Supplementary Materials for "Integrated study of copy number states and genotype calls using high density SNP arrays"

A HapMap samples

Originally, Illumina performed 73 CEU samples, 77 YRI samples, and 75 CHB+JPT samples on Human 610-Quad arrays. Since one of our criteria for results evaluation is the overlap of the CNVs between offspring and parents, we only use the CEU and YRI samples that are form complete parents-child trios, which correspond to 14 CEU trios and 20 YRI trios. Two CEU trios and three YRI trios are excluded for further analysis due to high noise of the data or chromosome-wide copy number aberrations, which are not expected in normal tissue. CHB+JPT samples are independent individuals and we do not observe serious low-quality arrays, so all the 75 CHB+JPT samples are used in our study.

Figure A-1 illustrate the BAF and LRR data of chromosome 2 of sample NA12006, which is one member of a deleted CEU trio. The LRR data appears to be quite noisy and in need of further normalization. Similar patterns are observed for other chromosomes of this sample. Another sample NA12264 suffers similar problem (Figure A-2). Figure A-3 and A-4 show two YRI individuals (NA19208 and NA19193) with chromosome wide amplification. Another YRI sample NA18870 has large variance for LRR, and many scattered SNPs with low or extremely low LRR (Figure A-5). These five samples and their corresponding families are excluded from our studies.

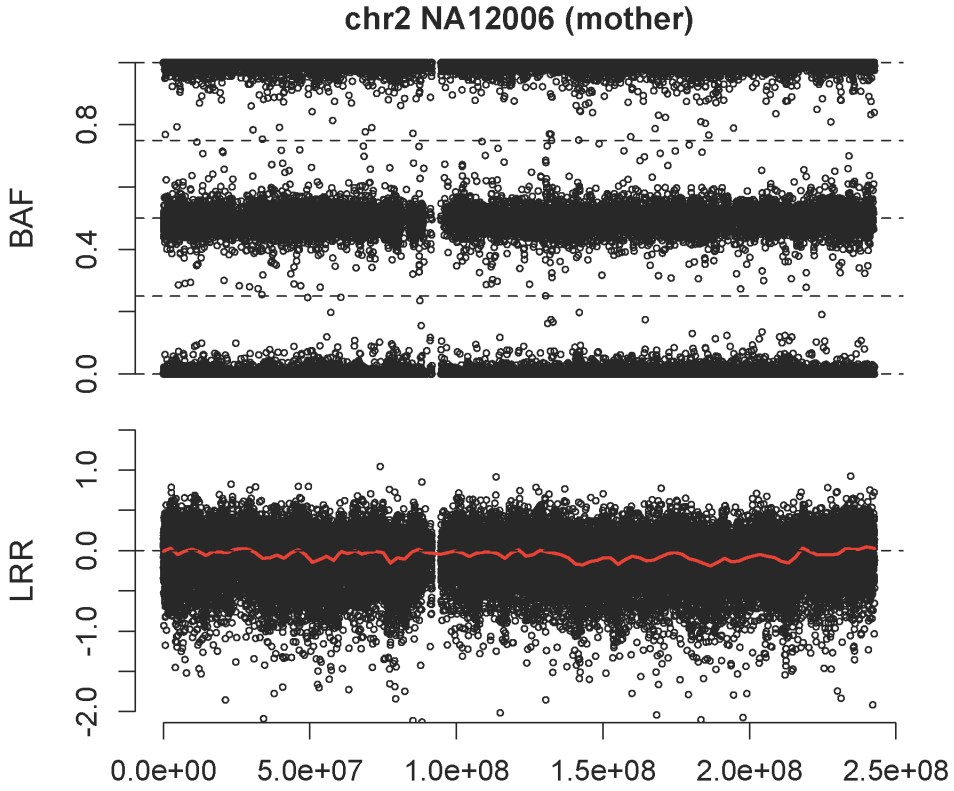

Figure A-1: BAF and LRR for chromosome 2 of HapMap sample NA12006 (mother of a CEU trio: NA12005, NA12006, and NA10839). Compared with other individuals, the LRR has large variance and the lowess fit of LRR (the solid red curve) fluctuates, which indicate either a bad array or insufficient normalization.

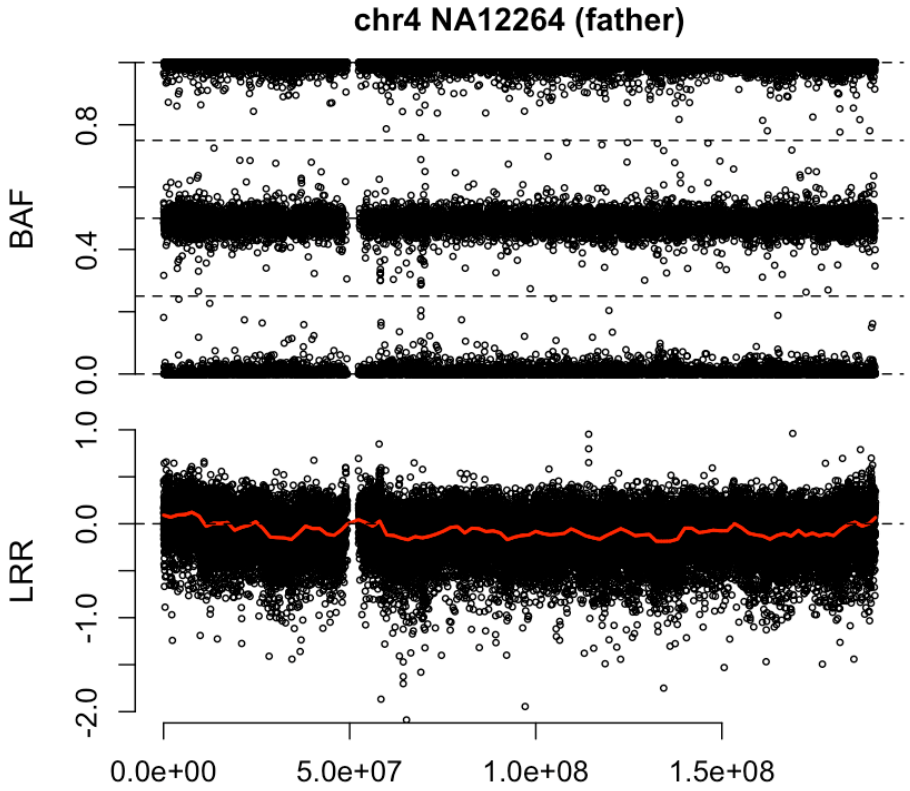

Figure A-2: BAF and LRR for chromosome 4 of HapMap sample NA12264 (father of a CEU trio: NA12264, NA12234, and NA10863). Similar to the sample NA12006, the LRR has large variance and the lowess fit of LRR (the solid red curve) fluctuates, which indicate either a bad array or insufficient normalization.

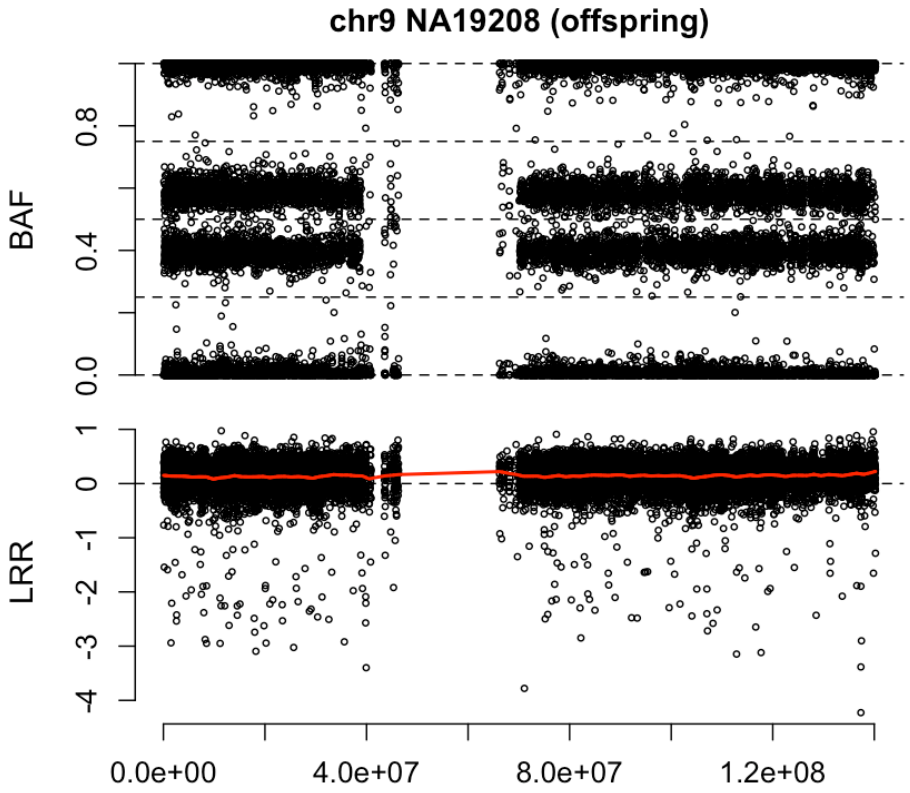

Figure A-3: BAF and LRR for chromosome 9 of HapMap sample NA19208 (offspring of a YRI trio: NA19207, NA19206, and NA19208). This is obviously a pattern of chromosome-wide amplification.

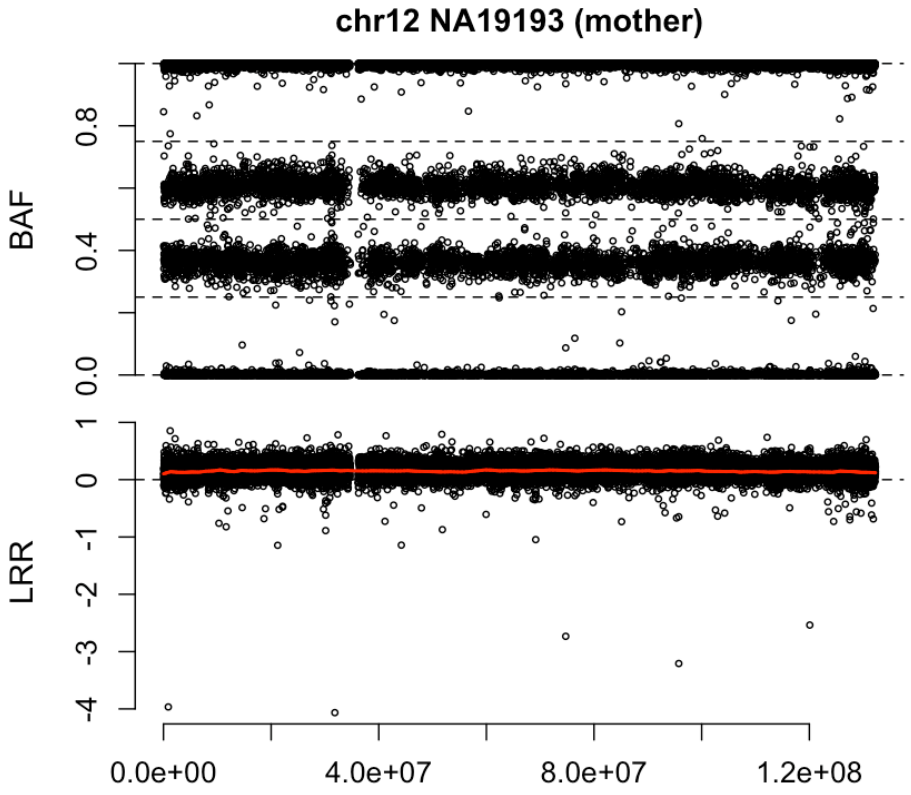

Figure A-4: BAF and LRR for HapMap sample NA19193 (mother of a YRI trio: NA19192, NA19193, and NA19194). This is obviously a pattern of chromosome-wide amplification.

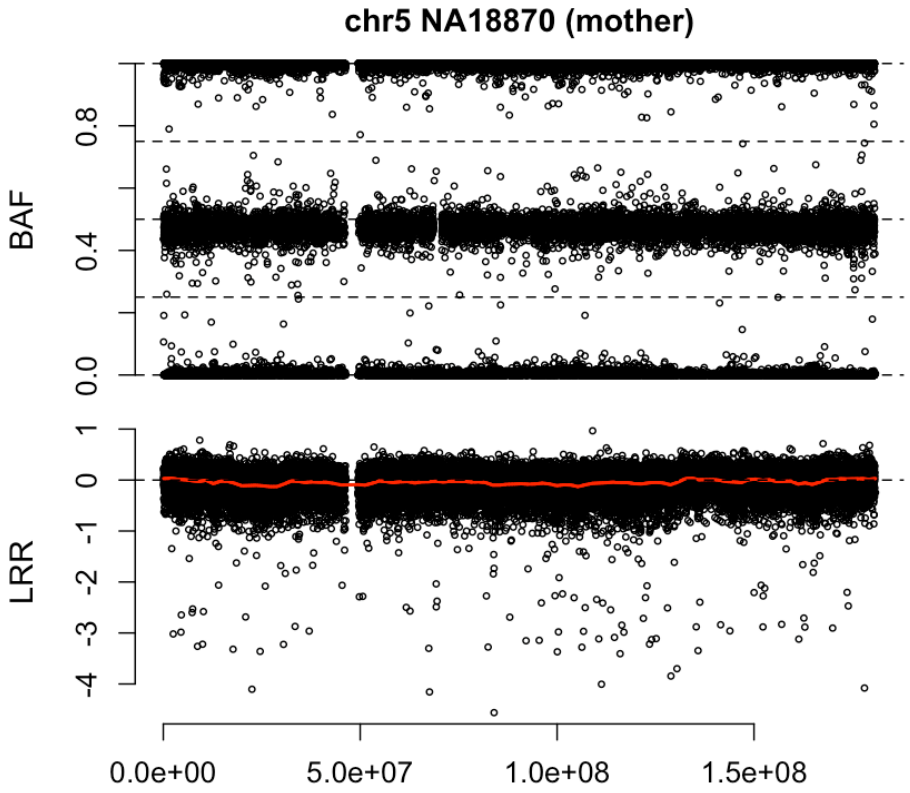

Figure A-5: BAF and LRR for HapMap sample NA18870 (mother of a YRI trio: NA18871, NA18870, and NA18872). The LRR has large variance, and many scattered SNPs have low or extremely low LRR.

Table A-1: Comparison of PennCNV (P) and xCNV (X) by the number/proportion of CNVs that are shared between the offspring and the parents in 12 CEU trios.

Family ID	Father	Mother	Offspring	Total		Matched		Proportion	
				P	X	P	X	P	X
1	NA12003	NA12004	NA10838	31	26	11	8	0.35	0.31
2	NA11992	NA11993	NA10860	34	35	14	12	0.41	0.34
3	NA07357	NA07345	NA07348	38	54	12	12	0.32	0.22
4	NA06994	NA07000	NA07029	34	38	13	15	0.38	0.39
5	NA11839	NA11840	NA10854	35	33	18	17	0.51	0.52
6	NA12264	NA12234	NA10863	46	46	19	17	0.41	0.37
7	NA12716	NA12717	NA12707	39	36	22	20	0.56	0.56
8	NA12891	NA12892	NA12878	53	47	18	16	0.34	0.34
9	NA12812	NA12813	NA12801	26	38	7	12	0.27	0.32
10	NA12874	NA12875	NA12865	48	48	21	21	0.44	0.44
11	NA12762	NA12763	NA12753	52	38	31	24	0.60	0.63
12	NA06993	NA06985	NA06991	38	41	20	20	0.53	0.49

Table A-2: Comparison of PennCNV (P) and xCNV (X) by the number/proportion of CNVs that are shared between the offspring and the parents in 17 YRI trios.

Family ID	Father	Mother	Offspring	Total		Matched		Proportion	
				P	X	P	X	P	X
1	NA18501	NA18502	NA18500	47	65	20	19	0.43	0.29
2	NA18504	NA18505	NA18503	36	43	23	20	0.64	0.47
3	NA18507	NA18508	NA18506	26	37	15	17	0.58	0.46
4	NA18516	NA18517	NA18515	46	34	22	14	0.48	0.41
5	NA18871	NA18870	NA18872	54	67	17	24	0.31	0.36
6	NA18853	NA18852	NA18854	59	91	16	15	0.27	0.16
7	NA18856	NA18855	NA18857	46	46	25	25	0.54	0.54
8	NA18913	NA18912	NA18914	29	28	20	16	0.69	0.57
9	NA19092	NA19093	NA19094	32	34	14	15	0.44	0.44
10	NA19138	NA19137	NA19139	55	36	26	23	0.47	0.64
11	NA19200	NA19201	NA19202	60	69	20	21	0.33	0.30
12	NA19171	NA19172	NA19173	51	44	21	21	0.41	0.48
13	NA19203	NA19204	NA19205	45	37	29	24	0.64	0.65
14	NA19160	NA19159	NA19161	35	32	16	16	0.46	0.50
15	NA19223	NA19222	NA19221	45	38	16	16	0.36	0.42
16	NA19119	NA19116	NA19120	32	40	12	18	0.38	0.45
17	NA19141	NA19140	NA19142	59	60	27	22	0.46	0.37

B HMM Algorithm

B.1 Emission probability for BAF

States 3 is a special case because it has only one normal component, so the weight is always 1. For xCNV, the weights (for any state other than state 3) are decided as follows. Let p_B be the population frequency of B allele in normal tissue, $w_{z,h} = \psi(n_B; n, p_B)$, where $\psi(n_B; n, p_B)$ is the binomial probability of choosing n_B of B alleles from a total of n alleles. For xCNA when genotype in normal tissue is not available, state 1 and 3 have the same weights as in xCNV. The weights for state 7 are $w_{7,1} = \psi(0; 2, p_B)$, $w_{7,2} = \psi(1; 2, p_B)$, and $w_{7,3} = \psi(2; 2, p_B)$. Assuming there is tissue contamination, the weights for state 2, 4, 5, 6, 8, and 9 are the same: $w_{z,1} = \psi(0; 2, p_B)$, $w_{z,2} = w_{z,3} = 0.5\psi(1; 2, p_B)$, and $w_{z,4} = \psi(2; 2, p_B)$. If there is no tissue contamination, weights for state 5 and 9 remain the same; for state 2, 4, 6, and 8, $w_{z,1} = p_A$, $w_{z,2} = w_{z,3} = 0$, and $w_{z,4} = p_B$.

For xCNA when genotype in normal tissue is available, the emission probability of BAF is as follows. If the genotype in normal tissue is homozygous,

$$\begin{aligned} p(b|z, g = AA) &= \pi_{b,z} I(0 < b < 1) + (1 - \pi_{b,z}) \left(q_e \varphi(b; \theta_{z,1}) + \sum_{h \neq 1} \frac{p_e}{1 - H_z} \varphi(b; \theta_{z,h}) \right), \\ p(b|z, g = BB) &= \pi_{b,z} I(0 < b < 1) + (1 - \pi_{b,z}) \left(q_e \varphi(b; \theta_{z,H_z}) + \sum_{h \neq H_z} \frac{p_e}{1 - H_z} \varphi(b; \theta_{z,h}) \right), \end{aligned}$$

where $\varphi(b; \theta_{z,h}) = \phi(b; \theta_{z,h})^{I(0<b<1)} \Phi(0; \theta_{z,h})^{I(b=0)} (1 - \Phi(1; \theta_{z,h}))^{I(b=1)}$, and p_e is the genotyping error (assume it is a known constant), and $q_e = 1 - p_e$. If the genotype in normal tissue is heterozygous and under the assumption of tissue contamination, for state $z = 1, 7$,

$$p(b|z, g = AB) = \pi_{b,1} I(0 < b < 1) + (1 - \pi_{b,1}) \left(q_e \varphi(b; \theta_{1,2}) + 0.5 p_e \sum_{h=1,3} \varphi(b; \theta_{1,h}) \right)$$

For state $z = 2, 4, 5, 6, 8, 9$,

$$\begin{aligned} p(b|z, g = AB) &= \pi_{b,z} I(0 < b < 1) + \\ &\quad (1 - \pi_{b,z}) \left(0.5 q_e \sum_{h=2,3} \varphi(b; \theta_{z,h}) + 0.5 p_e \sum_{h=1,4} \varphi(b; \theta_{z,h}) \right) \end{aligned}$$

If it is assumed that there is no tissue contamination, all the above equations hold except that for states $z = 2, 4, 6, 8$

$$p(b|z, g = AB) = \pi_{b,z} I(0 < b < 1) + (1 - \pi_{b,z}) \left(0.5 \varphi(b; \theta_{z,1}) + 0.5 \varphi(b; \theta_{z,H_z}) \right).$$

B.2 Notations

Let L be the number of SNP probes, and let N be the number of distinct states in the HMM. The input data, denoted as $\mathbf{X}$ ={pos, LRR, BAF, λ, G}, includes the probe positions (pos), the logR ratio (LRR, denoted by $r_i,\ 1 \leq i \leq L$), the B allele frequency (BAF, denoted by $b_i,\ 1 \leq i \leq L$), parameters of state duration $\lambda = \{\lambda_j,\ 1 \leq j \leq N\}$, and an optional input $G = \{g_i,\ 1 \leq i \leq L\}$, the genotypes of all the SNPs in normal tissue. The parameters of the HMM are $\Theta = \{\pi_{r,z}, \mu_r, \sigma_r, \pi_{b,z}, \mu_b, \sigma_b, A\}$. Specifically, $\{\pi_{r,z}, \mu_r, \sigma_r\}$ are the parameters for the emission probability of LRR, where $\pi_{r,z}$ are the mixture proportions of the uniform components, μ_r and σ_r indicate all the mean, standard deviation parameters of LRR. Similarly, $\{\pi_{b,z}, \mu_b, \sigma_b\}$ are the parameters for the emission probability of BAF. $A = \{a_{jk}\ (1 \leq j \neq k \leq N)\}$ is the transition probability matrix. We need to define some additional notations:

- q_i: the state at position i,
- κ_z: the probability that state of the first SNP probe is state z,
- $e(i, z)$: the emission probability of state z at position i,
- d_i: the distance from probe $i-1$ to i,
- $\alpha_i(j, k)$: the transition probability from state j to k, from probe $i-1$ to i. If $k \neq j$, $\alpha_i(j, k) = a_{jk}(1 - \exp(-\lambda_j d_i))$. If $k = j$, $\alpha_i(j, j) = \exp(-\lambda_j d_i) = 1 - \sum_{k \neq j} \alpha_i(j, k)$.

Most of the computations are carried out in log scale to avoid underflow or overflow. A utility function `logsumexp` is used to facilitate the computation. Specifically, it is defined as $\texttt{logsumexp}_j(v) = \log\left(\sum_j \exp(v_j)\right)$, where $v = \{v_j\}$ is a vector.

B.3 Viterbi Algorithm

Given all the parameters, find the best path.

Input

$\mathbf{X}, \Theta$

Output

`path`: the most likely path; `logPv`: the log probability of the most likely path;

Intermediate Variables

$v(i, z)$: p(the most likely path ending at position $i \mid q_i = z$); $\texttt{path.m}(i-1, z)$: the most likely state at position $i-1$, given $q_i = z$.

Algorithm

1. Initialization:

$$v(1, z) = \kappa_z e(1, z) \tag{B-1}$$

$$\log(v(1, z)) = \log(\kappa_z) + \log(e(1, z)) \tag{B-2}$$

2. Recursion, for $i \in (2:L)$[1] and for $z \in (1:N)$,

$$v(i,z) = \max_j (v(i-1,j)\alpha_i(j,z))e(i,z) \quad \text{(B-3)}$$

$$\log(v(i,z)) = \max_j [\log(v(i-1,j)) + \log(\alpha_i(j,z))] + \log(e(i,z)) \quad \text{(B-4)}$$

$$\texttt{path.m}(i-1,z) = \text{argmax}_j [\log(v(i-1,j)) + \log(\alpha_i(j,z))] \quad \text{(B-5)}$$

where $\max_j(v)$ returns the maximum value and $\text{argmax}_j(v)$ returns the index of the maximum value.

3. Termination

$$\texttt{path}(L) = \text{argmax}_z \log(v(L,z)) \quad \text{(B-6)}$$

$$\texttt{logPv} = \max_z \log(v(L,z)) \quad \text{(B-7)}$$

For $i \in ((L-1):1)$

$$\texttt{path}(i) = \texttt{path.m}(i, \texttt{path}(i+1)) \quad \text{(B-8)}$$

B.4 Forward Algorithm

Given all the parameters, find the forward probabilities.

Input

$\mathbf{X}, \Theta$

Output

$f(i,z)$: forward probability, $p(x_1, x_2, ..., x_i, q_i = z|\Theta)$; overall likelihood $\log(p(\mathbf{X}|\Theta))$.

Algorithm

1. Initialization:

$$f(1,z) = \kappa_z e(1,z) \quad \text{(B-9)}$$

$$\log(f(1,z)) = \log(\kappa_z) + \log(e(1,z)) \quad \text{(B-10)}$$

2. Recursion, for $i \in (2:L)$ and for $z \in (1:N)$,

$$f(i,z) = e(i,z) \sum_j f(i-1,j)\alpha_i(j,z) \quad \text{(B-11)}$$

$$\log(f(i,z)) = \log(e(i,z)) + \texttt{logsumexp}_j [\log(f(i-1,j)) + \log(\alpha_i(j,z))] \quad \text{(B-12)}$$

3. Termination

$$p(\mathbf{X}|\Theta) = \sum_z f(L,z) \quad \text{(B-13)}$$

$$\log(p(\mathbf{X}|\Theta)) = \texttt{logsumexp}_z \log(f(L,z)) \quad \text{(B-14)}$$

[1] we use $m:n$ to indicate a series of $m, m+1, ..., n$, if $m < n$, or a series of $m, m-1, ..., n$ if $m > n$, where m and n are both integers

B.5 Backward Algorithm

Input

$\mathbf{X}, \Theta$

Output

$b(i,z)$: backward probability, $p(x_{i+1}, ..., x_L, |q_i = z, \Theta)$.

Algorithm

1. Initialization:

$$b(L,z) = 1 \tag{B-15}$$
$$\log(b(L,z)) = 0 \tag{B-16}$$

2. Recursion, for $i \in (L:2)$ and for $z \in (1:N)$,

$$b(i-1,z) = \sum_j [\alpha_i(z,j)e(i,j)b(i,j)] \tag{B-17}$$
$$\log(b(i-1,z)) = \texttt{logsumexp}\,[\log(\alpha_i(z,j)) + \log(e(i,j)) + \log(b(i,j))] \tag{B-18}$$

B.6 Posterior Probability

Calculate the posterior probability based on forward and backward algorithm.

Input

Forward probability $f_{L\times N} = \{f(i,z)\}$, backward probability $b_{L\times N} = \{b(i,z)\}$, and overall likelihood $\texttt{logP} = \log(p(\mathbf{X}|\Theta))$.

Output

$\gamma(i,z)$: posterior probability $p(q_i = z|X, \Theta)$.

Algorithm

$$\gamma(i,z) = \frac{f(i,z)b(i,z)}{p(\mathbf{X}|\Theta)} \tag{B-19}$$
$$\log(\gamma(i,z)) = \log(f(i,z)) + \log(b(i,z)) - \texttt{logP} \tag{B-20}$$

B.7 Baum-Welch Algorithm

Here we only describe one-step update of the Baum-Welch Algorithm. The whole algorithm is simply repeats of this one-step update, plus the initial values of the parameters and a convergence criterion.

Input

$\mathbf{X}$ and Θ_0. Here Θ_0 is either the initial values of the parameters or the parameter estimates from previous iteration.

Output

Estimated parameters Θ_1; `logP`: the log likelihood from each iteration.

Algorithm

(1) *Estimate the posterior probabilities that one SNP belongs to one HMM state.*

$$\begin{aligned} f &= \texttt{forward}(\mathbf{X}, \Theta_0), && \text{(B-21)} \\ b &= \texttt{backward}(\mathbf{X}, \Theta_0), && \text{(B-22)} \\ \texttt{logP} &= \texttt{logsumexp}_z \log(f(L, z)). && \text{(B-23)} \end{aligned}$$

The initial probability κ_z is simply the posterior probability of being state z at position 1, therefore the new estimate of κ_z, denoted as $\overline{\kappa}_z$, is

$$\begin{aligned} \overline{\kappa}_z &= \frac{f(1,z)b(1,z)}{p(\mathbf{X}|\Theta)}, && \text{(B-24)} \\ \log(\overline{\kappa}_z) &= \log(f(1,z)) + \log(b(1,z)) - \texttt{logP}. && \text{(B-25)} \end{aligned}$$

(2) *Estimate the transition probability a_{jk}.* Denote the estimated a_{jk} as $\overline{a}_{jk}$, for $j \neq k$

$$\begin{aligned} \overline{a}_{jk} &= \frac{\sum_{i=2}^{L} f(i-1,j) a_{jk} (1-\exp(-\lambda_j d_i)) e(i,k) b(i,k)}{\sum_{i=2}^{L} \sum_{l \neq j} f(i-1,j) a_{jl} (1-\exp(-\lambda_j d_i)) e(i,l) b(i,l)} && \text{(B-26)} \\ &= \frac{a_{jk} \sum_{i=2}^{L} f(i-1,j) (1-\exp(-\lambda_j d_i)) e(i,k) b(i,k)}{\sum_{l \neq j} a_{jl} \sum_{i=2}^{L} f(i-1,j) (1-\exp(-\lambda_j d_i)) e(i,l) b(i,l)}. && \text{(B-27)} \end{aligned}$$

Let $c_{jk} = \texttt{logsumexp}_i[\log(f(i-1,j)) + \log(1-\exp(-\lambda_j d_i)) + \log(e(i,k)) + \log(b(i,k))]$, then

$$\log(\overline{a}_{jk}) = \log(a_{jk}) + c_{jk} - \texttt{logsumexp}_{l \neq j}[\log(a_{jl}) + c_{jl}]. \quad \text{(B-28)}$$

(3) *Estimate $\{\pi_{r,z}, \mu_{r,z}, \sigma_{r,z}\}$, the parameters for the emission probability of LRR.* Because the likelihoods for LRR and BAF are independent with each other given the states, we estimate $\{\pi_{r,z}, \mu_{r,z}, \sigma_{r,z}\}$ by maximizing the mixture density $p(r|z) = \pi_{r,z}/R_m + (1 - \pi_{r,z})\phi(r; \mu_{r,z}, \sigma_{r,z})$. For each observation r_i, we calculate the probability it belongs to the

uniform component ($U_{r,z}$) and the normal component ($N_{r,z}$) respectively:

$$\begin{aligned}
\gamma(i,z,U_{r,z}) &= p(q_i=z,\xi_i=U_{r,z}|\mathbf{X},\Theta_0) \\
&= p(\xi_i=U_{r,z}|q_i=z,\mathbf{X},\Theta_0)p(q_i=z|\mathbf{X},\Theta_0) \\
&= \frac{(\pi_{r,z}/R_m)\gamma(i,z)}{\pi_{r,z}/R_m+(1-\pi_{r,z})\phi(r_i;\mu_{r,z},\sigma_{r,z})}, \qquad \text{(B-29)} \\
\gamma(i,z,N_{r,z}) &= p(q_i=z,\xi_i=N_{r,z}|\mathbf{X},\Theta_0) \\
&= p(\xi_i=N_{r,z}|q_i=z,\mathbf{X},\Theta_0)p(q_i=z|\mathbf{X},\Theta_0) \\
&= \frac{(1-\pi_{r,z})\phi(r_i;\mu_{r,z},\sigma_{r,z})\gamma(i,z)}{\pi_{r,z}/R_m+(1-\pi_{r,z})\phi(r_i;\mu_{r,z},\sigma_{r,z})}, \qquad \text{(B-30)}
\end{aligned}$$

where ξ_i indicates the mixture component (of LRR) of probe i. Therefore the new estimates for $\pi_{r,z}$, $\mu_{r,z}$, and $\sigma_{r,z}$ are respectively

$$\overline{\pi}_{r,z} = \frac{\sum_{i=1}^{L}\gamma(i,z,U_{r,z})}{\sum_{i=1}^{L}\gamma(i,z)}, \qquad \text{(B-31)}$$

$$\overline{\mu}_{r,z} = \frac{\sum_{i=1}^{L}\gamma(i,z,N_{r,z})r_i}{\sum_{i=1}^{L}\gamma(i,z,N_{r,z})}, \qquad \text{(B-32)}$$

$$\overline{\sigma}_{r,z} = \frac{\sum_{i=1}^{L}\gamma(i,z,N_{r,z})(r_i-\mu_{r,z})^2}{\sum_{i=1}^{L}\gamma(i,z,N_{r,z})}. \qquad \text{(B-33)}$$

For those states that share the same number of copies, the parameters $\mu_{r,z}$ and $\sigma_{r,z}$ can be estimated by combining those states. For example, state 1 and 2 both have copy number 2, thus

$$\overline{\mu}_{r,1}=\overline{\mu}_{r,2} = \frac{\sum_{i=1}^{L}\sum_{z=1}^{2}\gamma(i,z,N_{r,z})r_i}{\sum_{i=1}^{L}\sum_{z=1}^{2}\gamma(i,z,N_{r,z})}, \qquad \text{(B-34)}$$

$$\overline{\sigma}^2_{r,1}=\overline{\sigma}^2_{r,2} = \frac{\sum_{i=1}^{L}\sum_{z=1}^{2}\gamma(i,z,N_{r,z})(r_i-\overline{\mu}_{r,1})^2}{\sum_{i=1}^{L}\sum_{z=1}^{2}\gamma(i,z,N_{r,z})}. \qquad \text{(B-35)}$$

(4) *Estimate* $\{\pi_{b,z},\mu_{b,z,h},\sigma_{b,z,h}\}$*, the parameters for the emission probability of BAF.* The distribution of BAF is a mixture of a uniform distribution and H_z normal distributions, where H_z varies from 1 to 5, depending on the state z. The estimation of $\mu_{b,z,h}$ and $\sigma_{b,z,h}$ is the most complicate part of our algorithm, mainly because of the truncation of BAF values at 0 and 1. Parameter estimation for truncated normal distribution is more computationally demanding and less stable than non-truncated normal distribution. Therefore we try to avoid the truncated normal distribution as long as it is possible.

First, state 3 is a special case because it is the null state with both alleles deleted. The likelihood of state 3 is composed of a uniform component and a normal component. As mentioned in the main text, we assume the mean value of its normal component is 0.5. We further assume the standard deviation is smaller than 0.15 so that the probability of truncation is smaller than 0.001, therefore as an approximation, we can estimate the standard deviation as if it is non-truncated normal distribution. The possible bias brought by the relatively small standard deviation can be compensated by employing a bigger weight for the uniform component. Furthermore, this null state have significantly lower LRR, so that limited degree of bias in its BAF emission probability will not bring big changes of the final posterior probability estimates.

Next, we consider all the other states except state 3. For the normal distributions corresponding to homozygous genotypes, since $\mu_{b,z,1} = 0.0$ and $\mu_{b,z,H_z} = 1.0$ are fixed, we only need to estimate the standard derivations, which is straightforward because the truncation point is exactly the the mean values. Specifically, we can simply use the observed b_is such that $0 < b_i < 1$ to estimate the standard deviations. For all the other normal components (which correspond to heterozygous genotypes), except for the tissue contamination mixtures such as (A, AB), it is safe to assume the mean values are far away from the boundary (0 or 1) so that the truncation effect can be neglected. For example, based on the parameters used in PennCNV [1], the minimum distance between a mean value of any heterozygous genotype class and the boundary is bigger than five standard deviations. In our method, we allow the data to estimate the parameters. But we add the restriction that for all the normal components corresponding to heterozygous genotypes, except for the tissue contamination mixtures, the minimum distance between the mean value and the boundaries (0 or 1) is 3.3 standard deviations, so that the probability of truncation is smaller than 0.001. Therefore, as an approximation, we can also estimate these normal components using only those b_is such that $0 < b_i < 1$.

Finally, for the two tissue contamination mixtures, the truncated normal distributions are inevitable. We will use all the observed b_is, including those that are exactly 0 or 1 to estimate the corresponding parameters.

Now we discuss more specifically the estimation algorithm. For each observation of BAF, denoted as b_i, we first calculate the probability it belongs to the uniform distribution of state z, (denoted as $U_{b,z}$) and one of the normal distribution of state z (denoted as $N_{b,z,h}$, where $h = 1, ..., H_z$). For all the genotype classes except tissue contamination mixtures, if $b_i = 0$,

$$\gamma(i, z, U_{b,z}) = 0,\ \gamma(i, z, N_{b,z,1}) = \gamma(i, z),\ \text{and}\ \gamma(i, z, N_{b,z,h}) = 0,\ \forall\ h > 1.$$

If $b_i = 1$,

$$\gamma(i, z, U_{b,z}) = 0,\ \gamma(i, z, N_{b,z,H_z}) = \gamma(i, z),\ \text{and}\ \gamma(i, z, N_{b,z,h}) = 0,\ \forall\ h < H_z.$$

For genotype classes except tissue contamination, the truncated values (0 or 1) may arise from either homozygous genotype classes or the tissue contamination mixtures. Specifically, if $b_i = 0$, for $z = 2, 4, 6, 8$,

$$\gamma(i, z, N_{b,z,1}) = \frac{0.5}{0.5 + \Phi(0, \mu_{b,z,2}, \sigma_{b,z,2})}, \tag{B-36}$$

$$\gamma(i, z, N_{b,z,2}) = \frac{\Phi(0, \mu_{b,z,2}, \sigma_{b,z,2})}{0.5 + \Phi(0, \mu_{b,z,2}, \sigma_{b,z,2})}. \tag{B-37}$$

where $\Phi(x, \mu, \sigma)$ is cumulative normal distribution with parameter μ and σ. If $b_i = 1$, for $z = 2, 4, 6, 8$,

$$\gamma(i, z, N_{b,z,3}) = \frac{1 - \Phi(1, \mu_{b,z,3}, \sigma_{b,z,3})}{1.5 - \Phi(1, \mu_{b,z,3}, \sigma_{b,z,3})}, \tag{B-38}$$

$$\gamma(i, z, N_{b,z,z}) = \frac{0.5}{1.5 - \Phi(1, \mu_{b,z,3}, \sigma_{b,z,3})}. \tag{B-39}$$

If $0 < b_i < 1$, for all the genotype classes

$$\begin{aligned} \gamma(i, z, U_{b,z}) &= p(q_i = z, \eta_i = U_{b,z} | \mathbf{X}, \Theta_0) \\ &= p(\eta_i = U_{b,z} | q_i = z, \mathbf{X}, \Theta_0) p(q_i = z | \mathbf{X}, \Theta_0) \\ &= \frac{\pi_{b,z} \gamma(i, z)}{\pi_{b,z} + (1 - \pi_{b,z}) \sum_{h=1}^{H_z} w_{z,h} \phi(b_i; \mu_{b,z.h}, \sigma_{b,z.h})}, \end{aligned} \tag{B-40}$$

$$\begin{aligned} \gamma(i, z, N_{b,z,h}) &= p(q_i = z, \eta_i = N_{b,z,h} | \mathbf{X}, \Theta_0) \\ &= p(\eta_i = N_{b,z,h} | q_i = z, \mathbf{X}, \Theta_0) p(q_i = z | \mathbf{X}, \Theta_0) \\ &= \frac{(1 - \pi_{b,z}) w_{z,h} \phi(b_i; \mu_{b,z.h}, \sigma_{b,z.h}) \gamma(i, z)}{\pi_{b,z} + (1 - \pi_{b,z}) \sum_{h=1}^{H_z} w_{z,h} \phi(b_i; \mu_{b,z.h}, \sigma_{b,z.h})}, \end{aligned} \tag{B-41}$$

where η_i indicates the mixture component (of BAF) of probe i, and the weights $w_{z,h}$ have been described in main text.

With the above posterior probability estimates, let $\Omega = \{i; 0 < b_i < 1\}$, the new estimate for $\pi_{b,z}$ is

$$\overline{\pi}_{b,z} = \frac{\sum_{i \in \Omega} \gamma(i, z, U_{b,z})}{\sum_{i \in \Omega} \sum_{h=1}^{H_z} (\gamma(i, z, N_{b,z,h}) + \gamma(i, z, U_{b,z}))} = \frac{\sum_{i \in \Omega} \gamma(i, z, U_{b,z})}{\sum_{i \in \Omega} \gamma(i, z)}. \tag{B-42}$$

Next, we seek to update the estimation of μ_b and σ_b. For all the states where the mean parameters are not fixed, except the tissue contamination mixtures, the mean parameter can be estimated by

$$\overline{\mu}_{b,z,h} = \frac{\sum_{i\in\Omega}\gamma(i,z,N_{b,z,h})b_i}{\sum_{i\in\Omega}\gamma(i,z,N_{b,z,h})}. \tag{B-43}$$

Then given the estimated mean value or the fixed mean value, which are both denoted as $\overline{\mu}_{b,z,h}$ to simplify the notation, the variance can be estimated by

$$\overline{\sigma}^2_{b,z,h} = \frac{\sum_{i\in\Omega}\gamma(i,z,N_{b,z,h})(b_i-\overline{\mu}_{b,z,h})^2}{\sum_{i\in\Omega}\gamma(i,z,N_{b,z,h})}. \tag{B-44}$$

If one genotype class is consistent with more than one HMM states, the corresponding parameters should be estimated by combining the data. For example, both state 1 and 2 have genotype AA and BB. The corresponding parameters should be estimated by merging the data ($\mu_{b,z,h}$ is already set as 0 or 1):

$$\overline{\sigma}^2_{b,1,1} = \overline{\sigma}^2_{b,2,1} = \frac{\sum_{i\in\Omega}\sum_{z=1}^{2}\gamma(i,z,N_{b,z,h})b_i^2}{\sum_{i\in\Omega}\sum_{z=1}^{2}\gamma(i,z,N_{b,z,h})}, \tag{B-45}$$

$$\overline{\sigma}^2_{b,1,3} = \overline{\sigma}^2_{b,2,2} = \frac{\sum_{i\in\Omega}\sum_{z=1}^{2}\gamma(i,z,N_{b,z,h})(b_i-1)^2}{\sum_{i\in\Omega}\sum_{z=1}^{2}\gamma(i,z,N_{b,z,h})}. \tag{B-46}$$

The parameters of the tissue contamination mixtures are estimated under the assumption of truncation at 0 or 1, respectively. Note we assume the distribution is only truncated at one side because the probability of truncation at the other side is extremely small. We first briefly introduce the parameter estimations of truncated normal in regular situations with i.i.d. observations and then describe a modified version for our HMM.

Following Halperin [2], suppose $(x_1, x_2, ..., x_n)$ are n independent samples from $N(\mu, \sigma^2)$ distribution and the value of the s samples $(x_1, x_2, ..., x_s)$ that are smaller than T are observed while for the remaining $n-s$ samples, we only know they are no less than T, but do not know their values. Thus the sample likelihood is given by

$$p(x_1,...,x_s) = \frac{n!}{(n-s)!}\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^s\left[\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{s}(x_i-\mu)^2\right)\right] \left[\frac{1}{\sqrt{2\pi}}\int_{(T-\mu)/\sigma}^{\infty}\exp(-z^2/2)dz\right]^{n-s}. \tag{B-47}$$

Let $h = (T - \mu)/\sigma$ and let $\overline{x}$ be the sample mean of $(x_1, x_2, ..., x_s)$,

$$\log p(x_1, ..., x_s) = \text{const} - s \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{s} (x_i - T)^2 - \frac{sh^2}{2} - \frac{sh}{\sigma}(\overline{x} - T) + (n - s) \log \left(\frac{1}{\sqrt{2\pi}} \int_h^\infty \exp(-z^2/2) dz \right). \quad \text{(B-48)}$$

Then the MLE can be obtained by solving the following two equations:

$$\frac{\partial \log p(x_1, ..., x_s)}{\partial \sigma} = -\frac{s}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{s} (x_i - T)^2 + \frac{sh}{\sigma^2}(\overline{x} - T) = 0, \quad \text{(B-49)}$$

$$\frac{\partial \log p(x_1, ..., x_s)}{\partial h} = -sh - \frac{s(\overline{x} - T)}{\sigma} - (n - s) g(h) = 0, \quad \text{(B-50)}$$

where

$$g(h) = \frac{\frac{1}{\sqrt{2\pi}} \exp(-h^2/2)}{\frac{1}{\sqrt{2\pi}} \int_h^\infty \exp(-z^2/2) dz}. \quad \text{(B-51)}$$

Equation (B-49) and (B-50) can be further written as

$$s\sigma^2 + sh(T - \overline{x})\sigma - \sum_{i=1}^{s} (x_i - T)^2 = 0, \quad \text{(B-52)}$$

$$\frac{s(T - \overline{x})}{sh + (n - s) g(h)} = \sigma. \quad \text{(B-53)}$$

Therefore

$$\sigma = \frac{s(T - \overline{x})}{sh + (n - s) g(h)} = \frac{T - \overline{x}}{2} \left(-h + \sqrt{h^2 + V_{pn}^2} \right), \quad \text{(B-54)}$$

where

$$V_{pn}^2 = \frac{4 \sum_{i=1}^{s} (x_i - T)^2}{s(T - \overline{x})^2} = \frac{4\overline{x^2} - 8\overline{x}T + 4T^2}{(T - \overline{x})^2}, \quad \text{(B-55)}$$

and

$$\overline{x^2} = \frac{\sum_{i=1}^{s} x_i^2}{s}. \quad \text{(B-56)}$$

After some further simplification,

$$g(h) = \frac{p}{(1 - p) V_{pn}^2} \left((2 - V_{pn}^2) h + 2\sqrt{h^2 + V_{pn}^2}, \right), \quad \text{(B-57)}$$

where $p = s/n$ is the proportion of un-truncated data. Solution of Equation (B-57), $\hat{h}$ can be found by numerical method. Then μ can be estimated by $T - \hat{h}\sigma$, and σ can be estimated by Equation (B-54) by plugging in $\hat{h}$.

If the samples are left-truncated. The above procedure can still be used. After we obtain $\hat{h}$, just flip its sign to get MLE of h. The standard deviation can be estimated by

$$\sigma = \frac{s(T-\overline{x})}{s\hat{h} - (n-s)g(-\hat{h})} = \frac{T-\overline{x}}{2}\left(-\hat{h} - \sqrt{\hat{h}^2 + V_{pn}^2}\right). \tag{B-58}$$

All the above discussion are based on observations $(x_1, x_2, ..., x_n)$ that are definitely generated from a truncated normal distribution. However, in the EM algorithm of HMM, we do not have a group of b_is that are exactly from the truncated normal distribution. Instead, we have the posterior probability that each b_i is from the truncated normal distribution. This problem can be circumvented as follows. In order to estimate the mean and the standard deviation, the sufficient statistics needed are simply the first/second moments of the un-truncated data ($\overline{x^2}$ and $\overline{x}$ in equation (B-55)) and the proportion of un-truncated data (p), which can all be estimated easily by b_is and the related posterior probabilities. For example, in order to estimate the parameters for the mixture (A, AB), we first estimate $\gamma(i, 4, N_{b,4,2})$, the posterior probability that b_i arises from the mixture (A, AB). Then let $\Omega = \{i; 0 < b_i < 1\}$ and $\Omega_0 = \{i; b_i = 0\}$,

$$\overline{b} = \frac{\sum_{i\in\Omega}\gamma(i,4,N_{b,4,2})b_i}{\sum_{i\in\Omega}\gamma(i,4,N_{b,4,2})}, \tag{B-59}$$

$$\overline{b^2} = \frac{\sum_{i\in\Omega}\gamma(i,4,N_{b,4,2})b_i^2}{\sum_{i\in\Omega}\gamma(i,4,N_{b,4,2})}, \tag{B-60}$$

$$p = \frac{\sum_{i\in\Omega}\gamma(i,4,N_{b,4,2})}{\sum_{i\in\Omega}\gamma(i,4,N_{b,4,2}) + \sum_{i\in\Omega_0}\gamma(i,4,N_{b,4,2})}. \tag{B-61}$$

Now given the sample moments of the un-truncated data ($\overline{b}$ and $\overline{b^2}$), and estimation of the proportion of un-truncated data (p), we can use the above approach to estimate the mean and variance of the genotype class (A, AB).

C Empirical measurements of tumor purity

Let β_{O} and β_{T} be the observed mean value of BAF and expected BAF in pure tumor tissue. Given the assumption that BAF can be approximated by the ratio of the number of B alleles and the total number of alleles, we have

$$\beta_{\mathrm{T}} = \frac{n_B}{n_A + n_B}, \tag{C-1}$$

$$\beta_{\mathrm{O}} = \frac{1 - p_{\mathrm{T}} + p_{\mathrm{T}} n_B}{2 - 2p_{\mathrm{T}} + p_{\mathrm{T}}(n_A + n_B)}, \tag{C-2}$$

where n_A and n_B are the number of A or B alleles in pure tumor tissue, respectively. Therefore p_T can be estimated as

$$p_T = \frac{1 - 2\beta_O}{\beta_O(n_A + n_B - 2) + 1 - n_B}. \tag{C-3}$$

Let $\beta_{O,G}$ be the observed mean BAF value for genotype or genotype mixture G. When copy number is 1, we update $\beta_{O,(A,\underline{AB})}$ and $\beta_{O,(B,\underline{AB})}$ by taking into account of systematic dye bias as follows:

$$\hat{\beta}_{O,(A,\underline{AB})} = 0.5\beta_{O,(A,\underline{AB})}/\beta_{O,AB}, \tag{C-4}$$

$$\hat{\beta}_{O,(B,\underline{AB})} = 0.5 + 0.5(\beta_{O,(B,\underline{AB})} - \beta_{O,AB})/(1 - \beta_{O,AB}). \tag{C-5}$$

Then we estimate $\beta_{O,(A,\underline{AB})}$ by averaging $\hat{\beta}_{O,(A,\underline{AB})}$ and $1 - \hat{\beta}_{O,(B,\underline{AB})}$:

$$\begin{aligned} \beta_{O,1} &= 0.5\left(\hat{\beta}_{O,(A,\underline{AB})} + 1 - \hat{\beta}_{O,(B,\underline{AB})}\right) \\ &= 0.25\left(\beta_{O,(A,\underline{AB})}/\beta_{O,AB} + 1 - (\beta_{O,(B,\underline{AB})} - \beta_{O,AB})/(1 - \beta_{O,AB})\right). \end{aligned} \tag{C-6}$$

Similarly, $\beta_{O,(AA,\underline{AB})}$ is estimated by

$$\beta_{O,2} = 0.25\left(\beta_{O,(AA,\underline{AB})}/\beta_{O,AB} + 1 - (\beta_{O,(BB,\underline{AB})} - \beta_{O,AB})/(1 - \beta_{O,AB})\right). \tag{C-7}$$

$\beta_{O,(AAB,\underline{AB})}$ is estimated by

$$\beta_{O,3} = 0.25\left(\beta_{O,(AAB,\underline{AB})}/\beta_{O,AB} + 1 - (\beta_{O,(BBA,\underline{AB})} - \beta_{O,AB})/(1 - \beta_{O,AB})\right). \tag{C-8}$$

Then we can separately plug in $\beta_{O,1}$, $\beta_{O,2}$ and $\beta_{O,3}$ into equation (C-3), together with the corresponding n_A and n_B, to estimate p_T. We denote the estimates of p_T from $\beta_{O,1}$, $\beta_{O,2}$, and $\beta_{O,3}$ as p_{T1}, p_{T2}, and p_{T3}, respectively. As shown in Figure 5 (a) in the main text, p_{T1} and p_{T2} are highly consistent. Overall, p_{T1} and p_{T3} are also consistent (Figure C-1). However p_{T1} and p_{T3} have larger discrepancy than p_{T1} and p_{T2}. This may be due to the dye bias, which has been completely corrected by equation C-4 and C-5.

The clinically estimated tumor purity tends to be very high despite the apparent pattern in the data which indicates a relatively low tumor purity. Figure C-2 shows an example.

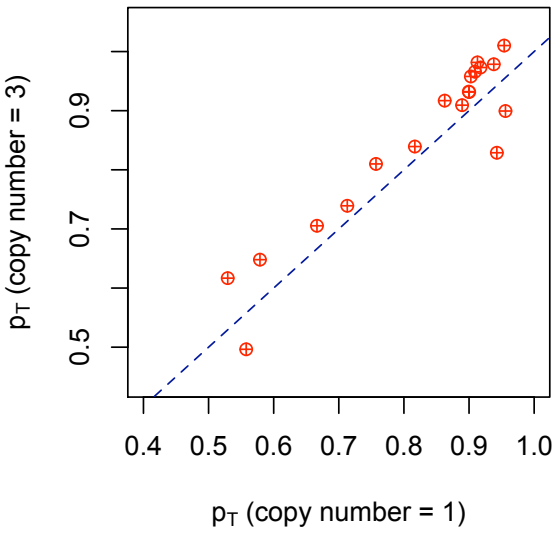

Figure C-1: Comparison of the proportion of tumor sample (p_{T}) estimated using mean BAF values when copy number is one (genotype (A, AB)) and three (genotype (AAB, AB)). Compared with (a), the p_{T3} in (b) are adjusted by subtracting a constant so that the maximum of p_{T3} is 0.99

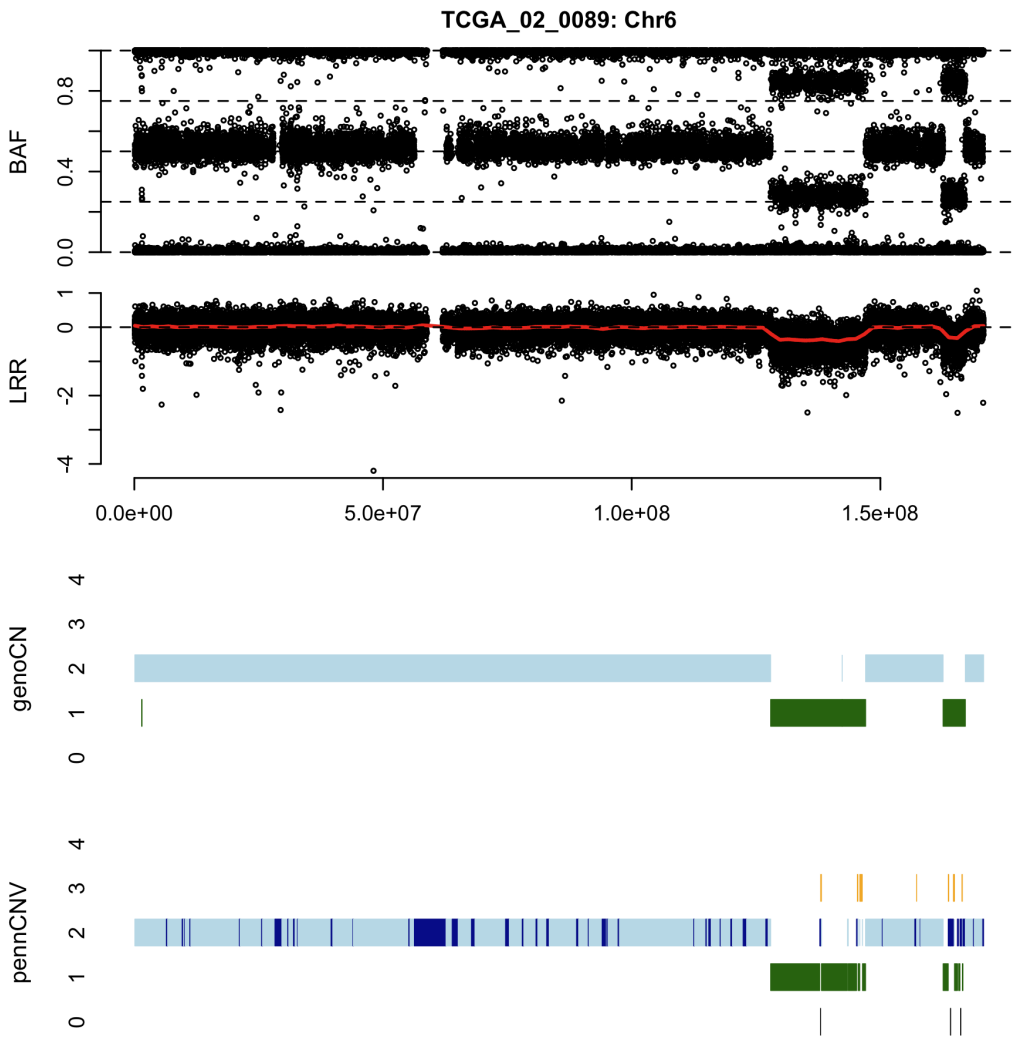

Figure C-2: An example that the extra band in the BAF plot (when copy number is one) indicates a relatively low tumor purity (71% from our data-driven estimates), while the clinically estimated tumor purity is 100%.

Table C-1: Comparison of data-driven tumor purity estimates (estimated using CNA with copy number 1) and clinical tumor purity estimates.

Sample	Tumor Purity Estimate	
	Data-driven	Clinical
TCGA_02_0003	0.9	1
TCGA_02_0007	0.94	1
TCGA_02_0009	0.92	1
TCGA_02_0014	0.96	1
TCGA_02_0021	0.89	1
TCGA_02_0028	0.94	1
TCGA_02_0033	0.56	1
TCGA_02_0034	0.58	1
TCGA_02_0037	0.86	1
TCGA_02_0038	0.82	1
TCGA_02_0046	0.91	1
TCGA_02_0054	0.53	0.95
TCGA_02_0064	0.67	1
TCGA_02_0083	0.9	1
TCGA_02_0089	0.71	1
TCGA_02_0099	0.76	0.975
TCGA_02_0102	0.9	0.975
TCGA_02_0114	0.95	1
TCGA_02_0116	0.91	1

References

[1] Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S., Hakonarson, H., and Bucan, M. (Nov, 2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.,* **17**, 1665–1674.

[2] Halperin, M. (1952) Estimation in the truncated normal distribution. *Journal of the American Statistical Association,* **47**(259), 457–465.