Integrative analysis of genome-wide loss of heterozygosity and mono-allelic expression at nucleotide resolution reveals disrupted pathways in triple negative breast cancer
Supplemental Material

Gavin Ha[1,2], Andrew Roth[1,2], Daniel Lai[2,3], Ali Bashashati[1], Jiarui Ding[1,3], Rodrigo Goya[2,5], Ryan Giuliany[1,2], Jamie Rosner[1], Arusha Oloumi[1], Karey Shumansky[1], Suet-Feung Chin[6], Gulisa Turashvili[1], Martin Hirst[5], Carlos Caldas[6], Marco A Marra[5], Samuel Aparicio[1,4], and Sohrab P Shah[1,3,4,*]

[1]Department of Molecular Oncology, British Columbia Cancer Agency, 675 West 10th Avenue, Vancouver, Canada
[2]Bioinformatics Training Program, University of British Columbia, Vancouver, Canada
[3]Department of Computer Science, University of British Columbia, Vancouver, Canada
[4]Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, Canada
[5]Genome Sciences Centre, British Columbia Cancer Agency, 675 West 10th Avenue, Vancouver, Canada
[6]Cancer Research UK, Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge, CB2 0RE, UK
[*]Corresponding Author: sshah@bccrc.ca

Contents

1 Supplemental Methods

1.1 Biospecimen collection and ethical consent

Tumour specimens were obtained from three tumour banks (BCCA Vancouver, breast tumour tissue repository; Alberta CBCF Breast Tumour Bank Edmonton, Cambridge UK, Addenbrooke's Hospital breast tumour bank) each with local REB/IRB approval for genomic studies of nucleic acids from breast cancer patients. This project was conducted under local BCCA REB/IRB projects H06-00289, H08-1230, H06-3199. The source of germline DNA was from peripheral blood lymphocytes in all but 4 cases. In these 4 cases histologically normal adjacent breast tissue was used. Initial case selection was based on clinical immunohistochemistry to define primary triple negative breast cancers obtained from surgical specimens, prior to the initiation of any chemotherapy or radiotherapy. Tumours typed as ER-, HER2- and PR- were initially selected for further review and re-validation of the IHC. Cases found to be ERBB2 amplified on copy number analysis, but IHC -ve for ERBB2, were rejected. The complete sequence level genome landscapes of these tumours will be described elsewhere (Shah et al, submitted).

1.2 Histopathological review

Tissue sections were subject to expert histopathological review (GT) to assess the presence of invasive tumour, pre-malignant or benign changes, lymphocytic infiltration, necrosis and tumour cellularity. Tumour cellularity was scored visually in a semiquantitative fashion on sections taken from the cryosectioning runs used to isolate nucleic acids from each tumour. Cellularity values were binned such that 'low cellularity' corresponds to samples with $<$40% malignant cells, 'moderate cellularity' corresponds to 40% - 70% malignant cells, and samples with $>$70% malignant cells were considered to have 'high cellularity'. All but one sample classified as low cellularity were excluded from further analysis. The ER-, PR- HER2- immunophenotype was reconfirmed on sections or TMA cores from the cases included for analysis and additionally CK5/6 and EGFR were assessed by IHC. Subsequently SNP6.0 copy number analysis was also used to confirm the absence of HER2 amplification in each case.

1.3 Library construction and sequence data generation

SOLiD whole genome shotgun libraries for 17 tumour/normal pairs were generated as previously described (McKernan et al., 2009) and aligned to the human reference genome (hg18, NCBI36) using BioScope. Illumina libraries were prepared as described in (Morin et al., 2011) and aligned using BWA (Li and Durbin, 2009). Paired end RNAseq libraries were generated as described in Wiegand et al (Wiegand et al., 2010). Sequence reads were aligned using a modified version of BWA (base version 0.5.5 (Li and Durbin, 2009) to a reference consisting of the human genome reference (NCBI build 36, hg18) and a database of known exon-exon junctions obtained from different annotation databases (Ensembl (Flicek et al., 2010), RefSeq (Pruitt et al., 2007), AceView (Thierry-Mieg and Thierry-Mieg, 2006)). Sequences representing exon-exon junctions were designed to require at least a 4 base pair overlap for split-reads. Considering a read length of 50 base pairs, 46 base pairs on either side of the exon-exon junction were concatenated to represent each exon-exon junction.

1.4 APOLLOH probabilistic framework description

A full representation of the APOLLOH framework as a probabilistic graphical model is given in Figure S1

1.4.1 Hidden state space

The full state space consists of 18 genotype states, $\mathbf{K}$. At each position t, the state space is restricted to K_{c_t} given copy number c_t (Table 1, main text). The initial state distribution, π, is conjugate Dirichlet distributed with hyperparameter δ^{π} (Table S1), acting as the prior on G_0. π is initialized with the maximum a posterior (MAP) using δ^{π} only. Positions that have are homozygous deleted are assumed to have zero depth and ignored if they managed to pass the initial depth filter. Similarly, positions with hemizygous deletion (i.e. 1 copy) status are re-labeled to copy neutral in order to guard the model against unexpected mixtures of both a and b reads. Remaining LOH regions are categorized as deletion and copy-neutral LOH by post-processing, referring back to original copy number status.

1.4.2 Emission model

Each read at a given position $t \in \mathbf{P}$ can be modeled as a Bernoulli trial. Given N_t independent Bernoulli trials, the number of reads mapping to the reference base, a_t, is modeled using a binomial distribution. The observed likelihood is a mixture of 18 univariate binomial distributions modeling input data reference read counts $a_{1:T}$ and total depth $N_{1:T}$ with parameter $\bar{\mu}_g$ conditioned on genotype $g \in \mathbf{K}$.

$$p(a_t|G_t = g) = Binomial(a_t|\bar{\mu}_g, N_t) \tag{1}$$

The parameter $\bar{\mu}_g$ is modeled using a two-component mixture,

$$\bar{\mu}_g = s\mu_N + (1-s)\,\mu_g \tag{2}$$

where μ_g is the unobserved reference allelic ratio for *tumour* cells in genotype state $g \in \mathbf{K}$, and μ_N is a fixed global allelic ratio of *normal* cells. μ_g are unobserved parameters; μ_N is set to 0.5 but is adjustable to accommodate the possibility of reference skew in the data. The observed likelihood becomes

$$\begin{aligned} p\left(a_t|G_t = g, \mu_N\right) &= \binom{N_t}{a_t} \left(\bar{\mu}_g\right)^{a_t} \left(1-\bar{\mu}_g\right)^{N_t - a_t} \\ &= \binom{N_t}{a_t} \left(s\mu_N + (1-s)\,\mu_g\right)^{a_t} \left(1 - s\mu_N - (1-s)\,\mu_g\right)^{N_t - a_t} \end{aligned} \tag{3}$$

The normal proportion parameter is prior Beta distributed with hyperparameters α_s and β_s and the tumour reference parameter is prior Beta distributed with hyperparameters α_{μ_g} and β_{μ_g},

$$p(\mu_g|\alpha_{\mu_g}, \beta_{\mu_g}) = Beta(\mu_g|\alpha_{\mu_g}, \beta_{\mu_g}) \tag{4}$$

$$p\left(s|\alpha_s, \beta_s\right) = Beta\left(s|\alpha_s, \beta_s\right) \tag{5}$$

The application of the priors help prevent over-fitting and avoid problems for the binomial due to fewer data points in genotype states of higher copy number regions. s and μ_g are both initialized as the MAP estimate using the pseudocounts alone.

1.4.3 Position-specific transition model

The transition component of the framework uses a position-specific transition matrix $\mathbf{A_t}$ which is an 18×18 matrix specifying probabilities for distance-dependent (Colella et al., 2007) and copy-number permitted transitions between genotypes at each position t. We employ a unique set of transition probabilities for each

position t to capture two key ideas which are each encoded in matrix $\mathbf{T_t}$ and indicator function C_t. Rows of $\mathbf{A_t}$ are normalized such that they sum to 1.

$$\begin{aligned} p(G_t = j | G_{t-1} = i, c_t) &= T_t(i,j) \times C_t(j) \\ &= A_t(i,j) \end{aligned} \tag{6}$$

1.4.4 Genotype transitions

The genotype state transition is specified by a position-specific stochastic transition matrix, $\mathbf{T_t} \in \mathbb{R}^{18\times 18}$, for each position t. There are two design goals that require modeling probabilities at each site:

1. The genomic distance between adjacent positions in $\mathbf{P}$ are non-uniform, thus, we employ a distance dependent strategy used in (Colella et al., 2007) whereby the transition probabilities are generated by an exponential function modeling a priori knowledge for transitioning between genetic events. The distance d_t between positions t and $t-1$ in base-pairs, and the expected length between transitions, L are used to define a function ρ_t. L was determined empirically by observing the average length of segments in 104 breast tumours (Shah *et al.* submitted), which was 2 Megabases (Mb) rounded to the nearest Mb.
$$\rho_t = 1 - \frac{1}{2}\left[1 - e^{(\frac{-d_t}{2L})}\right] \tag{7}$$

2. The genotype transition matrix applies strong probabilities (ρ_t) for self-transitions and transitions between genotypes of same zygosity status. For example, genotype states AA, BBB, and AAAAA have LOH zygosity status, and therefore should have similar transition probabilities. More formally, a transition from genotype $G_{t-1} = i$ to a genotype $G_t = j$ such that i and j have same zygosity status, should have probability ρ_t.
$$T_t(i,j) = \begin{cases} \rho_t & i = j \ or \ sameZS(i,j) \\ \frac{1-\rho_t}{|K_{c_t}|-1} & otherwise \end{cases} \tag{8}$$
where $sameZS(i,j)$ is a function that returns *true* if genotype states i and j have same zygosity status.

1.4.5 Copy number transitions

In order to capture the transitions between genotypes of different copy number, we use a position-specific indicator function, C_t, which defines allowable genotype transitions from $G_{t-1} = i$ into $G_t = j$ such that j can only be one of K_{c_t} for the given c_t at position t.

$$C_t(j) = \begin{cases} 1 & j \in K_{c_t} \\ 0 & otherwise \end{cases} \tag{9}$$

If uncertainty in copy number is provided in the form of a vector of probabilities at each position, then the copy number transition matrix can be remodeled to incorporate this information. Effectively, rather than using binary values, C_t becomes a soft weighting matrix that introduces probability mass into all genotypes in $\mathbf{K}$,

$$C_t(j) = p(c_t), j \in K_{c_t}, \forall c_t \tag{10}$$

1.4.6 Learning and inference

We employ the expectation maximization (EM) approach for estimating model parameters, $\theta = \{s, \mu_{1:18}, \pi_{1:18}\}$. In the expectation step, we compute the expectation of the complete-data likelihood resulting in the posterior marginal probabilities ${\gamma_\mathbf{t}}^{(n)}$,

$$p(\mathbf{G_t}|\mathbf{a}, \mathbf{N}, \theta^{(n-1)}) = \frac{p(\mathbf{a}, \mathbf{N}|\mathbf{G_t}, \theta^{(n-1)})p(\mathbf{G_t}|\theta^{(n-1)})}{p(\mathbf{a}, \mathbf{N}|\theta^{(n-1)})} \tag{11}$$

$\forall t \in P$ using the current settings of parameters, $\theta^{(n-1)}$. This calculation is done efficiently using the scaled version of the forwards-backwards algorithm (Bishop, 2007),

$$\gamma_t(g) = \frac{f_t(g)b_t(g)}{\sum_{t=1}^{T} \log w_t} \tag{12}$$

where f_t and b_t are the forward and backward probabilities, respectively, at position t and genotype state $g \in \mathbf{K}$. Using scaled forward/backward probabilities at each position t such that they are normalized avoids probabilities that quickly decrease to zero. Keeping track of the normalizing constant, w_t, at each position of the forward propagation conveniently gives us the log-likelihood, $p(\mathbf{a}, \mathbf{N}|\theta^{(n-1)}) = \sum_{t=1}^{T} \log w_t$.

The sum of the expected complete log-likelihood and the log priors gives the objective function for iteration n of the M-step where the expectation is taken with respect to $\mathbf{G}|\mathbf{a}, \mathbf{N}$,

$$\begin{aligned}
Q^{(n)} &= \sum_{k=1}^{K} p\left(G_0 = k|\mathbf{a}, \mathbf{N}, \theta^{(n-1)}\right) \log Multinomial\left(G_0|\pi^{(n-1)}\right) \\
&+ \sum_{t=1}^{T}\sum_{i=1}^{K}\sum_{j=1}^{K} p(G_t = i, G_{t-1} = j|\mathbf{a}, \mathbf{N}, \theta^{(n-1)}) \log \mathbf{A_t}(i,j) \\
&+ \sum_{t=1}^{T}\sum_{k=1}^{K} p\left(G_t = k|\mathbf{a}, \mathbf{N}, \theta^{(n-1)}\right) \log Binomial\left(a_t|\mu_k^{(n-1)}, N_t\right) \\
&+ \log Beta\left(s^{(n-1)}|\alpha_s, \beta_s\right) \\
&+ \sum_{k=1}^{K} \log Beta\left(\mu_k^{(n-1)}|\alpha_{\mu_k}, \beta_{\mu_k}\right) \\
&+ \log Dirichlet\left(\pi^{(n-1)}|\delta^\pi\right)
\end{aligned} \tag{13}$$

In the maximization step, we use the maximum a posteriori (MAP) estimate to update the parameters, $\theta^{(n)} = \arg\max_\theta \left\{p(\theta^{(n-1)}|\mathbf{a}, \mathbf{N})\right\} = \left\{s^{(n)}, \mu_{1:18}^{(n)}, \pi_{1:18}^{(n)}\right\}$. $\mathbf{A_t}$ is fixed and not re-estimated. The update equation for the initial state distribution π is

$$\pi_k^{(n)} = \frac{\gamma^{(n)}\left(G_0 = k\right) + \delta^\pi\left(k\right) - 1}{\sum_{k'=1}^{K}\left(\gamma^{(n)}\left(G_0 = k'\right) + \delta^\pi\left(k'\right) - 1\right)} \tag{14}$$

The normal proportion and tumour allelic ratio parameters of the binomial observation model are derived by maximizing Q^n and taking partial derivatives w.r.t s and μ_k for a given genotype state $k \in K$, equating

to zero, and solving for the parameters.

$$\frac{\partial Q^{(n)}}{\partial \mu_k} = (1-s)\left(\frac{\bar{a}_k}{s\mu_N + (1-s)\,\mu_k} - \frac{\bar{b}_k}{1 - s\mu_N - (1-s)\,\mu_k}\right) + \left(\frac{\alpha_{\mu_k} - 1}{\mu_k} - \frac{\beta_{\mu_k} - 1}{1-\mu_k}\right) \tag{15}$$

$$\frac{\partial Q^{(n)}}{\partial s} = \sum_{k=1}^{K}\left((\mu_N - \mu_k)\left(\frac{\bar{a}_k}{s\mu_N + (1-s)\,\mu_k} - \frac{\bar{b}_k}{1 - s\mu_N - (1-s)\,\mu_k}\right)\right) + \left(\frac{\alpha_s - 1}{s} - \frac{\beta_s - 1}{1-s}\right) \tag{16}$$

where $\bar{a}_k = \sum_{t=1}^{T} \gamma^{(n)}\,(G_t = k)\,a_t$ and $\bar{b}_k = \sum_{t=1}^{T} \gamma^{(n)}\,(G_t = k)\,(N_t - a_t)$. The EM convergence criteria is met when $F^{(n)} - F^{(n-1)} < threshold$, where F is sum of the log-likelihood and the log priors,

$$\begin{aligned} F^n &= log\left(p(\mathbf{a}, \mathbf{N}|\theta^{(n)})\right) \\ &+ \log Beta\left(s^{(n-1)}|\alpha_s, \beta_s\right) \\ &+ \sum_{k=1}^{K} \log Beta\left(\mu_k^{(n-1)}|\alpha_{\mu_k}, \beta_{\mu_k}\right) \\ &+ \log Dirichlet\left(\pi^{(n-1)}|\delta^{\pi}\right) \end{aligned} \tag{17}$$

The converged parameters $\hat{\theta}$ are used to infer the optimal hidden state path of genotypes using the Viterbi algorithm,

$$G_{1:T} = \arg\max_{G}\left\{p\left(\mathbf{G}|\mathbf{a}, \mathbf{N}, \hat{\theta}\right)\right\} \tag{18}$$

Finally, the zygosity state can be decoded, resulting in the final sequence of zygosity status, $ZS_{1:T}$.

1.4.7 Implementation

APOLLOH is implemented in Matlab; the forward-backward and Viterbi algorithms are implemented in C. The run-time ($\mathcal{O}\left(K^2T\right)$) and memory ($\mathcal{O}\left(KT\right)$) usage of the algorithm for about 1.5 million positions on a single-core is about 20 minutes and 3GB based on an average of 41 iterations.

The source code and compiled executable (usable on Linux x64 architecture) can be downloaded at `http://compbio.bccrc.ca/software/apolloh/`

1.5 Copy number analysis of WGSS data

The WGSS-derived copy number results used as input in APOLLOH were generated using a modified version of a paired tumour-normal strategy described previously (Chiang et al., 2009; Shah et al., 2009). This algorithm, called HMMcopy, was developed in-house and can be downloaded at `http://compbio.bccrc.ca/software/hmmcopy/`. The genome was divided into fixed windows of 1kb, and read depth is extract for each window in the tumour and normal. In this study, we applied two additional preprocessing steps prior to segmentation in order to achieve more accurate copy number estimates. First, we applied a filter to remove repetitive regions that are highly mappable. Second, we corrected GC content bias to remove wave-like patterns in the tumour and normal, separately, using a loess curve fit between GC content and read depth. A log ratio is computed for each window by computing proportion between the GC corrected tumour value and GC corrected normal value.

Divide the genome into fixed genomic windows

First, the genome was divided into fixed genomic windows of 1kb, reducing the analysis to a set of approximately 2.5 to 3 million loci, $\mathbf{R}$.

Extract read depth for normal and tumour

Next, separately for the normal and tumour genomes of each patient, we extracted the total read depth for each window in $\mathbf{R}$ using BAMtools (Barnett et al., 2011), resulting in a vector of read count data from the normal, $\mathbf{N_R} = \left(n_1, \ldots, n_{|R|}\right)$ and tumour, $\mathbf{T_R} = \left(t_1, \ldots, t_{|R|}\right)$, respectively.

Removing windows that are highly mappable

Using "ENCODE Duke Uniqueness of 35bp sequences" track from UCSC (`http://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg18&g=wgEncodeMapability`), we filtered windows that were within repetitive regions, resulting in being highly mappable by aligners. Windows that had mappability score of ≥ 0.9 were excluded. This removed extreme amplified positions which would have otherwise posed as confounding outliers in downstream segmentation analysis.

GC content correction of normal and tumour read counts

We performed GC content bias correction to the tumour and normal of each patient separately. We applied a global loess fit between GC content and read depth for windows in $\mathbf{R}$. Due to computational restrictions of fitting 3 million data points, we further excluded outlier windows based on read depths being in the upper and lower 1% quantile, and randomly sampled 20,000 of the remaining windows for generating the loess curve. Finally, the read depth of all windows, $\mathbf{N_R}$ and $\mathbf{T_R}$, are corrected by scaling the observed value by the loess fitted value (Equation 19).

$$corrected\, read\, depth = \frac{observed\, read\, depth}{loess\ fitted\ value} \tag{19}$$

Normalizing copy number in tumours

The GC corrected normal $\mathbf{N_R}$ and tumour $\mathbf{T_R}$ counts are normalized independently to generate $\bar{\mathbf{N}}_\mathbf{R} = \left(\bar{n}_1, \ldots, \bar{n}_{|R|}\right)$ and $\bar{\mathbf{T}}_\mathbf{R} = \left(\bar{t}_1, \ldots, \bar{t}_{|R|}\right)$, respectively, where $\bar{n}_i = \frac{n_i}{\sum_j n_j}$ and $\bar{t}_i = \frac{t_i}{\sum_j t_j}$, $i \in \{1, \ldots, |R|\}$. To obtain the final tumour copy number observed at each loci $r \in R$, we applied another normalization step by taking the log2 ratio between tumour and normal copy number, $\mathbf{T_N}(\mathbf{i}) = log_2\left(\frac{\bar{t}_i}{\bar{n}_i}\right),\ i \in \{1, \ldots, |R|\}$.

Segmentation and copy number prediction via HMM

A 6-state version of a hidden Markov model (HMM) approach (Shah et al., 2006) was used to segment the input data $\mathbf{T_N}$. Initialization and hyperparameters for the prior means of the Student's-t distribution used for this analysis were $log2((1, 1.4, 2, 2.7, 3, 4.5])/2)$.

1.6 Application of APOLLOH to 23 triple negative breast cancers.

The analysis workflow is described in Methods and Figure S2. For extracting heterozygous positions from the normal genomes, default settings were used for GATK's UnifiedGenotyper (McKenna et al., 2010). Heterozygous genotypes were accepted based on "PASS" or "." in the reported UnifiedGenotyper 'Quality' field. In the tumours, pileups were generated using SAMtools (Li et al., 2009) and reads in the corresponding normal heterozygous positions were filtered by base quality of 10 and mapping quality of 20. Low depth (>10 reads) and read-sink (<200 reads) positions were excluded.

APOLLOH parameters (described in Table S1) used in this analysis are given in a configuration file packaged with the software (`http://compbio.bccrc.ca/software/apolloh/`).

1.7 OncoSNP analysis of Affymetrix SNP6.0 analysis

Affymetrix SNP6.0 genotyping arrays were analyzed for the 23 breast cancers. We determined regions of loss of heterozygosity (LOH) using the OncoSNP software v1.0 Beta (Internal Release v2.19) (Yau et al., 2010). Due to the absence of a complete set of matched normals, the unpaired tumour analysis was used. OncoSNP was adapted for the Affymetrix SNP6 platform by using initial parameters settings obtained by training 45 COSMIC (Bignell et al., 2010) breast cancer samples hybridized to SNP6 arrays. Log ratios and B-allele frequencies were obtained from PennCNV-Affy (Wang et al., 2007) normalization results. OncoSNP hyperparameters used in the analysis was provided as a standard configuration file with the downloaded software. Other OncoSNP settings included using 15 iterations of expectation maximization, 30 sub-sampling, and stromal setting activated. Two to eight copy 'Somatic' LOH predictions made by OncoSNP were consolidated as simply LOH while 'Mono-allelic amplification' states were relabelled as 'Allele-specific copy number amplifications' (ASCNA).

1.8 Analyses for comparing APOLLOH results and Affymetrix SNP6 data

1.8.1 WGSS and SNP6 platform comparison

For each predicted APOLLOH segment x with boundaries x_{start} and x_{end} and segment median allelic ratio a_x, we computed the SNP6 median BAF y_x for probes that overlapped x as,

$$y_x = median(BAF(p)),\ \{p : p_{start} \geq x_{start}, p_{end} \leq x_{end}\} \tag{20}$$

Spearman's rank correlation is computed on $y_{1:X}$ and $a_{1:X}$ where X is the total number APOLLOH predicted segments.

To measure association with the dynamic range of clusters as shown in Figure **??** and Figure S3, we computed Euclidean distance of the class centroids between LOH and HET and calculated the Spearman rank correlation statistic with APOLLOH estimates of normal proportion s (Figure S4).

1.8.2 Model evaluation using SNP6 predictions

SNP6 LOH results, predicted using the OncoSNP (Yau et al., 2010) software, were used as truth for evaluating APOLLOH model variants and SNVMix. OncoSNP was run using parameters and settings designed for the Affymetrix SNP6.0 platform. Predicted states were redefined into comparable classes: deletion, neutral and amplified LOH; allele-specific copy number amplification; and heterozygous. Positions that intersected between the loci used in APOLLOH for each sample and the probes of the SNP6 array were used for evaluation. Homozygous positions predicted by OncoSNP or HMMcopy (Supplemental Methods)

were excluded from the evaluation. True positives (TP) are defined as positions that were predicted as LOH by both APOLLOH and OncoSNP; false positives (FP) are positions where APOLLOH predicted LOH but were predicted as HET or ASCNA by OncoSNP; false negatives (TN) make up positions that APOLLOH called HET/ASCNA but OncoSNP predicted as LOH. Precision (TP/(TP+FP)), recall (TP/(TP+FN)), and F-measure (Equation 21) were calculated for each sample and APOLLOH model variant. For ASCNA performance, positives were ASCNA for APOLLOH (full model) and HET for APOLLOH-noCN.

$$F\ measure = \frac{2 \times precision \times recall}{precision \times recall} \tag{21}$$

Evaluation using exome data was computed the same as above. LOH predictions for the exome data were generated using the full APOLLOH model.

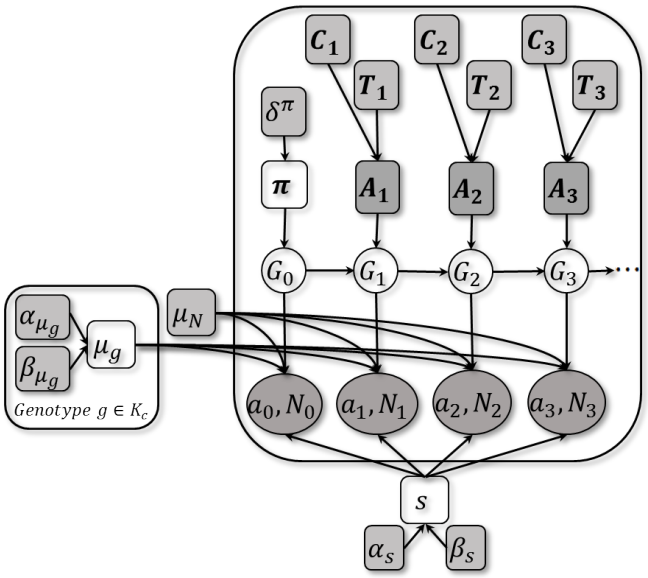
1.9 Comparison of transcriptome allelic ratios (TAR)

RNAseq pileups for each sample were generated using SAMtools (Li et al., 2009), and positions that intersected the loci used in the APOLLOH analysis (i.e. heterozygous positions in the normal genome) of each WGSS sample were extracted. Transcriptome allelic ratios were first converted to symmetric counts,

$$AR_i = \frac{max(a_i, b_i)}{a_i + b_i} \tag{22}$$

where a is the reference count and b is the non-reference count for each position i. Each position of the RNAseq is then classified with the corresponding zygosity based on the APOLLOH call of the same loci in the genome.

2 Supplemental Figures



$$p(G_0|\pi) = Multinomial(G_0|\pi)$$
$$P(\pi|\delta^\pi) = Dirichlet(\pi|\delta^\pi)$$

$$p(a_t|G_t = g, c_t) = Binomial(a_t|\bar{\mu}_g, N_t)$$
$$p(\mu_g|\alpha_{\mu_g}, \beta_{\mu_g}) = Beta(\mu_g|\alpha_{\mu_g}, \beta_{\mu_g})$$
$$p\left(s|\alpha_s, \beta_s\right) = Beta\left(s|\alpha_s, \beta_s\right)$$
$$\bar{\mu}_g = s \cdot \mu_N + (1 - s) \cdot \mu_g$$

$$p(G_t = j|G_{t-1} = i) = A_t(i, j) = T_t(i, j) \times C_t(i, j)$$

$$C_t(i, j) = \begin{cases} 1 & i \in K_l, j \in K_k \\ 0 & otherwise \end{cases}$$

$$T_t(i, j) = \begin{cases} \rho & i = j \ or \ sameZS(i, j) \\ \frac{1-\rho}{|K_{c_t}|} & otherwise \end{cases}$$

Figure 1: Probabilistic graphical model of APOLLOH. Shaded nodes are known or observed quantities; open nodes are random variables of unknown quantities. Arrows represent conditional dependence between random variables. The latent variables $G_t \in \mathbf{K}$ represent the genotype state at position t. π is the initial state distribution on G_0 and is Dirichlet distributed with hyperparameter δ^π. APOLLOH is an HMM that employs a position-specific transition model $\mathbf{A_t}$ that is fixed with transition probabilities in T_t that are restricted based on copy number from C_t. The emission component models the number of symmetric reference counts a_t and depth N_t, using a mixture of binomial distributions conditional on $G_t = g$. μ_g and s are Beta (prior) distributed with hyperparameters α_g and β_g. Parameters μ, s and π are estimated using expectation maximization, and the genotype sequence $G_{1:T}$ is inferred using Viterbi algorithm. See Table S1 for variable definitions.

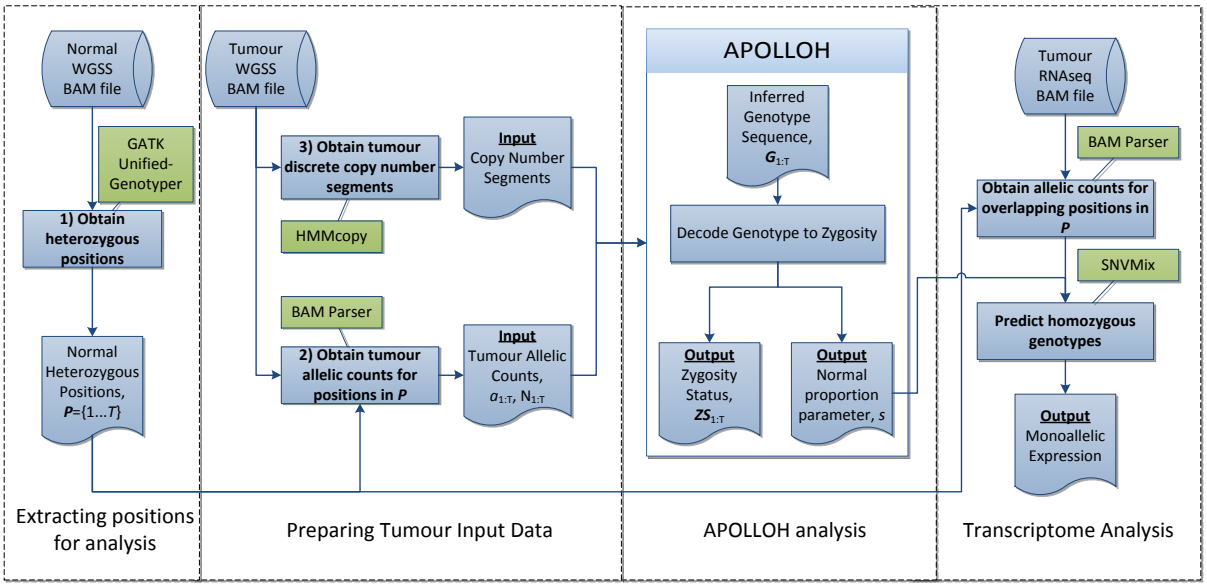


Figure 2: Workflow of the analysis for APOLLOH. Three inputs are required: 1) Heterozygous positions found in the normal DNA predicted by genotyping tools such as SNVMix(Goya et al., 2010); these genomic positions are the sites of interest in the analysis; 2) Reference counts at these positions in the tumour DNA sequencing data are obtained by extracting alignment read counts using SAMtools (Li et al., 2009); 3) Copy number status for the tumour are predicted by HMM-Dosage. APOLLOH uses the inputs to infer the genotype and subsequently zygosity status is determined for each position of interest. Transcriptome RNAseq data was analyzed for expressed allelic imbalance, and mono-allelic expression (MAE) was determined as homozygous genotypes using SNVMix (Goya et al., 2010).

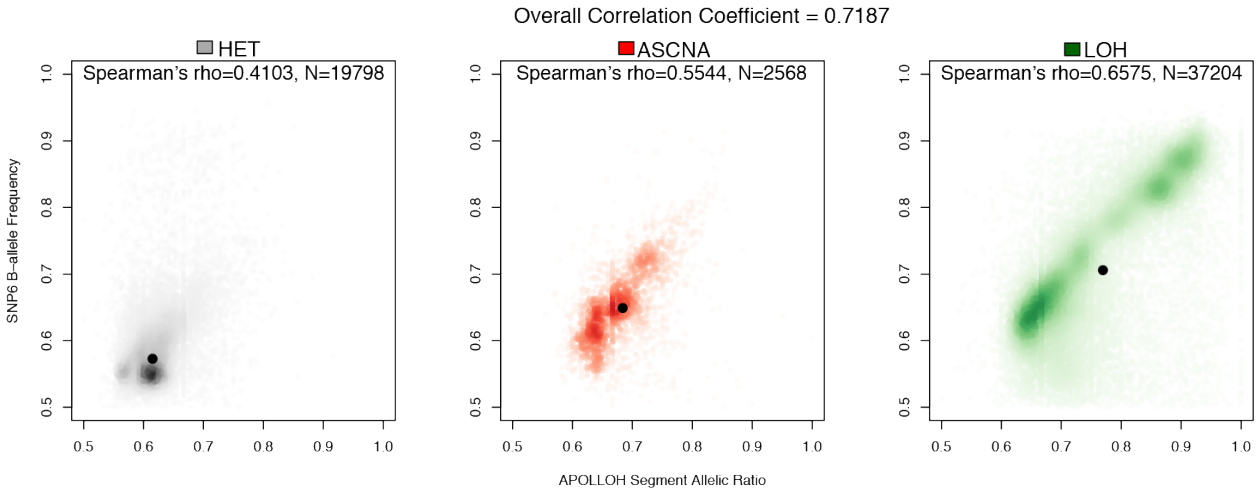


Figure 3: Benchmarking of WGSS allelic ratios against SNP6 genotyping array B-allele frequencies (BAF) for 23 breast cancer samples. Each datapoint represents an APOLLOH segment whose median allelic ratio is plotted against median BAF across probes that overlap the segment (Methods). WGSS allelic ratios ($max(refCount, nonRefCount)/depth$) and BAF ($max(A-intensity, \_B-intensity)/Total-intensity$) were computed as symmetric values. LOH allelic ratios and B-allele frequencies are distributed within a range, which is likely due to differing normal cell contamination proportions. The Spearman rank correlation coefficient was 0.72 ($p < 0.001$).

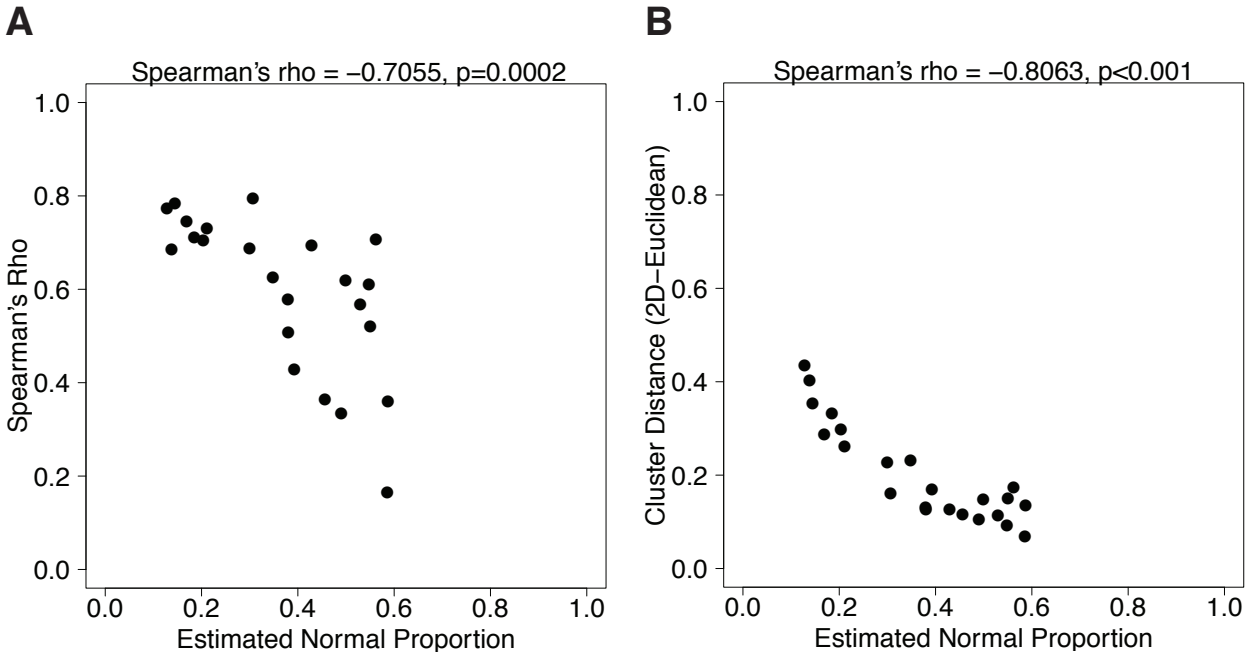


Figure 4: Correlation of estimated normal proportion with cluster distances (A) and correlation (B) from APOLLOH WGSS allelic ratios and SNP6 B-allele frequency (BAF). (A) For each of the 23 breast cancer samples, Euclidean distance was computed between cluster centroids (2-dimensional median) of APOLLOH predicted LOH and HET classes. Example clusters are shown in Figure 3A. Spearman rank correlation (rho = -0.8063, $p < 0.001$) was computed between euclidean distances and estimated normal proportion parameter s across the 23 samples. (B) The correlation between APOLLOH segment allelic ratio and SNP6 BAF are plotted against normal proportion for the 23 samples. Examples of correlations for 3 samples are shown in Figure 3A. The association between APOLLOH-SNP6 correlation and estimated normal contamination is significantly, negatively correlated (rho = -0.7055, $p < 0.001$).

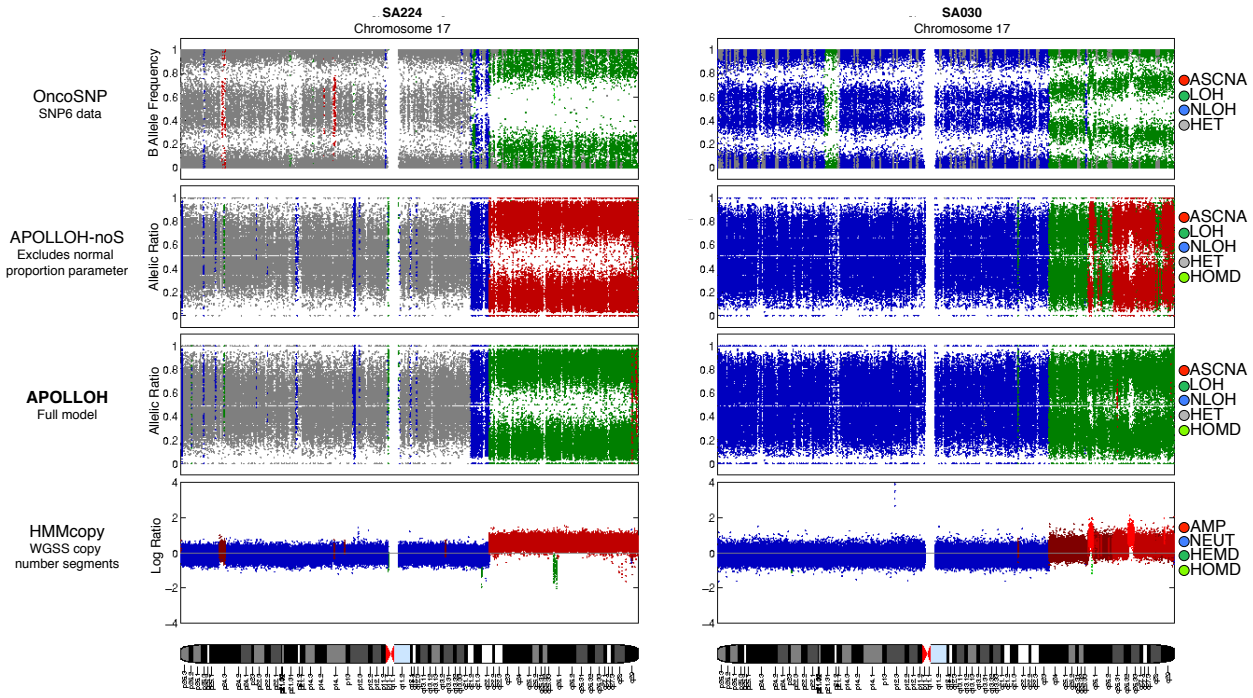


Figure 5: Examples of improved results when accounting for normal contamination in APOLLOH. OncoSNP (Yau et al., 2010) results was used as ground truth. APOLLOH-noS is the model that does not estimate the normal proportion parameter. APOLLOH (full model) models normal contamination. HMMcopy is tool, designed in-house, to predict and segment copy number in tumour-normal samples (Supplemental Methods).

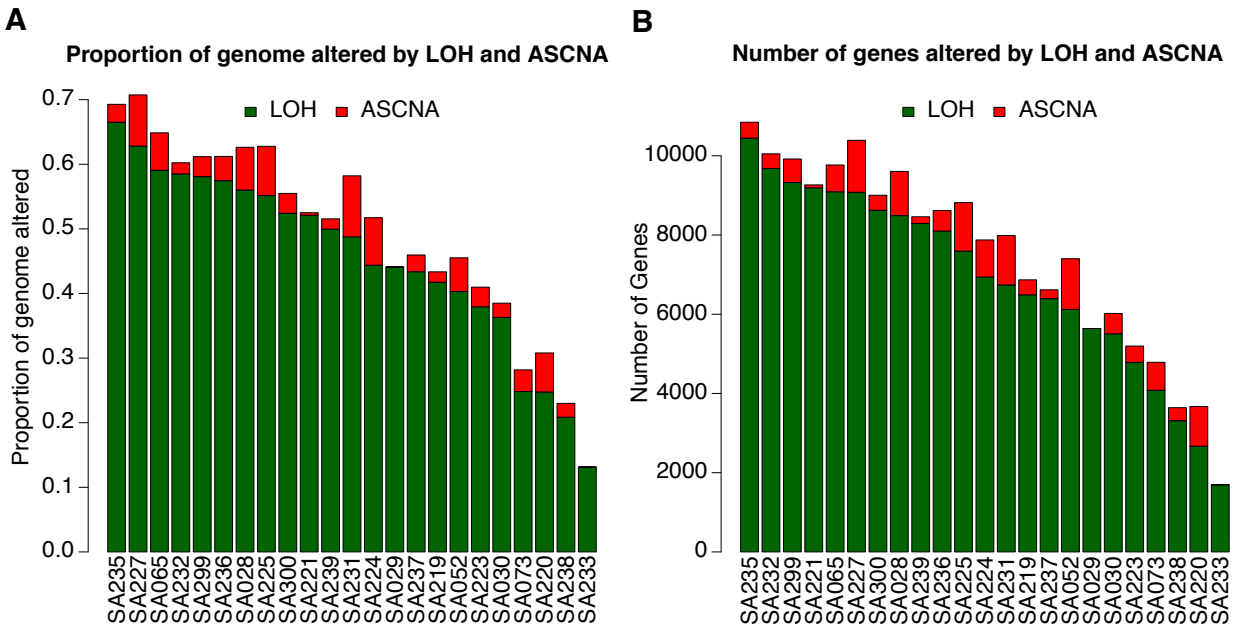


Figure 6: Distribution of the proportion of (A) genome altered (Table S6B) and (B) number of genes (Table S6C) by APOLLOH predicted LOH (green) and ASCNA regions (red). The proportion of the genome altered by LOH ranges from 13-67%, with a median of 49%. The number of genes altered by LOH ranges from 1694 to 10,446, with a median of 6941.

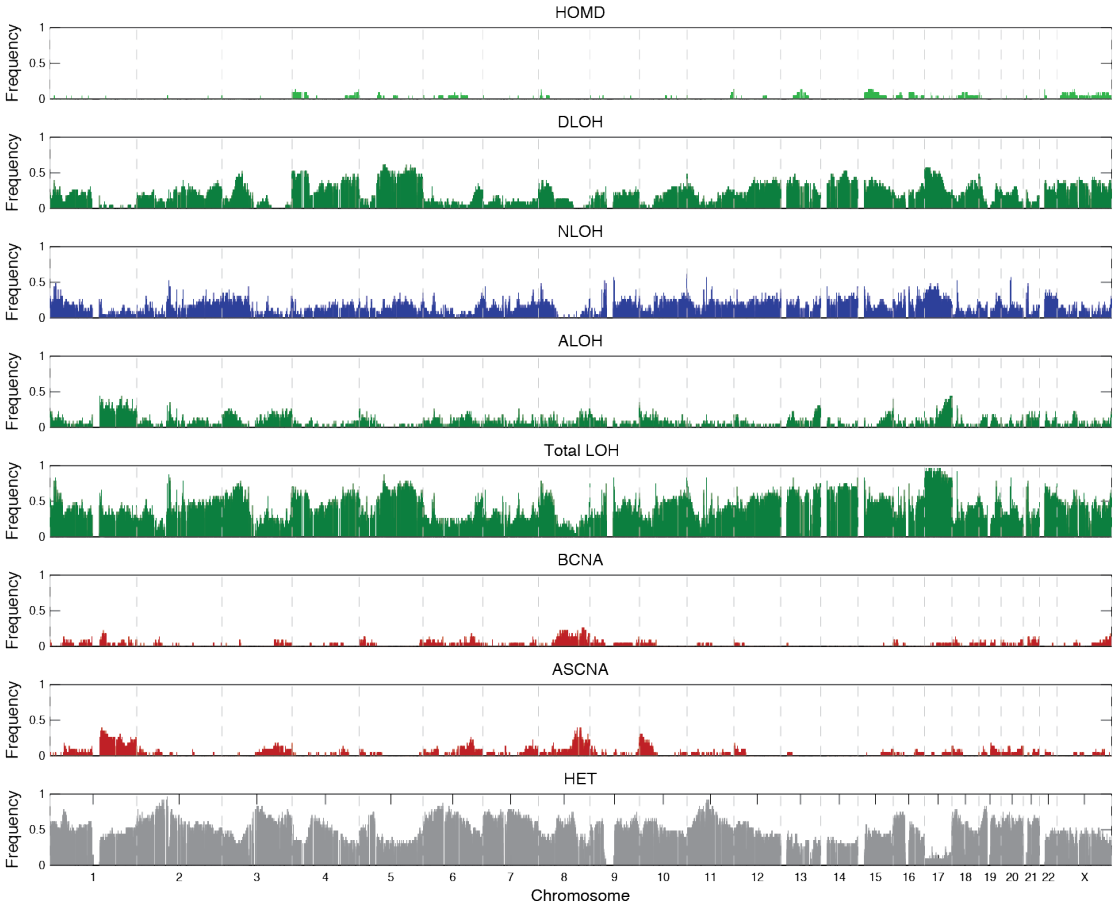


Figure 7: Genome-wide gene frequency landscape of APOLLOH loss of heterozygosity (LOH) predictions for 23 TNBC samples. Events are categorized into homozygous deletion (HOMD), deletion LOH (DLOH), copy neutral LOH (NLOH), amplification LOH (ALOH), overall LOH (Total LOH), balanced copy number amplification (BCNA), allele-specific copy number amplification (ASCNA), and heterozygous or retention (HET).

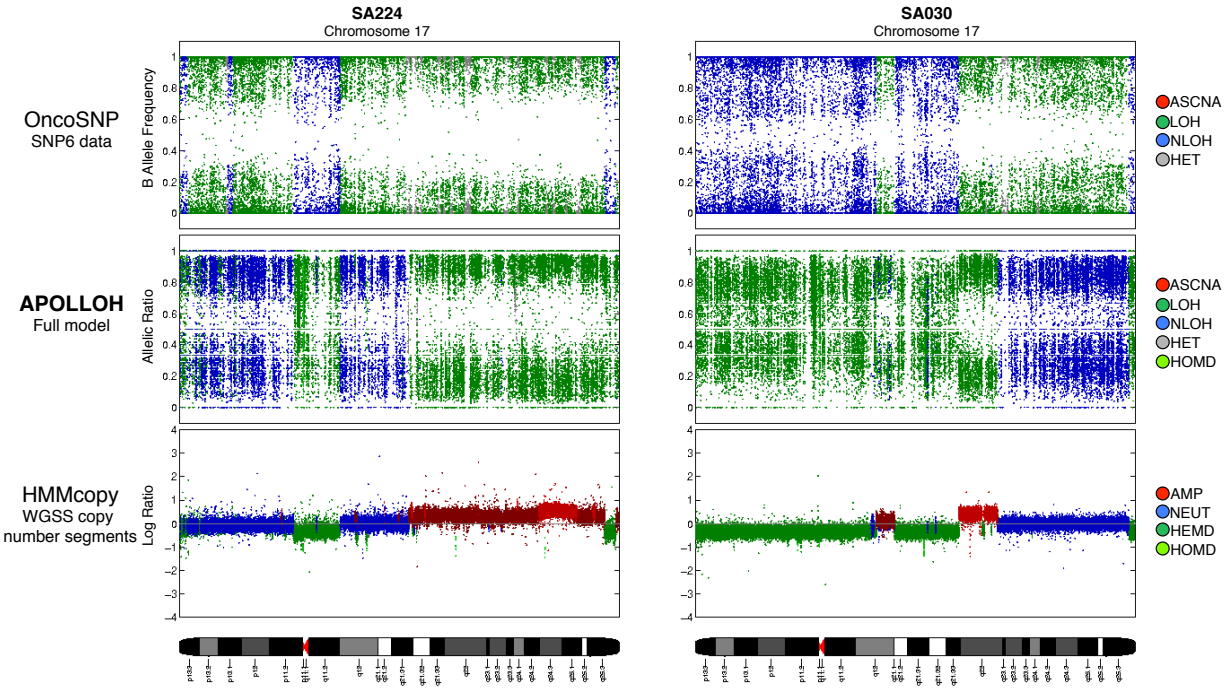


Figure 8: Examples of LOH events predicted within amplifications in chromosome 17. Sample SA224 contains a large amplification spanning q21.32 to q35 where only one allele is observed. SA030 contains a more focal amplification in q22 undergoing LOH.

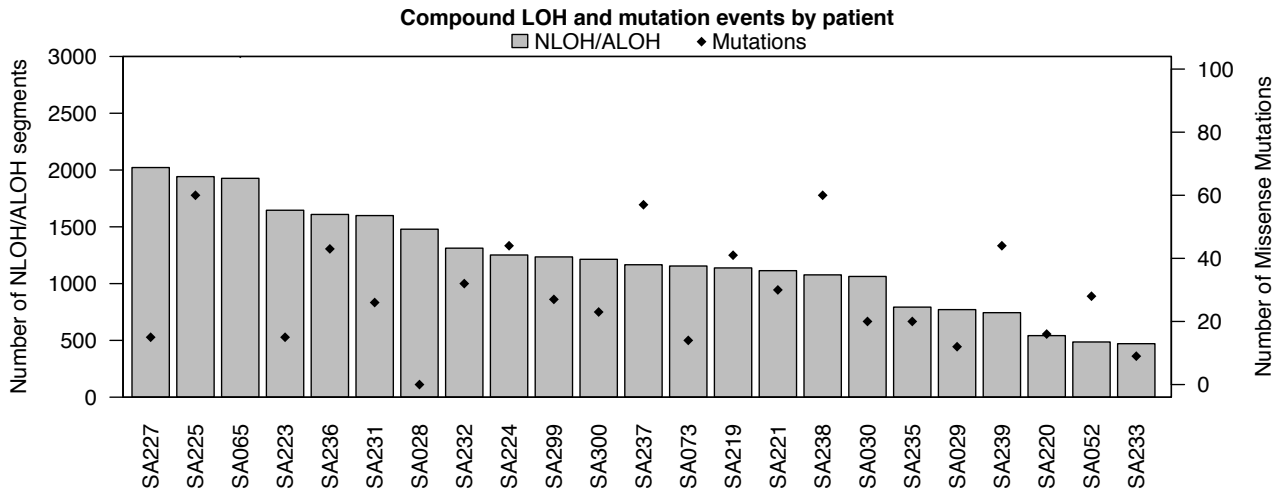


Figure 9: Number of compound copy number events and mutations. Compound events are defined as sequential copy number changes starting with a deletion and then duplication (NLOH) or amplification (ALOH) of the remaining allele. The left axis denotes the number of compound events. Mutations included in this figure are missense and nonsense mutations that have been validated (Shah et al., 2012) and predicted (see Methods). The right axis denotes the number of mutations.

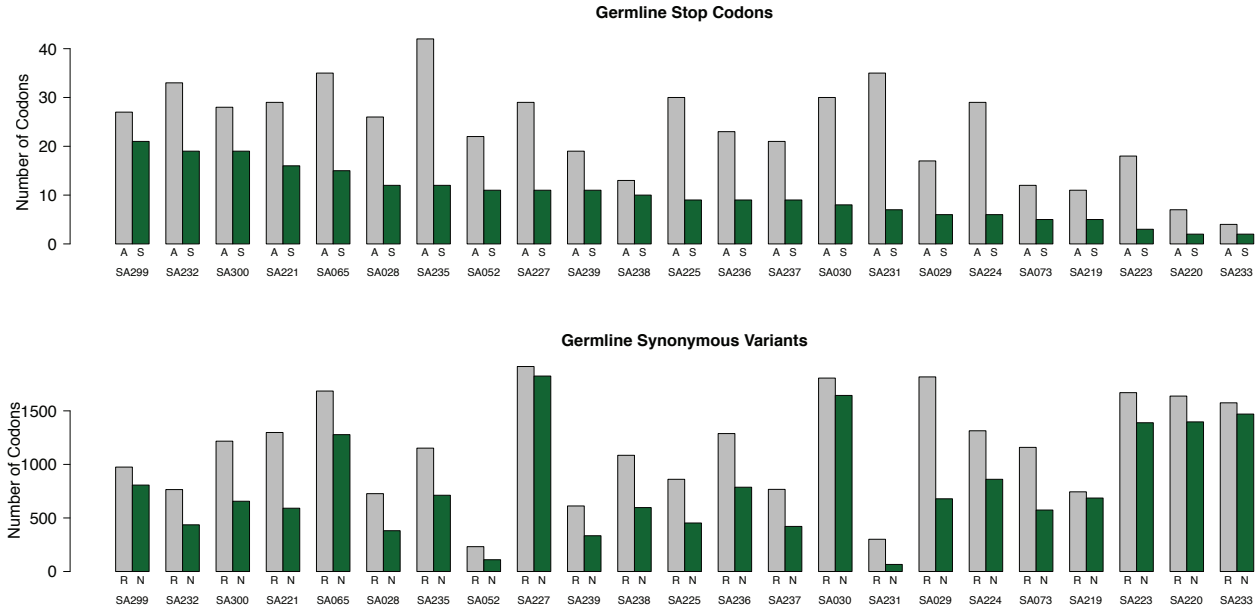


Figure 10: The number of genes with germline stop codon and synonymous variants that were affected by LOH is shown for each of the 23 tumour samples. Normal heterozygous positions in determined by GATK (and used in the APOLLOH analysis were annotated with codon effects using snpEff (Cingolani, 2012). For all variant loci that was annotated with having a stop codon and overlapped LOH regions, the remaining allele corresponding to the amino acid or the stop codon were labeled as ”A” and ”S”, respectively. For loci annotated as synonymous, the remaining allele corresponding to the reference and non-reference were labeled as ”R” and ”N”, respectively.

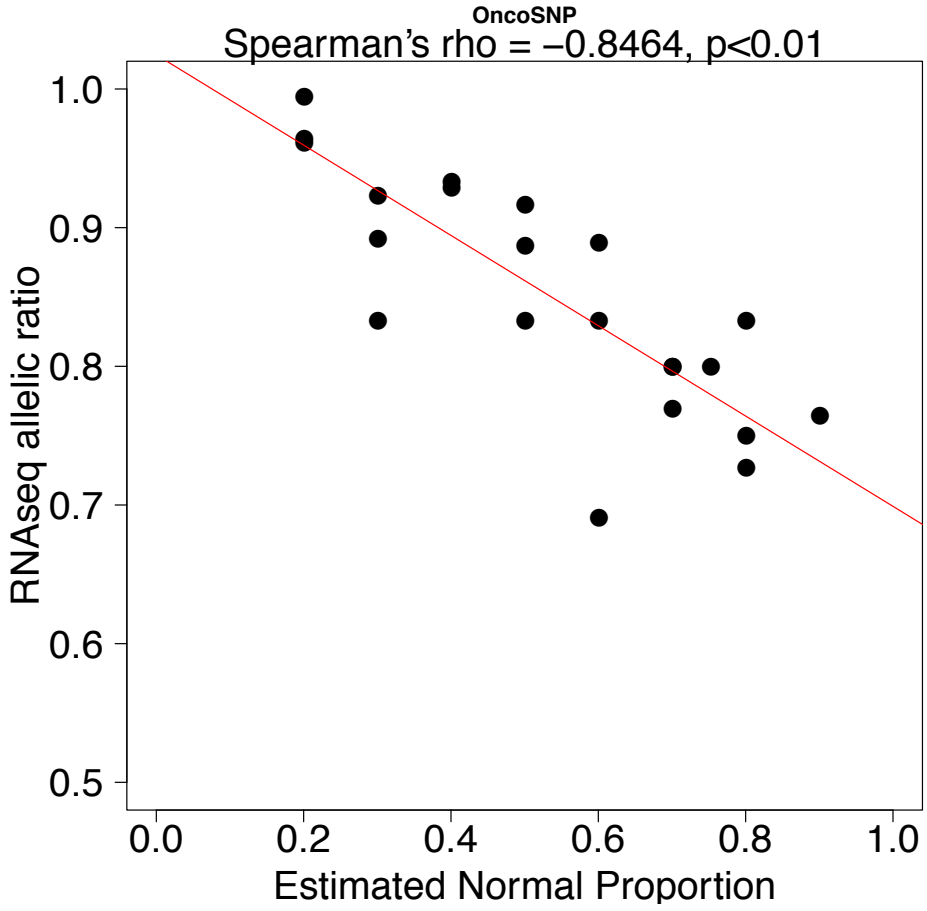


Figure 11: Correlation of OncoSNP predicted normal contamination and RNAseq allelic ratio. The median symmetric $(max(refCount, nonRefCount)/depth)$ allelic ratio of RNAseq data within predicted LOH segments for each sample is correlated (Spearman's rho = -0.85, $p < 0.01$) with the normal proportion parameter s estimated by OncoSNP (Yau et al., 2010). The first principal component line is shown in red.

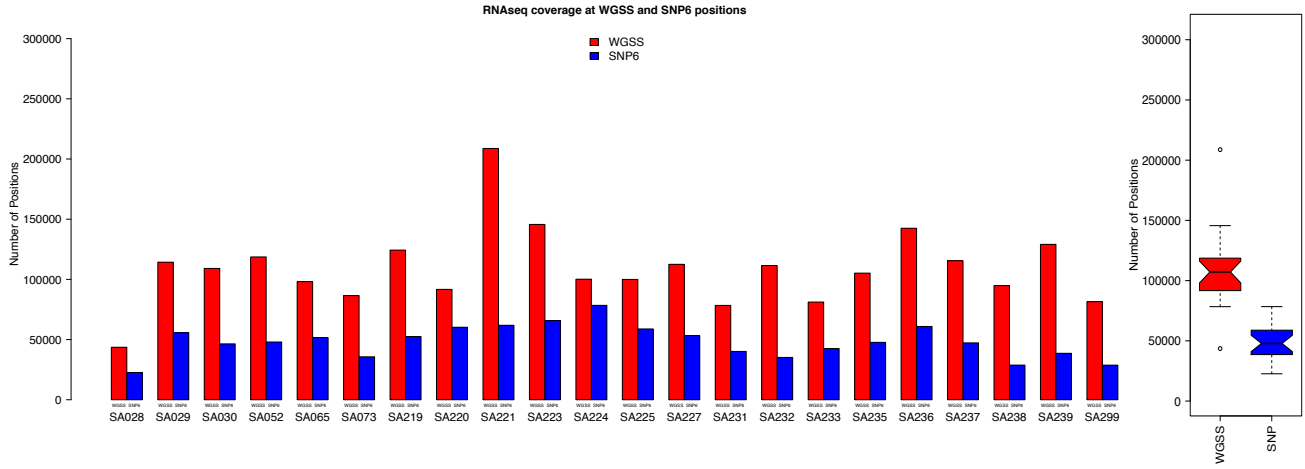


Figure 12: Number of WGSS and SNP6 probe positions with RNAseq coverage. The number of normal heterozygous positions predicted by GATK (McKenna et al., 2010) from the WGSS data (red) with coverage in the RNAseq data is consistently higher across the cohort when compared to the number of SNP probes on the Affymetrix SNP6 platform.

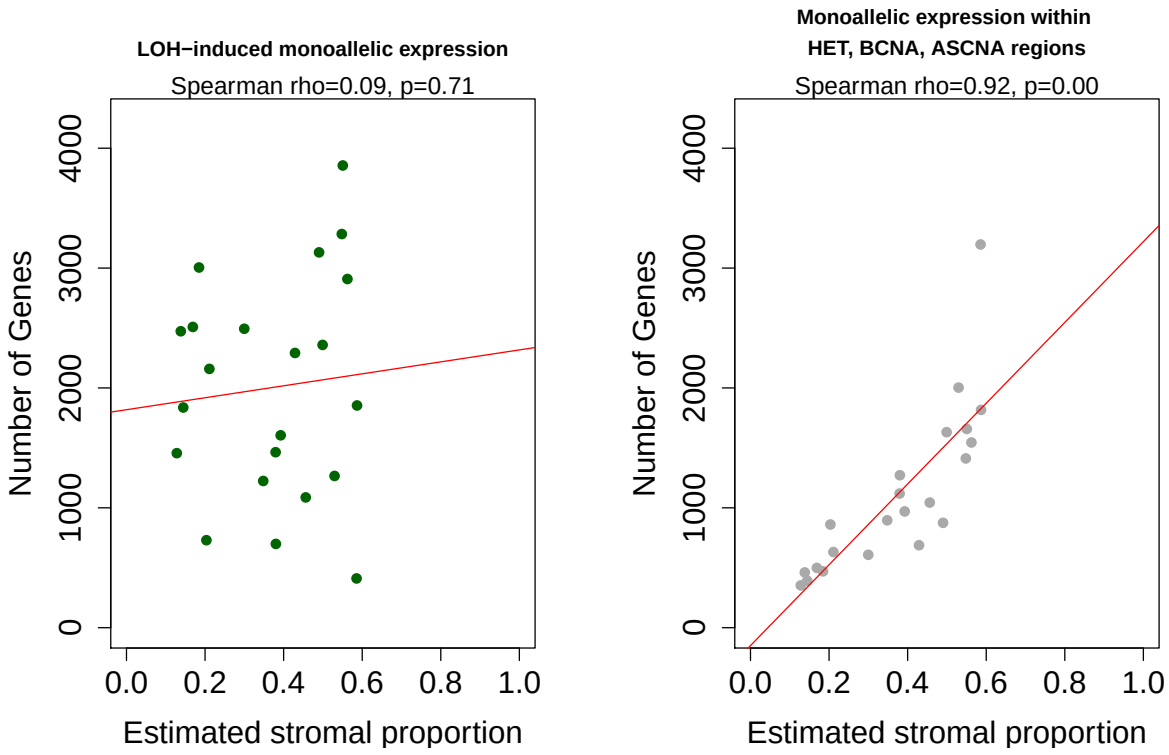


Figure 13: Number of monoallelic expression (MAE) of genes associated with estimated normal (stromal) contamination. The left plot represents MAE genes induced by genomic LOH. The right plot represents MAE genes that have balanced copy number amplification (BCNA), allele-specific copy number amplification (ASCNA), or heterozygous or retention (HET) genomic allelic imbalance states. The line of best fit is shown for each plot. The strong correlation for HET/BCNA/ASCNA MAE genes indicates that MAE may be due to germline epigenetic events that become easier to detect as normal cell content increases in the samples.

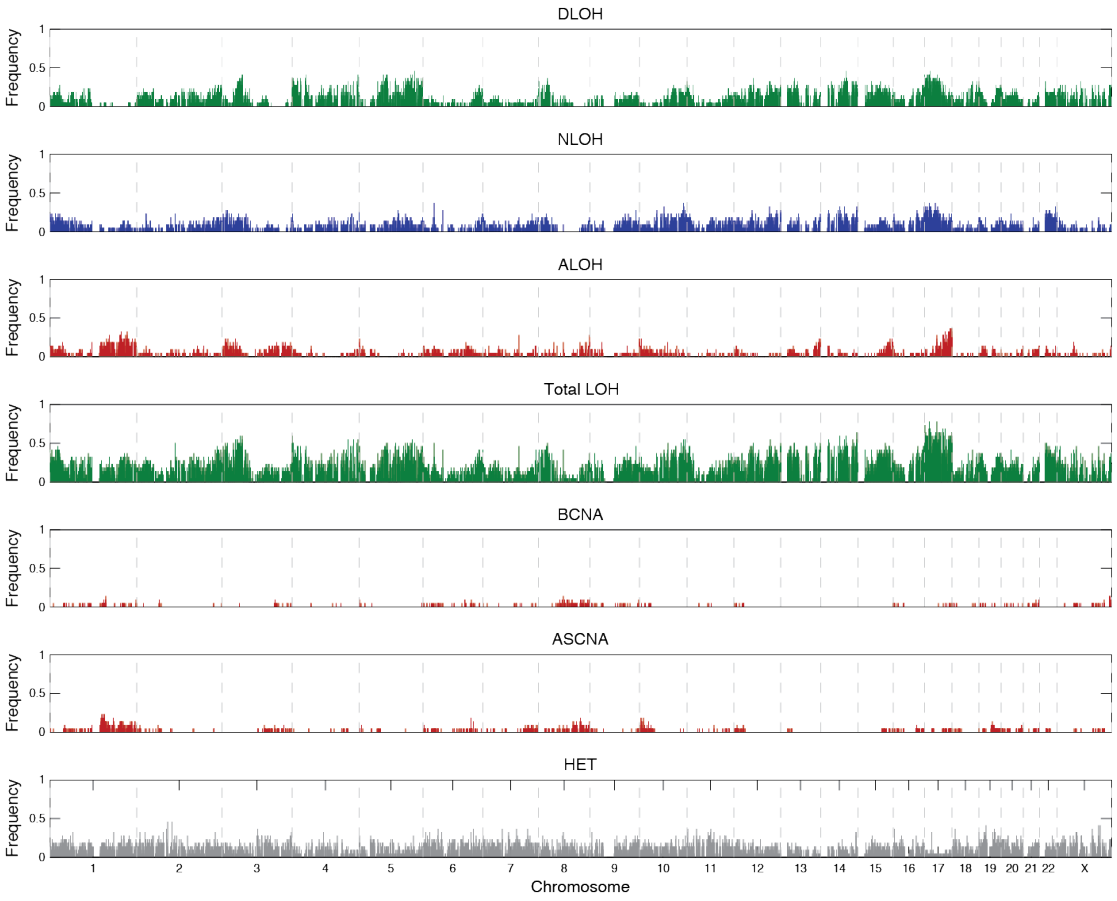


Figure 14: Genome-wide gene frequency landscape of monoallelic expression (MAE) as a consequence of loss of heterozygosity (LOH). MAE landscapes are categorized into the corresponding genomic events of deletion LOH (DLOH), copy neutral LOH (NLOH), amplification LOH (ALOH), overall LOH (Total LOH), balanced copy number amplification (BCNA), allele-specific copy number amplification (ASCNA), and heterozygous or retention (HET).

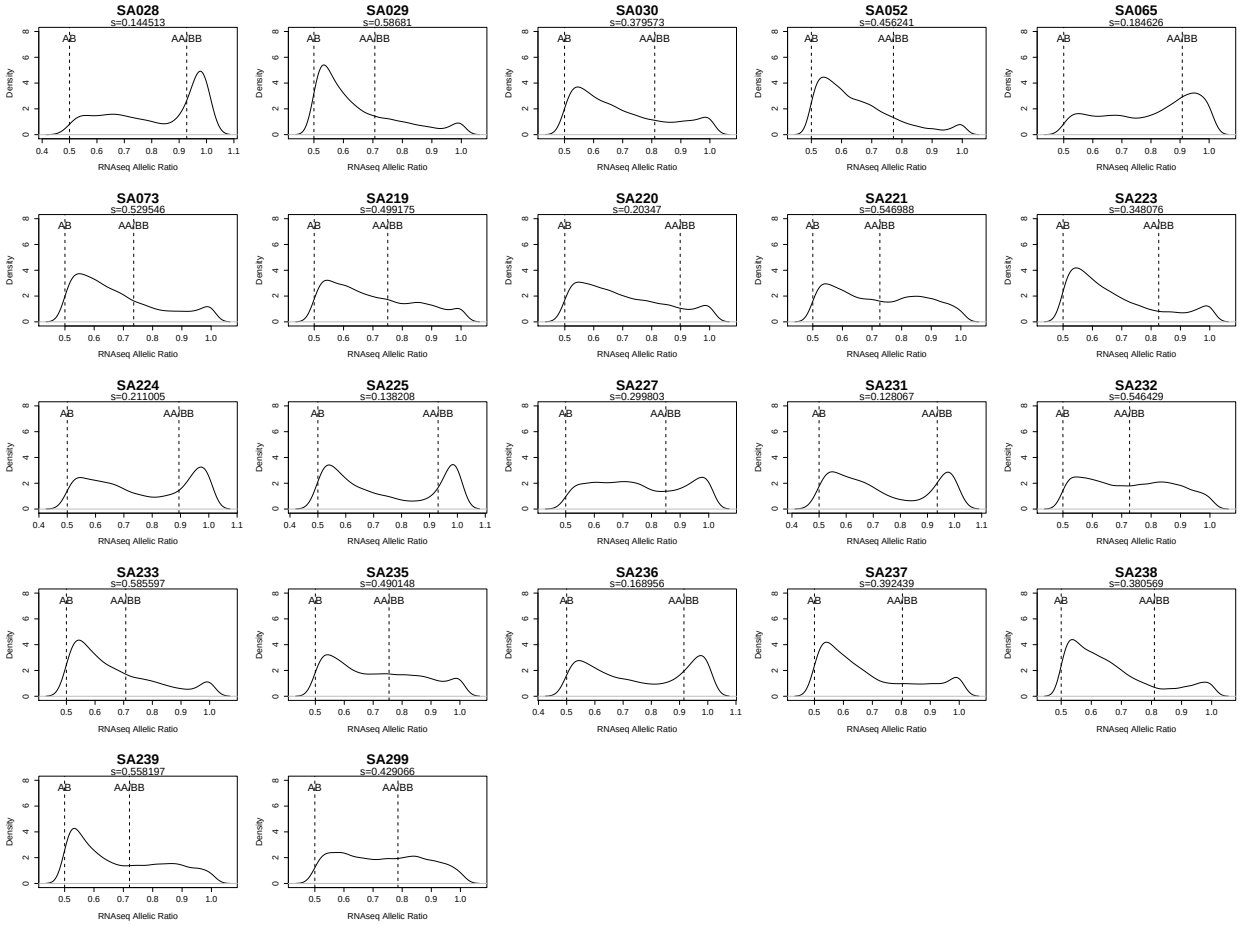


Figure 15: Transcriptome allelic ratio distribution and SNVMix parameters used for determining MAE. The distribution is for symmetric RNAseq allelic ratio, which is defined as $max(refCount, nonRefCount)/depth$. Binomial parameters are fixed using APOLLOH-inferred normal proportion parameters s such that $\mu_{aa} = s * 0.5 + (1 - s) * 1.0$, $\mu_{bb} = 1 - \mu_{aa}$, and $\mu_{ab} = s * 0.5 + (1 - s) * 0.5$. The dotted lines AA/AB and AB represent μ_{aa}/μ_{bb} and μ_{ab} for each case given s.

3 Supplemental Tables

Variable	Description	Source
π	Initial state distribution	Estimated by EM in M-step
δ^{π}	Prior counts; parameter of Dirichlet for π	User-defined
G_t	Latent variable for genotype at position t	Estimated by EM in E-step
a_t	Symmetric reference count at position t	Observed
N_t	Total read depth at position t	Observed
c_t	Copy number status at position t	Observed
μ_N	Normal reference allelic ratio genotype g	User-defined
μ_g	Tumour reference allelic ratio genotype g	Estimated by EM in M-step
α_{μ_g}	Hyperparameter of Beta prior on μ_g	User-defined
β_{μ_g}	Hyperparameter of Beta prior on μ_g	User-defined
s	Global stromal contamination proprotion parameter	Estimated by EM in M-step
α_s	Hyperparameter of Beta prior on s	User-defined
β_s	Hyperparameter of Beta prior on s	User-defined
$\mathbf{C_t}$	18×18 copy number transition matrix at t; determines allowable transition based on c_t	Fixed
$\mathbf{T_t}$	18×18 genotype transition matrix at position t; genomic distance-dependant probabilities	Fixed
$\mathbf{A_t}$	18×18 combined, copy-number restricted transition matrix, $\mathbf{C_t} \times \mathbf{T_t}$ at position t	Fixed
L	Expected length of chromosomal regions altered in breast tumours	User-defined

Table 1: Description of random variables and fixed quantities in the APOLLOH framework depicted in Figure 1). $a_{1:T}$, $N_{1:T}$ and $c_{1:T}$ are observed input quantities. All hyperparameters are user-defined and used to help initialize model parameters. The position-specific HMM transition probabilities are fixed quantities. $\pi_{1:18}$ and $\mu_{1:18}$ are unknown variables estimated by expectation maximization (EM).

Table 2: Coverage statistics for 6 Illumina HiSeq and 17 Life/ABI SOLiD tumour-normal paired genomes.

Table 3: Normal contamination estimates predicted by APOLLOH and OncoSNP and transcriptome allelic ratios for LOH predicted regions in 23 breast cancer samples. Quality measures of tumour cellularity or content is given as moderate or high (see Methods for histopathological review). Median transcriptome allelic ratios of positions overlapping all LOH regions, predicted by APOLLOH and OncoSNP, for 22 breast samples with corresponding RNAseq data is shown. Wilcoxon one-tailed significance tests were performed between APOLLOH and OncoSNP RNAseq distributions.

Table 4: Performance evaluations for APOLLOH and model variants, using SNP6 data as ground truth. A) Precision, recall and F-measure performance metrics for loss of heterozygosity (LOH) predictions made by the APOLLOH model variants and the naive model (SNVMix) for each of the 23 breast cancer samples. OncoSNP (Yau et al., 2010) predictions made on Affymetrix SNP6 data was used as the benchmark in the evaluation. B) Comparison of performance between APOLLOH model variants and the naive model (SNVMix). Results are presented for all 23 triple negative breast cancer samples. Wilcoxon one-tailed significance p-values and performance distributions for pairwise comparisons between the models are shown. C) Precision, recall and F-measure performance metrics allele-specific copy number amplification (ASCNA) predictions made by the APOLLOH model variants and the naive model (SNVMix) for each of the 23 breast cancer samples. D) Performance for 5 cases was computed using Exon Capture (EXCAP) sequencing data, published in (Shah et al., 2012), as ground truth data. Calculations were performed as described in Methods, similar to evaluation using SNP6 data as truth.

Table 5: Inferred normal proportion and F-Measure performance for the 30$\times$ and 60$\times$ tumour-normal mixture experiment. Mixture proportions are based on the sampling proportions extracted from the tumour and normal BAM files. The adjusted theoretical proportions, which factors 15% normal cell content (85% cellularity), is the best approximation to the true tumour-normal mixture. This is computed as $adjusted\ Normal\ Proportion = 1 - (tumour\ mixture * 0.85)$. A comparison of the F-measure between the APOLLOH-noS model (not accounting for normal contamination) and the full APOLLOH model is provided.

Table 6: Summary statistics for each of 23 breast cancer samples analyzed with APOLLOH. A) Number APOLLOH segments categorized by zygosity and copy number informed status. B) Proportion of genome altered by segments classified under each APOLLOH state. C) Number of LOH genes predicted per case. D) Proportion of amplified LOH (ALOH), balanced copy number amplification (BCNA), and ASCNA within amplified regions. E) Number and proportion of genes exhibiting monoallelic expression within LOH, ASCNA and HET predicted genomic regions.

Table 7: Full list of APOLLOH predicted segments and overlapping genes (in non-heterozygous segments) across the 23 breast cancer samples. 'Symmetric allelic ratio' column was computed as the median symmetric allelic ratio $[max(refCount, nonRefCount)/depth]$ across the positions in the segment. APOLLOH calls are defined as deletion LOH (DLOH), copy neutral LOH (NLOH), amplified LOH (ALOH), heterozygous or retention (HET), balanced amplification (BCNA), allele-specific amplification (ASCNA).

Table 8: Full list of Ensembl 54 genes with LOH and monoallelic expression (MAE) gene frequencies across the 23 breast cancer samples analyzed using APOLLOH. MAE events are categorized by APOLLOH predictions on the 22 genomes (with available RNAseq data): deletion LOH (DLOH), copy neutral LOH (NLOH), amplified LOH (ALOH), allele-specific copy number amplification (ASCNA), and heterozygous or retention (HET).

Table 9: (A) Germline truncating (stop codon) variant analysis summary. Analysis of codon effect annotations were performed using snpEff (Cingolani, 2012). Column definitions: "Stop_HET", number of codons for which both alleles coding for stop codon and amino acid were observed in the tumour; "Stop-LOH-WT", number of variants homozygous for the stop codon, coded on the reference allele, after LOH; "Stop-LOH-MUT", number of variants homozygous for the stop codon, coded on the non-reference allele, after LOH; "AA-LOH-MUT", number of variants homozygous for an amino acid, coded on the reference allele, after LOH; "AA-LOH-WT", number of variants homozygous for an amino acid, coded on the non-reference allele, after LOH; "Synon_HET", number of synonymous variants for which both alleles are present in tumour; "Synon_LOH-WT", number of synonymous variants that was homozygous for the reference allele after LOH; "Synon_LOH-MUT", number of synonymous variant that was homozygous for the non-reference allele after LOH. χ^2 tests were performed for codon counts from "Stop_LOH-WT + Stop_LOH-MUT", "AA_LOH-WT + AA_LOH-MUT", "Synon_LOH-WT", "Synon_LOH-MUT". Multiple test correction was applied using Benjamini and Hochberg.
(B) List of truncated genes as a result of somatic mutation. Read counts for reference and variant allele were obtained from WGSS or ultra-deep amplicon sequencing data (Shah et al., 2012). Genes that have $APOLLOH\_call \in \{DLOH, NLOH, ALOH\}$ are candidates for complete inactivation. snpEff (Cingolani, 2012) was used to annotate codon effects in the "Effect" column. "Remaining allele" contains one of $\{Mutation,\ Stop\ Codon,\ Complex\}$. "Mutation" denotes that the position is homozygous for the missense mutation found on the non-reference (variant) allele. "Stop Codon" denotes that the position is homozygous for the stop codon mutation found on the non-reference allele. "Complex" denotes that the allelic ratio is skewed towards the reference; this observation may be due to more intricate events such as sub-clonality and/or more complex ordering of mutation and LOH events.
(C) Somatic truncating mutations summary. Column definitions are the same as in (A).

Table 10: Number of germline heterozygous positions predicted in the normal genome using GATK UnifiedGenotyper (McKenna et al., 2010). Positions used for analysis in APOLLOH were pre-processed to exclude loci with depth < 10 and > 200 reads. The number of heterozygous positions and SNP6 probes with RNAseq coverage is presented (also see Figure S12).

Table 11: Number of focal segments across 23 breast cancers. "overallStats" represents statistics for all APOLLOH-predicted LOH segments (with length ¿ 1bp). "exclusiveStats" represents statistics for all APOLLOH-predicted LOH segments that did not overlap any OncoSNP-predicted LOH segments. "betProbeStats" represents statistics for all APOLLOH-predicted segments that were found between or outside of Affymetrix SNP6 probe scaffold. "exclusiveBetProbeStats" represents statistics for APOLLOH-predicted segments that did not overlap OncoSNP-predicted LOH segments and were found between/outside of SNP6 probes.

Table 12: Full gene by patient matrix for monoallelic expression (MAE) within LOH, ASCNA, BCNA, and HET regions. Genes that have all overlapping positions predicted as MAE in RNAseq and APOLLOH calls of DLOH, NLOH, ALOH, HET, BCNA or ASCNA are represented by 1, 2, 3, 4, 5 and 6, respectively. A zero represents a gene that contains at least 1 position that is HET or balanced allelic expression (BAE). NA is given to a gene that has no overlapping positions in RNAseq data or the APOLLOH analysis.

Table 13: Enriched pathways within modules determined using Reactome Functional Interaction software (Wu et al., 2010). This table was exported directly from the Reactome plugin in Cytoscape (Smoot et al., 2011). Six modules contain enriched pathways at false discovery rate (FDR) of 0.05. The 'GeneSet' column lists the pathways; 'Nodes' in this analysis are the genes that belong to a specific module and found within the pathways.

References

Barnett, D. W., Garrison, E. K., Quinlan, A. R., Strömberg, M. P., and Marth, G. T., 2011. Bamtools: a c++ api and toolkit for analyzing and managing bam files. *Bioinformatics*, **27**(12):1691–1692.

Bignell, G. R., Greenman, C. D., Davies, H., Butler, A. P., Edkins, S., Andrews, J. M., Buck, G., Chen, L., Beare, D., Latimer, C., *et al.*, 2010. Signatures of mutation and selection in the cancer genome. *Nature*, **463**(7283):893–898.

Bishop, C. M., 2007. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1st ed. 2006. corr. 2nd printing edition.

Chiang, D. Y., Getz, G., Jaffe, D. B., O'Kelly, M. J. T., Zhao, X., Carter, S. L., Russ, C., Nusbaum, C., Meyerson, M., and Lander, E. S., *et al.*, 2009. High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat Methods*, **6**(1):99–103.

Cingolani, P., 2012. snpeff: Variant effect prediction. *http://snpeff.sourceforge.net*, .

Colella, S., Yau, C., Taylor, J. M., Mirza, G., Butler, H., Clouston, P., Bassett, A. S., Seller, A., Holmes, C. C., and Ragoussis, J., *et al.*, 2007. Quantisnp: an objective bayes hidden-markov model to detect and accurately map copy number variation using snp genotyping data. *Nucleic Acids Res*, **35**(6):2013–2025.

Flicek, P., Aken, B. L., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., *et al.*, 2010. Ensembl's 10th year. *Nucl. Acids Res.*, **38**(suppl_1):D557–562.

Goya, R., Sun, M. G. F., Morin, R. D., Leung, G., Ha, G., Wiegand, K. C., Senz, J., Crisan, A., Marra, M. A., Hirst, M., *et al.*, 2010. Snvmix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics*, **26**(6):730–736.

Li, H. and Durbin, R., 2009. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, **25**(14):1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, . G. P. D. P., *et al.*, 2009. The sequence alignment/map format and samtools. *Bioinformatics*, **25**(16):2078–2079.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., *et al.*, 2010. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome Res*, **20**(9):1297–1303.

McKernan, K. J., Peckham, H. E., Costa, G. L., McLaughlin, S. F., Fu, Y., Tsung, E. F., Clouser, C. R., Duncan, C., Ichikawa, J. K., Lee, C. C., *et al.*, 2009. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Research*, **19**(9):1527–1541.

Morin, R. D., Mendez-Lago, M., Mungall, A. J., Goya, R., Mungall, K. L., Corbett, R. D., Johnson, N. A., Severson, T. M., Chiu, R., Field, M., *et al.*, 2011. Frequent mutation of histone-modifying genes in non-hodgkin lymphoma. *Nature*, **476**(7360):298–303.

Pruitt, K. D., Tatusova, T., and Maglott, D. R., 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucl. Acids Res.*, **35**(suppl_1):D61–65.

Shah, S. P., Morin, R. D., Khattra, J., Prentice, L., Pugh, T., Burleigh, A., Delaney, A., Gelmon, K., Guliany, R., Senz, J., *et al.*, 2009. Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature*, **461**(7265):809–13.

Shah, S. P., Roth, A., Goya, R., Oloumi, A., Ha, G., Zhao, Y., Turashvili, G., Ding, J., Tse, K., Haffari, G., *et al.*, 2012. Primary triple negative breast cancers exhibit a continuous spectrum of clonal and mutational evolution. *Nature*, **In press.**

Shah, S. P., Xuan, X., DeLeeuw, R. J., Khojasteh, M., Lam, W. L., Ng, R., and Murphy, K. P., 2006. Integrating copy number polymorphisms into array cgh analysis using a robust hmm. *Bioinformatics*, **22**(14):e431–9.

Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P. L., and Ideker, T., 2011. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, **27**(3):431–432.

Thierry-Mieg, D. and Thierry-Mieg, J., 2006. AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome Biology*, **7**(Suppl 1):S12.

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S. F., Hakonarson, H., and Bucan, M., 2007. Pennncnv: an integrated hidden markov model designed for high-resolution copy number variation detection in whole-genome snp genotyping data. *Genome Res*, **17**(11):1665–1674.

Wiegand, K. C., Shah, S. P., Al-Agha, O. M., Zhao, Y., Tse, K., Zeng, T., Senz, J., McConechy, M. K., Anglesio, M. S., Kalloger, S. E., *et al.*, 2010. Arid1a mutations in endometriosis-associated ovarian carcinomas. *N Engl J Med*, **363**(16):1532–1543.

Wu, G., Feng, X., and Stein, L., 2010. A human functional protein interaction network and its application to cancer data analysis. *Genome Biol*, **11**(5).

Yau, C., Mouradov, D., Jorissen, R. N., Colella, S., Mirza, G., Steers, G., Harris, A., Ragoussis, J., Sieber, O., and Holmes, C. C., *et al.*, 2010. A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data. *Genome Biol*, **11**(9).