D968–D974 *Nucleic Acids Research, 2011, Vol. 39, Database issue* *Published online 23 October 2010*
doi:10.1093/nar/gkq997

CaSNP: a database for interrogating copy number alterations of cancer genome from SNP array data

Qingyi Cao[1,2], Meng Zhou[3], Xujun Wang[3], Cliff A. Meyer[2], Yong Zhang[3], Zhi Chen[1], Cheng Li[2,*] and X. Shirley Liu[2,*]

[1]State Key Laboratory for Diagnosis and Treatment of Infectious Diseases, The First Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, 310012, China, [2]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Harvard School of Public Health, Boston, MA 02115, USA and [3]Department of Bioinformatics, School of Life Science and Technology, Tongji University, Shanghai, 200092, China

Received August 16, 2010; Accepted October 6, 2010

ABSTRACT

Cancer is known to have abundant copy number alterations (CNAs) that greatly contribute to its pathogenesis and progression. Investigation of CNA regions could potentially help identify oncogenes and tumor suppressor genes and infer cancer mechanisms. Although single-nucleotide polymorphism (SNP) arrays have strengthened our ability to identify CNAs with unprecedented resolution, a comprehensive collection of CNA information from SNP array data is still lacking. We developed a web-based CaSNP (http://cistrome.dfci.harvard.edu/CaSNP/) database for storing and interrogating quantitative CNA data, which curated ~11 500 SNP arrays on 34 different cancer types in 104 studies. With a user input of region or gene of interest, CaSNP will return the CNA information summarizing the frequencies of gain/loss and averaged copy number for each study, and provide links to download the data or visualize it in UCSC Genome Browser. CaSNP also displays the heatmap showing copy numbers estimated at each SNP marker around the query region across all studies for a more comprehensive visualization. Finally, we used CaSNP to study the CNA of protein-coding genes as well as LincRNA genes across all cancer SNP arrays, and found putative regions harboring novel oncogenes and tumor suppressors. In summary, CaSNP is a useful tool for cancer CNA association studies, with the potential to facilitate both basic science and translational research on cancer.

INTRODUCTION

Cancer is a complex genetic disease, whose initiation and progression are often accompanied by genome alterations. A great amount of copy number alterations (CNAs) is known to occur in the malignant neoplasm at full scale of human genome. Of the different types of genome variations, CNA has been the most implicated in oncogenesis and cancer progression, and many CNAs are known to be characteristic of specific types of cancers (1–3). There is a growing demand to understand the nature of CNA in cancer, as CNAs not only serve as biomarkers to predict cancer malignancy and prognosis, but also often harbor tumor suppressors and oncogenes (4,5), the studies of which could shed light on the sequence and mechanism of oncogenesis. In addition, there is increasing evidence that some CNAs could target noncoding RNA (ncRNA) genes such as miRNAs (6), suggesting ncRNAs might be extensively involved in oncogenesis.

Array comparative genomic hybridization (aCGH) has long been the standard platform to investigate the relative gains and losses of genomic DNA by measuring the relative signal ratios of the differentially labeled array hybridization between tumor and normal samples. Several repositories focusing on CNAs detected from aCGH are

*To whom correspondence should be addressed. Tel: +1 617 632 3012, +1 617 632 3498; Fax: +1 617 632 2444; Email: xliu@hsph.harvard.edu
Correspondence may also be addressed to Cheng Li. Tel: +1 617 632 3012, +1 617 632 3498; Fax: +1 617 632 2444; Email: cli@hsph.harvard.edu
Present address:
Qingyi Cao, State Key Laboratory for Diagnosis and Treatment of Infectious Diseases, The First Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, 310012, China.

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

already publicly available (7–9). However, most of these aCGH studies use BAC or cDNA probes, which have a coarse resolution for CNA detection.

In the last few years, single nucleotide polymorphism (SNP) arrays have gradually become the major platform for SNP genotyping and CNA detection (10,11). SNP array has probes on known SNPs that are densely distributed in the human genome, and allows accurate SNP genotyping at these loci for individual biological samples. In addition, by comparing the signal intensities at the SNP loci between cancer and normal reference samples, one can also gain high-resolution CNA knowledge about the cancer of interest. It has been reported that SNP arrays outperform traditional aCGH in CNA detection resolution (12), enabling high-resolution SNP and CNA detection at individual gene level (13–15). Currently, some studies use websites to display the results of their own (e.g. http://www.broadinstitute.org/tumorscape). However, a comprehensive resource of CNA data from all cancer SNP array experiments is still unavailable.

We present CaSNP as a comprehensive collection of CNA information inferred from cancer SNP array data. We analyzed ~11 500 Affymetrix SNP arrays on 34 different cancer types in 104 studies to profile the genome-wide CNAs. This includes all the publicly available cancer SNP profiles using Affymetrix SNP arrays, mostly from Gene Expression Omnibus (GEO) (16). We also developed a data extraction and annotation schema to interrogate copy number on user-specified genomic region by cancer type and across different array platforms (from SNP 10K to 6.0) and studies. CaSNP is available at http://cistrome.dfci.harvard.edu/CaSNP/.

DESIGN AND IMPLEMENTATION

Data analysis and curation

Among the 104 studies collected, 100 are from GEO, one is from GlaxoSmithKline (https://cabig.nci.nih.gov/tools/caArray_GSKdata) and three are from individual publication's supplementary websites (17,18). The raw data (.cel file) of array experiments and accompanying genotype files (if available) for samples were collected. dCHIP-SNP (19), a widely used and referenced SNP array analysis algorithm (cited by 238 accordingly to Google Scholar), was applied to each data set. Array raw data within each study were normalized in

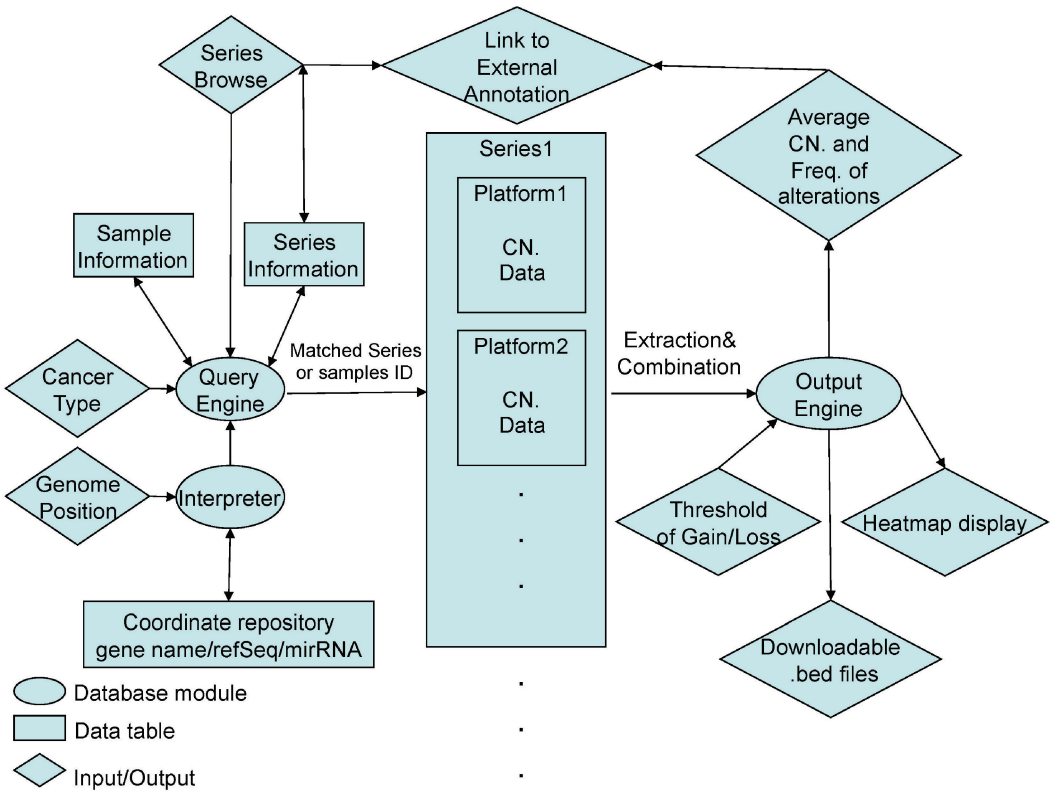
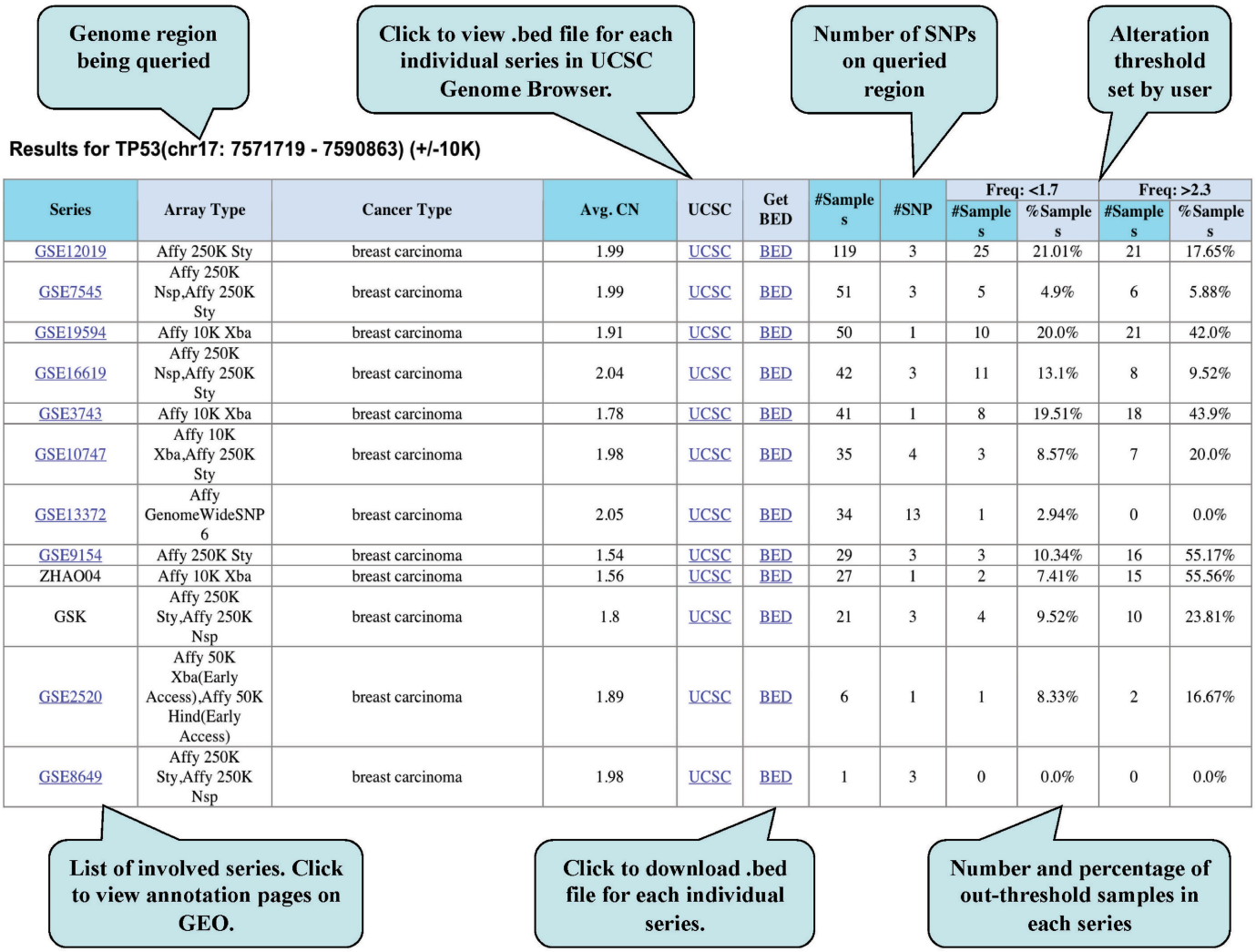


Figure 1. Overview of internal structure of CaSNP. The input for genome regions, whatever of its kind, will be uniformly translated into genomic coordinates, by querying coordinate tables of miRNA or refSeq gene, and then sent to the query engine. The input for cancer type will be checked against sample information table, to extract the names of samples qualifying this cancer type, which will further be used by the query engine to search the CNA data tables. The CNA data are stored in tables of each series and grouped by platform type. After having been extracted from data tables, relevant copy number data are combined and grouped by the output engine to calculate average copy numbers and the percentage of threshold-passing samples, which will be further displayed on the result page. Besides, a graphic display is available within which the signals of each series on the region of query will be represented as heatmaps. In addition, the returned CNA data are coordinated and written to .bed files for users to download. Detailed information for each study could be viewed on the 'Browse Data' page by linking to their corresponding annotations on GEO.



Series	Array Type	Cancer Type	Avg. CN	UCSC	Get BED	#Samples	#SNP	Freq: <1.7		Freq: >2.3	
								#Samples	%Samples	#Samples	%Samples
GSE12019	Affy 250K Sty	breast carcinoma	1.99	UCSC	BED	119	3	25	21.01%	21	17.65%
GSE7545	Affy 250K Nsp,Affy 250K Sty	breast carcinoma	1.99	UCSC	BED	51	3	5	4.9%	6	5.88%
GSE19594	Affy 10K Xba	breast carcinoma	1.91	UCSC	BED	50	1	10	20.0%	21	42.0%
GSE16619	Affy 250K Nsp,Affy 250K Sty	breast carcinoma	2.04	UCSC	BED	42	3	11	13.1%	8	9.52%
GSE3743	Affy 10K Xba	breast carcinoma	1.78	UCSC	BED	41	1	8	19.51%	18	43.9%
GSE10747	Affy 10K Xba,Affy 250K Sty	breast carcinoma	1.98	UCSC	BED	35	4	3	8.57%	7	20.0%
GSE13372	Affy GenomeWideSNP 6	breast carcinoma	2.05	UCSC	BED	34	13	1	2.94%	0	0.0%
GSE9154	Affy 250K Sty	breast carcinoma	1.54	UCSC	BED	29	3	3	10.34%	16	55.17%
ZHAO04	Affy 10K Xba	breast carcinoma	1.56	UCSC	BED	27	1	2	7.41%	15	55.56%
GSK	Affy 250K Sty,Affy 250K Nsp	breast carcinoma	1.8	UCSC	BED	21	3	4	9.52%	10	23.81%
GSE2520	Affy 50K Xba(Early Access),Affy 50K Hind(Early Access)	breast carcinoma	1.89	UCSC	BED	6	1	1	8.33%	2	16.67%
GSE8649	Affy 250K Sty,Affy 250K Nsp	breast carcinoma	1.98	UCSC	BED	1	3	0	0.0%	0	0.0%

Figure 2. A screenshot of CaSNP's query result page.

dCHIP-SNP with invariant set normalization, and signal values for individual SNP loci were further computed with the model-based expression index method (20). Relative copy number value for each SNP was calculated as the signal ratio of tumor samples versus the average of normal reference samples within the same data set, and was exported and stored in CaSNP. For data sets with no normal reference samples, the average 'normal reference' was calculated for each SNP from the tumor samples bearing the middle 50% of signals (i.e. 25% outlier signals from both sides were excluded). We did not choose normal samples from other experiments of the same array type as reference to avoid potential microarray batch effect.

To treat and query copy number data from different array platforms in a unified manner, we updated the genome coordinate system to the latest human genome assembly (UCSC hg19). In addition, all SNP IDs were converted to dbSNP129. We also manually extracted and curated information on sample clinical background and organized them at two levels: the top level on the tissue origin (e.g. lung cancer), while the second on cancer subtypes (e.g. small cell lung carcinoma).

Query parameters

The only required field in a user query is the genome region where a user inputs a genomic coordinate range (limited to 2-MB size), a gene name, a RefSeq ID or an miRNA name—all of which will be internally converted to a genomic coordinate range. The user could optionally specify the cancer type and subtype to limit the query. Alternatively, one could also go to the 'Browse Data' page to select a subset of the data sets/series for analysis. The 'Browse Data' option allows the user to focus on specific studies or conduct joint analysis in two or more studies and/or across multiple cancer types. When the user specifies a cancer type or subtype or data sets/series, CaSNP will consult the sample information table to extract the matching samples for analysis. In addition, the user can specify the upper and lower CNA thresholds (default 2.2 and 1.8, respectively) for CaSNP to calculate the percentage of samples beyond the thresholds within

each study. A flowchart depicting the internal table schema of CaSNP is shown in Figure 1.

Data output and visualization

A screenshot of CaSNP's result output page is shown in Figure 2. The most important results from a CaSNP query is the average copy number of the queried region for each of the series involved, and the percentage of samples exceeding the copy number thresholds. This value is calculated as the mean of all biological samples in each series. If there are multiple array platforms for a sample, all data for the sample will be combined before the calculation. If user specifies the upper or lower CNA threshold at input, the frequency of threshold-passing samples will also be displayed for each series. This could help the user to determine whether an observed CNA is prevalent in many samples or only caused by outlier ones. The percentage values of threshold-passing samples at the SNP loci in the region are also coded in the bedGraph file format, which is the standard for displaying continuous-valued data as a track in the UCSC genome browser. The bed files generated could be directly viewed in UCSC genome browser (21) via a link or downloaded. Also displayed on the result page are statistics of sample and SNP number for each series, links to their corresponding GEO entries at NCBI and other relevant information.

A graphic display of the results is also provided through the 'HeatMap' query page (Figure 3). The series returned are grouped by array platforms, with CNAs (loss to gain) expressed in color gradient (blue to red), and white for normal diploid (copy number 2) which gives users a comprehensive view of the copy number data in the queried region and cancer types. The heatmaps are dynamically generated from the data in the database.

Database implementation

CaSNP is running on an Apache web server and the data resides in a MySQL server. The scripts for query processing and data analysis are written in Python and the user interface is based on a django frame.

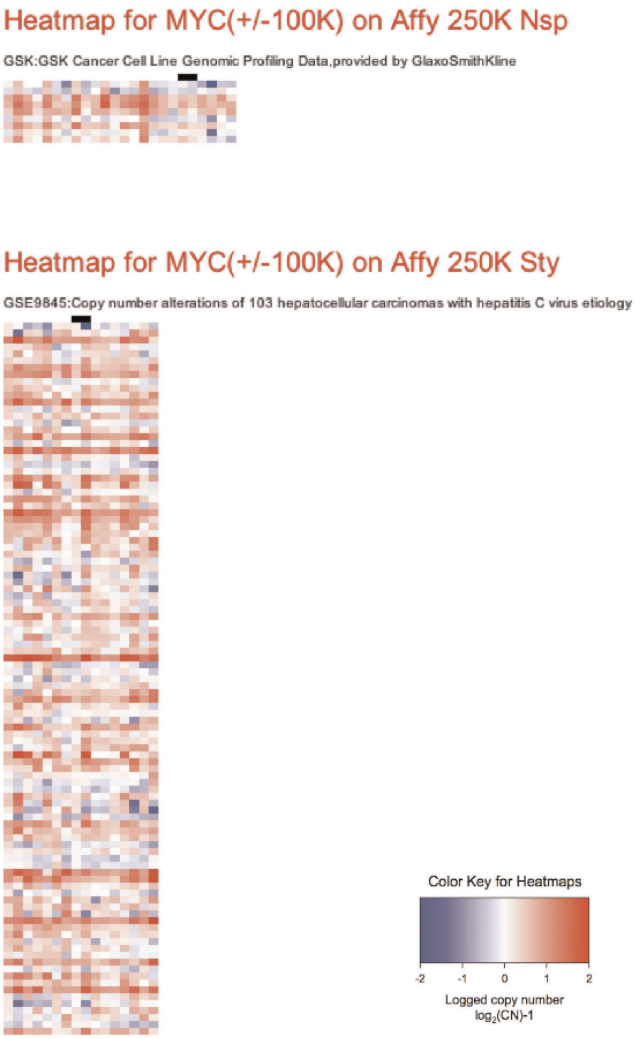


Figure 3. A screenshot of CaSNP' s heatmap query result page. Red represents higher copy number and blue represents lower copy number, and white for normal. Rows are samples involved, and columns are individual SNP markers detected by their corresponding array platforms along the queried region.

A CASE STUDY USING CaSNP

As an example of how CaSNP can be used for cancer biomarker or oncogene/tumor suppressor detection, we systematically exctracted the copy number of all 20 221 RefSeq genes from CaSNP. We then calculated a G-score, which is a component of the GISTIC methodology (22) for each gene to summarize both the frequency and amplitude of its copy number alteration in all 11 500 cancer samples. When comparing with known annotated database of oncogenes (http://www.sanger.ac.uk/genetics/CGP/Census/) and tumor suppressor genes (http://cbio.mskcc.org/CancerGenes/), we found that regions of highest or lowest G-scores often harbor known oncogenes and tumor suppressor genes, respectively (Figure 4). This partially validated the quality of the data and the accuracy of our copy number estimation. Interestingly, we observed that many chromosome ends show strong deletions in cancer, and harbor some of the well-known tumor suppressors such as STK11, TSPAN32, MAPK9 and PTGES.

A very striking exception is a strong amplified region on the left tip of chromosome 5, with no previously annotated tumor suppressors and oncogenes. The region was implicated in breast cancer risk (23), and a recent cancer CNA study (24) identified the putative target amplification gene as TERT, but did not experimentally validate its function in breast caner. Checking Oncomine (25), we found that TERT is not highly expressed in breast cancers. Instead, a nearby gene IRX2 not only shows gene amplification and enhanced expression in breast cancers, but also has some literature support for playing a role in mammary gland neoplasia (26). Alternatively, the oncogene in the Chr5 left tip might be an ncRNA, so we investigated the CNAs of all 4013 newly identified LincRNAs (27) in mammalian genomes (Supplementary Figure S1). Although gene expression of LincRNA in breast cancers is still lacking, our analysis did generate

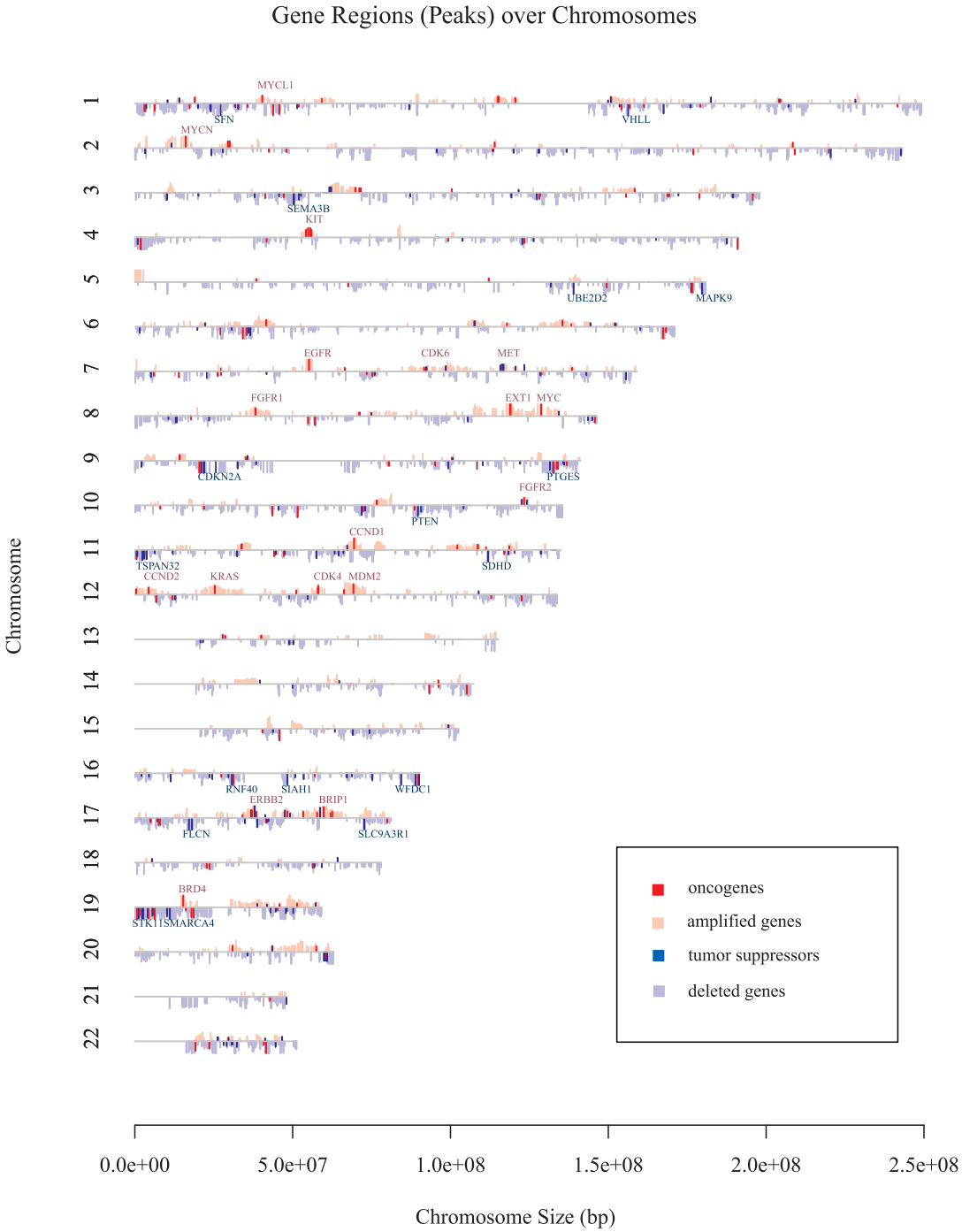


Figure 4. The distribution of amplified/deleted genes over the whole genome. The height of the bar represents the relative value of G-score. Top 50 oncogenes/tumor suppressors in G-score ranking were denoted.

interesting leads for potential follow up validations of LincRNAs as tumor suppressors and oncogenes and demonstrate the value of CaSNP.

DISCUSSION

Here, we have presented the CaSNP database for identifying and visualizing CNAs in cancers at any specific region within the human genome. CaSNP stores pre-computed raw copy numbers, and dynamically generates viewable and downloadable summaries of CNA status in response to user queries. A schema for uniformly processing, storing, annotating and presenting data sets across different data sets or platforms was successfully implemented, making CaSNP a useful tool for cancer genomic meta-study. The query results contain numerical values of cancer copy numbers and the frequencies of CNA events, which are well suited for more detailed analysis

by other software or methods. Besides the tabular display, the heatmap view displays SNP copy numbers in colors, enabling users to intuitively and comprehensively visualize the results and facilitating finding novel CNA regions in subset of samples. Besides, we provided a scenario of using CaSNP to explore cancer biomarkers or genes through a meta-analysis, and proved CaSNP's ability in suggesting novel oncogenes/tumor suppressors, whether a protein coding gene or a ncRNA.

Benefited from the abundance of SNP array data sets in recent years, CaSNP is the largest repository of SNP array-oriented CNA data among all the databases of the similar type. The amount of public-accessible SNP array data on cancer is still expanding, so will be the data collection in CaSNP. Such a large-scale analysis will be extremely valuable when correlating CNA data with a genomic location with specific diagnostic, prognostic or therapeutic value found in other studies, or to reduce noise from individual studies via meta-analysis. Nowadays, when high-throughput methods as ChIP–chip or ChIP-seq could generate hundreds of thousands of regions of interest in a single run, CaSNP will be powerful for independent validation purpose, such as screening the regions which might be related to oncogenesis and might go unnoticed in ChIP experiments alone. Besides collecting more data, we will commit our work to make better use of them. The loss-of-heterozygosity (LOH) information deduced from genotype data will be added, and the CNA status will be compared across different cancer types for specified regions and across the genome.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We greatly appreciate Yi Wang, Scott Taing, Len Taing, Tao Liu and Luhua Zhang's help on the design and deployment of CaSNP.

FUNDING

The National Institutes of Health R01 (HG004069 to X.S.L., GM077122 to C.L.); Chinese Scholarship Council (2008632067 to Q.C.); State S & T Projects (11th Five Year) of China (2008ZX10002-007 to Z.C.); National Basic Research Program of China (973 Program No. 2010CB944904 to M.Z., X.W. and Y.Z.). Funding for open access charge: National Basic Research Program of China (973 Program No. 2010CB944904).

Conflict of interest statement. None declared.

REFERENCES

1. Myllykangas,S., Böhling,T. and Knuutila,S. (2007) Specificity, selection and significance of gene amplifications in cancer. *Semin. Cancer Biol.*, **17**, 42–55.
2. Stark,M. and Hayward,N. (2007) Genome-wide loss of heterozygosity and copy number analysis in melanoma using high-density single-nucleotide polymorphism arrays. *Cancer Res.*, **67**, 2632–2642.
3. Weir,B.A., Woo,M.S., Getz,G., Perner,S., Ding,L., Beroukhim,R., Lin,W.M., Province,M.A., Kraja,A., Johnson,L.A. *et al.* (2007) Characterizing the cancer genome in lung adenocarcinoma. *Nature*, **450**, 893–898.
4. Hu,X., Stern,H.M., Ge,L., O'Brien,C., Haydu,L., Honchell,C.D., Haverty,P.M., Peters,B.A., Wu,T.D., Amler,L.C. *et al.* (2009) Genetic alterations and oncogenic pathways associated with breast cancer subtypes. *Mol. Cancer Res.*, **7**, 511–522.
5. Tchatchou,S. and Burwinkel,B. (2008) Chromosome copy number variation and breast cancer risk. *Cytogenet. Genome Res.*, **123**, 183–187.
6. Calin,G.A., Sevignani,C., Dumitru,C.D., Hyslop,T., Noch,E., Yendamuri,S., Shimizu,M., Rattan,S., Bullrich,F., Negrini,M. *et al.* (2004) Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. *Proc. Natl Acad. Sci. USA*, **101**, 2999–3004.
7. Baudis,M. and Cleary,M.L. (2001) Progenetix.net: an online repository for molecular cytogenetic aberration data. *Bioinformatics*, **17**, 1228–1229.
8. Knutsen,T., Gobu,V., Knaus,R., Padilla-Nash,H., Augustus,M., Strausberg,R.L., Kirsch,I.R., Sirotkin,K. and Ried,T. (2005) The interactive online SKY/M-FISH & CGH database and the Entrez cancer chromosomes search database: linkage of chromosomal aberrations with the genome sequence. *Genes Chromosomes Cancer*, **44**, 52–64.
9. Scheinin,I., Myllykangas,S., Borze,I., Böhling,T., Knuutila,S. and Saharinen,J. (2008) CanGEM: mining gene copy number changes in cancer. *Nucleic Acids Res.*, **36**, D830–D835.
10. Heinrichs,S. and Look,A.T. (2007) Identification of structural aberrations in cancer by SNP array analysis. *Genome Biol.*, **8**, 219.
11. Carter,N.P. (2007) Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat. Genet.*, **39**, S16–S21.
12. Bignell,G.R., Huang,J., Greshock,J., Watt,S., Butler,A., West,S., Grigorova,M., Jones,K.W., Wei,W., Stratton,M.R. *et al.* (2004) High-resolution analysis of DNA copy number using oligonucleotide microarrays. *Genome Res.*, **14**, 287–295.
13. Dutt,A. and Beroukhim,R. (2007) Single nucleotide polymorphism array analysis of cancer. *Curr. Opin. Oncol.*, **19**, 43–49.
14. Beaudet,A.L. and Belmont,J.W. (2008) Array-based DNA diagnostics: let the revolution begin. *Annu. Rev. Med.*, **59**, 113–129.
15. Leary,R.J., Lin,C., Cummins,J., Boca,S., Wood,L.D., Parsons,D.W., Jones,S., Sjöblom,T., Park,B.H., Parsons,R. *et al.* (2008) Integrated analysis of homozygous deletions, focal amplifications, and sequence alterations in breast and colorectal cancers. *Proc. Natl Acad. Sci. USA*, **105**, 16224–16229.
16. Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.
17. Zhao,X., Li,C., Paez,J.G., Chin,K., Jänne,P.A., Chen,T.H., Girard,L., Minna,J., Christiani,D., Leo,C. *et al.* (2004) An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Res.*, **64**, 3060–3071.
18. Zhao,X., Weir,B.A., LaFramboise,T., Lin,M., Beroukhim,R., Garraway,L., Beheshti,J., Lee,J.C., Naoki,K., Richards,W.G. *et al.* (2005) Homozygous deletions and chromosome amplifications in human lung carcinomas revealed by single nucleotide polymorphism array analysis. *Cancer Res.*, **65**, 5561–5570.
19. Lin,M., Wei,L.J., Sellers,W.R., Lieberfarb,M., Wong,W.H. and Li,C. (2004) dChipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data. *Bioinformatics*, **20**, 1233–1240.
20. Li,C. and Wong,W.H. (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA*, **98**, 31–36.

21. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
22. Beroukhim,R., Getz,G., Nghiemphu,L., Barretina,J., Hsueh,T., Linhart,D., Vivanco,I., Lee,J.C., Huang,J.H., Alexander,S. *et al.* (2007) Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc. Natl Acad. Sci. USA*, **104**, 20007–20012.
23. Vachon,C.M., Sellers,T.A., Carlson,E.E., Cunningham,J.M., Hilker,C.A., Smalley,R.L., Schaid,D.J., Kelemen,L.E., Couch,F.J. and Pankratz,V.S. (2007) Strong evidence of a genetic determinant for mammographic density, a major risk factor for breast cancer. *Cancer Res.*, **67**, 8412–8418.
24. Beroukhim,R., Mermel,C.H., Porter,D., Wei,G., Raychaudhuri,S., Donovan,J., Barretina,J., Boehm,J.S., Dobson,J., Urashima,M. *et al.* (2010) The landscape of somatic copy-number alteration across human cancers. *Nature*, **463**, 899–905.
25. Rhodes,D.R., Kalyana-Sundaram,S., Mahavisno,V., Varambally,R., Yu,J., Briggs,B.B., Barrette,T.R., Anstet,M.J., Kincead-Beal,C., Kulkarni,P. *et al.* (2007) Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia*, **9**, 166–180.
26. Lewis,M.T., Ross,S., Strickland,P.A., Snyder,C.J. and Daniel,C.W. (1999) Regulated expression patterns of IRX2, and Iroquois-class homeobox gene, in the human breast. *Cell Tissue Res.*, **296**, 549–554.
27. Huarte,M., Guttman,M., Feldser,D., Garber,M., Koziol,M.J., Kenzelmann-Broz,D., Khalil,A.M., Zuk,O., Amit,I., Rabani,M. *et al.* (2010) A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell*, **142**, 409–419.