Supplementary Tables

Table 1: Targeted sequencing coverage by population

Population	Total Genomes Studied	Average coverage	Average Coverage (Microsat Loci)
Central European in Utah, USA (CEU)	57	64.5x	15.7x
Han Chinese in Beijing, China (CHB)	95	32.6x	13.3x
Tuscan in Florence, Italy (TSI)	4	73.4x	35.5x
Chinese in Denver, Colorado, USA (CHD)	107	29.7x	15.3x
Luhya in Webuye, Kenya (LWK)	108	21.4x	12.4x
Yoruba in Ibadan, Nigeria (YRI)	84	57.0x	14.8x
Japanese in Tokyo, Japan (JPT)	96	59.8x	20.8x
All	551	45.5x	15.5x

The individuals studied where those which were sequenced on the Illumina or 454 platforms at a minimum read length (reads of at least 45 base pairs in length). Coverage in microsatellite regions only considers those reads which completely span the microsatellite loci in targeted regions and were specifically aligned to a particular microsatellite locus through enough unique flanking sequence on both sides of the repeat.

Table 2: Variations at microsatellite loci by population

(A)

Region	CEU Genomes: 57		CHB Genomes: 95		TSI Genomes: 4		CHD Genomes: 107	
	Reliable Genotype	% Diff	Reliable Genotype	% Diff	Reliable Genotype	% Diff	Reliable Genotype	% Diff
Upstream	124	4.8%	136	7.4%	124	4.8%	136	7.4%
5'UTR	230	5.7%	235	8.1%	230	5.7%	235	8.1%
Exon	1,273	4.8%	1,274	6.8%	1,273	4.8%	1,274	6.8%
Intron	5,317	6.3%	5,573	9.0%	5,317	6.3%	5,573	9.0%
3'UTR	350	4.6%	359	8.9%	350	4.6%	359	8.9%
Downstream	112	5.4%	112	8.9%	112	5.4%	112	8.9%
Total	7,406	5.9%	7,689	8.6%	7,406	5.9%	7,689	8.6%

(B)

Region	LWK Genomes: 108		YRI Genomes: 84		JPT Genomes: 96	
	Reliable Genotype	% Diff	Reliable Genotype	% Diff	Reliable Genotype	% Diff
Upstream	130	15.4%	129	6.2%	134	4.5%
5'UTR	230	16.5%	231	5.2%	246	4.5%
Exon	1,256	10.6%	1,278	6.7%	1,287	4.4%
Intron	5,602	15.6%	5,498	7.4%	5,510	5.7%
3'UTR	363	13.8%	360	8.6%	345	4.6%
Downstream	111	14.4%	112	6.3%	117	6.0%
Total	7,692	14.7%	7,608	7.2%	7,639	5.4%

The total number of reliable genotypes which could be called in at least one sample for a specific microsatellite loci are shown along with the percent of those microsatellite loci which differed from the human reference sequence in at least one individual in the population. All reads were processed as though they were from the 454 platform and so homopolymer regions were removed for all to provide consistency.

Table 3: Average variation at microsatellite loci per individual by population

Region	CEU	CHB	TSI	CHD	LWK	YRI	JPT	ALL
Total Genomes Analyzed	54	94	1	107	108	67	96	530
Upstream	1.8% (±0.3%)	0.6% (±0.1%)	-	0.5% (±0.1%)	1.8% (±0.0%)	2.0% (±0.3%)	0.6% (±0.1%)	1.1% (±0.0%)
5'UTR	1.4% (±0.0%)	1.6% (±0.2%)	-	1.2% (±0.1%)	1.2% (±0.0%)	0.6% (±0.1%)	1.1% (±0.1%)	1.2% (±0.1%)
Exon	0.5% (±0.0%)	0.5% (±0.0%)	-	0.5% (±0.0%)	0.7% (±0.0%)	0.6% (±0.1%)	0.4% (±0.0%)	0.5% (±0.0%)
Intron	1.2% (±0.0%)	1.1% (±0.0%)	-	1.2% (±0.0%)	1.6% (±0.0%)	1.2% (±0.0%)	1.0% (±0.1%)	1.2% (±0.0%)
3'UTR	0.8% (±0.1%)	0.9% (±0.0%)	-	0.9% (±0.0%)	1.0% (±0.0%)	1.0% (±0.1%)	0.8% (±0.0%)	0.9% (±0.0%)
Downstream	0.6% (±0.1%)	0.8% (±0.1%)	-	1.7% (±0.0%)	0.8% (±0.1%)	0.4% (±0.0%)	1.0% (±0.1%)	0.9% (±0.1%)
Total	1.1% (±0.1%)	0.9% (±0.0%)	-	1.0% (±0.0%)	1.2% (±0.1%)	1.0% (±0.0%)	0.8% (±0.0%)	1.0% (±0.0%)

The total variations at microsatellite loci on average for a single individual in the population are presented. Only those samples which had at least 10% of the total microsatellite allelotypes found (300) were considered. The TSI population is not shown as there were less than two individuals remaining with at least 300 microsatellite allelotypes after processing out all homopolymer regions which are error prone for the 454 platform to obtain consistency between the Illumina and 454 results. The number in parenthesis is the standard deviation.

Table 4: Primers for Sanger sequencing validation

Samples	Gene	Forward Primer	Reverse Primer	Variation
NA06994	TARS2	AGGGAGGGATGGTGATAAGG	AGCTTCTTCTAGTCCCTCAGC	T/A
NA06994	WNK3	CAATATCCTCTGGCGCATC	TCCATTTGCAGTGCCTCTTC	T/C
NA07048	HEATR6	GAGATCGCCTCAAATCAAGG	AGCTTGGAGAGGTGATTTGC	T/C
NA11918	BAIAP3	CATGCTAAGTAGGGCATCTGG	GTGACCTGGATGCACTTGG	C/T
NA12249	NPC1L1	TGTGGCACATGGAGTAGAGG	GATCACTCGAGGTGTTGTGG	G/T
NA12249	PLXDC1	CTCATGTTGGCTGACTGTGG	AACACAGCGAGACCTCATCC	G/T
NA12272	GIGYF2	AGCATACCAGGCCTTTGTGG	CCTCTGGCGCATTAACTCC	G/T
NA12414	GPR114	AGAACTGGGTGTGCTTCTCC	GCGAGGAGAGATGATGTGC	C/A
NA12717	COL6A3	CCAGGTGTACAAGCCCAGAT	AAGCCAGCAGCTGTAAGACC	-/GCT
NA12750	TBX18	CATGGATAAGCTGGTCTGTGG	CTCCAAGGTACTGGGAATGG	T/G
NA12873	RCL1	GGTGCTGTTGGTCTTGAAGG	AACCTCGGGAAGAAGTAAAGC	G/A
NA18539	UTRN	TGCTACCCTCAGTGGAAAGG	TGAGGGTGTGTTTCACTTGC	A/G
NA18596	FBXO24	CCAAGCTCCAACTCATCTCC	AATACAGGTGGGCTCTCAGC	G/A
NA18641	TEX14	AGGCACAGAGCTGAATAGCC	GCAGGTTCTGTGTGGAGACC	G/A
NA18985	FTSJ3	GGGAACTCCAAGCTCTACCC	AGAACAAGCAAAGGCACTGG	C/A
NA19011	OR10H3	TGTTCACTGTTGCCATCACC	ACACACACCAGCATCACACC	C/A
NA19011	TIAM2	GAAGAATTCCTCCAGCAACG	CGGAAATCCTCAGTCTCTGG	G/A
NA19220	CDHR2	CCAATGGCTCCATAGTCCTC	GAGCACTGGATGGTGAAGGT	-/CAA
NA19462	ETS2	GCTTCGAGAGCTCAGACTCC	CGGTGAATGTGGTACTGTGG	C/A
NA19556	ZSCAN2	CAGACATCCCGAGAGTGACC	TGACCCTGCTCATGGTACG	G/C
NA20763	CLSPN	TGCAGTGCTTTGGCTGTAAC	ACAATCCAGGCCATGCTAAC	-/GAA
NA20763	KANK1	TCCGAGTCAGATGACGAGTG	AGGGAAGGGGAATGTACCAC	-/GGA

The non-synonymous variations which could not be validated in HapMap were validated by Sanger capillary sequencing using the primers shown.