Characterization of the human genome: codon bias, gene density, GC content, recombination, and CpG islands

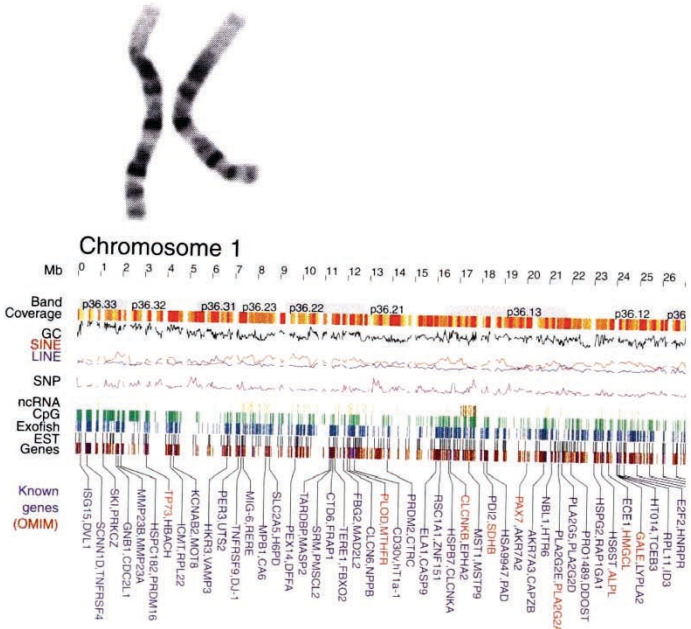
Biosciences 741: Genomics

Fall, 2013

Week 9

Outline

- Chromosome bands and regional variations in GC content
- CpG islands as an indicator of gene density, which is also correlated with promoter type
- Regional variations in recombination, cytosine deamination, and GC content
- Codon bias in mammals, and regional variations in codon bias



Chromosome bands

- Chromatin is composed of about 50% protein and 50% DNA.
- Chromosome bands that stain more darkly with Giemsa contain more protein and DNA than interband regions. This is because the chromatin is more tightly coiled (at metaphase of mitosis).
- The location of Geimsa bands is correlated with the local %GC content, as well as the relative GC content (relative to nearby areas) and the rate of change of GC content along the chromosome.
- In any case, computer programs do a reasonably good job of predicting the size and location of Giemsa bands from the DNA sequence.
- Each chromosome has a different staining pattern, but different people generally have the same staining pattern.
- Giema bands are weak in reptiles and absent in fish and amphibians, probably because they have limited heterogeneity in GC content.

Variation in GC content

- GC content correlates with chromatin coiling (dark G bands tend to be AT-rich) because of factors such as DNA methylation and histone acetylation.
- GC content also correlates with gene density, and CpG island density.
- GC content also correlates with chromosomal location (neighboring genes of unrelated function tend to have similar GC content).
- GC content also correlates with the activity of transposable elements.
- A histogram of GC content in the human genome shows a GC-rich tail.

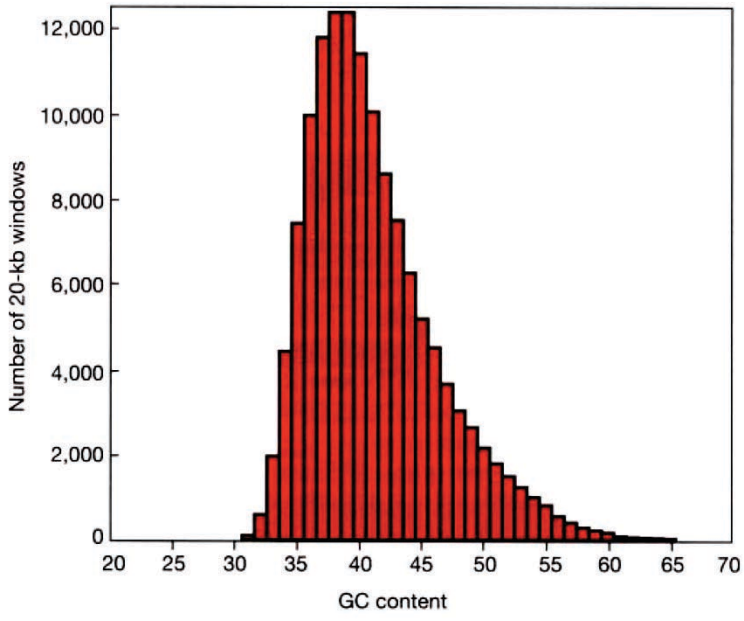
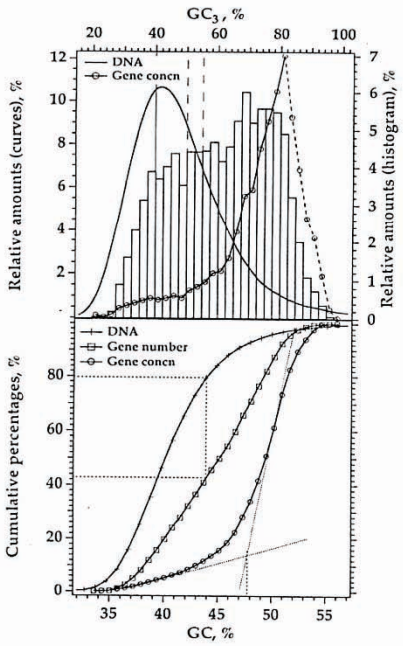
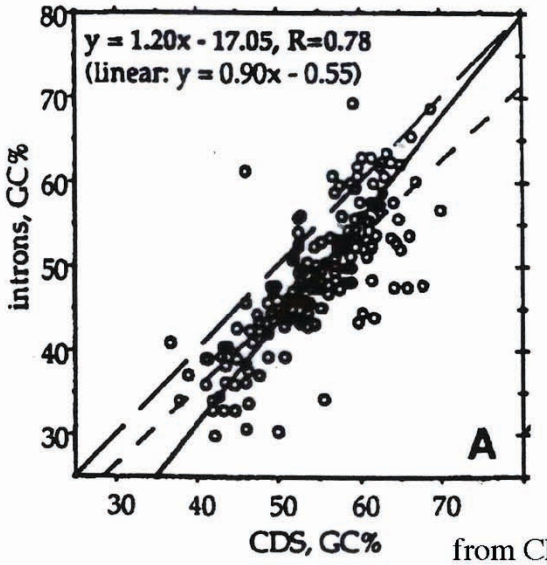


Figure 12 Histogram of GC content of 20-kb windows in the draft genome sequence.

Distribution of human genes by GC3 content

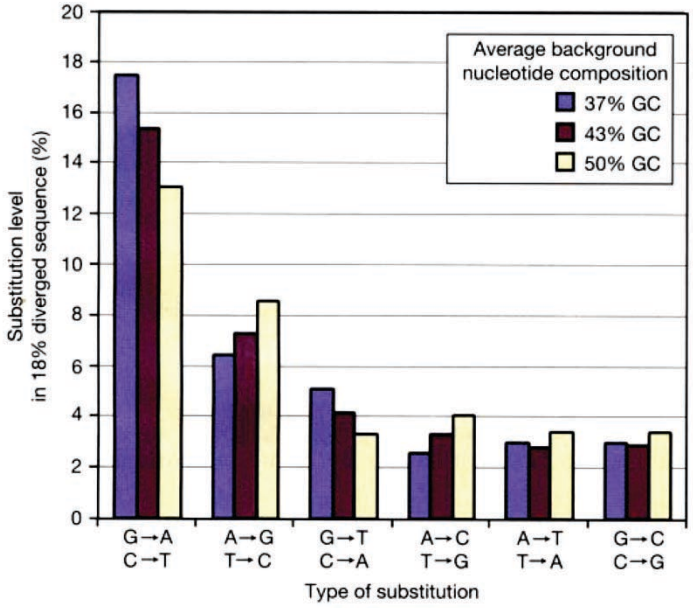


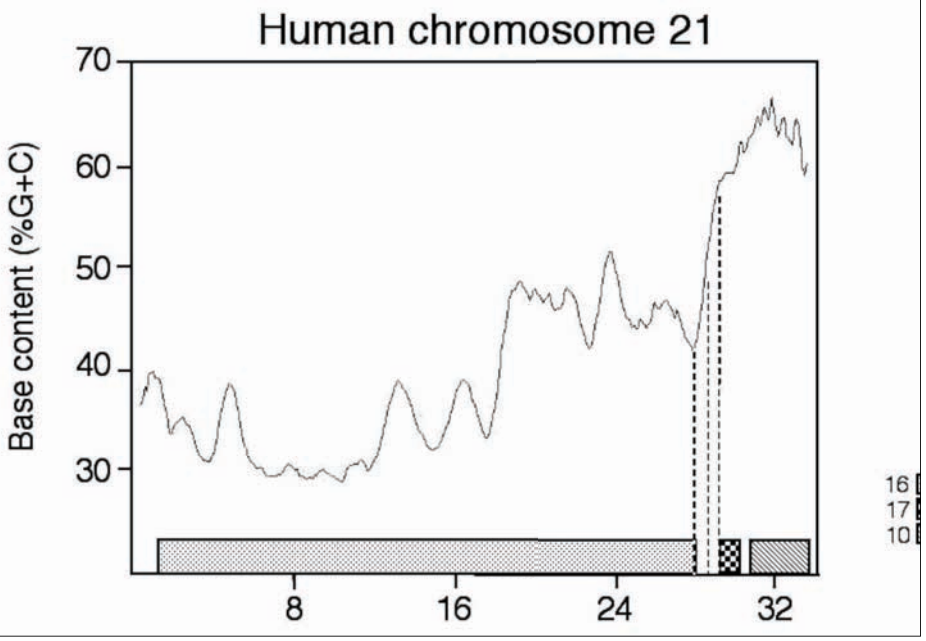
The GC content of exons is correlated with the GC content of the adjacent intron



from Clay et al. (1996)

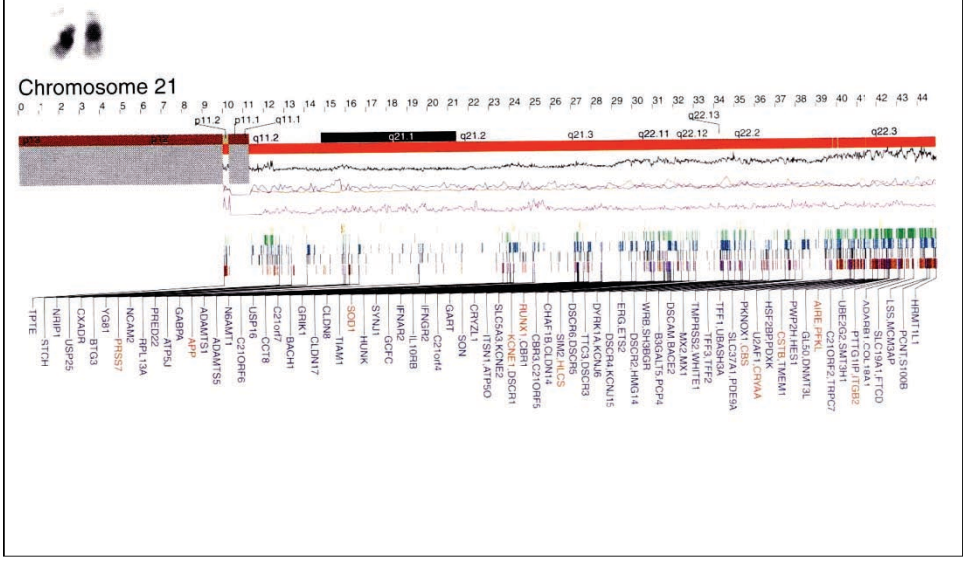
Base substitutions in human transposable elements





Outline

- Chromosome bands and regional variations in GC content
- CpG islands as an indicator of gene density, which is also correlated with promoter type
- Regional variations in recombination, cytosine deamination, and GC content
- Codon bias in mammals, and regional variations in codon bias



Why does gene density correlate with GC content?

- Protein-coding sequences are generally more GC-rich than noncoding sequences. This will cause a higher GC content in regions that happen to be gene dense.
- Genes are associated with CpG islands. This will also cause a higher GC content in regions that happen to be gene dense.
- GC-rich regions tend to have a more GC-rich mutation pressure, because of cytosine deamination. This may increase the previous two effects.
- Transposable elements may be selected against more efficiently in regions of higher recombination (clear in Drosophila, unclear in humans), which also tend to be more GC-rich.
- Transposable elements tend to insert more often in AT-rich sequences, thus lowering their gene density.

Comparison to the chicken genome (Gordon et al., 2007)

- Human chromosome 19 has a gene-dense, GC-rich end, and a gene-poor, AT-rich end.
- The former is orthologous to chicken minichromosome 28, the latter is homologous to a portion of chicken chromosome 11.
- The authors found that most of the genes in both regions were orthologous to human genes on chromosome 19.
- The authors mapped 31 human-chicken syntenic breakpoints (gaps in the human-chicken alignment). All were in the GC-rich region!
- 72 lineage-specific genes were identified (mostly paralogs). These lineage-specific genes were over-represented at or near syntenic breakpoints.
- The results are consistent with a model in which recombination (including illegitimate recombination) is more common in GC-rich regions, and the formation of lineage-specific genes (such as paralogs) often occurs via illegitimate recombination.

DNA methylation and gene regulation

- Actively expressed genes, and particularly their promoters, are generally under-methylated in the specific tissues in which they are expressed.
- This has been implicated as being both cause and effect - in other words, methylation interferes with expression, and expression interferes with methylation.
- Methylation may help to keep inappropriate genes (and transposable elements) turned off.
- DNA methylation helps to regulate X chromosome inactivation in female mammals (myoD, azaC, etc).

CpG islands contain high frequencies of the dinucleotide CpG

- Methyl groups are added to cytosine in most CpG sequences in the human genome.
- Deamination of methyl-C -> T is less efficiently repaired (than C -> U), therefore CpG -> TpG is a hotspot for mutations in the human genome.
- Therefore, CpG is under-represented in the human genome (average ~0.25 of expected frequency).
- One exception to the rule is "CpG islands", which are GC-rich sequences (~60-90% GC) in which CpG occurs at nearly the expected frequency (~0.6 - 1.0 of expected).

CpG islands contain high frequencies of the dinucleotide CpG

- Methyl groups are added to cytosine in most CpG sequences in the human genome.
- Deamination of methyl-C -> T is less efficiently repaired (than C -> U), therefore CpG -> TpG is a hotspot for mutations in the human genome.
- Therefore, CpG is under-represented in the human genome (average ~0.25 of expected frequency).
- One exception to the rule is "CpG islands", which are GC-rich sequences (~60-90% GC) in which CpG occurs at nearly the expected frequency (~0.6 - 1.0 of expected).

Outline

- Chromosome bands and regional variations in GC content
- CpG islands as an indicator of gene density, which is also correlated with promoter type
- Regional variations in recombination, cytosine deamination, and GC content
- Codon bias in mammals, and regional variations in codon bias

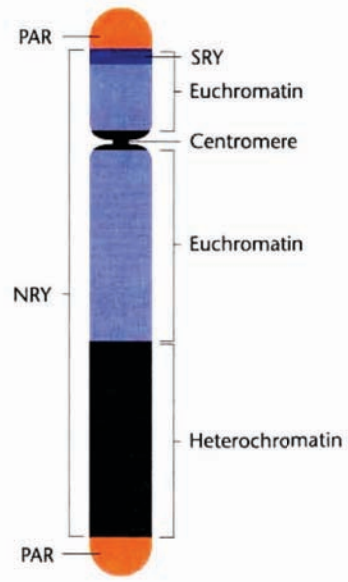


Figure 7–8 The various regions of the human Y chromosome.

Gene conversion occurs because of heteroduplexes formed during recombination

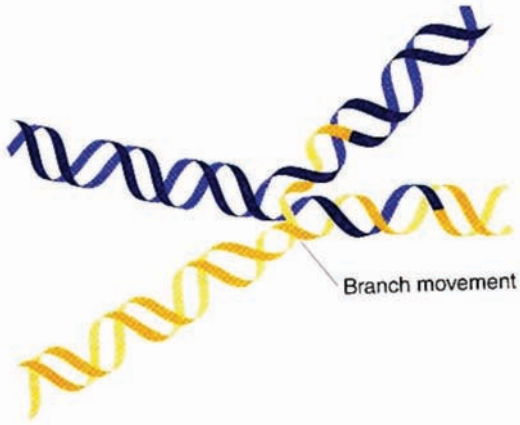


Figure 19-10 Branch migration, the movement of the crossover point between DNA complexes. (After T. Broker, *Journal of Molecular Biology* 81, 1973, 1; from J. D. Watson et al., *Molecular Biology of the Gene*, 4th ed. Copyright © 1987 by Benjamin Cummings.)

Detecting recombination events in African-Americans

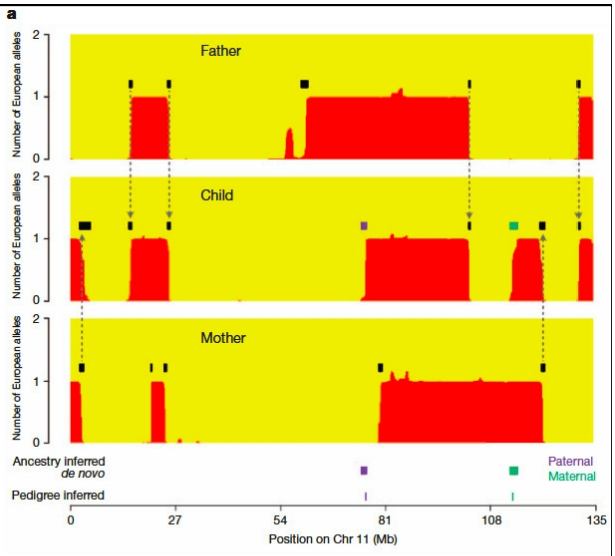
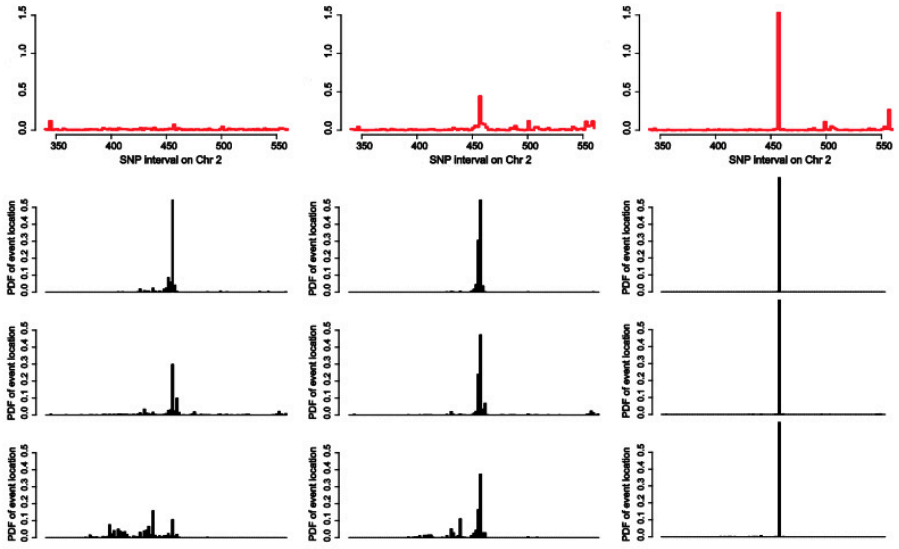


Figure 1 | Building an African-American genetic map. a, HAPMIX detection of crossovers between segments of inferred ancestry is illustrated in a father–mother–child trio. Black segments show inferred crossovers; arrows show transmission of ancestral crossovers from parent to child; purple/green segments show *de novo* events (paternal/maternal origin, respectively) corresponding to events identified directly using two additional children (bottom, 'pedigree inferred').

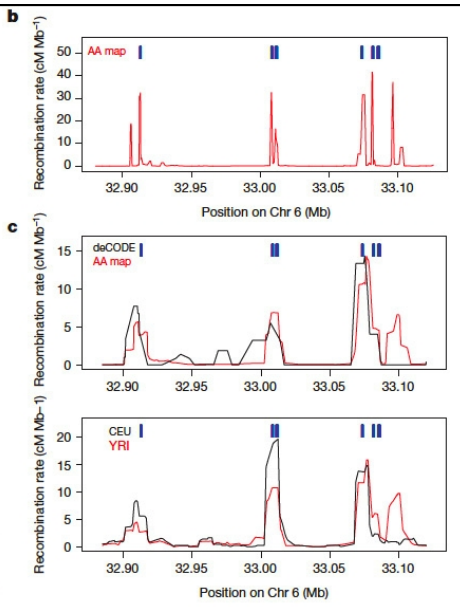
Hinch et al. (2011)
Nature 476, 170-175.

Figure S1
Localization of hotspots by the MCMC



Legend: The top panels show the recombination rate (cM) estimated by the MCMC at different stages in the chain for a small region in Chromosome 2. The bottom panels show three events occurring in different individuals in the same region of the genome. The x-axis shows SNP position and the y-axis shows the estimated probability distribution function (PDF) of localization of each event. The chain was started in the left panel using a uniform recombination rate per base of 1.1cM/Mb. The middle panel shows the state of the chain after 100 iterations and the final panel shows the state of the chain after 10,000 iterations.

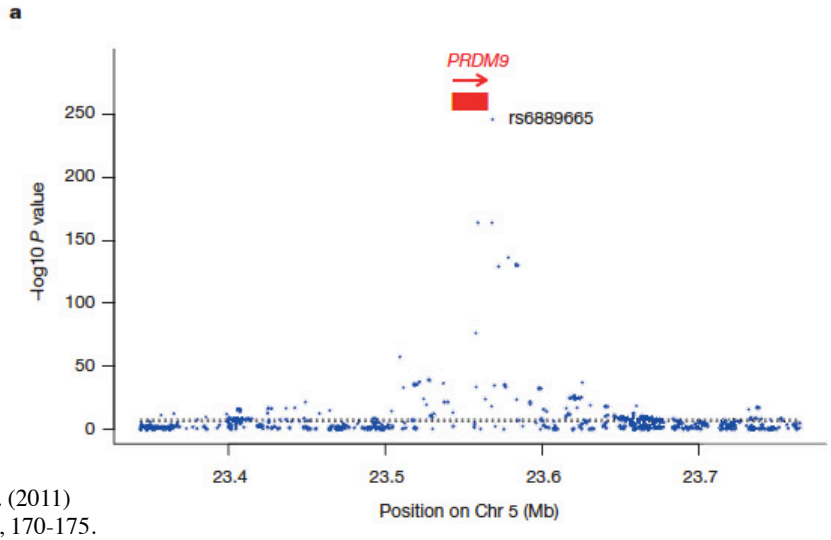
Recombination hotspots in the MHC – one hotspot is African-specific



b, The AA map localizes five hotspots in a region of the MHC whose positions (blue) were previously mapped by sperm typing[1]. c, Comparison of maps shows a hotspot at 33.1 Mb in the African-derived AA and YRI maps, but not the deCODE and CEU maps (all maps smoothed to 10 kb).

Hinch et al. (2011)
Nature 476, 170-175.

GWAS identifies PRDM9 as main locus that was associated with African-specific recombination events.

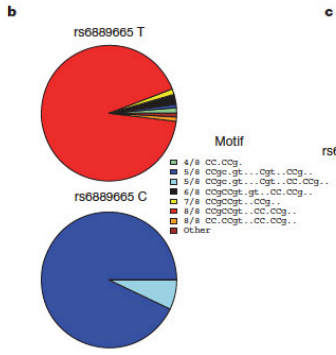


Hinch et al. (2011)
Nature 476, 170-175.

PRDM9 & RNF212

- PRDM9 (PR domain containing protein 9) contains multiple zinc finger (DNA binding) domains, plus multiple PR and Kruppel box (protein interaction) domains.
- PRDM9 has been characterized biochemically as a histone H3 lysine methyl transferase and helps to regulate crossovers during meiosis (among other things).
- RNF212 (ring finger protein 212), like other ring finger proteins, is believed to function as a ubiquitin ligase, in other words it adds ubiquitin side groups to lysine residues.
- Mouse RNF212 is haplo-insufficient and essential for meiotic crossovers. Immunocytological experiments have shown that RNF212 functions to couple chromosome synapsis to the formation of crossover-specific recombination complexes – localization of RNF212 to a subset of synapsis sites is a key early step in the crossover designation process.

rs6889665 T is the derived allele and predicts usage of African-specific recombination hotspots. Other "hot" alleles were derived from it.



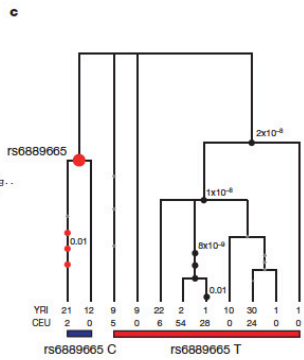


Figure 2 | Association of *PRDM9* genetic variation with hotspot activity. a, A genome-wide association study measuring association of the AE phenotype shows a single genome-wide significant peak at *PRDM9*, with rs6889665 the best-associated SNP. b, Relationship between alleles of rs6889665 and predicted binding target of the *PRDM9* ZF array[9] for West African and European samples. The binding predictions are grouped into 8 clusters according to their best-matching region to the 13-bp motif, and annotated by the number of bases matching the motif. The African-enriched rs6889665 C allele always co-occurs with motifs with a poor (5/8) match to the 13-bp motif. c, Gene tree[25] of the linkage disequilibrium block containing the *PRDM9* ZF array (Methods); numbered circles show SNPs and significant *P* values for association, after conditioning on rs6889665.

Hinch et al. (2011) *Nature* 476, 170-175.

African vs. European recombination hotspots for PRDM9.

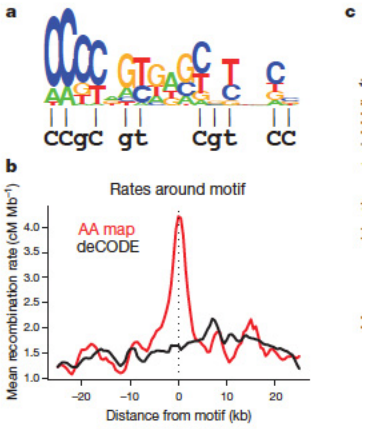
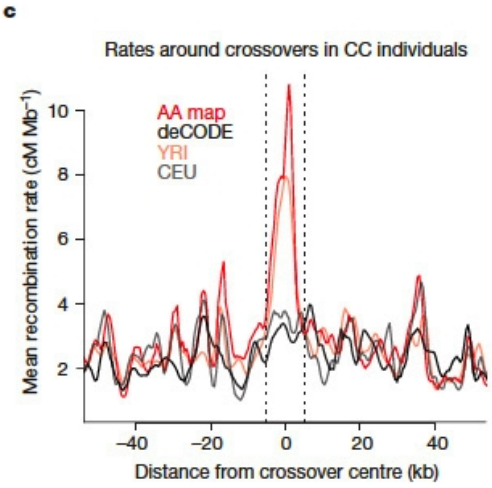


Figure 3 | A sequence motif specifying the positions of African-enriched hotspots. **a**, Logo plot showing a degenerate 17-bp hotspot motif, with stack height proportional to −log *P* value, and relative letter height proportional to the mean crossover rate increase given each base. Below is the bioinformatic *PRDM9* binding prediction for the alleles associated with rs6889665 allele C (from Fig. 2b), matching this motif at 10/11 bases (lines). **b**, Average crossover rate (in 2-kb sliding windows) in the AA (red line) and deCODE (black line) maps surrounding the 500 strongest motif matches.

Hinch et al. (2011) *Nature* 476, 170-175.

Agreement between AA vs. YRI hotspots



c, In seven rs6889665 CC individuals from the pedigree study, we localized 82 crossovers to within 10 kb, and plot average AA, YRI, deCODE and CEU map rates. There is no strong peak above local background in the deCODE or CEU maps.

Hinch et al. (2011) *Nature* 476, 170-175.

Outline

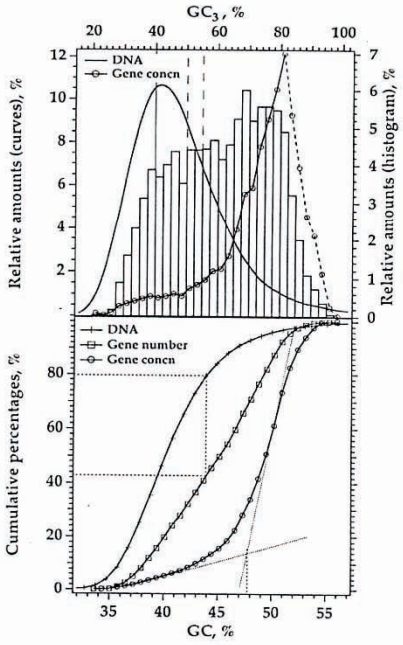
- Chromosome bands and regional variations in GC content
- CpG islands as an indicator of gene density, which is also correlated with promoter type
- Regional variations in recombination, cytosine deamination, and GC content
- Codon bias in mammals, and regional variations in codon bias

Codon bias

- There are two types of codon bias: biased use of non-synonymous codons (amino acids), and biased use of synonymous codons (third codon position). Most published studies focus on the latter.
- A slightly confusing point - published data tables are often stated as usage per 10,000 codons, which includes both types of codon bias.
- Both types of codon bias are influenced by both natural selection and mutational pressures.
- Synonymous codon bias is greater in genes that are expressed at high levels.
- Synonymous codon bias in the human genome varies substantially, depending on the GC content of the surrounding DNA.
- In *E. coli*, codon bias varies with the level of expression (high bias correlates with genes expressed in higher levels).

Phe	171	UUU	AAA	0	Ser	147	UCU	AGA	10	Tyr	124	UAU	AUA	1	Cys	99	UGU	ACA	0
	203	UUC	GAA	14		172	UCC	GGA	0		158	UAC	GUA	11		119	UGC	GCA	30
Leu	73	UUA	UAA	8		118	UCA	UGA	5	stop	0	UAA	UUA	0	stop	0	UGA	UCA	0
	125	UUG	CAA	6		45	UCG	CGA	4	stop	0	UAG	CUA	0	Trp	122	UGG	CCA	7
Leu	127	CUU	AAG	13	Pro	175	CCU	AGG	11	His	104	CAU	AUG	0	Arg	47	CGU	ACG	9
	187	CUC	GAG	0		197	CCC	GGG	0		147	CAC	GUG	12		107	CGC	GCG	0
	69	CUA	UAG	2		170	CCA	UGG	10	Gln	121	CAA	UUG	11		63	CGA	UCG	7
	392	CUG	CAG	6		69	CCG	CGG	4		343	CAG	CUG	21		115	CGG	CCG	5
Ile	165	AUU	AAU	13	Thr	131	ACU	AGU	8	Asn	174	AAU	AUU	1	Ser	121	AGU	ACU	0
	218	AUC	GAU	1		192	ACC	GGU	0		199	AAC	GUU	33		191	AGC	GCU	7
	71	AUA	UAU	5		150	ACA	UGU	10	Lys	248	AAA	UUU	16	Arg	113	AGA	UCU	5
Met	221	AUG	CAU	17		63	ACG	CGU	7		331	AAG	CUU	22		110	AGG	CCU	4
Val	111	GUU	AAC	20	Ala	185	GCU	AGC	25	Asp	230	GAU	AUC	0	Gly	112	GGU	ACC	0
	146	GUC	GAC	0		282	GCC	GGC	0		262	GAC	GUC	10		230	GGC	GCC	11
	72	GUA	UAC	5		160	GCA	UGC	10	Glu	301	GAA	UUC	14		168	GGA	UCC	5
	288	GUG	CAC	19		74	GCG	CGC	5		404	GAG	CUC	8		160	GGG	CCC	8

Distribution of human genes by GC3 content



Discussion questions - week 9

- Why do some human genes have much higher third codon GC content (GC3) than the flanking DNA, but other genes do not? Would you expect such genes to be associated with CpG islands? What does this tell us about codon bias in the human genome?
- What is "biased" about gene conversion in the human genome? Is biased gene conversion good for you? If so, why? If not, why not?
- Does the "pseudoautosomal region" of the human Y chromosome have a higher or lower recombination rate than the rest of the Y chromosome? Why? How does this appear to affect the GC content along the Y chromosome?
- Suggest several plausible reasons (as many as possible) why AT-rich regions in mammalian (and bird) genomes tend to have low to extremely low gene densities.
- Suggest several plausible reasons (as many as possible) why AT-rich regions in mammalian (and bird) genomes tend to correlate with properties such as chromatin structure, replication time, recombination rate, etc.

Discussion questions (continued)

- Discuss the evidence that a derived PRDM9 SNP controls African-specific recombination hotspots. How do we know that this allele is derived? Is this SNP allele likely to be functional? How did African-Americans provide unique evidence in this study? How does this study alter our understanding of the factors affecting the length of haplotypes in African populations? Why is the number of zinc fingers in PRDM9 important?