BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 19 2013, pages 2482–2484
doi:10.1093/bioinformatics/btt416

Genome analysis

Advance Access publication August 7, 2013

OncoSNP-SEQ: a statistical approach for the identification of somatic copy number alterations from next-generation sequencing of cancer genomes

Christopher Yau

Department of Mathematics, South Kensington Campus, Imperial College London, London SW7 2AZ, UK

Associate Editor: Gunnar Ratsch

ABSTRACT

Summary: Recent major cancer genome sequencing studies have used whole-genome sequencing to detect various types of genomic variation. However, a number of these studies have continued to rely on SNP array information to provide additional results for copy number and loss-of-heterozygosity estimation and assessing tumour purity. OncoSNP-SEQ is a statistical model-based approach for inferring copy number profiles directly from high-coverage whole genome sequencing data that is able to account for unknown tumour purity and ploidy.

Availability: MATLAB code is available at the following URL: https://sites.google.com/site/oncosnpseq/.

Contact: c.yau@imperial.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on November 1, 2012; revised on June 5, 2013; accepted on July 15, 2013

1 INTRODUCTION

Many cancers are characterized by complex structural rearrangements involving copy number alterations (CNAs) and/or losses of heterozygosity (LOH) events (Beroukhim *et al.*, 2010). Genetic abnormalities in specific regions may be related to the aggressiveness of a cancer and be associated with clinical outcomes (Collisson *et al.*, 2012). OncoSNP-SEQ is a statistical model-based approach for identifying CNAs and LOH events directly from next-generation sequencing data that builds on the OncoSNP tool for SNP array analysis (Yau *et al.*, 2010). The statistical model used differs from previous methods such as BIC-seq (Xi *et al.*, 2010), Control-FREEC (Boeva *et al.*, 2012) and APOLLOH/HMMcopy (Ha *et al.*, 2012) by addressing three major issues in the analysis of cancer samples: (i) normal cell contamination, (ii) intra-tumour heterogeniety and (iii) polyploidy, within a fully integrated statistical framework for using both read depth and allele-specific information in high-coverage experiments. In the following, we describe the statistical model and its application to a publicly available normal-cancer cell line dataset, simulated mixtures and a heterogeneous cancer.

2 METHODS

Assume that the data consists of a pair of allele-specific read count measurements (a_i, b_i) for $i = 1, \ldots, N$ SNPs. Furthermore, assume that the sequencing coverage is sufficiently high ($>60\times$) so that it is possible to approximate the total read counts as a continuous measurement. This approximation provides benefits in terms of simplified mathematical manipulation, flexibility and computational speed. Any necessary pre-processing to remove GC content and mapability related biases are assumed to have been applied (see Supplementary Methods).

2.1 Statistical model

The total read count $r_i = a_i + b_i$ is distributed according to a Student t-distribution with variance λ^2 and degrees of freedom ν:

$$r_i | x_i, u_i, u_0, \lambda^2, \nu \sim \text{Student}(m_{x_i}, \lambda^2, \nu). \quad (1)$$

The quantity m_i represents the expected read count at a locus, given the copy number aberration state x_i with corresponding normal copy number $c^{(n)}_{x_i}$ and tumour copy number $c^{(t)}_{x_i}$,

$$m_{x_i} = (u_0 + u_i(1-u_0))c^{(n)}_{x_i} h + (1-u_0)(1-u_i)c^{(t)}_{x_i} h,$$

which is formed by contributions from (i) the normal cell population, $0 \leq u_0 < 1$, (ii) the proportion of tumour cells harbouring the aberration $(1-u_0)(1-u_i), 0 \leq u_i < 1$ and (iii) the proportion of tumour cells, which do not have the aberration $(1-u_0)u_i$. The term h corresponds to the haploid read coverage.

We model the relative proportion of one of the alternate alleles b_i to the total read count r_i using a mixture of a Uniform distribution on $(0, r_i)$ and a Binomial distribution with r_i trials and success probability, p_i, similar to (Ha *et al.*, 2012). The success probability that the b_i reads come from the B allele is given by

$$\pi(b_i|\cdot) = e\text{Uni}(0, r_i) + (1-e)\text{Bn}(r_i, p_{z_i, x_i}), \quad (2)$$

where e corresponds to the proportion of SNPs for which the sequencing technology gives erroneous results.

For the Binomial distribution, the success probability p_{z_i, x_i} is given by $p_{z_i, x_i} = \epsilon(1-\tilde{p}) + (1-\epsilon)\tilde{p}$ where

$$\tilde{p} = \frac{(u_0 + (1-u_0)u_i)B^{(n)}_{z_i, x_i} + (1-u_0)(1-u_i)B^{(t)}_{z_i, x_i}}{(u_0 + (1-u_0)u_i)c^{(n)}_{x_i} + (1-u_0)(1-u_i)c^{(t)}_{x_i}}$$

And ϵ is the probability that any individual read will be erroneous, i.e. allele b will be misread as allele a and vice-versa, and $B^{(n)}_{z_i, x_i}, B^{(t)}_{z_i, x_i}$ correspond to the number b alleles for the z_i-th normal and tumour genotypes of the x_i-th copy number aberration state. The observation likelihood is formed by the product of (1) and (2).

The copy number aberration states ($x_i \in \{1, \ldots, S\}$) form a discrete-time Markov chain:

$$\pi(x_i = k | x_{i-1} = j) = A_{j,k}, (j,k) \in \{1, \ldots, S\}^2, i = 2, \ldots, N,$$

with transition matrix A and initial state distribution ν. The transition matrix is symmetric with state-independent switching probability ρ. We have not used spatial scaling of the transition rates for non-uniform

markers, as the density of the markers is extremely high (see Supplementary Methods).

2.2 Parameter estimation

It is typical with SNP and aCGH data to use expectation-maximization or full Bayesian techniques (via Monte Carlo simulations) to estimate model parameters. In this instance, the sequence lengths and dynamic range means the iterative application of the forward-backward algorithm for HMMs is not computationally trivial. As a consequence, we estimate model parameters off-line. The variance parameter λ^2 is obtained from the data using $\hat{\lambda}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(r_i - \hat{r}_i)^2$ where $\hat{r}_i$ is the 'smoothed' version of the total read counts r using a moving window average. The sequencing error parameters ϵ (≈ 0.01 in our experiments) and e (≈ 0.01) are fixed and estimated off-line from the analysis of sequence data obtained from germ-line samples using SNP genotypes obtained from SNP arrays. For the haploid read coverage h and normal contamination level u_0, we scan a range of values for both parameters and calculate the log-likelihood $p(r, b|h, u_0)$ with other parameters fixed. We identify local modes in the log-likelihood profile and treat each local mode as a potentially plausible ploidy-purity configuration and report copy number profiles for each configuration (see Supplementary Fig. S2).

2.3 Inference

The sequence of copy number aberrations states is estimated using a multi-step Viterbi segmentation. A range of values for the transition rate parameter is selected, $\rho_1 < \rho_2 < \ldots < \rho_J$. The Viterbi solution $\hat{x}^{(1)} = \arg\max_{x_1} p(x^{(1)}|r, b)$ is obtained conditional on the smallest transition parameter ρ_1. Then $\hat{x}^{(2)}$ is found by conditioning on $\hat{x}_1$ and the next transition parameter ρ_2 and so on for $j = 3, \ldots, J$ ($J = 7$ for our experiments). The sequence of Viterbi solutions $\hat{x}_1, \ldots, \hat{x}_J$ provides a multi-scale representation of the copy number profile with $\hat{x}_1$ containing only the largest events, whereas $\hat{x}_J$ incorporates finer scale genomic aberrations (see Supplementary Fig. S1).

3 RESULTS

OncoSNP-SEQ was used to analyse a publicly available paired normal-cancer cell line (HCC1187) that was sequenced using Complete Genomics (Complete Genomics, 2012) for which Affymetrix SNP 6.0 SNP array data were also available via the Cancer Cell Line Encyclopaedia project (Barretina *et al.*, 2012). A series of simulated normal-cancer mixtures was generated from the reads of the normal and cancer cell lines in the following ratios (T:N): 100:0, 81:19, 68:32, 50:50, 35:65 and 25:75. Sequence-based copy number calls from OncoSNP-SEQ were compared with OncoSNP (Yau *et al.*, 2010) calls derived from the Affymetrix SNP data.

Figure 1 shows that OncoSNP-SEQ was able to approximately estimate the proportion of normal contamination used in the simulated mixtures and, for mixtures involving <50% normal contamination, give CNA and LOH profiles that are >90% concordant with those derived from the SNP array data. Visual inspection of the copy number profiles showed that the majority of copy number estimation differences (for tumour content greater than 50%) involved small events with a slight bias towards copy number gains that maybe attributed to the differing resolutions and dynamic ranges of the platforms (see Supplementary Fig. S4–S9).

Below 50% tumour content, copy number prediction becomes more error prone, as different configurations of normal contamination and base read depth become compatible with the data (see Supplementary Fig. S3). For complex genomically unstable tumours, it is typical to see a number of configurations that plausibly fit the data. This arises because of the linearity of the sequencing data, which produce an invariance in the observation likelihood. For example, if we halve the haploid read depth and double the copy number, the observation likelihood would remain the same. Copy number analysis of complex tumours with high levels of normal contamination should always be treated with caution.

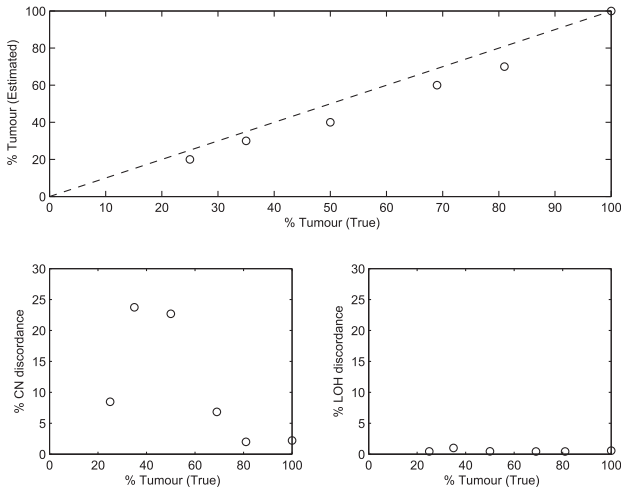

Fig. 1. Tumour purity estimation and copy number/LOH discordance for simulated normal-cancer mixtures

The ability to detect heterogeneous events was tested using a sequencing dataset from a recently published study (Schuh, 2012) of heterogeneous clonal evolution patterns in chronic lymphocytic leukemia (CLL) patients. The data consist of samples taken from the same patient (CLL003) at six different time points during the evolution of the disease. OncoSNP-SEQ was about to identify deletion events spanning the *ATM* gene and characterize the fall in sample fraction harbouring the deletion (100–50%) at the time point taken in sample CLL003-P1 (see Supplementary Fig. S4).

4 CONCLUSION

OncoSNP-SEQ is a tool for automatic detection of CNAs and LOH regions using high-coverage (60× and above) whole-genome sequencing data. High-coverage data allows allele-specific information to be utilized, and, in the case of tumour samples, OncoSNP-SEQ uses this information to evaluate the level of contamination by normal cells as well as allowing for degrees of intra-tumour heterogeneity. The formal assessment of the accuracy of this latter capability will need to be ascertained using independent assays, e.g. FISH, in future studies. For polyploid tumours, OncoSNP-SEQ provides a probabilistic means of evaluating different ploidy configurations. The software is written in MATLAB, and the source codes are freely available for modification, re-engineering or incorporation into genomic analysis pipelines.

ACKNOWLEDGEMENTS

Thanks to A. Halpern (Complete Genomics) who provided the simulated normal-cancer mixtures sequence data and A. Schuh for the CLL data. Thanks to J–B. Cazier for comments and discussions.

Conflict of Interest: none declared.

REFERENCES

Barretina,J. *et al.* (2012) The Cancer Cell Line Encyclopaedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**, 603–607.

Beroukhim,R. *et al.* (2010) The landscape of somatic copy-number alteration across human cancers. *Nature*, **463**, 899–905.

Boeva,V. *et al.* (2012) Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics*, **28**, 423–425.

Collisson,E.A. *et al.* (2012) What are we learning from the cancer genome? *Nat. Rev. Clin. Oncol.*, **9**, 621–630.

Complete Genomics. (2012) http://www.completegenomics.com/public-data/cancer-data/ (10 October 2011, date last accessed).

Ha,G. *et al.* (2012) Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer. *Genome Res.*, **22**, 1995–2007.

Schuh,A. *et al.* (2012) Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns. *Blood*, **120**, 4191–4196.

Yau,C. *et al.* (2010) A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data. *Genome Biol*, **11**, R92.

Xi,R. *et al.* (2010) BIC-seq: a fast algorithm for detection of copy number alterations based on high-throughput sequencing data. *Genome Biol.*, **11** (**Suppl. 1**), O10.