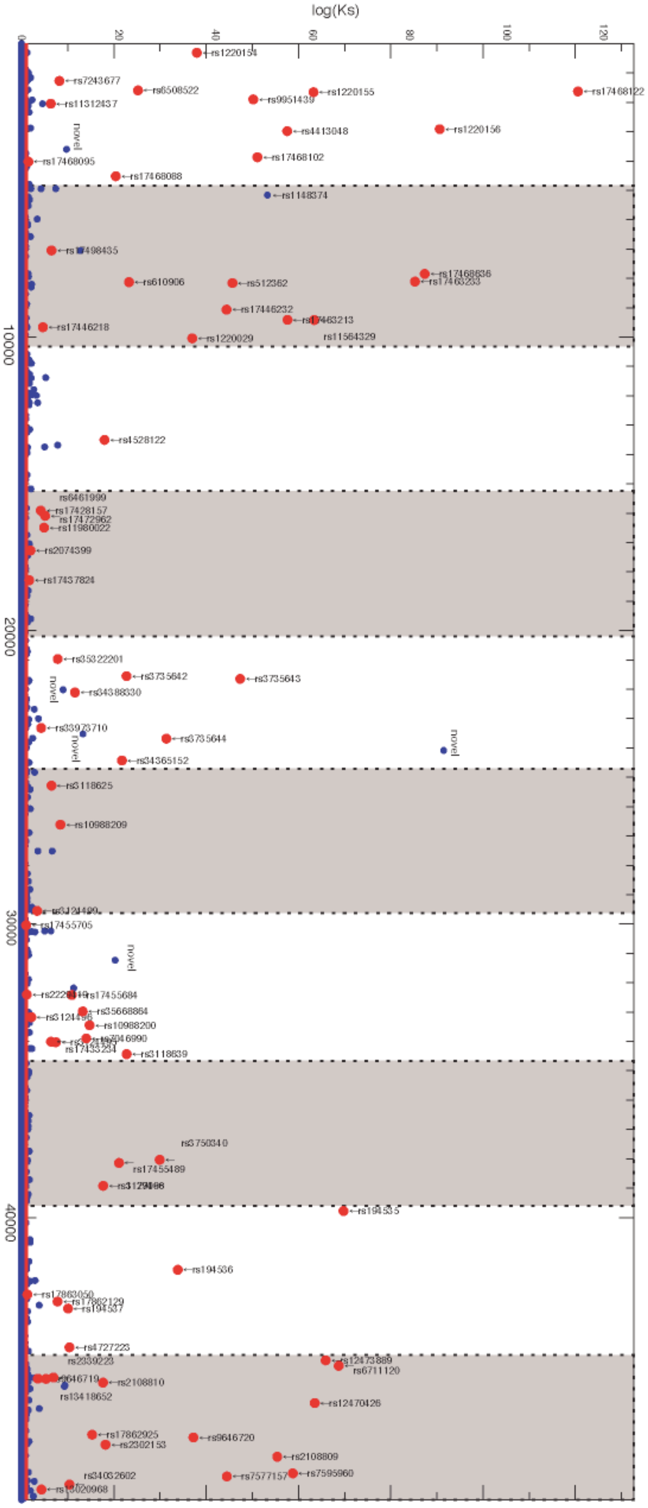
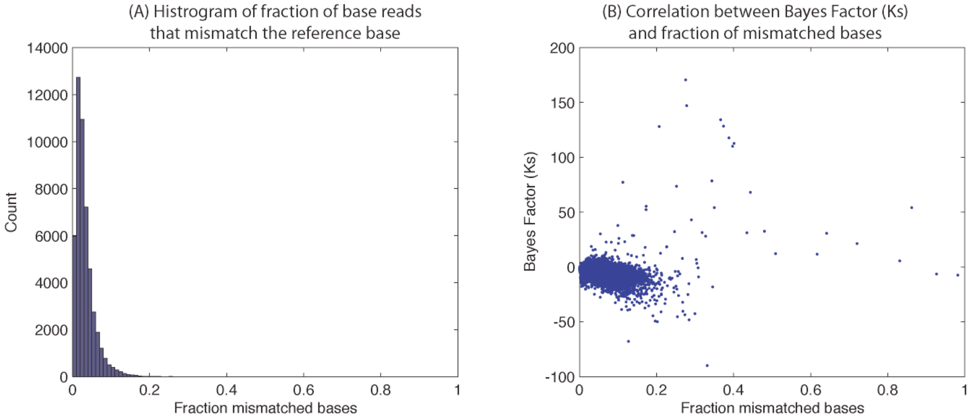
Supplementary Figure 1.



Supplemental Figure 1. Library B polymorphism discovery. Similar to Figure 4, variant discovery plot for Library B. As with Library A, inspection of traces from NCBI Trace Archive confirmed polymorphic bases that were previously not annotated.

Supplementary Figure 2



Supplementary Figure 2. Bayes factors vs. non-reference bases alignments. (A) A histogram of the fraction of mismatches at each base indicating that a substantial percentage of bases have 15% of the aligned bases disagreeing with the reference base (B.) Scatterplot of K_s vs. fraction of mismatched. The lack of a positive correlation indicates that most bases favor the model of a binomial distribution even as the fraction of mismatched bases increases. Taken together, these data suggest that most bases sample error, as expected, from a binomial distribution across individuals. Those bases that do not are possible polymorphisms.

Supplementary Table 1

Region #	Amplicon Primer Sequence	Region (Build 36.1)
Region 1	Gagatgggtctcctgagtgc	chr2:234189304+234194258
	Aaccaactgcatgcttttc	
Region 2	Ccagccaaacttgacgtacc	chr7:113839924+113844896
	Aacagcccaattcaacttgc	
Region 3	gaccaaggaaacaaccaacc	chr20:33747673-33752665
	caggaaagcccacatacacc	
Region 4	ctaactcagcggctttgtcc	chr2:220085623+220090673
	tcacgtccttttggagacc	
Region 5	gggagctcacgatatcaagg	chr11:130743490+130748475
	caggaagcagcagctctagg	
Region 6	gctgggtgttggatatttgc	chr1:149678673-149683603
	cagtaggcaaggacacatgc	
Region 7	taacagtggggctgaaaagc	chr5:142043190-142048159
	cttgggagtctccaggtagc	
Region 8	gttgaaaccagggacaatgg	chr5:141971045-141976067
	ctcctctcacctgcagaacc	
Region 9	gccatgggagttaacagagg	chr7:126531683-126536662
	ttgctaccatttgccattcc	
Region 10	tggccagttttgttttcagg	chr11:116250173-116255189
	tatttgggacgaaggattgc	
Region 11	agtagggtgagcttggatgc	chr13:112347206+112352113
	ccgcatgacttttgtttgg	
Region 12	gaccttgtgatctgccttcc	chr16:26053412+26058370
	gaggggctcctaaagtttgc	
Region 13	atggagttttgcttttgttgc	chr11:63970822+63975969
	agttgtccctgtggctatgg	
Region 14	tcagtggtgctgtactcatgc	chr16:41060-46001
	aggacacctggggattacg	

Supplementary Table 1. Primers. Primers for amplification of targeted regions within Library A.

Supplementary Table 2.

Encode Regions	Multiplex Pool	Chr	Physical Position Region (36.1)	Region Features
ENr231	A,B	1	149678673-149683603	Moderate non-exonic conservation, High Gene density
ENr331	A	2	220085623-220090673	High-exonic conservation, High Gene density
ENr131	A,B	2	234189304-234194258	Low-exonic conservation, High Gene density
ENr212	A	5	141971045-141976067	Moderate non-exonic conservation, Moderate Gene density
ENr212	A	5	142043190-142048159	Moderate non-exonic conservation, Moderate Gene density
ENm014	A	7	126531683-126536662	7q31.33
ENm017	B	7	127008375-127012881	7q32.1
ENm017	B	7	127013024-127018100	7q32.1
ENm017	B	7	127016306-127021248	7q32.1
ENm010	B	7	27246985-27251931	HOXA Cluster
ENm012	B	7	89709475-89714560	FOXP2
ENr335	B	9	130810649-130815588	High-exonic conservation, High Gene density
ENr335	B	9	130908740-130913660	High-exonic conservation, High Gene density
ENm003	A	11	116250173-116255189	Apo_cluster
ENr312	A	11	130743490-130748475	High-exonic conservation, Low Gene density
ENr123	B	12	38507675-38512629	Low-exonic conservation, High Gene density
ENr123	B	12	38783227-38788147	Low-exonic conservation, High Gene density
ENr123	B	12	38903255-38908283	Low-exonic conservation, High Gene density
ENr132	A	13	112347206-112352113	Low-exonic conservation, High Gene density
ENr213	B	18	23783347-23788366	Moderate-exonic conservation, Low gene density
ENr213	B	18	23808628-23813466	Moderate-exonic conservation, Low gene density
ENr333	A	20	33747673-33752665	High-exonic conservation, High Gene density

Supplementary Table 2. Selected Regions. Two different indexed pools were sequenced across a series of runs (referred to as multiplex Library A and B). The ENCODE region name, and key ENCODE attributes are listed.

Supplementary Table 3

Phos	5'-3' index	Sequence	5'-3' index
P-	TTT TTA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	AAA AAT
P-	GGT TGA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	CAA CCT
P-	CCT TCA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	GAA GGT
P-	AAT TAA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	TAA TTT
P-	GTG TGA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	CAC ACT
P-	TGG TTA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	AAC CAT
P-	ACG TAA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	TAC GTT
P-	CAG TCA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	GAC TGT
P-	CTC TCA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	GAG AGT
P-	AGC TAA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	TAG CTT
P-	TCC TTA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	AAG GAT
P-	GAC TGA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	CAG TCT
P-	ATA TAA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	TAT ATT
P-	CGA TCA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	GAT CGT
P-	GCA TGA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	CAT GCT
P-	TAA TTA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	AAT TAT
P-	GTT GGA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	CCA ACT
P-	TGT GTA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	ACA CAT
P-	ACT GAA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	TCA GTT
P-	CAT GCA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	GCA TGT
P-	TTG GTA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	ACC AAT
P-	GGG GGA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	CCC CCT
P-	CCG GCA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	GCC GGT
P-	AAG GAA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	TCC TTT
P-	ATC GAA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	TCG ATT
P-	CGC GCA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	GCG CGT
P-	GCC GGA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	CCG GCT
P-	TAC GTA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	ACG TAT
P-	CTA GCA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	GCT AGT
P-	AGA GAA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	TCT CTT
P-	TCA GTA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	ACT GAT
P-	GAA GGA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	CCT TCT
P-	CTT CCA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	GGA AGT
P-	AGT CAA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	TGA CTT
P-	TCT CTA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	AGA GAT
P-	GAT CGA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	CGA TCT
P-	ATG CAA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	TGC ATT
P-	CGG CCA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	GGC CGT
P-	GCG CGA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	CGC GCT
P-	TAG CTA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	AGC TAT
P-	TTC CTA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	AGG AAT
P-	GGC CGA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	CGG CCT
P-	CCC CCA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	GGG GGT
P-	AAC CAA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	TGG TTT
P-	GTA CGA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	CGT ACT
P-	TGA CTA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	AGT CAT
P-	ACA CAA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	TGT GTT
P-	CAA CCA	GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG	
		ACACTCTTTCCCTACACGACGCTCTTCCGATCT	GGT TGT

Supplementary Table 3. Full oligionucleotide sequences for indexed adapters.

Supplementary Table 4.

DNA Indexes Appended to Adapter Sequence			
AAAAAT	CAACCT	GAAGGT	TAATTT
AACCAT	CACACT	GACTGT	TACGTT
AAGGAT	CAGTCT	GAGAGT	TAGCTT
AATTAT	CATGCT	GATCGT	TATATT
ACACAT	CCAACT	GCATGT	TCAGTT
ACCAAT	CCCCCT	GCCGGT	TCCTTT
ACGTAT	CCGGCT	GCGCGT	TCGATT
ACTGAT	CCTTCT	GCTAGT	TCTCTT
AGAGAT	CGATCT	GGAAGT	TGACTT
AGCTAT	CGCGCT	GGCCGT	TGCATT
AGGAAT	CGGCCT	GGGGGT	TGGTTT
AGTCAT	CGTACT	GGTTGT	TGTGTT

Supplementary Table 4. DNA Indexes Appended to Each Adapter. A total of 48 different 6-mer index sequences were chosen from the 4096 different possible 6-mers. Each index is designed so that the 1st and 5th base are identical and represent an XOR-based checksum of bases 2-4 so that the index sequence remains identifiable even with an uncalled or low-quality base.

Supplementary Table 5.

Library A dbSNP identifier	log(Ks)	Coverage
rs12533005	51.1	355
rs11167787	74.0	713
rs11960262	47.7	455
rs11167786	55.7	467
rs10210058	18.7	452
rs34012	8.1	417
rs1042597	4.6	297
rs13027376	2.9	297
rs13400017	7.9	397
rs11167785	48.9	528
rs1361963	63.8	859
rs6975798	34.1	443
rs152524	5.4	372
rs1557644	24.0	774
rs10791140	55.6	1645
rs907676	5.1	179
rs11222591	22.7	761
rs250108	5.1	390
rs1872858	14.1	168
rs2237790	31.9	686
rs4114768	16.4	642
rs11563720	10.0	541
rs249926	33.5	1514
rs17092980	0.5	601
rs4291502	4.0	533
rs4528122	23.5	58
rs12705964	4.5	950
rs7727653	4.4	398
rs17099096	3.2	362
rs17864670	5.0	405
rs9919599	2.4	40
rs11955552	4.3	825
rs681524	6.9	611
rs17864673	13.1	1009
rs17868309	4.6	815
rs10221563	9.3	374
rs11975039	1.2	832
rs17099102	1.7	709
rs17099100	3.2	598
rs7119590	2.1	1707
rs10241421	-1.7	0
rs17099249	-1.5	17
rs12292614	-1.3	0
rs1042590	-1.7	191
rs17867764	-1.3	0
rs3088078	-0.7	0
rs11222590	-2.5	5
rs17863762	0.1	0
rs1126806	-1.5	0

Library B dbSNP	Log(Ks)	Coverage
rs13289043	0	0
rs17445742	0	0
rs17863844	0	0
rs17437810	0.5724	1
rs3814492	0.3001	1
rs17468650	0.138	2
rs17468601	-0.0202	3
rs17446218	4.6375	4
rs2229419	1.1383	4
rs9945664	-0.3719	4
rs17437824	1.6533	6
rs4291502	-0.1137	6
rs17445496	-0.1526	6
rs17468095	1.4706	7
rs10988209	8.4219	15
rs4727223	10.3836	16
rs3118625	6.4789	16
rs17428157	4.1719	16
rs11980022	4.8702	18
rs7046990	14.0197	20
rs17862129	7.7949	20
rs4528122	17.9849	26
rs3124495	7.4348	27
rs10988200	14.7299	29
rs3118639	22.777	30
rs13020968	4.3019	40
rs4363925	5.6175	50
rs17468088	20.4266	110
rs194537	10.06	116
rs13418652	5.2823	137
rs3750340	29.9307	144
rs7243677	8.1997	152
rs17446232	44.4083	170
rs7577157	44.4728	172
rs4800835	24.5798	174
rs2302153	18.1967	212
rs17468102	51.1039	219
rs7595960	58.7743	227
rs194536	33.8634	239
rs9951439	50.2121	243
rs17463213	57.6308	260
rs1220155	63.2521	271
rs512362	45.6942	272
rs194535	69.7629	275
rs1220029	36.9822	278
rs12470426	63.5206	288
rs11564329	63.3966	291
rs1530380	62.8385	291
rs12473889	65.888	298
rs9646720	37.2413	343
rs2108809	55.3984	358
rs4768189	44.4388	358
rs17463233	85.2413	367
rs1220154	37.9816	376
rs7962260	64.8659	385
rs17468636	87.3576	392
rs4413048	57.5525	453
rs17468122	120.5659	469
rs610906	23.2828	470
rs1148374	53.2736	571
rs6508522	25.2233	583
rs6711120	68.73	608
rs1220156	90.5985	678

Supplementary Table 5. Polymorphism coverage and K_s. List of SNPs containing known variants, their coverage, and their K_s in sequenced regions for Library A (left) and Library B (right).

Supplementary Methods

Amplification of targeted regions. 46 HapMap individuals were whole genome amplified by RepliG (Qiagen) and the concentration determined by replicate measures using the Quant-iT PicoGreen dsDNA kit (Invitrogen). For each HapMap individual, the ENCODE regions were amplified by long-range PCR in a 25ul reaction volume using 75ng template DNA, 1X PfuUltra Buffer, 2mM dNTPs (total), 400nM each of the forward and reverse primers, and 0.5ul PfuUltra II HS DNA polymerase (Stratagene) per reaction. The thermal cycler conditions were: a denaturation step at 95°C for 2 minutes, 30 cycles consisting of 95°C for 20 seconds, an annealing temperature specific to each primer for 20 seconds, and 68°C for 3 minutes, and a final extension of 68°C for 5 minutes. To generate a sufficient amount of amplicons, the initial PCR products were put through a second PCR reaction. This second-round reaction used 2uL of the initial PCR product, 2mM dNTPs (total), 400nM each of the forward and reverse primers, 1.5uL PfuUltra II HS DNA polymerase (Stratagene), and 1x PfuUltra buffer in a 100ul reaction. The thermal cycler conditions were the same for both the first and second round PCR. Products were purified on QiaQuick 96 well columns (Qiagen) and quantified by taking optical density and Picogreen measurements. A 4-fold range of concentrations was observed after the second round of PCR so the products were titrated such that equimolar amplicons from each individual were pooled and the combined total DNA was ~5μg. Pooled amplicons were digested to 200-300 bp fragments using DNAse I enzyme (NEB). Fragmented pools were then blunt end repaired using T4 DNA Polymerase, DNA polymerase I Klenow fragment, and T4 polynucleotide kinase enzyme (NEB). Subsequently, dATP incorporation was performed to the blunt ended amplicons with Klenow Fragment 3'-5' exo minus enzyme (NEB). DNA was purified after each step using Illumina-recommended Qiagen 96 well purification columns.

Indexing adapter preparation. A total of 48 different adapters were produced to index sequenced fragments (See supplementary tables 2-6). The adapter sequence began with the oligonucleotide sequences provided by, and shared with the permission of, Illumina, (© 2006 Illumina, Inc.; 5' P-GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG and 5' ACACTCTTTCCCTACACGACGCTCTTCCGATCT) and were followed by the index in the forward and reverse directions respectively. Lyophilized, indexed adapters were dissolved in 10mM Tris-HCl pH 7.8 (Sigma) to a 100uM stock concentration. A 10X annealing buffer was made containing 100mM Tris-HCl ph 7.8 and 0.5M NaCl (Gibco). Indexed adapter pairs were combined to a 50uM final concentration with 10X annealing buffer for a final concentration of 1X. Subsequently, a step down annealing reaction was performed, where the indexed adapter mix was incubated at 95°C for 2 minutes followed by a series of cooling steps of 1°C per minute from 95°C to 25°C.

Adapter ligation, PCR enrichment and sequencing of indexed amplicons. A unique indexed-adapter sequence was ligated to each HapMap individuals' adenylated amplicon pool. Ligation was performed at 20°C for 2 hours followed by 16°C for 16 hours using T4 DNA ligase enzyme (NEB). Ligated amplicons were then pooled for all individuals to be sequenced on the same flow cell lane. The pooled ligated amplicons (referred to as a library) were loaded onto a 4% agarose gel and 150-200 bp fragments were gel-purified using Qiagen gel purification columns. Libraries were PCR-enriched using Phusion DNA polymerase PCR mix and adapter compatible primers 1.1 and 2.1 (Illumina). Each PCR product was loaded onto a 4% agarose gel and fragments of 150-200bp were gel-purified using Qiagen gel purification columns. Each library was quantified using Nanodrop ND-1000 and diluted to 10nM working concentration using EB buffer. Each library was loaded on a lane of the flowcell for cluster generation and libraries were sequenced on an Illumina GA. Initial cluster counts typically ranged from 1M to 2M clusters per flow cell lane. In later runs, cluster counts of approximately 8-10M clusters per flow cell lane were typically observed during indexed sequencing. Also in later runs (not shown), we found that the same indexing strategy of ligating barcoded adapters was effective in paired-end sequencing on the Genome Analyzer II upgraded system. In these runs, 5-base indexes were utilized. With 5-base paired-end indexing runs, indexes from both the forward and reverse reads are sequenced for a total of 10 indexed bases.

Calculation of false-positives. False negative rates were determined by calculating if a base known to be polymorphic in our library of HapMap individuals reached previously specified K_s thresholds. Within Library A, 152 SNPs were listed in dbSNP. Of these SNPs, 50 SNPs have at least one individual varying in genotype from the other sequenced individuals. We observed 41 of the 50 SNPs had a base-level K_S>10, with 40 of 41 exceeding K_B>100. Of the 9 false-negative SNP positions, 1 SNP was triallelic and not entirely compatible with our biallelic analysis model, 1 SNP was found only in 2 individuals and the other 7 SNPs were found in only 1 individual. Similar results were obtained for Library B (table 1). The false negatives appear to result from lack of coverage (figure 5a) and not simply because only 1 individual had the variant: of the 7 single-event false negatives, 3 individuals had less than 2 reads at the variant position. False positives are more difficult to quantify since not all polymorphic sites are known, even in previously resequenced regions. In our analysis, to be defined as a false positive, the base must reach a pre-specified K_s, must not exactly match the location of variants within dbSNP, and must not have trace sequencing data indicating a missed variant. Library B, which is entirely composed of regions with existing ENCODE sequencing data, was used to specifically estimate false-positives (table 1). Immediately evident by visual inspection (Figure 4) is an overall low false positive rate for bases with large K_s values; most high-ranking bases (K_s>100) are also listed in dbSNP (see Figure 4a). Particularly within the two ENCODE sequenced regions, 25 of the top 25 ranked bases were at the exact position of the SNP. In previously non-sequenced regions, we could identify many candidate variants. For instance, in region 2, there were 20 bases with $K_s > 3$. For the 8 highest ranking bases, 4 were unambiguously confirmed as novel SNPs. Of the remaining 4 bases, 1 did not have sufficient trace data, 1 was neighboring an identified SNP and neighboring a repeat region, 1 was in a repeat region, and 2 were in regions highly homologous to other regions in the genome. For the remaining 12 high-ranking bases, we were not able to confirm the existence of a SNP. For these bases, trace data frequently exhibited multiple reads, was poor in quality in both read directions, or was within a location with high sequence homology (see examples of difficult-to-assess traces in Figure 4d and 4e). In some cases, high quality sequence data was not available for the exact individuals driving the ranking. However, it is likely that these 12 unconfirmed candidate variants are false-positives and we treat them as such for our analysis. Indeed, the most prominent source of false positives with our approach appears to arise from homology to other regions or polymorphic sites elsewhere in the genome. In practice, one would expect lower false positive rates for less complex regions.