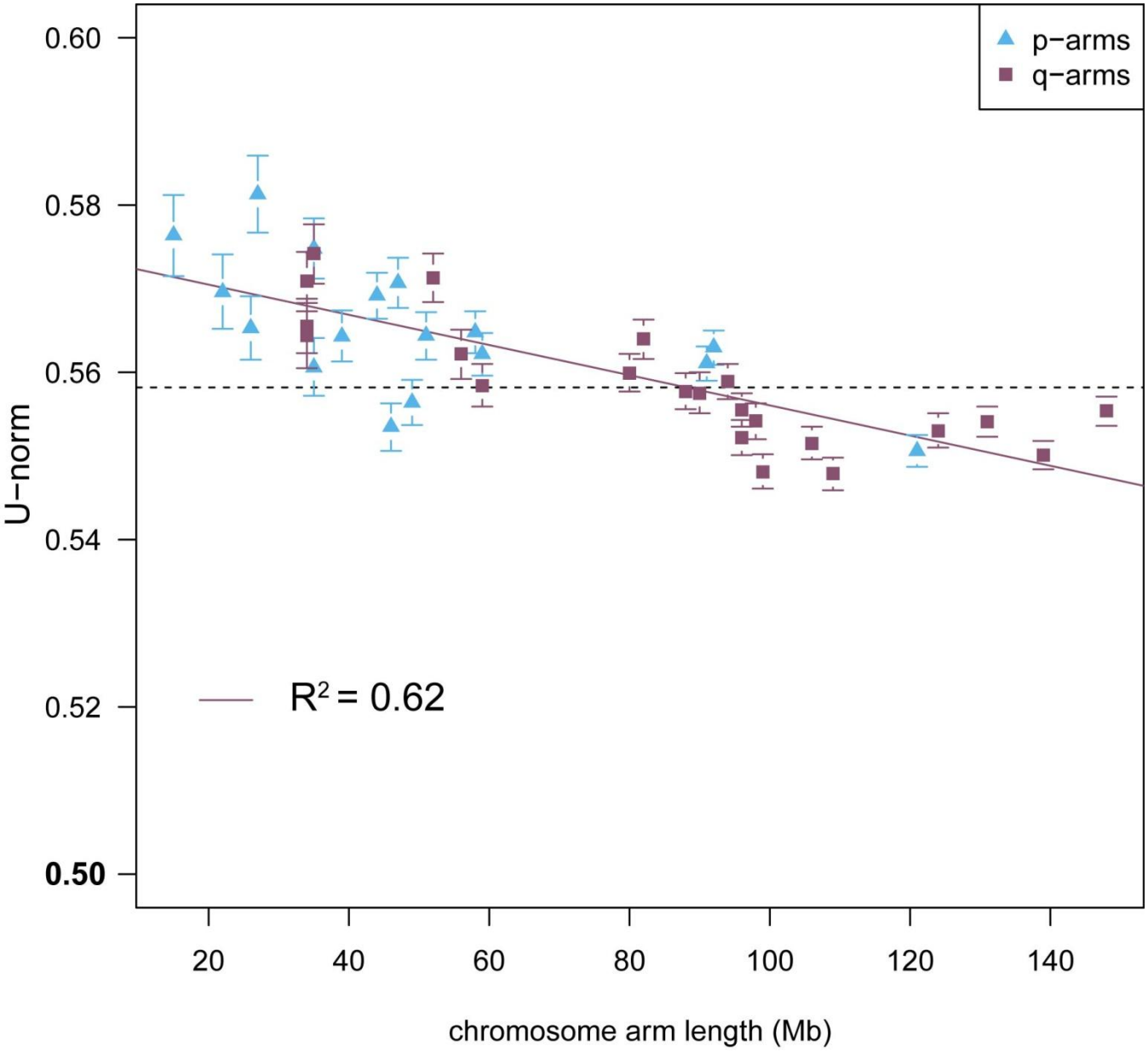
U-norm

- Quantify DAF skew to compare regions with different numbers of SNPs
- Normalized Mann-Whitney U
  - U/Umax (Umax = #W2S x #S2W)
  - Range 0 to 1
  - Interpretation: P(W2S DAF > S2W DAF)
    + 0.5 P(W2S DAF = S2W DAF)
  - 0.50 in absence of bias
- For YRI genome-wide: 0.558

U-norm by chromosome arm



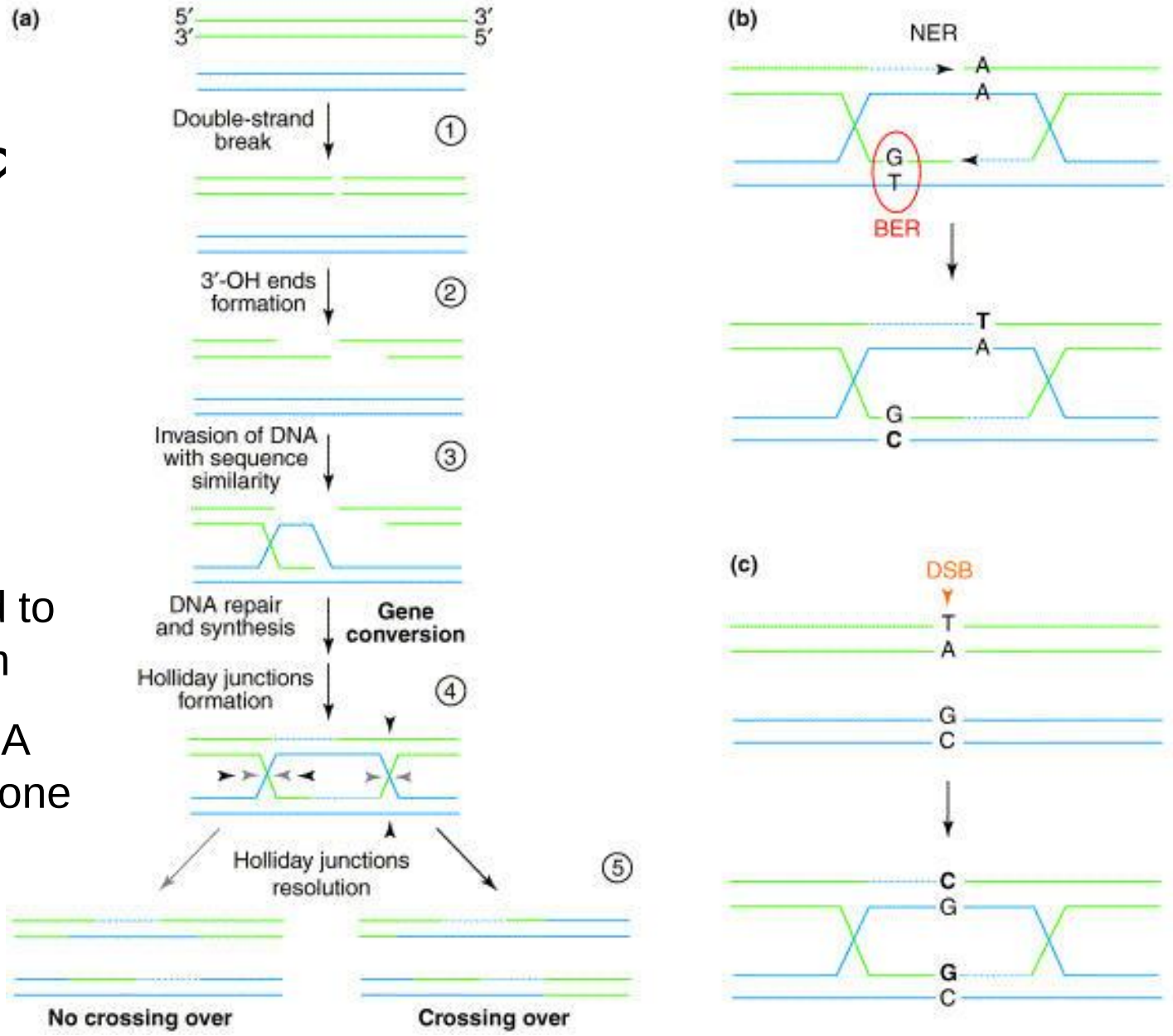
U-norm by chrom-arm length



What is gene conversion?

Why is it biased?

- DNA synthesized to match other chrom
- heteroduplex DNA repaired to match one chrom

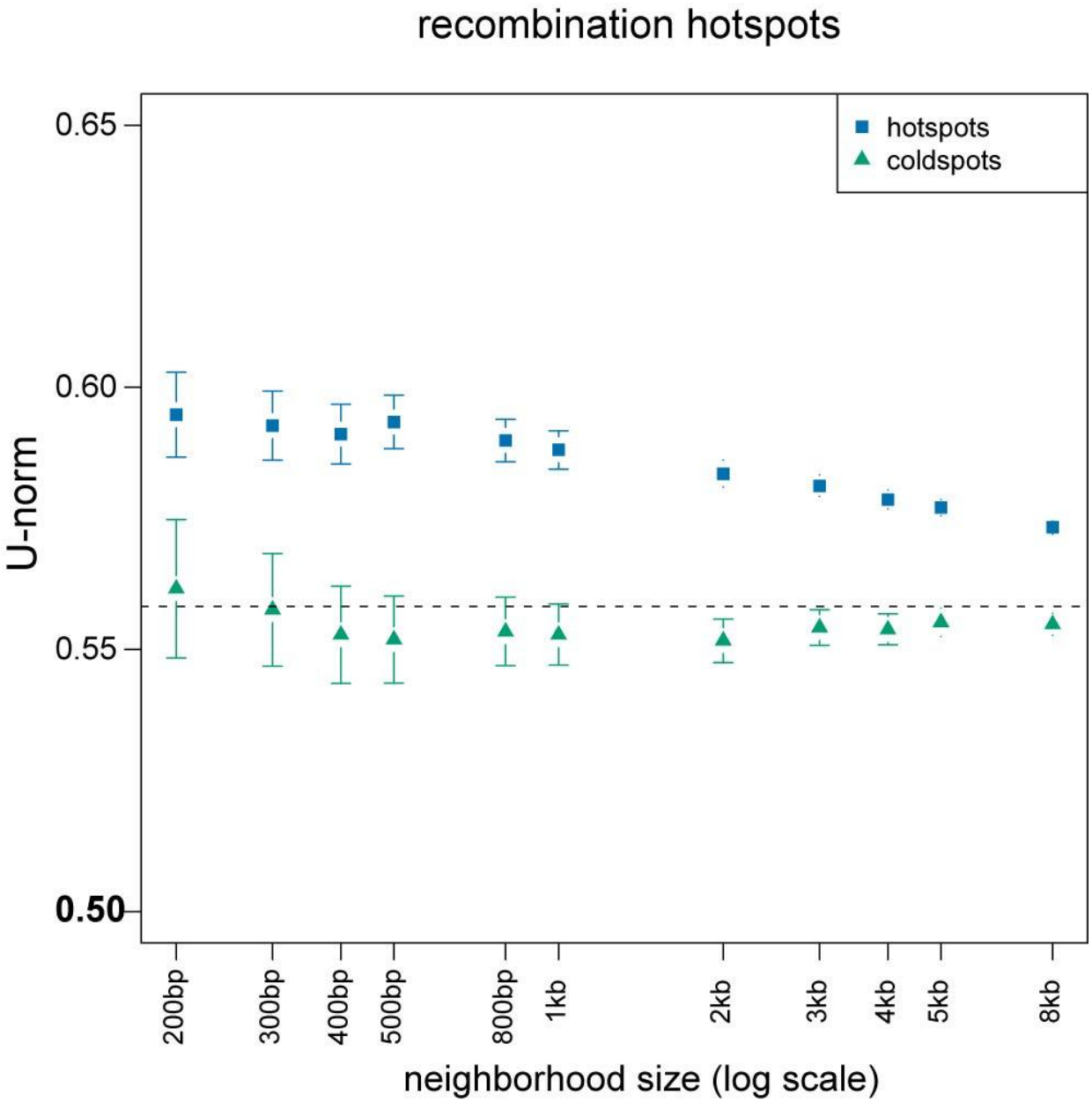


Marais et al. *Trends in Genetics* (2003)

Recombination?

- Limited number of recomb events per chrom-arm? More events **per nucleotide** on short arms, so higher U-norm
- If GC-Biased Gene Conversion (gBGC) is its cause, W2S-bias should be more evident at recombination hotspots
- Data from Oxford genotype study: 25,000 hotspots (and 9,000 coldspots) Myers et al. *Science* (2005)
- Analyze all SNPs aggregated from regions of various sizes centered on each hotspot (or coldspot)

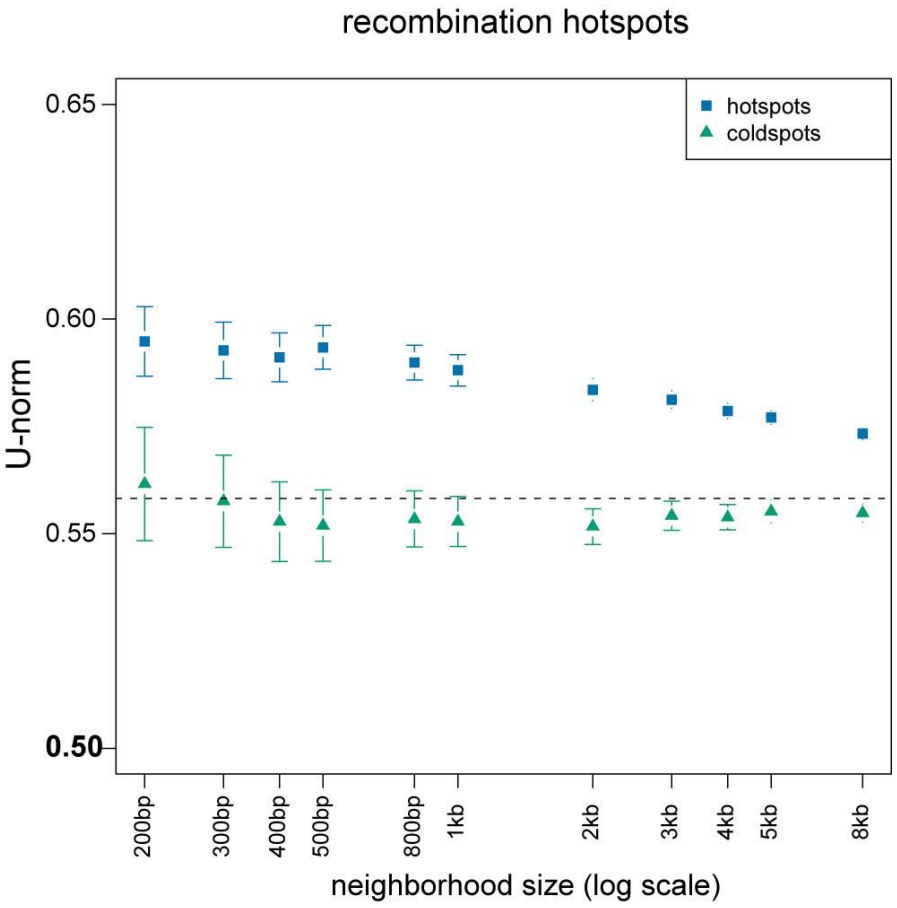
U-norm in neighborhoods of…

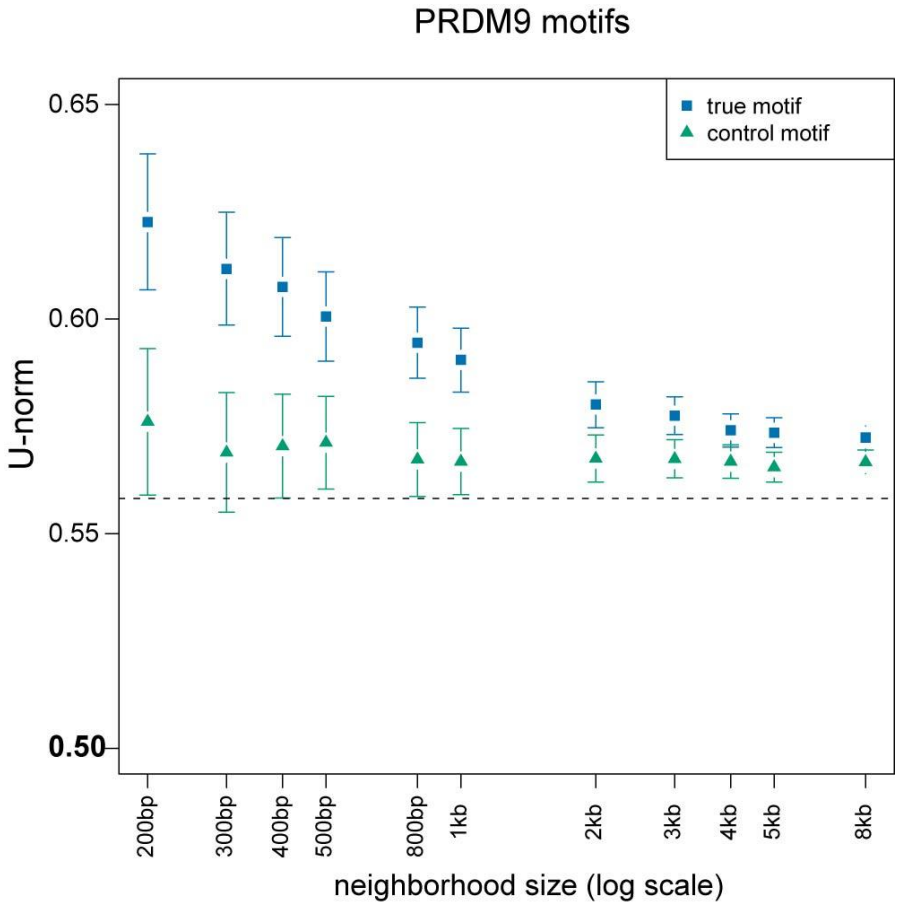


PRDM9

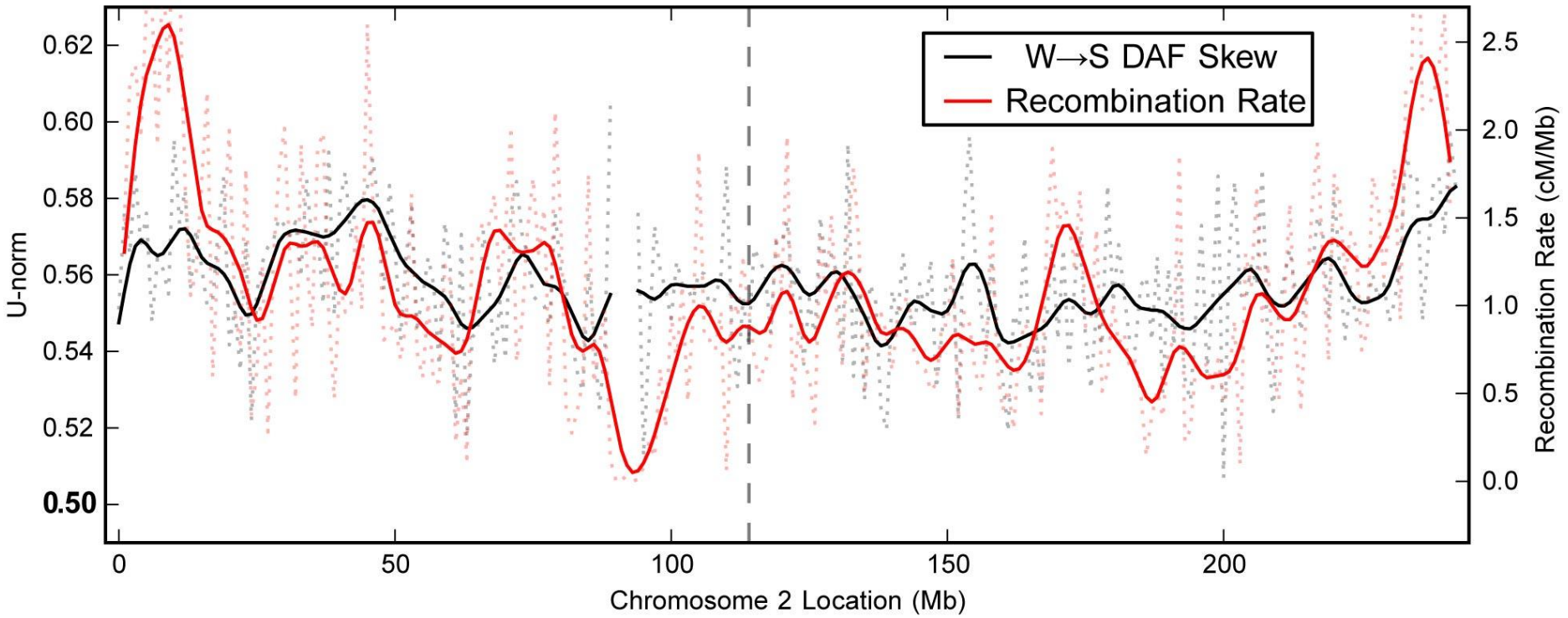
- Histone methyltranferase associated with recombination hotspots Myers et al. *Science* (2010)
- 13bp binding motif (elevated recomb rate)
  - CCTCCCTNNCCAC Myers et al. *Nat.Genet.* (2008)
- control motif (no elevated recomb rate)
  - CTTCCCTNNCCAC 1000Genomes.org *Nature* (2010)
- About 7,000 of each type of sites in the human genome. Aggregate the SNPs…

U-norm in neighborhoods of…



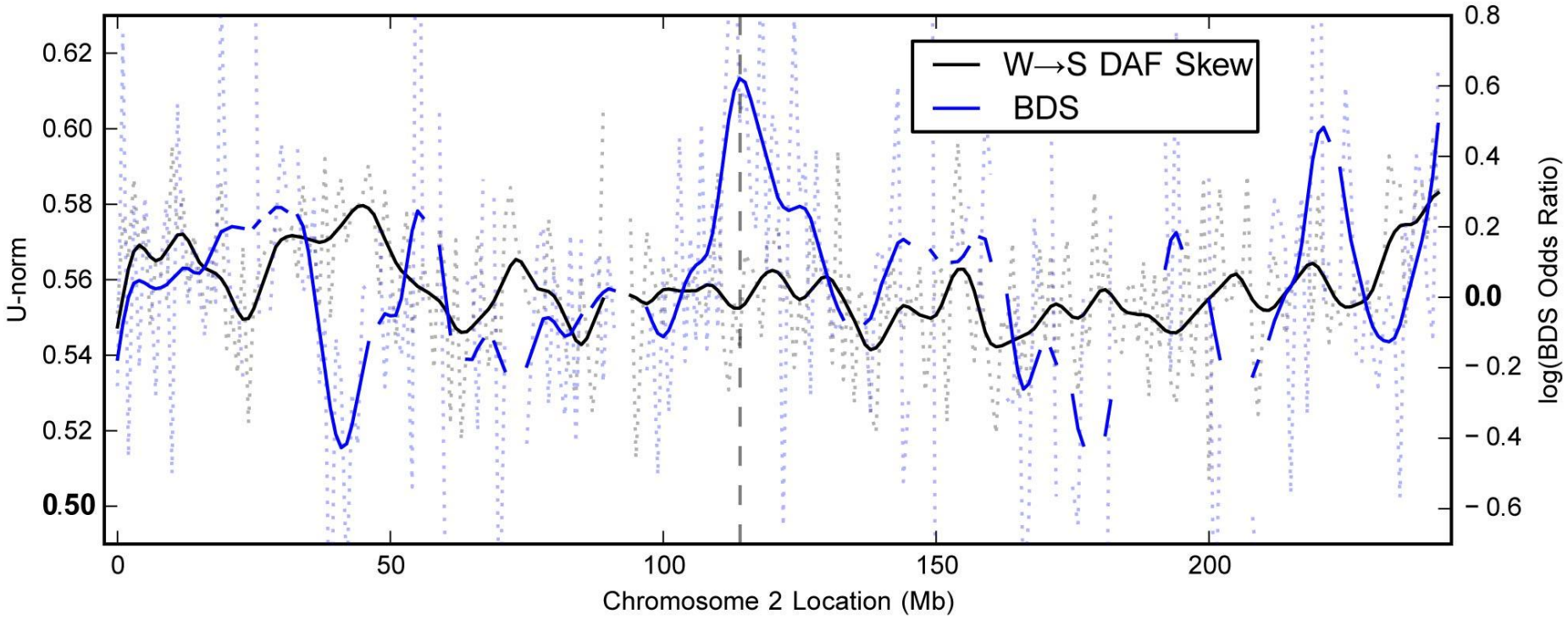


Correlation with recombination rate



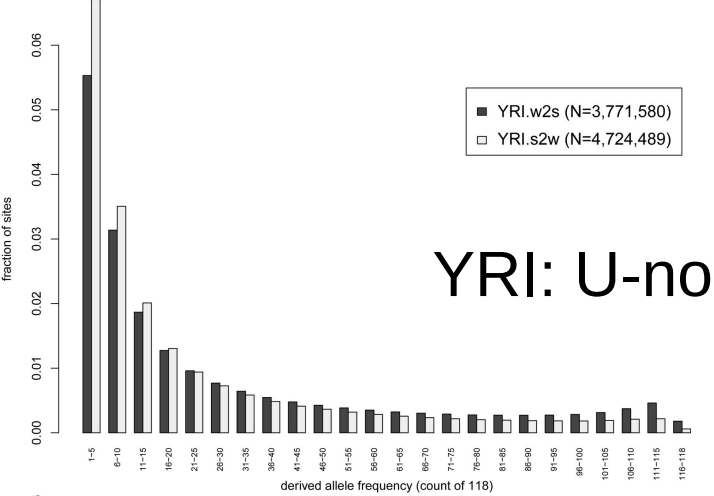
- Hapmap LD-based YRI recomb rate map
- whole genome 40kb windows: Spearman's ρ = 0.20
- whole genome 1Mb windows: Spearman's ρ = 0.53
- chrom2 1Mb windows ρ = 0.43

(weak) Correlation with historical GC-bias

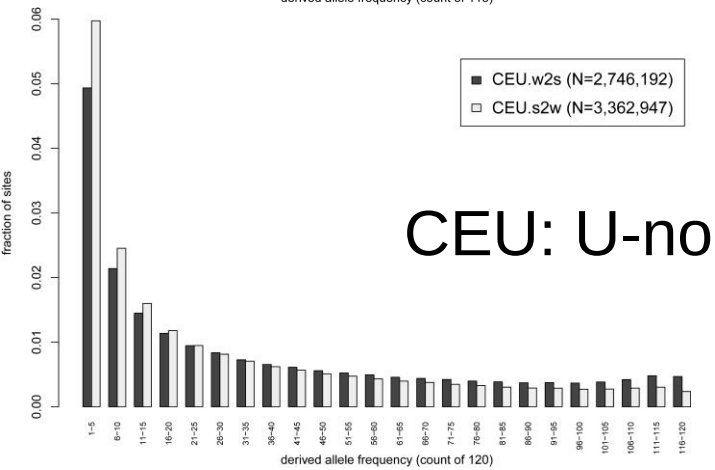


- Bias in Divergent Sequence (log Odds ratio)
- whole genome 1Mb windows: Spearman's ρ = 0.18
- chrom2 1Mb windows ρ = 0.09
- No peak in U-norm near chrom2 fusion event !

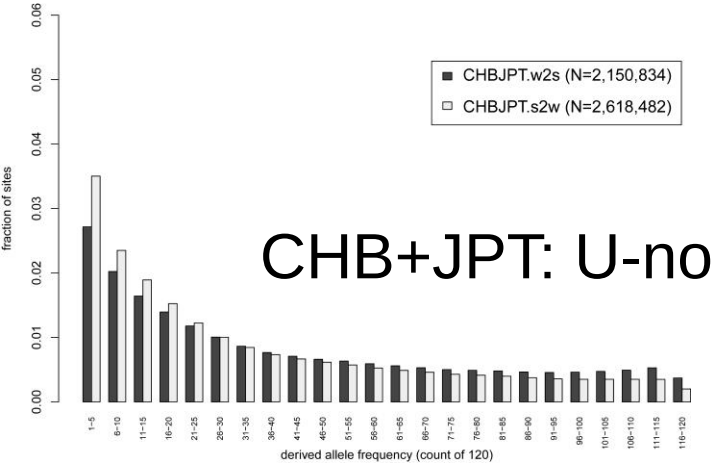
Population comparisons



YRI: U-norm = 0.558

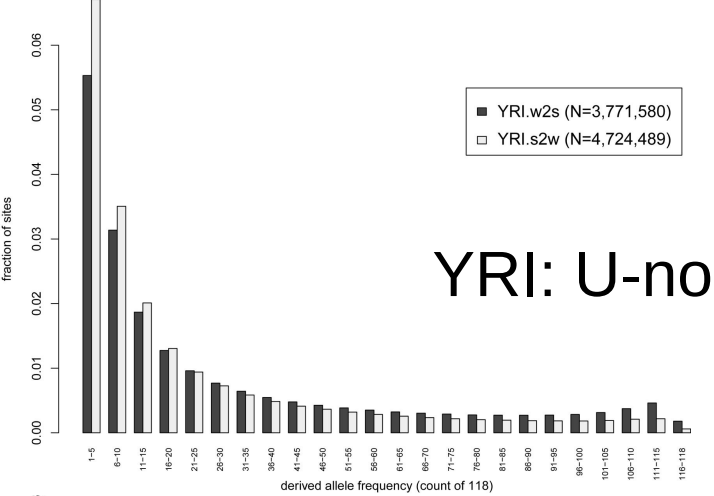


CEU: U-norm = ?.???

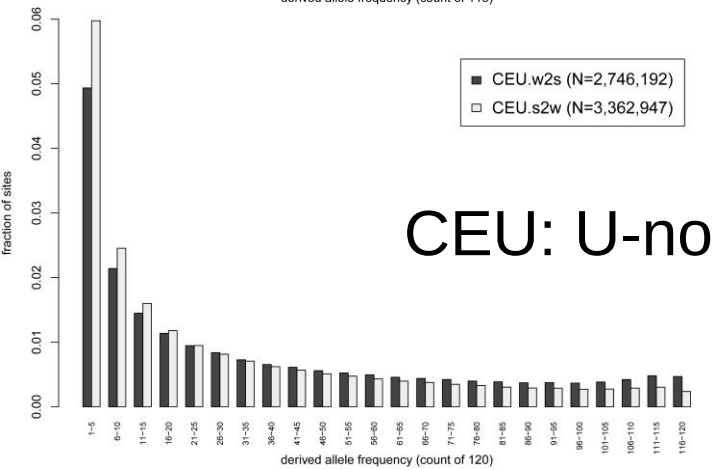


CHB+JPT: U-norm = ?.???

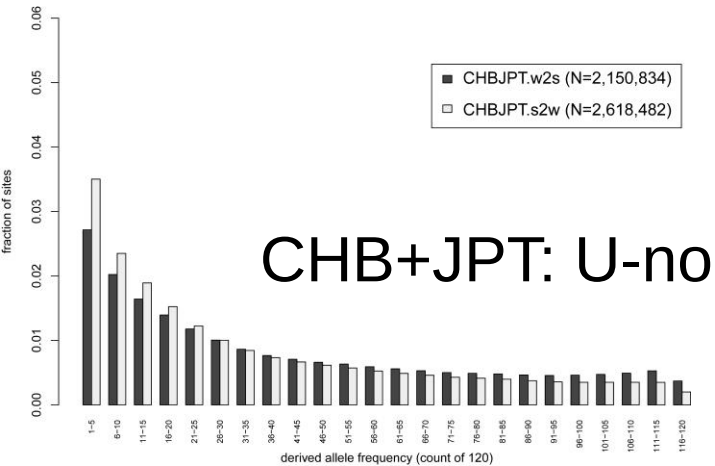
Population comparisons



YRI: U-norm = 0.558



CEU: U-norm = 0.552



CHB+JPT: U-norm = 0.553

Cross-population correlations at 1Mb scale, whole genome:

- YRI vs CEU: $\rho = 0.67$
- YRI vs CHB+JPT: $\rho = 0.54$
- CEU vs CHB+JPT: $\rho = 0.60$

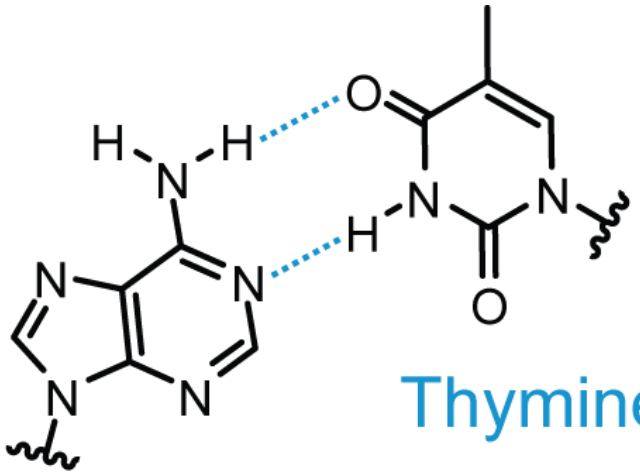
Conclusions

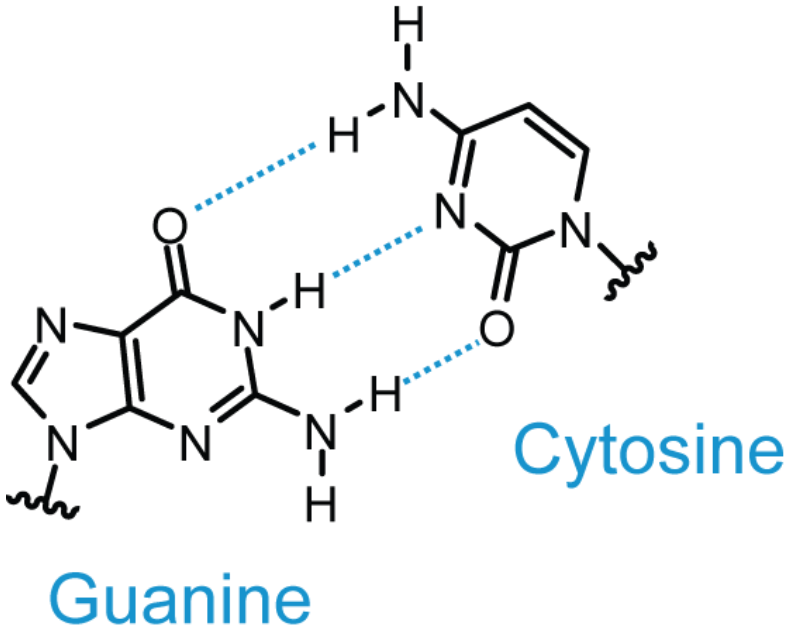
- Throughout the genome, W2S SNPs are at generally higher derived allele freq than S2W SNPs (heading towards fixation?)
  - This effect could have driven changes that were capitalized on by selection.
  - Or it could have made it harder to get to fitness improvements.
  - Or it could have caused fixation of deleterious alleles.
- The effect is correlated with recombination rate and is pronounced in very close neighborhoods of recomb hotspots and PRDM9 binding sites. gBGC a likely cause.
- Our genome is not at GC% equilibrium.
- **When does it stop? Where is it inhibited?**

Thanks

- The 1000 Genomes Project !!

Weak vs. Strong





You say you're lookin' for someone who's never ***weak*** *but always* ***strong…***

-- Bob Dylan *It ain't me babe* (1964)