Losing balance: Hardy–Weinberg disequilibrium as a marker for recurrent loss-of-heterozygosity in cancer

Katherine Wilkins1,2 and Thomas LaFramboise1,2,3,*

1Department of Genetics, Case Western Reserve University School of Medicine, Cleveland, Ohio 44106, USA, 2Department of Electrical Engineering and Computer Science, Case Western Reserve University, Cleveland, Ohio 44106, USA and 3Genomic Medicine Institute, Lerner Research Institute, Cleveland Clinic Foundation, Cleveland, Ohio 44106, USA

Received July 14, 2011; Revised September 2, 2011; Accepted September 12, 2011

Identifying regions of loss-of-heterozygosity (LOH) in a tumor sample is a challenging problem. State-of-the-art computational approaches can infer LOH from single-nucleotide polymorphism (SNP) array data, but calling precise boundaries is complicated by normal-cell contamination and markers that are homozygous in the germline and therefore non-informative. More recently, the focus has shifted to pinpointing the loci recurrently affected by LOH events across multiple tumors. Recurrent LOH regions often harbor genes important for tumor suppression. Here, we propose a method that infers LOH rates across an entire sample set on an SNP-by-SNP basis. Our method achieves this by leveraging the straightforward principle that, by definition, LOH depletes heterozygotes, thereby disrupting Hardy–Weinberg equilibrium. We apply a statistical test for such LOH-influenced disruptions, and derive a maximum-likelihood estimator for the LOH rate based on the observed number of heterozygotes. This accounts for LOH in both its hemizygous deletion and copy-neutral forms, and does not make use of matched normal genotypes. Power simulations show high levels of sensitivity for the statistical test, and application to a control normal-tissue data set demonstrates a low false-discovery rate. We apply the method to three large publicly available tumor SNP array data sets, where it is able to localize tumor-suppressor gene targets of the LOH events. Inferred LOH rates are quite concordant across platforms/laboratories and between cell lines and tumors, but in a tumor type-dependent fashion. Finally, we produce rate estimates that are generally higher than previously published, and provide evidence that the latter are likely underestimates.

INTRODUCTION

Human cancers are characterized by the acquisition of somatic mutations. These mutations include single-base changes, structural alterations such as inversions and translocations, and gains and losses of chromosomal segments. One crucial class of somatic mutation is loss-of-heterozygosity (LOH), wherein a portion of one of the parental homologs is lost, resulting either in hemizygous deletion (one copy per tumor cell of the locus) or copy-neutral LOH where the deleted segment of one parental homolog is replaced with a duplication of the other homolog. The latter phenomenon is also known as uniparental disomy or gene conversion, and results in two copies of the locus per tumor cell. In cancer, LOH events are often manifestations of Knudson’s ‘two-hit’ hypothesis (1) in which LOH is the ‘second hit’ after the retained parental homolog is neutralized via mutation, DNA methylation or other means. Alternatively, LOH may mitigate the activity of a gene in the region because of haploinsufficiency. All of these mechanisms can confer a selective growth advantage to the affected cell by inactivating tumor-suppressor genes. It follows that tumor-suppressor genes may be identified by querying large data sets of tumors for recurrent regions of LOH. This

*To whom correspondence should be addressed at: 10900 Euclid Avenue, Cleveland, OH 44106, USA. Tel: +1 2163680150; Fax: +1 2163683432; Email: thomas.laframboise@case.edu

© The Author 2011. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com
underlying paradigm—that by querying large tumor data sets, recurrent ‘driver’ lesions may be separated from ‘passengers’—is central to large efforts such as the Cancer Genome Atlas (USA), the Cancer Genome Project (UK) and others.

In the last decade, the single-nucleotide polymorphism (SNP) array has emerged as a tool to scan the cancer genome in a high-throughput manner, revealing multiple classes of somatic changes (2). The arrays provide genotypes at hundreds of thousands of SNPs spread across the genome. The two alleles of each SNP are usually labeled as A and B for convenience, and a DNA sample is therefore genotyped at an SNP as homozygous AA, homozygous BB or heterozygous AB. Although tumor DNA often harbors chromosomal regions deviating from two copies per cell, the genotyping software will call SNPs in such regions as homozygous or heterozygous. The calls are useful for LOH detection (3) since, in theory, LOH implies absence of heterozygotes (though the converse is not true).

Precisely calling LOH regions on a sample-by-sample basis from SNP genotypes is challenging. Without matched normal data, most algorithms rely on unusual stretches of homozygos- ity as a sign of LOH. This approach is very imprecise and depends heavily on hidden Markov model (HMM) (4) or segmentation (5) approaches. Even when matched normal genotypes are available, SNPs that are homozygous in the germline will be uninformative regarding LOH, again necessi- tating imprecise HMM/segmentation procedures. Entire segments of true loss may be missed because of contamination by stromal tissue surrounding tumor cells, or even by a single SNP in the middle of the segment miscalled as a hetero-zygote. Furthermore, highly focal regions of LOH are very likely to go undetected. Sample-specific lack of sensitivity will affect downstream multiple-sample querying for recurrent LOH, which is typically performed by intersecting the sample-specific regions across the entire cohort. As a result, many recent large-scale cancer genome studies do not prominently feature LOH analysis despite its biological importance. For example, the Cancer Genome Atlas study barely mentions LOH in its study of glioblastoma multiforme (GBM) (6), and only considers samples with matched normal genotypes for this purpose.

In this manuscript, we present a method that avoids these problems by treating each SNP individually. The key observation is that recurrent LOH will result in a relative depletion of heterozygotes for SNPs in the affected region. If an SNP is in Hardy–Weinberg equilibrium (HWE) in the population germ- line, recurrent LOH events in the tumor will disrupt this equi- librium (Fig. 1). The Hardy–Weinberg principle has been central in the population genetics field for most of the past century but has not previously been applied in somatic cancer genomics. At each SNP in a data set of tumor geno- types, we apply a one-sided test—termed the HWE-LOH test—for the alternative hypothesis that there are fewer hetero-zygotes than expected. The null distribution of the number of heterozygotes under HWE was described in Wigginton et al. (7), where the $P$-value appropriate for our one-sided test was termed $P_{\text{low}}$. In addition to adapting this test to our setting, we extend the concept to derive a maximum-likelihood esti- mator based on the observed number of heterozygotes and allelic composition at the SNP.

To assess the performance of our test, we perform simula- tion experiments to gauge its power. In addition, we apply the test and maximum-likelihood estimator to three large-scale SNP array genotype data sets from tumors (Table 1). The first comprises 166 patient samples from The Cancer Genome Atlas (TCGA) study of glioblastoma (6). These samples were run on the Illumina Infinium HumanHap550 SNP Chip, which interrogates 547,458 autosomal SNPs. Matched
normal DNA for each sample was also run on the array, providing a convenient negative control. The second data set has Affymetrix Genome-Wide Human SNP Array 6.0 genotypes generated by the Wellcome Trust Sanger Institute’s Cancer Genome Project for each of the 841 cancer cell lines derived from various tumor types (8). For the third data set, we obtained 1767 array samples from a recently published study across multiple cancer types (9). These samples were run on the Affymetrix 250K Sty array, which interrogates 222 838 autosomal SNPs. In the current study, we refer to the three data sets as the TCGA data, the Cancer Genome Project data and the 250K data, respectively.

RESULTS

HWE-LOH test is well-powered to detect recurrent LOH

To assess the power of the HWE test, we first performed an analysis using simulated data (see Materials and Methods). The parameters that affect power in our setting are sample size, LOH rate and SNP minor allele frequency (MAF). We ran our simulations for sample sizes 100, 500 and 1000. Supplementary Material, Figure S1 shows power as a function of LOH rate and MAF for the various sample sizes. As can be seen, we have >80% power to detect LOH rates as low as 20%, unless the sample size or MAF is quite low. Since LOH regions often encompass multiple SNPs, a high-MAF SNP being affected is usually likely. Therefore, sample size would be more likely to adversely impact power in practice. For the larger sample sizes, however, we are well-powered to detect even LOH rates as low as 10–20%. As we show below, rates this high are quite common in tumor collections.

HWE-LOH test has a low false-positive rate

Matched normal genotype data were available for each sample in the TCGA data set, providing a convenient negative control (note that our method neither requires nor uses matched normal genotypes—they are used here solely to gauge specificity). The HWE-LOH test results for the tumor and normal TCGA genotypes are shown in Figure 2. Among the 510 932 array SNPs tested, only 76 (<0.015%) in the matched normal genotypes (Fig. 2A) yielded \( P \)-values < 10\(^{-6} \). In contrast, 20 710 (4%) of the SNPs show \( P \)-values < 10\(^{-6} \) for the tumor genotypes (Fig. 2B), implying a false discovery rate < 0.4% for this \( P \)-value threshold. The

<table>
<thead>
<tr>
<th>Table 1. Data sets</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Identifier in current study</strong></td>
</tr>
<tr>
<td>TCGA</td>
</tr>
<tr>
<td>Cancer Genome Project</td>
</tr>
<tr>
<td>250K</td>
</tr>
</tbody>
</table>

\(^a\)Prior to sample filtering.

Figure 2. HWE-LOH \( P \)-values in the TCGA data. (A) Illumina 550K array matched normal genotype \( P \)-values from 166 GBM patients. Genotype counts for rs7857074 in the \( PTPRD \) gene on chromosome 9p are shown in the inset. (B) Illumina 550K array tumor genotype \( P \)-values from the same patients. The inset shows genotype counts for rs7857074 in the tumor.
HWE-LOH test identifies PTPrD in the peak region

Figure 2 shows very strong signals across all of chromosome 10, as well as a much more focal peak on chromosome 9p. The top-scoring SNP in the peak is rs7857074. The insets in Figure 2 show genotype counts in the normal and the tumor at rs7857074. This SNP is within the transcribed region of the PTPrD gene. Intriguingly, PTPrD is emerging as an important tumor-suppressor gene in a variety of cancer types, including glioblastoma (10–12). The gene is not highlighted in the TCGA glioblastoma (6) paper (PTPrD alterations are termed ‘infrequent’), however, despite the fact that their study used the very same data that we present here.

Maximum-likelihood estimator of LOH rates recapitulates tumor-suppressor genes highlighted in the TCGA study

Although the HWE-LOH P-value is a sensitive and specific measure to detect recurrent LOH, it is influenced not only by the local LOH rate, but also by the underlying allele frequencies of the SNP (Supplementary Material, Fig. S1). As such, the P-values are not locally constant—that is, their values may change abruptly between two genomically consecutive SNPs, even though the LOH rate may be nearly identical at the two SNPs. Since LOH events occur in segments, on the other hand, the LOH rate is locally constant. It is desirable to assess the rate across a sample set at a specific locus. To address this, we developed a maximum-likelihood approach to estimate the LOH rate at each SNP. Our approach constructs the expected probability distribution for the number of tumor heterozygotes given an LOH rate. Since the numbers of heterozygotes are observed, the maximum-likelihood estimator for LOH rate may then be computed (see Materials and Methods for details).

One of the remarkable findings in the original TCGA paper was the identification of three frequently altered pathways—RTK/RAS/PI(3)K, p53 and RB—in the interrogated samples (Supplementary Material, Table S2) represented in this cell line collection, particularly 3p, 9p, 10p, 13p and 17p. Each of these locations harbor tumor-suppressor genes crucial in a variety of tissue types—MLH1, CDKN2A, PTEN, RB1 and TP53, respectively. The result is therefore consistent with the data being an amalgamation of tumor types, with the strongest LOH signals coming from regions harboring genes critical across tumor classes. Indeed, there are 26 different tissue types (Supplementary Material, Table S2) represented in this cell line collection, based on the ~80% for which this information was available. For comparison, we also computed the LOH estimates on a subset of the data comprising 83 lung cell lines (Fig. 4B). Overall, the LOH peaks are considerably higher in the lung-only set than in the whole data set, likely owing to greater homogeneity of lung-specific tumor-suppressor genes. On the other hand, the smaller sample size for lung-only gives a larger variance in the maximum-likelihood estimates, which is reflected in the wider vertical range of the plot. Since maximum-likelihood estimators are well
known to be statistically unbiased, however, the average rates across arms should be quite accurate.

Like the Cancer Genome Project data set, the 250K set comprises dozens of tumor types (Supplementary Material, Table S3). To test whether the LOH estimator is affected by platform- or laboratory-specific effects, we applied it to the non-small cell lung subset (198 samples) of the 250K data set and performed a correlation analysis with the lung-only subset of the Cancer Genome Project data (Fig. 5A). For each chromosome arm, we computed average estimated LOH rates within the Cancer Genome Project and 250K groups, and examined the correlation between the two. The correlation was quite strong here ($r^2 = 0.82; P = 2.06 \times 10^{-11}$). We conclude that our methods are not overly susceptible to platform- or laboratory-specific artifacts.

There is some controversy in the cancer research community as to whether cell lines are a reliable model for in vivo human tumors. To consider this in the context of our own data, we divided the 250K sample set (the portion for which the information was available) into 109 cell lines and 415 primary tumors. As shown in Figure 5B, the correlation is again quite strong ($r^2 = 0.79; P = 3.64 \times 10^{-10}$), lending support to the validity of the cell line as a model. Note the outlying key peaks 9p and 17p arms in the plot, likely reflecting the high rate of LOH at tumor-suppressor genes CDKN2A and TP53, respectively, in multiple tumor types. It should also be noted that the LOH rate estimates are systematically higher in the cell lines than in the primary tumors, but this is not surprising since LOH is well known to be obscured due to contamination by the normal cells that are typically present in primary tumors.

Genomic LOH profiles vary by tumor type but the overall rate higher than previously reported

We noticed that, in general, our maximum-likelihood approach yields LOH rate estimates considerably higher than has typically been reported. No specific estimates of LOH rates were provided by the three studies that produced the three data sets we analyze here. However, for example, Weir et al. (14) estimate LOH rates in a data set of 371 lung adenocarcinomas run on the 250K array. In that study, the most frequent LOH region was on chromosome 17p, where the estimated rate was $\sim14\%$ (17 of 124 of the least stromally

Figure 4. Estimates of LOH rates in 471 Cancer Genome Project cell lines. (A) The maximum-likelihood estimates of LOH rate at each SNP derived from the Affymetrix 6.0 genotypes of 471 tumor cell lines are mapped to the SNP’s genomic coordinates and smoothed. (B) Results from the same estimation procedure performed on a subset of 83 lung cancer samples. Note that the smaller sample size in (B) yields a larger estimator variance, accounting for the wider vertical spread.
contaminated samples). In contrast, our 17p estimates in the lung samples were \(\approx 78\%\) for the Cancer Genome Project Data and 24% for the 250K set (Fig. 5A). This raises the possibility that ours are overestimates. To investigate more closely, we examined germline heterozygosity rates of the Affymetrix array SNPs. We reasoned that SNP heterozygosity across the arm should drop in the tumor at a rate approximately equal to the actual LOH rate. According to the manufacturer, Affymetrix array SNPs as a whole had a mean heterozygosity rate of 26.7% in western European (HapMap CEU) samples. This corresponds closely with the mean 27.5% heterozygosity rate in the TCGA-matched normal samples on 17p. It follows that we would expect an approximate \((1-0.78) \times 0.267 = 5.9\%\) average heterozygosity rate at the chromosome 17p SNPs in the Cancer Genome Project lung tumors and \((1-0.24) \times 0.267 = 20\%\) in the 250K lung data. In fact, the actual averages are even slightly lower at 4.7 and 19.8%, respectively, giving evidence that our approach does not overestimate LOH rates.

**Novel candidate tumor-suppressor genes identified across data sets**

Regions of recurrent LOH (see Materials and Methods), the genes harbored therein and the SNPs giving strong LOH signals are provided in Supplementary Material, Tables S1, S4 and S5. Although the overall genomic profiles differ considerably among the three data sets, genes with strong signals across data sets do emerge. For example, the TCGA data set is clearly dominated by the loss of chromosome 10 (Fig. 2B), although chromosome 10 does not feature prominently in the results of the other two data sets. However, the SORCS1 gene on 10q is among the most significant genes in both the TCGA (HWE-LOH test \(P = 1.25 \times 10^{-26}\)) and 250K (\(P = 7.6 \times 10^{-84}\)) data sets, ranking among the top 25 in both. Outside of chromosome 10, GLP2R on chromosome 17 shows an estimated LOH rate \(>30\%\) in both the 250K and Cancer Genome Project data sets. Despite using genotypes from different platforms and tumor types, these rates rank among the top 25 in their respective data sets. Neither SORCS1 nor GLP2R has been previously identified as a tumor-suppressor gene. The results of our analysis, however, mark them as strong candidates in this regard.

**SORCS1** does map to the same chromosome arm as the well-known and important tumor-suppressor gene \(PTEN\), \(\approx 19\) Mb away. Although this is a considerable genomic distance, it does raise the possibility that the recurrent SORCS1 loss revealed in our study is simply a consequence of \(PTEN\) being targeted in broad chromosomal loss events. However, the maximum-likelihood-based LOH rate estimates are higher for SORCS1 than for \(PTEN\) in both the TCGA data (74 versus 54%) and the 250K data (25 versus 19%). As an additional line of evidence, the availability of matched normal genotypes for the TCGA data set allows us to examine LOH events more closely, on a sample-by-sample basis. Indeed, SNPs with heterozygous genotypes in a normal sample that change to homozygotes in the matched tumor indicate LOH (note, however, that the method presented in our study is designed to work without the presence of matched normal genotypes; we use them here strictly as an independent validation). Examination of matched normal genotypes shows that 101 of the TCGA samples carry a germline heterozygote within SORCS1 that becomes homozygous in the matched tumor. On the other hand, only 37 samples carry germline heterozygotes in \(PTEN\) that lose heterozygosity in the tumor, and these 37 samples are all included among the 101 showing LOH at SORCS1. This indicates that samples undergoing LOH at \(PTEN\) are predominantly a subset of those undergoing LOH at SORCS1. This is perhaps unsurprising, as many LOH events affect the ends of chromosomes (i.e. are not interstitial), and SORCS1 is distal to \(PTEN\).

**DISCUSSION**

We have presented a new method to detect recurrent LOH from SNP genotypes. This straightforward approach applies...
a classical population genetics principle—HWE—in the novel setting of somatic cancer genomics. Analyzing genotypes from three large tumor SNP array data sets shows that the approach can reliably identify regions harboring important tumor-suppressor genes, without relying on matched normal genotype data. By inferring LOH rates on an SNP-by-SNP basis, we entirely circumvent the problem of calling sample-specific LOH regions. Furthermore, our approach can detect LOH in both of its forms—hemizygous deletion and copy-neutral LOH. We have also shown that the method is consistent across platforms and laboratories.

To our knowledge, there is only one previous study (Seroussi et al.) (15) that applies the Hardy–Weinberg principle for the purpose of detecting chromosomal deletion. That study uses HWE to infer germline deletions in populations of cattle, and the authors use the term LOH to denote inheritance of a deletion on one parental homolog. In our study, the individual inherits both copies of the locus intact in the germline, but loses one copy somatically. The difference between these two notions of LOH is crucial here, since the germline deletion in the Seroussi et al. study is assumed to segregate independently of the two SNP alleles, thereby setting up a tri-allelic case of HWE. In our case, the LOH affects the genome sporadically, and the resulting three alleles need not be in HWE (and are not assumed to be). In short, the Seroussi et al. approach assumes HWE (although in the tri-allelic setting), whereas our approach exploits the deviation from HWE, as is appropriate in the cancer genomics setting.

One advantage of the single-SNP approach is highlighted here by the localization of PTPRD, in particular, in the TCGA data set. Close examination of the genotypes surrounding the top-scoring SNP rs7857074 underscores the difficulty of calling a discrete segment of LOH in many of the samples (Supplementary Material, Fig. S2). Few of the samples have runs of homozygosity unbroken by heterozygotes or non-informative SNPs, likely presenting challenges to segmentation/HMM approaches. Indeed, the TCGA paper’s Supplement (6) states that segments fewer than 10 SNPs long were ignored. As we have shown, recurrent LOH provides clues regarding important genes/pathways that are complementary to those gleaned from recurrent inactivating mutation and deletion events.

As mentioned above, the two genes emerging from our analysis have not been previously implicated as tumor-suppressor genes. GLP2R does harbor a variant that has recently been associated with levels of HbF-containing erythrocyte cells (16). However, a previous study (17) found no association between GLP2R levels and intestinal tumor cell growth or survival, and no other publications link the gene with malignancy. Therefore, its precise role in cancer remains unclear. The other gene, SORCS1, is part of a family of neuropeptide receptor genes associated with neurogenesis (18). SORCS1 has been linked to neurological disorders—Alzheimer disease (19) and attention-deficit hyperactivity disorder (20). The gene’s functional importance in the brain is interesting in the context of our study since the TCGA data set is derived exclusively from brain tumors, suggesting that disruption of SORCS1 may subvert normal growth inhibition mechanisms in the brain.

Our method does have some drawbacks. Its goal is to pinpoint regions of recurrence. Although recurrence is often a sign of cancer relevance, this is not always the case. Furthermore, the method relies on Hardy–Weinberg disequilibrium as a signal for recurrent LOH, and therefore SNPs out of HWE in the germline could, in theory, give rise to false-positive signals—it is well known that population stratification results in fewer heterozygotes than expected under HWE. However, although the majority of samples in the TCGA data set are of unspecified ancestry, population stratification does not seem to result in a substantial inflation of P-values in the matched normal data (Fig. 2A). It therefore seems that the effect of population stratification in disrupting HWE is not sufficiently strong to yield P-value that falls below our stringent thresholds. Nonetheless, since the other samples sets could be more ancestrally stratified, we restricted our analyses in the Cancer Genome Project and 250K data to individuals with predominantly western European ancestry.

Looking toward the future, national and international consortia such as the Cancer Genome Atlas (6) and the International Cancer Genome Consortium (21) are generating ever-larger data sets of tumor genotypes from both SNP arrays and newer ‘next-generation’ sequencing platforms. Since the method presented here takes genotypes as input, it would work equally well with deep sequencing data. However, it will be several years before costs decrease enough to permit sequencers to match the SNP array in its ability to affordably interrogate SNP genotypes globally andagnostically in a large number of samples. In any case, the ongoing deluge of genotype data—regardless of platform—ensures that approaches such as the one we propose will become increasingly important.

MATERIALS AND METHODS

Quality control

For data quality purposes, we filtered out SNPs whose population HWE P-value in Caucasians (the ancestry comprising the majority in all three sample sets) as reported in the HapMap database online (22) is $<10^{-6}$. We also removed SNPs with $<1\%$ MAF. These filtering steps left 711 774, 510 932 and 185 956 SNPs for the Cancer Genome Project, TCGA and 250K data sets, respectively.

Since we did not have access to matched normal genotypes for the Cancer Genome Project or 250K data sets, we could not assess whether P-values in these data were inflated due to population stratification. As such, we restricted the sample sets to individuals with at least 85% western European ancestry (the predominant background for all three data sets), as determined by the STRUCTURE software (23). The samples were clustered along with the HapMap samples having western European, African and Asian ancestries (the CEU, YRI and CHB/JPT cohorts), thereby forcing the algorithm to output percentage identity with each of these three ancestries.

Simulation procedure

For each sample size N, we first simulated 2N chromosomes with alleles randomly selected based on MAF. The simulated
chromosomes were then randomly paired, yielding a germline genotype for each pair (representing an individual). We then simulated LOH events by converting each heterozygote to a homozygote with probability dictated by LOH rate (lost allele chosen at random), yielding tumor genotypes. We ran our HWE-LOH deviation test on these simulated tumor genotypes, and recurrent LOH was deemed detected if the $P$-value was $<10^{-6}$. This threshold corresponds to a false-positive rate $<0.015\%$ in the TCGA sample set (see Results). The power was computed as the proportion of 1000 replicates for which LOH was detected.

**LOH rate maximum-likelihood estimator**

To derive the maximum-likelihood estimator for the number of samples undergoing LOH at an SNP, we first computed the probability distribution, $P(H_t = h | L, N)$, of the number, $H_t$, of heterozygotes observed at that SNP in the tumor, conditional on the number of samples, $N$, and the number, $L$, of them that undergo LOH at the SNP. The probability depends on the number, $n_{AB}$, of heterozygotes at the SNP in the matched germline samples. Since this number is typically unobserved, we sum over its probability mass function, $f(n_{AB}|N, N_A)$, under HWE for $N$ samples:

$$f(n_{AB}|N, N_A) = \frac{2^{n_{AB}}N!}{n_{AA}!n_{AB}!n_{BB}!} \times \frac{N_A!N_B!}{(2N)!},$$

where $N_A$ and $N_B$ denote allele counts, and $n_{AA}$, $n_{AB}$ and $n_{BB}$ denote genotype counts (7). Although $N_A$ and $N_B$ are also unobserved, random allelic loss in the tumors allowed the use of their (observed) allelic counts to estimate $N_A$ and $N_B$ in an unbiased manner. The conditional probability distribution of $H_t$ could therefore be expressed as

$$P(H_t = h | L, N) = \sum_{n_{AB}=h}^{h+L} P(H_t = h | n_{AB}, L, N) \times f(n_{AB}|N, N_A).$$

The first term in the product can be derived from a hypergeometric distribution, assuming that the $L$ LOH events affect the individuals independent of germline heterozygosity status:

$$P(H_t = h | n_{AB}, L, N) = \frac{\binom{n_{AB}}{n_{AB} - h} \binom{N - n_{AB}}{L - n_{AB} + h}}{\binom{N}{L}}.$$

From the distribution derived in Eq. (1), we generated the maximum-likelihood estimates, $\hat{L}$, of the number of samples, $L$, undergoing LOH at each SNP. This is achieved by computing the value of the probability for each possible value of $L$; that which gives the maximum value is $\hat{L}$.

**Selection of significant regions of LOH**

Mapping the maximum-likelihood estimates to the SNPs' genomic coordinates, we smoothed the values using a kernel average smoother (24). Cubic splines were fitted to these smoothed LOH frequency estimates, and the result was then searched for peak regions using a method described in Yavas et al. (25). We flagged as significant those regions in LOH frequency peaks that crossed a threshold (20% by default) and had SNPs with HWE-LOH test $P$-values $<10^{-6}$.

**Algorithm runtime**

All of the computational procedures described herein are relatively computationally straightforward, with very short runtime (even with large sample sizes across hundreds of thousands of SNPs), with the exception of the maximum-likelihood estimation procedure. This procedure was slowed substantially by the necessity of computing probabilities for each possible value of LOH count, $L$. Each of these computations, in turn, necessitates computing the value probability mass function, $f$, for each possible value of the number, $n_{AB}$, of observed heterozygotes. To avoid performing these nested computations for each SNP, we a priori generate a table of probabilities for all possible numbers of observed heterozygotes in a given sample size. Although the table is generated slowly, it only needs to be done once for each data set. Storing the table increases the speed of the algorithm and takes far less total time than generating these probabilities multiple times at runtime. This results in a runtime of $\sim 2$ min to generate the matrix of probabilities and $\sim 700$ min to run the algorithm on the 166-sample set for the 42 072 chromosome 1 SNPs. This runtime could easily be decreased by running the algorithm on batches of SNPs in parallel, as the calculations are independent for each SNP.

**SUPPLEMENTARY MATERIAL**

Supplementary Material is available at HMG online.

**ACKNOWLEDGEMENTS**

The Cancer Genome Project data set was obtained from the Wellcome Trust Sanger Institute Cancer Genome Project web site, http://www.sanger.ac.uk/genetics/CGP.

Conflict of Interest statement. None declared.

**FUNDING**

This work was supported by the National Institutes of Health (CA131341 to T.L.) and the National Science Foundation (IIS-0916102 to T.L.).

**REFERENCES**


