www.sciencemag.org/cgi/content/full/science.1198056/DC1

Supporting Online Material for

The Genetic Landscape of the Childhood Cancer Medulloblastoma

D. Williams Parsons, Meng Li, Xiaosong Zhang, Siân Jones, Rebecca J. Leary,
Jimmy Cheng-Ho Lin, Simina M. Boca, Hannah Carter, Josue Samayoa,
Chetan Bettegowda, Gary L. Gallia, George I. Jallo, Zev A. Binder, Yuri Nikolsky,
James Hartigan, Doug R. Smith, Daniela S. Gerhard, Daniel W. Fults,
Scott VandenBerg, Mitchel S. Berger, Suely Kazue Nagahashi Marie,
Sueli Mieko Oba Shinjo, Carlos Clara, Peter C. Phillips, Jane E. Minturn,
Jaclyn A. Biegel, Alexander R. Judkins, Adam C. Resnick, Phillip B. Storm,
Tom Curran, Yiping He, B. Ahmed Rasheed, Henry S. Friedman, Stephen T. Keir,
Roger McLendon, Paul A. Northcott, Michael D. Taylor, Peter C. Burger,
Gregory J. Riggins, Rachel Karchin, Giovanni Parmigiani, Darell D. Bigner, Hai Yan,
Nick Papadopoulos, Bert Vogelstein,* Kenneth W. Kinzler,* Victor E. Velculescu*

*To whom correspondence should be addressed. E-mail: velculescu@jhmi.edu (V.E.V.);
kinzlke@jhmi.edu (K.W.K.); vogelbe@gmail.com (B.V.)

Published 16 December 2010 on *Science* Express
DOI: 10.1126/science.1198056

This PDF file includes:

Materials and Methods
References

Other Supporting Online Material for this manuscript includes the following:
(available at www.sciencemag.org/cgi/content/full/science.1198056/DC1)

Tables S1 to S10

Supporting Online Material

Materials and Methods

Medulloblastoma (MB) DNA samples

Tumor DNA was obtained from MB xenografts, cell lines, and primary tumors, as previously described (*1*). The Discovery Screen consisted of 22 tumor samples (17 primary tumors, 4 xenografts, and 1 cell line), with the Prevalence Screen including another 66 primary tumors. Clinical data regarding Discovery Screen and Prevalence Screen samples are available in Table S2, and summarized in Table S1. All samples had been given a diagnosis of MB (WHO grade IV) by institutional report. All samples with available hematoxylin and eosin-stained (H+E) slides or available tissue blocks from which new H+E slides could be produced were subjected to central review by a pediatric neuropathologist (PB) (scanned images of H+E slides of samples are available at http://cgap.nci.nih.gov/Data_Access). For each slide the percentage of tumor cells present was estimated, and the MBs were subclassified as large cell/anaplastic (LCA), nodular/desmoplastic (ND), or classic, non-nodular (C) when possible. All tumor samples were obtained at the time of the original surgery except one Discovery Screen sample (MB106X) and 6 Prevalence Screen samples (MB107PT, MB116PT, MB157PT, MB211PT, MB230PT, MB239PT), which were obtained at the time of MB recurrence. One sample (MB122PT) was obtained from a patient with Li-Fraumeni syndrome (germline mutation of *TP53*).

Identification of Transcripts for Sequence Analysis

Protein encoding transcripts were derived from three sources. The majority of protein encoding transcripts (46,482) were derived from the 61,043 transcripts present in the Ensembl database downloaded from the UCSC Genome Bioinformatics site (ensGene.txt, File Date 8/27/2008). The Ensembl transcripts were then compared to 20,025 transcripts present in the CCDS database downloaded from the UCSC Genome Bioinformatics Site (ccdsGene.txt, File Date 2/2/2009). This comparison identified 132 protein encoding transcripts not represented in Ensembl which were added to the list of transcripts to be considered for sequencing. The above 46,614 protein encoding transcripts were then compared to 29,996 transcripts present in the RefSeq database downloaded from the UCSC Genome Bioinformatics Site (refGene.txt, File Date 1/18/2009). This analysis identified a further 4,407 protein encoding transcripts that were unique to RefSeq bringing the total number of transcripts under consideration to 51,021. 446 Ensembl derived transcripts were eliminated because they lacked uninterrupted open reading frames. Finally, 1,099 transcripts that mapped to the mitochondrial genome, chromosome Y or alternate haplotypes were eliminated bringing the total number of protein encoding transcripts targeted for sequencing to 49,476.

The protein encoding transcripts were supplemented with microRNA (miRNA) transcripts. Coordinates for 718 miRNAs were downloaded from the Sanger miRBase Sequence Database (Release 13.0) and 715 were added to the list of transcripts targeted for sequencing after excluding 3 miRNAs mapped to the mitochondrial genome. This addition brought the total number of transcripts targeted for sequencing to 50,191. The combined set of transcripts represented 24,893 genes (24,178 protein encoding and 715 miRNA) and comprised 226,467 unique exons (225,752 protein encoding and 715 miRNA) covering 36,909,796 bases. For the purposes of considering protein encoding genes, transcripts were grouped into genes using their Ensembl gene names. CCDS and RefSeq transcripts not present in Ensembl were assumed to represent distinct genes and were designated with their transcript names. For miRNA, each distinct transcript was assumed to represent a different gene.

Primer Design and Sequence Analysis

A total of 36,909,796 bases were identified within the regions of interest (ROIs) of the 50,191 targeted transcripts. The ROIs comprised the entire transcribed portion of the 715 miRNA exons and the protein encoding portion plus 4 bases of flanking sequence for the 225,752 protein encoding exons. For clarity, the 4 bases of flanking sequence for the protein encoding exons would thus encompass sequences upstream of the start codon, downstream of the stop codon, and splice acceptors and splice donors. A total of 228,907 primer pairs were designed that could amplify 35,190,701 (95.3%) bases of the ROIs (table S1). These primer pairs were then used to amplify and sequence DNA from the 22 medulloblastoma samples and one normal sample as previously described (*2,3*). The vast majority of these primers (219,532; 95.9%) yielded PCR products and high quality sequencing results in 18 or more of the 23 samples sequenced. A total of 735,126,675 bases were evaluated for mutations in the 22 medulloblastomas (average of 31,962,029 bases per sample, range 28,031,708 to 32,395,730) (sequence trace data are available at http://cgap.nci.nih.gov/Data_Access). Of the evaluated bases, 99.3% had a Phred score of 20 or more and 97.9% had a score of 30 or more. All coordinates listed in the Supplementary Tables correspond to the human reference genome hg18 release (NCBI 36.1, March 2006).

The sequencing data were analyzed using Mutation Surveyor (SoftGenetics, State College, PA) coupled to a relational database (Microsoft SQL Server). Following automated and manual curation of the sequence traces, regions containing potential single base mutations and small insertions and deletions (indels) not present in the reference genome or single nucleotide polymorphism (SNP) databases (dbSNP release 125 variants that had been validated by the HapMap project) were re-amplified in both the tumor and matched normal tissue DNA and analyzed either through sequencing by synthesis on an Illumina GAII instrument or by conventional Sanger sequencing. This process allowed us to confirm the presence of the mutation in the tumor sample and determine whether the alteration was somatic (i.e. tumor-specific). BLAT and In Silico PCR (http://genome.ucsc.edu/cgi-bin/hgPcr) were used to perform homology searches in the human and mouse genomes and to remove variants present in related genomic regions. Additionally, mutations identified in the xenografts were confirmed to be present in the corresponding primary tumors at this stage of the analysis.

We further evaluated a set of 15 mutated genes that were mutated twice or more in the Discovery Screen samples (either by two sequence alterations or a sequence and copy number alteration) or were mutated once in the Discovery Screen and had previously been reported to be mutated in other tumor types in a second (Prevalence) screen, which included an additional 67 MBs (table S2). The primers used (table S1) and methods of analysis and curation of potential mutations were the same as described for the Discovery Screen.

Copy Number Alterations

The Illumina Infinium II Whole Genome Genotyping Assay employing the BeadChip platform was used to analyze tumor samples at 1,199,187 (1M-Duo) SNP loci. All SNP positions were based on the hg18 (NCBI Build 36, March 2006) version of the human genome reference sequence. The genotyping assay begins with hybridization to a 50 nucleotide oligo, followed by a two-color fluorescent single base extension. Fluorescence intensity image files were processed using Illumina BeadStation software to provide normalized intensity values (R) for each SNP position. For each SNP, the normalized experimental intensity value (R) was compared to the intensity values for that SNP from a training set of normal samples and represented as a ratio (called the "Log R Ratio") of log2(Rexperimental/Rtraining set) (Illumina copy number data are available at http://cgap.nci.nih.gov/Data_Access).

The SNP array data were analyzed using modifications of a previously described method (*4*). Homozygous deletions (HDs) were defined as two or more consecutive SNPs with a Log R Ratio value of ≤ -2. The first and last SNPs of the HD region were considered to be the boundaries of the alteration for subsequent analyses. To eliminate chip artifacts and potential copy number polymorphisms, we removed all HDs that were observed with identical boundaries in two or more samples. Adjacent homozygous deletions separated by two or fewer SNPs were considered to be part of the same deletion. To identify the target genes affected by HDs, we compared the location of coding exons in the RefSeq, CCDS and Ensembl databases with the genomic coordinates of the observed HDs. Any gene with a portion of its coding region contained within a homozygous deletion was considered to be affected by the deletion.

As outlined in (*4*), amplifications were defined by regions with an average LogR ratio ≥ 0.9, containing at least one SNP with a LogR ratio ≥ 1.4 and at least one SNPwith a LogR ratio ≥ 1 every ten SNPs. As focal amplifications are more likely to be useful in identifying specific target genes, a second set of criteria were used to remove complex amplifications, large chromosomal regions or entire chromosomes that showed copy number gains. Amplifications > 3Mb in size and groups of nearby amplifications (within 1 Mb) that were also > 3Mb in size were removed. Amplifications or groups of amplifications that occurred at a frequency of ≥4 distinct amplifications in a 10 Mb region or ≥5 amplifications per chromosome were removed. The amplifications remaining after these filtering steps were considered to be focal amplifications and were the only ones included in subsequent statistical analyses. To identify protein coding genes affected by amplifications, we compared the location of the start and stop positions of each gene within the RefSeq, CCDS and Ensmbl databases with the genomic coordinates of the observed amplifications. As amplifications containing only a fraction of a gene are less likely to have a functional consequence, we only considered genes whose entire coding regions were included in the observed amplifications.

Statistical Analysis

Overview of Statistical Analysis

The statistical analyses focused on quantifying the evidence that the mutations in a gene or a biologically defined set of genes reflect an underlying mutation rate that is higher than the passenger rate. In both cases, the analysis integrates data on point mutations with data on copy number alterations (CNA). The methodology for the analysis of point mutations is based on that described in (*3*) while the methodology for integration across point mutations and CNA's is based on (*2*). This methodology was used before in both (*2*) and (*3*). We provide a self-contained summary herein, as some modifications to the previously described methods were required.

Statistical Analyses of *CAN*-genes

The mutation profile of a gene refers to the number of each of the twenty-five context-specific types of mutations defined earlier (*5*). The evidence on mutation profiles is evaluated using an Empirical Bayes analysis (*6*) comparing the experimental results to a reference distribution representing a genome composed only of passenger genes. This is obtained by simulating mutations at the passenger rate in a way that precisely replicates the experimental plan. Specifically, we consider each gene in turn and simulate the number of mutations of each type from a binomial distribution with success probability equal to the context-specific passenger rate. The number of available nucleotides in each context is the number of successfully sequenced nucleotides for that particular context and gene in the samples studied. When considering non-synonymous mutations other than indels, we focus on nucleotides at risk, as defined previously (*5*).

Using these simulated datasets, we evaluated the passenger probabilities for each of the genes that were analyzed in this study. These passenger probabilities represent statements about specific genes rather than about groups of genes. Each passenger probability is obtained via a logic related to that of likelihood ratios: the likelihood of observing a particular score in a gene if that gene is a passenger is compared to the likelihood of observing it in the real data. The gene-specific score used in our analysis is based on the Likelihood Ratio Test (LRT) for the null hypothesis that, for the gene under consideration, the mutation rate is the same as the passenger mutation rate. To obtain a score, we simply transform the LRT to s = log(LRT). Higher scores indicate evidence of mutation rates above the passenger rates. This general approach for evaluating passenger probabilities follows that described by Efron and Tibshirani (*6*). Specifically, for any given score s, F(s) represents the proportion of simulated genes with scores higher than s in the experimental data, F0 is the corresponding proportion in the simulated data, and p0 is the estimated overall proportion of passenger genes (discussed below). The variation across simulations is small but nonetheless we generated and collated 250 datasets to estimate F0. We then numerically estimated the density functions f and f_0 corresponding to F and F0 and calculated, for each score s, the ratio $p_0 \cdot f_0(s)/f(s)$, also known as "local false discovery rate" (*6*). Density estimation was performed using the function "density" in the R statistical programming language with default settings. The passenger probability calculations depend on an estimate of p_0, the proportion of true passengers. Our implementation seeks to give an upper bound to p_0 and thus provide conservatively high estimates of the passenger probability. To this end we set $p_0=1$. We also constrained the passenger probability to change monotonically with the score by starting with the lowest values and recursively setting values that decrease in the next value to their right. We similarly constrain passenger probabilities to change monotonically with the passenger rate.

An open source package for performing these calculations in the R statistical environment, named CancerMutationAnalysis, is available at http://astor.som.jhmi.edu/~gp/software/CancerMutationAnalysis/cma.htm. A detailed mathematical account of our specific implementation is provided in (*7*) and general analytic issues are discussed in (*8*). The only difference in the present study is that a gene passed into the Prevalence Screen if it had at least two non-silent alterations in at least two tumor samples in the Discovery Screen or at least one nonsynonymous mutation in the Discovery Screen and had also been previously altered in other tumor types. Under the null hypothesis, the assumptions were that a gene passed into the Prevalence Screen if it had at least two nonsynonymous mutations in the Discovery Screen or it had at least one nonsynonymous mutation in the Discovery Screen and it was on a fixed list of known candidate cancer genes.

Statistical Analysis of CNA

For each of the genes involved in amplifications or deletions, we further quantified the strength of the evidence that they drive tumorigenesis through estimations of their passenger probabilities. In each case, we obtain the passenger probability as an *a posteriori* probability that integrates information from the somatic mutation analysis above with the data presented in this article. The passenger probabilities derived from the point mutation analysis serve as *a priori* probabilities. Then, a likelihood ratio for "driver" versus "passenger" was evaluated using as evidence the number of samples in which a gene was found to be amplified (or deleted). The passenger term is the probability that the gene in question is amplified (or deleted) at the frequency observed. For each sample, we begin by computing the probability that the observed amplifications (and deletions) will include the gene in question by chance. Inclusion of all available SNPs is required for amplification, while any overlap of SNPs is sufficient for deletions. Specifically, if in a specific sample N SNPs are typed, and K amplifications are found, whose sizes, in terms of SNPs involved, are A_1 ... A_K, a gene with G SNPs will be included at random with probability

$(A_1\text{-}G+1)/N + ... + (A_K\text{-}G+1)/N$ for amplifications and

$(A_1+G\text{-}1)/N + ... + (A_K+G\text{-}1)/N$ for deletions.

We then compute the probability of the observed number of amplifications (or deletions) assuming that the samples are independent but not identically distributed Bernoulli random variables, using the Thomas and Taub algorithm (*9*). Our approach to evaluating the likelihood under the null hypothesis is highly conservative, as it assumes that all the deletions and amplifications observed only include passengers. The driver term of the likelihood ratio was approximated as for the passenger term, after multiplying the sample-specific passenger rates above by a gene-specific factor reflecting the increase (alternative hypothesis) of interest. This increase is estimated by the ratio between the empirical deletion rate of the gene and the expected deletion rate for that gene under the null. Genes that occurred in the same amplification or deletion as known cancer genes were excluded from this analysis.

This combination approach makes an approximating assumption of independence of amplifications and deletions. In reality, amplified genes cannot be deleted, so independence is technically violated. However, because of the relatively small number of amplification and deletion events, this assumption is tenable for the purposes of our analysis.

Analysis of mutated gene pathways and groups

Three types of data were obtained from the MetaCore database (GeneGo, Inc., St. Joseph, MI): pathway maps, Gene Ontology (GO) processes, and GeneGo process networks. The memberships of each of the analyzed transcripts in these categories were retrieved from the databases using RefSeq, Entrez, and Ensembl identifiers. In GeneGo pathway maps, 46,226 relations were identified, involving 6,071 transcripts and 1,063 pathways. For Gene Ontology processes, a total of 586,952 pairwise relations were identified, involving 14,361 transcripts and 8,947 GO groups. For GeneGo process networks, a total of 25,431 pairwise relationships, involving 6,419 transcripts and 155 processes, were identified.

For each of the gene sets considered, we quantified the strength of the evidence that they were altered in a higher-than-average proportion of samples from the Discovery Screen, calculating p-values using a patient-oriented gene-set analysis (the permutation null without heterogeneity method from (*10*). We then corrected for multiplicity by the q-value method with an alpha of 0.2 (*11*). An open source R package for the implementation of this method, PatientGeneSets, is currently in Version 2.7 of Bioconductor and is available at http://bioconductor.org/packages/2.7/bioc/html/PatientGeneSets.html .

Bioinformatics Analysis

CHASM uses a supervised machine learning method called Random Forest (*12,13*) to distinguish putative driver mutations on the basis of their similarity to a positive class of driver missense mutations versus a negative class of passenger missense mutations. The Random Forest is an ensemble of CART decision trees (*14*), each of which is trained on a different subset of training examples and features. The training set used here is larger than the set used in (*15*). The positive class consists of all missense mutations in the COSMIC database (*16*) that occur in genes meeting criteria to be considered as tumor suppressors or oncogenes (3299). Tumor suppressor genes are required to harbor at least 6 mutations and to have a ratio of truncating (nonsense, splice site, frameshift) to other non-silent mutations > 0.2. Oncogenes are required to have at least one amino acid position that is mutated in at least two tumors.

We generated 5000 random passenger missense mutations for training and another 5000 for feature selection, according to base substitution rates estimated from the medulloblastoma sequencing data, in eight di-nucleotide contexts.

Base substitution rates in 8 di-nucleotide contexts in Medulloblastoma.

	C in CpG	G in CpG	C in TpC	G in GpA	A	C	G	T
A	0.07	1.73	0.09	0.12	-	0.06	0.10	0.03
C	-	0.04	-	0.06	0.03	-	0.04	0.04
G	0.04	-	0.08	-	0.05	0.04	-	0.03
T	1.49	0.09	0.11	0.13	0.03	0.08	0.08	-

We selected 73 predictive features for each missense mutation, which passed a minimum threshold of 0.001 bits of mutual information with class labels. These features included general and position-specific properties of amino acid substitution, predicted protein local structure, evolutionary conservation and curated annotations from the UniProt Knowledgebase (*15,17*) According to the Random Forest feature importance criterion (*13*), the most discriminatory features are:

- Location in an enzymatic domain involved in post-translational modification;
- Compatibility with observed amino acid residues in an alignment of protein orthologs;
- Frequency of SNPs in the exon in which the mutation occurs;
- Average PhastCons (*18*) nucleotide-level conservation in the exon in which the mutation occurs;
- Change in amino acid polarity resulting from the substitution;

- Negative entropy in the column of amino acids that align to the mutated position in a protein superfamily multiple sequence alignment.

The CHASM score for a missense mutation is the fraction of decision trees in the Random Forest that vote for the passenger class. The score ranges from 0 (unanimous vote for driver) to 1 (unanimous for passenger). We compute P-values and Benjamini-Hochberg false discovery rate (*11*) using an empirical null score distribution of ~5000 random mutations generated in a set of genes unlikely to be involved in cancer, based on the Atlas of Genetics and Cytogenetics in Oncology and Haematology http://atlasgeneticsoncology.org/Genes/Geneliste.html, COSMIC, and the MSigDB C4 gene set collection (*19*). The score does not consider whether the gene in which a mutation occurs is expressed, but rather predicts whether the mutation would behave as a driver if the gene were expressed.

We applied CHASM to the 148 unique somatic missense mutations detected in this study to assess their role in medulloblastoma (Table S4). Fifteen of the mutations scored as putative drivers (FDR ≤ 0.20), one of which occurred in *TP53* and was previously known to act as a driver. Three of the mutations occurred in *PTCH1*, a gene in the sonic hedgehog signaling pathway that has previously been implicated in medulloblastoma. We estimated the fraction of non-silent mutations predicted to alter protein function in the Discovery and Prevalance sequencing phases for medulloblastoma, glioblastoma multiforme (GBM), pancreatic, breast and colorectal cancer studies by summing the predicted driver missense changes (with FDR ≤ 0.20) together with the nonsense and frameshift mutations for each tumor type (*1-3,5*).

Nonsense mutations in MB. To assess whether nonsense mutations were overrepesented in the 24 medulloblastomas, we compared the number of somatic nonsense mutations observed with the number expected to occur due to tumor-specific background mutation rates. We performed a similar comparison for missense mutations. This analysis was also repeated for an additional four tumor types from previous studies: breast, colorectal and pancreatic cancers and glioblastoma multiforme (*1-3,5*).

We made several simplifying assumptions for this analysis. Each sequenced gene was represented by its longest transcript and only single base changes were considered as possible mutations. We define X_i as the nucleotide base at position i in codon X. $X_{i'}$ is a substituted base at position i that can take on values $j \in \{A,C,G,T\}/X_i$. Finally, $c(X_i)$ is the context of the base at position i, with possible contexts $c \in$ {C*pG, CpG*,G*pA,TpC*,A,C,G,T}. C*pG indicates a C in a CpG dinucleotide. The last four contexts cover all A and T nucleotides and C and G nucleotides that do not match any of the first four contexts.

The substitution rate of $c(X_i) \text{->} X_{i'}$ in tumor type t is

$$\lambda^{t}_{c(x_i)\rightarrow x_{i'}}$$

which we estimated from the number of non-silent mutations per megabase detected during tumor sequencing, the context in which each mutation occurred and the fraction of nucleotide substitutions seen in each context.

The expected number of nonsense mutations in a single codon is

$$E[X \rightarrow X_{\mathrm{STOP}}] = \sum_i \sum_j \lambda^t_{c(x_i) \rightarrow x_{i'} = j} \delta(x_i)$$

where

$$\delta(x_i) = \begin{cases} 1 & \text{if } x_i \rightarrow x_{i'} \text{ results in stop codon} \\ 0 & \text{otherwise} \end{cases}$$

The expected number of nonsense mutations in a gene is the sum of expectations over its constituent codons. The expected number in a single sequenced tumor exome is the sum over the sequenced genes. Finally, the expected number of nonsense mutations seen in all sequenced tumors of type *t* is

$$n \sum_{i=1}^{m} \sum_{j=1}^{l_i} E[X_{i,j} \rightarrow X_{\mathrm{STOP}}]$$

where *n* is the number of sequenced tumors, *i* indexes the *m* genes in each sequenced tumor, *j* indexes the codons in each gene and l_i is the number of codons in gene *i.*

The expected number of missense mutations were computed with the same method.

Finally, we computed the proportions of expected and observed nonsense mutations with respect to the total number of nonsilent (nonsense plus missense) for each tumor. Because expectations were computed for all genes in the sequenced exomes, we only considered observed mutations from the whole exome Discovery Phase. We used R statistical software (*21*) to evaluate the null hypothesis that the expected and observed proportions were statistically identical.

Supplemental References

1. T. Sjoblom *et al.*, *Science* **314**, 268 (2006).
2. S. Jones *et al.,* Science **321**, 5897 (2008).
3. D.W. Parsons *et al., Science* **321**, 5897 (2008).
4. R. J. Leary *et al.*, *Proc Natl Acad Sci U S A.* **105**,16224 (2008).
5. L. D. Wood *et al.*, *Science* **318**, 1108 (2007).
6. B. Efron, R. Tibshirani, *Genet Epidemiol* **23**, 70 (2002).

7. G. Parmigiani *et al.*, “Statistical Methods for the Analysis of Cancer Genome Sequencing Data” (Johns Hopkins University, 2006).

8. G. Parmigiani *et al.*, *Genomics* **93**, 17 (2009).

9. M. A. Thomas, A. E. Taub, *Journal of Statistical Computation and Simulation* **14**, 125 (1982).

10. S.M. Boca *et al., Genome Biology, in press* (2010).

11. Y. Benjamini, Y. Hochberg, *Journal of the Royal Statistical Society. Series B (Methodological)* **57** 289 (1995).

12. Y. Amit, D. Geman, *Neural Computation* **9**, 1545 (1997).

13. L. Breiman, *Machine Learning* **45**, 5 (2001).

14. L. Breiman, “Classification and regression trees: Regression Trees, The Wadsworth Statistics/Probability Series” (Wadsworth International Group, 1984).

15. H. Carter *et al*., *Cancer Res* **69**, 6660 (2009).

16. S. Forbes *et al*., *Br J Cancer* **94**, 318 (2006).

17. C.H. Wu *et al*., *Nucleic Acids Res* **34**, D1897 (2006).

18 A. Siepel *et al*., *Genome Res* **15**, 1034 (2005).

19. A. Subramanian *et al*., *Proc Natl Acad Sci U S A* **102**, 15545 (2005).

20. E. Parzen, *Ann Math Stat* **33**, 1065 (1962).

21. R Development Core Team. (Vienna, Austria, 2006).