

Model-Based Genotyping and Allele-Specific Copy Number Analysis Using SNP Arrays

Thomas LaFramboise,¹ Barbara Weir,² Xiaojun Zhao,³ Matthew Meyerson,⁴

Keywords: SNP array, copy number, probe-level modeling, EM algorithm

1 Introduction.

Genomic alterations are believed to be the major underlying cause of cancer [5]. These alterations include various types of mutations, translocations, and copy number alterations. The latter category includes chromosomal regions with more than two copies in the cell (amplifications), and regions with one (heterozygous deletions) or zero copies (homozygous deletions). Genes contained in amplified regions are natural candidates for oncogenes, while those in regions of deletion are potential tumor-suppressor genes. Thus, the identification of specific sites of these alterations in cell lines and tumor samples is a central aim in cancer research.

Affymetrix has recently produced the the GeneChip Human Mapping 100K Set [1], which we hereafter refer to simply as the SNP array. SNP arrays use 40 oligonucleotide probes to interrogate each each SNP. The arrays aim to identify which of the two alleles — arbitrarily labeled allele A and allele B — occurs for each chromosome at each SNP site. Thus, each individual can ideally be genotyped as either homozygous A, homozygous B, or heterozygous A/B. The current generation of SNP arrays is actually a pair of arrays able to interrogate over 100,000 human SNPs. It has been demonstrated that these arrays may be used to identify copy number alterations, producing a measure of genomic copy number at each SNP [6].

SNP arrays have also successfully been employed in loss-of-heterozygosity (LOH) identification [4]. However, typically LOH is inferred when an imbalance in allelic signal occurs in a tumor sample where its matched normal is heterozygous. A complicating issue is that the imbalance may be due to the amplification of one of the alleles rather than the deletion of the other. This difficulty would be overcome if allele-specific copy number could be determined. Allele-specific copy number is also applicable in studies tying certain SNP haplotypes to deleterious germline mutations via linkage disequilibrium.

In this work, we present an algorithm that infers allele-specific copy numbers from SNP array data. We present a single model relating raw probe intensity to allele-specific copy number. This model is based on the observation that the 40 probes interrogating each SNP may be classified by the number of bases the probe mismatches each SNP allele. Fitting our model to SNP array data from various samples using an expectation-maximization (EM) algorithm [3], we are able to: a) very accurately genotype normal samples; and b) infer allele-specific copy numbers in cancer cells after model calibration on normal samples.

¹Department of Medical Oncology, Dana-Farber Cancer Institute, 44 Binney Street, Boston, MA 02115. E-mail: thomas_laframboise@dfci.harvard.edu

²Department of Medical Oncology, Dana-Farber Cancer Institute, 44 Binney Street, Boston, MA 02115. E-mail: barbara_weir@dfci.harvard.edu

³Department of Medical Oncology, Dana-Farber Cancer Institute, 44 Binney Street, Boston, MA 02115. E-mail: xiaojun_zhao@dfci.harvard.edu

⁴Department of Medical Oncology, Dana-Farber Cancer Institute, 44 Binney Street, Boston, MA 02115. E-mail: matthew_meyerson@dfci.harvard.edu

2 Model Specification.

In the SNP array, each probe is a 25-mer designed to either perfectly or nearly match some 25-nucleotide sequence containing the SNP. Restricting our attention to a specific but arbitrarily chosen SNP, let C_A and C_B denote the number of copies of the allele A and allele B , respectively, in the sample. Assuming that the intensities are a linear function of copy number [2], with slope depending upon the number of mismatch bases, the resulting model is

$$Y_i = \alpha + \beta_{A_i} C_A + \beta_{B_i} C_B + e$$

where Y_i denotes the (normalized) intensity of the i^{th} probe and A_i and B_i are the number of bases (either 0, 1, or 2) at which the probe is not perfectly complementary to the A and B targets, respectively. The first term α represent background signal, which can arise from optical noise and non-specific binding. Hence the model parameters are α , β_0 , β_1 , and β_2 .

In general, neither the copy numbers nor the values of the parameters are known *a priori*. Therefore, we first calibrate the model on a set of normal samples, using an EM algorithm for fitting. In this way, we obtain highly accurate genotype calls along with estimates of the model parameters. The parameter estimates are then used in the model to infer copy numbers in aberrant tumors and cell lines.

3 Results.

To evaluate our method for genotyping normal samples, we applied the approach to data from samples analyzed as part of the International HapMap Project (<http://www.hapmap.org>). We had very strong agreement (> 99%) with the HapMap calls.

We applied our allele-specific copy number algorithm to SNP array data derived from tumor and cancer cell lines. Interestingly, our result indicates that for the overwhelming majority of amplifications, the gain occurs on one of the chromosomes but not the other. We tested our allele-specific copy number inferences using allele-specific PCR. The agreement between our estimates and the PCR results is very strong, though this agreement degrades for very high-level amplifications.

References

- [1] Affymetrix 2004. *GeneChip Human Mapping 100K Set Data Sheet*, Santa Clara, CA.
- [2] Bignell, G. R., Huang, J., Greshock, J., Watt, S., Butler, A., West, S., Grigiriva, M., Jones, K. W., Wei, W., Stratton, M. R., Futreal, P. A., Weber, B., Shaper, M. H., and Wooster, R. 2004. High-resolution analysis of DNA copy number using oligonucleotide microarrays. *Genome Research* 14:287–295.
- [3] Dempster, A. P., Laird, N. M., and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* 39:1–38.
- [4] Wang, Z. C., Lin, M., Wei, L. J., Li, C., Miron, A., Lodeiro, G., Harris, L., Ramaswamy, S., Tanenbaum, D. M., Meyerson, M., Inglehart, J. D., and Richardson, A. 2004. Loss of heterozygosity and its correlation with expression profiles in subclasses of invasive breast cancers. *Cancer Research* 64:64–71.
- [5] Weir, B., Zhao, X., and Meyerson, M. 2004. Somatic alterations in the human cancer genome. *Cancer Cell* 6(5):433–438.
- [6] Zhao, X., Li, C., Paez, J. G., Chin, K., Jänne, P. A., Chen, T. H., Girard, L., Minna, J., Christiani, D., Leo, C., Gray, J. W., Sellers, W. R., and Meyerson, M. 2004. An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Research* 64:3060–3071.