FunciSNP: an R/bioconductor tool integrating functional non-coding data sets with genetic association studies to identify candidate regulatory SNPs

Simon G. Coetzee[1,2], Suhn K. Rhie[1,2], Benjamin P. Berman[1,2,3], Gerhard A. Coetzee[1,2,4,*] and Houtan Noushmehr[1,2,3,*]

[1]Norris Cancer Center, [2]Department of Preventive Medicine, [3]Epigenome Center and [4]Department of Urology, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA

Received March 20, 2012; Revised April 26, 2012; Accepted May 13, 2012

ABSTRACT

Single nucleotide polymorphisms (SNPs) are increasingly used to tag genetic loci associated with phenotypes such as risk of complex diseases. Technically, this is done genome-wide without prior restriction or knowledge of biological feasibility in scans referred to as genome-wide association studies (GWAS). Depending on the linkage disequilibrium (LD) structure at a particular locus, such tagSNPs may be surrogates for many thousands of other SNPs, and it is difficult to distinguish those that may play a functional role in the phenotype from those simply genetically linked. Because a large proportion of tagSNPs have been identified within non-coding regions of the genome, distinguishing functional from non-functional SNPs has been an even greater challenge. A strategy was recently proposed that prioritizes surrogate SNPs based on non-coding chromatin and epigenomic mapping techniques that have become feasible with the advent of massively parallel sequencing. Here, we introduce an R/Bioconductor software package that enables the identification of candidate functional SNPs by integrating information from tagSNP locations, lists of linked SNPs from the 1000 genomes project and locations of chromatin features which may have functional significance.

Availability: FunciSNP is available from Bioconductor (bioconductor.org).

INTRODUCTION

Genome-wide association studies (GWAS) have yielded numerous single nucleotide polymorphisms (SNPs) significantly associated with many phenotypes (P-value $< 9e^{-06}$). In some cases, dozens of SNPs, called tagSNPs, mark each of a number of distinct loci in complex diseases, such as prostate (>50 loci), breast (>20 loci), ovarian (>10 loci), colorectal (>10 loci) and brain cancers (>5 loci), but the mechanisms by which these polymorphisms exert their functions remains largely unknown (1–3). Since most of the tagSNPs (>80%) are found in non-protein coding regions (intergenic and intron regions; Figure 1), identifying functional and/or causal variants has been an important limitation of GWAS data interpretation (4). Several hypotheses have been proposed to explain this phenomenon (3,5), such as effects via largely unannotated transcripts (e.g. non-coding RNAs and rare splice variants) or gene regulatory sequences (e.g. insulators, enhancers or silencers). Testing for the effects of risk SNPs at regulatory sequences has been successful; for example, genomic enhancers identified by histone mapping were shown to be highly enriched across a number of diverse GWAS studies (6), and at the gene desert of chromosome 8q24 where specific enhancers were identified as differentially affected by SNP alleles (7). However, assigning putative functionality to many

*To whom correspondence should be addressed. Tel: +1 310 570 2362; Fax: +1 323 442 7880; Email: houtan@usp.br
Correspondence may also be addressed to Gerhard A. Coetzee. Tel: +1 323 865 0631; Fax: +1 323 865 0634; Email: coetzee@usc.edu
Present addresses:
Simon G. Coetzee, Department of Genetics, Medical School of Ribeirão Preto, University of São Paulo, Ribeirão Preto, São Paulo, 14049-900, Brasil.
Houtan Noushmehr, Department of Genetics, Medical School of Ribeirão Preto, University of São Paulo, Ribeirão Preto, São Paulo, 14049-900, Brasil.

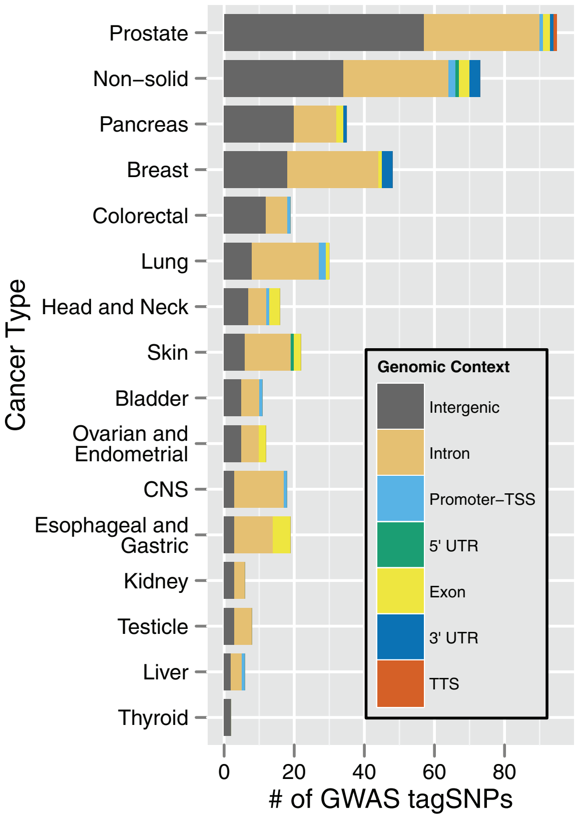

Figure 1. Summary of all known GWAS SNPs across a number of different cancer types. GWAS SNPs were annotated using known genomic features supplied by HOMER (version 3.9) (10), using build hg19 as reference. GWAS SNPs were extracted from the UCSC Genome table browser, track name 'gwasCatalog' with a *P*-value cut-off of 9e^{-06} (1).

other GWAS tagSNPs has only been successful when fine mapping around a known risk region was performed (8,9).

Linkage disequilibrium (LD) is defined as the non-random association of alleles at two or more loci. In population genetic studies, LD is influenced by the rate of recombination, mutation, genetic drift or selection. Due to the nature of LD, each individual locus identified in an association study can yield up to hundreds of surrogate SNPs for each linked group of tagSNPs. The first step in any *in silico* analysis is to take a genomic window around each tagSNP and extract all known variants (at least with a minor allele frequency of $\geq$1%) with the assumption that the functional and/or causal variant(s) is likely contained within this window (3,4). Within this genomic window, LD structure between populations and genotype can be used to subsequently refine estimates of risk, but the number of linked SNPs can generally still be quite large. To aid in identifying a full spectrum of variants in the genome, the 1000 genomes project recently released a catalog of human genomic variants (minor allele frequency of $\geq$1%) across many different ethnic populations (2,11). Initially, the 1000 genomes project goal was to sequence up to 1000 individuals, but has since sequenced more than 2000 individuals, thereby increasing our current knowledge of known genomic variations, which currently is at just over 50 million SNPs genome wide (~2% of the entire genome and on average 1 SNP every 60 bp) (2).

Ascertaining biological function for each SNP requires well-planned, and often expensive and time-consuming, molecular biology experiments (9). Thus, analyzing the large number SNPs linked to any particular locus in practice requires a systematic bioinformatic evaluation and prioritization to narrow the set of likely functional candidate variants. In a recent perspective paper, we and others recently formulated a well-ordered approach in assigning functionality to coding and non-coding risk regions (3). In this approach, a set of molecular (*in vitro* and *in vivo*) as well as bioinformatic (*in silico*) tools and strategies are used to prioritize regions of interest within identified loci for subsequent functional follow-up studies. Several bioinformatic tools have been reported previously which take into account the LD structure and gene expression effect (eQTL) or likely impact on mutation(s) that causes amino acid substitution and protein function or other known observable phenotypic associations, such as clinical features (12,13). However, these tools rely primarily on gene annotations and do not incorporate the critical non-coding genomic features known to have regulatory function and likely to underlie many of the tagSNPs identified outside of gene regions in complex diseases.

Recently, with the advent of advanced sequencing technologies (next-generation sequencing, NGS), genomic architecture and regulatory elements in non-coding regions are becoming well characterized and annotated via sequence-based chromatin mapping/epigenomics techniques (14–17). Coupling high-throughput sequencing to chromatin immunoprecipitation for regulatory proteins (e.g. transcription factor or histone marks), known as ChIP-seq, has provided us with a much more comprehensive view of the genomic regions that regulate transcription (Supplementary Figure S1A). Specifically, it enables mapping of genomic regions of DNA fragments bound by transcription factors, epigenetic histone marks or other proteins at an amazing resolution. Many ChIP-seq algorithms are currently available to assist in identifying these enriched regions or 'peaks' (18). Since the current prevailing consensus is that identified ChIP-seq peaks are highly correlated to biological function, we define the collection of peaks for an experimental type as a 'biofeature'.

In addition to profiling genomic regions marked by distinct protein:DNA interactions, the advancement in sequencing technologies have also been used to demarcate regions of the genome defined as either euchromatin (lightly packed form of chromatin) or heterochromatin. Even mapping of subtle nucleosome-depleted regions (NDR) as found at promoters and enhancers is now possible. Specifically, it has been noted that a variety of different histone modifications exists on well-positioned nucleosomes surrounding NDRs in a variety of cell types depending on their differentiated state (19). Currently, two distinct technical methods are used to profile NDR in an unbiased manner. DNase I hypersensitive regions are generally regions demarcate 'sites of action' in the genome which includes active promoters and enhancers. Formaldehyde-assisted isolation of regulatory elements (FAIRE) is a similar approach used to characterize open chromatin structures genome-wide when coupled with deep sequencing; however, the method does not include any enzymatic steps and is therefore simpler to use. Recently, Song *et al.* identified regulatory elements that shape cell-type identity and found that FAIRE-seq and DNaseI-seq identify distinct but overlapping profiles of NDR (20).

Work by large consortia groups such as the Encyclopedia of DNA Elements (ENCODE) (14), the Roadmap Epigenomics Mapping Consortium (21) and The Cancer Genome Atlas (TCGA) (22), have made publicly available a growing catalog of many different histone marks, transcription factors and genome-wide sequencing data sets for a variety of different diseases and cell lines, including well-characterized normal and cancer cell lines, such as IMR90 (fibroblast), MCF7 (breast cancer), HCT116 (colon cancer), U87 (brain cancer) and LNCaP (prostate cancer). Integrating and correlating many of these publicly available data with unpublished genomics and epigenomics data was recently described in a study of the first colon cancer methylome (17). This study illustrated the power of integrating whole-genome DNA methylation data with publicly available ChIP-seq data sets to gain novel insights into the biology of the cancer epigenomic architecture, specifically with respect to the 3D organization of chromosomes with the cell nucleus that lead to changes in gene expression. The number of cell line with whole-genome chromatin maps is rapidly increasing, along with the diversity of mapping techniques—innovative new techniques include ChIA-PET (23), ChIP-exo (24), ChIRP (25) and NOMe-seq (26). This wealth of chromatin and epigenomics data will be invaluable in interpreting disease polymorphisms, but tools to exploit it do not currently exist.

Here, we describe a new bioinformatic tool, called Functional Identification of SNPs (FunciSNP) to aid in the identification of candidate functional SNPs associated with a phenotype by integrating and correlating knowledge obtained from three whole-genome sequencing data types (1000 genomes, GWAS SNPs and sequence-based chromatin maps). Integrating non-coding regions as annotated by chromatin mapping helps inform and prioritize candidate regulatory regions for follow up molecular experiments. Using FunciSNP, we test the hypothesis that there may be many more putative functional SNPs associated with a phenotype that are in LD to the original tagSNP. To introduce and describe FunciSNP's utility and application, we used glioblastoma multiforme (brain cancer) as an example GWAS phenotype (27–30). We extract ENCODE ChIP-seq data for binding of two distinct transcription factors (TFs) in a glioma cell line, U87 (14), as these sites may have functional significance for the GBM cancer phenotype. We identified several putative functional SNPs overlapping the transcription factor (TF) binding sites, and promoter regions of well-annotated genes, which are in strong LD to the GWAS tagSNP. Many, but not all, of the genes near these candidate regulatory SNPs have been previously reported to be important in glioblastoma development and risk.

MATERIALS AND METHODS

FunciSNP package

FunciSNP is an R package, which is licensed under the General Public License (GPLv3) and is freely available through the Bioconductor repository (31). By developing the package in R and conforming to the strict guidelines for package submission to Bioconductor, we are able to utilize and incorporate existing R packages and statistics to assist in identifying candidate functional SNPs. FunciSNP builds upon and integrates the following Bioconductor core packages: Rsamtools, GGtools, VariantAnnotation, snpStats and ChIPpeakAnno. A schematic overview is described in Figure 2.

In order to successfully run FunciSNP, two inputs are required: GWAS tagSNP information and a set of user-defined NGS peak files (biofeatures). GWAS tagSNPs and NGS peaks were discussed in the 'Introduction' above. Specifically, biofeatures are defined

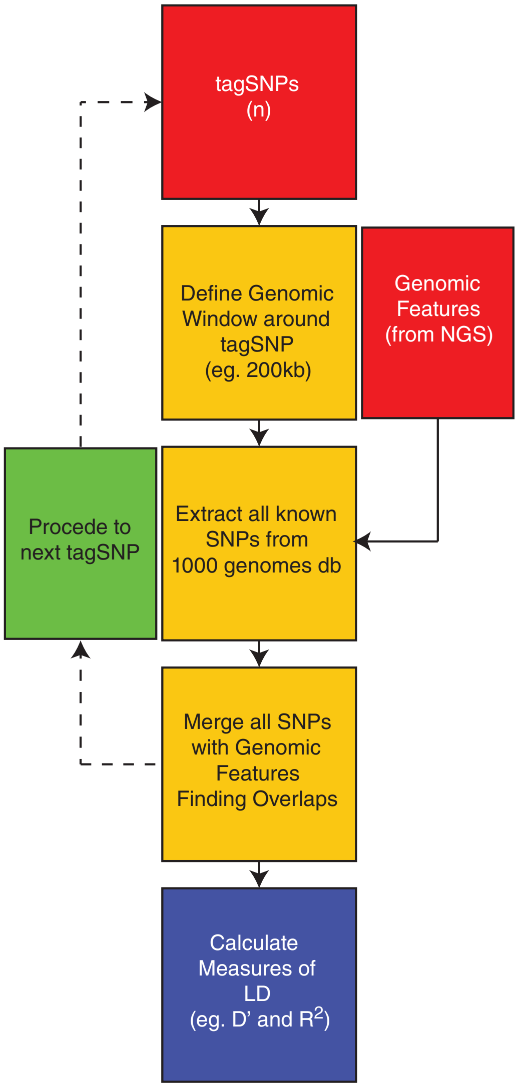

Figure 2. Schematic flowchart to describe FunciSNP. Purple boxes represent process before integration with biofeature. Red boxes represent information after integration with biofeature.

here as a collection of genomic regions identified by deep sequencing of a particular experimental type. These genomic regions were computationally identified using currently available 'peak calling' algorithms as discussed in a recent report (32). During an initial run, FunciSNP extracts all available 1000 genomes SNPs (1kgSNP) around a user-defined window centered on each tagSNP (Figure 2). By default, this window size is defined as 200 000 base pairs (100 000 bp on either side of the tagSNP). User can easily define this window by setting the window.size argument in getFSNPs(). The information on each 1kgSNP is then used to find overlaps with each defined biofeature. Only those 1kgSNPs that overlap a biofeature annotation are used to calculate the R^2, D′ and genomic annotations (distance to nearest TSS, nearest lincRNA, genomic characterization, such as gene bodies and promoters, see Section 3 of Supplemental Text for more detail). R^2 and D′ are useful calculations to access the degree of LD between two alleles. In this case, we use these calculations to evaluate the degree of correlation or association between the original GWAS tagSNP and the identified 1kgSNP overlapping a biofeature. FunciSNP will use these two parameters to filter the list even further to identify the most likely candidate functional SNP associated with the phenotype. In addition, FunciSNP can provide R^2, D′ and genomic annotations for all 1kgSNPs extracted from the 1000 genomes project, independent of any available biofeature. For more fine grained study, FunciSNP will output this data in an annotated table (Supplemental Table S1) that contains an entry for each unique combination of tagSNP, YAFSNP and biofeature peak. From this table, users are free to appropriate the totality data created by FunciSNP, for export or use in the wide variety of packages in the Bioconductor and R ecosystems. This application is useful when knowledge about all SNPs in LD to the tagSNP is important, regardless of the overlapping biofeature. And, finally, FunciSNP contains several custom plots and summary functions to aid in making informed hypothesis of the newly identified candidate functional SNPs (see Supplemental Text for more details).

Glioma GWAS SNP data

Seven tagSNPs associated with glioma were obtained from recent GWAS (27–30). Four of the seven tagSNPs were randomly selected as an example set. The 'snp.regions' file is organized in a tab or whitespace separated file with three columns: genomic position, rs number and population for each tagSNP. See Supplemental Text for more information on how to organize a 'snp.regions' file for input. Using correctly matching genome reference coordinates is critical, and hg19 coordinates were used throughout this study.

Biofeature annotation data

ChIP-seq peaks, regions in the genome identified by deep sequencing of immunoprecipation of a specific DNA:protein complex, for all available ENCODE data associated with glioma cell lines were obtained through ENCODE's public repository (14). Only two distinct ChIPseq types were available as biofeatures for U87, a glioma cell line: neuron-restrictive silencer factor (NRSF) and RNA polymerase II (Pol2). These two peak files (biofeatures) were used as FunciSNP biofeature inputs in this example. Each ChIP-seq data set was collected and translated into a standard BED format. In addition to user-defined biofeatures, FunciSNP also contains a list of all known annotated promoters as defined by a window around a known transcription start site (TSS). The window parameter for promoters is −1000 to +100 bp from the TSS. FunciSNP also contains information on known CTCF binding sites (33) as well as DNaseI hypersensitive location across a number of different cell lines (14). CTCF and DNaseI sites make up a large fraction of open chromatin regions throughout the genome and thus represent a set of regions likely to be highly enriched in gene regulatory elements (6). We also included FAIRE-seq peaks and we believe the combination of NDR peaks will assist in making informed decision of candidate functional SNPs (see 'Introduction' section). FAIRE-seq peaks were extracted from more than 100 different cell types, collected from the UNC FAIRE ENCODE data. Peaks were filtered such that only peaks with a *P*-value below 0.01 were considered significant. The remaining peaks from each cell type were then merged into a single file representing clusters, across cell types, of FAIRE peaks. Each peak data set is defined in FunciSNP as a biofeature. See Supplemental Material for more information as well as how to load publicly available peak files from ENCODE.

1000 genomes SNPs

All SNPs within the specified window surrounding each tagSNPs were extracted directly from FTP servers from the 1000 genomes public repository. From time to time, based on server load, connection to one of the FTP servers may become interrupted, leading to corruption of 1000 genomes data downloaded by FunciSNP. Several checks are in place to check the server status and the reliability of the data. Data is initially requested from either EBI or NCBI server at random, with a short wait time between requests. If a problem is detected (either the connection cannot be made, or the data is incomplete), the request is resubmitted to the alternate server. This repeats until successful, and allows for the process to run unattended to completion. The GenomicRanges package (version ≥ 1.6.7) from Bioconductor allows for efficient downloading of only those 1kgSNPs overlapping the selected biofeatures. This allows for much shorter execution times.

Statistical and data analysis

LD is defined as the non-random association between allelic markers on a chromosome and is classically measured using one of two estimators, D′ or R^2 (34). Each correlated SNP pooled from the 1000 genomes database along with available population identification is used to calculate the R^2 and D′. All plots and summary outputs are generated using *R* (version 2.14) (31) and most of FunciSNP's plots are generated using ggplot2 (http://had.co.nz/ggplot2/).

The two respective measures of LD are generated between all SNPs (1000 genomes project SNPs and the tagSNP) that overlap at least one biofeature. The snpStats package (available through bioconductor) provides an efficient mechanism to calculate LD and store SNP genotypes. We extend this structure to provide the 3 × 3 genotype table and 2 × 2 haplotype table required to perform further calculations. The *P*-value is the result of a Fisher's Exact Test performed on the 2 × 2 table of haplotype frequencies, and can be useful as a guide to evaluating the R^2 and D′ measures of LD. Crucially, all calculations are done solely within the population of the selected tagSNP, allowing population specific measures of LD. For the cases wherein tagSNP's population is given as ALL, all calculations are repeated for each group (AFR, AMR, ASN and EUR). Since multiple *P*-values are calculated, we corrected the *P*-values by Benjamini-Hochberg (BH) method. BH is set by default in FunciSNP, but user may invoke any one of the available multiple testing correction (see 'p.adjust()' function in R).

All data is contained in a nested collection of objects that can be exposed to allow the user to extend their analysis in ways that are not built into FunciSNP itself. The primary object output by FunciSNP is a TSList object that contains a list of the TagSNP objects. Each TagSNP object is represented by its rs number followed by the population in which it was examined (ex: rs10936599:EUR). This object contains data representing the chromosomal position, the reference SNP cluster 'rs' ID, the population in which it was studied, the reference and alternate allele, the genotype information across the population for the tagSNP, the R^2 and D′ measures of LD between the tagSNP and all 1000 genome SNPs within the examined window that overlap at least one biofeature, and a CorrelatedSNP object. The CorrelatedSNP object contains similar information, but for all 1000 genomes SNPs within the examined window. Additionally, it contains data that identify which potentially correlated SNPs overlap biofeatures, which biofeatures those include, and the genomic location of that biofeature. Handles to all data are exposed, for example, to access the genotype information for all SNPs within the examined window for SNP rs6010620:EUR that overlap at least one biofeature in our example data set one would enter the following command: > EUR. overlapping.snps.geno(glioma[“rs6010620”]).

Gene annotations close to each surrogate SNPs overlapping at least one biofeature are used to annotate the genomic context of the SNP, using existing Bioconductor packages. VariantAnnotation (version ≥ 1.0.5) was used to annotated genomic context with respect to gene annotations—exon, intron, 5′UTR, 3′UTR, promoter, lincRNA or intergenic. TxDb.Hsapiens.UCSC. hg19.knownGene (version ≥ 2.6.2) provides the list of known genes from the UCSC genome browser. ChIPpeakAnno (version ≥ 2.2.0) (35) is used to identify the nearest gene(s) or long non-coding RNAs (lincRNAs). LincRNA information was obtained directly from the UCSC Genome table browser. This genomic context information was used for all summary plots, summary tables and to generate BED files used to visualize results in a genome browser such as UCSC Genome Browser (see 'Results' and 'Supplemental Material' sections for more information).

Computing time

Running FunciSNP takes on average 3.8 s per 10 000 bp window size centered on a tagSNP using one single central processing unit (CPU) core (Supplementary Figure S1B). In order to increase the processing speed, we incorporated *R*'s base package (parallel) to run processes across multiple CPUs. Depending on the total number of original tagSNPs, users can specify as many CPUs necessary to run analysis. For example, if user has four available CPU and wants to only analyze two tagSNP, then the maximum number of CPU is 2. Using maximum number of CPU cores to tagSNP ($n = 4$), FunciSNP completed the analysis on average 1.5 s per 10 000 bp window size centered on a tagSNP (Supplemental Figure S1B). If we increase the total number of biofeatures, it does not significantly change the run time (data not shown). Since FunciSNP downloads information from the 1000 Genomes FTP server for each tagSNP surrounding a predetermined window size, the final time can vary significantly.

Vignette

A user guide is provided which details each command and output. See Supplemental Text for more information.

RESULTS

Identification of several new candidates functional SNPs overlapping genomic biofeatures associated with glioma (brain cancer)

To test FunciSNP, we used four known brain cancer tagSNPs derived from recent GWAS (27–30) and available biofeatures specific to glioma cell lines (U87) as reported by the ENCODE public data release (14). Five different biofeatures were incorporated; NRSF and Pol2 regions from U87 cells, along with CTCF binding sites and DNaseI hypersensitive regions derived from multiple cell lines, and promoter regions surrounding all annotated TSSs. Using four tagSNPs position and five different biological features (ChIPseq for 'NRSF', 'PolII', CTCF, DNaseI, DNaseI + CTCF and promoters of approximately 38 000 genes) as two types of input, FunciSNP identified 1205 candidate SNPs that overlap at least one biofeature (Supplementary Figure S1C and D).

Each candidate SNP contains an R^2 value to the associated tagSNP (Figure 3A). Using R^2 cut-off at 0.5, we identified 48 candidate SNPs overlapping at least one biofeature. This value represents 3.98% of the total available candidate SNP in LD to all four tagSNPs. In addition, we found three biological features for which three candidate SNPs overlap. Interestingly, tagSNP rs6010620 (28) is associated with 40 different candidate

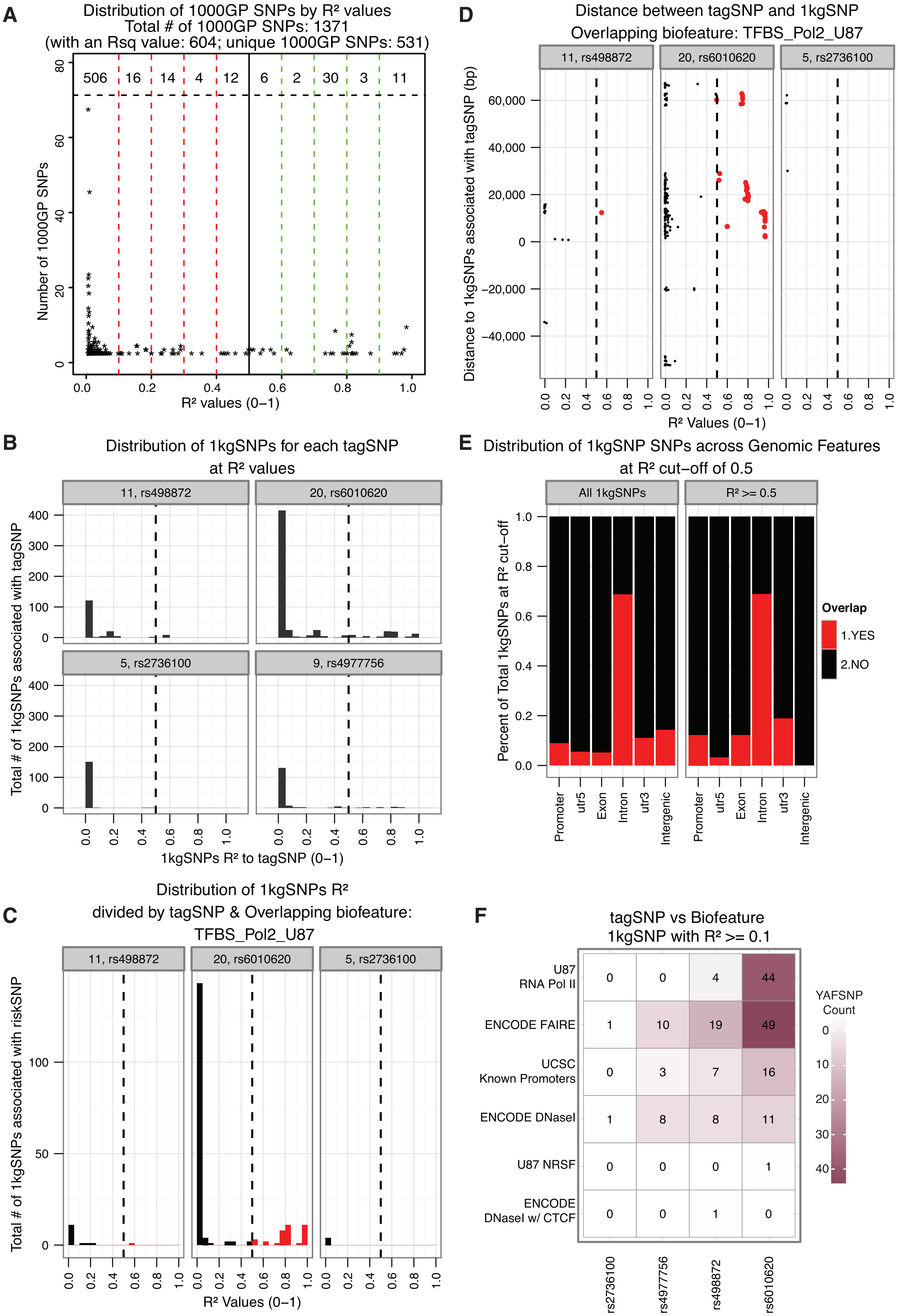

Figure 3. (**A**) Distribution of R^2 values of all YAFSNPs. Each marked bin contains the total number of YAFSNPs. The sum of all the counts would total the number of correlated SNPs. (**B**) Distribution of R^2 values of all YAFSNPs divided by the tagSNP and by its genomic location. (**C**) Histogram distribution of R^2 value for all 1kgSNP extracted and overlaps PolII. R^2 values are determined by its association to the tagSNP. (**D**) Scatter plot of the R^2 and distance to tagSNP for all 1kgSNP extracted and overlap PolII. (**E**) Stacked bar chart summarizing all correlated SNPs for each of the identified genomic features: exon, intron, 5UTR, 3UTR, promoter, lincRNA or in gene desert. R^2 cut-off at 0.5. This plot is most informative if used with a rsq value. (**F**) Heatmap of the number of 1kgSNPs by relationship between tagSNP and biofeature. Total number of YAFSNPs is listed within each quadrant to represent the number of potential candidate functional SNPs overlapping a biofeature (*y*-axis) which are in LD to the original tagSNP (*x*-axis).

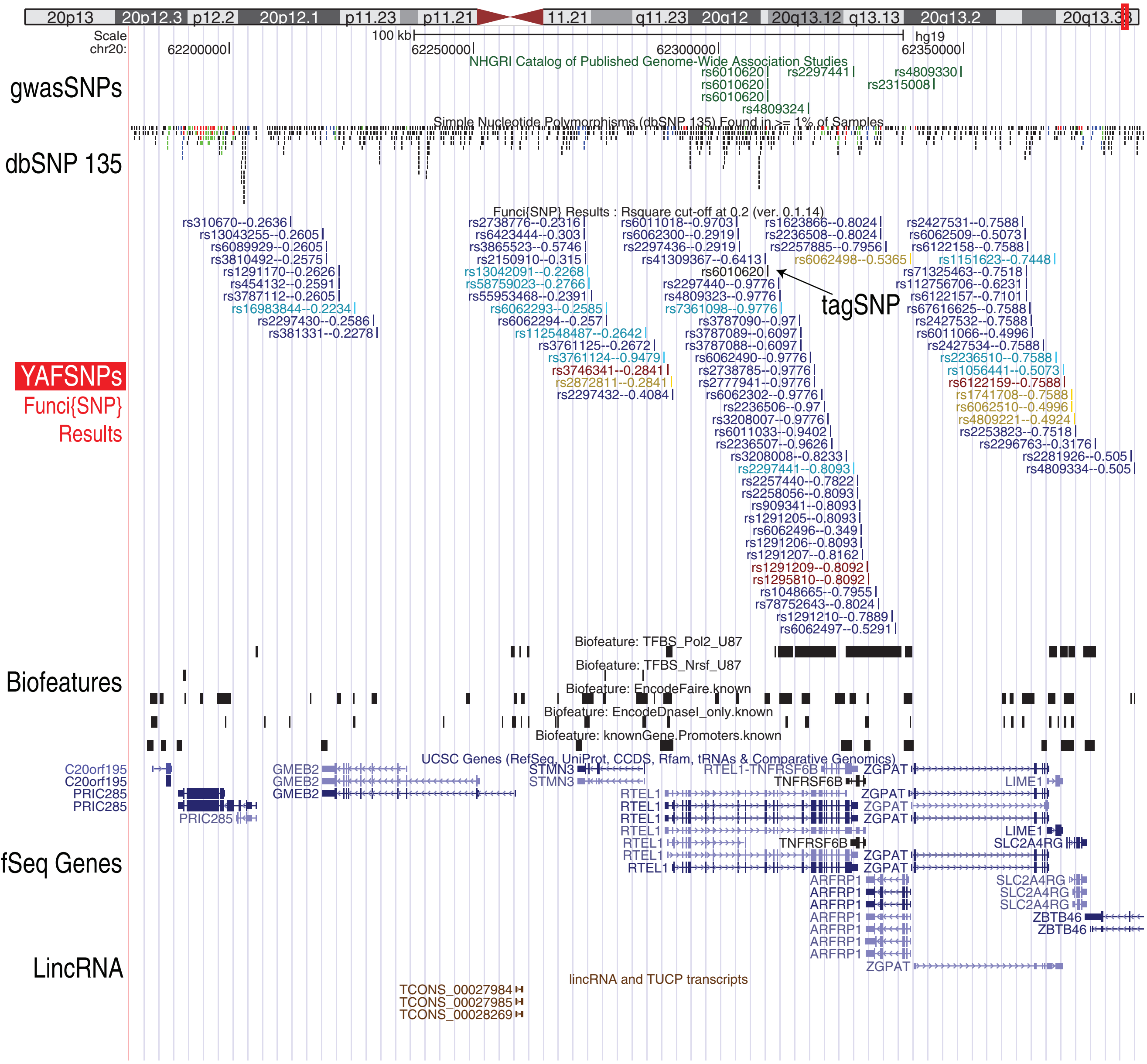

Figure 4. FunciSNP results viewed in UCSC genome browser. Tracks are ordered in the following manner: known GWAS hits, dbSNP135, FunciSNP result, biofeatures, refseq genes and known lincRNA. TagSNP is highlighted in the FunciSNP result track and each YAFSNP is color coded to reflect the number of biofeatures which it overlaps. The color ranges from blue (low number of biofeature overlap) to red (high number of overlap). Each YAFSNP is highlighted by its known rsID and the calculated R^2 value. The results are saved in a UCSC genome session: http://goo.gl/xrZPD.

SNPs which overlap at least one biofeature, and four of them overlap at least three biofeatures (Figure 3B and C, Supplementary Figure S1D and 2A–C). We refer to each newly identified SNPs as a 'YAFSNP' (yet another candidate functional SNP), since they are now known to overlap a number of different biofeatures and are in LD to the tagSNP or phenotype, in this case brain cancer.

In addition, we annotated each YAFSNPs to the nearest gene by using 'geneSum' set to TRUE (Supplementary Figure S1E). Interestingly, CDKN2B and TNFRSF6B (RETL1) were reported previously to have a functional role in glioma development and high-risk association in brain cancer (28,36). Figure 3D describes the relative position of all newly identified YAFSNPs to the associated tagSNP. TagSNP 'rs6010620' (28) contains many more YAFSNPs with $R^2 \geq 0.5$, the majority of which are contained in a small cluster between +0 and +20 kb from the tagSNP. Interestingly, another significant set of YAFSNPs in strong LD lies about 60 kb downstream (Figure 3D and Supplementary Figure S2D–F).

Genomic annotation of candidate functional SNPs for glioma

The majority of our identified YAFSNPs with $R^2 > 0.5$ to the associated tagSNP are enriched in introns but depleted in intergenic regions and promoters (Figure 3E). The cross-indexed heatmap (Figure 3F) highlights the relationship between each tagSNP and the number of associated biofeatures of each type. This figure is informative because it assists in highlighting specific tagSNPs with the most number of correlated SNPs overlapping a number of distinct biofeatures. Again, it is clearly visible that rs6010620 contains the most number of associated YAFSNPs which overlap several biofeatures, whereas rs2736100 contains very limited number of YAFSNPs which overlap only one biofeature (Figure 3F). In addition to visualizing the data in this heatmap, user can also extract information directly from the data matrix outputted from FunciSNPAnnotateSummary() (see Supplemental Text for additional insight into extracting information from the results).

Another feature we developed into FunciSNP is the ability to output the entire FunciSNP results in a standard UCSC BED file which can then be loaded into any genome browser (Figure 4). In this case, we extracted all YAFSNPs associated with tagSNP rs6010620 and highlighted all YAFSNPs in red along with the overlapping biofeatures. This provides genomic context to the final results, which illustrates all associated YAFSNPs overlapping a number of different promoters of genes and the relative genomic position to each other. In addition, using UCSC genome browser to visualize FunciSNP results offers the opportunity to add additional publicly available tracks (e.g. ENCODE TF binding motifs and conservation; Figure 4) to assist in formulating hypotheses and in selecting candidate functional SNPs for follow up studies. In our example, it is now clear that the central cluster of YAFSNPs overlap two large regions of PolII that mark the transcripts for RTEL1-TNFRSF6B and ZGPAT. At about 60 kb downstream of tagSNP, another set of interesting YAFSNPs overlap PolII marking in the promoter regions of the SLC2A4RG and LIME1 transcripts. A third and fourth set of YAFSNP loci occur about 20 and 40 kb upstream—interestingly, one of these (rs6062293) overlaps one of the rare NRSF sites within the STMN3 gene.

DISCUSSION

Because many GWAS were performed using microarray technologies that contained only a small fraction of SNPs which were determined using a limited number of populations, it is likely that they often do not contain the functional SNP responsible for risk, but rather a linked surrogate. As described in Freedman *et al.* (3), methods are needed to specifically identify and prioritize functional candidate SNPs in non-coding regions that may be more likely to confer risk to the disease than the GWAS-identified tagSNP. However, to our knowledge, no open source or freely available tool exists to perform these functions. We developed FunciSNP to fully integrate information derived from GWAS, 1000 genomes database and chromatin mapping/epigenomics data in order to identify candidate functional SNPs. We expect FunciSNP will better assist molecular epidemiologist and biologist in characterizing candidate markers for risk in complex diseases such as cancer, diabetes, obesity, Alzheimer's disease and others.

In order to describe a proof-of-principle case for FunciSNP, we used glioma (brain cancer) as an example because the biological significance of GWAS tagSNPs is not currently understood. We integrated five distinct biological features with four glioma-associated tagSNPs. We identified a region containing SNPs that are highly linked to tagSNP rs6010620 and overlap at least one biofeature. We expect most, if not all, of these candidate SNPs to have highly correlated functional relevance in the context of brain cancer. Follow-up molecular experiments and epidemiological studies are required to validate their putative function and associated risk in brain cancer.

We have performed analyses using FunciSNP in breast, prostate, colon and ovarian cancer with very high success in validating candidate functional SNPs/regions in potential enhancer region by using *in-vitro* enhancer assays (manuscripts submitted/or in preparation). Thus, the identification, characterization and possible clinical link of these newly identified putative functional SNP and the associated genomic regions should become a lasting legacy of GWAS and ultimately justify the initial substantial investment into these studies.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Figures 1–2 and Supplementary Methods.

ACKNOWLEDGEMENTS

We like to thank Joshua Millstein for suggesting the name, 'FunciSNP'. We like to thank the members of GES meeting for suggestions and comments. We like to thank Timothy Triche Jr for suggestions and tips for bioconductor package submission.

FUNDING

National Institutes of Health (NIH) [P30CA014089 to B.P.B.; CA109147, U19CA148537 and U19CA148107 to G.A.C.; 5T32CA009320-27 to H.N.]. The scientific development and funding of this project were in part supported by the Genetic Associations and Mechanisms in Oncology (GAME-ON): a NCI Cancer Post-GWAS Initiative. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute or the National Institutes of Health. Funding for open access charge: NIH.

Conflict of interest statement. None declared.

REFERENCES

1. Hindorff,L.A., Sethupathy,P., Junkins,H.A., Ramos,E.M., Mehta,J.P., Collins,F.S. and Manolio,T.A. (2009) Potential etiologic and functional implications of genome-wide association

loci for human diseases and traits. *Proc. Natl Acad. Sci. USA*, **106**, 9362–9367.

2. Consortium,T.G.P. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.
3. Freedman,M.L., Monteiro,A.N., Gayther,S.A., Coetzee,G.A., Risch,A., Plass,C., Casey,G., De Biasi,M., Carlson,C., Duggan,D. *et al.* (2011) Principles for the post-GWAS functional characterization of cancer risk loci. *Nat. Genet.*, **43**, 513–518.
4. Wang,X., Prins,B.P., Sober,S., Laan,M. and Snieder,H. (2011) Beyond genome-wide association studies: new strategies for identifying genetic determinants of hypertension. *Curr. Hypertens. Rep.*, **13**, 442–451.
5. Coetzee,G.A., Jia,L., Frenkel,B., Henderson,B.E., Tanay,A., Haiman,C.A. and Freedman,M.L. (2010) A systematic approach to understand the functional consequences of non-protein coding risk regions. *Cell Cycle*, **9**, 47–51.
6. Ernst,J., Kheradpour,P., Mikkelsen,T.S., Shoresh,N., Ward,L.D., Epstein,C.B., Zhang,X., Wang,L., Issner,R., Coyne,M. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–49.
7. Jia,L., Landan,G., Pomerantz,M., Jaschek,R., Herman,P., Reich,D., Yan,C., Khalid,O., Kantoff,P., Oh,W. *et al.* (2009) Functional enhancers at the gene-poor 8q24 cancer-linked locus. *PLoS Genet.*, **5**, e1000597.
8. Chang,B.L., Cramer,S.D., Wiklund,F., Isaacs,S.D., Stevens,V.L., Sun,J., Smith,S., Pruett,K., Romero,L.M., Wiley,K.E. *et al.* (2009) Fine mapping association study and functional analysis implicate a SNP in MSMB at 10q11 as a causal variant for prostate cancer risk. *Hum. Mol. Genet.*, **18**, 1368–1375.
9. Chung,C.C., Magalhaes,W.C., Gonzalez-Bosquet,J. and Chanock,S.J. (2010) Genome-wide association studies in cancer: current and future directions. *Carcinogenesis*, **31**, 111–120.
10. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.
11. Pennisi,E. (2010) Genomics. 1000 Genomes Project gives new map of genetic diversity. *Science*, **330**, 574–575.
12. Conde,L., Vaquerizas,J.M., Dopazo,H., Arbiza,L., Reumers,J., Rousseau,F., Schymkowitz,J. and Dopazo,J. (2006) PupaSuite: finding functional single nucleotide polymorphisms for large-scale genotyping purposes. *Nucleic Acids Res.*, **34**, W621–W625.
13. De Baets,G., Van Durme,J., Reumers,J., Maurer-Stroh,S., Vanhee,P., Dopazo,J., Schymkowitz,J. and Rousseau,F. (2012) SNPeffect 4.0: on-line prediction of molecular and structural effects of protein-coding variants. *Nucleic Acids Res.*, **40**, D935–D939.
14. Myers,R.M., Stamatoyannopoulos,J., Snyder,M., Dunham,I., Hardison,R.C., Bernstein,B.E., Gingeras,T.R., Kent,W.J., Birney,E., Wold,B. *et al.* (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.*, **9**, e1001046.
15. Hawkins,R.D., Hon,G.C. and Ren,B. (2010) Next-generation genomics: an integrative approach. *Nat. Rev. Genet.*, **11**, 476–486.
16. Kalhor,R., Tjong,H., Jayathilaka,N., Alber,F. and Chen,L. (2011) Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat. Biotechnol.*, **30**, 90–98.
17. Berman,B.P., Weisenberger,D.J., Aman,J.F., Hinoue,T., Ramjan,Z., Liu,Y., Noushmehr,H., Lange,C.P., van Dijk,C.M., Tollenaar,R.A. *et al.* (2011) Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. *Nat. Genet.*, **44**, 40–46.
18. Pepke,S., Wold,B. and Mortazavi,A. (2009) Computation for ChIP-seq and RNA-seq studies. *Nat. Methods*, **6**, S22–S32.
19. Siersbaek,R., Nielsen,R. and Mandrup,S. (2012) Transcriptional networks and chromatin remodeling controlling adipogenesis. *Trends Endocrinol. Metab.*, **23**, 56–64.
20. Song,L., Zhang,Z., Grasfeder,L.L., Boyle,A.P., Giresi,P.G., Lee,B.K., Sheffield,N.C., Graf,S., Huss,M., Keefe,D. *et al.* (2011) Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Res.*, **21**, 1757–1767.
21. Bernstein,B.E., Stamatoyannopoulos,J.A., Costello,J.F., Ren,B., Milosavljevic,A., Meissner,A., Kellis,M., Marra,M.A., Beaudet,A.L., Ecker,J.R. *et al.* (2010) The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.*, **28**, 1045–1048.
22. Noushmehr,H., Weisenberger,D.J., Diefes,K., Phillips,H.S., Pujara,K., Berman,B.P., Pan,F., Pelloski,C.E., Sulman,E.P., Bhat,K.P. *et al.* (2010) Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell*, **17**, 510–522.
23. Handoko,L., Xu,H., Li,G., Ngan,C.Y., Chew,E., Schnapp,M., Lee,C.W., Ye,C., Ping,J.L., Mulawadi,F. *et al.* (2011) CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat. Genet.*, **43**, 630–638.
24. Rhee,H.S. and Pugh,B.F. (2011) Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell*, **147**, 1408–1419.
25. Chu,C., Qu,K., Zhong,F.L., Artandi,S.E. and Chang,H.Y. (2011) Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. *Mol. Cell*, **44**, 667–678.
26. Han,H., Cortez,C.C., Yang,X., Nichols,P.W., Jones,P.A. and Liang,G. (2011) DNA methylation directly silences genes with non-CpG island promoters and establishes a nucleosome occupied promoter. *Hum. Mol. Genet.*, **20**, 4299–4310.
27. Shete,S., Hosking,F.J., Robertson,L.B., Dobbins,S.E., Sanson,M., Malmer,B., Simon,M., Marie,Y., Boisselier,B., Delattre,J.Y. *et al.* (2009) Genome-wide association study identifies five susceptibility loci for glioma. *Nat. Genet.*, **41**, 899–904.
28. Wrensch,M., Jenkins,R.B., Chang,J.S., Yeh,R.F., Xiao,Y., Decker,P.A., Ballman,K.V., Berger,M., Buckner,J.C., Chang,S. *et al.* (2009) Variants in the CDKN2B and RTEL1 regions are associated with high-grade glioma susceptibility. *Nat. Genet.*, **41**, 905–908.
29. Liu,Y., Shete,S., Hosking,F., Robertson,L., Houlston,R. and Bondy,M. (2010) Genetic advances in glioma: susceptibility genes and networks. *Curr. Opin. Genet. Dev.*, **20**, 239–244.
30. Simon,M., Hosking,F.J., Marie,Y., Gousias,K., Boisselier,B., Carpentier,C., Schramm,J., Mokhtari,K., Hoang-Xuan,K., Idbaih,A. *et al.* (2010) Genetic risk profiles identify different molecular etiologies for glioma. *Clin. Cancer Res.*, **16**, 5252–5259.
31. Gentleman,R.C., Carey,V.J., Bates,D.M., Bolstad,B., Dettling,M., Dudoit,S., Ellis,B., Gautier,L., Ge,Y., Gentry,J. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.
32. Micsinai,M., Parisi,F., Strino,F., Asp,P., Dynlacht,B.D. and Kluger,Y. (2012) Picking ChIP-seq peak detectors for analyzing chromatin modification experiments. *Nucleic Acids Res.*, **40**, e70.
33. Xie,X., Mikkelsen,T.S., Gnirke,A., Lindblad-Toh,K., Kellis,M. and Lander,E.S. (2007) Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proc. Natl Acad. Sci. USA*, **104**, 7145–7150.
34. Slatkin,M. (2008) Linkage disequilibrium–understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.*, **9**, 477–485.
35. Zhu,L.J., Gazin,C., Lawson,N.D., Pages,H., Lin,S.M., Lapointe,D.S. and Green,M.R. (2010) ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics*, **11**, 237.
36. Roth,W., Isenmann,S., Nakamura,M., Platten,M., Wick,W., Kleihues,P., Bahr,M., Ohgaki,H., Ashkenazi,A. and Weller,M. (2001) Soluble decoy receptor 3 is expressed by malignant gliomas and suppresses CD95 ligand-induced apoptosis and chemotaxis. *Cancer Res.*, **61**, 2759–2765.