Accurate estimation of homologue-specific DNA concentration-ratios in cancer samples allows long-range haplotyping

Scott L. Carter[1,3,*], Matthew Meyerson[2,3], Gad Getz[3,*]
1 Division of Health Science and Technology, MIT, Cambridge, MA, USA
2 Dana-Farber Cancer Institute, Boston, MA, USA
3 Broad Institute, Cambridge, MA, USA
∗ **E-mail: scarter@broadinstitute.org, gadgetz@broadinstitute.org**

Abstract

Interpretation of allelic copy measurements at polymorphic markers in cancer samples presents distinctive challenges and opportunities. Due to frequent gross chromosomal alterations occurring in cancer (aneuploidy), many genomic regions are present at homologus-allele imbalance. Within such regions, the unequal contribution of alleles at heterozygous markers allows for direct phasing of the haplotype derived from each individual parent. In addition, genome-wide estimates of homologue specific copy-ratios (HSCRs) are important for interpretation of the cancer genome in terms of fixed integral copy-numbers. We describe HAPSEG, a probabilistic method to interpret bi-allelic marker data in cancer samples. HAPSEG operates by partitioning the genome into segments of distinct copy number and modeling the four distinct genotypes in each segment. We describe general methods for fitting these models to data which are suitable for both SNP microarrays and massively parallel sequencing data. In addition, we demonstrate a specially tailored error-model for interpretation of systematic variations arising in microarray platforms. The ability to directly determine haplotypes from cancer samples represents an opportunity to expand reference panels of phased chromosomes, which may have general interest in various population genetic applications. In addition, this property may be exploited to interrogate the relationship between germline risk and cancer phenotype with greater sensitivity than is possible using unphased genotype. Finally, we exploit the statistical dependency of phased genotypes to enable the fitting of more elaborate sample-level error-model parameters, allowing more accurate estimation of HSCRs in cancer samples.

Author Summary

The human genome typically exists in two copies of each chromosome, with one copy, or *homologue* derived from either parent. One of the most fundamental hallmarks of human cancers is their tendency to have *aneuploid* genomes, that is, unbalanced copy-number alterations in the genetic material at various blocks of the normal human genome. This often results in unequal contributions of the homologues derived from either parent to the genomes of cancer cells. Estimation of the precise contribution of each homologue in a DNA sample obtained from cancer tissue is crucial to understand the genetic alterations occurring specifically in the cancer cells. Such estimation requires the identification and interpretation of alterations in the genetic sequence of each homologue, such as they appear in the cancer sample, which contains a mixture of DNA derived from both

cancerous and normal cells. We have developed a statistical method to accurately model the resulting data, producing summarizations of the information from hundreds of thousands of data points into hundreds of segments, describing the concentration of both homologues for large genomic regions of equal copy-number. In addition, we show that the tendency of cancer samples to be aneuploid may be exploited to attribute contiguous chromosomal blocks of genetic variation to one or the other homologue, representing the contribution of a single parent. Because they can help reduce the dimensionality of the data, such summarizations will be useful to further understand the interaction between inherited genetic variation and cancer development.

Introduction

The genomes of human cancer cells frequently harbor copy-number alterations, ranging from focal gain and loss of small regions to widespread chromosomal aneuploidy [1], [2], in many cases exacerbated by DNA ploidy increases followed by predominant attrition of the fixed DNA in the evolving somatic clone [3]. Allelic analysis of cancer genomes offers several advantages in analysis of cancer genomes [4], [5]. Detection of genomic regions with fixed somatic loss of heterozygosity - (LOH) helps identify recessively inactivated tumor suppressors, carrying mutations on the retained allele [6]. In particular, allelic measurements are required for detection of copy-neutral LOH, which may occur either due to compensatory gain of the retained alleles, or by homologous recombination. Finally, genome-wide HSCR estimates provide the foundation for inference of tumor-nuclei percentage, cancer-genome ploidy, and integral allelic copy-numbers [7], [8], [9], [3].

Human genomes are normally diploid, with one haploid genome inherited from each parent (although tetraploid cells are also involved in physiological processes [10].) As a result of widespread chromosomal aneuploidy, many genomic regions are in homologos-allele imbalance, with the two homologues fixed at unequal copy-numbers within the somatic clone [3]. Methods for genotyping diploid samples are unsuitable for analysis of aneuploid cancer samples. For example, discovery of both germline and somatic point-mutations in cancer samples using massively paralell sequencing typically requires analysis of paired cancer-normal DNAs [11], [12]. For SNP microarrays, probes for each SNP allele are calibrated using diploid control samples [13], making genotype calls unreliable in aneuploid samples [14].

Accurate estimation of homologue-specific DNA concentrations in tumor samples is challenging due to substantial marker-level noise occurring on a background of complex biases related to the genotypes, sequence context, allelic copy-numbers at which loci are observed. The particular manifestation of these factors in the resulting data depends on physical properties of the technology used. Development of explicit generative models describing these effects in particular samples can provide greater sensitivity to detect true alterations in the underlying cancer genome. We extend previous physically grounded error models for gene expression microarrays [15], [16] for use with Affymetrix SNP arrays, facilitating our ability to estimate HSCRs in multiple cancer-derived datasets.

Early work with SNP microarrays demonstrated the feasibility of allelic cancer-genome analysis [17, 18, 4], setting the stage for development of methods to detect allele-specific amplification [19], and LOH in unpaired samples [20]. Development of high-resolution Affymetrix SNP6.0 microarrays was enabled by the development of a method to accurately calibrate allelic probes, allowing for highly accurate genotyping without the need for mismatch probes [13]. These calibrations became

the basis for modern cancer total copy-ratio analysis using these arrays [21], [22], [23].

The development of methods for allelic analysis in cancer samples has been further pursued to identify genomic regions of LOH [24], and as part of solutions to the tumor purity/ploidy problem using Illumina bead-arrays [7], [8], and in Affymetrix SNP6.0 microarrays [25], [14], [9]. Our view is that this problem is best treated separately, allowing for detailed treatment of the HSCR problem and facilitating comparison of various methods specifically addressing this problem [26], [27], [28].

We present a novel computational method, HAPSEG, for accurate estimation of haplotype-specific DNA copy-ratios (HSCRs), offering several specific advantages over existing methods: (i) implementation of an advanced error model tailored to the basic physics of Affymetrix SNP microarray measurements; (ii) Internal recomputation of genomic segmentation using error-model fit; (iii) utilization of LD information from phased haplotype panels [29], improving inference of genotypes by exploiting statistical dependencies between the genotypes of adjacent markers, and improving HSCR estimation.

We present a novel demonstration of direct haplotype phasing using allelic imbalances in cancer-tissue samples. These are of interest to further characterize the influence of germline risk in cancer development. In addition, this capability may be used to generate densley typed reference panels of phased chromosomes for use with imputation of rare alleles in whole exome sequencing

Results

Method overview

Consider a defined set of polymorphic bi-allelic markers (SNPs), for which data proportional to the concentration of each allele (channels a and b) has been generated. For microarray data, the set of SNPs is fixed by the array design. We define the *copy-ratio* as the ratio of allelic concentration in a cancer-derived DNA sample to that of the haploid locus concentration in an equivalent DNA aliquot derived from diploid cells. The copy-ratio of a given allele depends on both the (germline) genotype and on the concentration in the cancer-derived DNA of the homologue on which it resides (the location).

Calibration is the process by which SNP measurements are standardized to copy-ratios. The specific manner in which calibration is performed depends upon the measurement technology being considered (e.g. microarray or sequencing). After calibration, the expected a and b values for a SNP in a diploid sample (at a region of 2 copies) are (0,2), (1,1), or (2,0), corresponding to genotypes BB, AB/BA, and AA, respectively. These values are observed with substantial noise, the distribution of which depends upon the measurement platform being used.

Because DNA copy-number is expected to be locally constant along the genome, superior HSCR estimates can be obtained by pooling datapoints within segments of constant total copy-ratio. For the purpose of initializing such a segmentation, calibrated signal from the a and b channels is added together genome-wide to produce copy-ratio estimates independently of marker genotypes. These values are input to the a segmentation algorithm, (e.g. CBS [30]), which fits a piecewise constant regression function to these values with respect to their location along the reference genome.

Figure 1 shows an example of HAPSEG on a cancer sample with pervasive allelic imbalance. Calibrated allelic copy-ratio estimates for each marker (fig. 1a) were modeled using HAPSEG,

which performed haploid genotyping and segmentation (fig. 1b). The resulting HSCR estimates are summarized at segmental level (fig 1c), revealing discrete levels occurring at regular copy-ratio intervals, consistent with fixed SCNAs in the cancer clone (fig 1d). These estimates formed the basis of a large-scale inference of absolute copy-numbers in cancer [3].

We assume that, within such a segment, the copy-ratios of both haplotypes will nearly always remain constant. Violations from this assumption would correspond to coincident breakpoints, with compensatory changes in the homologous copy numbers such that the total is unchanged. We note that such cases are theoretically possible (homologous recombination), and have been previously reported using related methods [28]; we leave the treatment of such alterations to future development.

A maximum of two distinct homologous copy-ratios can exist in a given segment, and the sum of A and B alleles at (germline) heterozygous sites must equal the number of A and B alleles at homozygous AA and BB sites, respectively. The allelic concentration ratios therefore generally correspond to four clusters the possible phased genotype of each SNP. The locations of these components, denoted $\boldsymbol{\mu}_{\mathrm{AA}_i}, \boldsymbol{\mu}_{\mathrm{AB}_i}, \boldsymbol{\mu}_{\mathrm{BA}_i}, \boldsymbol{\mu}_{\mathrm{BB}_i}$, are specified by two free parameters: the total copy-ratio τ_i, and the difference of homologous copy-ratios, denoted δ_i. The locations of the homozygous components are then:

$$\boldsymbol{\mu}_{\mathrm{AA}_i} = (0,\, \tau_i)\,,\; \boldsymbol{\mu}_{\mathrm{BB}_i} = (\tau_i,\, 0)\,. \tag{1}$$

The locations of the heterozygous components are:

$$\boldsymbol{\mu}_{\mathrm{AB}_i} = \left(\frac{\tau_i - \delta_i}{2},\, \frac{\tau_i + \delta_i}{2}\right),\; \boldsymbol{\mu}_{\mathrm{BA}_i} = \left(\frac{\tau_i + \delta_i}{2},\, \frac{\tau_i - \delta_i}{2}\right). \tag{2}$$

As segments approach allelic-balance ($\delta_i \to 0$), the two heterozygous clusters become superimposed, with the natural corollary that no information is provided regarding the phase of SNPs in that segment (as when genotyping diploid samples). To model observed allelic data, we represent the four possible phased genotypes using a mixture model with components for each genotype {AA, AB, BA, BB}. Figure 2 demonstrates HSCR estimation on 3 example segments (from a single sample,) at differing homologous copy-numbers (fig 2a,b). The relationship between homologous imbalance and phased genotyping is demonstrated (fig 2c-h).

The calibrated data for M measured markers, $\mathbf{X}$, are a $2 \times M$ matrix of bi-allelic (haploid) copy-ratio point estimates. Segmentations of $\mathbf{X}$ are denoted by $\mathcal{S}$, which specify partitions of $\mathbf{X}$ into N successive matrices $\mathbf{X}_{1\ldots N}$. For each segment i, $\mathbf{X}_i$ has dimension $2 \times M_i$, where M_i refers to the number of SNPs in the segment. The model parameter $\mathbf{C}$ represents the 4 phased genotypes of SNPs $\mathbf{X}$. Denote by $\mathbf{X}_{ij}$ the a and b channels for SNP j in segment i, and by $\mathbf{C}_{ij}$ the genotype of SNP j. Prior information about the genotype of each SNP may be available from population allele-frequencies, from the analysis of a paired-normal sample, or from phased haplotype panels, and is represented as $\mathcal{G}$.

The conditional distribution of $\mathbf{X}$, representing the mapping from segment locations $\boldsymbol{\mu}_i$ to probability densities over observed $\mathbf{X}_i$, is denoted $\mathcal{P}\left(\mathbf{X}_i | \boldsymbol{\mu}_i, \boldsymbol{\Theta}\right)$, where $\boldsymbol{\Theta}$ represents the set of sample-level parameters representing specific sources of experimental fluctuation. Estimation of $\boldsymbol{\Theta}$ at the sample-level increases our power to fit realistically complex error-models without the risk of over-fitting small segments. The function $\mathcal{P}$ and parameters $\boldsymbol{\Theta}$ are specific to the measurement platform in use. We developed a novel error-model for Affymetrix SNP microarrays (Methods) and applied it throughout.

We specify the probability density of an observed point, conditional on g as:

$$P\left(\mathbf{X}_{ij}|\mathbf{C}_{ij}=g,\delta_i,\tau_i,\mathbf{\Theta}\right)\equiv\mathcal{P}\left(\mathbf{X}_{ij}|\boldsymbol{\mu}_{ig},\mathbf{\Theta}\right). \tag{3}$$

The complete likelihood of segment i is therefore:

$$\mathcal{L}_i\left(\mathbf{X}_i|\delta_i,\tau_i,\mathbf{\Theta},\mathcal{G}\right)=\prod_{j=1}^{M_i}Z_{ij}, \tag{4}$$

where Z_{ij} denotes the complete likelihood of SNP j:

$$Z_{ij}=\sum_{g=1}^{4}P\left(\mathbf{X}_{ij}|\mathbf{C}_{ij}=g,\delta_i,\tau_i,\mathbf{\Theta}\right)P\left(\mathbf{C}_{ij}=g|\mathcal{G}\right). \tag{5}$$

The full sample-likelihood is:

$$\mathcal{L}_f\left(\mathbf{X}|\boldsymbol{\delta},\boldsymbol{\tau},\mathbf{\Theta},\mathcal{S},\mathcal{G}\right)=\prod_{i\in\mathcal{S}}\mathcal{L}_i\left(\mathbf{X}_i|\delta_i,\tau_i,\mathbf{\Theta},\mathcal{S},\mathcal{G}\right). \tag{6}$$

We implemented model-fitting using general numerical optimizations which are independent of the specific likelihood; support for alternate measurement platforms may be added by implementing an appropriate density for $\mathcal{P}\left(\mathbf{X}_i|\boldsymbol{\mu}_i,\mathbf{\Theta}\right)$. We describe an general algorithm for joint estimation of sample-level error-model parameters Θ and segmental allelic copy-ratios δ,τ, denoted $\mathbf{C^*}$, allowing updates of these parameters based on $\{\mathbf{X},\hat{\mathcal{S}},\mathcal{G}\}$: $\{\hat{\mathbf{\Theta}},\hat{\boldsymbol{\delta}},\hat{\boldsymbol{\tau}}\}=\mathbf{C^*}\left(\hat{\mathbf{\Theta}}^{(n)}|\mathbf{X},\hat{\mathcal{S}},\mathcal{G}\right)$.

We use the **C***algorithm to iteratively update the genomic segmentation $\mathcal{S}$, genotype probabilities $\mathcal{G}$, sample error-model parameters $\mathbf{\Theta}$, and HSCR locations $\boldsymbol{\delta},\boldsymbol{\tau}$.

Method HAPSEG:

1. **Initialize:** Compute $\hat{\mathcal{S}}^{(1)}$ from $\mathbf{X}$. Set $\mathbf{\Theta}^{(0)},\boldsymbol{\delta}^{(0)},\boldsymbol{\tau}^{(0)},\mathcal{G}^{(0)}$.
2. **Fit error-model / segment-locations:** $\{\hat{\mathbf{\Theta}}^{(1)},\hat{\boldsymbol{\delta}}^{(1)},\hat{\boldsymbol{\tau}}^{(1)}\}=\mathbf{C^*}\left(\mathbf{\Theta}^{(0)}|\mathbf{X},\hat{\mathcal{S}}^{(1)},\mathcal{G}^{(0)}\right)$.
3. **Refine segmentation:** Compute $\hat{\mathcal{S}}^{(2)}$ from $\{\hat{\mathbf{\Theta}}^{(1)},\hat{\boldsymbol{\delta}}^{(1)},\hat{\boldsymbol{\tau}}^{(1)}\}$ Estimation of $\hat{\mathbf{\Theta}}^{(1)}$ allows for more sensitive and specific evaluation of proposed segmental breakpoints. Consider a pair of adjacent segments in the reference genome. Because the initial segmentation $\hat{\mathcal{S}}^{(1)}$ may have introduced a spurious breakpoint, we develop a probabilistic criterion for their merger using Bayesian model comparison (Methods).
4. **Update:** $\{\hat{\mathbf{\Theta}}^{(2)},\hat{\boldsymbol{\delta}}^{(2)},\hat{\boldsymbol{\tau}}^{(2)}\}=\mathbf{C^*}\left(\hat{\mathbf{\Theta}}^{(1)}|\mathbf{X},\hat{\mathcal{S}}^{(2)},\mathcal{G}^{(0)}\right)$.
5. **Haploid genotype estimation:** compute $\hat{\mathbf{C}}^{(1)}$ from $P(\mathbf{C}|\mathbf{X},\hat{\mathbf{\Theta}},\hat{\boldsymbol{\delta}},\hat{\boldsymbol{\tau}},\mathcal{G}^{(0)})$. For a segment i, posterior distributions $\hat{\mathbf{C}}_i$ are computed for each SNP j as follows:

$$\hat{\mathbf{C}}_{ij}=\frac{1}{Z_{ij}}P\left(\mathbf{X}_{ij}|\mathbf{C}_{ij}=g,\hat{\delta}_i,\hat{\tau}_i,\hat{\mathbf{\Theta}}\right)P\left(g|\mathcal{G}^{(0)}\right), \tag{7}$$

where Z_{ij} is computed as in (5). $\hat{\mathbf{C}}^{(1)}$ contains the probability distribution of the haploid genotypes, with the phase information provided by homologous copy-imbalance in the somatic clone.

6. **Reconciliation of phased genotype estimates with reference panels:** compute $\hat{\mathcal{G}}^{(1)}$ from $\hat{\mathbf{C}^{(1)}}$, $\mathbf{D}$, $\mathbf{m}$.

   This procedure examines the evidence for the phased genotypes in a given segment in a panel of phased reference chromosomes characterizing population diversity of haplotypes Consortium:2010en. This is accomplished using the statistical program BEAGLE [31] to phase diploid genotypes estimates computed from $\hat{\mathbf{C}^{(1)}}$. This procedure utilizes phased panels $\mathbf{D}$, and an estimate of the genetic recombination rate $\mathbf{m}$ to compute a maximum-likelihood estimate of phase. HAPSEG then identifies and corrects 'switch-errors' in this phasing by reconciling the estimates with those based on homologous imbalance ($\hat{\mathbf{C}^{(1)}}$). The phasing ability of HAPSEG was validated by examining the local concordance with maximum likelihood phase estimates produced via BEAGLE [31], a statistical method based on a population reference-panel of phased chromosomes (fig 3). We demonstrate that phase estimates produced by BEAGLE are locally concordant with those implied by HAPSEG, but that switch-errors tend to occur at regions of high recombination rate (fig 3a,b).

7. **Final estimates:**

$$\{\hat{\mathbf{\Theta}}, \hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\tau}}\} = \mathbf{C*}\Big(\hat{\mathbf{\Theta}}^{(2)} | \mathbf{X}, \hat{\mathcal{S}}^{(2)}, \hat{\mathcal{G}}^{(1)}\Big) \tag{8}$$

$$\hat{\mathbf{C}}^{(2)} = P\left(C_{ijk} | \mathbf{X}_{ij}, \hat{\delta}_i, \hat{\tau}_i, \hat{\mathbf{\Theta}}, \mathcal{G}^{(1)}\right) \tag{9}$$

8. **Standard error of location estimates.** With the Hessian $\mathbf{A}_i$ as in eq.(14), standard errors on the segment locations are:

$$\sigma_{\delta_i} = \sqrt{|\mathbf{A}^{-1}_{i_{11}}|},\ \sigma_{\tau_i} = \sqrt{|\mathbf{A}^{-1}_{i_{22}}|}.$$

   This approximation is valid given that the posterior distribution of δ and τ is multivariate normal.

Affymetrix error model. Following a classic microarray error-model [15] we consider the observed calibrated signal X to be distributed according to:

$$X = \mu e^{\eta} + \epsilon, \tag{10}$$

$$\eta \sim \mathcal{N}(0, \sigma_\eta),\ \epsilon \sim \mathcal{N}(0, \sigma_\epsilon),$$

where μ represents the true copy-ratio, η and ϵ represent multiplicative and additive errors, respectively. Importantly, this allows for specifying cluster variance as a function of the mean, without the need to fit additional segment-level variance parameters, as in [28].

We generalized this error model to the two-dimensional case appropriate for fitting two-channel microarray data (Methods, eq. (17)). Because an explicit formula for likelihood calculations could not be obtained, we used an approximation whereby a transformation is applied to the data, after which its distribution becomes approximately bi-variate Gaussian. Specifically, we extended the variance stabilization technique described for the one-dimensional case [32] to the 2D error model, accounting for additional observed correlations in heterozygous SNP clusters, which were presumed to arise from the PCR amplification. Additional parameters are included in $\mathbf{\Theta}$ taking into account attenuation and background fluctuation (Methods). The quality of the model-fit to the marker-

level data is shown in figure 4, demonstrating that much of the variation in the data is accurately captured.

Discussion

We have demonstrated the ability of HAPSEG to accurately estimate haplotype-specific DNA concentrations in tumor samples, including those for which no patient-matched normal sample exists. We demonstrated the utility of HSCR estimates from HAPSEG to estimate tumor purity, ploidy, and absolute copy-numbers using ABSOLUTE [3]. By factorizing general inference methods from platform-specific error models, greater generality was achieved, allowing us to easily adapt HAPSEG to the analysis of massively parallel sequencing data. Although promising results have been obtained from whole-exome hybrid capture sequencing (WES)[33] data (not shown), further development is needed to derive calibrated copy-estimates from multiplexed WES using bar-coded reads.

In addition to providing accurate estimates of haplotype-specific copy-ratios, HAPSEG can produce partially phased haplotypes by exploiting the widespread allelic-imbalance common in aneuploid cancer genomes, without the need for methods based on phased reference panels. This will have several applications in cancer genetics, such as establishing compound heterozygosity of germline alleles. In addition, when used with sequencing data, long-range haplotyping using HAPSEG will aid reconstruction of complete somatic karyotype and history of aberrations transforming from the diploid state[34]. Furthermore, these methods may aid detection of specific SCNAs subject to allelic bias in cancer [35, 36, 8].

The availability of partially phased genotypes will allow improvements in panel-based imputation methods, which attempt to model observed diploid genotypes as a mosaic of phased reference haplotypes [37], [38], [39], [40], [31], (reviewed[41]). Optimal exploitation of will require extension of these methods to handle partially phased genotype data containing variable gaps (due to somatic allele-balance.)

Similar considerations apply to admixture mapping, whereby long-range phasing may be inferred by differences in parental allele-frequency along the genome, due to recent mixing of outbred populations [42], [43]. The phased homologues provided by HAPSEG, each of which is derived from a single parent, will allow for more sensitive detection of recent admixture, possibly supporting mapping with less divergent populations than permitted in the purely diploid case. Since admixture inferences allow for inter-chromosomal phasing of homologues, they may be of interest in the analysis of germline risk loci.

Long-range haplotyping has utility to uncover interactions between germline and somatic genetics in cancer. Examples include variants at the 8q24 locus, affecting inherited risk of development of multiple cancer types [42], and later shown to interact with MYC [44]. Furthermore, these methods may facilitate identification of additional germline alleles mechanistically predisposed to somatic alteration in cancer, such as JAK2 in myeloproliferative neoplasms [45]. In contrast, to the above examples, the genetic basis by which EGFR mutations in lung adenocarcinoma occur predominately in female non-smoking patients of East Asian ethnicity [46] remains unclear.

As whole-genome sequencing of aneuploid cancer samples becomes increasingly routine, directed long-range haplotyping using HAPSEG will increase the available panels of densely genotyped chromosomes, without the need for genotyping parent-child trios. This resembles the precedent of geno-

typing uniparental-disomic samples derived from complete hydatidiform moles [47], which formed an important component of the phased Japanese HAPMAP panel [48]. HAPSEG generalizes this requirement to allow direct phasing from homologous imbalance, extending the samples on which it may be applied to tens of thousands of cancer samples [49].

Long-range haplotyping is an important technique for human population genetics, with wide ranging applications including understanding haplotype diversity and genetic recombination rates [29], illuminating recent positive selection in the human lineage [50] and human history [51].

Partially phased genotyping by analysis of cancer-derived DNA using HAPSEG represents a novel technique to obtain such data. This may be particularly useful in the study of poorly characterized populations, for which phased reference panels do not exist, and for which parent-child trio data may be impossible to obtain. In such cases, this strategy may present a cost-effective alternative to recently described molecular methods for long-range haplotyping [52], [53].

Methods

Algorithm C*

C* uses the data, segmentation $\mathcal{S}$, and genotype priors $\mathcal{G}$ to update estimates of the sample error-model $\boldsymbol{\Theta}$ and segment locations $\{\boldsymbol{\delta}, \boldsymbol{\tau}\}$. This is accomplished by iterating two conditional updates until convergence of $\mathcal{L}_f$:

1. **Conditional update of locations $\{\boldsymbol{\delta}, \boldsymbol{\tau}\}$:**

$$\forall_i,\ \left\{\delta_i^{(t+1)}, \tau_i^{(t+1)}\right\} = \underset{\{\delta_i,\tau_i\}}{\operatorname{argmax}} \log \mathcal{L}_i \left(\mathbf{X}_i | \delta_i, \tau_i, \boldsymbol{\Theta}^{(t)}, \hat{\mathcal{S}}, \mathcal{G}\right) \tag{11}$$

2. **Conditional update of sample error-model $\boldsymbol{\Theta}$.** $\boldsymbol{\Theta}$ is a vector of platform-specific parameters upon which conditional updates are performed serially. We found that the results are insensitive to ordering, and suppress indexing:

$$\boldsymbol{\Theta}^{(t+1)} = \underset{\boldsymbol{\Theta}}{\operatorname{argmax}} \log \mathcal{L}_f \left(\mathbf{X} | \boldsymbol{\delta}^{(t)}, \boldsymbol{\tau}^{(t)}, \boldsymbol{\Theta}, \hat{\mathcal{S}}, \mathcal{G}\right) P(\boldsymbol{\Theta}). \tag{12}$$

   Note the superscripts (t) above denote the internal iterations of the **C***algorithm.

3. **Stop if** the full likelihood $\mathcal{L}_f$ does not change significantly.

Probabilistic segment merging

Let $\mathcal{H}_0$ and $\mathcal{H}_1$ denote the seperate and merged segment hypotheses. In order to compare the evidence supporting each model, we compute an approximation to the Bayesian evidence (Laplace) by approximating the normalizing constant of the posterior distribution $P(\delta_i, \tau_i | \mathbf{X}_i, \hat{\boldsymbol{\Theta}})$. If the posterior is multivariate Gaussian, the normalization constant may be calculated from the likelihood

and curvature at the mode:

$$\mathrm{Ev}(\mathbf{X}_i) = \int_{\delta_i}\int_{\tau_i} \mathcal{L}_i(\mathbf{X}_i|\delta_i,\tau_i,\hat{\mathbf{\Theta}},\mathcal{G}^{(0)}) \tag{13}$$

$$\simeq \hat{\mathcal{L}}_i \times \left|\frac{\det \mathbf{A}_i}{2\pi}\right|^{-1/2} \times (5/2)^{-2}, \tag{14}$$

where $\hat{\mathcal{L}}_i$ denotes the maximum likelihood (at $\hat{\delta}_i, \hat{\tau}_i$), and $\mathbf{A}_i = -\nabla\nabla \log \mathcal{L}_i(\mathbf{X}_i|\delta_i,\tau_i,\hat{\mathbf{\Theta}})$ denote the Hessian matrix around the mode of a segment location: $\{\hat{\delta}_i, \hat{\tau}_i\}$. We verified the accuracy of the approximation using quadrature on a subset of segments (not shown).

The constant third term gives the volume of the domain: $\delta_i \times \tau_i$. Note that the first and second-two terms correspond to the 'Best fit likelihood' × 'Occam factor' formulation of Evidence, as described in [54].

Following Bayes' rule: $P(\mathcal{H}|\mathbf{X}_i) \propto P(\mathbf{X}_i|\mathcal{H})P(\mathcal{H})$, we compute the Bayes factor $\mathrm{BF}(\mathcal{H}_0)$:

$$\frac{P(\mathcal{H}_0|\mathbf{X}_{1,2})}{P(\mathcal{H}_1|\mathbf{X}_{12})} = \frac{\mathrm{Ev}(\mathbf{X}_1)\mathrm{Ev}(\mathbf{X}_2)}{\mathrm{Ev}(\mathbf{X}_{12})}. \tag{15}$$

Our experience suggests that the cancer genome is often heavily over-segmented using the initial segmentation $\hat{\mathcal{S}}^{(1)}$ described above. Adjacent segments are therefore merged if $\mathrm{BF}(\mathcal{H}_0) < 1 \times 10^{-10}$. $\mathrm{BF}(\mathcal{H}_0)$ is computed for all breakpoints in $\hat{\mathcal{S}}^{(1)}$ and segments are merged greedily by joining the adjactent pair with the lowest $\mathrm{BF}(\mathcal{H}_0)$ value. The merge-probability for any breakpoints adjacent to the resulting combined segment are computed at each step. The procedure is finished when no pairs exist with $\mathrm{BF}(\mathcal{H}_0)$ below the threshold, resulting in a refined segmentation $\hat{\mathcal{S}}^{(2)}$. We note that the Bayes factor defined above could equivalently be applied to accept or reject novel proposed breakpoints in $\mathcal{S}$. We leave the development of an efficient method to identify valid breakpoints to the future.

Reconciliation of phased genotype estimates with reference panels

1. Compute diploid genotype estimates $\hat{d}$ from haploid estimates $\hat{\mathbf{C}}^{(1)}$ by collapsing (marginalizing) the two heterozygous clusters.

2. Compute the maximum likelihood statsitical phasing of d using BEAGLE [31]:

   $$\hat{\mathbf{C}}^{(\mathrm{B})} = \mathrm{BEAGLE}(\hat{d}, \mathbf{D}, \mathbf{m})$$

   If evidence exists for the haploytype configurations $\hat{\mathbf{C}}^{(1)}$ as a mosaic of phased chromosomes in $\mathbf{D}$ with recombination rate $\mathbf{m}$, then these configurations will be present in $\hat{\mathbf{C}}^{(\mathrm{B})}$, up to occasional spurious phase reversals (switch-errors), which are expected to occur at frequency related to $\mathbf{m}$.

3. Reconcile haploid estimates from allelic-imbalance $\hat{\mathbf{C}}^{(1)}$ with $\hat{\mathbf{C}}^{(\mathrm{B})}$ by allowing for occasional switch-errors:

   $$\hat{\mathcal{G}}^{(1)} = \mathrm{Viterbi}(\hat{\mathbf{C}}^{(1)}, \hat{\mathbf{C}}^{(\mathrm{B})}, \mathbf{\Phi}_{\mathbf{m}})$$

   This is accomplished using the Viterbi algorithm, which finds the best path through adjacent-marker phased-genotype probabilities $\hat{\mathbf{C}}^{(1)}$ such that they are concordant with $\hat{\mathbf{C}}^{(\mathrm{B})}$, with

occasional switches to the phase-reversed $\hat{\mathbf{C}}^{(-\mathrm{B})}$ (obtained by swapping of the AB and BA genotype probabilities). Switching dynamics are governed by the transition-probability matrix $\mathbf{\Phi_m}$:

$$\mathbf{\Phi_m} = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix}.$$

Thus, switches between $\hat{\mathbf{C}}^{(\mathrm{B})}$ and $\hat{\mathbf{C}}^{(-\mathrm{B})}$ are expected to occur with marginal probability 0.1. We note the adjusting the switch-error probability according to the recombination rate $\mathbf{m}$ would likely be more sensitive, however these values appear to work well on our data (fig. 3). Upon termination, $\hat{\mathcal{G}}^{(1)}$ contains updated knowledge about haploid genotypes $\mathbf{C}$. If no phase information is implied in $\hat{\mathbf{C}}^{(1)}$ (due to allelic balance), $\hat{\mathcal{G}}^{(1)} = \hat{\mathbf{C}}^{(\mathrm{B})}$, since the phase of either configuration in $\hat{\mathbf{C}}^{(1)}$ will be equally probable. Likewise, if $\hat{\mathbf{C}}^{(1)}$ reflects a spurious phasing from over-fitting the locations of a given segment, then the switch probability 0.1 is not adequate to form a path with expected switches at every other marker. In such cases, the locations will be corrected by recomputing the segment HSCRs using $\hat{\mathcal{G}}^{(1)} = \hat{\mathbf{C}}^{(\mathrm{B})}$.

Calibration

(Note: The remaining sections describe methods specific to SNP microarrays)

Calibration of Affymetrix SNP-array measurements was accomplished using the Birdseed algorithm [13], which combines population allele-frequency estimates with location estimates on training samples with known genotypes. These prior estimates are adjusted by Birdseed to fit data from batches of samples. These adjusted location estimates are then used to define the three diploid genotype hybridization-intensity clusters, with locations denoted I_{AA}, I_{AB}, I_{BB}. A large amount of measurement error is thought to arise from variability in the conditions under which the PCR reaction is run. We therefore define batches of PCR (eg., 96-well plates for which the reactions were run in parallel.) Calibrations are performed independently for each of such batches. Genotype locations are estimated using only the normal (non-cancer) samples in the batch, which need not be paired to the cancer samples. Because Birdseed jointly determines the genotypes of the normal samples along with the locations of the genotype clusters for the batch, it is not necessary to use controls with known genotypes. Because normal samples are essential for calibration, we generally recommend that at least 10 such samples be included in each batch.

With the intensity of the 3 genotype-clusters determined by Birdseed, background and scale parameters for the a and b channels of each probe-set can be obtained. The background for each channel is estimated as the observed intensity corresponding to zero copies of that allele:

$$I_{0_a} = I_{BB_a},\ I_{0_b} = I_{AA_b}.$$

The scale for each channel is estimated as the difference in observed intensities corresponding to one and zero copies of that allele:

$$d_{1_a} = I_{AB_a} - I_{BB_a},\ d_{1_b} = I_{AB_b} - I_{AA_b}.$$

Calibrations of observed intensities in cancer samples are then computed as:

$$X_a = \frac{I_a - I_{0_a}}{d_{1_a}},\ X_b = \frac{I_b - I_{0_b}}{d_{1_b}}. \tag{16}$$

The calibrated signals may therefore be interpreted as ratios of the locus-concentration in the tumor-sample to the concentration corresponding to 1 copy in a sample derived from diploid cells.

The above calibration procedure is strictly valid given linear responses throughout the domain of locus-concentration. In fact, attenuation effects are observed for many probe-responses at locus-concentrations corresponding to 2-copies / diploid cell, and become more pronounced with increasing concentration. By performing calibration using only locus concentrations corresponding to 0 and 1 copy, we have sought to base our estimates on values from the linear response range in karyotypically normal samples. Attenuation effects occurring in tumor samples are dealt with in a subsequent section.

Following the procedure described above, we perform an additional probe-set level calibration step designed to remove spurious correlation between the a and b channels arising due to cross-hybridization. This procedure is described in a subsequent section.

Error model

Modern SNP microarrays measure on the order of 100-1000K polymorphic sites present in an input DNA aliquot. Specifically, these arrays allow for the measurement of hybridization occurring between input DNA and substrate-bound oligonucleotides corresponding to each of two forms for each polymorphic locus interrogated (an a and a b channel for each SNP). In ideal cases, the experimental conditions under which the array is run result in hybridization proportional to the concentration of the specific locus being interrogated, plus some unknown background (presumed to arise from non-specific hybridization.)

The motivation for the model in eq. (10) stems from the consideration of two major noise sources inherent in the measurement procedure. A molecule with concentration proportional to μ is subjected to n rounds of PCR amplification, to yield 2^n molecules. Assume $n \sim \mathcal{N}(m, \sigma_m)$, then $\log(\mu) \sim \mathcal{N}$. Hybridization of species to the microarray then results in measured signal $X \sim \mathcal{N}(\mu e^{\eta}, \sigma_{\epsilon})$. The distribution of X is therefore the convolution of a normal and a log-normal distribution, representing the measurement of a process subject to multiplicative noise using a process (hybridization) subject to additive noise.

Although the error-model of eq. 10 fits the observed data reasonably well, consistent positive correlation is observed between the a and b channels of the heterozygous clusters AB and BA. Furthermore, the magnitude of this correlation is directly proportional to $\hat{\sigma}_{\eta}$ (not shown). Noting that the two alleles of any given SNP lie on restriction fragments that typically differ by a single base, we surmise that any variations in PCR efficiency arising due to sequence composition are likely to be shared. This phenomenon affects only the heterozygous clusters because, by definition, one allele is absent for any homozygous SNP, and is not subject to PCR amplification. We therefore generalize the error-model of equation 10 to the 2-dimensional case as follows:

$$\begin{bmatrix} X_a \\ X_b \end{bmatrix} = \alpha + \begin{bmatrix} \mu_a e^{\eta_a} \\ \mu_b e^{\eta_b} \end{bmatrix} + \begin{bmatrix} \epsilon_a \\ \epsilon_b \end{bmatrix}$$

$$\begin{bmatrix} \eta_a \\ \eta_b \end{bmatrix} \sim \mathcal{N}(0, \mathbf{\Sigma}_\eta)\,;\ \mathbf{\Sigma}_\eta = \begin{bmatrix} \sigma_\eta & \rho\sigma_\eta \\ \rho\sigma_\eta & \sigma_\eta \end{bmatrix} \tag{17}$$

$$\begin{bmatrix} \epsilon_a \\ \epsilon_b \end{bmatrix} \sim \mathcal{N}(0, \mathbf{\Sigma}_\epsilon)\,;\ \mathbf{\Sigma}_\epsilon = \begin{bmatrix} \sigma_\epsilon & 0 \\ 0 & \sigma_\epsilon \end{bmatrix}$$

Note that the assumption of diagonal additive covariance (Σ_ϵ) in the above equation is reasonable only because of the calibration procedure which removes cross-hybridization signal prior to fitting the model. This procedure is described in a subsequent section.

Attenuation

The heterozygous segment locations are overdetermined with respect to $\boldsymbol{\delta}$ and $\boldsymbol{\tau}$ (2). For each segment, 8 genotype-cluster coordinates are determined by two free parameters. Observed discrepancies from these dependencies can be exploited for improved estimates of microarray attenuation.

As in [55], we adopt the Langmuir isothermal adsorption model [56] to deal with attenuation effects observed as the concentration of a given hybridized species increases:

$$\theta = \frac{\phi C}{1 + \phi C},$$

where θ indicates the proportion of bound species (presumed to be equivalent to the observed hybridization signal), C indicates the concentration of a particular species, and ϕ is a constant related to the binding affinity of the species to the oligonucleotide probes targeting it. This model has been previously shown to accurately model the attenuation characteristics of microarray probes [16].

Due to the calibration procedure described in eq (16), the observed copy-ratio X is actually the ratio of two isotherms:

$$X = \frac{\phi C_T}{1 + \phi C_T} / \frac{\phi C_N}{1 + \phi C_N},$$

where C_T and C_N are proportional to concentration in the tumor and normal sample, respectively. Let $C_R = C_T / C_N$. Because C_N is defined to be the hybridization intensity corresponding to 1 copy for a given probe-set, we can rewrite the above equation as:

$$X = \frac{C_R(1 + \phi)}{(1 + C_R\phi)}.$$

We can then define a transformation of genotype-cluster coordinates:

$$g(\mu) = \frac{\mu(1 + \phi)}{(1 + \mu\phi)} \tag{18}$$

This transformation is applied to the locations of the 4 genotype-clusters defined in equations 1, 2. This ensures that the physical constraints defined by the model are compatible with the observed data.

Noting that

$$\lim_{C_R \to \infty} \frac{C_R(1 + \phi)}{(1 + C_R\phi)} = \frac{1 + \phi}{\phi}$$

corresponds to the asymptotic saturation copy-ratio, m, gives a direct physical interpretation for the value of ϕ; $\phi = \frac{1}{m-1}$.

Variance-stabilizing transformation

Following [32], we make use of a variance-stabilizing transformation h, such that

$$h(x) \simeq h(\mu) + \epsilon_h;\ \epsilon_h \sim \mathcal{N}(0, \sigma_h), \tag{19}$$

where σ_h is constant over the domain of X. Writing the variance of X as a function of the mean μ : $\mathrm{var}(X) = v(\mu)$, a general form for h can be expressed as [57]:

$$h(x) = \int^x \frac{1}{\sqrt{v(u)}} du.$$

The specific form of $h(x)$ for use with the error-model defined by eq. (10) is derived in [32]:

$$\begin{aligned} h(x) &= \log\left(bx + \sqrt{1 + bx^2}\right) \\ &= \sinh^{-1}(bx). \end{aligned} \tag{20}$$

The transformation has a parameter b that is calculated from σ_η and σ_ϵ as follows:

$$b = \frac{\sqrt{e^{\sigma_\eta^2} - 1}}{\sigma_\epsilon}. \tag{21}$$

The transformed data, denoted $\mathbf{X}'$, will have variance, denoted σ_h^2, approximately independent of $\mathbf{X}$, with

$$\begin{aligned} \sigma_h &= \sigma_\epsilon \frac{d}{dx} h(x)|_{x=0} \\ &= \sigma_\epsilon b. \end{aligned} \tag{22}$$

Using simulated data, we verified that the approximation in eq. 19 is accurate for values of σ_η and σ_ϵ within the range of typical estimates (data not shown).

Likelihood calculation

We construct an approximation to the model defined in eq. (17) based on application of the variance-stabilizing transformation (eq. 20) applied to both data channels. We use the scaled bivariate t-distribution on the transformed data $\mathbf{X}'$, which has an additional parameter $\nu \in \mathbf{\Theta}$ allowing excess density in the tails.

$$\begin{aligned} \mathcal{P}\left(\mathbf{X}_{ij} | \boldsymbol{\mu}_{ik}, \mathbf{\Theta}\right) &\equiv P(\mathbf{X}_{ij} | \boldsymbol{\mu}_{ik}, \nu, \mathbf{\Sigma}_i) \\ &= \frac{1}{2\pi |\mathbf{\Sigma}_i|^{1/2}} \left(1 + \nu \left(\mathbf{X}'_{ij} - \boldsymbol{\mu}'_{ik}\right)^{\mathrm{T}} \mathbf{\Sigma}_i^{-1} (\mathbf{X}'_{ij} - \boldsymbol{\mu}'_{ik})\right)^{-(\nu/2+1)} |\mathbf{J}_{ij}|, \end{aligned} \tag{23}$$

where

$$\mathbf{X}'_{ij} = h(\mathbf{X}_{ij}),$$

and

$$\boldsymbol{\mu}'_{ik} = h(g(\boldsymbol{\mu}_{ik})),$$

with g and h given by eqns. 18 and 20, respectively. Define

$$\mathbf{X}'_{ij} = \begin{bmatrix} X'_a \\ X'_b \end{bmatrix},$$

The Jacobian of the transformation h is:

$$\begin{aligned} |\mathbf{J}_{ij}| &= \begin{vmatrix} \frac{\partial h(X'_a)}{\partial X'_a} & \frac{\partial h(X'_a)}{\partial X'_b} \\ \frac{\partial h(X'_b)}{\partial X'_a} & \frac{\partial h(X'_b)}{\partial X'_b} \end{vmatrix} \\ &= \frac{\partial h(X'_a)}{\partial X'_a} \frac{\partial h(X'_b)}{\partial X'_b} \\ &= \frac{b}{\sqrt{1+(bX'_a)^2}} \frac{b}{\sqrt{1+(bX'_b)^2}}, \end{aligned}$$

where b and σ_h may be calculated from σ_η and σ_ϵ via equations (21) and (22), respectively.

Introducing a sample-level parameter $0 < \rho_h < \frac{3}{4}\sigma_h^2 \in \boldsymbol{\Theta}$, we approximate the covariance-matrix of the transformed data in a given segment i as:

$$\boldsymbol{\Sigma}_i = \begin{bmatrix} \sigma_h^2 & s\rho_h \\ s\rho_h & \sigma_h^2 \end{bmatrix}, \text{ with } s = \begin{cases} \mu_{ik_a}\mu_{ik_b} & \text{if } \mu_{ik_a}\mu_{ik_b} < 1 \\ 1 & \text{otherwise.} \end{cases}$$

Thus, the covariance is spherical for the homozygous clusters. The scaling of ρ_h by s provides a consistent approximation to the covariance implied by eq. (17) for heterozygous clusters.

Substitution of eq. (23) into the generic form for $\mathcal{P}(\mathbf{X}_{ij}|\boldsymbol{\mu}_{ik}, \boldsymbol{\Theta})$ in eq. (3) completes the specification of our HSCR estimation based on SNP microarray data. The sample-level error model parameters which must be estimated from the data are: $\boldsymbol{\Theta} = \{\sigma_\epsilon, \sigma_\eta, \rho_h, \nu, \phi\}$.

Calibration of cross-hybridization effects

We develop a calibration step to remove allelic-crosstalk from copy-ratio data $\mathbf{X}$ in a SNP-specific manner as follows. Starting with large collection of normal samples, genotyped using Birdseed [13], we compute the cluster-centers for each genotype for each SNP. We attempt to estimate the covariance matrix for each SNP, denoted $\boldsymbol{\Sigma}_i$. We make use of the factorization $\boldsymbol{\Sigma} = \mathbf{SRS}$, where

$$\mathbf{R} = \begin{pmatrix} 1 & r_{AB} \\ r_{AB} & 1 \end{pmatrix}, \mathbf{S} = \begin{pmatrix} \sigma_2 & 0 \\ 0 & \sigma_0 \end{pmatrix},$$

with r_{AB} denoting the correlation between the A and B allele. σ_0 and σ_2 denote the standard-deviation of genotype-cluster axes corresponding to 0 and 2 copies, respectively.

Because we are attempting to estimate a large number of parameters from data which may be limited, we make use of a Bayesian regularization technique. Population priors describing the distribution of scales and correlations are estimated from all SNPs, and these are then used to compute the most probable (MP) estimates for each SNP. This follows standard use of a hierarchical

model, as in [58].

We use scaled-inverse-χ^2 prior distributions for σ_0 and σ_2, defined as follows:

$$\text{Let } X \sim \chi^2(\nu) \text{ and } Y = \frac{\sigma^2 \nu}{X}, \text{ then } Y \sim \text{Scale-inv-}\chi^2(\nu, \sigma^2).$$

We use a normal distribution for the Fisher-transform of r_{AB}, $z_{AB} = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right)$. We denote the standard error of z as $z_\sigma = \frac{1}{\sqrt{N-3}}$. We define population-level distributions on the SNP covariance parameters as:

$$z_{AB} \sim \mathcal{N}(\pi_{\mu_z}, \pi_{\sigma_z})$$

$$\sigma_0^2 \sim \text{Scaled-Inverse-}\chi^2(\pi_{s_0}, \pi_{\nu_0})$$

$$\sigma_2^2 \sim \text{Scaled-Inverse-}\chi^2(\pi_{s_2}, \pi_{\nu_2})$$

π_{μ_z} and π_{σ_z} are estimated via maximum likelihood from the calculated SNP correlation-coefficients.

The MP estimate of z is the mode of the posterior (Gaussian) distribution:

$$\hat{z}_{\text{MP}} = \frac{\frac{\pi_{\mu_z}}{\pi_{z_\sigma}^2} + \frac{\hat{z}_{\text{ML}}}{z_\sigma^2}}{\frac{1}{\pi_{z_\sigma}^2} + \frac{1}{z_\sigma^2}}$$

The MP estimate of r is the inverse of the Fisher transform on $\hat{z}_{\text{MP}}$:

$$\hat{r}_{\text{MP}} = \frac{e^{2\hat{z}_{\text{MP}}} - 1}{e^{2\hat{z}_{\text{MP}}} + 1}$$

The posterior distribution for the scales σ_0 and σ_2 is:

$$\sigma_0^2 | X \sim \text{Scaled-Inverse-}\chi^2\left(\theta_\sigma, \theta_\nu\right),$$

with

$$\theta_\sigma^2 = \frac{\pi_{\nu_0}\pi_{s_0}^2 + N\hat{\sigma}_{0\text{ML}}^2}{\pi_{\nu_0} + N},\ \theta_\nu = \pi_{\nu_0} + N$$

The posterior mode is therefore

$$\hat{\sigma}_{0\text{MP}} = \sqrt{\frac{\theta_\nu \theta_\sigma^2}{\theta_\nu + 2}}.$$

$\hat{\sigma}_{2\text{MP}}$ is computed in a similar fashion.

We then construct the MP covariance matrix for each SNP:

$$\hat{\mathbf{\Sigma}}_{\text{MP}} = \begin{pmatrix} \hat{\sigma}_{2\text{MP}} & 0 \\ 0 & \hat{\sigma}_{0\text{MP}} \end{pmatrix} \begin{pmatrix} 1 & \hat{r}_{\text{MP}} \\ \hat{r}_{\text{MP}} & 1 \end{pmatrix} \begin{pmatrix} \hat{\sigma}_{2\text{MP}} & 0 \\ 0 & \hat{\sigma}_{0\text{MP}} \end{pmatrix}.$$

Denote the slope of the 1st eigen-vector of $\hat{\mathbf{\Sigma}}_{\text{MP}}$ as $\hat{M}_{BB}$. We construct an affine transformation as follows:

$$\mathbf{T} = \begin{pmatrix} 1 & -\hat{M}_{BB} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -\hat{M}_{BB} & 1 \end{pmatrix}.$$

The corrected data is then:

$$\mathbf{X}' = \mathbf{TX}.$$

Acknowledgments

We are indebted to Joshua Korn for several helpful conversations.

References

1. Albertson D, Collins C, McCormick F, Gray J (2003) Chromosome aberrations in solid tumors. Nature Genetics 34: 369–376.

2. Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, et al. (2010) The landscape of somatic copy-number alteration across human cancers. Nature 463: 899–905.

3. Carter S (2011) Ph.D Thesis, Chapter 4: Absolute quantification of somatic DNA alterations in cancer reveals frequent genome doublings in human cancers .

4. Zhao X, Li C, Paez J, Chin K, Jänne P, et al. (2004) An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. Cancer Research 64: 3060.

5. Weir B, Zhao X, Meyerson M (2004) Somatic alterations in the human cancer genome. Cancer Cell 6: 433–438.

6. Knudson A (1971) Mutation and cancer: statistical study of retinoblastoma. Proceedings of the National Academy of Sciences 68: 820.

7. Popova T, Manié E, Stoppa-Lyonnet D, Rigaill G, Barillot E, et al. (2009) Genome Alteration Print (GAP): a tool to visualize and mine complex cancer genomic profiles obtained by SNP arrays. Genome Biology 10: R128.

8. Van Loo P, Nordgard S, Lingjærde O, Russnes H, Rye I, et al. (2010) Allele-specific copy number analysis of tumors. Proceedings of the National Academy of Sciences 107: 16910.

9. Yau C, Mouradov D, Jorissen RN, Colella S, Mirza G, et al. (2010) A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data. Genome Biology 11: R92.

10. Storchova Z, Pellman D (2004) From polyploidy to aneuploidy, genome instability and cancer. Nature Reviews Molecular Cell Biology 5: 45–54.

11. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics .

12. Cibulskis K (2011) muTect: Reliable and Accurate Somatic Mutation Detection in Next Generation Cancer Genome Sequencing, manuscript in preparation .

13. Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, et al. (2008) Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nature Genetics 40: 1253–1260.

14. Greenman CD, Bignell G, Butler A, Edkins S, Hinton J, et al. (2010) PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data. Biostatistics 11: 164–175.

15. Rocke DM, Durbin B (2001) A model for measurement error for gene expression arrays. Journal of computational biology : a journal of computational molecular cell biology 8: 557–569.

16. Hekstra D (2003) Absolute mRNA concentrations from sequence-specific calibration of oligonucleotide arrays. Nucleic acids research 31: 1962–1968.

17. Mei R (2000) Genome-wide Detection of Allelic Imbalance Using Human SNPs and High-density DNA Arrays. Genome research 10: 1126–1137.

18. Lindblad-Toh K, Tanenbaum DM, Daly MJ, Winchester E, Lui WO, et al. (2000) Loss-of-heterozygosity analysis of small-cell lung carcinomas using single-nucleotide polymorphism arrays. Nature Biotechnology 18: 1001–1005.

19. LaFramboise T, Weir BA, Zhao X, Beroukhim R, Li C, et al. (2005) Allele-Specific Amplification in Cancer Revealed by SNP Array Analysis. PLoS Computational Biology 1: e65.

20. Beroukhim R, Lin M, Park Y, Hao K, Zhao X, et al. (2006) Inferring Loss-of-Heterozygosity from Unpaired Tumors Using High-Density Oligonucleotide SNP Arrays. PLoS Computational Biology 2: e41.

21. McLendon R, Friedman A, Bigner D, van Meir E (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature .

22. Bignell GR, Greenman CD, Davies H, Butler AP, Edkins S, et al. (2010) Signatures of mutation and selection in the cancer genome. Nature 463: 893–898.

23. Network TCGAR, Network TCGAR, sites Dwg, tissue source, Medicine GscBCo, et al. (2011) Integrated genomic analyses of ovarian carcinoma. Nature 474: 609–615.

24. Staaf J, Lindgren D, Vallon-Christersson J, Isaksson A, Goransson H, et al. (2008) Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome SNP arrays. Genome Biology 9: R136.

25. Attiyeh EF, Diskin SJ, Attiyeh MA, Mosse YP, Hou C, et al. (2008) Genomic copy number determination in cancer cells from single nucleotide polymorphism microarrays based on quantitative genotyping corrected for aneuploidy. Genome research 19: 276–283.

26. Bengtsson H, Neuvial P, Speed TP (2010) TumorBoost: normalization of allele-specific tumor copy numbers from a single pair of tumor-normal genotyping microarrays. BMC bioinformatics 11: 245.

27. Olshen AB, Bengtsson H, Neuvial P, Spellman P, Olshen RA, et al. (2011) Parent-specific copy number in paired tumor-normal studies using circular binary segmentation. Bioinformatics .

28. Chen H, Xing H, Zhang NR (2011) Estimation of Parent Specific DNA Copy Number in Tumors using High-Density Genotyping Arrays. PLoS Computational Biology 7: e1001060.

29. Consortium TIH, investigators P, leaders Pc, group Mw, QC G, et al. (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467: 52–58.

30. Venkatraman ES, Olshen AB (2007) A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics 23: 657–663.

31. Browning BL, Yu Z (2009) Simultaneous Genotype Calling and Haplotype Phasing Improves Genotype Accuracy and Reduces False-Positive Associations for Genome-wide Association Studies. The American Journal of Human Genetics 85: 847–861.

32. Huber W, von Heydebreck A, Sültmann H, Poustka A, Vingron M (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics 18 Suppl 1: S96–104.

33. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, et al. (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nature Biotechnology 27: 182–189.

34. Ozery-Flato M, Shamir R (2009) Sorting Cancer Karyotypes by Elementary Operations. Journal of Computational Biology 16: 1445–1460.

35. LaFramboise T, Dewal N, Wilkins K, Pe'er I, Freedman ML (2010) Allelic Selection of Amplicons in Glioblastoma Revealed by Combining Somatic and Germline Analysis. PLoS Genetics 6: e1001086.

36. Dewal N, Freedman ML, LaFramboise T, Pe'er I (2010) Power to detect selective allelic amplification in genome-wide scans of tumor data. Bioinformatics 26: 518–528.

37. Li Y (2006) Li: Mach 1.0: rapid haplotype reconstruction and... - Google Scholar. Am J Hum Genet.

38. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, et al. (2007) PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics 81: 559–575.

39. Marchini J, Howie B, Myers S, McVean G, Donnelly P (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. Nature Genetics 39: 906–913.

40. Howie B, Donnelly P, Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 5: e1000529.

41. Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. Nature Publishing Group 11: 499–511.

42. Freedman ML, Haiman CA, Patterson N, McDonald GJ, Tandon A, et al. (2006) Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. Proceedings of the National Academy of Sciences of the United States of America 103: 14068–14073.

43. Price A, Tandon A, Patterson N, Barnes K, Rafaels N, et al. (2009) Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genetics 5.

44. Ahmadiyeh N, Pomerantz MM, Grisanzio C, Herman P, Jia L, et al. (2010) 8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with MYC. Proceedings of the National Academy of Sciences 107: 9742–9746.

45. Kilpivaara O, Mukherjee S, Schram AM, Wadleigh M, Mullally A, et al. (2009) A germline JAK2 SNP is associated with predisposition to the development of JAK2V617F-positive myeloproliferative neoplasms. Nature Genetics 41: 455–459.

46. Yatabe Y, Mitsudomi T (2007) Epidermal growth factor receptor mutations in lung cancers. Pathology International 57: 233–244.

47. Fan JB, Surti U, Taillon-Miller P, Hsie L, Kennedy GC, et al. (2002) Paternal Origins of Complete Hydatidiform Moles Proven by Whole Genome Single-Nucleotide Polymorphism Haplotyping. Genomics 79: 58–62.

48. Kukita Y (2005) Genome-wide definitive haplotypes determined using a collection of complete hydatidiform moles. Genome research 15: 1511–1518.

49. Hudson Chairperson TJ, Anderson W, Aretz A, Barker AD, Bell C, et al. (2010) International network of cancer genome projects. Nature 464: 993–998.

50. Grossman SR, Shylakhter I, Karlsson EK, Byrne EH, Morales S, et al. (2010) A Composite of Multiple Signals Distinguishes Causal Variants in Regions of Positive Selection. Science 327: 883–886.

51. Li H, Durbin R (2011) Inference of human population history from individual whole-genome sequences. Nature : 1–5.

52. Fan HC, Wang J, Potanina A, Quake SR (2010) Whole-genome molecular haplotyping of single cells. Nature Biotechnology 29: 51–57.

53. Kitzman JO, MacKenzie AP, Adey A, Hiatt JB, Patwardhan RP, et al. (2010) Haplotype-resolved genome sequencing of a Gujarati Indian individual. Nature Biotechnology 29: 59–63.

54. MacKay DJC (2003) Information theory, inference, and learning algorithms. Cambridge Univ Pr.

55. Chiang D, Getz G, Jaffe D, O'Kelly M, Zhao X, et al. (2009) High-resolution mapping of copy-number alterations with massively parallel sequencing. Nature methods 6: 99.

56. Langmuir I (1916) . Journal of the American Chemical Society 38: 2221–2295.

57. Tibshirani R (1988) Estimating transformations for regression via additivity and variance stabilization. Journal of the american statistical association .

58. Gelman A (2004) Bayesian data analysis. CRC Press.

Figures

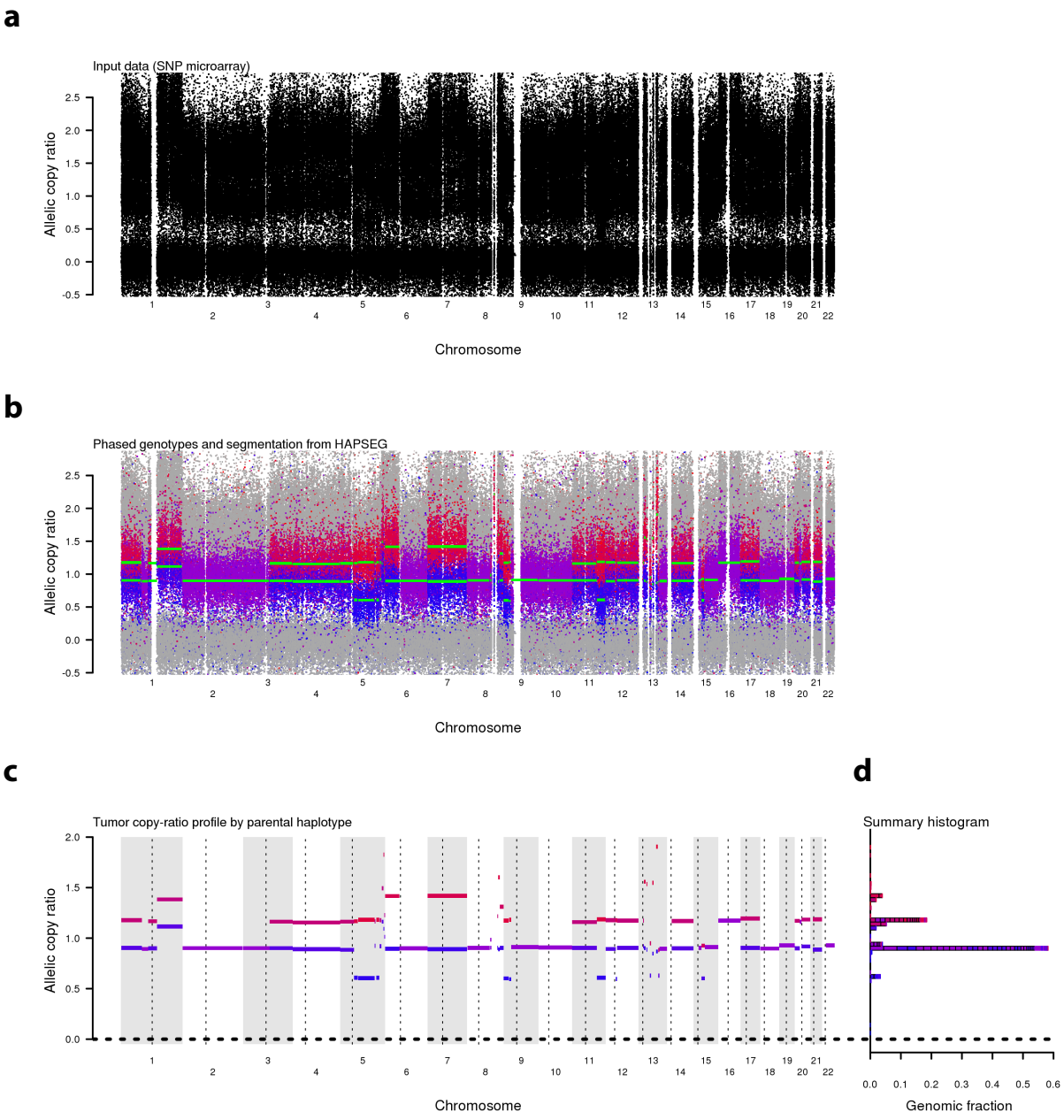


Figure 1

Figure 1. Overview of analysis with HAPSEG a., Input data from an Affymetrix SNP microarray hybridization. The calibrated allelic copy-ratios of A and B channels is shown for each SNP vs. its genomic location. **b**., The data in (a), after processing with HAPSEG. SNPs are colored according to their phased genotype: grey homozygous, red/blue heterozygous (phased), purple heterozygous (unphased). The HSCR segmentation is indicated by green horizontal lines. **c**.,

Summary of the data in (b), where the individual SNPs have been removed and each HSCR segment is colored by the average phasability of the heterozygous SNPs from which it was estimated. **d**., A histogram summarizing the data in (c), by marginalizing over the genome. Although the HSCR locations are independent between segments, only four discrete levels are apparent in the histogram. These corresponding to fixed SCNAs fixed in the tumor sample, and are the basis for analysis with ABSOLUTE [3].

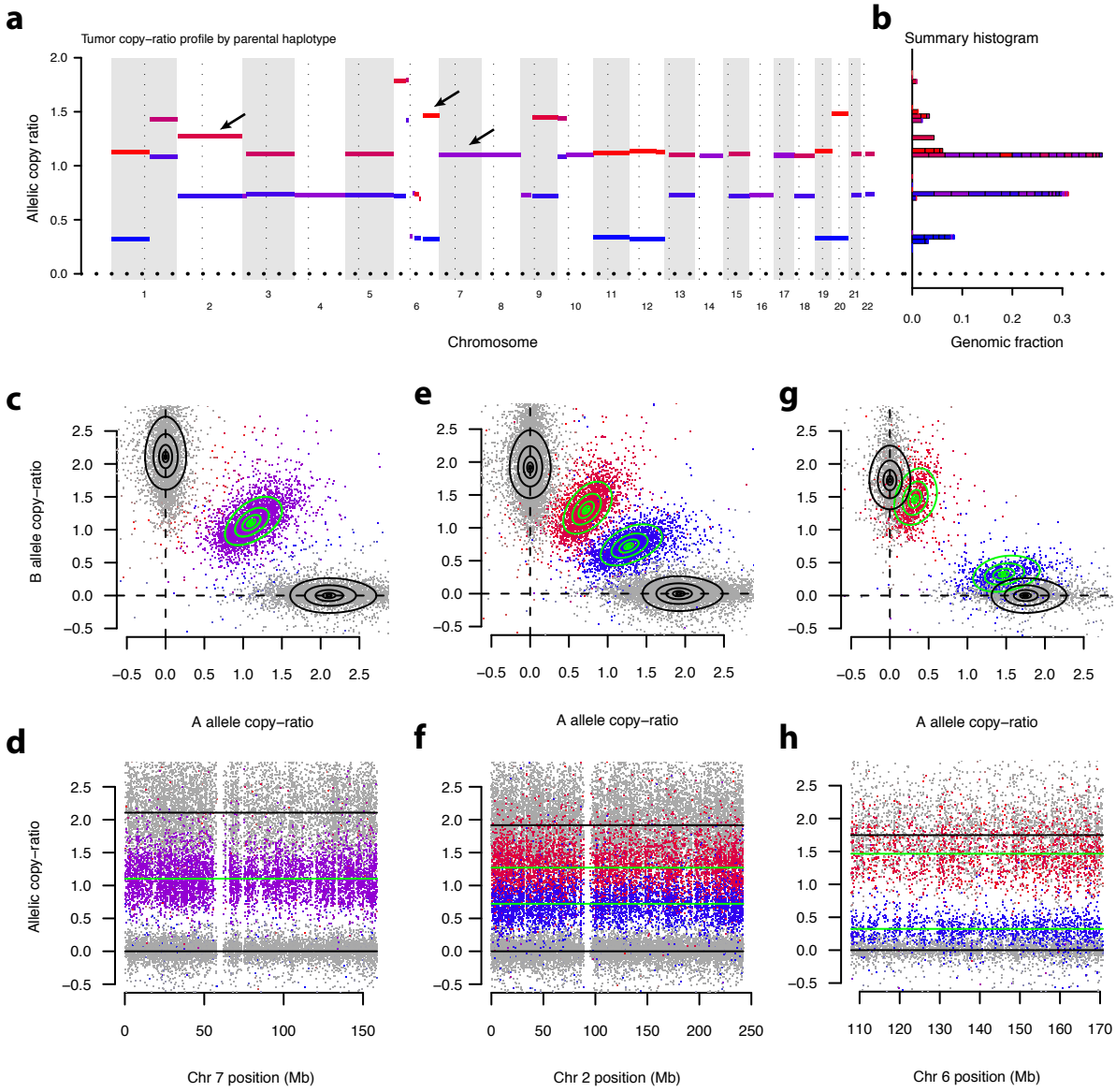
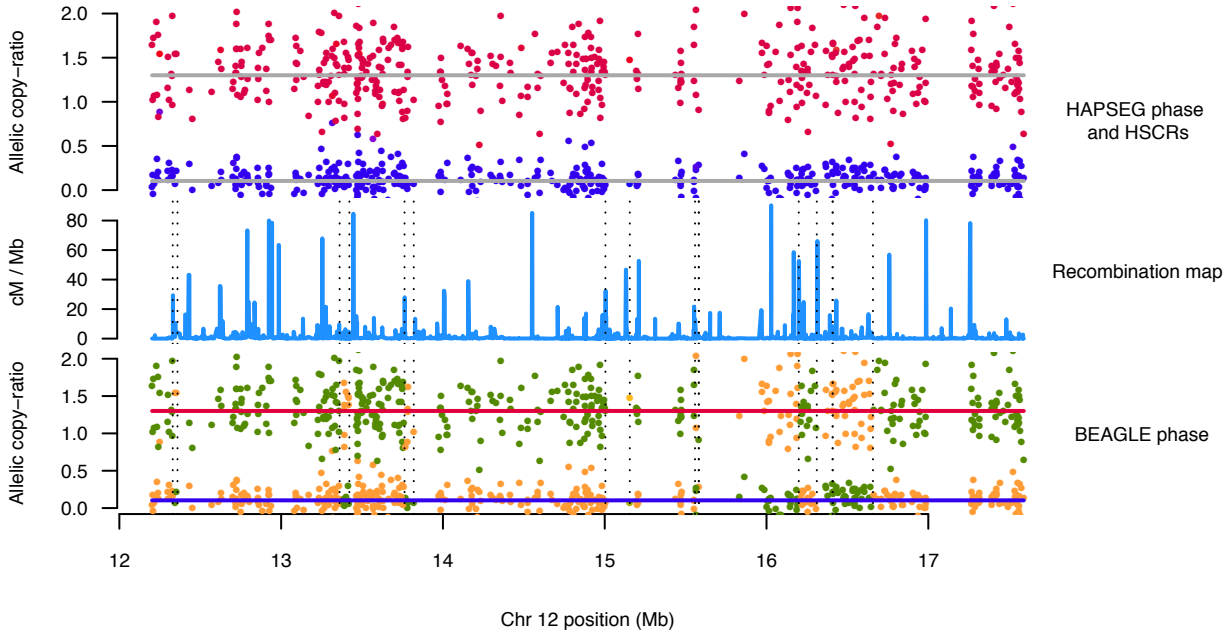


Figure 2

Figure 2. Examples of segmental HSCR inference by HAPSEG **a**., A tumor sample with genome-wide HSCRs as inferred by HAPSEG, as in fig. 1c. **b**., Summary histogram of estimated HSCR values, as in fig. 1d. **c**., **e**., **g**., Plots of calibrated A vs. B-allele copy-ratios for SNPs in three genomic locations indicated in (a) (arrows). SNPs are colored by inferred (haploid) genotype, as in fig. 1b. Contours denote the error-model fit for each genotype cluster: black - homozygous, green - heterozygous. **d**.,**f**.,**h**., Plots of A and B-allele copy-ratios vs. genomic position for the segments in (c,e,g). SNPs are colored as in (c,e,g). Horizontal lines denote genotype cluster locations, colored as in (c,e,g) The heterozygous locations (HSCRs) correspond to the allelic copy-ratios in (a). **c**.,**d**., A segment (chr 7) at allelic balance, with equal copy-numbers of both homologues. No phase information is available for heterozygous SNPs in this segment. **e**.,**f**., A segment (chr 2) at allelic imbalance, with unequal homologous copy-numbers. Note that the SCNA affecting this segment

was predicted to be subclonal by ABSOLUTE [3]. **g**.,**h**., A segment (chr 6) at more extreme allelic imbalance. Note that the lower HSCR here corresponds to LOH in this tumor sample [3]; the DNA contributing to the heterozygous alleles is derived from normal contaminating cells.

a



b

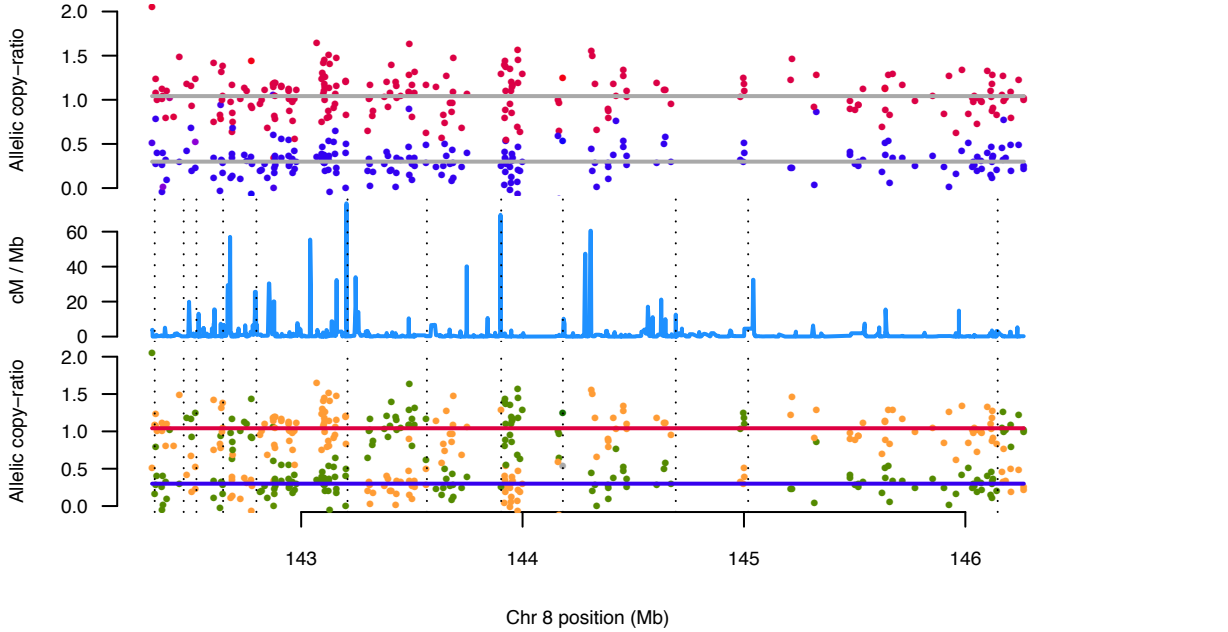


Figure 3. Demonstration of direct haplotyping by homologous imbalance a.,**b**., Comparison of HAPSEG and BEAGLE phase in example segments. Top - Allelic copy-ratios of heterozygous SNPs are shown at their genomic coordinates. Color indicates phase as estimated by HAPSEG ($\hat{\mathbf{C}}^{(2)}$), eq. (9). Grey horizontal lines indicate the HSCR estimates for the segment, estimated by HAPSEG, eq. (8). Middle - The genetic recombination rate is plotted vs. the genome. Bottom - As in (**a**), but replacing the HAPSEG phase estimates by those obtained using the statistical phasing program BEAGLE [31]. Switch errors in the BEAGLE phasing, detected by HAPSEG, are indicated by dotted vertical lines.

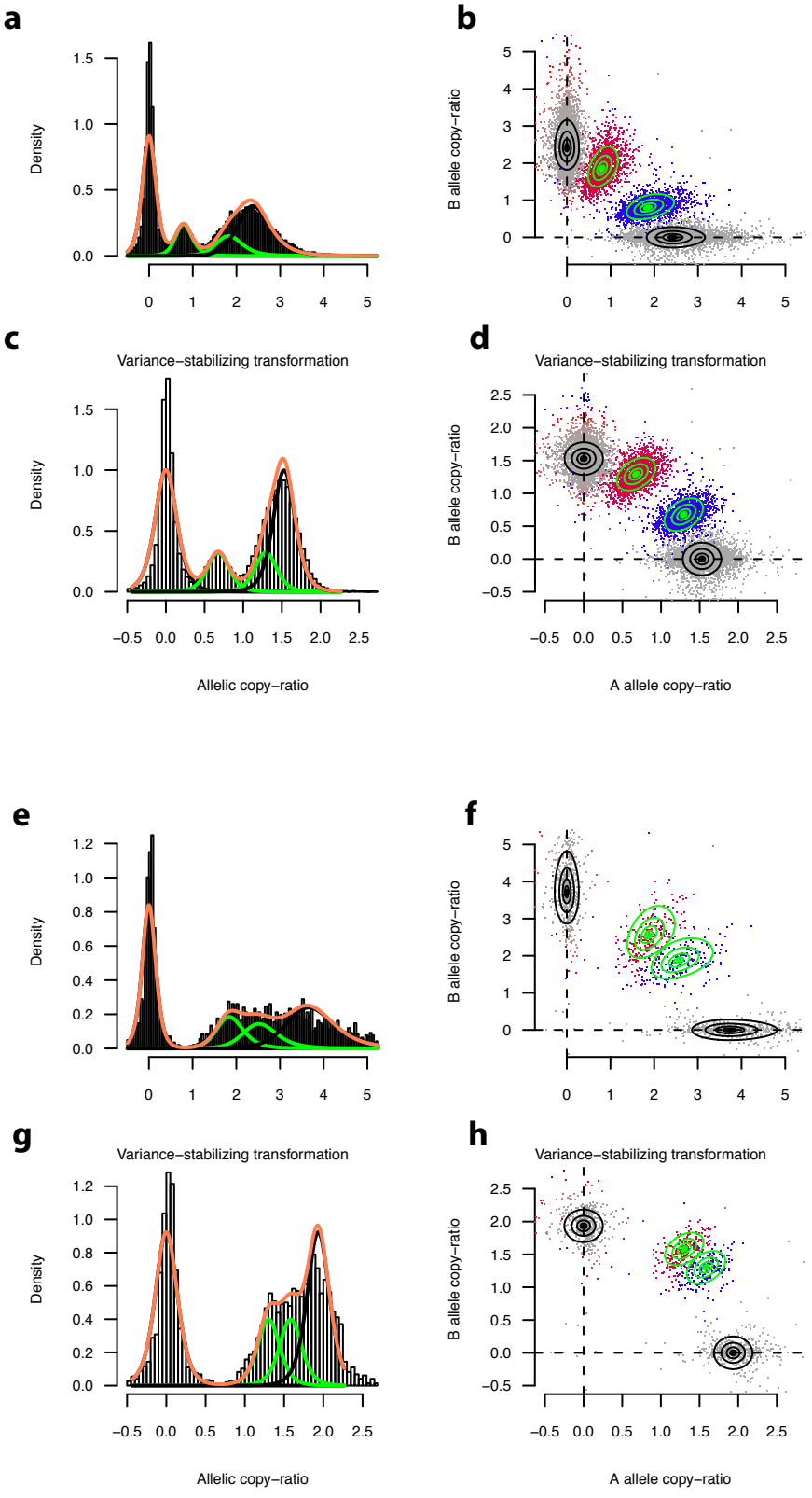


Figure 4. Demonstration of error-model for Affymetrix SNP microarrays. a-d., The fit of the error-model to marker-level data is shown in a single example segment. **a**., The histogram summarizes the A and B channels for assayed markers within the genomic segment. The marginal fit of the A/B marker copy-ratios is denoted by the green (het) and coral (het+hom) curves. **b**., Marker copy-ratios in (**a**) are shown separately for the A vs. B channel. Contours denote fit of the error model to the 4 modeled genotype clusters, eq. (1), (2) (green - het, black - hom). **c,d**., as in **a,b**., but using the variance stabilizing transformation, eq. (20). Likelyhood calculations are performed in this space using a bi-variate t distribution, eq. (23). Positive covariance of the A/B channels is modeled in the heterozygous clusters (green), this aspect of the fit cannot be dispalyed in one dimension, as in (**a,c**). Contours show the fit of this density to the data, as in (**b**), which were derived by inverting the variance stabalizing transformation. We note that the covariance matrix $\mathbf{\Sigma}_i$ in the density used here is fully determined by the segment HSCR locations, conditional on error model parameters $\mathbf{\Theta}$, which are fit at the sample level, eq. (12). Only two degrees of freedom are fit to the data specifically shown here, eq. (2). **e-h**., an additional example segment is shown, as in **a-d**.