# BIOINFORMATICS

# dChipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data

Ming Lin[1,2], Lee-Jen Wei[1,2], William R. Sellers[3,4],
Marshall Lieberfarb[5], Wing Hung Wong[1,2,6] and Cheng Li[1,2,*]

[1]Department of Biostatistics, Harvard School of Public Health, [2]Department of
Biostatistical Sciences and [3]Department of Medical Oncology, Dana-Farber Cancer
Institute, [4]Department of Medicine, Harvard Medical School, [5]Department of Radiation
Oncology, Brigham and Women's Hospital, Boston, MA 02115, USA and
[6]Department of Statistics, Harvard University, Cambridge, MA, USA

## ABSTRACT

**Motivation:** Oligonucleotide microarrays allow genotyping of thousands of single-nucleotide polymorphisms (SNPs) in parallel. Recently, this technology has been applied to loss-of-heterozygosity (LOH) analysis of paired normal and tumor samples. However, methods and software for analyzing such data are not fully developed.

**Result:** Here, we report automated methods for pooling SNP array replicates to make LOH calls, visualizing SNP and LOH data along chromosomes in the context of genes and cytobands, making statistical inference to identify shared LOH regions, clustering samples based on LOH profiles and correlating the clustering results to clinical variables. Application of these methods to prostate and breast cancer datasets generates biologically important results.

**Availability:** The software module dChipSNP implementing these methods is available at http://biosun1.harvard.edu/complab/dchip/snp/

**Contact:** cli@hsph.harvard.edu

**Supplementary information:** The breast cancer data are provided by Andrea L. Richardson, Zhigang C. Wang and James D. Iglehart.

## INTRODUCTION

Oligonucleotide microarrays have been used widely to generate data for gene expression analysis (Lipschutz *et al.*, 1999; Li and Wong, 2001). This technique has also been used to detect genetic variations of single nucleotide polymorphisms (SNPs; Chee *et al.*, 1996; Wang *et al.*, 1998; Cargill *et al.*, 1999) and to link SNPs to complex human diseases and drug susceptibilities (Hacia *et al.*, 1999; Halushka *et al.*, 1999). Recently, oligonucleotide-based SNP arrays (HuSNP, Affymetrix Inc., 2000) containing 1494 human SNP markers have been used to identify loss-of-heterozygosity (LOH) of chromosomal regions based on paired normal and tumor samples

from the same patient (Mei *et al.*, 2000; Lindblad-Toh *et al.*, 2000; Schubert *et al.*, 2001).

Affymetrix genotyping software analyzes the scanned image data of SNP arrays and generates SNP calls (Cutler *et al.*, 2001). The SNP calls of paired normal and tumor samples can then be combined to make LOH calls in dChipSNP (Fig. 1). The existing methods in the literature for analyzing such LOH data are largely exploratory. In this paper, we present the quantitative methods and software specifically developed to analyze SNP-array-based LOH data, which include automated reading SNP calls from SNP call text files, pooling SNP array replicates, making LOH calls, making statistical inference for identifying shared LOH regions and using LOH profiles for sample clustering.

In this paper, we use the dataset in Lieberfarb *et al.* (2003) to illustrate the methods and software. For this dataset, there are 176 Affymetrix HuSNP arrays hybridized to the normal and tumor samples from 52 prostate cancer patients. Replicate arrays are hybridized to most tumor samples to alleviate the effects of normal sample contaminations and to obtain more accurate SNP calls. The replicate arrays are at split-*in vitro* transcription (IVT) or split-DNA level. The data are available at http://biosun1.harvard.edu/complab/dchip/snp/

## SYSTEMS AND METHODS

### dChipSNP software module

We develop a dChipSNP module based on the dChip software (Li and Wong, 2001) to perform the aforementioned SNP-array-based LOH analysis. The HuSNP arrays have a format similar to that of oligonucleotide expression arrays, and thus the existing functions in dChip such as 'CEL Image View' and 'PM/MM Data View' are immediately available for HuSNP array analysis. After dChipSNP reads in HuSNP CEL files and the corresponding SNP call files, the array images and probe intensity data for individual SNPs can be visualized. Then the normal and tumor SNP calls are combined to make

---

*To whom correspondence should be addressed.
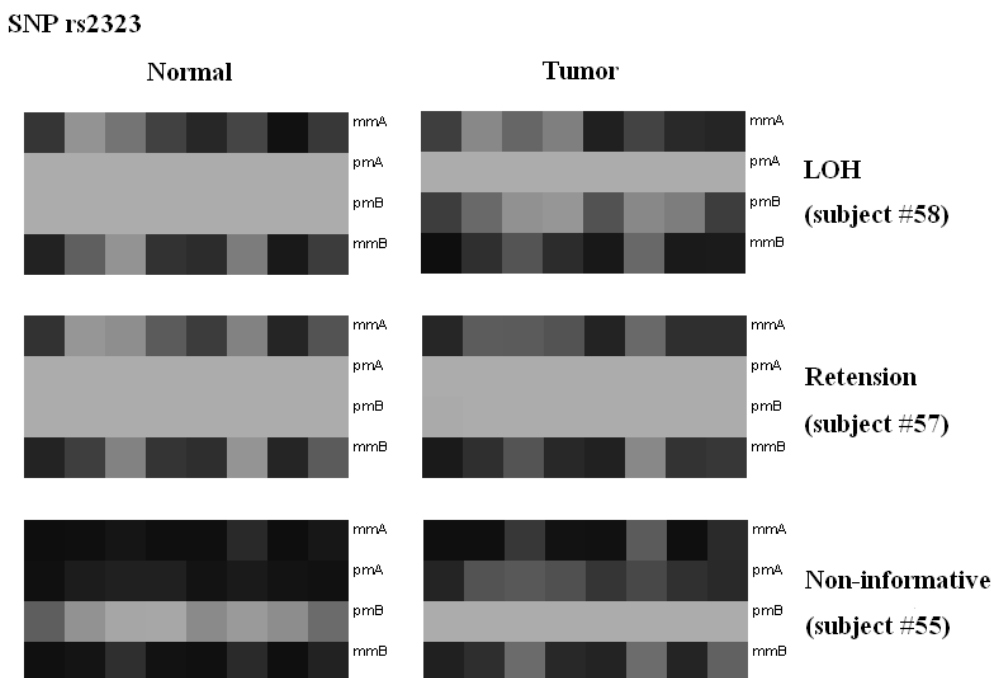
**Fig. 1.** The probe intensity data of SNP marker rs2323 of three subjects are displayed on three rows. This SNP genotype is interrogated by eight mini-blocks (each consisting of four vertically aligned mmA, pmA, pmB and mmB probes), and they contain 20 bp oligonucleotides complementary to the reference sequences covering the SNP position. The four vertical probes in a mini-block have the same sequence except the central position, where four different nucleotides of A, T, G, C are placed to distinguish the SNP genotype. Based on the probe intensity patterns, Affymetrix software makes an A, B or AB call, and the SNP calls of a pair of normal and tumor samples for the same patient can be used to infer the LOH status of the tumor sample at this SNP position. The colour version of this figure can be obtained at http://biosun1.harvard.edu/complab/dchip/snp/

**Table 1.** Making LOH calls based on the SNP calls of paired normal and tumor samples of the same individual

| LOH call | | Tumor SNP call | | | |
| --- | --- | --- | --- | --- | --- |
| | | A | B | AB | No call |
| Normal SNP call | A | Non-informative | No call | No call | Non-informative |
| | B | No call | Non-informative | No call | Non-informative |
| | AB | Loss | Loss | Retention | No call |
| | No call | No call | No call | Retention | No call |

LOH calls as described in Table 1. If sample replicates exist, the replicate SNP calls are pooled by the 'Majority-Voting' scheme before making LOH calls (described in the 'Pooling replicate arrays and making LOH calls' section).

In the dChipSNP 'Chromosome View', users can choose to display the LOH calls (Fig. 2) and inferred LOH calls (Fig. 3, described in the 'Inferring non-informative markers' section) or the original SNP calls. One may also enlarge a chromosomal region with shared LOH events to see the genes located in this region. Moreover, users may search a particular gene and examine whether some tumors have LOH in the nearby regions. The software and the manual can be obtained at http://biosun1.harvard.edu/complab/dchip/snp/

## Significance curve for shared LOH regions

After obtaining and visualizing LOH calls, we are often interested in defining the regions of LOH loss shared by multiple tumors because such regions are likely to contain tumor-suppressor genes. But this is typically done by simple methods such as visualization. Here, we use permutation methods to answer the following questions: where are the significant shared LOH regions, and how likely is an observed shared LOH region due to chance? The resulting *p*-value curves are displayed next to the LOH data to help investigators locate interesting shared LOH regions (Fig. 2).

Specifically, for a particular chromosomal region, we define a score for each individual to quantify the region's likelihood of being 'Loss'. The scores of all individuals are then summed up to give a summary score for this chromosomal region. Suppose all the observed LOH events are due to call errors and thus are not cancer-related; then the paired normal and tumor samples are conceptually indistinguishable, and the observed differences between them represent the background noise from which we would like to distinguish the real LOH
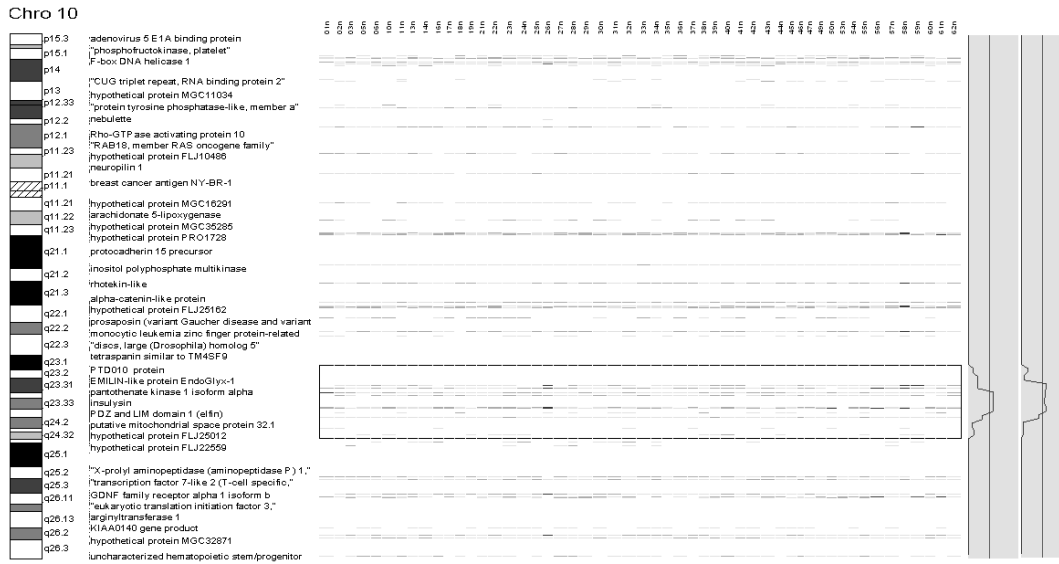
**Fig. 2.** The 'Chromosome View' in dChipSNP. The SNP markers in chromosome 10 are drawn proportional to their chromosomal positions. The dark colour represents observed 'Loss', the light colour represents 'Retention' and 'Non-informative' LOH status, and the white color indicates 'No Call' or no markers. Each row represents a SNP marker, and each column represents a patient. The genes and cytobands are displayed on the left, with small blue dots representing omitted names due to limited space. One can zoom in to view more details. The blue curves on the right side are the multiple-testing adjusted $p$-value curves ($-\log 10$ transformed) using a simple score (left) and HMM score (right). The region with a curve peak exceeding the user specified significant threshold (0.05 in this picture, shown with vertical red line) is considered as a significant region of shared LOH events. A larger 14 Mb region is outlined in a black rectangle in the LOH data area since the LOH data in this region are used to compute the $p$-value for the chromosomal region exceeding the threshold. The colour version of this figure can be obtained at http://biosun1.harvard.edu/complab/dchip/snp/

events. Therefore, we can simulate the background noise by permuting the paired normal and tumor samples. Specifically, for each individual, we randomly assign one of the paired samples as the tumor sample and treat the other as the normal. We then compare the LOH events in the original data with the LOH events in a large number of such simulated data to assess the statistical significance of the former. This permutation method can be applied for any reasonable scores, and we propose two scoring methods here.

*Permutation using simple scores.* For a SNP marker at the chromosomal position $t$-megabases, we define $C_i(t) = 1$ for the $i$-th individual if 'Loss' is observed, $-1$ if the normal sample has a homozygous SNP call but the tumor sample has a heterozygous SNP call (this is most likely due to measurement error) and 0 if 'Retention' or 'Non-informative' is observed. We also define $D_i(t) = 0$ if this SNP is 'Non-informative' and 1 otherwise. LOH NO_With the observed data $\{[C_i(t), D_i(t)], i = 1, \Lambda, N, 0 \leq t \leq L\}$, where $L$ is the length of the chromosome in megabase and N is the number of individuals, we consider a summary score $R(x)$ for the chromosomal region $(x - b, x + b)$,

$$R(x) = \sum_{i=1}^{N} \frac{\sum_{\{x-b \leq t \leq x+b\}} C_i(t)}{\sum_{\{x-b \leq t \leq x+b\}} D_i(t)}, \quad b \leq x \leq L - b.$$

The $i$-th summand in $R(x)$ can be viewed as the proportion of 'Loss' events among all informative markers in this region for the $i$-th individual, with penalty given to measurement errors and intervening 'Retention' markers. We use the proportion of 'Loss' markers rather than the actual counts of 'Loss' markers to partially alleviate the influence of different marker densities at different chromosomal regions.

Under the null hypothesis that there are no real 'Loss' regions for the entire chromosome (all the observed 'Loss' are assumed to come from measurement error), one can generate the null distribution of $R(x)$ by permuting the paired SNP samples and then obtain a simulated $R(x)$ value ($b \leq x \leq L - b$) based on the permuted dataset. From a large number of such permutations, we obtain the estimate for the null distribution of $R(x)$ and the raw $p$-value of a specific region $(x - b, x + b)$, which is the proportion of the permuted $R(x)$ that are equal or greater than the observed $R(x)$.

We then use either the maxT procedure (Westfall and Young, 1993) or the false discovery rate (FDR) controlling procedure (Benjamini and Hochberg, 1995) to adjust the $p$-values for multiple testing. The maxT procedure is performed as follows: for each permuted dataset, we obtain $\text{MAX}_{b \leq y \leq L-b} R(y)$ (i.e. the maximum score among all the regions on the genome or a chromosome), and the adjusted $p$-value of a specific region $(x - b, x + b)$ is the proportion of the permuted $\text{MAX}_{b \leq y \leq L-b} R(y)$ that are greater than the observed $R(x)$.
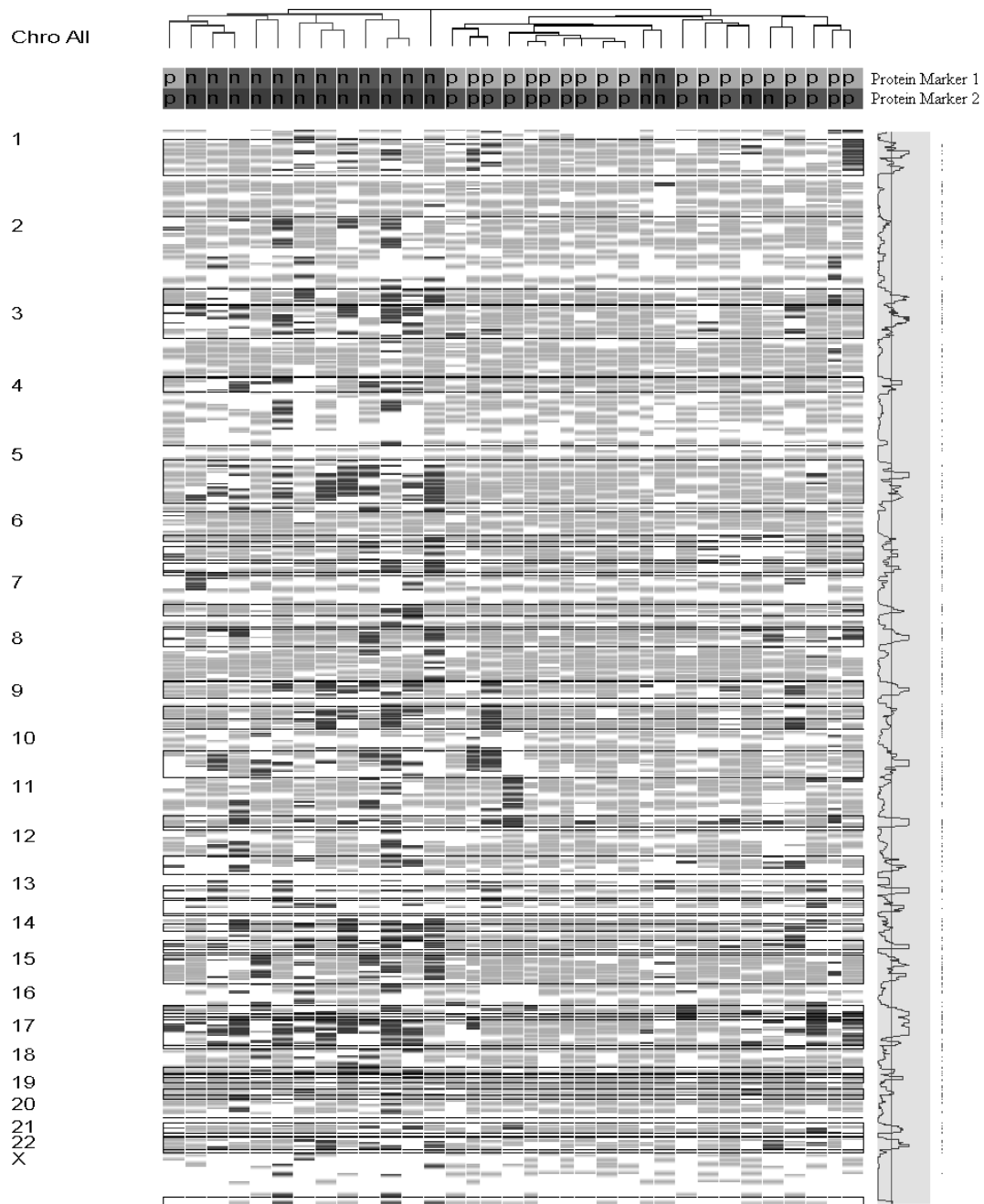
**Fig. 3.** Whole-genome LOH patterns of breast cancer data. Right: the multiple-testing adjusted *p*-value curve (−log 10 transformed) using the simple score method. Top: the sample clustering tree based on the LOH data in the significant regions and the colored (n: negative, p: positive) clinical status of two protein markers. In this figure, the non-informative markers are inferred by the 'Regions with Same Boundary' method for better visualization. The default extension size of 10 Mb is used here. The colour version of this figure can be obtained at http://biosun1.harvard.edu/complab/dchip/snp/

The first curve on the right-hand side of Figure 2 is the *p*-value curve generated by applying the above method with maxT adjustment (the maximum is taken over the whole genome). Here, we use $b = 7$ Mb and discretize $x$ by increments of 1 Mb. That is, for each chromosome, we move a window 14 Mb in length from one end to the other in 1 Mb step. Each window overlaps with several of its neighboring windows, and therefore the *p*-values for these overlapping windows are positively correlated. The window size parameter, $b$, can be adjusted in dChipSNP to tune to datasets using particular

tissues or arrays. The significant region of shared LOH shown in Figure 2 harbors the known PTEN tumor suppressor gene, and we discuss more of its biological implication elsewhere (Lieberfarb *et al.*, 2003). We also use the raw *p*-values and the FDR controlling procedure to find the *p*-value threshold that corresponds to the nominal FDR. However, in our application, this procedure is conservative in that it controls FDR at a level lower than the nominal level because the *p*-values are positively correlated.

*Permutation using hidden Markov model (HMM) scores.* LOH events on chromosomes are spatially correlated because the chromosomal locations near a known LOH locus are very likely to be LOH, and this likelihood decreases for farther chromosomal locations. A Markov chain is suitable to model this spatial correlation among the unobserved real LOH status of a chromosome.

We use the HMM to derive a more sophisticated score to capture such underlying biological process of real LOH events. For each individual, the unobserved real LOH status ('Loss' or 'Retention') of each SNP within a specified region of the chromosomes can be modeled by a bi-directional Markov chain. Given a chromosomal position $x$ Mb and its neighboring region $(x - b, x + b)$, we have the LOH status of $n$ SNPs in the region $(x - b, x)$ and $m$ SNPs in the region $(x, x + b)$. For each individual, we denote the real LOH status of position $x$ by $y_0$, and the real LOH status of the $n$ and $m$ SNPs on either side of $x$ by $y_1, y_2, \Lambda, y_n$ and $y_1', y_2', \Lambda, y_m'$, where $y_i$ (or $y_i'$) $= 1$ ('Loss') or $-1$ ('Retention'). We also denote their observed LOH status by $z_1, z_2, \Lambda, z_n$ and $z_1', z_2', \Lambda, z_m'$, where $z_i$ (or $z_i'$) $= 1$ ('Loss'), $-1$ ('Retention') or $0$ ('Non-informative'). In the above notations, the ones with smaller subscripts are closer to the center, $x$.

In the HMM, the prior probability of LOH 'Loss' at position $x$ is $p_0 = P(y_0 = 1) = 1 - P(y_0 = -1)$, which is estimated by the overall 'Loss' rate among all informative markers in the data. The emission probabilities (the distribution of $z_i$ conditioned on $y_i$) reflect the probabilities of observing correct 'Loss' and 'Retention' calls, 'Non-informative' calls and measurement errors. We estimate them by Bayesian estimators of the 'Non-informative' probabilities and the conflict call rate in region $(x - b, x + b)$. The transition probabilities are modeled to depend on the distance between two neighboring markers:

$$\frac{P(y_i = 1 | y_{i-1})}{P(y_i = -1 | y_{i-1})} = \left( \frac{p_0}{1 - p_0} \right) \exp \left( \frac{\beta y_{i-1}}{d_{i-1,i}} \right),$$

where $\beta$ is positive and is the same for all individuals and all chromosomal regions, and $d_{i-1,i}$ is the distance between SNP $i - 1$ and $i$ in megabases. Note that as $d_{i-1,i} \to \infty$, $P(y_i = 1 | y_{i-1}) = P(y_i = 1) = p_0$ and that as $d_{i-1,i} \to 0$, $P(y_i = 1 | y_{i-1} = 1) = 1$ and $P(y_i = 1 | y_{i-1} = -1) = 0$. Thus, the transition probabilities agree with our intuition that close chromosomal positions tend to have the same

LOH status, while far away positions have independent LOH status. Similar emission probabilities and transition probabilities are used for the $m$ SNPs in the region $(x, x + b)$. To estimate the parameter $\beta$, we model the LOH status of all the SNP markers on each chromosome by a HMM with the same HMM probabilistic settings and estimate $\beta$ by the maximum-likelihood method. After we get the initial, emission and transition probabilities, the marginal probability, $P(z_1, \Lambda, z_n, z_1', \Lambda, z_m' | \text{LOH})$, can then be calculated by the forward–backward algorithm (Durbin *et al.*, 1998).

Under the null hypothesis, we hypothesize that there is no real LOH event within the region $(x - b, x + b)$ for this patient, and the observed LOH events are due to measurement errors. The likelihood of the data under the null hypothesis can be calculated as a special case of the HMM with all real LOH status being 'Retention'. The HMM-based score of region $(x - b, x + b)$ for the *i*-th individual is then defined as the log-likelihood ratio,

$$R_i(x) = \log \left[ \frac{P(z_1, \Lambda, z_n, z_1', \Lambda, z_m' | \text{LOH})}{P(z_1, \Lambda, z_n, z_1', \Lambda, z_m' | \text{Null})} \right],$$

and the overall score of the region $(x - b, x + b)$ as $R(x) = \sum_{i=1}^{N} R_i(x)$. To avoid $R(x)$ being driven by a single individual, it is truncated if it is higher than a pre-specified cutoff.

Permutation can then be performed in the same manner as in the previous section. The second curve on the right-hand side of Figure 2 shows the *p*-value curve using HMM scores. It identifies the same shared LOH region as the simple score method. For this dataset, we find that these two scoring methods identify similar shared LOH regions across the whole genome.

Considering the sparseness of LOH events in the prostate cancer data, we also test the permutation method on independent breast cancer data where LOH events occur much more frequently (Wang *et al.*, 2003). The *p*-value curve for all chromosomes generated using the simple score method is shown in Figure 3. The curve is able to capture the regions where LOH events occur across multiple tumors, and therefore, it can help investigators to focus on regions that are most likely to be really involved in the underlying biological process of tumor formation. In addition, filtering out non-significant regions improves the result of sample clustering by reducing the noise in the data, which is discussed in the next section.

## Sample clustering based on significant LOH regions

Researchers are often interested in the co-occurrence of LOH events, or subclasses of tumor samples harboring similar LOH events across the genome. To this end, we applied the hierarchical clustering algorithm (Eisen *et al.*, 1998) to tumor samples using LOH data of one chromosome or all chromosomes. We find that when using the data of all the chromosomal regions

for clustering, the result tends to be driven by the 'Retention' patterns in the non-significant regions. So, we perform hierarchical clustering using only the LOH data in the identified significant LOH regions. We make LOH call for each of the significant regions in each individual: an individual is classified as 'Loss' if there are one or more 'Loss' SNP markers in the region, 'Retention' if there is no 'Loss' SNP marker but one or more 'Retention' SNP markers in the region and 'Non-informative' if all the SNP markers in the region are 'Non-informative'. The distance between any two individuals is defined as the proportion of discordant regions among all the significant regions for which both individuals have informative LOH calls. The average-linkage algorithm is used to merge samples and clusters of samples during the clustering procedure. This method is applied to the breast cancer dataset, and Figure 3 shows the result.

After the samples are clustered based on LOH profiles, we can correlate the clustering results with the sample clinical information. There are two main clusters in Figure 3: one of them contains 12 patients, among which 11 are negative for protein markers 1 and 2 (branch highlighted in blue color), while the other cluster contains 21 patients whose status are mostly positive for protein markers 1 and 2. This suggests that there is an association between patients' LOH pattern and their status of these two protein markers (Fisher's exact test, $p < 0.001$ for each cluster, dChipSNP automatically performs such tests for all clusters). Further investigation of the shared LOH regions specific to both sample clusters and genes contained in these regions may reveal the biological underpinning of the relationship between LOH defined clusters and clinical variables and will be presented elsewhere (Wang *et al.*, 2003).

## Pooling replicate arrays and making LOH calls

The possible SNP calls made by Affymetrix genotyping software are A, B, AB, AB_A (meaning the allele type is either AB or A), AB_B and 'No call'. There may be inconsistent calls obtained for the replicate arrays hybridized to the same tumor sample. It is time consuming to resolve such inconsistence by visually checking the array images (Fig. 1).

We adopt a 'Majority-Voting' scheme to determine the pooled SNP call of a sample based on all the replicates. Unambiguous observed SNP calls A, B and AB vote 1 for themselves only, while ambiguous SNP call AB_A (or AB_B) votes 0.5 for AB and 0.5 for A (or B). As an example, for the calls 'AB_A, A, A' of a SNP in three replicates, the final vote for (A, B, AB) is (2.5, 0, 0.5). We then define the 'Voting Score' (VS) as the positive difference between the largest two of the three votes for (A, B, AB). The pooled SNP call is the one in (A, B, AB), with the largest vote if VS $\geq 1$ and 'No call' if VS $< 1$. For the above example, VS = 2, and the pooled call is A. When there is no replicate (such as the normal samples), the net effect of this method is to regard the observed unambiguous call as real call and declare 'AB_A' and 'AB_B' as 'No call'.

**Table 2.** The three percentages used to assess the 'Majority-Voting' method when pooling two or more replicate SNP calls

| % | 52 single tumors | 1 single tumor, 51 duplicate tumors | 1 single tumor, 30 duplicate tumors, 21 triplicate tumors |
|---|---|---|---|
| I | 20.7 | 15.5 | 14.0 |
| II | 1.22 | 0.30 | 0.35 |
| III | 7.1 | 11.6 | 12.5 |

Tumor 'No call' percentage: the percentages of 'No call' in tumor samples; Conflict percentage (II): when the normal call is A or B, among the tumor calls not equal to 'No call', the percentage of tumor calls that are in conflict with the corresponding normal call (e.g. normal call is A but tumor call is AB or B); Retention inference percentage (III): when the normal call is 'No call' (in this dataset, 18.1% of normal calls are 'No call'), the percentage of tumor calls is equal to AB. In such a case, we can still infer the 'Retention' status of a SNP marker in the tumor sample.

We use three percentages to assess how pooling replicates helps the analysis and by what magnitude (Table 2 legend). All percentages are computed using the data of all patients after applying the 'Majority-Voting' method to normal and tumor samples. The three percentages when not using tumor replicates and using tumor duplicates or triplicates are shown in Table 2. As we would have expected, using duplicates decreases percentages for I and II while increasing the percentage for III. However for percentages II and III, triplicating tumor samples does not have as much improving effect as duplicating them.

After pooling replicates to make pooled, we use the rules in Table 1 to make LOH calls from the SNP calls of paired normal and tumor samples. These LOH calls are the main data used in the aforementioned analysis.

## Inferring non-informative markers

It is often useful to infer the true status of the 'Non-informative' calls. Lindblad-Toh *et al.* (2000) adopted a simple extension method. The drawback of this method is that it does not consider the relative chromosomal positions of the SNP markers. We implement the 'Nearest Neighbor' and 'Regions with Same Boundary' methods in dChipSNP to infer the LOH status of 1 Mb apart pseudo markers along the whole chromosome. The 'Nearest Neighbor' method infers the LOH status of a pseudo marker as the LOH status of its nearest informative real marker. For the 'Regions with Same Boundary' method, the LOH status of all pseudo markers bounded by two real markers with the same LOH status ('Loss' or 'Retention') are inferred as the LOH status of its two boundaries and are not inferred ('Non-informative') if they are not bounded in this way. The color intensities of inferred pseudo markers decline to the white color as their distances from the nearest real markers increase, and so the credibility of the inferred LOH calls can be visualized (Fig. 3). We also specify an extension limit (10 Mb as the default) so that pseudo markers are

not inferred if their distances to the closest informative real markers are larger than this distance.

## DISCUSSIONS

In this paper, we developed several methods for SNP-array-based LOH data analysis: pooling SNP calls from replicate arrays and making LOH calls, visualizing LOH data, identifying shared LOH regions by statistical significance, clustering samples based on the identified shared LOH regions and correlating LOH-based sample clusters with clinical variables. The next generation Affymetrix SNP arrays consist of much denser (11 500 SNP) markers, and they have a higher heterozygosity rate on average. We have found that most methods presented here can be readily applied to the data generated by the new SNP arrays.

There are many directions deserving further studies. First, our two methods for identifying shared LOH regions do not use the probe level intensity data (Fig. 1). This is part of the reason why the more sophisticated HMM score method does not render a superior result over the simple score method. An HMM-type model on the probe level data may make better use of the information in the data and give better results. Second, LOH-based sample clustering is unsupervised, and as the data accumulate, we may develop a supervised classification method to predict tumor subtypes or survival time based on LOH patterns, in a way similar to classification based on gene expression data (Golub *et al.*, 1999). Third, in many LOH studies, the tumor samples are not homogeneous but vary in a set of clinical behaviors. It will be useful to develop a statistical method that can automatically identify regions that exhibit different LOH patterns in different subgroups defined by clinical behaviors with adjustment for potential confounders. Lastly, sometimes in addition to SNP data, we may also have data generated by gene expression microarrays and comparative genomic hybridization for the same set of samples. How to properly integrate all these related genomics data to identify chromosomal changes and gene regulations underlying diseases is an exciting and challenging problem. We will be actively working on these aspects and will report the progress in future work.

## ACKNOWLEDGEMENTS

## REFERENCES

Affymetrix Inc. (2000) GeneChip® HuSNP™ Mapping Assay User's Manual. *Technical Report*.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.*, **B57**, 289–300.

Cargill,M., Altshuler,D., Ireland,J., Sklar,P., Ardlie,K., Patil,N., Shaw,N., Lane,C.R., Lim,E.P., Kalyanaraman,N. *et al.* (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.*, **22**, 231–238.

Chee,M., Yang,R., Hubbell,E., Berno,A., Huang,X.C., Stern,D., Winkler,J., Lockhart,D.J., Morris,M.S. and Fodor,S.P. (1996) Accessing genetic information with high-density DNA arrays. *Science*, **274**, 610–614.

Cutler,D.J., Zwick,M.E., Carrasquillo,M.M., Yohn,C.T., Tobin,K.P., Kashuk,C., Mathews,D.J., Shah,N.A., Eichler,E.E., Warrington,J.A. *et al.* (2001) High-throughput variation detection and genotyping using microarrays. *Genome Res.*, **11**, 1913–1925.

Durbin,R., Eddy,S., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis*. Cambridge University Press, Cambridge.

Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci., USA*, **95**, 14863–14868.

Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J.P., Coller,H., Loh,M.L., Downing,J.R., Caligiuri,M.A. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Hacia,J.G., Fan,J.B., Ryder,O., Jin,L., Edgemon,K., Ghandour,G., Mayer,R.A., Sun,B., Hsie,L., Robbins,C.M. *et al.* (1999) Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. *Nat. Genet.*, **22**, 164–167.

Halushka,M.K., Fan,J.B., Bentley,K., Hsie,L., Shen,N., Weder,A., Cooper,R., Lipshutz,R. and Chakravarti,A. (1999) Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat. Genet.*, **22**, 239–247.

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.

Lieberfarb,M., Lin,M., Lechpammer,M., Li,C., Tanenbaum,D.M., Wright,R., Shim,J., Kantoff,P.W., Loda,M., Meyerson,M. *et al.* (2003) Genome-wide loss-of-heterozygosity analysis from laser-capture microdissected prostate cancer using SNP arrays and a novel bioinformatics platform dChipSNP. *Cancer Res.*, **63**, 4781–4785.

Lindblad-Toh,K., Tanenbaum,D.M., Daly,M.J., Winchester,E., Lui,W.O., Villapakkam,A., Stanton,S.E., Larsson,C., Hudson,T.J., Johnson,B.E. *et al.* (2000) Loss-of-heterozygosity analysis of small-cell lung carcinomas using single-nucleotide polymorphism arrays. *Nat. Biotechnol.*, **18**, 1001–1005.

Lipschutz,R.J., Fodor,S.P.A., Gingeras,T.R. and Lockhart,D.J. (1999) High density synthetic oligonucleotide arrays. *Nat. Genet.*, **21**, 20–24.

Mei,R., Galipeau,P.C., Prass,C., Berno,A., Ghandour,G., Patil,N., Wolff,R.K., Chee,M.S., Reid,B.J. and Lockhart,D.J. (2000) Genome-wide detection of allelic imbalance using human SNPs and high-density DNA arrays. *Genome Res.*, **10**, 1126–1137.

Schubert,E.L., Malone,K., Hsu,L., Daling,J.D., Cousens,L.G. and Porter,P. (2001) Whole genome LOH analysis of lobular and ductal breast cancers by Hu-SNP array. *Breast Cancer Res. Treatment*, **69**, 232.

Wang,D.G., Fan,J.B., Siao,C.J., Berno,A., Young,P., Sapolsky,R., Ghandour,G., Perkins,N., Winchester,E., Spencer,J. *et al.* (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, **280**, 1077–1082.

Wang,Z.C., Lin,M., Wei,L.J., Li,C., Miron,A., Lodeiro,G., Harris,L., Ramaswamy,S., Tanenbaum,D.M., Meyerson,M., Iglehart,J.D. and Richardson,A. (2003) Loss of heterozygosity and its correlation with expression profiles in subclasses of invasive breast cancers. *Cancer Res.*, in press.

Westfall,P.H. and Young,S.S. (1993) *Resampling-Based Multiple Testing: Examples and Methods for P-value Adjustment*. Wiley, New York.