Supplementary Information for Berger *et al*., Melanoma genome sequencing reveals frequent *PREX2* mutations

I. List of Supplementary Figures and Tables

II. Supplementary Methods

A. Sample Attributes, DNA Preparation, and Quality Control

Description of the clinical cohort

All of the melanoma samples were collected under an IRB approved protocol, and informed consent was obtained from all subjects. The 25 melanoma samples that were subjected to Illumina sequencing came from metastatic tissue surgically excised and histopathologically reviewed at the Medical University of Vienna, as described previously[1,2]. Associated clinical information for these samples is provided in **Supplementary Tables S1**. The validation and extension panel that was subjected to capillary sequencing of *PREX2* consisted of an additional 45 metastatic melanomas from the same Vienna cohort as well as 62 patient-derived melanoma short term cultures obtained at the German Cancer Research Center (Deutsches Krebsforschungszentrum, DKFZ), as described previously[3]. Associated clinical information for these samples is provided in **Supplementary Table S10**. All normal DNA was obtained from patient matched blood samples. Genomic DNA from tissues was extracted using DNeasy Tissue Kit (Qiagen, Valencia, CA). Genomic DNA from short term cultures was extracted using the Puregene DNA purification kit (Gentra Systems, Minneapolis, MN).

Quality assessment of DNA and tumor

All DNA samples (tumor and germline) were evaluated for several quality criteria prior to Illumina sequencing. DNA concentration was measured using PicoGreen® dsDNA Quantitation Reagent (Invitrogen, Carlsbad, CA). Samples less than 60 ng/μl were concentrated by ethanol precipitation and re-suspension, as required for Illumina library construction. Structural integrity of DNA was monitored by gel electrophoresis, and degraded samples were removed from consideration. To ensure that each tumor and germline DNA pair was derived from a single individual, the identities all DNA samples were confirmed by mass spectrometric fingerprint genotyping of 24 common SNPs (Sequenom, San Diego, CA).

The 25 DNA pairs subjected to whole genome shotgun (WGS) sequencing were selected according to quantitative estimates of tumor purity and ploidy as inferred from SNP microarrays. Tumor DNA was hybridized to genome-wide human SNP microarrays (Affymetrix SNP Array 6.0) and analyzed as described previously[4]. We fit the observed allele-specific copy number levels to a model dependent on the tumor purity and average ploidy using a novel algorithm, ABSOLUTE (Carter S.L. *et al*., manuscript in press, *Nature Biotechnology*). We then calculated the "allelic index" for each tumor, indicative of the fraction of sequence reads expected to harbor the non-reference allele for a somatic mutation existing at a single copy per nucleus; we required an allelic index >0.1 for WGS. All melanomas subjected to WGS sequencing exhibited a purity of >20%, as indicated in **Supplementary Table 1**.

DNA sample pairs submitted for capillary sequencing of *PREX2* were prepared by whole genome amplification (WGA) of the native material and were also confirmed by fingerprint genotyping as described above.

B. Sequence Data Generation and Processing

We sequenced the complete genomes of 25 melanoma tumor/normal pairs according to the manufacturer's protocols (Illumina, San Diego, CA) and as described previously[5,6]. A summary is provided below.

Whole genome shotgun (WGS) and exon capture library construction

For whole genome shotgun (WGS) libraries, 3 μg of native DNA from each tumor and germline sample was sheared using the Covaris E210 instrument (Covaris, Woburn, MA) to a range of 100-700 base pairs. The resulting DNA fragments were end-repaired, phosphorylated, and adenylated at the 3' ends. Standard paired end adaptors were ligated to both ends, and fragments were purified by gel electrophoresis (4% agarose, 85 volts, 3 hours) and size selected by gel excision of two bands (500-520 bp and 520-540 bp). Purified fragments were enriched by PCR amplification (10 cycles). This produced two WGS libraries for each sample, with inserts averaging 380 bp and 400 bp, respectively. Qiagen Min-Elute columns were used for DNA purification and clean-up after each step.

Illumina sequencing

Sequence libraries were quantified by qPCR to determine the concentration of fragments with properly ligated adapter, and libraries were normalized to 2 nM. For ME009, ME012, ME015, ME032 and ME045, cluster amplification was performed according to Illumina's protocols using v2 chemistry and v2 flowcells. We performed paired-end sequencing on the Illumina Genome Analyzer II using v3 sequencing-by-synthesis kits and the Illumina v1.3.4 analysis pipeline. For the remaining 20 sequenced genomes, cluster amplification was performed according to Illumina's protocols using v3 chemistry and v3 flowcells. We performed paired-end sequencing on the Illumina HiSeq 2000 using v3 sequencing-by-synthesis kits and the HCS 1.1.37.8 and HCS 1.4.5 analysis pipelines.

WGS libraries were sequenced as 2 × 101 bp pairs, resulting in an average haploid coverage of 58x for each tumor genome and 32x for each germline genome. Tumors ME009, ME012, ME015, ME032, and ME045 were sequenced to an average of 30x haploid coverage, while the other 20 cases were sequenced to an average of 65x haploid coverage.

Data processing pipeline (Picard)

Pre-processing, alignment, and post-filtering of Illumina sequence data was performed using the "Picard" pipeline developed by the Broad Institute Sequencing Platform.

Individual tools in the Picard pipeline are available for download at http://picard.sourceforge.net/.

The output of Picard is a single BAM file[7] (http://samtools.sourceforge.net/SAM1.pdf) storing sequences, base quality scores, and the corresponding alignment information for all read pairs from a given sample. Read pairs were aligned to the human genome (hg19) using BWA. Base quality scores initially reported by the Illumina pipeline were recalibrated based on the read-cycle, the lane, the flow cell tile, the base in question, and the preceding base. (The original quality scores are kept in the BAM file in the OQ tag for each read.) This recalibration method was developed in collaboration with the Broad Institute's Medical and Population Genetics group as part of the Genome Analysis Tool Kit[8] (GATK; http://www.broadinstitute.org/gatk). Data from separate lanes and libraries are aggregated to a single sample-level BAM file, with lane and library identifiers captured in the read group tag and the BAM header. Artifactual molecular duplicates are identified based on the mapping position of read pairs and flagged to indicate artifacts of PCR amplification. BAM files may be loaded directly and viewed in the Integrative Genomics Viewer[9] (http://www.broadinstitute.org/igv).

C. Identification of Somatic Mutations

The Cancer Genome Analysis group at the Broad Institute has developed the "Firehose" pipeline to process the "flood" of tumor and normal sequence data emerging from massively parallel sequencing projects in cancer. This analysis includes (but is not limited to) data quality assessment, identification of somatic point mutations and indels, discovery of somatic structural rearrangements and breakpoints, and identification of genes significantly mutated across many tumors. These analyses are described below. Firehose manages the collection of analysis tools, their respective input and output files, and the overall workflow (Voet D. *et al*., unpublished). Firehose uses GenePattern[10] as its execution engine, ensuring control of parameters and versions and that the analysis results are reproducible. The following analyses were performed, as also described elsewhere[5,6].

Quality control and identity checks

Quality control checks were performed on the sequence data from each Illumina flow cell lane to (1) confirm sample identity and (2) prevent mix-ups between the tumor and normal samples for the same individual.

To confirm that the sequence data matched their corresponding patient, base calls were compared to independent genotypes obtained from either Affymetrix SNP 6.0 microarrays[11] or the genetic fingerprint at 24 common SNPs described above (Sequenom). For samples where SNP array experiments had been performed (including 24/25 WGS tumor/normal pairs), homozygous non-reference genotypes were compared to the observed bases at the corresponding genomic positions for each separate lane. Illumina lanes with <95% concordance were removed from the BAM files and excluded

from the analysis. If no SNP array experiment was performed, as was the case for certain exon capture samples, baits targeting the 24-SNP footprint were spiked into the hybrid selection reaction, and concordance between the Illumina and Sequenom data was established for these positions.

To identify possible mix-ups between the tumor and normal samples for the same patient, we determined the copy number profile of each lane using the depth of coverage in genomic windows. For sample pairs where SNP array experiments had been performed, we compared the tumor's microarray-based copy number profile to the sequence-derived copy number profile for each Illumina lane of tumor and normal DNA. Tumor lanes that did not match the expected profile, and normal lanes that deviated from the expected flat profile, were excluded. For all other sample pairs, we confirmed that each tumor lane harbored a greater number of copy number alterations than its corresponding normal lane.

For WGS samples, we performed an additional quality check based on the observed insert size distributions. Each sequencing library has a characteristic insert size distribution with a precisely defined mean and standard deviation. Illumina lanes whose insert size distribution did not match the corresponding distributions for the other lanes harboring the same sequencing library were excluded.

Local realignment around indels

Due to the presence of somatic and germline indels with respect to the reference genome (hg19), we performed a multiple sequence alignment for each set of reads in the vicinity of a putative indel site along the genome. Putative indel sites were denoted based on the presence of gaps and/or consecutive mismatches in individual reads. This "cleaning" process removed spurious mismatches at the ends of reads due to incorrect ungapped alignments, and it ensured that all reads mapping to the same locus harbored the same read structure. This analysis was performed for each tumor and normal BAM file using the local realignment module of the Genome Analysis Tool Kit (http://www.broadinstitute.org/gatk).

Identification of base pair substitutions

Following local realignment, somatic base pair substitutions were identified using a novel algorithm developed by the Cancer Genome Analysis group of the Broad Institute, called muTect. First, reads with low quality scores and/or many mismatches are eliminated in a pre-filtering step. Second, candidate somatic mutations are identified based on a confidence score indicating the presence of the variant in the tumor and absence in the normal sample. Third, candidate mutations are subject to a set of empirical filters, including a strand filter that requires that the orientations of reads harboring the variant allele and the orientations of all reads mapping to the locus exhibit similar distributions. Finally, mutations are annotated according to their genomic region (e.g., exon, intron, promoter, intergenic region), amino acid change, and protein change. All predicted somatic base pair substitutions from WGS sequencing are listed in **Supplementary Tables S3** and **S4**.

muTect utilizes a Bayesian statistical framework to compare the probabilities of generating the observed sequence data given underlying reference or non-reference genotypes. Two LOD scores (log odds) are calculated for each sample pair at a given position: one expressing the likelihood that the *tumor* is *non-reference* and the other expressing the likelihood that the *normal* is *reference*. (Each LOD score is calculated based on the local sequence coverage, observed allele counts, and base quality scores for the reads mapping to that genomic position.) The tumor LOD score and normal LOD score are compared to separate cutoffs reflecting the prior probabilities of false positives in the tumor (variants called somatic that are actually germline) and false negatives in the normal. The full details of the algorithm will be presented elsewhere (Cibulskis K. *et al*., manuscript in preparation).

Given the high mutation rates observed in these melanomas, it is impractical to validate all candidate mutations. However, we can infer a specificity of 95% for mutation calling based on independent validation results presented in several published and unpublished studies that utilized the same mutation calling strategy. Overall, 11,023 candidates have been tested at the Broad Institute, 10,434 of which have been validated[5,6,12-15].

We reviewed rejected muTect calls at base pairs coding for BRAF V600 and NRAS Q61, given the known recurrent mutations at these loci. Four tumor samples (ME007, ME029, ME037 and ME041) for which our algorithm did not detect BRAF codon 600 mutations nonetheless showed sufficient evidence for the mutant alleles in the tumor sample by manual inspection (these had been rejected due to insufficient coverage in the normal sample). The presence of somatic mutations resulting in BRAF V600E in ME007, ME029, ME037 and ME041 was also confirmed independently by whole exome sequencing (data not shown).

Identifying significantly mutated genes

Global mutation rates for the 25 melanomas subjected to WGS sequencing were calculated for "covered" base pairs (*i.e.*, at least 14 reads overlapped the position in the tumor and at least 8 reads overlapped the position in the normal). Genes with observed somatic mutation rates greater than expected by chance represent candidate driver genes in melanoma.

Significantly mutated genes were identified using the MutSig algorithm (Lawrence M.L. *et al.*, manuscript in preparation) based partly on methods published elsewhere[4,16]. The mutations observed in each gene were divided into categories of different mutation contexts: (1) C>T in context 5'[T]C[N]3', (2) C>T in context 5'[A/C/G]C[N]3', (3) A>G in context 5'[N]A[N]3', (4) any transversion and (5) any indel, nonsense or splice site mutation. For each gene, we calculated the probability of obtaining the observed set of nonsilent mutations (or a more extreme one) given the background mutation rates calculated per-tumor and given the observed number of synonymous mutations per-gene (as a measure of gene-specific background mutation rates), based on the algorithm as described previously[17]. P-values were converted to Q-values according to the Benjamini-

Hochberg procedure for controlling False Discovery Rate (FDR). Significantly mutated genes (Q<0.01) are displayed in **Table 1**.

Identification of short insertions and deletions

Following local realignment around putative indel sites (see above), candidate indels were predicted from the tumor BAM file based on the fraction of reads mapping to the locus that support the insertion or deletion. Candidates were discarded if the supporting reads exhibited low base quality scores and/or exceeded a threshold number of mismatches. Somatic and germline events were distinguished based on the presence of supporting reads in the normal BAM file. A version of this algorithm has been adopted into the Genome Analysis Tool Kit[8] (http://www.broadinstitute.org/gatk). In other projects, our method has exhibited a high and variable false positive rate (up to 40%) based on independent validation experiments (Sequenom). While consistent with other groups, this is not nearly as accurate as our method for detecting single base substitutions. However, visual inspection of candidate indels using the Integrative Genomics Viewer[9] (http://www.broadinstitute.org/igv) has been a highly successful means to eliminate the vast majority of false positive calls. Therefore, we manually reviewed all indels predicted within protein coding exons, and we report our high confidence candidates in **Supplementary Table S6**.

Identification of chromosomal rearrangements

Candidate chromosomal rearrangements were identified from the observation of multiple discordant read pairs using a novel algorithm, dRanger, developed by the Cancer Genome Analysis group of the Broad Institute (Lawrence M. *et al.*, manuscript in preparation). Discordant read pairs are defined as those that map to different chromosomes or on the same chromosome in different genomic positions (>600 bp apart, depending on the average insert size in the sequenced DNA library) or in improper orientations (*i.e.*, on opposite strands out of order or on the same strand). Based on the mapping positions and orientations of the supporting read pairs, events were annotated as interchromosomal, long range intrachromosomal (>1 megabase apart), inversions, deletions, or tandem duplications. Candidate rearrangements supported by clusters of discordant read pairs were removed if there were also any supporting read pairs in the corresponding matched normal and/or in a panel of additional normal genomes sequenced at the Broad Institute. A quality score from 0 to 1 was calculated for each candidate rearrangement based on: (1) the fraction of nearby reads with a mapping quality of zero; (2) the number and diversity of other discordant pairs near these breakpoints; (3) the standard deviation of the mapping positions of the supporting read pairs. Each candidate rearrangement was assigned an overall score equal to the number of supporting read pairs multiplied by this quality score. Events with an overall score greater than or equal to 4.0 were considered high confidence and are reported in **Supplementary Table S7**. The CIRCOS program was used to visualize intra- and interchromosomal rearrangements in **Figure 2a** (http://mkweb.bcgsc.ca/circos). Breakpoints were further categorized as intronic, exonic, or intergenic according to RefSeq gene annotations. Events with 2 intragenic breakpoints were annotated as to whether they were consistent with a gene fusion and whether it

would be in-frame or out-of-frame. A full description of the dRanger algorithm will be presented elsewhere (Lawrence M. *et al.*, manuscript in preparation).

Approximate locations of rearrangement breakpoints were determined based on the boundaries of the reads in discordant read pairs at each locus. When possible, breakpoints were mapped precisely to base pair resolution using BreakPointer (Drier Y. *et al.*, manuscript in preparation), a complementary algorithm also developed in the Cancer Genome Analysis group of the Broad Institute. Individual reads spanning the fusion point were identified from read pairs where one read mapped wholly on one side of the breakpoint and the mate pair was partly mapped or failed to align anywhere. These partly mapped and unmapped reads were subjected to a modified Smith-Waterman local alignment with the ability to jump between the two reference sequences. This procedure enabled the identification of the most probable fusion point. Using BreakPointer, we mapped the precise breakpoints for 73% of events.

We observed a dense clustering of rearrangement breakpoints near *ETV1* and *PREX2* in acral melanoma ME032 (**Figure 2b,c**). We manually reviewed the sequence data at these loci and detected additional candidate rearrangements that did not meet our criteria for high-confidence events in dRanger. Candidates that were validated by multiplex PCR and 454 sequencing (discussed below) are depicted in **Figure 2b and 2c**.

D. Experimental Validation of Structural Alterations

PCR and massively parallel sequencing of structural rearrangements

As discussed elsewhere[5], rearrangements predicted by dRanger were validated in a process involving PCR followed by massively parallel sequencing. PCR primers spanning each chimeric fusion junction were designed using Primer 3 (http://frodo.wi.mit.edu/primer3) to produce amplicons approximately 300–350 bp long. PCRs were performed on whole genome amplified DNA from both tumor and normal specimens. PCR products were pooled at normalized concentrations following quantitation by a NanoDrop spectrophotometer (Thermo Scientific). (For normal DNA products, we used the same volumes as calculated for the corresponding tumor DNA products.) Matching tumor and normal products were placed in separate pools. Sequencing libraries were prepared from each pool and sequenced in separate regions of a 454 Genome Sequencer FLX System (454 Life Sciences, Branford, CT). A rearrangement was judged to be somatic if the predicted chimeric product was detectable in tumor DNA and not normal DNA. Out of 371 predicted rearrangements tested, we confirmed 319 as somatic (86%). However, we estimate that the overall sensitivity of the PCR validation assay is only 85-90%. (In 24/209 cases where we designed two different primer pairs, only one successfully amplified the chimeric product.) Thus, the true accuracy of dRanger predictions is close to 100%.

Fluorescence in situ hybridization (FISH) for PREX2 amplification

The detection of *PREX2* rearrangements in archival formalin-fixed paraffin embedded sections (4-5 μm), as shown in **Figure 2d**, was performed by using a break-apart probe set that spans the entire PREX2 gene locus. The probe set consists from DNA of the following BAC clones: RP11-252K16, RP11-953G20, RP11-78H16, and RP11-463I8 labeled red (SpectrumOrange), RP11-984M8, RP11-66P19, and RP11-728C12 labeled green (SpectrumGreen) by nick-translation (Abbott Molecular). An interphase FISH analysis was performed as described previously[18] with a few modifications. Briefly, tissue sections were baked in a dry oven at 65°C for 30 minutes and subsequently deparaffinized. The sections were further processed through successive treatments in order to reduce auto-fluorescence: 70% alcohol/ammomia for 20 minutes followed by sodium borohydride for 40 minutes. Slides were then washed, tissue was treated with proteinase K, and DNA was denatured 80°C for 10 minutes. Pre-hybridization was performed at 37°C in 50% formamide / 2X SSC for 15 minutes followed by hybridization with 1 μg probe for 18 hours at 37°C in a humidified chamber in the dark. Post-hybridization washes were conducted at 42°C on a shaker in the dark as follow: 50% formamide / 2X SSC for 20 minutes, and 2X SSC and 0.5X SSC for 15 minutes each. After a brief wash with PBS, slides were counterstained with DAPI and mounted with Antifade (Abbott Molecular). Image data were acquired using an E8000 microscope (Nikon) and analyzed with FISHView software (Applied Spectral Imaging).

To verify the *PREX2* amplification as shown in **Supplementary Figure S5**, FISH was performed using a probe that spans the entire PREX2 gene locus, consisting of the following BAC clones: RP11-252K16, RP11-953G20, RP11-78H16, and RP11-463I8. The probe was directly labeled by Cy3 (Abbott Molecular) and co-hybridized with a chromosome 8 centromere probe (CEP 8, Abbott Molecular).

E. Capillary Sequencing of *PREX2* in an Extension Cohort

40 exons of *PREX2* were sequenced by PCR and bidirectional capillary sequencing (Beckman Coulter Genomics) in 131 melanomas: 24 melanomas that were subjected to Illumina sequencing, and 107 additional melanomas from 2 cohorts (described above). Whole genome amplified DNA product was used as input to sequencing reactions. SNPs with respect to the reference human genome were identified by two methods: PolyPhred (http://droog.mbt.washington.edu/poly_doc50.html) and Sequencher (http://www.genecodes.com/). Known SNPs present in the dbSNP database were filtered out. Putative insertion/deletion (indel) mutations were also identified using PolyPhred. We only considered variants that were called in reads in both directions. (High quality reads in both directions were available for 89% of amplicons.) Raw sequence traces were manually inspected for all novel candidate SNPs and indels. Matched normal DNA was sequenced for all candidates passing manual review to determine whether variants were somatic or germline in origin.

To estimate the sensitivity of the bidirectional capillary sequencing assay and analysis, we considered the 14 somatic mutations previously detected by Illumina sequencing of

native DNA. Only 9/14 mutations were detected by at least 1 of the 2 calling algorithms. The other 5 were clearly visible from manual inspection of the sequence traces but were missed by both methods. (Incidentally, these 5 mutations exhibited among the lowest allele frequencies: 14%, 14%, 21%, 22%, and 26%, underscoring the limited sensitivity of Sanger sequencing for low purity or heterogeneous tumors.) We identified 15 novel non-silent somatic *PREX2* mutations (14 single base substitutions and 1 frameshift insertion) in 15/107 additional samples (14%).

There are several reasons why a mutation frequency of 14% may be an underestimate: (1) only 9/14 mutations detected in the discovery samples by Illumina sequencing were called in the matched Sanger data (despite clear evidence for the remaining 5 from manual inspection of the sequence traces); (2) only a fraction of amplicons harbored the required 2 bidirectional Sanger reads for sequence analysis; (3) the extension samples exhibited lower tumor purity than the discovery samples (as evident in the relative peak heights in the Sanger sequence traces), making them more prone to false negatives. This may partly explain the discrepancy in the PREX2 mutation frequencies we observed in the discovery set and extension set.

F. Functional Studies of *PREX2*

Cell culture

Human primary melanocytes (pMEL/hTERT/CDK4(R24C)/p53DD) expressing either $NRAS^{G12D}$ were cultured, as previously described[2], in Ham's F10 medium (Cellgro) containing 10% heat inactivated fetal bovine serum (FBS) and 1% Penicillin/Streptomycin. 293T cells were grown in DMEM (Cellgro) containing 10% FBS. All cell lines were propagated at 37°C in a humidified atmosphere of 5% CO_2 and routinely tested for mycoplasma contamination.

Plasmids and lentiviral transduction

PREX2 cDNA, generously provided by R. Parsons, was subcloned into pLenti6.3-V5-DEST via Gateway recombination cloning (Invitrogen). *PREX2* mutations were generated using the Quickchange Site-Directed Mutagenesis Kit (Stratagene) according to the manufacturer's instructions. Lentiviral stocks were prepared by co-transfecting 293T cells with *PREX2* expression constructs and standard virus packaging systems. Viral supernatants were collected 48 and 72 hours post-transfection and subsequently used to generate stable pMEL-NRAS* cell lines.

Western immunoblotting

For immunoblotting, cells were lysed in NP-40 buffer (20 mM Tris-HCl, pH 8.0, 150 mM NaCl, 2 mM EDTA, 1% NP40) containing 1 mM PMSF, 1x Protease Inhibitor Cocktail (Roche) and 1x Phosphatase inhibitor (Calbiochem), separated on NuPAGE 4-12% Bis-Tris gels (Invitrogen), and blotted onto PVDF (Millipore). The following

antibodies were used following the manufacturer's recommendations: AKT (1:1000; Cell Signaling), phospho-AKT (Ser473; 1:1000; Cell Signaling), PTEN (1:1000; Cell Signaling), and V5 (1:5000; Invitrogen).

Xenograft studies

All animal studies were approved by Harvard Medical School Internal Animal Care and Use Committee (IACUC). pMEL-NRAS* cells expressing GFP, WT, or mutant PREX2 were mixed 1:1 with Matrigel (BD Bioscience) and subcutaneously implanted (1×10^6 cells) in female NCR-NUDE mice (Taconic). Animals were monitored thrice weekly until tumors were visible. Tumor volume was determined weekly by measuring in two directions with vernier calipers and formulated as tumor volume = (length × width^2) / 2. Animals were sacrificed when tumor volume approached 1.5 cm^3. Two-tailed t-test calculations were performed using Prism 5 (Graphpad).

G. Determination of Global Gene Expression Profiles

Gene expression microarray experiments were performed for 5 of the melanomas subjected to Ilumina WGS sequencing using Affymetrix Human Gene 1.0 ST arrays. Experiments were analyzed using Bioconductor (http://www.bioconductor.org/packages/devel/bioc/html/affy.html) and normalized using the RMA (robust multi-array) normalization procedure. For each sample, genes were ranked by expression level and binned into quintiles. Intronic and exonic C>T mutations were grouped by quintile and assigned to the transcribed or non-transcribed DNA strand of the associated gene according to RefSeq annotations (**Supplementary Fig. S1**). Normalized gene expression values are listed in **Supplementary Table S11**.

III. Supplementary Figures

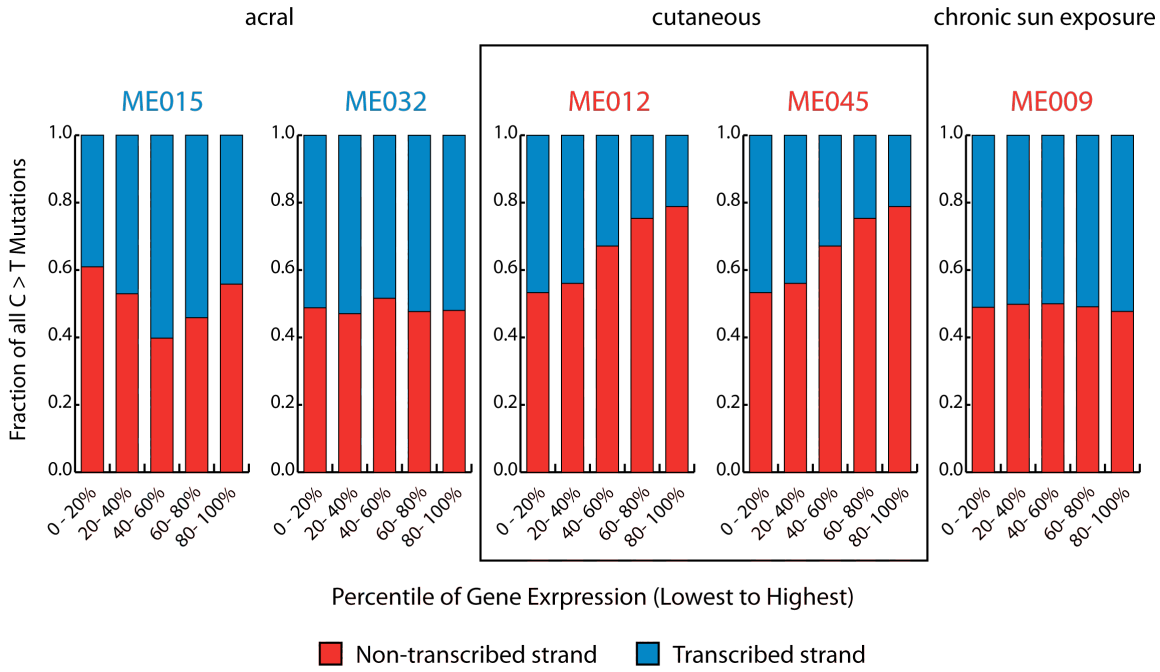


Figure S1: C>T strand bias in transcribed genes. All exonic and intronic C>T mutations identified from WGS sequencing were annotated as to whether they were observed on the transcribed or non-transcribed DNA strand. Genes were binned for each melanoma according to their expression levels determined by Affymetrix Human Gene 1.0 ST microarrays. Acral melanomas of glabrous skin are labeled in blue, and cutaneous melanomas of non-glabrous skin are labeled in red. For most cutaneous melanomas (exemplified by ME012 and ME045), there is a depletion of C>T mutations on the transcribed strand that is dependent on the gene expression level, supportive of transcription-coupled repair for UV-induced C>T transitions. Here, 74% of mutations occurred on the non-transcribed strand compared to 26% on the transcribed strand ($P < 10^{-30}$). No such bias was detected in the acral melanomas or the hypermutated sample ME009. Further, the overall mutation rate for ME009 was equivalent in transcribed and intergenic regions, whereas ME012 and ME045 exhibited an overall depletion of mutations in introns and exons (data not shown). Conceivably, the TCR mechanism itself may be deficient in some contexts, consistent with the presence of non-synonymous mutations in 3 genes in the TCR pathway in ME009 (*ERCC4*, *LIG3*, and *PARP1*). Alternatively, the mutation rates may be too low to permit detection of this bias (acral tumors), or they may supercede the repair capacity (ME009). In support of the latter, we observed that the overall mutation rate in transcribed regions was significantly lower than in intergenic regions for ME009 when C>T transitions were excluded (data not shown).

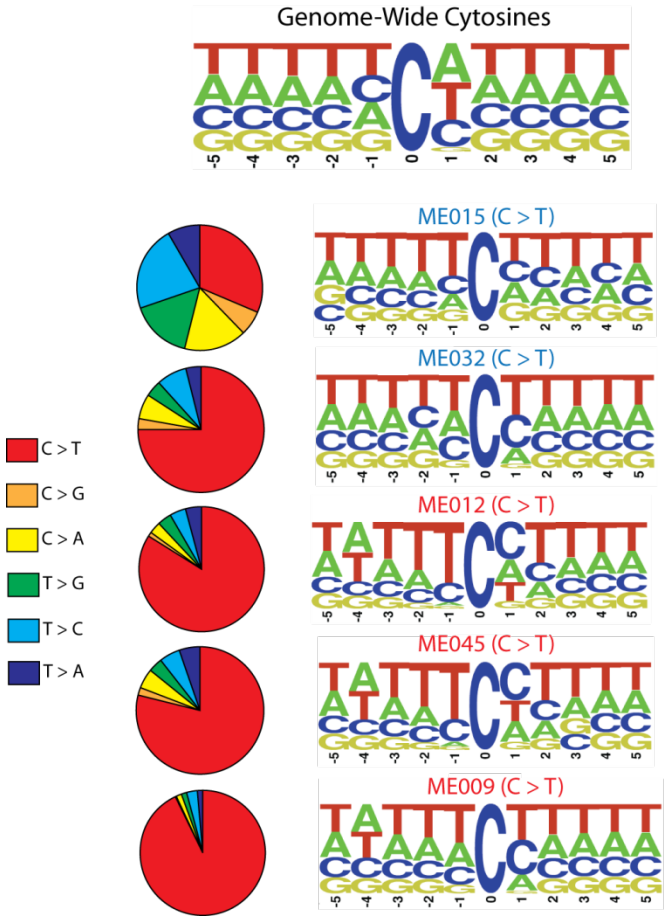


Figure S2: C>T mutation frequency and sequence context for representative cutaneous and acral genomes. Samples exhibit distinct patterns, as described in the text. The genome-wide sequence context of all cytosines is shown at the top. Logos were created using enoLOGOS (http://biodev.hgen.pitt.edu/enologos/).

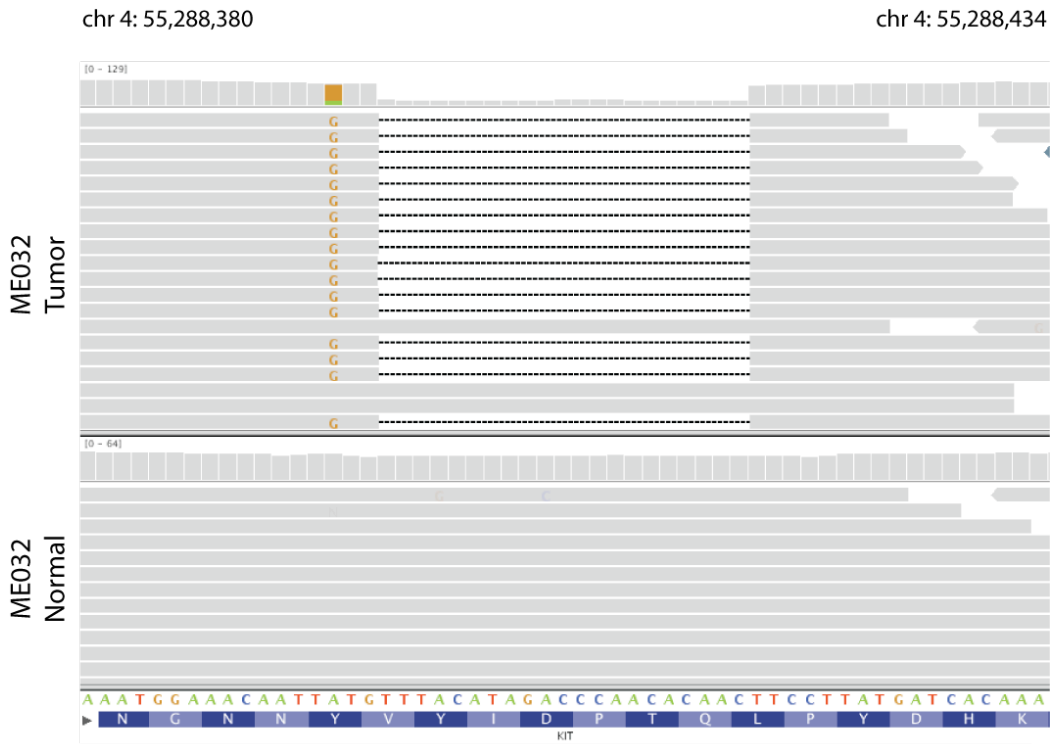


Figure S3: In-frame somatic deletion and point mutation in *KIT*. Melanoma ME032 harbors a 21 bp deletion, which is evident in the sequence reads from the tumor genome but not the matched normal genome. Sequence data are visualized using the Integrative Genomics Viewer[9]: http://www.broadinstitute.org/igv.

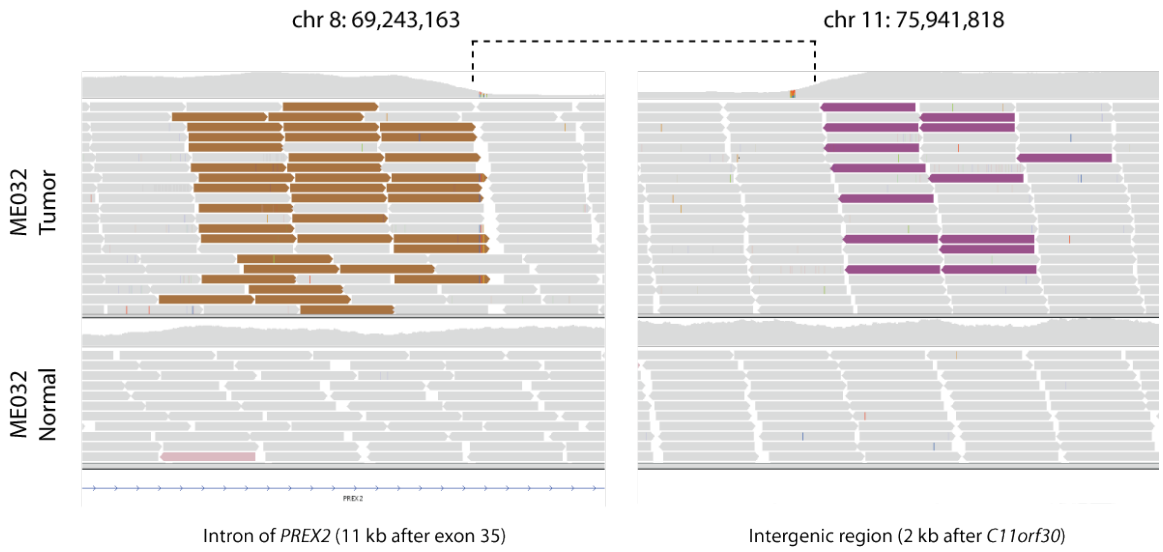


Figure S4: Interchromosomal translocation disrupting *PREX2* in melanoma ME032. A rearrangement joining chromosomes 8 and 11 is implicated by the presence of discordant read pairs in the tumor genome but not the normal genome (brown and purple bars), indicative of an intragenic breakpoint in *PREX2* and an intergenic breakpoint 2 kb downstream of *C11orf30*. This particular rearrangement is accompanied by chromosomal amplifications, as evidenced by the change in coverage apparent in the coverage track (top).

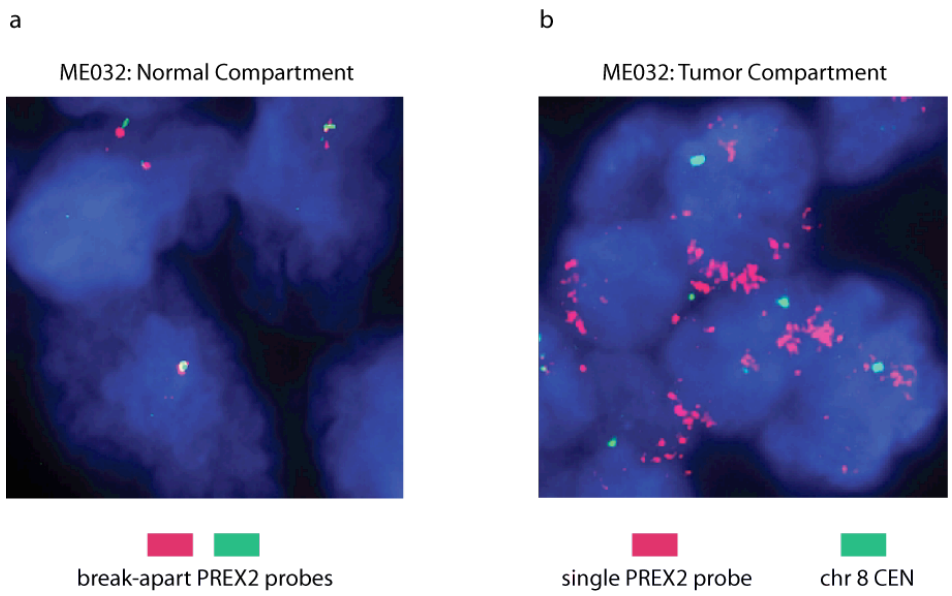
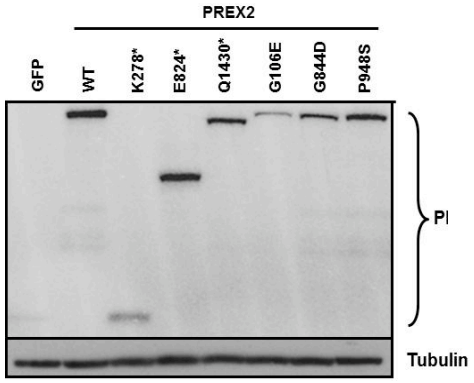


Figure S5: Amplification of PREX2 in melanoma ME032. (a) Normal stroma compartment FISH stained by dual-color break-apart PREX2 probe, as in Figure 2d. (b) Tumor compartment FISH, co-hybridized using a probe spanning entire the PREX2 gene locus (red) and the chromosome 8 centromere probe (CEP 8, Abbott Molecular, green), as described in Supplementary Methods.

a



b

PMEL-NRAS	Tumor Incidence *(n=10 mice)*	Comparison to Wt PREX2
GFP	3/10	*P=0.6*
WT	2/10	
K278*	**7/10**	***P=0.022***
E824*	**7/10**	***P=0.018***
Q1430*	**9/10**	***P=0.0007***
G844D	**8/10**	***P=0.006***
P948S	4/10	*P=0.344*
G106E	5/10	*P=0.168*

c

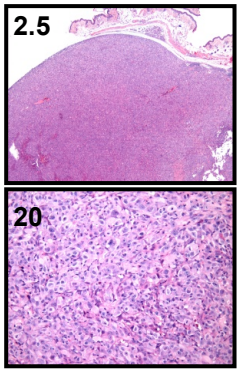


Figure S6: Ectopic expression of mutant PREX2 is tumorigenic in primary immortalized melanocytes. (a) pMEL-NRAS* expressing GFP (control), wild type (WT), truncated, or mutated PREX2 were grown in media containing 10% serum and lysates prepared and immunoblotted with the indicated antibodies. (b) Table showing tumor free survival of NUDE mice (n=10) injected with pMEL-NRAS* cells expressing GFP, WT, truncated, and mutated PREX2 subcutaneously. (c) Representative histological section of one PREX2-dependent xenograft (stained with hematoxylin and eosin).

IV. References

1. Kabbarah, O. et al. Integrative genome comparison of primary and metastatic melanomas. *PLoS One* **5**, e10770 (2010).
2. Garraway, L.A. et al. Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. *Nature* **436**, 117-22 (2005).
3. Ugurel, S. et al. B-RAF and N-RAS mutations are preserved during short time in vitro propagation and differentially impact prognosis. *PLoS One* **2**, e236 (2007).
4. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061-8 (2008).
5. Berger, M.F. et al. The genomic complexity of primary human prostate cancer. *Nature* **470**, 214-20 (2011).
6. Chapman, M.A. et al. Initial genome sequencing and analysis of multiple myeloma. *Nature* **471**, 467-72 (2011).
7. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-9 (2009).
8. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-303 (2010).
9. Robinson, J.T. et al. Integrative genomics viewer. *Nat Biotechnol* **29**, 24-6 (2011).
10. Reich, M. et al. GenePattern 2.0. *Nat Genet* **38**, 500-1 (2006).
11. Korn, J.M. et al. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet* **40**, 1253-60 (2008).
12. Wang, L. et al. SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. *N Engl J Med* **365**, 2497-506 (2011).
13. Stransky, N. et al. The mutational landscape of head and neck squamous cell carcinoma. *Science* **333**, 1157-60 (2011).
14. Bass, A.J. et al. Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion. *Nat Genet* **43**, 964-8 (2011).
15. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609-15 (2011).
16. Ding, L. et al. Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455**, 1069-75 (2008).
17. Getz, G. et al. Comment on "The consensus coding sequences of human breast and colorectal cancers". *Science* **317**, 1500 (2007).
18. Capodieci, P. et al. Gene expression profiling in single cells within tissue. *Nat Methods* **2**, 663-5 (2005).