Supporting Information

Wang et al. 10.1073/pnas.1114669108

SI Materials and Methods

Patients. For all primary cutaneous squamous cell carcinoma (cSCC) and basal cell carcinoma (BCC) samples P6–8, 14 men and 2 women were enrolled in the skin cancer study between 2006 and 2010, ranging from 61 to 87 y of age. Two patients were immunosuppressed following organ transplants and had histories of multiple nonmelanoma skin cancers. All subjects provided informed consent according to procedures approved by the University of California, San Francisco Committee on Human Research, including that for DNA sequencing and array-based genetic analysis. Consents enable sharing of information obtained from these studies with other scientists as long as patient identity is not shared. The diagnosis of cSCC was confirmed for all tumors via histological examination of a standard biopsy specimen by a board-certified dermatopathologist. Tumor samples were obtained by curettage before Mohs micrographic surgery. Paired control samples were obtained from peritumoral normal skin removed during reconstruction.

For BCCs P1–5, tumor tissue was removed by Mohs micrographic surgery, during which the central tumor mass was first debulked, and then the tumor margins were excised by examining serial frozen sections under the microscope. The debulked material, which should contain minimal contaminating normal tissue, was shown by histological examination to contain >50% tumor cells. Tumor tissue was snap-frozen in liquid nitrogen and stored at −80 °C. Blood samples (10 mL) were obtained from each source as a source for constitutional DNA. Clinical information on each tumor included location, size, histologic subtype, and whether the tumor was primary or recurrent and sporadic or hereditary. For BCCs P1–5, the study was approved by the Yale University School of Medicine Human Investigation Committee.

Patient information corresponding to SCCs of the lung has been described as part of the TCGA sequencing effort in the Database of Genotypes and Phenotypes (dbGAP).

Molecular Profiling. For cSCCs and BCCs P6–8, tumor and normal tissue were either snap frozen and stored on liquid nitrogen or stored in Ambion RNALater solution at −80 °C or snap frozen. DNA was extracted from tumor and control samples by using the QIAamp DNA Mini Kit or the Qiagen DNEasy Kit as per manufacturer's protocol. DNA quality was assessed by running samples on a 1% agarose gel and quantitated by using a NanoDrop 1000 Spectrophotometer. Quality of cDNA was confirmed by using the Agilent 2100 Bioanalyzer.

For BCCs P1–5, tumor tissue was pulverized in liquid nitrogen, then resuspended in 5 mL of TNE buffer (10 mM Tris, pH 8.0, 100 mM NaCl, and 25 mM EDTA) with 1 mg/mL proteinase K (Boehringer Mannheim) and 1% SDS, and incubated at 37 °C for at least 2 h. After two phenol-chloroform extractions, DNA was precipitated with the addition of 1/10 volume 3 M sodium acetate (pH 5.2) and 2 volumes of 100% ethanol and resuspended in TE buffer (10 mM Tris, pH 8.0, and 1 mM EDTA). Blood samples were suspended in a red blood cell lysis buffer (1.6 M NH_4Cl, 0.1 M $KHCO_3$, and 1 mM EDTA, pH 7.5) and centrifuged at 500 × *g* for 5 min at 4 °C. DNA was extracted from the leukocyte pellet by the guanidine-hydrochloride method (1). The quantity and quality of DNA were analyzed as described for the cSCC samples.

For exome sequencing (cSCCs P1–12, BCC P1–3), ∼1.0 μg of genomic DNA was from tumor, and normal tissue was sheared by sonication to a target length of 200 bp. About 40 megabases of coding sequence were targeted by using oligonucleotide-based hybrid capture using Agilent SureSelect Exome Capture kits (2). Sequencing-by-synthesis using the Illumina GAIIx or HiSeq2000 systems resulted in >85% of targeted regions receiving 14× fold coverage at >90% of bases. Validation sequencing used BigDyeTerminator Version 3.1 chemistry and was run on the Genetic Analyzer GA3730 platform, all from Applied Biosystems.

For transcriptome sequencing (BCC P4,5) RNA was isolated with mirVana miRNA Isolation Kit (Ambion), and poly(A) RNA was then enriched using MicroPoly(A)Purist Kit (Ambion). Five hundred nanograms of RNA was fragmented with RNA Fragmentation Reagent (Ambion) and then used to prepare a paired end library (Illumina PE kit) and sequenced on the Illumina GAII.

Mutation Calling for Exome Sequencing. For cSCCs, raw sequencing data in Illumina's fastq data format were converted into fastq files with base quality scores encoded in the Sanger basecall format. Next, the reads were aligned by using the BWA aligner developed at Sanger (3). This aligner is based on the Burrows–Wheeler transformation, aligns paired-end reads, and handles indels robustly. The output of BWA are the aligned reads in SAM format (currently standard file format for aligned sequence data). Reads stored in SAM format were then converted to the binary BAM format by using the samtools software (4). Once reads are in the sorted and indexed BAM file format, position-based retrieval of reads becomes fast and data storage requirements are minimized.

Next, to remove erroneous mutation calls due to PCR duplication, all duplicate reads were removed by using MarkDuplicates, an analysis tool included in the Picard software package developed by the Broad Institute (http://sourceforge.net/projects/picard/). After removal of the duplicate reads, the base quality scores were recalibrated by using the CountCovariates and TableRecalibration tools included in the GATK software, also developed by the Broad Institute (http://www.broadinstitute.org/gsa/wiki/index.php/Base_quality_score_recalibration#Introduction).

Mutations were called from raw Illumina sequencing reads using the muTect software package (https://confluence.broadinstitute.org/display/CGATools/MuTect), which in brief, consists of three steps:

1. Preprocessing aligned reads in tumor and normal sequencing data. This step ignores reads with too many mismatches or very low quality scores because these represent noisy reads that introduce more noise than signal.
2. Statistical analysis identifying sites that are likely to carry somatic mutations with high confidence. The statistical analysis predicts a somatic mutation by using two Bayesian classifiers—the first aims to detect whether the tumor is nonreference at a given site, and, for those sites that are found as nonreference, the second classifier makes sure the normal does not carry the variant allele. In practice, the classification is performed by calculating a LOD score (log odds) and comparing it to a cutoff determined by the log ratio of prior probabilities of the considered events. For each site in the tumor sample, we calculate

$$\mathrm{LOD}_T = \log_{10}\left(\frac{P(\text{observed data in tumor}|\text{site is mutated})}{P(\text{observed data in tumor}|\text{site is reference})}\right)$$

and for each site in the normal we calculate

$$\mathrm{LOD}_N = \log_{10}\left(\frac{P(\text{observed data in normal}|\text{site is reference})}{P(\text{observed data in normal}|\text{site is mutated})}\right)$$

A site is called mutated in the tumor if both the tumor sample is called different from the reference and the normal sample is called as equal to the reference (no tumor mutations are allowed at sites where the normal sample is not called identical to the reference). In other words, a site is called as mutated in the tumor sample if both

$$LOD_T > \theta_T \text{ and } LOD_N > \theta_N$$

for prespecified cutoffs θT and θN. Since we expect somatic mutations to occur at a rate of ?1 in a Mb, we set

$$\theta_T = \log_{10}(2x10^6) \approx 6.3$$

which guarantees that our false positive rate, due to noise in the tumor, is less than half of the somatic mutation rate. In the normal (not in dbSNP) sites, we require

$$\theta_N = \log_{10}(2x10^2) \approx 2.3$$

because non-dbSNP germ-line variants occur at a rate of ~100 in a Mb. This cutoff guarantees that the false positive somatic call rate, due to missing the variant in the normal, is also less than half the somatic mutation rate.

3. Postprocessing of candidate somatic mutations to eliminate artifacts of next-generation sequencing, short-read alignment, and hybrid capture. For example, sequence context can cause hallucinated alternate alleles but often only in a single direction. Therefore, we test that the alternate alleles supporting the mutations are observed in both directions.

As muTect attempts to call mutations, it also generates a coverage file [in a wiggle file format (5), which indicates for every base whether it is sufficiently covered in the tumor and normal to be sensitive enough to call mutations].

After muTect processing, raw mutation calls were filtered for acceptable coverage depth in the tumor and normal sample and annotated in detail, including gene name, affected amino acid, and COSMIC annotation if a mutation has been cataloged previously. All mutations known in dbSNP were subtracted unless present in COSMIC. In parallel with mutation calling, all known SNP positions in sequence data were interrogated in comparison with dbSNP130 to determine SNP alleles and their frequencies. Mutations were only called as present in this study when 14 independent reads were detected in the tumor and 10 reads in the normal sample.

For BCCs, P1–2 and matching normal tissue were sequenced on Illumina GAII. The sequences were aligned to hg18 by using Maq, and then Samtools was used to determine single nucleotide variations. An alignment and identification of indels was done using BWA, and then samtools was used to annotate indels. P3 and matching normal tissue was sequenced on Illumina HiSeq, and Eland was in the Illumina pipeline to align to hg18. BWA was used to identify indels. Samtools was used both to determine single-nucleotide variations and to annotate indels.

Study Oversight. The academic investigators were solely responsible for study design. The investigators collected samples and data and one investigator wrote a first draft of the manuscript. All investigators reviewed and approved the manuscript. All authors had full access to the data, contributed to the interpretation, and affirm both the accuracy of findings and adherence to the clinical protocol. The protocol was approved by the institutional review board at all study sites. All patients provided written consents before procedures specific to the study began.

1. Sambrook J, Fritsch E, Maniatis T (1989) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY), 2nd Ed.
2. Gnirke A, et al. (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* 27:182–189.
3. Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26:589–595.
4. Li H, et al.; 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079.
5. Rhead B, et al. (2010) The UCSC Genome Browser database: Update 2010. *Nucleic Acids Res* 38(Database issue):D613–D619.

Table S1. Identified amino acid substitution mutations in Notch receptors and pathway genes in noncutaneous carcinoma cell lines, primary lung SCCs sequenced by TCGA, and primary BCCs

Sample	N1	N2	N3	N4	RBJP	MAML1	MAML2	MAML3
TCGA lung SCCs, primaries								
21-1070-01		D1733Y						
33-4583-01	Splice site							
39-5031-01		Q1392*						
46-3769-01	I1440T							
66-2789-01	C429S, R353C							
Noncutaneous carcinomas, cell lines								
TT								
TE10	A683*							
SW900								
HCC95								
A549								
H549								
BCCs, cell lines								
BCC P1			P2036L			P585S		
BCC P2								
BCC P3								
BCC P4								
BCC P5								
BCC P6								
BCC P7								
BCC P8								

*denotes stop codons. Shaded grid denotes genes that were not analyzed for mutation. For TCGA data, only mutated samples are shown; 40 samples were analyzed on whole-exome level.

Table S2. Clinical characteristics of cSCCs sequenced for Notch pathway mutations

Sample	Sex	Age	Site	Immune status	TP53
cSCCs, primaries					
cSCC P1	M	76	Scalp	+	R248W
cSCC P2	M	84	Left dorsal hand	+	E224 (splice site)
cSCC P3					
cSCC P4	M	83	Left cheek	+	H179Y, P278S
cSCC P5	F	61	Left cheek	+	Y220N
cSCC P6	M	87	Scalp	+	E285K
cSCC P7	M	85	Right temple	+	P142N, H179Y
cSCC P8	M	58	Left helix	–	E286K, T329I, E349*
cSCC P9	M	59	Lip	+	Splice site
cSCC P10	F	80	Right vertex	–	R196*
cSCC P11	M	86	Left ear	+	G245D, Q104*
cSCC P12	M	66	Upper back	+	N.A.
cSCCs, cell lines					
SCC4	M	54	Floor of mouth	+	P151S
SCC12B	M	60	Face	–	V216G
SCC12F	M	60	Face	–	V216G
SCC25	M	74	Base of tongue	+	208FS
SCCRDEB2	M	54	Arm	+	V173L
SCCRDEB3	F	36	Left forearm	+	R273H
SCCRDEB4	F	32	Hand	+	P152L
SCCT1	M	61	Forearm	–	Y234S
SCCT2	M	66	Hand	–	P278F
SCCT3	M	55	Hand	–	V216M
SCCT8	M	67	Ear	–	Y91G
SCCIC1	M	77	Right temple	+	H179Y/p.R248Y
SCCIC8	F	51	Buttock	+	N.A.
SCCIC12	F	87	Left calf	+	N.A.

N.A., not applicable.

Table S3. PolyPhen analysis of identified *NOTCH1* and *NOTCH2* mutations

	Probably damaging		Probably benign	
Sample	*NOTCH1*	*NOTCH2*	*NOTCH1*	*NOTCH2*
cSCCs, primaries				
cSCC P1	Q610*	R1838*, R452C, W330*, P224L		
cSCC P2	C478F			
cSCC P3				
cSCC P4	W1768*	Q1634*, G313C		T2278I
cSCC P5				
cSCC P6	P1770S		R1594Q	
cSCC P7	Splice site	S1836F		E297K
cSCC P8	Q1923*	Q1616*, G488D		
cSCC P9	R353C			
cSCC P10	C423F			
cSCC P11	E1446*			N46S, E38K
cSCC P12				
cSCCs, cell lines				
SCC4				
SCC12B				
SCC12F			S137L	
SCC25				
SCCRDEB2				
SCCRDEB3	R353C			
SCCRDEB4	D1517N			
SCCT1		C861Y		
SCCT2	E415D, C409F, D469G	C433Y		G1751D, P2343S
SCCT3		D1451N		
SCCT8		C616F		
SCCIC1	N1809H	P1913S		
SCCIC8	Q1687*			
SCCIC12				

Mutations resulting in stop codons (*) or involving splice sites were not analyzed by PolyPhen and were categorized as probably damaging.