Supplementary Figure 1. Mutation spectrum in sequenced cutaneous squamous cell cancers shows UV-radiation spectrum damage and active transcription-coupled repair. (A) Summative analysis of possible substitution events: transition events (G>A and C>T), indicative of UV radiation damage, comprise greater than 80% of detected substitutions. (B) C > T transition mutation frequencies are depleted by about 30% on the transcribed strand, confirming the persistence of transcription-coupled repair.

Supplementary Figure 2. Chromosomal duplications discretely increase mutation copy number. For representative samples, mutant allele frequency (*y*-axis) shown plotted against physical position on chromosome 17 (*x*-axis) for cutaneous squamous cell cancers. Frequencies of single nucleotide polymorphisms are displayed in grey scatter, heterozygous mutations in regions of CN-LOH as blue circles, and mutations in regions without CN-LOH as yellow X's. Homozygous mutations, as determined by allele frequencies greater than two standard deviations from the regional mean, are represented as red triangles. Estimated allele frequencies are shown as lines for heterozygous (solid) and homozygous (dotted) mutations (See Supplementary Methods and Appendix for details). A histogram of allele frequencies for each event is shown in the panel to right.

Supplementary Figure 3. Extrapolation of regional chromosomal findings to the genome. The genome was binned into approximately 0.18 MB regions of coding sequence, each represented by a single point. The number of mutations per bin was counted, taken over the width of the bin to yield the mutation rate (*y*-axis), and plotted against physical position (*x*-axis). A proportion of heterozygote mutations are lost with CN-LOH, greater for later events. CN-LOH 2, which was ordered as later than CN-LOH 17, shows a heterozygous mutation rate that is lower than median, in agreement with the claim that CN-LOH 2 occurs later in cancer development. CN-LOH 14 heterozygous mutation rate is the most dramatically low of the three. It is clear from the plot that its heterozygous mutation rate is significantly lower than average. (See Supplementary Methods and Appendix for details).

Supplementary Figure 4. Sequence read depth for approximately 27MB of nucleotides both targeted by exome exon capture and mapping to the Ensembl database, per sample. The *x*-axis plots read depth and the *y*-axis displays proportion of targeted bases; tb = % for tumor sample bases, nb = % for normal sample bases.

Patients

Seven men and one woman were enrolled in the skin cancer study between 2006 and 2010, ranging from 61 to 87 years of age. One patient was immunosuppressed following a lung transplant and had a history of multiple non-melanoma 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 sequence of all non-synonymous mutations from each sample will be deposited to dbGAP. The diagnosis of cutaneous squamous cell carcinoma was confirmed for all tumors via histological examination of a standard biopsy specimen by a board-certified dermatopathologist. Tumor samples were obtained by curettage prior to Mohs micrographic surgery. Paired control samples were obtained from peritumoral normal skin removed during reconstruction. Patient information corresponding to ovarian cancers analyzed here has been described previously(1).

Molecular profiling

Cutaneous squamous cell cancer and normal tissue were either snap-frozen and stored on liquid nitrogen or stored in Ambion RNALater solution (Austin, Texas, USA) at -80° C or snap frozen. DNA was extracted from tumor and control samples using the QIAamp DNA Mini Kit or the Qiagen DNEasy Kit (Valencia, California, USA) as per manufacturer's protocol. DNA quality was assessed by running samples on a 1% agarose gel and quantitated using a NanoDrop 1000 Spectrophotometer (Wilmington, Delaware, USA). Quality of cDNA was confirmed using the Agilent 2100 Bioanalyzer. Copy number analysis of DNA was performed using Affymetrix

Genome-Wide Human SNP Array 6.0 chips(2). All samples were processed on Affymetrix microarrays, fluidics stations, and scanner (Santa Clara, California, USA) according to the manufacturer's instructions. The pairs of SNP arrays were processed using the allele-specific CRMA v2 method(3) and segmented using circular binary segmentation as provided in the package by the function CbsModel. Then, TumorBoost(4) was applied to normalize the allele fractions obtained.

For sequencing, approximately 1.0 µg 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 using oligonucleotide-based hybrid capture using Agilent SureSelect Exome Capture kits(5). Sequencing-by-synthesis using the Illumina GAIIx or HiSeq2000 systems resulted in more than 85% of targeted regions receiving 14x fold coverage at >90% of bases. Validation sequencing utilized BigDyeTerminator Version 3.1 chemistry and was run on the Genetic Analyzer GA3730 platform, all from Applied Biosystems (Foster City, CA).

Mutation calling

Raw sequencing data in Illumina's fastq data format was converted into fastq files with base quality scores encoded in the Sanger basecall format. Next, the reads were aligned using the BWA aligner developed at Sanger(6). 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 using the samtools software(7). 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 are removed using MarkDuplicates, an analysis tool included in the Picard software package developed by the Broad Institute(8). After removal of the duplicate reads, the base quality scores are recalibrated using the CountCovariates and TableRecalibration tools included in the GATK software, also developed by the Broad Institute(9).

Mutations were called from raw Illumina sequencing reads using the muTect software package (Cibulskis K. *et al*., in preparation), 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 since 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 non-reference at a given site and, for those sites that are found as non-reference, 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

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

for prespecified cutoffs θ_T and θ_N. 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 $\theta_N = \log_{10}(2x10^2) \approx 2.3$ since non-dbSNP germline variants occur roughly 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. Post-processing 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(10), 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 catalogued previously. All mutations known in dbSNP are subtracted unless present in COSMIC. In parallel with mutation calling, all known SNP positions in sequence data were interrogated in comparison to 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. Mutations meeting these criteria are those referenced in the text and listed in Supplementary Table 2.

Sequence coverage and comments on mutational spectrum

On average, for bases targeted for capture aligning to the Ensembl database, more than 95% of bases in tumor sample were read at > 14x coverage, and more than 90% of bases at > 10x in the normal sample (Supplementary Figure 4). In total, the characteristics of base substitutions identified with at least 14 reads in tumor and 10 in normal, before thresholding for mutant allele frequency, was as follows:

Sample	Non-synonymous/	Mean mutant allele	STD mutant allele

	total substitutions	allele frequency	allele frequency
1	1027/1557	0.34380	0.07687
2	1297/1920	0.17530	0.04695
3	472/714	0.18710	0.06461
4	465/701	0.3200	0.07240
5	1224/1781	0.27120	0.06563
6	1116/1658	0.28290	0.11820
7	608/896	0.34470	0.11590
8	789/1177	0.4506	0.11990

All substitutions are listed in Supplementary Table 2. 86% of mutations were C > T transitions reflective of UV DNA damage (Supplementary Figure 1), indicating that the vast majority of called mutations are real. Further supporting specificity, of all non-synonymous base substitutions detected in *TP53* and *CDKN2A* across all samples, 13/13 were at positions previously catalogued in COSMIC; of the two synonymous mutations, both were C > T transitions and one occurred at a *TP53* splice site. PCR confirmation of 50 mutations with mutant allele frequency > 0.5 is described in Supplementary Table 1.

We detect, on average, a 15% relative depletion of C > T substitutions on the coding strand, signaling persistence of transcription-coupled repair(11). The Multiple Sequence Alignment realignment tool from the Broad Institute(12), which utilizes BAM files to locally realign reads, was employed to detect small insertions and deletions (indels). After screening against normal sample and requiring both 14 definitive, discrete reads in tumor and 10 reads in normal, between 10 and 26 indels were recognized per cSCC sample. Because of the ascertainment bias in indel calling, these mutations were not further confirmed and analyzed for reconstruction of genomic aberration history for the purposes of this study. All mutations sequenced in this study will be deposited in dbGAP.

Loss of wild-type p53 might herald such diverse forms of somatic aberration either by directly impairing DNA repair or replication efficiency (13), or reducing an apoptotic response to DNA damage. Because transcription-coupled repair depends on both dedicated machinery and global genomic repair activity(14), its detection indicates, the substantial UV mutagenesis in these cSCCs does not appear to result primarily from compromised DNA repair. Rather, loss of residual p53 is likely required for DNA-damaged cells to evade apoptosis at the G1 or S checkpoints and survive in renewing tissues.

Other than mutations in *NOTCH1* and *NOTCH2*, deleted alleles and potential biallelic mutation was also detected for *PKHD1*, a ciliary structural protein with loss of function in recessive polycystic kidney disease (15). Notably, 3/6 of these substitutions were observed at amino acids also mutated in the heritable disease(16).

Modeling effects of contamination with normal (non-tumor) tissue

For all plots of mutant allele frequency in regions of CN-LOH, a sample of pure tumor (containing no normal tissue) should show i) for SNPs, a shift to frequencies of either 0 or 1 and ii) for mutant alleles, a frequency of either 0.5 (occurring after duplication) or 1.0 (occurring before duplication). The presence of contamination of tumor with normal tissue, the level of which is sample-dependent, affects both SNP and mutant allele frequencies. For duplicated originally heterozygous SNPs (gray scatter in all plots), contamination lowers frequencies below 1.0. SNPs that were eliminated in the tumor are still visible from the contaminating cells, usually at a frequency between 0.1 and 0.25. For mutant allele frequences, normal contamination decreases their relative fraction for both homozygote and heterozygote mutations, effectively compressing their dynamic range on all plots.

The effect of normal contamination is indistinguishable from the effect of heterogeneity; as such, a model with both as parameters is unidentifiable. Instead, we estimated the combined effect of contamination and heterogeneity. As it turned out, in sample 1, the expected allele frequencies for LOH regions obtained under this model matched up well with the observed mean allele frequencies in those regions, with little difference between the regions. There did not appear to be an effect of regional heterogeneity, therefore a calculation of expected allele frequency offset using only average effect of contamination and heterogeneity was sufficient for our purposes.

To estimate the effect of heterogeneity and contamination, a Poisson model was used to model the read depth in regions without chromosomal aberration. It was assumed that read coverage at a particular allele was distributed $Poi(\lambda_{reg}(p^{i}_{ref,T} + p^{i}_{mut,T} + \mu_{NHet} + \varepsilon^{i}_{NHet}))$, where λ_{reg} is the average number of reads in a region (region endpoints obtained from sequencing technology documentation); $p^{i}_{ref,T}$ and $p^{i}_{mut,T}$, allele-specific proportions for seeing the non-mutated base or the mutated base, respectively; and $\mu_{NHet} + \varepsilon^{i}_{NHet}$, average combined effect of normal contamination and heterogeneity, with an allele-specific additive error term.

Restricting attention to normal regions simplified the problem so that the ratio of $p^i_{ref,T}$ to $p^i_{mut,T}$ would be the same for all alleles considered, allowing us to impose the constraint that our estimates for $p^i_{mut,T}$ and $p^i_{ref,T}$ must be equal. This yields the intuitive estimate $\mu_{NHet} = 1 - \frac{2}{N}\sum_{i=1}^{N}\frac{x_i}{n_i}$, where x_i and n_i are the number of reads showing the mutation and the total number of reads at allele *i*, respectively, and *N* is the total number of alleles sequenced in regions without chromosomal aberration. Essentially, μ_{NHet} is the difference between 0.5 (the expected allele frequency without normal contamination or heterogeneity) and the mean observed allele frequency. For cSCC sample #1, the MLE estimate of combined contamination and heterogeneity was 0.30.

Given an estimate of μ_{NHet}, the estimated homozygous allele frequency is $1-\mu_{NHet}$ (represented as a dotted line in graphs), and the estimated heterozygous allele frequency is $(1-\mu_{NHet})/2$ (represented as a solid line in graphs). Not surprisingly, simple substitution into the equation above shows that the estimated heterozygous allele frequency is the mean allele frequency in regions without chromosomal aberration; the estimated homozygous allele frequency is twice this mean allele frequency.

Model for reconstruction of genomic aberration history

For all chromosomal regions undergoing aberrant duplication, including copy neutral loss-of-heterozygosity (CN-LOH)(17) and simple copy gains, those mutations occurring before duplication are doubled in copy number. Those mutations occurring after duplication are observed in their original heterozygote state. By comparing mutant allele frequencies, this principle can be used, in aberrant regions, to distinguish earlier and later mutations in the evolution of a cancer(11). Not all chromosomal abnormalities are informative. For example, mutations in regions of simple copy loss, resulting in a monoploid state, cannot be evaluated in this manner. It is also formally possible that chromosomal loss and gain do not occur simultaneously in CN-LOH. However, the losses and gains we observe are physically matched, making it far more likely these events were the event of a crossing-over event: a somatic analogue of uniparental disomy, as is currently believed(17). Moreover, in cSCCs did we observe a unbalanced gain or loss at the most common site of CN-LOH, 17p, that might suggest temporal dissociation.

Regional CN-LOH and copy gain events were first determined conservatively by evaluating the processed SNP array data using aroma.affymetrix(18) for raw copy number depth at baseline relative to other chromosomes in a given sample. These events were visually confirmed based on their bimodal SNP frequencies and agreement with theoretical estimates along the region both by SNP array and whole exome analysis. For regions determined to harbor CN-LOH, the ratio of sequenced mutant allele to total independently sequenced alleles was determined for all mutations with more than 50 independent reads. Only frequencies meeting this threshold are displayed in figures and include synonymous and non-coding mutations. Mutant allele frequencies two or more standard deviations from the mean average frequency of heterozygote mutations on neighboring normal diploid regions were called as homozygote. This method identified a total of 177 homozygous and 311 heterozygous mutations in the 21 CN-LOH regions in the 8 samples (Supplementary Table 1).

The very large mutation rate in cSCCs enables these analyses on the whole exome level; identifying enough mutations to achieve comparable resolution many other solid cancers requires nearly 50-fold greater (whole genome) sequence coverage.

Ordering timing of CN-LOH events

Given the number of mutations we see in the CN-LOH region, we wish to estimate the true fraction of events that are early in a way that takes into account read depth, which is a measure of our confidence in the allele frequency (low read depth alleles are trusted less than high read depth alleles). If n reads cover a particular allele, the number of reads showing mutation can be described by a mixture model, where with probability π, the true chance of an early mutation, the reads showing mutation are distributed $Bin(n, p_{early})$, and with probability $1-\pi$, they are distributed $Bin(n, p_{late})$. Here, p_{early} and p_{late} are the expected allele frequencies for early and late mutations in an LOH region, respectively, corrected for the offset by average contamination and heterogeneity.

For example, without contamination or heterogeneity, p_{early} should be 1, but with 20% heterogeneity and contamination, p_{early} is 0.80. Using expectation-maximization, we arrive at three different estimates $\hat{\pi}$, one per CN-LOH region. Bootstrapping the distribution of $\hat{\pi}$ yields the 95% confidence interval given in the paper: for each CN-LOH region, 1000 vectors of $Ber(\hat{\pi})$ random deviates were generated, and $\hat{\pi}$ recalculated for each random deviate. This modified binomial distribution function distinguishes the proportion of heterozygote and homozygote mutant frequencies between chromosome 17 (0.00-0.12), chromosome 2 (0.17-0.45), and 14 (0.65-0.94) at $p < 0.05$. The estimates $\hat{\pi}$ for chromosome 17, 2, and 14 are 0.04, 0.31, and 0.80, respectively. This yields homozygous:heterozygous ratios of 0.04, 0.45, and 4, respectively.

Furthermore, the method provides the probability that a mutation is early, given our estimate $\widehat{\pi}$. Those with conditional probability at least 0.5 were called as early, otherwise they were called as late. In the plots of Figure 2 and Supplementary Figure 2, the point size for mutations in the CN-LOH region is proportional to the conditional probability (the closer the conditional probability is to 0.5, the smaller the point-size). In most cases, the early and late allele frequencies were well-separated, so the conditional probabilities were close to either 0 or 1 and there are no discernable differences in point size.

The ovarian sample of Figure 1 (panel D) was called using a different method because fewer data points were available. In these cases, the mean and standard deviation (SD) of heterozygous allele frequency in neighboring regions without chromosomal aberration were computed. Mutations above 2 SDs of the mean were called as early.

When the calls from this method were compared to those of the binomial mixture model for the three CN-LOH events in cSCC #1, they were found to agree completely.

Ordering timing of chromosome 9/11 copy gain events

In sample 1, for both chromosomes 9 and 11, raw copy depth analysis indicates a copy gain has occurred relative to the predominant diploid state of these samples. In chromosome 9, the highest SNP allele frequencies are consistently comparable to those observed in CN-LOH regions on chromosome 2, 14, and 17, which requires loss of one of the original two chromosomal copies (single gain without loss of wild-type would generate maximum allele frequencies at approximately 0.5). The most parsimonious interpretation is gain of one chromosome, twice, and loss of the complementary copy.

The mutation allele frequencies of the three peaks observed on chromosome 9 are phase-shifted relative to peaks observed in the CN-LOH events, instead consistent with three chromosomes each harboring independent mutant allele frequencies. A triploid mutation frequency (centered at ~0.8) can only occur for mutations that were duplicated prior to the initial of two gains. Diploid mutation frequencies (centered at about ~0.55) can only occur for mutations that occurred after the first gain but prior to the second. Therefore these frequencies can be used to estimate occurrence of both copy gain in tumor evolution, although not loss. The mutant allele frequencies for chromosome 11 demonstrate consistency with only the haploid and diploid peaks in chromosome 9, indicating a simple copy gain that has given rise to homozygous mutations, but none in triploid copy-number.

Extrapolation of regional chromosomal findings to the genome

Sample 1 was used to model relative mutation levels across the exome (Supplementary Figure 3). For this analysis the genome was binned into regions of approximately 0.18 MB of coding sequence; in other words, each bin contains several concatenated exonic regions within one chromosome, and the width of the bin is the sum of the widths of these exonic regions. This corresponded to bins with median genomic width of between 6.1 MB (25th percentile) to 18.4 MB (75th percentile). The number of mutations per bin was counted, and taken over the width of the bin to yield the mutation rate (Supplementary Figure 3, *y*-axis). The number of mutations per bin were in the range of 5 (25th percentile) to 10 (75th percentile); similarly, most mutation rates fell within the rate 3.2 x 10^{-5} (25th percentile) and 5.62 x 10^{-5} (75th percentile).

Outside of the CN-LOH events shown, only regions without chromosomal aberration were plotted; hence the lack of points from chromosome 7, 9, or 11, which showed chromosome-wide deletions/gains. Therefore, the *x*-axis on Supplementary Figure 3 is a rough measure of the position of the bins: apart from the omission of several regions with aberration, it was mentioned earlier as well that the genomic width of each bin varies greatly, though the amount of coding sequence represented by each bin is consistent.

The figure shows both the overall mutation rate for the CN-LOH regions of interest as well as the rate for only the heterozygous mutations. It is presumed that CN-LOH that occurs late in cancer development would have a heterozygous mutation rate that is less than the normal mutation rate, since less time was allowed for the accumulation of mutations post-CN-LOH.

As shown in Figure 2, we conclude CN-LOH 17 occurs early in cancer evolution; CN-LOH 2 occurs later, and CN-LOH 14 occurs even later. In concordance with this idea, the plot shows that the heterozygous mutation rate for CN-LOH 17 is comparable to that for normal regions, suggesting that CN-LOH 17 must have occurred early enough to accumulate as many mutations as regions without aberration. CN-LOH 2, which was ordered as later than CN-LOH 17, shows a heterozygous mutation rate that is lower than median - but not strikingly low - in agreement with the claim that CN-LOH 2 occurs later in cancer development, but not nearly as late as CN-LOH 14. CN-LOH 14 heterozygous mutation rate is the most dramatically low of the three. It is clear from the plot that its heterozygous mutation rate is significantly lower than average.

An initial pass at this analysis sought to eliminate any effect of GC content on the mutation rate. With bins of this size, it was found that there was minimal correlation between GC content and mutation rate, probably because of the bin size, and the computation was therefore not GC-adjusted.

Study oversight

The study was supported by the National Institutes of Health, the U.S. Departments of Energy and Defense, the Stand Up To Cancer-American Association of Cancer Research Dream Team Translational Cancer Research Grant, and the Samsung Advanced Institute of Technology. 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.

References

1. http://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp.

2. Lindblad-Toh K, Tanenbaum DM, Daly MJ, Winchester E, Lui WO, Villapakkam A, et al. Loss-of-heterozygosity analysis of small-cell lung carcinomas using single-nucleotide polymorphism arrays. Nat. Biotechnol. 2000 Sep;18(9):1001-1005.

3. Bengtsson H, Wirapati P, Speed TP. A single-array preprocessing method for estimating full-resolution raw copy numbers from all Affymetrix genotyping arrays including GenomeWideSNP 5 & 6. Bioinformatics. 2009 Sep 1;25(17):2149-2156.

4. Bengtsson H, Neuvial P, Speed TP. TumorBoost: normalization of allele-specific tumor copy numbers from a single pair of tumor-normal genotyping microarrays. BMC Bioinformatics. 2010;11:245.

5. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat. Biotechnol. 2009 Feb;27(2):182-189.

6. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010 Mar 1;26(5):589-595.

7. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078-2079.

8. http://sourceforge.net/projects/picard/.

9. http://www.broadinstitute.org/gsa/wiki/index.php/Base_quality_score_recalibration#Introduction.

10. Rhead B, Karolchik D, Kuhn RM, Hinrichs AS, Zweig AS, Fujita PA, et al. The UCSC Genome Browser database: update 2010. Nucleic Acids Res. 2010 Jan;38(Database issue):D613-619.

11. Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, et al. A comprehensive catalogue of somatic mutations from a human cancer genome. Nature. 2010 Jan 14;463(7278):191-196.

12. http://www.broadinstitute.org/gsa/wiki/index.php/Local_realignment_around_indels.

13. Ford JM, Hanawalt PC. Li-Fraumeni syndrome fibroblasts homozygous for p53 mutations are deficient in global DNA repair but exhibit normal transcription-coupled repair and enhanced UV resistance. Proc. Natl. Acad. Sci. U.S.A. 1995 Sep 12;92(19):8876-8880.

14. Hanawalt PC, Spivak G. Transcription-coupled DNA repair: two decades of progress and surprises. Nat. Rev. Mol. Cell Biol. 2008 Dec;9(12):958-970.

15. Harris PC, Torres VE. Polycystic kidney disease. Annu. Rev. Med. 2009;60:321-337.

16. http://www.humgen.rwth-aachen.de/index.php?page=database.

17. O'Keefe C, McDevitt MA, Maciejewski JP. Copy neutral loss of heterozygosity: a novel

chromosomal lesion in myeloid malignancies. Blood. 2010 Apr 8;115(14):2731-2739.

18. Bengtsson H, Simpson K, Bullard J, Hansen K. aroma.affymetrix: A generic framework in R for analyzing small to very large Affymetrix data sets in bounded memory. Berkeley, CA: Department of Statistics, University of California, Berkeley; 2008.