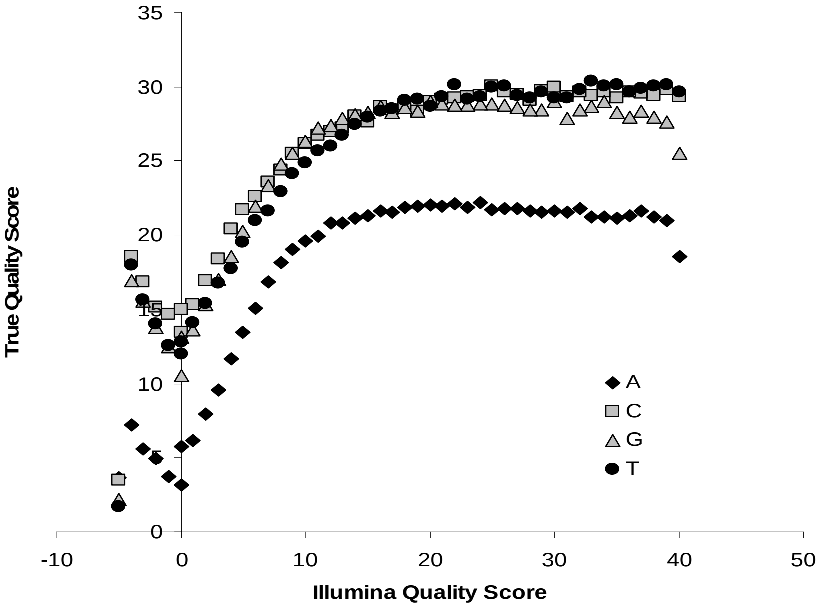
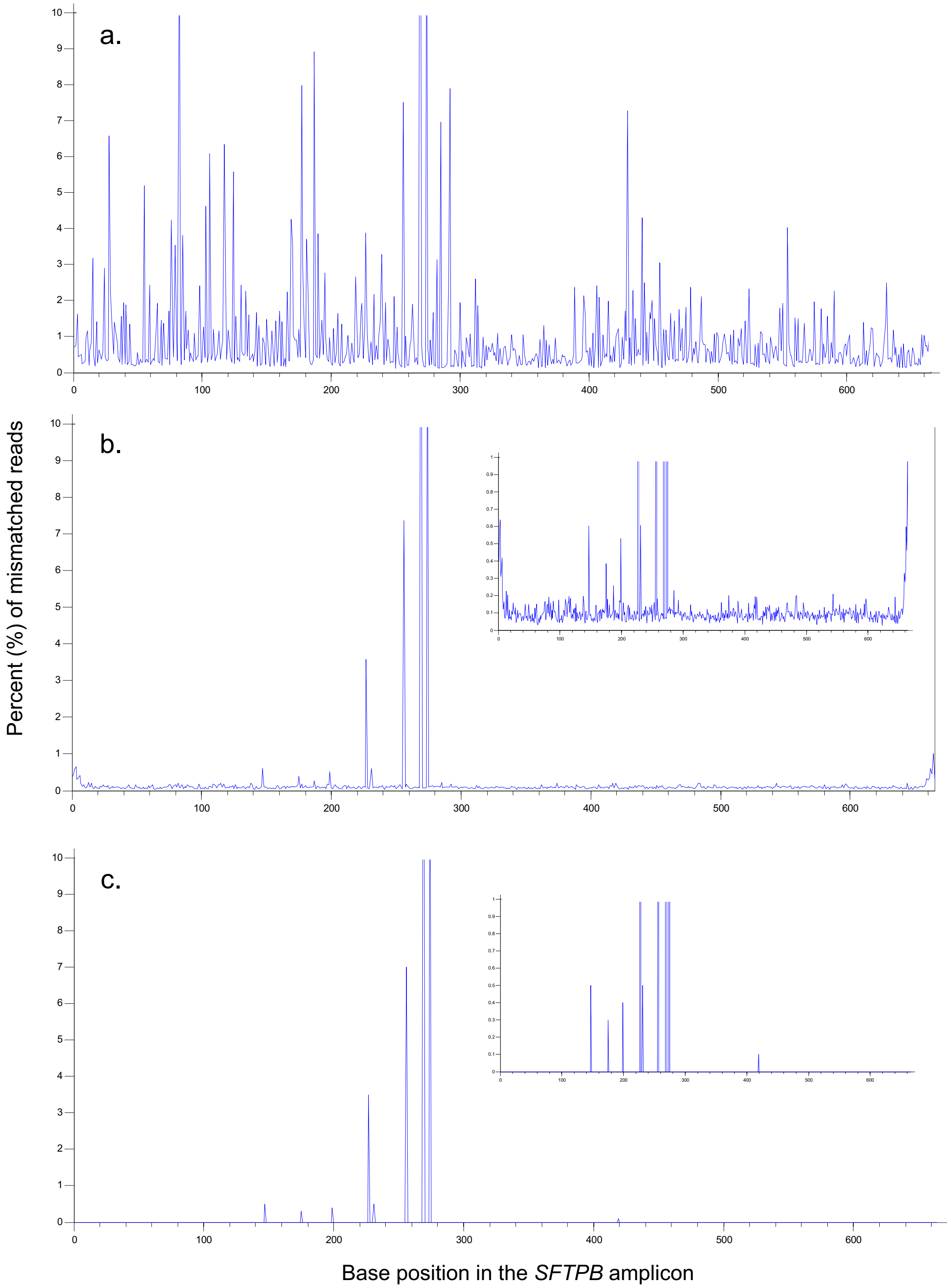
Supplementary Figure 1. Illumina Genome Analyzer I quality score analysis.



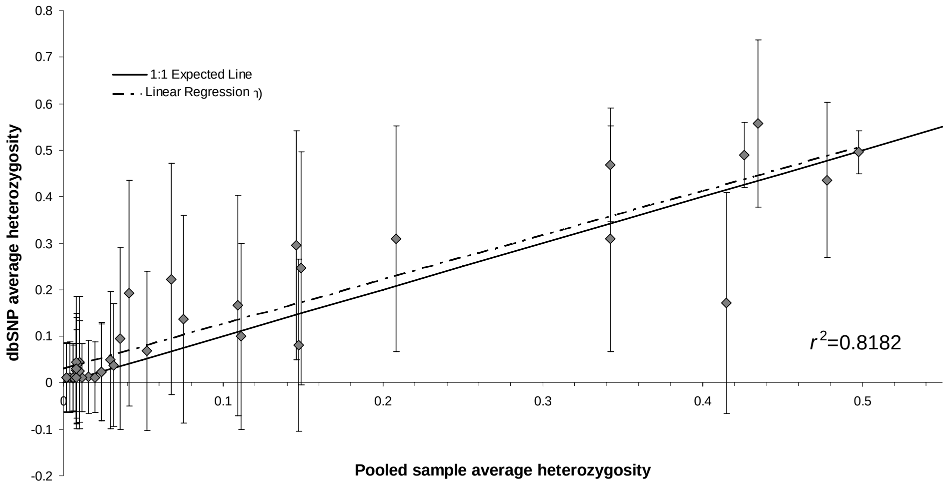
Supplementary Figure 1. Illumina Genome Analyzer I quality score analysis. The quality scores generated by the Illumina Genome Analyzer I are plotted along the X-axis while the true quality scores, defined as *–log*$_{10}$ of the probability of incorporating a nucleotide different from the reference sequence bases (as calculated from the negative control vector as described).

Supplementary Figure 2. How SNPSeeker improves SNP calling



Supplementary Figure 2. How SNPSeeker improves SNP calling. In each panel, the X-axis depicts the sequential 665 bp of the *SFTPB* amplicon and the Y-axis is the percent likelihood of a mismatched base when compared to the reference sequence. (**a**) When using 32 bases per read to perform SNP calling, there are 218 positions in the *SFTPB* amplicon that are considered likely to contain a SNP. (**b**) When using only the first 12 bases of each read and a frequency cutoff calculated by the error model generated from pUC19 data, there are 19 base positions considered to have potential SNPs. INSET: Ten-fold higher resolution plot showing a maximum 1% likelihood of mismatch. (**c**) When implementing SNPSeeker on the data from cycles 1-12, only 9 SNP positions remain. INSET: Ten-fold higher resolution plot showing a maximum 1% likelihood of mismatch.

Supplementary Figure 3. Pooled sample vs. dbSNP average heterozygosity



Supplementary Figure 3. Pooled-sample versus dbSNP average heterozygosity. Of the 44 SNPs that were described in dbSNP, 37 had average heterozygosity values listed. These values (with standard error bars) are plotted on the Y-axis against the corresponding average heterozygosity value as determined from the SNP frequency in the pooled-sample (correlation coefficient r^2 = 0.82). The solid line is a plot of the idealized 1:1 correlation between data sets and the dashed line is the actual linear regression.

Supplementary Table 1. PCR primers and conditions.

		PRIMER COMBO				
	TEMPLATE	LEFT (5'→3')	RIGHT (3'←5')	AMPLICON SIZE (bp)	ANNEALING TEMP	EXTENSION TIME
1	pUC19	gacctgcaggcatgcaag	gtatcaacggactgaggggcagcac	1276	62 C	4 min
2	*ACTB* , exons 2/3	gcctccgaccagtgtttgcctttta	acgcgggaaagagtgaccaagagag	755	60 C	90 sec
3	*ACTB* , exons 4/5	ccccagcacacttagccgtgttctt	acagaaaggacggactcgactggac	853	60 C	90 sec
4	*ACTB* , exons 6	gctgtcacatccagggtcctcactg	cacactgaaacaccacaccgaccca	821	60 C	90 sec
5	*SFTPB* , exons 7/8	agtggaggcttgccaagtgaaggtc	ctgtggatcctacccggtcttacct	665	58 C	90 sec
6	*TP53* , exon 1	cctccccaactccatttcctttgct	ccactcaccctaccttcgaaccgat	423	58 C	90 sec
7	*TP53* , exon 2-4	tgctggatccccacttttcctcttg	gtagggtagtgtgggagtcgtagag	884	60 C	90 sec
8	*TP53* , exon 5/6	ttgctgccgtcttccagttgcttta	agagaccctcctccccaattcccac	530	58 C	90 sec
9	*TP53* , exon 7-9	ggcctcatcttgggcctgtgttatc	tacggagtttctgttaccgaggacc	928	58 C	4 min
10	*TP53* , exon 10	acttctccccctcctctgttgctgc	aaggtaagagtaggacggaagtacc	223	56 C	60 sec
11	*TP53* , exon 11	ttgaattcccgttgtcccagcctta	ggactcagacgttactcacacccga	1495	60 C	5 min
12	*APC* 1, exon 15	cccacccccctgcaaatgttttaagc	gttatgggtcggctggatcgggtat	1332	62 C	5 min
13	*APC* 2, exon 15	caatacccagccgacctagcccata	tacgacgtcaagtctcccaggtcca	1447	62 C	5 min
14	*APC* 4, exon 15	taccagacagaggggcagcaactga	ggttcttcactcagacggaggtttcc	1381	62 C	5 min
15	*APC* 5, exon 15	ccaagaagtgagtctgcctccaaagg	acaatctcccaaaaacaagaccttcg	1500	62 C	5 min
				14513		

Supplementary Table 1. PCR primers and conditions. For each amplified region of genomic DNA, the specific primer combination, size of the resulting amplicon, PCR annealing temperature, and extension time are listed.

Supplementary Table 2. Taqman assay primers and probes.

a

SNP	Forward primer	Reverse Primer
TP53; rs17881850	proprietary by ABI	proprietary by ABI
TP53; rs17880847	proprietary by ABI	proprietary by ABI
TP53; chr17:7512477	TCCCCTCCTTCTCCCTTTTTATATCC	GTTCACCCCTCAGACACACA
ACTB; chr7:5534814	CAGCCATGTACGTTGCTATCCA	CCATCACGATGCCAGTGGTA
APC; rs2229994	proprietary by ABI	proprietary by ABI
APC; chr5:112202922	CCATCCAAGTTCTGCACAGAGTAG	TCTACACAATAAGTCTGTATTGTTTCTTGGTT
APC; chr5:112207052	CCGCAAAAGGAACATGGAGAAAA	TGAGGAAACGGTCTGAGAAGTACTA
SFTPB; chr7:85744364	CCTGGCTGAGCGCTACTC	GGGCAGTGGGCTCACTT
SFTPB; chr7:85744391	GCGGGCGGCATCTG	CCCAGCAGCGTGTCGA

b

SNP	Probe 1	Probe 2
TP53; rs17881850	proprietary by ABI	proprietary by ABI
TP53; rs17880847	proprietary by ABI	proprietary by ABI
TP53; chr17:7512477	CTTATTTTACAATAAAACTTTG	TTATTTTACAATACAACTTTG
ACTB; chr7:5534814	CCCTGTACGCCTCTG	TCCCTTTACGCCTCTG
APC; rs2229994	proprietary by ABI	proprietary by ABI
APC; chr5:112202922	CCTCAAAAGGCTGCC	CTCAGAAGGCTGCC
APC; chr5:112207052	TTGTGGGAGAAAAT	TTGTGGGACAAAAT
SFTPB; chr7:85744364	CTGGGCCGCATGC	CTGGGCCACATGC
SFTPB; chr7:85744391	CTCCGTCATCCTGC	CTACTCCATCATCCTGC

Supplementary Table 2. Taqman assay primers and probes. The specific primers (**a**) and probes (**b**) for the nine sites tested in this study are listed. Commercially available primers and probes for specific dbSNP sites are proprietary to Applied Biosystems (ABI).

Supplementary Table 3a. *SFTPB* SNP comparison between Sanger and pooled sequencing

Genomic Position	Substitution/ Gene Location	Function	Percent frequency from Sanger data[1]	Percent frequency from pooled sample data
Chr2: 85744339	CGC(R)→CAC(H) Exon 7	Coding, nonsynonymous	0.5	0.4
Chr2: 85744311	GAC(D) →GAT(D) Exon 7	Coding, synonymous	4.2	3.5
Chr2: 85744307	GCT(A)→ACT(T) Exon 7	Coding, nonsynonymous	0.5	0.5
Chr2: 85744282	C →A Intron 7	Non-coding	5.8	8.0
Chr2: 85744269	A →G Intron 7	Non-coding	21.2	21.9
Chr2: 85744264	T →C Intron 7	Non-coding	7.5	11.8
Chr2: 85744119	CCG(P) →CTG(P) Exon 8	Coding, synonymous	0.05	0.10

Supplementary Table 3b. Positions validated by Taqman assay.

Gene / Genomic Position	dbSNP Accession Number	Substitution/ Gene Location	Function	Percent frequency from pooled sample data	Percent frequency from Taqman assay
ACTB Chr7: 5534814	n/a	CTG(L)→CTT(L) Exon 4	Coding, synonymous	0.5	0.59
TP53 Chr17: 7514622	rs17880847	A →T Intron 10	Non-coding	0.6	0.77
TP53 Chr17: 7513782	rs17881850	C →T Intron 10	Non-coding	1.0	0.86
TP53 Chr17: 7512477	n/a	A →C Exon 11	Non-coding 3' UTR	0.7	0.81
APC Chr5: 112202922	n/a	CAA(Q) →CAG(Q) Exon 15	Coding, synonymous	0.8	0.50
APC Chr5: 112206391	rs2229994	CTA(L) →TTA(L) Exon 15	Coding, synonymous	1.2	1.31
APC Chr5: 112207052	n/a	TCT(S) →TGT(C) Exon 15	Coding, nonsynonymous	1.1	0.86

Supplementary Table 3. Validated SNPs. (**a**) SNP comparison between Sanger and pooled-sample sequencing at the *SFTPB* locus. (**b**) The 7 rare SNP positions validated by Taqman assay at the *TP53*, *APC* and *ACTB* loci. The correlation between called and actual frequencies for these 14 SNP positions shown in the main text **Fig. 2**.

Supplementary Table 4. Known and putative SNP positions identified by pooled-sample sequencing

Gene	Position in reference sequence	Reference → Observed	Frequency in data set (%)	Bonferroni-corrected P(plus), $\log_{10}$	Bonferroni-corrected P(minus), $\log_{10}$	dbSNP ref # (build 128)	dbSNP average heterozygosity	Amino Acid change
TP53	7520526 (intron 2)	C → G	78.1	-4547.5	-13062.7	rs1642785	0.468 ± 0.122	intronic
Chr17	7520344 (intron 3)	C → A	5.9	-481.6	-233.6	rs17883323	0.1 ± 0.2	intronic
	7520304 (exon 4)	G → A	0.8	-29.1	-11.0	rs1800370	0.013 +/- 0.079	synonymous
	7520273 (exon 4)	C → T	0.4	-4.9	-4.2	rs1800371	0.027 +/- 0.114	CCG(P)→TCG(S)
	7520197 (exon 4)	C → G	69.2	-16138.0	-13782.3	rs1042522	0.49 ± 0.07	CCC(P)→CGC(R)
	7519200 (exon 5)	C → T	0.1	-2.4	-5.0	n/a		CCG(P)→CTG(L)
	7519072 (intron 5)	C → T	0.1	-1.4	-1.6	n/a		intronic
	7518935 (exon 6)	A → G	1.2	-111.0	-172.2	rs1800372	0.024 +/- 0.106	synonymous
	7518274 (exon 7)	C → T	0.1	-9.3	-5.8	n/a		synonymous
	7518272 (exon 7)	G → A	0.4	-89.8	-49.8	n/a		GGC(G)→GAC(D)
	7518152 (intron 7)	C → T	8.1	-2118.2	-4662.7	rs12947788	0.246 +/- 0.250	intronic
	7518132 (intron 7)	T → G	7.9	-2800.5	-2665.4	rs12951053	0.296 +/- 0.246	intronic
	7517566 (intron 9)	T → C	1.5	-175.9	-99.8	rs1800899	unknown	intronic
	7514622 (intron 10)	A → T	0.6	-103.8	-95.0	rs17880847	0.011 +/- 0.074	intronic
	7513806 (intron 10)	G → C	0.2	-3.7	-4.3	rs17883043	0.012 +/- 0.076	intronic
	7513782 (intron 10)	C → T	1	-98.4	-96.7	rs17881850	0.012 +/- 0.076	intronic
	7513447 (exon 11)	G → A	0.4	-4.8	-30.7	rs16956880	0.019 +/- 0.095	UTR
	7513324 (exon 11)	G → A	0.3	-6.3	-4.5	rs17881366	0.010 +/- 0.071	UTR
	7513243 (exon 11)	C → A	0.5	-23.7	-22.6	n/a		UTR
	7513167 (exon 11)	G → A	1.8	-328.8	-255.0	rs4968187	0.095 +/- 0.196	UTR
	7512826 (exon 11)	G → A	3.9	-1122.7	-404.3	rs17884306	0.137 +/- 0.223	UTR
	7512544 (exon 11)	A → C	0.2	-2.8	-2.5	n/a		UTR
	7512539 (exon 11)	A → C	0.5	-17.0	-39.7	n/a		UTR
	7512477 (exon 11)	A → C	0.7	-109.2	-23.6	n/a		UTR
	7512431 (intergenic)	G → C	0.5	-69.5	-24.9	rs17883782	0.044 +/- 0.142	intergenic
ACTB	5535853 (intron 1)	C → T	29.4	-17920.7	-7530.5	rs2908425	0.172 +/- 0.237	intronic
Chr7	5535824 (intron 1)	G → T	0.3	-18.3	-1.6	rs13447394	0.011 +/- 0.074	intronic
	5535625 (intron 2)	A → G	0.4	-2.9	-29.0	n/a		intronic
	5535302 (intron 3)	C → T	0.5	-39.9	-36.9	n/a		intronic
	5535288 (intron 3)	C → G	0.1	-1.4	-3.0	rs13447399	0.011 +/- 0.074	intronic
	5534892 (intron 3)	T → C	46.6	-37145.2	-39328.9	rs852423	0.496 +/- 0.046	intronic
	5534814 (exon 4)	G → T	0.5	-89.3	-77.1	n/a		synonymous
	5534731 (exon 4)	C → T	0.1	-4.3	-7.7	n/a		GCC(A)→GTC(V)
	5534242 (exon 5)	C → T	0.1	-4.0	-1.7	rs13447407	0.011 +/- 0.075	synonymous
	5534203 (exon 5)	G → A	2.7	-1085.8	-1080.7	rs11546939	0.068 +/- 0.171	synonymous
	5534010 (exon 6)	C → T	0.5	-3.4	-56.5	rs11546907	unknown	synonymous
	5533695 (exon 6)	C → T	0.4	-1.7	-25.2	rs13447411	0.011 +/- 0.074	UTR
	5533645 (exon 6)	G → A	0.4	-24.3	-32.9	rs11546906	unknown	UTR
	5533638 (exon 6)	G → A	31.9	-14272.0	-9855.0	rs7612	0.558 +/- 0.180	UTR
	5533625 (exon 6)	G → A	0.3	-17.3	-5.9	rs11546905	unknown	UTR
	5533452 (exon 6)	C → G	1.5	-294.6	-82.7	rs3210032	0.049 +/- 0.148	UTR
APC	112201308 (intron 14)	A → T	0.7	-13.5	-7.8	n/a		intronic
Chr5	112201798 (exon 15)	C → T	0.5	-35.7	-6.2	rs33974176	0.025 +/- 0.109	CCA(P)→TCA(S)
	112202576 (exon 15)	T → C	0.3	-9.9	-1.9	n/a		TTG(L)→TCG(S)
	112202661 (exon 15)	G → A	0.4	-8.1	-24.5	n/a		synonymous
	112202922 (exon 15)	A → G	0.8	-28.6	-82.3	n/a		synonymous
	112203139 (exon 15)	G → C	1.3	-419.9	-96.7	rs1801166	unknown	GAA(E)→CAA(Q)
	112203516 (exon 15)	T → A	0.5	-70.7	-43.9	n/a		synonymous
	112205070 (exon 15)	G → A	60.5	-11973.1	-15140.9	rs465899	0.436 +/- 0.167	synonymous
	112206111 (exon 15)	G → A	0.4	-26.5	-14.9	rs2229993	0.044 +/- 0.142	synonymous
	112206391 (exon 15)	C → T	1.2	-66.5	-56.4	rs2229994	0.023 +/- 0.104	synonymous
	112206694 (exon 15)	G → A	1.6	-148.5	-368.0	rs2229995	0.038 +/- 0.132	GGT(G)→AGT(S)
	112206894 (exon 15)	A → G	2.1	-141.0	-321.2	rs35043160	0.193 +/- 0.243	synonymous
	112207052 (exon 15)	C → G	1.1	-151.5	-26.8	n/a		TCT(S)→TGT[C]
	112207808 (intergenic)	C → A	5.8	-293.4	-247.8	rs1804197	0.166 +/- 0.236	UTR
SFTPB	85744391 (exon 7)	G → A	0.5	-81.4	-108.8	n/a		GTC(V)→ATC(I)
Chr2	85744363 (exon 7)	G → A	0.3	-46.7	-28.6	n/a		CGC(R)→CAC(H)
	85744339 (exon 7)	G → A	0.4	-60.0	-78.8	rs3024809	0.030 +/- 0.119	CGC(R)→CAC(H)
	85744311 (exon 7)	C → T	3.5	-1377.3	-1554.2	rs3024810	0.223 +/- 0.249	synonymous
	85744307 (exon 7)	G → A	0.5	-98.1	-96.0	rs36210375	unknown	GCT(A)→ACT(T)
	85744282 (intron 7)	C → A	8	-3819.7	-3600.3	rs893159	0.081 +/- 0.185	intronic
	85744269 (intron 7)	A → G	21.9	-20234.0	-13347.0	rs2304566	0.310 +/- 0.243	intronic
	85744264 (intron 7)	T → C	11.8	-8190.4	-6199.5	rs762548	0.310 +/- 0.243	intronic
	85744119 (exon 8)	C → T	0.1	-2.4	-3.8	rs35076740	unknown	synonymous

Supplementary Table 4. Known and putative SNP positions identified by pooled-sample sequencing. All positions identified as SNPs from all genes sequenced are listed. The Bonferroni-corrected *P*-value ($\log_{10}$) for each substitution is listed for both the sense (plus) and antisense (minus) strands in columns 5 and 6, respectively. A *P*-value < 0.05 was considered significant (equivalent to -1.3 on a $\log_{10}$ scale). A given substitution was required to have a significant P-value on both strands to be identified as a SNP. The accession numbers and corresponding average heterozygosity values for positions listed in dbSNP are shown in columns 7 and 8, respectively. The absence of an accession number means that SNP was not found in dbSNP. Non-synonymous SNPs that were included in the comparative genomics analysis are highlighted in yellow. UTR = untranslated region.

Supplementary Methods

I. DNA Preparation and sequencing

Genomic DNA Samples. We extracted genomic DNA from 1,111 random, anonymous Guthrie cards collected for newborn screening between 1993 and 2000 by the Missouri Department of Health and Senior Services (DHSS)[1,2]. Both the Missouri DHSS Institutional Review Board and the Washington University Human Research Protection Office reviewed the project and approved waiver of individual consent for use of de-identified DNA samples under regulation 45CFR467.116d (for the regulation, see http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.htm#46.116). Each individual DNA sample was anonymously linked to clinical characteristics in a vital statistics (birth-death certificate) database maintained by the Missouri DHSS to determine ethnicity. Ethnicities within the population were as follows: European-American = 871 (78.4%), African-American = 196 (17.6%), Hispanic = 34 (3.1%), Asian = 5 (0.5%) and unknown = 5 (0.5%).

Genomic DNA Extraction. Genomic DNA extraction was adapted from previously described methods by Hamvas et al [1]. Modifications to this process included sequential incubation in 200 μL distilled water at room temperature for 60 min and 45 min on an orbital shaker at 400 rpm. DNA was extracted in a solution of 200 μL of 10 mM Tris with 1 mM EDTA, pH 6.8 and 2% (wt/vol) Chelex 100 Molecular Biology Grade chelating resin (Bio-Rad, Hercules, CA, U.S.A.).

DNA Quantification. High-throughput DNA quantification was performed using a fluorescent nucleic acid stain in a 384-well format. To mimic the fragmented and denatured quality of the sample DNA, commercial human genomic DNA at 163 ng/μL (Promega, Madison, WI, U.S.A.) was sonicated for 15 seconds at maximal power using a Misonix XL2020 Ultrasonic Processor sonicator (Misonix, Farmingdale, NY, U.S.A.) and then heated at 100^{o}C for 25 minutes. This DNA was then serially diluted by 50% eight times. DNA concentrations of 20.38, 10.19, 5.09, 2.55, 1.27 and zero ng/μL were added, as described below, to the 384 plate in duplicate to be used as a DNA standard. SYBR Gold (Molecular Probes, Eugene, OR, U.S.A.) was the fluorescent nucleic acid stain used due to its ability to bind both single and double-stranded DNA. Immediately prior to each quantification experiment, the stain was diluted 1:1000 in 10 mM Tris with 1 mM EDTA, pH 7.5 and protected from light. For quantification determination, each well on a black 384-well polystyrene Fluotrac 200 microtiter plate (Greiner Bio-One, Monroe, NC, U.S.A.) contained a final volume of 20 μL consisting of 2 μL DNA suspension, 8 μL of 10 mM Tris with 1 mM EDTA, pH 7.5 and 10 μL SYBR Gold 1:1000 suspension. The plate was protected from light until processing. Fluorescent detection was done on a Synergy HT (Biotek, Winooski, VT, U.S.A.) plate reader. With excitation at 485/20 and emission 528/20, the machine automatically determined the optimum sensitivity level for each experiment by scaling fluorescence against the negative control wells. The fluorescent plate reader then determined the fluorescent level in each well on the plate. A mean fluorescent value was determined for each concentration of the DNA standard and a linear regression was then generated. The equation of the linear regression was then used to calculate the concentration of DNA in each sample well.

Pooling of DNA We elected to pool 80 ng of DNA per individual. This amount was chosen simply to have a pool of DNA large enough to perform multiple PCR and sequencing reactions. Pooling was done manually and the final volume was over 29 mL. Concentration of pooled DNA was done using the Qiagen QIAvac 96-well vacuum manifold and QIAquick 96 PCR Purification Kit (Qiagen, Valencia, CA, U.S.A.). This kit is intended to purify DNA from 100 bp to 10 kb. Gel electrophoresis of the highly fragmented genomic DNA isolated from blood spots shows a mean size of approximately 3 kb with no visible smear above 10 kb (not shown). The resulting DNA suspension was approximately 2 mL with a concentration, determined by SYBR Gold staining, of 2.6 ng/μL.

PCR. PCR primers were designed using Primer 3 (http://frodo.wi.mit.edu/primer3/input.htm) with parameters previously described[3]. All primers were purchased through Integrated DNA Technologies (Coralville, IA, U.S.A.). Reference sequence for each human gene was obtained from the UCSC Genome Browser (http://genome.ucsc.edu/index.html) and, for the pUC19 vector, from New England Biolabs NEBcutter website (http://tools.neb.com/NEBcutter2/index.php). Refseq accession numbers for human reference sequences were: *TP53*, NM000546; *SFTPB*, NM000542; *APC*, NM000038; *ACTB*, NM001101. Each individual PCR reaction was designed to include an average of 30 genome equivalents per individual. Assuming a Poisson distribution for the number of molecules input into the pool, this amount was determined to minimize the likelihood that one individual DNA sample would be omitted in a PCR reaction, while keeping the volume of PCR reactions within common standards as well as optimizing the allocation of the pooled DNA. The thermostable polymerase utilized was PfuUltra High-Fidelity DNA Polymerase (Stratagene, La Jolla, CA, U.S.A.) due to its reported extreme accuracy. PCR reaction contents were as follows: 1X final concentration of 10X PfuUltra Buffer, 200 μM dNTPs (Invitrogen, Carlsbad, CA, U.S.A.), 400 nM forward primer, 400 nM reverse primer, 1M betaine (Sigma-Aldrich/Fluka, St. Louis, MO, U.S.A.), 2.5 units PfuUltra DNA polymerase and 150.8 ng of pooled genomic DNA. The final reaction volume was 100 μL. Annealing temperatures and extension times varied slightly between reactions and are listed, along with primer sequences, in **Supplementary Table 1**. In general, all PCR reactions were as follows: 1) 93^{o}C x 2 minutes; 2) 93^{o}C x 30 seconds; 3) 56-62^{o}C x 30 seconds; 4) 65^{o}C x 1-5 minutes; 5) steps 2-4 for 28 cycles; 6) 65^{o}C x 10 minutes; 7) hold at 4^{o}C. In total, there were 14 human PCR amplicons covering 13,237 bp and 19 exons in the four genes samples as well as a single amplicon from the pUC19 vector. All PCR products then went through the QIAquick PCR purification protocol (Qiagen, Valencia, CA, U.S.A.) and quantified using the Nanodrop ND-1000 spectrophotometer.

Random Amplicon Ligation. Next-generation sequencing was designed to sequence major portions of or even whole genomes in a single machine run. Equal coverage across the genome is dependent upon random fragmentation followed by aligning and tiling the millions of small DNA sequences against the reference genome. Our computational analysis suggested that random fragmentation by sonication cannot be achieved for small DNA sequences such as PCR amplicons less than about 1500 bp (not shown). A majority of such fragments would be fragmented in the middle of the sequence resulting in overrepresentation of the unfragmented end pieces of the amplicon in the final sequencing output. To overcome this, we randomly ligated a normalized number of pooled PCR-amplified products and sonicated the resulting concatamers. From each PCR product, 40x10^{10} molecules of each amplicon were pooled. A blunt ended ligation (PfuUltra generates blunt ended products) was performed as follows: 1X of 10X T4 DNA Ligase Buffer (New England Biolabs, Ipswitch, MA, U.S.A.), 2400 units of 400 units/μL T4 DNA Ligase (New England Biolabs, Ipswitch, MA, U.S.A.), 120 units of 10 units/μL T4 Polynucleotide Kinase (New England Biolabs, Ipswitch, MA, U.S.A.), 15% (wt/vol) polyethylene glycol 8000 MW (Sigma-Aldrich, St. Louis, U.S.A.) and the PCR amplicon pool brought the total volume to 600 μL. This volume was aliquoted into four equal parts and incubated for 17 hours at 22^{o}C, followed by 20 minutes at 65^{o}C and held at 4^{o}C thereafter. Agarose gel electrophoresis was performed with a small amount of the resulting product to confirm concatenation. Electrophoresis confirmed concatenated products >10kb (not shown).

Sonication. Random fragmentation of the ligated concatemers was performed using the Covaris S2 sonicator (Covaris, Woburn, MA, U.S.A.). Each 150 μL ligation aliquot was diluted in 350 μL of sterile, distilled water and transferred to a 13mm x 65mm borosilicate glass tube with polypropylene screw-top (Covaris, Woburn, MA, U.S.A.). The samples were sonicated individually with the sonicator programmed as follows: 15 cycles, duty cycle 20%, intensity 10, cycles/burst 1000, time 60 seconds, bath temperature limit 20^{o}C. Agarose gel electrophoresis confirmed fragmentation of concatenated DNA as a smear from approximately 75-3000 bp with maximal ethidium bromide staining intensity between 150-200 bp (not shown).

DNA Library Preparation for Sequencing. Following fragmentation, a DNA library for sequencing was prepared according to the protocol described in the Preparing Samples for Sequencing Genomic DNA document provided by Illumina for the Genome Analyzer and starting with the end repair step. A difference from the described protocol included gel excision and purification of fragments from 125-400 bp. The concentration of PCR-enriched, adapter-ligated DNA fragments following clean up was determined by

Nanodrop ND-1000 spectrophotometer and found to be 29.3 ng/μL.

Sequencing. While we could have generated all data from a single flowcell, we were sharing flowcell lanes with other investigators. Therefore, the same DNA library was sequenced in a total of 12 flowcell lanes from four different dates. The protocol for preparing DNA samples for sequencing is described on page 12 of the Illumina document entitled Preparing Sample DNA for Cluster Generation. Sample DNA was diluted to 10 nM in Qiagen EB buffer as recommended. In the initial sequencing run, a titration of DNA concentrations of 0.5, 1 and 2 pM was performed in three separate lanes to determine optimum cluster generation. Two pM was determined to be optimal and all subsequent sequencing was performed at this concentration. From the same library, we performed an additional 9 lanes of sequencing on three separate dates.

II. TAQMAN Assays

For independent, individual validation of putative SNPs, we performed Taqman assays on all individuals in our pool. The primers / probes design, manufacture, and testing is done at Applied Biosystems (ABI) manufacturing facilities. Each designed probe set contains 2 probes, with VIC (allele 1) and FAM (allele 2) reporter dyes linked to the 5' end of each respective probe, a MGB, and a nonfluorescent quencher (NFQ) at the 3' end of both probes. The primers and probes for the nine sites tested in this study are listed in **Supplementary Table 2**. Sites found in dbSNP had pre-existing primers and probes that are commercially available and proprietary to ABI. For each SNP site of interest, we submit a sequence file identifying the polymorphic base, along with 300 flanking bases both 5' and 3' of each SNP site to ABI. ABI then designs an assay using proprietary algorithms that minimize adverse assay effects, such as base runs or secondary structure formation. Primers and probe sequences are matched by melting temperature, permitting universal PCR assay conditions. They also use mass spectrometry to verify the oligonucleotide sequence and perform further testing to ensure proper formulation of the primer and probe mix. The assays are also functionally tested using an allelic detection test prior to delivery, then delivered in a single tube format. We assayed 12 96-well plates with no-template (blank) controls in wells H10, H11, H12 of each plate for the nine SNP positions listed in **Supplementary Table 2**. For each plate, a master mix was made of 1,250 μl 2X Taqman genotyping master mix (ABI, part # 4371357) and 125 μl 20X SNP genotyping assay mix (ABI, part # 4332027). Next, 13.75 μl master mix is aliquoted into each well of a 96-well optical reaction plate (ABI, part # 4346906). Then, 1 μl of genomic DNA from each test plate, along with 10.25 μl DNase-free water, is added to each well of the optical reaction plate. Plates are then covered with an optical adhesive film (ABI, part # 4311971) and PCR is performed on an MJ PCT-225 thermocycler. An initial enzyme activation is required at 95C for 10 minutes, followed by 40 cycles of denaturing at 92C for 15 seconds and anneal/extension at 60C for 1 minute. Upon completion of the PCR amplification, endpoint plate reading and genotype calls are performed on the ABI 7500 FAST Real Time PCR system.

III. SNPSeeker: SNP detection using Large Deviation Theory

Large Deviation Theory applied to Illumina Genome Analyzer output. We found that existing second-generation base calling programs were unable to detect and quantify rare variants in a large pool of multiple individuals (Vallania and Mitra, unpublished results), so we developed a new base calling method based on Large Deviation Theory and named SNPSeeker. All the software and sequence data used for the analysis is available at (http://www.genetics.wustl.edu/rmlab/). Sequences were mapped back to reference by using an ungapped alignment algorithm and allowing 2 mismatches (this allowed unambiguous mapping) and then positions were considered separately. For each position i in the reference, all the sequence bases aligned to that position are analyzed. This analysis relies on the assumption that the sequencing process generates sequences independent from one another. A second assumption is that sequences are formed by nucleotides (corresponding to sequencing cycles) that are independent from one another as well.

Given these assumptions, it is possible to define subsets of nucleotides for each cycle j, sequencing run d and strand s. These nucleotides are drawn from the same background probability distribution and, as part of our assumptions, are independent from one another. Each set can be therefore defined as a series of n i.i.d. random variables (where n is the number of total bases for each considered subset)

$$X_{j,d,s,1}, X_{j,d,s,2}, X_{j,d,s,3}, \ldots, X_{j,d,s,n}$$

drawn according to a distribution $Q_{j,d,s}(x)$ with symbols $\chi = \{A, C, G, T, N\}$. For each subset, it is possible to define an empirical probability distribution, also known as *type*, or $P_{j,d,s}$, as

$$P_{j,d,s} = \left(\frac{As}{n}, \frac{Cs}{n}, \frac{Gs}{n}, \frac{Ts}{n}, \frac{Ns}{n} \right)$$

If we assume the null hypothesis of observing no polymorphism at position i, $i-1$ and $i-2$ in the reference sequence, then $Q_{j,d,s}(x)$ will exactly correspond to the distribution of errors derived from pUC19 as described in the previous paragraph. However, this can be generalized into

$$Q_{j,d,s}(x) = \sum_{n \in \aleph} Pr(x|M_i = n, j, d) * Pr(M_i = n|s, \tau) \tag{1}$$

where:

- $Pr(x|M_i = n, j, d)$ corresponds to the distribution of errors computed from pUC19 and indicates the probability of seeing a base x in the sequence at cycle position j on run d given that the original base at position i in the reference, M_i, is equal to n, where $x \in B$ and $n \in \aleph$ (see *Definition of the Error Model* section).
- $Pr(M_i = n|S, s, \tau)$ corresponds to the probability of observing nucleotide n in the reference sequence at position i, $M_i = n$, given the strand s and the true allele frequency vector τ.

To determine the likelihood of observing a SNP at position i, we compute the amplitude of the deviation between $P_{j,d,s}$ and $Q_{j,d,s}$, formulated as the probability of observing a number of nucleotides different from the major allele (the nucleotide in the reference at position i) equal or greater than the number observed in the analyzed subset. According to *Sanov's theorem* [4], this is equivalent to

$$Q^n_{j,d,s}(E) = 2^{-n_{j,d,s} D(P_{j,d,s} \| Q_{j,d,s})} \tag{2}$$

id est the probability of generating a set of *types* E satisfying the above described conditions. $D(P_{j,d,s} \| Q_{j,d,s})$ corresponds to the relative entropy or *Kullback–Leibler* distance [4] between $P_{j,d,s}$ and $Q_{j,d,s}$. Since we expect true SNPs to be detected on both strands and errors to behave independently with respect to the strand, we calculated a cumulative p-value for each strand as

$$Q_s(E) = \prod_d \prod_j Q^n_{j,d,s}(E) \tag{3}$$

given the previous independence assumptions. Each probability value $Q^n_{j,d,s}(E)$ is Bonferroni-corrected for the total number of tests performed at position i and $Q_s(E)$ is additionally corrected for the total number of tests performed at each position in the reference sequence (corresponding to its length). In order for position i to be called as a SNP, $Q_s(E)$ must be below a significancy cutoff ($\alpha = 0.05$) given the appropriate corrections. If this is true for both strands then position i contains at least one allele variant.

Estimation of the true SNP frequencies by Non-Linear Least Square Fit. For every position in which a SNP was found, we estimated its true allele frequencies by performing a non-linear least square fit. Given the relative entropy term in (2), we can decompose it as

$$n_{j,d,s} D(P_{j,d,s} \parallel Q_{j,d,s}) = -n_{j,d,s} \left[H(P_{j,d,s}) + \sum_{x \in B} P_{j,d,s}(x) log_2 Q_{j,d,s}(x) \right] \tag{4}$$

Given $P_{j,d,s}$ and $Q_{j,d,s}$, their relative entropy will converge to 0 the closer they will get to each other. That will correspond to a descrease in the difference between $n_{j,d,s} H(P_{j,d,s})$ and $-n_{j,d,s} \sum_{x \in B} P_{j,d,s}(x) log_2 Q_{j,d,s}(x)$ so that eventually they will converge to

$$n_{j,d,s} H(P_{j,d,s}) = -n_{j,d,s} \sum_{x \in B} P_{j,d,s}(x) log_2 Q_{j,d,s}(x) \tag{5}$$

If we repeat this consideration for each cycle j, day d and strand s, we can define an observation vector y and a least-square vector $\widehat{y}$ such that

$$y = \begin{bmatrix} n_{1,1,+} H(P_{1,1,+}) \\ \cdots \\ \cdots \\ \cdots \\ \cdots \\ n_{j,d,s} H(P_{j,d,s}) \end{bmatrix} \widehat{y} = \begin{bmatrix} -n_{1,1,+} \sum_{x \in B} P_{1,1,+}(x) log_2 Q_{1,1,+}(x) \\ \cdots \\ \cdots \\ \cdots \\ \cdots \\ -n_{j,d,s} \sum_{x \in B} P_{j,d,s}(x) log_2 Q_{j,d,s}(x) \end{bmatrix}$$

As defined in (1), $Q_{j,d,s}$ will depend on $Pr(M_i = n|s, \tau)$ which is conditioned on τ. τ is defined as a stochastic vector with 4 entries, each representing the probability of seeing a particular nucleotide in a given position in the reference sequence (with respect to the forward strand as the adopted convention). The τ vector that minimizes $|y - \widehat{y}|$ is defined $\widehat{\tau}$ and corresponds to

$$\widehat{\tau} = \arg\min_{\tau} \sum_j \sum_d \sum_s \left\{ n_{j,d,s} \left[H(P_{j,d,s}) + \sum_{x \in B} P_{j,d,s}(x) log_2 \sum_{n \in \aleph} Pr(x|M_i = n, j, d) * Pr(M_i = n|S, s, \tau) \right] \right\}^2 \tag{6}$$

This is computed iteratively, initially defining all possible τ probability vectors starting with a resolution ρ of 1 significant digit (0.1) and then selecting the vector that minimizes $|y - \widehat{y}|$. After this step, ρ is decreased 10-fold (0.01 or 2 significant digits) and only vectors located within a range defined as $\widehat{\tau} \pm 20 * \rho$ for each vector entry are then analyzed. This allows the values of $\widehat{\tau}$ to be refined without requiring massive computational power. These steps are iterated progressively until the estimate at each step is refined to resolution of 3 significant digits (the final resolution is a user defined parameter).

IV. Comparative Genomics Analysis using SIFT, PolyPhen, and the Likelihood Ratio Test

While most common SNPs are likely neutral, 50% of rare (< 5% in the general population) nonsynonymous SNPs have been estimated to be deleterious[5], many of which may produce significant phenotypic effects, even if heterozygous. We thus sought to determine if any of the identified nonsynonymous SNPs are deleterious. The genome sequences of multiple vertebrate species make it possible to identify functional sequences by their conservation across species. Given enough evolutionary time, even a signle amino acid position has a very small probability of being conserved by chance.

To test whether any of the nonsynonymous SNPs are deleterious, we used three different prediction algorithms. Two of the algorithms, SIFT[6] and PolyPhen[7], make predictions based on conservation and structural motifs, respectively. However, without a formal probabilistic framework the rate of false positive prediscitons is difficult to know. The third algorithm is based on a likelihood ratio test (LRT) that compares the probability of conservation across species at a single amino acid position under a neutral model and a model of selective constraint[8]. Under the neutral model, the probability of amino acid conservation is calculated using the synonymous substitution rate for each gene. Under the model of selective contraint, the nonsynonymous substitution rate is allowed to be a fraction of the synonymous rate. This codon based test is similar to a test developed for noncoding sequences.

For each gene, coding sequences were downloaded from ENSEMBL (www.ensembl.org), the translated protein sequences were aligned using ClustalW, and then translated back into their corresponding DNA sequences. The number of species ranged from 15 (*ACTB*) to 21 (*APC*) with the most distant species ranging from Platypus (*SFTPB*) to Zebrafish (*TP53*). The maximum likelihood phylogenetic relationship of each gene was obtained using Phylip and the maximum likelihood estimate of the synonymous substitution rate was obtained using HyPhy and the MG94xHKY85_3x4 nucleotide substitution model. The total synonymous substitution rate ranged from 4.37 for SFTPB to 14.98 for TP53. The likelihood ratio test compares two models in order to determine whether a nonsynonymous SNP is deleterious and disrupts a conserved amino acid position. Under the null model, the likelihood of the data was calculated assuming no constraints such that both the synonymous and nonsynonymous substitution rate of the SNP containing codon were equal to the synonymous rate for the entire gene. Under the alternative model, the likelihood of the data was calculated assuming the SNP containing codon was constrained such that the nonsynonymous rate was a fraction of the synonymous rate. Deleterious SNPs were defined as those for which nonsynonymous rate was significantly less than the synonymous rate.

ACTB

- Species: 16
- Tree order: human, chimp, pongo, macaca, treeshrew, mouselemur, dog, elephant, squirrel, rat, mouse, opposum, platypus, chicken, xenopus
- dS: 5.76853
- dN: 0.24075

APC

- Species: 21
- Tree order: human, chimp, macaque, mouselemur, elephant, tenrec, cat, dog, cow, microbat, hedgehog, shrew, guineapig, mouse, rat, treeshrew, pika, rabbit, oppossum, platypus, xenopus
- dS: 8.48990
- dN: 0.68431

SFTBP

- Species: 15

- Tree order: human, chimp, orangutan, macaque, elephant, tenrec, cow, bat, dog, horse, rat, mouse, bushbaby, mouselemur, hedgehog, pika, opposum, platypus
- dS: 6.49362
- dN: 1.36192

TP53

- Species: 18
- Tree order: human, chimp, macaque, mouselemur, cat, cow, microbat, shrew, armadillo, mouse, rat, rabbit, pika, xenopus, Fugu, stickleback, medaka, daniorerio
- dS: 28.94836
- dN: 1.54825

Supplementary Results

I. Error Model Generation using pUC19

Definition of the error model. For all sequencing performed, an internal control consisting of a 1,276 bp PCR amplified sequence from *E.coli*-cloned pUC19 vector was included in order to model the likelihood of observing errors in a SNP free context. Bases 1-800 of this amplicon were used to parameterize the algorithm with each machine run. Bases 801-1,276 were then used as a negative control test sequence. This model specifies the probability of observing an error in a sequencing read as a function of 1) the true identity of the base being sequenced, 2) the identity of the observed base, 3) the identities of the two reference bases immediately upstream of the base being sequenced, and 4) the current cycle number of the sequencing read (i.e. cycle 1 to 36).

For each sequencing cycle j in the read, we calculated the probability of observing a base x, where $x \in B, B = \{A, C, G, T, N\}$, given a base n in the reference sequence such that $n \in \aleph, \aleph = \{A, C, G, T\}$. Due to observed variability in sequencing errors between machine runs (**Fig. 1B**), we created a new error model for each machine run.

SNPSeeker uses a 2^{nd} Order Model which assumes a 2^{nd} order dependency between reference sequence nucleotides. Thus, we assume that the likelihood of i being sequenced correctly depends on $i - 1$ and $i - 2$ (this model is computed starting from read cycle 3). We found that the first twelve bases of each Illumina read contained significantly fewer errors than later bases (**Fig. 1A**) so we only used these bases to identify sequence variation. Since we use a 2nd order dependency model for SNP identification, only mismatches at bases 3-12 of each read were used to identify SNPs. On average, each position in the reference sequence was observed 30,593 times in bases 3-12 of the Illumina sequencing reads (i.e. 13.8-fold coverage of 1,111 diploid genomes).

Illumina Quality Scores. Our error models do not take into account the ILLUMINA quality scores since we did not find any improvement by including them in our error models. As shown in **Supplementary Fig. 1**, we have plotted the quality scores (QS) generated by the Illumina Genome Analyzer against the true quality scores, defined as the $-log_{10}$ of the probability of incorporating a nucleotide different from the reference sequence base (calculated from the negative control vector as described). As shown in the **Supplementary Fig. 1**, quality scores less than 10 are informative of poor bases, but higher quality scores do not accurately reflect true error rates, as is evident by the plateau in the trend line for quality scores above 10. Since most bases with quality scores less than 10 occured at the ends sequencing reads, and this information is already accounted for by our read position parameter, incorporating QS into the algorithm did not improve SNP calling. However, for other applications needing to utilize QS data, these measurements are easily integrated into the Large Deviation Theory framework.

Non-specific PCR amplification analysis. To assure that non-specific PCR amplification across the human genome was not a source of error in our SNP calling, we performed the following analyses. First, all primer combinations were designed to avoid repetitive regions of the human genome. Second, we performed *in silico* PCR for each primer combination (via the UCSC Genome Browser) against the entire human genome to demonstrate that a single, unique PCR product, spanning the region of interest, was the only product expected. Finally, we used our alignment algorithm to map back all sequencing reads against the entire human genome. We found that 544,195 reads (1.45% of the total) mapped back to more than one location in the human genome. We then excluded these reads from further SNP calling analysis and found that none of our 64 SNP calls were significantly altered in terms of identity or frequency. These results indicate that non-specific PCR amplification is not adversely affecting SNP identification.

II. How SNPSeeker improves SNP calling

A single allele, occurring in a population of 2222 alleles, has a frequency of 0.00045. Despite limiting the data for SNP calling at bases 1-12, the average likelihood of an error across these bases was 0.00065 and rose dramatically as more bases were included in the analysis. In order to identify, with a high degree of certainty, true polymorphisms that occurred in the pool at a frequency less than the incipient error rate of the sequencing platform, we designed SNPSeeker, an algorithm based on Large Deviation Theory, and implemented SNPSeeker into the analysis. **Supplementary Fig. 2** demonstrates how SNPSeeker further refines SNP identification above and beyond simply using the first 12 bases of each read.

When considering pUC19 and trying to identify a single allele in the pool, using 32 bases per read identified 785 bases (out of 800 in the training set) as potential SNPs. When only considering the first 12 bases per read without using the algorithm, 705 bases out of 800 were identified as SNPs.

By implementing SNPSeeker, zero positions were identified as SNPs. **Supplementary Fig. 2** demonstrates how each of these conditions would affect SNP calling in the *SFTPB* amplicon. Using 12 bases and a frequency cutoff calculated on pUC19, we identified 19 potential SNP sites. By applying SNPSeeker, over 50% of the sites are eliminated and only 9 SNP positions remain.

III. Comparative Genomics Analysis Demonstrates that Rare Non-synonymous SNPs are Deleterious

Twelve nonsynonymous SNPs from all four genes tested were identified in this analysis. Of these 12 sites, SIFT identified 7 as deleterious, PolyPhen 3, and the LRT predicted that 5 would disrupt highly significant positions (dN/dS <1 and LRT, P <0.001) (**Supplementary Table 5**). Five of the 7 sites identified by SIFT are not found in dbSNP. One of these five, position 112207052 in *APC* (marked with †), was validated by Taqman assay. Four of the five non-dbSNP sites (2 in *TP53* and 2 in *APC*, marked with asterisks), including *APC* 112207052, were previously published in the germline of individuals with cancer [9–12]. Four of the five evolutionarily conserved amino acid positions identified by the LRT are perfectly conserved across all species. If recessive, the phenotypic effects of the deleterious SNPs should rarely be observed and may be quite severe.

Supplementary References

1. Hamvas,A. *et al.* Population-based screening for rare mutations: high-throughput DNA extraction and molecular amplification from Guthrie cards. *Pediatr. Res.* 50, 666- 668 (2001).

2. Hamvas,A. *et al.* Comprehensive genetic variant discovery in the surfactant protein B gene. *Pediatr. Res.* 62, 170-175 (2007).

3. Mitra R.D. et al. Digital genotyping and haplotyping with polymerase colonies. *Proc. Natl. Acad. Sci. U.S.A.* 100, 5926-5931(2003)

4. Cover T. and Thomas J.A. *Elements of Information Theory.* Wiley Interscience (1991)

5. Fay J.C., Wyckoff G.J. and Wu C.I. Positive and negative selection on the human genome. *Genetics* 158, 1227-1234 (2001)

6. Ng P.C. and Henikoff S. Accounting for human polymorphisms predicted to affect protein function. *Genome Res.* 12, 436-446 (2002)

7. Ramensky V., Bork P. and Sunyaev S. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.* 30, 3894-3900 (2002)

8. Doniger S.W. *et al.* A catalog of neutral and deleterious polymorphism in yeast. *PLoS. Genet.* 4, e100183 (2008)

9. Azzopardi D. *et al.* Multiple rare nonsynonymous variants in the adenomatous polyposis coli gene predispose to colorectal adenomas. *Cancer Res.* 68, 358-363 (2008)

10. Srivastava S. *et al.* Germ-line transmission of a mutated p53 gene in a cancer-prone family with Li-Fraumeni syndrome. *Nature* 348, 747-749 (1990)

11. Varley J.M. *et al.* Are there low-penetrance TP53 alleles? Evidence from childhood adrenocortical tumors. *Am. J. Hum. Genet.* 65, 995-1006 (1999)

12. Zhou X.L. *et al.* Definition of candidate low risk APC alleles in a Swedish population. *Int. J. Cancer* 110, 550-557 (2004)

SNPSeeker Readme

This software allows you to analyze sequencing data generated by the Illumina GA machine from Pooled DNA sequencing experiments to detect Single Nucleotide Polymorphisms (SNPs) and estimate their frequency in the pool. The pipeline is structured as follows:

1. Reads are aligned to a reference sequence (1 or more DNA sequences)
2. Aligned reads are tagged in order to identify them according to the sequencing run/flowcell they were generated from (this is important for a good error model as each run will have a significant variability in the error profile)
3. Aligned and tagged reads are then used for building an error model (each run is required to contain a negative control sample that is SNP free, e.g. plasmid DNA)
4. SNPSeeker can then be used to analyze the aligned tagged reads using the error model file. Its output will correspond to the positions in each DNA sequence that will contain a SNP and its estimated pool frequency (fraction of As, Cs, Gs, Ts in the pool).

   The output is structured as follows:

   [DNA sequence name] [position] [log10 pvalue for + strand] [log10 pvalue for - strand] [A] [C] [G] [T]

1 Read Alignment

ALIGNER accepts reads in SCARF format and reference sequences in FASTA format (Unix newline).SCARF is one of the 3 default formats produced by the Illumina pipeline. Each line corresponds to a single read and has the following format:

```
HWI-EAS158_11:5:1:117:369:GCAAAGAACACGGCTAAGTGTGCTGGGGACCT:40 40 40 40 40 40 40 40 40 40 33 40 40 40 40 2 15 20 40 40 40 40 35 40 13 10 21 1 1 18 11 40
```

The program returns in output a tab-delimited file containing the name of the reference sequence, read id, position in the reference sequence and error profile for each read. command line:

```
./ALIGNER [Read file in SCARF format] [Reference sequences in FASTA format] [number of accepted mismatches] [number of bases to align starting 5']
```

2 Error Model Generation

Before generating the error model the alignment output files coming from a single run need to be concatenated together (use UNIX cat command) and tagged. The tag can be added using alignment_tagger.pl on the alignment file.
command line:

```
./alignment_tagger.pl [ALIGNMENT file] [TAG]
```

In order to generate an error model run perl script error_model_generator.pl on ALIGNER output and Error Model sequence [WARNING make sure the sequence was included in the FASTA file used to align the reads]. It is possible to specify the first and last base to include in the error model generation so that the remaining positions can be used as a negative control. command line:

```
./error_model_generator.pl [ALIGNER tagged file] [Error Model Reference sequence] [number of mismatches allowed] [first base to include] [last base to include (in case you want to train only on a portion of the model and use the other as a control)] [use pseudocounts (y/n)]
```

3 SNP Detection

Once Error model file is generated, it can be used to analyze the alignments using our implementation of Large Deviation theory.

command line:

```
./SNPSeeker [ALIGNER output] [Reference Sequence] [Error model file] [number of cycles analyzed (reccomended 12)] [p-value cutoff in log10 scale (0.05 → -1.301)]
```

4 Example Files

This is an example of how to run the pipeline on the data used for the analysis published in Druley TE et al Nature Methods 2009. The original input files can be found at these links

=Lanes from November 15th 2007=

http://cgsweb.wustl.edu/~fvallania/1_nature_methods_2009/NM_2009_LANE1_20071115.txt
http://cgsweb.wustl.edu/~fvallania/1_nature_methods_2009/NM_2009_LANE2_20071115.txt
http://cgsweb.wustl.edu/~fvallania/1_nature_methods_2009/NM_2009_LANE3_20071115.txt

=Lanes from November 20th 2007=

http://cgsweb.wustl.edu/~fvallania/1_nature_methods_2009/NM_2009_LANE6_20071120.txt
http://cgsweb.wustl.edu/~fvallania/1_nature_methods_2009/NM_2009_LANE7_20071120.txt
http://cgsweb.wustl.edu/~fvallania/1_nature_methods_2009/NM_2009_LANE8_20071120.txt

=Lanes from January 10th 2008=

http://cgsweb.wustl.edu/~fvallania/1_nature_methods_2009/NM_2009_LANE7_20080110.txt
http://cgsweb.wustl.edu/~fvallania/1_nature_methods_2009/NM_2009_LANE8_20080110.txt

=Lanes from February 15ht 2008=

http://cgsweb.wustl.edu/~fvallania/1_nature_methods_2009/NM_2009_LANE5_20080215.txt
http://cgsweb.wustl.edu/~fvallania/1_nature_methods_2009/NM_2009_LANE6_20080215.txt
http://cgsweb.wustl.edu/~fvallania/1_nature_methods_2009/NM_2009_LANE7_20080215.txt
http://cgsweb.wustl.edu/~fvallania/1_nature_methods_2009/NM_2009_LANE8_20080215.txt

The reference sequence file used for the analysis can be found at this link
http://cgsweb.wustl.edu/~fvallania/1_nature_methods_2009/1111_SEQS.txt

The reference sequence for the error model [pUC19 plasmid fragment] can be found at
http://cgsweb.wustl.edu/~fvallania/1_nature_methods_2009/pUC19.fa

1. Align read file to reference using the aligner program [we are using lane 7 from January 10th as an example] and save the output into a file [for example ALIGNMENT_LANE7_20080110]

   ```
   ./ALIGNER_64bit NM_2009_LANE7_20080110.txt 1111_SEQS.txt 2 36 > ALIGNMENT_LANE7_20080110
   ```

2. Tag the aligner output using the perl script [for example ALIGNMENT_LANE7_20080110] and add a tag to it [for example JAN102008]. Save the output into a new file.

   ```
   ./alignment_tagger.pl ALIGNMENT_LANE7_20080110 JAN102008 > TAGGED_LANE7_20080110
   ```

   NOTE!!! Lanes generated in the November 15th and 20th run are already labeled, therefore this step has to be skipped for them

3. At this point all the aligned files that have been tagged can be merged together in one big file using the cat command and redirecting the output into a new file

   [for example NM_2009_1111_POOL_ALL_ALIGNED_TAGGED_2MIS.txt]

   ```
   cat TAGGED_LANE7_20080110 .... > NM_2009_1111_POOL_ALL_ALIGNED_TAGGED_2MIS.txt
   ```

   this file has already been generated by us and can be downloaded from

   http://cgsweb.wustl.edu/~fvallania/1_nature_methods_2009/
   NM_2009_1111_POOL_ALL_ALIGNED_TAGGED_2MIS.txt

4. At this point a new error model file can be generated using the Perl script error_model_generator.pl and saving the output [for example EM_071115~080205_pUC19_2M_0_800.txt]

   ```
   ./error_model_generator.pl NM_2009_1111_POOL_ALL_ALIGNED_TAGGED_2MIS.txt pUC19.fa 2 N
   800 y > EM_071115~080205_pUC19_2M_0_800.txt
   ```

   NOTE!!! Only the first 800bp of pUC19 were used as the second half contains a base (position 877) that differs from our reference. This allows us to check if the settings produced false negatives on the remaining portion of the vector and also to detect the mutation (that will have frequency of 100

   The error model file can also be downloaded from

   http://cgsweb.wustl.edu/~fvallania/1_nature_methods_2009/EM_071115~080205_pUC19_2M_0_800.txt

5. Now all the files are ready to be analyzed by SNPSeeker

   ```
   ./SNPSeeker_64bit NM_2009_1111_POOL_ALL_ALIGNED_TAGGED_2MIS.txt 1111_SEQS.txt
   EM_071115~080205_pUC19_2M_0_800.txt 12 -1.301 > SNPSEEKER_OUTPUT
   ```

   The analysis is now complete!