Project of 1000 genomes

A map of human genome variation from population-scale sequencing. The 1000 Genomes Project Consortium. 2010. Nature 467:1061-1073.

Journal club in bioinformatics

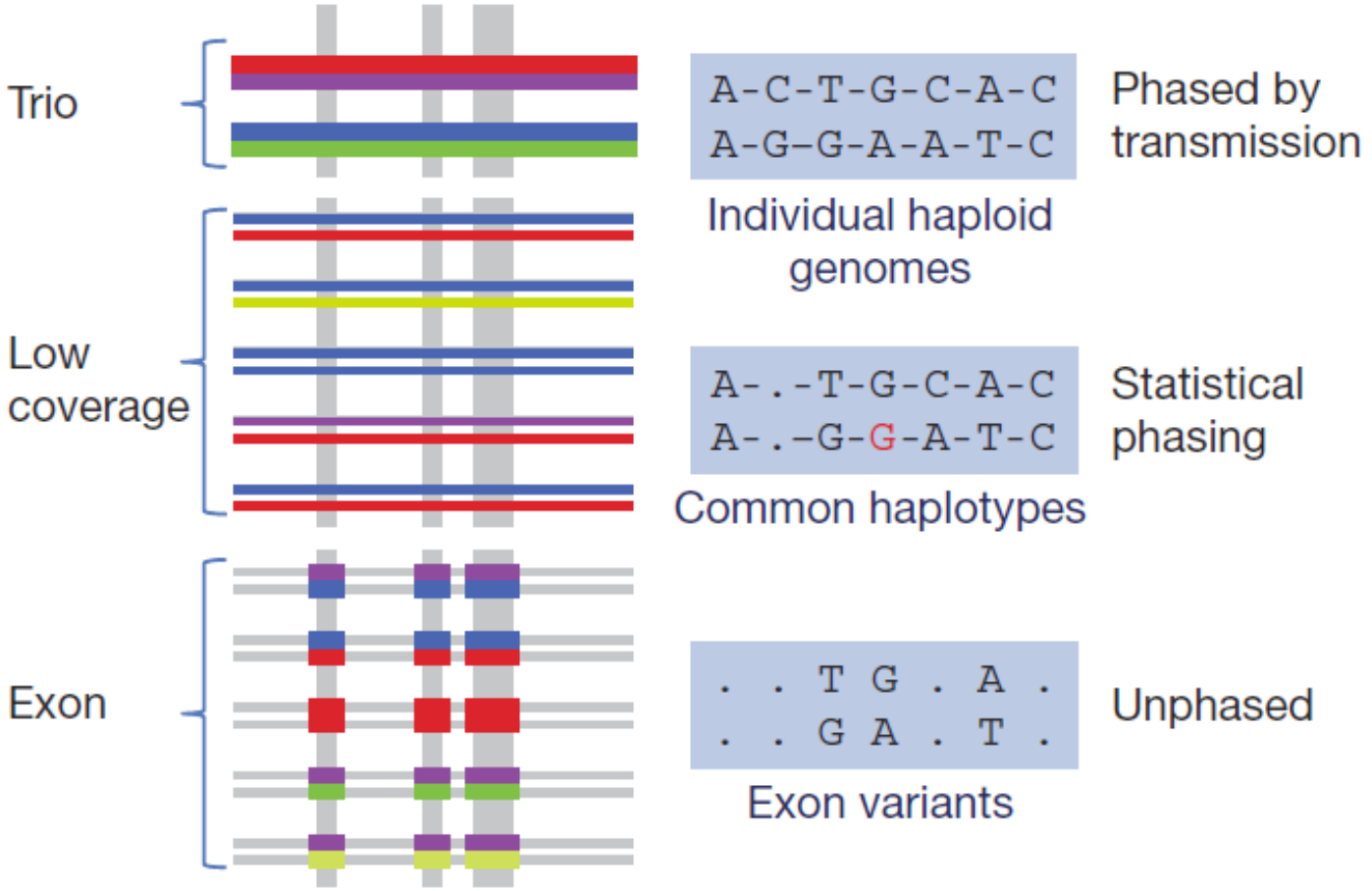
22.11.2010

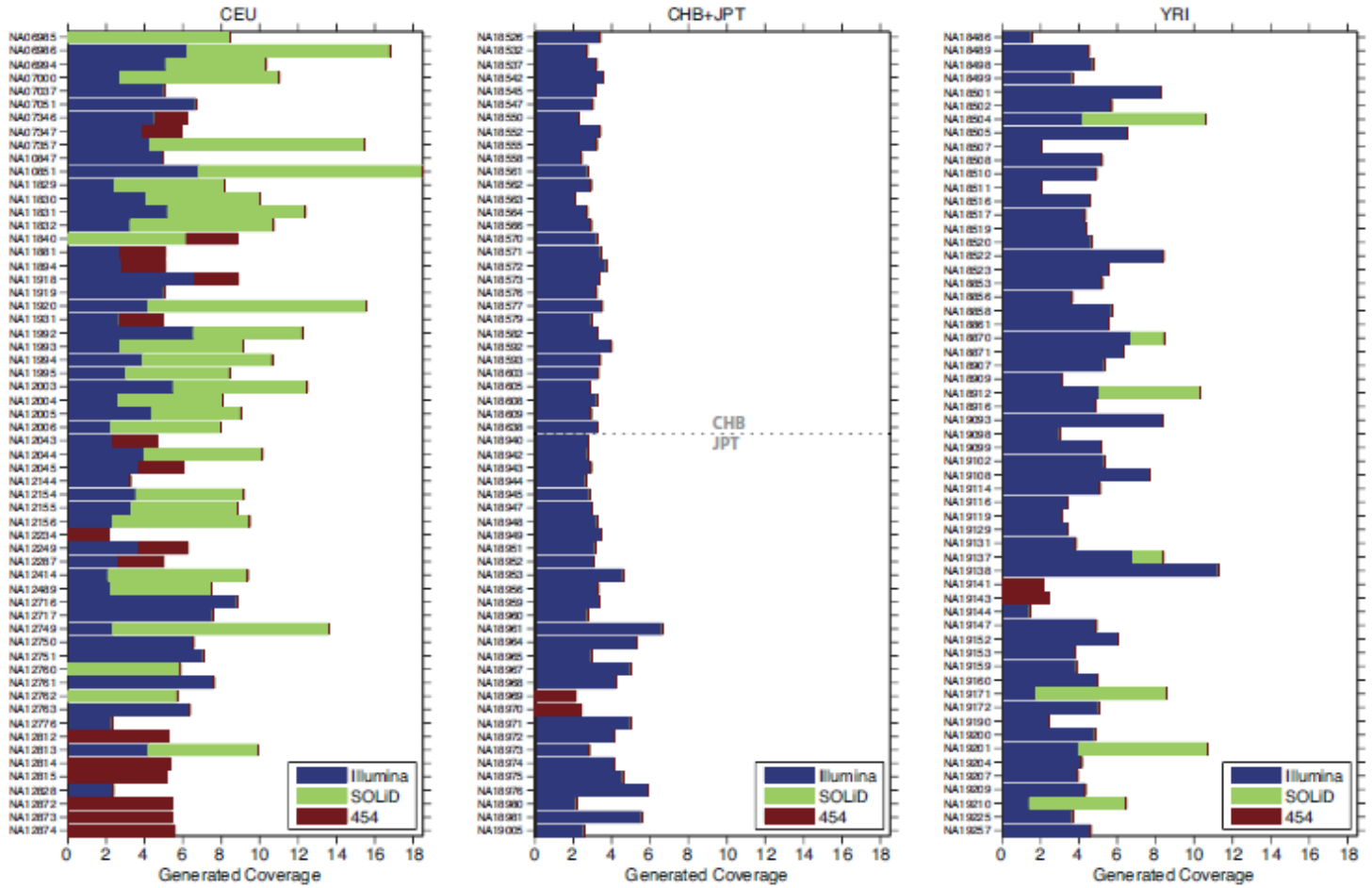
Tarmo Puurand

Projects

- Trio: Whole-genome shotgun sequencing at high coverage (average 42x) of two families wiht daugthers (CEU and YRI).
- Low-coverage: whole-genome shotgun sequencing at low coverage (2-6x). YRI (59), CEU (60), CHB (30) and JPT (30). All individuals are unrelated.
- Exon: targeted capture of 8140 exons from 906 randomly selected genes (average > 50x) in 697 individuals (YRI, LWK, CEU, TSI, CHB, JPT,CHD)

Three strategies in pilot project





Supplementary Figure 2. Amount of sequence coverage generated (mapped bases/2.85 Gb) in the low-coverage project by sample and sequencing technology; blue = Illumina, green = SOLiD, red = 454. Note that populations and samples differ considerably in coverage (CEU highest, CHB+JPT lowest, sample coverage from c. 2x to 18x) and the balance of technologies. Many samples have data from two technologies.

Projects workflow

- Discovery: alignment of sequence reads to the reference genome and identification of candidate sites or regions at which one or more samples differ from the reference sequence;
- Filtering: use of quality control measures to remove candidate sites that were probably false positives;
- Genotyping: estimation of the alleles present in each individual at variant sites or regions;
- Validation: assaying a subset of newly discovered variants using an independent technology, enabling the estimation of the false discovery rate (FDR).

Accessible genome

- Sequence reads were aligned to the NCBI36 reference genome and made available in the BAM file format.
- Accessible genome- in low-coverage analysis contains ca 85% of the reference genome and 93% of the coding sequences. HapMapII 99% of sites are included. Of inaccessible sites, over 97% are annotated as high-copy repeats or segmental duplications.
- Mapping: Illumina -> Maq v0.7, 454 -> SSAHA v2.4, SOLid -> Corona_Lite v.4.0r2.0

Calibration

- base quality scores reported by the image processing software were empirically recalibrated by tallying the proportion that mismatched the reference sequence (at non-dbSNP sites) as a function of the reported quality score, position in read and other characteristics.
- at potential variant sites, local realignment of all reads was performed jointly across all samples, allowing for alternative alleles that contained indels. This realignment step substantially reduced errors, because local misalignment, particularly around indels, can be a major source of error invariant calling.
- by initially analysing the data with multiple genotype and variant calling algorithms and then generating a consensus of these results, the project reduced genotyping error rates by 30–50% compared to those currently achievable using any one of the methods alone.

Local realignment and assembly

- Local realignment was used to generate canditate alternative haplotypes in the process of calling short (1-50 bp) indels, as well as local *de novo* assembly to resolve breakpoints for deletions greater than 50 bp.
- Full genome *de novo* assembly was performed, resulting in the identification of 3,7 MB of novel sequences not maching reference genome.

Pilot projects variants summary

Table 1 | Variants discovered by project, type, population and novelty

a Summary of project data including combined exon populations

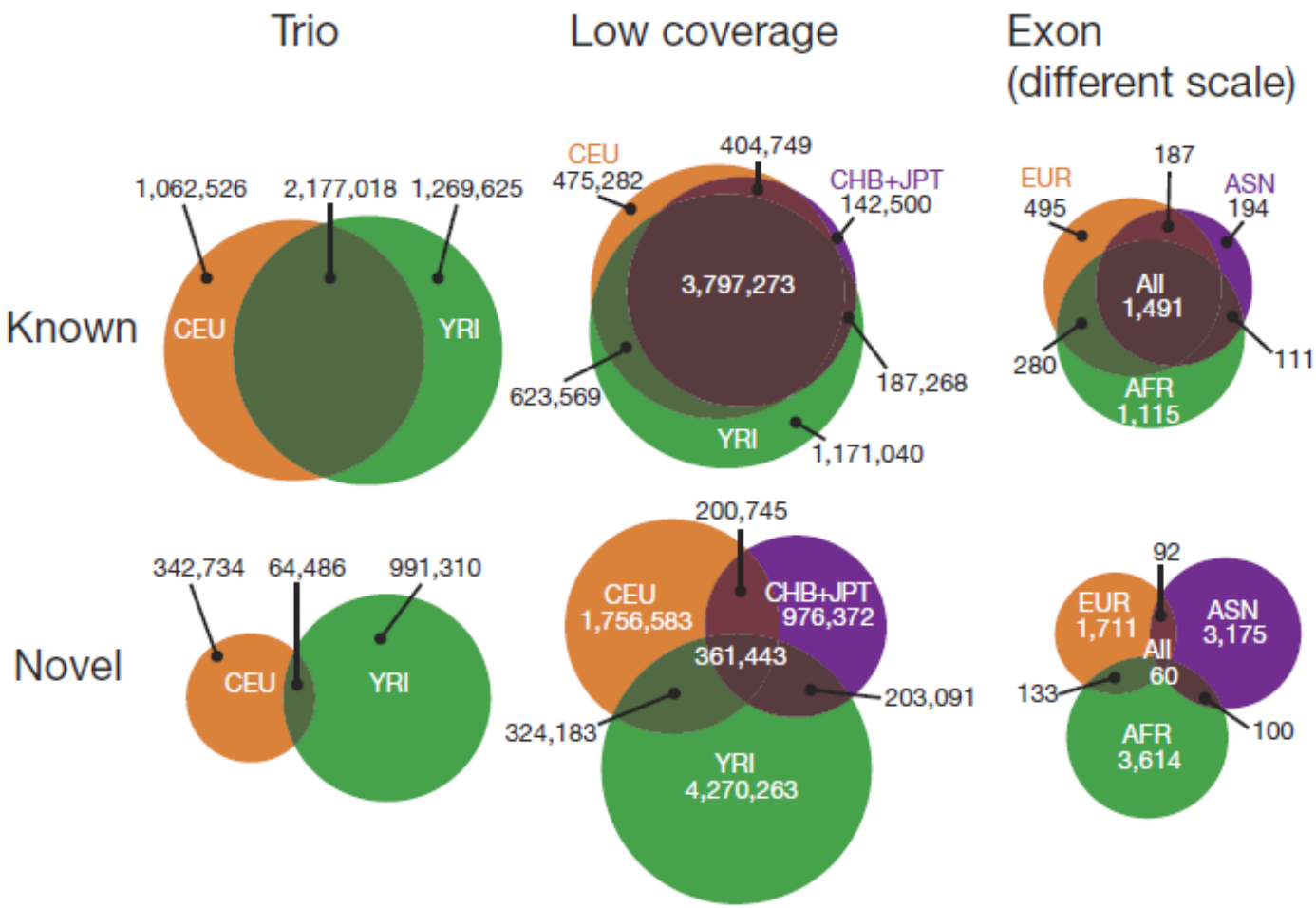
	Low coverage				Trios				
Statistic	CEU	YRI	CHB+JPT	Total	CEU	YRI	Total	Exon (total)	Union across projects
Samples	60	59	60	179	3	3	6	697	742
Total raw bases (Gb)	1,402	874	596	2,872	560	615	1,175	845	4,892
Total mapped bases (Gb)	817	596	468	1,881	369	342	711	56	2,648
Mean mapped depth (×)	4.62	3.42	2.65	3.56	43.14	40.05	41.60	55.92	NA
Bases accessed (% of genome)	2.43 Gb (86%)	2.39 Gb (85%)	2.41 Gb (85%)	2.42 Gb (86.0%)	2.26 Gb (79%)	2.21 Gb (78%)	2.24 Gb (79%)	1.4 Mb	NA
No. of SNPs (% novel)	7,943,827 (33%)	10,938,130 (47%)	6,273,441 (28%)	14,894,361 (54%)	3,646,764 (11%)	4,502,439 (23%)	5,907,699 (24%)	12,758 (70%)	15,275,256 (55%)
Mean variant SNP sites per individual	2,918,623	3,335,795	2,810,573	3,019,909	2,741,276	3,261,036	3,001,156	763	NA
No. of indels (% novel)	728,075 (39%)	941,567 (52%)	666,639 (39%)	1,330,158 (57%)	411,611 (25%)	502,462 (37%)	682,148 (38%)	96 (74%)	1,480,877 (57%)
Mean variant indel sites per individual	354,767	383,200	347,400	361,669	322,078	382,869	352,474	3	NA
No. of deletions (% novel)	ND	ND	ND	15,893 (60%)	6,593 (41%)	8,129 (50%)	11,248 (51%)	ND	22,025 (61%)
No. of genotyped deletions (% novel)	ND	ND	ND	10,742 (57%)	ND	ND	6,317 (48%)	ND	13,826 (58%)
No. of duplications (% novel)	259 (90%)	320 (90%)	280 (91%)	407 (89%)	187 (93%)	192 (91%)	256 (92%)	ND	501 (89%)
No. of mobile element insertions (% novel)	3,202 (79%)	3,105 (84%)	1,952 (76%)	4,775 (86%)	1,397 (68%)	1,846 (78%)	2,531 (78%)	ND	5,370 (87%)
No. of novel sequence insertions (% novel)	ND	ND	ND	ND	111 (96%)	66 (86%)	174 (93%)	ND	174 (93%)

b Exon populations separately

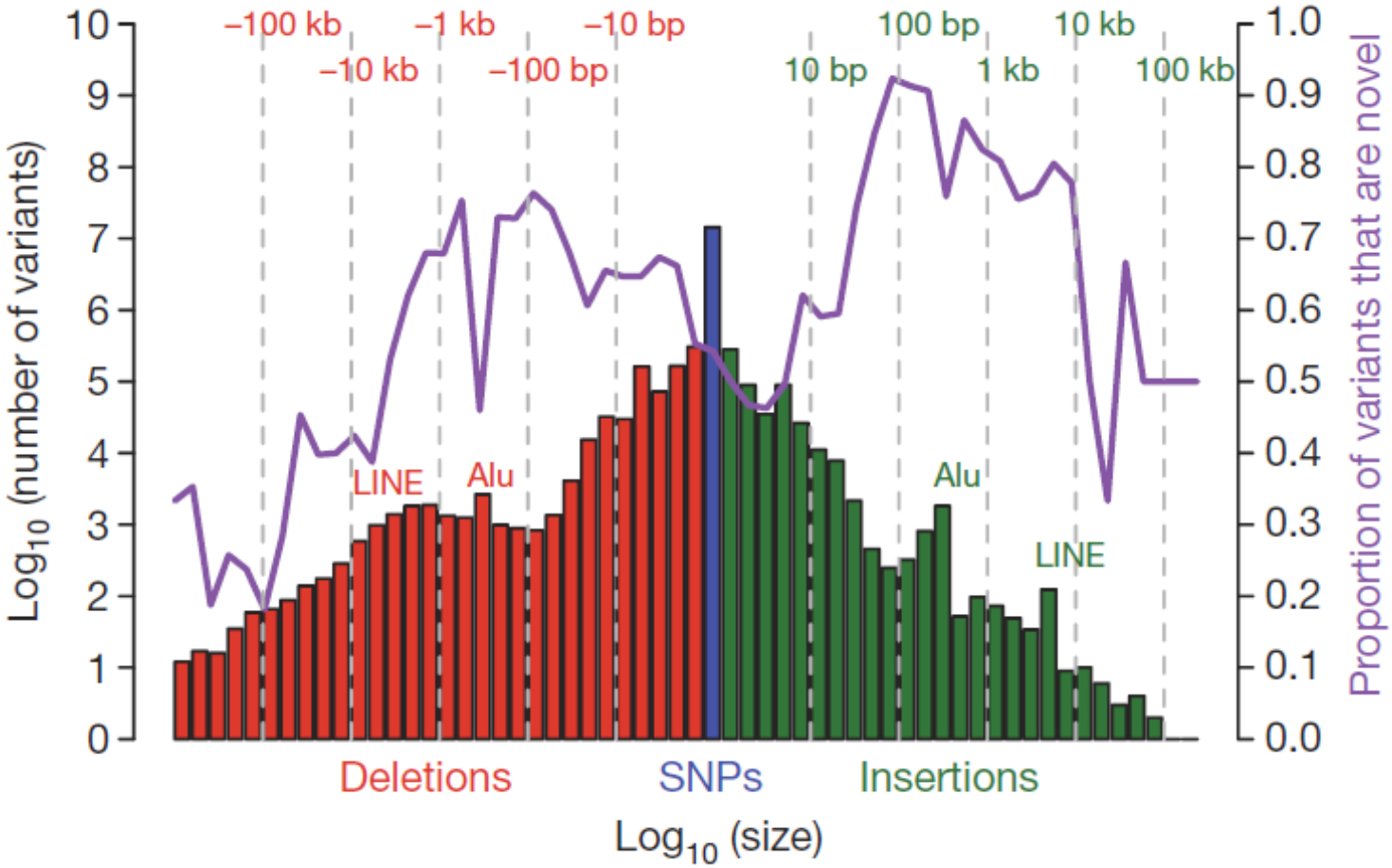
Statistic	CEU	TSI	LWK	YRI	CHB	CHD	JPT
Samples	90	66	108	112	109	107	105
Total collected bases (Gb)	151	64	53	147	93	127	211
Mean mapped depth on target (×)	73	71	32	62	47	62	53
No. of SNPs (% novel)	3,489 (34%)	3,281 (34%)	5,459 (50%)	5,175 (46%)	3,415 (47%)	3,431 (50%)	2,900 (42%)
Variant SNP sites per individual	715	727	902	794	713	770	694
No. of indels (no. novel)	23 (10)	22 (11)	24 (16)	38 (21)	30 (16)	26 (13)	25 (11)
Variant indel sites per individual	3	3	3	3	3	2	3

NA, not applicable; ND, not determined.

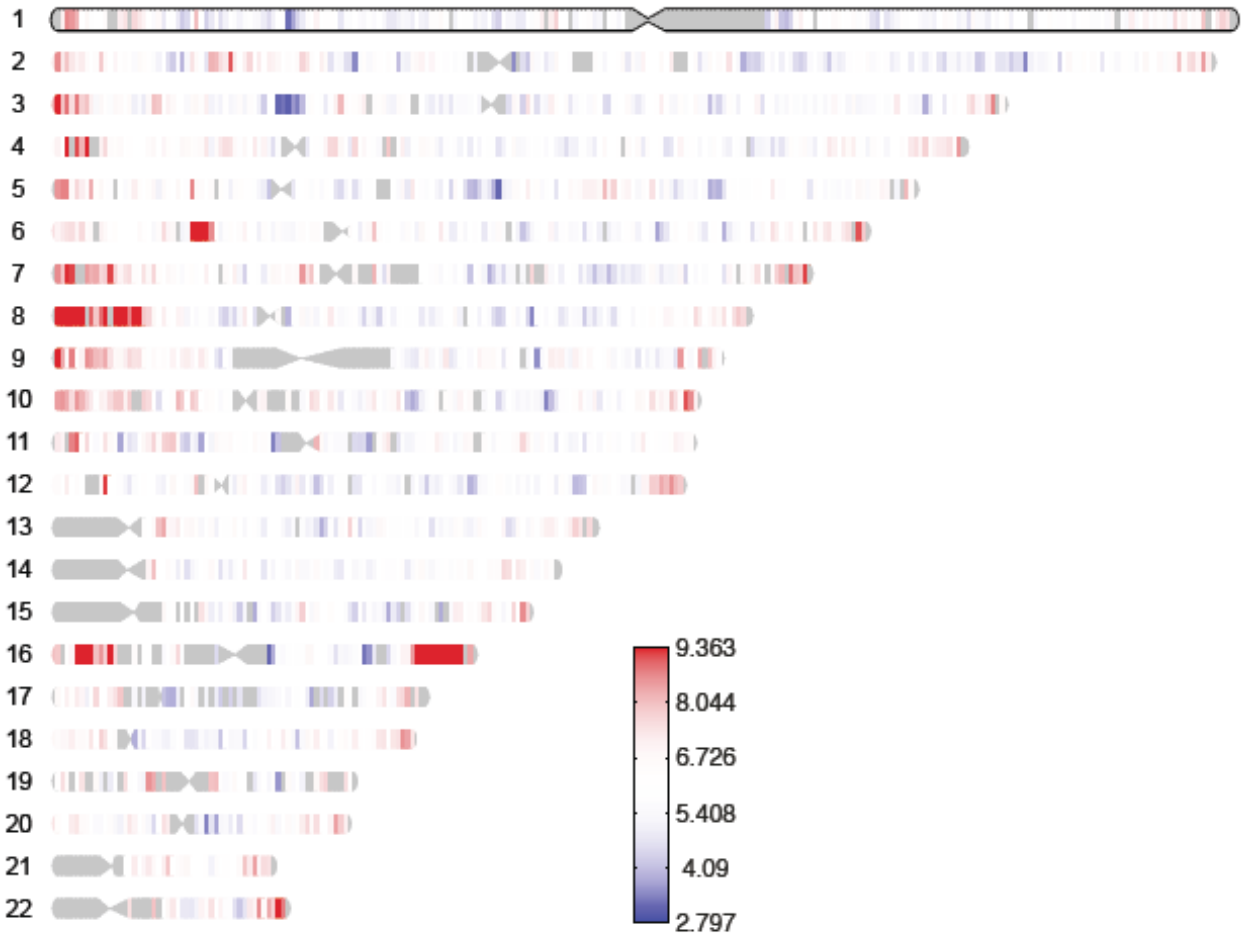
Variant novelty



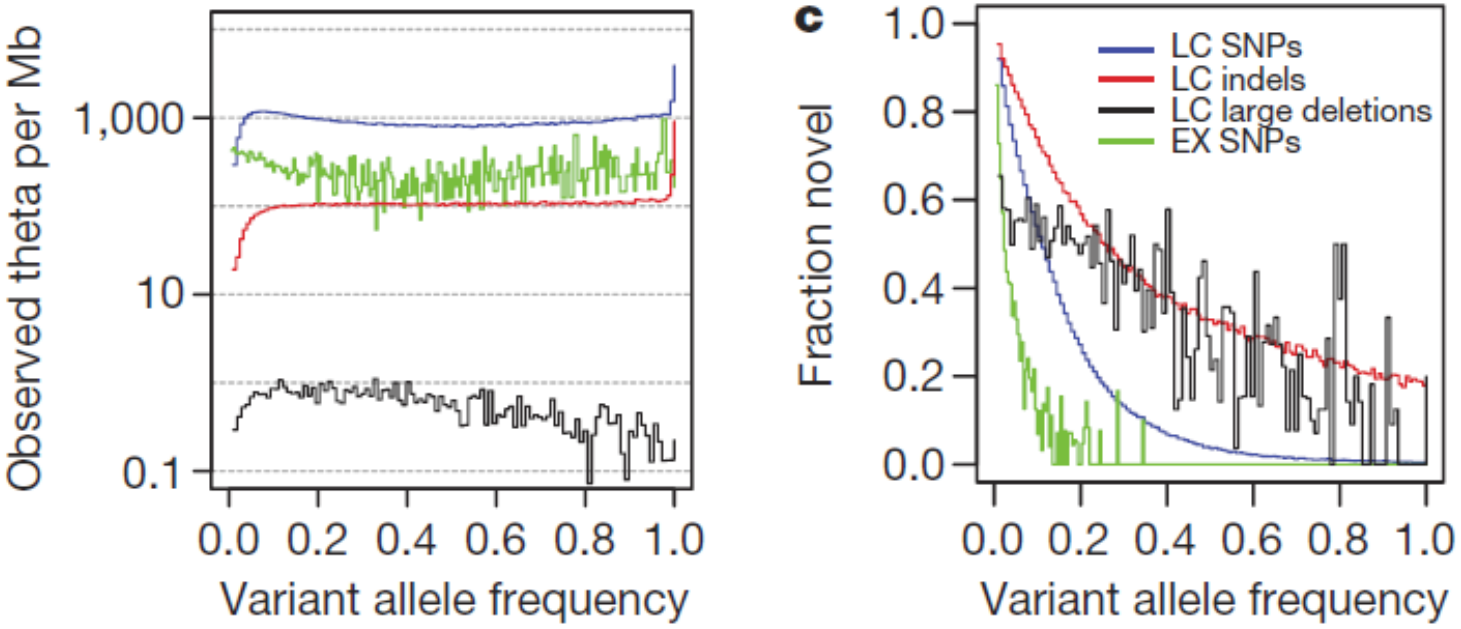
Variant novelty



SNP density


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
9.363
8.044
6.726
5.408
4.09
2.797

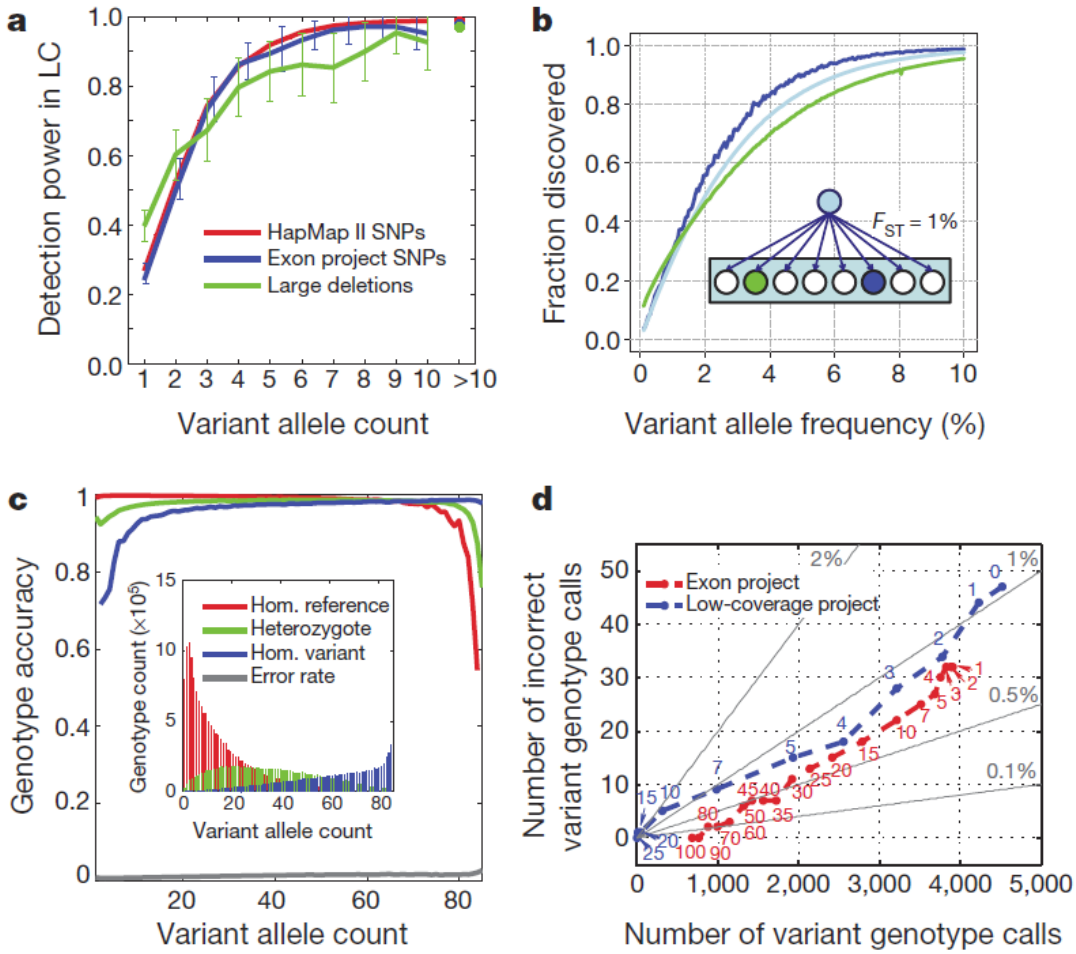
Variants distribution



ChrMT, chrY

- Deep coverage of the mitochondrial genome allowed manually curate sequences for 163 samples
- Length heteroplasmy was detected in 79% of individuals compared with 52% using capillary sequencing, largely in the control region. Base-substitution heteroplasmy was observed in 45% of samples, seven times higher than reported in the control region alone, and was spread throughout the molecule.
- The Y chromosome was sequenced at an average depth of 1.83 in the 77 males in the low-coverage project, and 15.23 depth in the two trio fathers. Using customized analysis methods, we identified 2,870 variable sites, 74% novel, with 55out of 56 passing independent validation.

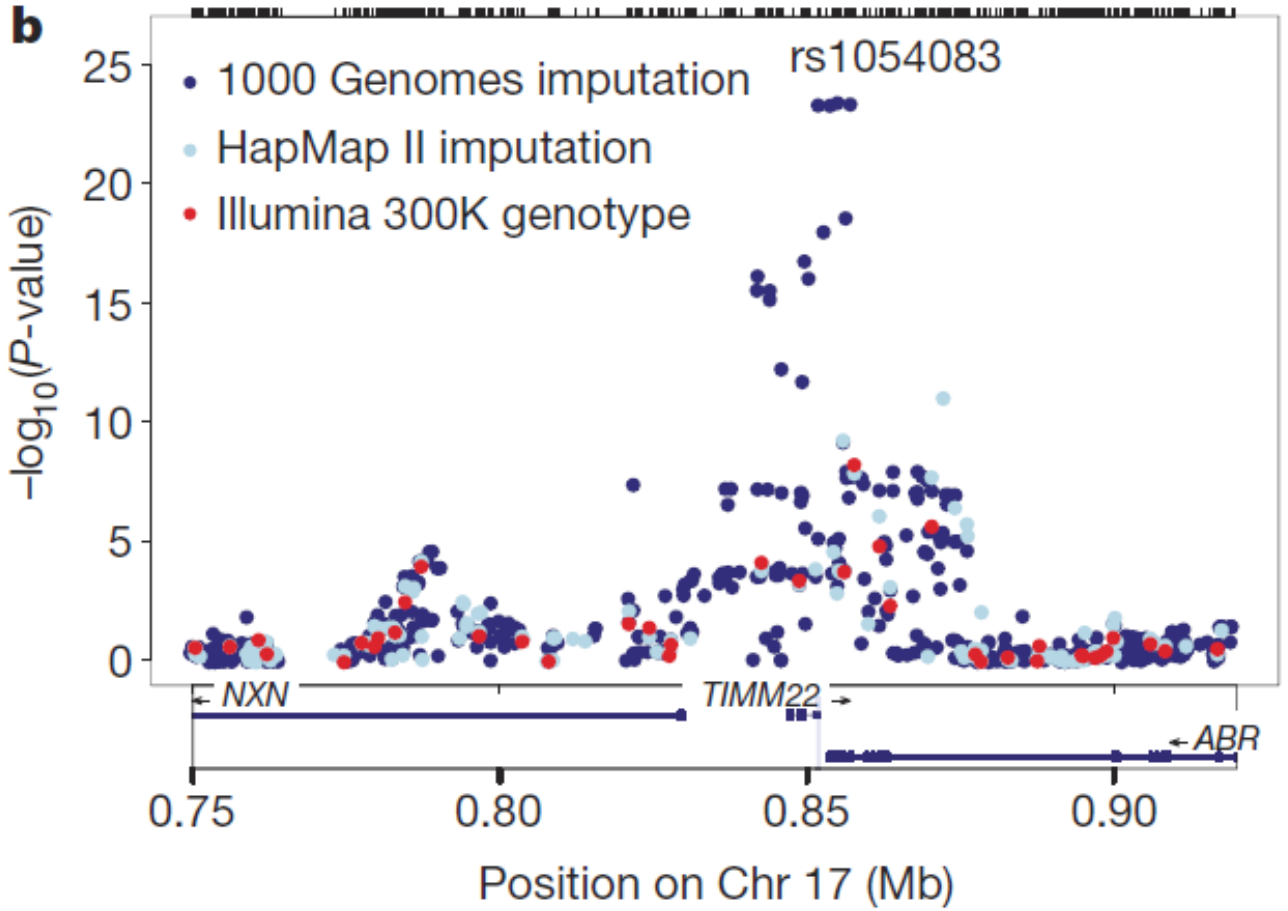
Power and accuracy of detected variants



Functional variants

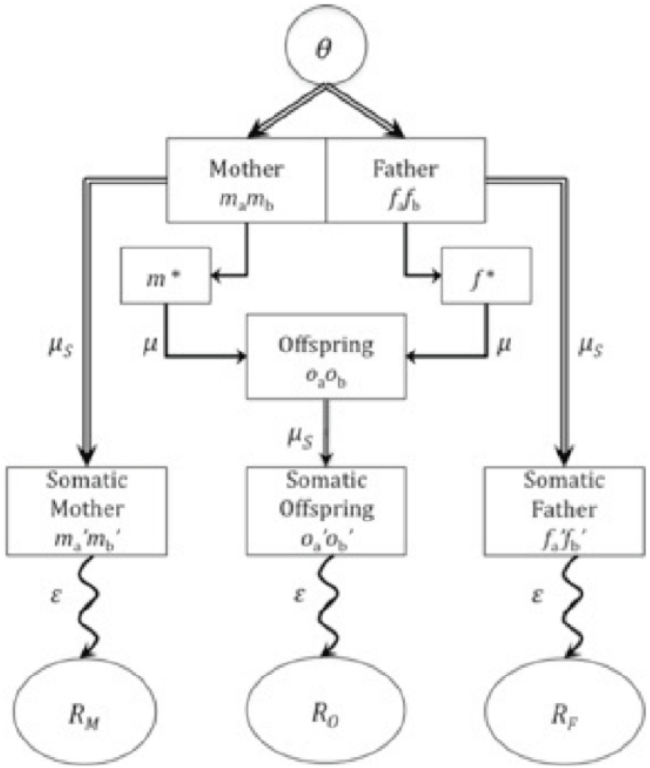
Class	Combined total	Combined novel	Low coverage		High-coverage trio		Exon capture		
			Total	Interquartile*	Total	Individual range	Total	Interquartile*	GENCODE extrapolation
Synonymous SNPs	60,157	23,498	55,217	10,572–12,126	21,410	9,193–12,500	5,708	461–532	11,553–13,333
Non-synonymous SNPs	68,300	34,161	61,284	9,966–10,819	19,824	8,299–10,866	7,063	396–441	9,924–11,052
Small in-frame indels	714	383	666	198–205	289	130–178	59	1–3	~25–75
Stop losses	77	40	71	9–11	22	4–14	6	0–0	~0–0
Stop-introducing SNPs	1,057	755	951	88–101	192	67–100	82	2–3	~50–75
Splice-site-disrupting SNPs	517	399	500	41–49	82	28–45	3	1–1	~50
Small frameshift indels	954	551	890	227–242	433	192–280	37	0–1	~0–25
Genes disrupted by large deletions	147	71	143	28–36	82	33–49	ND	ND	ND
Total genes containing LOF variants	2,304	NA	1,795	272–297	483	240–345	77	3–4	~75–100
HGMD ‘damaging mutation’ SNPs	671	NA	578	57–80	161	48–82	99	2–4	~50–100

Association studies and imputation

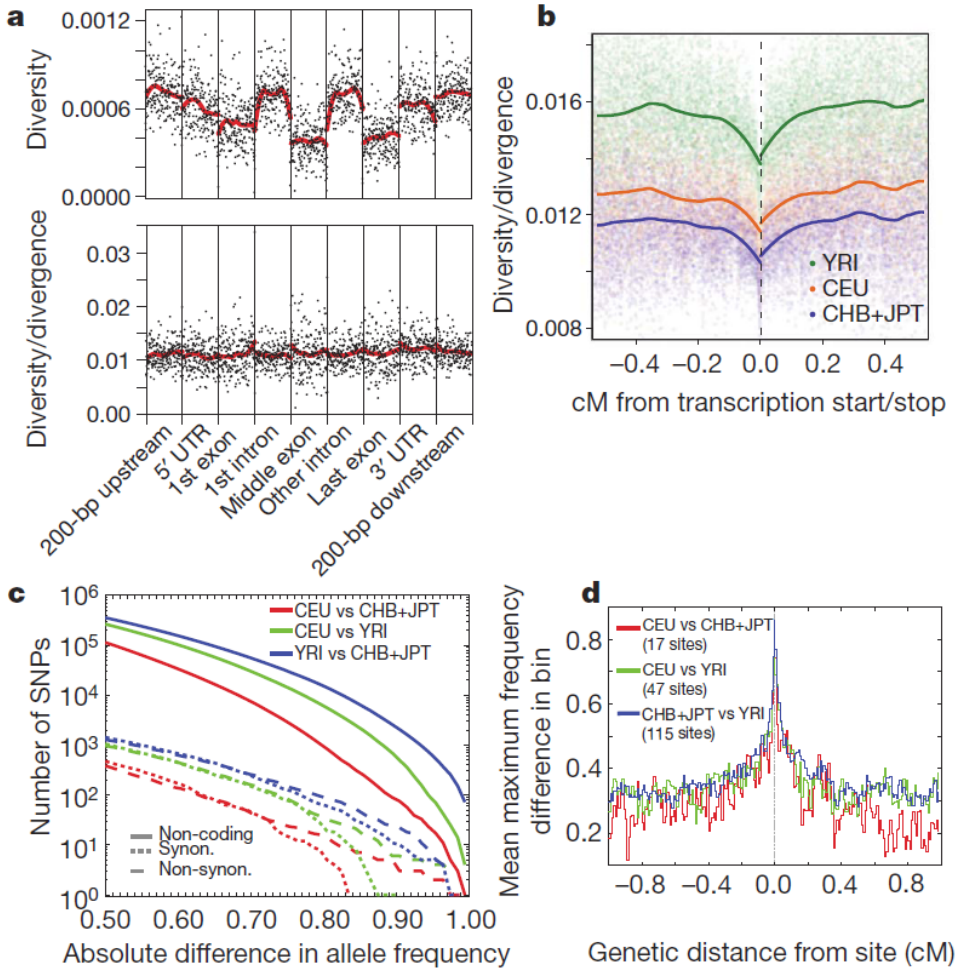


de novo mutations in trio samples

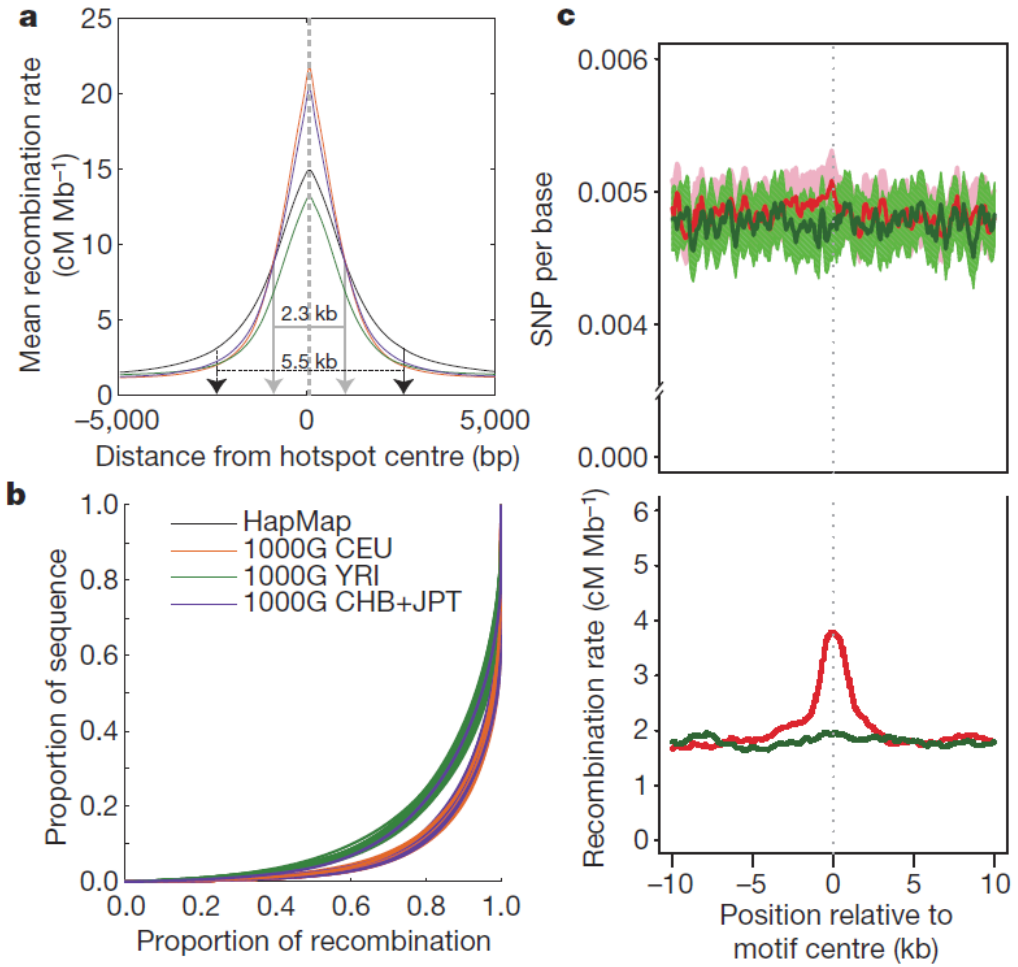
Population	Canditate mutations	Validated with re-sequencing	Confirmed true	Mutation rate per generation
CEU	3236	1001	49	1.2×10^{-8}
YRI	2750	669	35	1.0×10^{-8}



Variation around genes

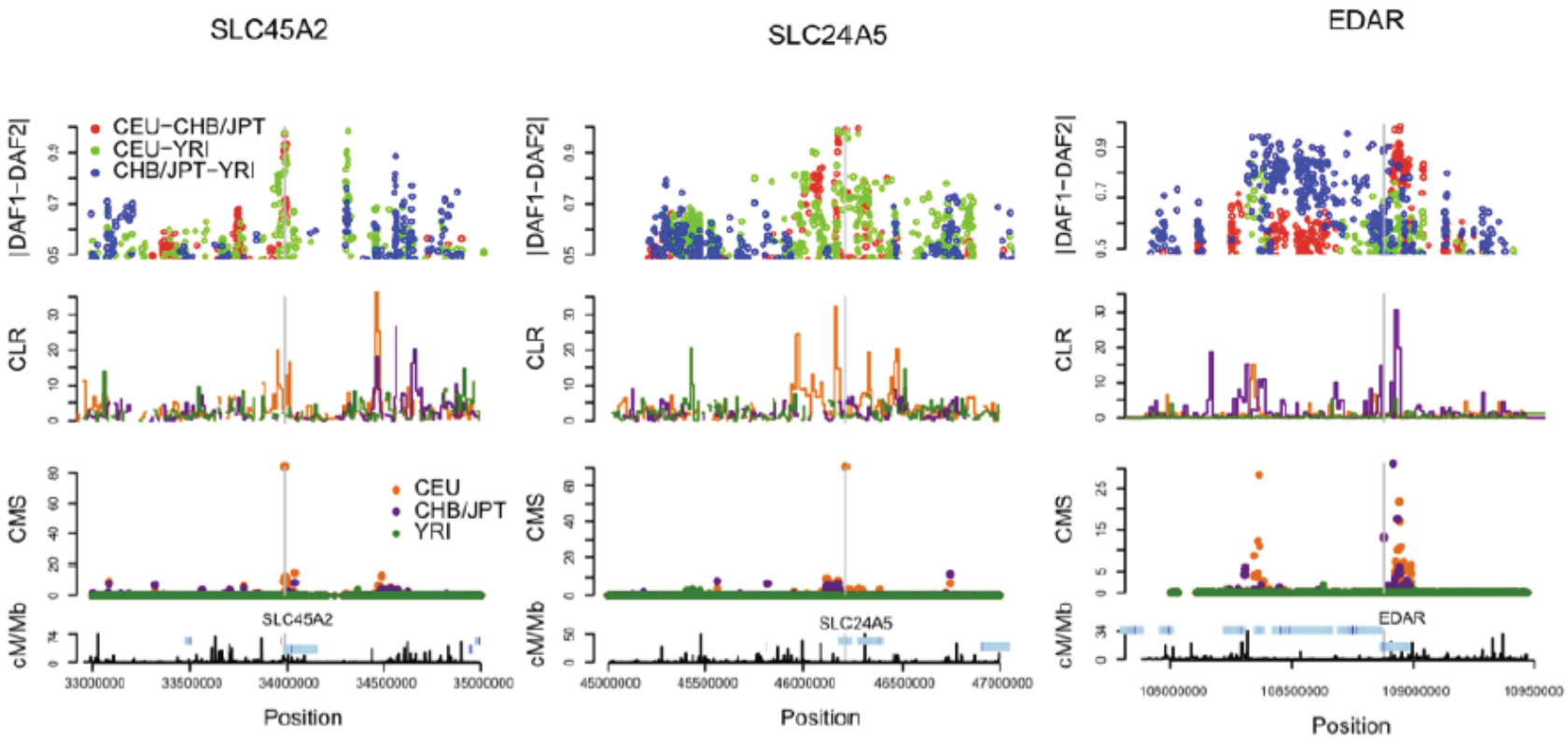


Recombination



Recombination. a, Improved resolution of hotspot boundaries. The average recombination rate estimated from low-coverage project data around recombination hotspots detected in HapMap II. Recombination hotspots were narrower, and in CEU (orange) and CHB1JPT (purple) more intense than previously estimated. See panel b for key. b, The concentration of recombination in a small fraction of the genome, one line per chromosome. If recombination were uniformly distributed throughout the genome, then the lines on this figure would appear along the diagonal. Instead, most recombination occurs in a small fraction of the genome. Recombination rates in YRI (green) appeared to be less concentrated in recombination hotspots than CEU(orange) or CHB1JPT (purple). HapMap II estimates are shown in black. c, The relationship between genetic variation and recombination rates in the YRI population. The top plot shows average levels of diversity, measured as mean number of segregating sites per base, surrounding occurrences of the previously described hotspot motif40 (CCTCCCTNNCCAC, red line) and a closely related, but not recombinogenic, DNA sequence (CTTCCCTNNCCAC, green line). The lighter red and green shaded areas give 95% confidence intervals on diversity levels. The bottom plot shows estimated mean recombination rates surrounding motif occurrences, with colours defined as in the top plot.

Selection


SLC45A2
SLC24A5
EDAR
|DAF1−DAF2|
CEU−CHB/JPT
CEU−YRI
CHB/JPT−YRI
CLR
CMS
CEU
CHB/JPT
YRI
cM/Mb
SLC45A2
SLC24A5
EDAR
Position
Position
Position

Full 1000 genomes project

Populations in the 1000 Genomes Project						
			Number of Samples			
Full Population Name	**Short Population Name**	**Abbreviation**	**Trio Pilot**	**LowCov Pilot**	**Exon Pilot**	**Full Project**
Han Chinese in Beijing, China	Han Chinese	CHB		30	109	100
Han Chinese South	Southern Han Chinese	CHS				100
Chinese Dai in Xishuangbanna, China	Dai Chinese	CDX				100
Chinese in Denver, Colorado	Denver Chinese	CHD			107	
Japanese in Tokyo, Japan	Japanese	JPT		30	105	100
Kinh in Ho Chi Minh City, Vietnam	Kinh Vietnamese	KHV				100
Utah residents (CEPH) with Northern and Western European ancestry	CEPH	CEU	3	60	90	100
Toscani in Italia	Tuscan	TSI			66	100
British in England and Scotland	British	GBR				100
Finnish in Finland	Finnish	FIN				100
Iberian populations in Spain	Spanish	IBS				100
Yoruba in Ibadan, Nigeria	Yoruba	YRI	3	59	112	100
Luhya in Webuye, Kenya	Luhya	LWK			108	100
Gambian in Western Division, The Gambia (possibly two populations)	Gambian	GWD				2 x 100 [1]
Malawian in Blantyre, Malawi	Malawian	MAB				100 [1]
African Ancestry in Southwest US	African-American SW	ASW				61
African American in Jackson, Mississippi	African-American MS	AJM				80
African Caribbean in Barbados	Barbadian	ACB				79
Mexican Ancestry in Los Angeles, California	Mexican-American	MXL				70
Colombian in Medellin, Colombia	Colombian	CLM				70
Peruvian in Lima, Peru	Peruvian	PEL				70
Puerto Rican in Puerto Rico	Puerto Rican	PUR				70
Ahom in Dibrugarh, India	Ahom	AHD				100 [1]
Kayastha in Kolkata, India	Kayastha	KAK				100 [1]
Reddy in Hyderabad, India	Reddy	RDH				100 [1]
Maratha in Mumbai, India	Maratha	MRM				100 [1]
Punjabi in Lahore, Pakistan	Punjabi	PJL				100
		Totals	6	179	697	2500

Note 1: the use of these populations in the full project has not been finalised

Full 1000 genomes project

- Low-coverage whole-genome sequencing
- Array-based genotyping
- Deep targeted sequencing of all coding regions