Genotyping techniques to address diversity in tumors.

Lindgren, David; Höglund, Mattias; Vallon-Christersson, Johan

Published in:
Advances in Cancer Research

DOI:
10.1016/B978-0-12-387688-1.00006-5

2011

Genotyping techniques to address diversity in tumors

David Lindgren,[a] Mattias Höglund,[b] and Johan Vallon-Christersson[b, c]

[a] Center for Molecular Pathology, Department of Laboratory Medicine, Lund University, SUS Malmö, Malmö, Sweden

[b] Department of Oncology, Clinical Sciences, Lund University, Lund, Sweden

[c] CREATE Health Strategic Center for Translational Cancer Research, Lund University, Lund, Sweden

Correspondence to: David Lindgren, Center for Molecular Pathology, Department of Laboratory Medicine, Lund University, SUS Malmö, Entrance 78, SE-205 02 Malmö, Sweden.

Abstract

Array based genotyping platforms have during recent years been established as a valuable tool for the characterization of genomic alterations in cancer. The analysis of tumor samples, however, presents challenges for data analysis and interpretation. For example, tumor samples are often admixed with nonaberrant cells that define the tumor microenvironment, such as infiltrating lymphocytes and fibroblasts, or vasculature. Furthermore, tumors often comprise subclones harboring divergent aberrations that are acquired subsequent to the tumor-initiating event. The combined analysis of both genotype and copy number status obtained by array based genotyping platforms provide opportunities to address these challenges. In this review, we present the basic principles for current array based genotyping platforms and how they can be used to infer genotype and copy number for acquired genomic alterations. We describe how these techniques can be used to resolve tumor ploidy, normal cell admixture, and subclonality. We also exemplify how genotyping techniques can be applied in tumor studies to elucidate the hierarchy among tumor clones, and thus, provide means to study clonal expansion and tumor evolution.

I. INTRODUCTION

Cancer development and tumor formation involves acquired genomic aberrations, such as sequence mutations and copy number changes. Molecular investigation of genomic alterations in tumors has traditionally been performed using methods such as loss of heterozygosity (LOH) analyses and comparative genomic hybridization (CGH). Conventional GCH, first described by Kallioniemi and coworkers (Kallioniemi *et al.*, 1992), use differentially fluorescently labeled DNA from tumor sample and reference DNA to reveal regions of loss and gain by competitive hybridization to immobilized normal metaphase chromosomes. With the advent of array-technology (Schena *et al.*, 1995), the analysis of cancer genomes advanced rapidly with greatly increased resolution and sensitivity. Array-based comparative genomic hybridization (aCGH) was first performed using gene-centered arrays originally developed for gene expression analysis, or using low-density arrays of large genomic segments cloned in bacterial artificial chromosomes (BACs) (Pollack *et al.*, 1999). Initial techniques were soon further developed for genome-wide investigation of copy number aberrations at high-resolution by tiling BAC arrays and subsequently by employing oligonucleotide probe arrays. In short, aCGH utilizes the same strategy as conventional metaphase CGH but DNA is hybridized to immobilized DNA probes mapped to known genomic locations. Current array platforms, comprising from tens of thousands up to millions of probes, allow for detection of breakpoints and copy number alterations at sub-gene resolution and have been widely used to screen for genomic alterations in cancer (Pinkel and Albertson, 2005). Such analyses have provided a depiction of copy number gain and loss frequencies across large

tumor cohorts in a variety of cancers, highlighting recurrent alterations important during oncogenesis and tumor development (Chin *et al.*, 2006). LOH analyses have, on the other hand, been widely used in cancer research to detect regions of allelic imbalances indicating regions of genomic deletion or copy number neutral LOH, and have been used to identify tumor suppressor genes inactivated by mutation followed by loss of the wild-type allele. Traditionally, LOH analysis use polymorphic markers, such as nucleotide repeat regions or single nucleotide polymorphisms, to detect regions of allelic imbalance.

Whole genome genotyping (WGG) arrays based on Single Nucleotide Polymorphisms (SNPs) (Wang *et al.*, 1998) were developed to analyze blood samples in association studies and have since its introduction successfully been used in numerous studies for identification of genetic susceptibility loci in a variety of diseases (Grant and Hakonarson, 2008). Progression of WGG arrays, or SNP arrays, has followed the identification of SNPs in the human genome derived from initiatives such as the international HapMap Project (http://www.hapmap.org), and platforms currently in use allow for genotyping of millions of SNPs simultaneously. Even though SNP arrays were not originally designed for analysis of tumor samples, it was soon demonstrated that these platforms are suitable for the analysis of cancer genomes (Lindblad-Toh *et al.*, 2000; Wang *et al.*, 2004; LaFramboise *et al.*, 2005; Zhao *et al.*, 2005; Peiffer *et al.*, 2006). Allele specific interrogation of tumor DNA using SNP arrays provides means to investigate the relative abundance of alleles and effectively combine the advantages of LOH analysis and aCGH analysis. Thus, SNP arrays enable researchers to detect copy neutral events in tumors along with copy number

aberrations. SNP arrays have therefore become a valuable tool for analysis of cancer genomes and have been used to provide detailed characterization of a variety of cancers (Wang and Armstrong, 2007; LaFramboise, 2009). The two most commonly used platforms for SNP arrays are Illumina BeadChip (Gunderson *et al.*, 2005; Steemers *et al.*, 2006) and Affymetrix (Wang *et al.*, 1998; Matsuzaki *et al.*, 2004). Experimental design and data analysis is somewhat different, although the main principles are applicable to both platforms as that they provide detection signals from individual alleles separately.

There are several inherent problems with analyses of tumor genomes. For example, a significant proportion of solid tumors are highly aneuploid and subjected to genome duplication events causing deviations from the normal diploid chromosome level (Rajagopalan and Lengauer, 2004). Therefore, a problem of determining the baseline for calling relative genomic copy number alterations becomes apparent. In cytogenetics, the most common chromosome number in a cell population – the modal number – determines if the state of a genomic region is regarded as neutral, gained, or lost relative to a fixed ploidy status. In aCGH, the absolute copy number cannot be resolved and copy number is presented as relative to a reference point, often approximated to the mean- or median copy number, or to the predominant relative copy number (Staaf *et al.*, 2007). As SNP arrays provide assessment of allelic composition in combination with abundance, it opens up for strategies that estimate absolute copy number and ploidy (LaFramboise, 2009). SNP array platforms have also successfully been applied to address problems regarding intermixture of nonaberrant cell populations. As analysis is performed on extracted DNA rather than on

individual cells *in-situ*, measured copy number changes will reflect overall net changes in the cell population from which DNA is extracted. Thus, the amplitude of signal associated with a copy number alteration is dependent on the fraction of cells harboring the alteration. When DNA from grossly dissected tumor biopsies is analyzed, presence of residual tumor-adjacent or infiltrating normal cells will reduce the dynamic range between segments of different copy number. Importantly, with aCGH data it is not straightforward to discriminate between contamination of normal genomes and varying magnitude of underlying net copy number changes, although there have been efforts aimed at resolving this issue (Tolliver *et al.*, 2010). Traditionally, normal-contamination issues have been addressed by excluding samples with low tumor cellularity from analysis or, when feasible, by microdissection of biopsies. However, the tumor microenvironment comprising tumor cells and other cell types such as immune cells, fibroblasts, and endothelial cells is integral in tumor development and progression. The interplay between cells within the tumor microenvironment has been highlighted as important hallmarks of cancer and its composition has been shown to represent an intrinsic property of tumors (Hanahan and Weinberg, 2011). Excluding samples due to cellularity may therefore bias tumor cohort composition. In this respect SNP arrays might offer an advantage as it has been demonstrated that interpretation of the readout of allelic-imbalances can be successfully used to estimate and correct for cellularity or the fraction of cells affected by an alteration (Nancarrow *et al.*, 2007; Assie *et al.*, 2008). The presence of genetic diversity within tumor samples, i.e., tumor subclonality, represents another source of sample heterogeneity that presents challenges

when analyzing tumor genomes. However, the combined analysis of both genotype and copy number status obtained by SNP array analysis provide opportunities to discern subclonal alterations against the background of the predominant clone. Likewise, genotype estimates from SNP array data will also provide increased possibilities to study clonal relationships between repeated tumor samples from the same individual.

Here we aim to provide the basic principles for array based genotyping platforms and the principles of how these techniques can be used to address sample heterogeneity and subclonality in tumors. We first provide a brief description of SNP array platforms, experimental procedures and data extraction. We then proceed to describe the calculation of B allele frequency and relative copy number, and how these values are affected by underlying acquired genetic alterations. We finally discuss how these data can be used and interpreted with the aim of deducing intermixture of nonaberrant cells within tumor biopsies, as well as subclonal events and intra-tumor heterogeneity.

II. THE BASIC PRINCIPLES OF SNP ARRAY PLATFORMS AND SNP ARRAY DATA INTERPRETATION

A. Platforms and Probe Design

There are two SNP array platforms predominantly in use, provided by Affymetrix and Illumina, respectively. Both platforms have been extensively used for genotyping blood samples in genetic linkage studies as well as for the analysis of cancer genomes. The underlying chemistry differs between the

platforms but both can be used to interrogate genotypes in a similar manner and both depend on classic base-pairing and hybridization of target DNA to nucleotide probes of complementary sequences immobilized on a solid surface (Fig 1). The principle relies on that probe intensities reflect the abundance of the respective alleles. There are numerous and detailed accounts of technical aspects of how the platforms work (Matsuzaki *et al.*, 2004; Peiffer *et al.*, 2006; Steemers *et al.*, 2006) and several comparisons of how they perform (Hehir-Kwa *et al.*, 2007; Baumbusch *et al.*, 2008; Gunnarsson *et al.*, 2008; Curtis *et al.*, 2009). Here, we will confine to describe the basic principles of the platforms and highlight some of the differences between them.

Since the first SNP array platforms were presented (Wang *et al.*, 1998), array density has increased by several orders of magnitude and the current platforms comprise millions of probes in a single assay. The earlier versions of Affymetrix SNP arrays utilized comprehensive collections of probes to interrogate each individual SNP, using paired perfect-match and miss-match probes to infer genotype, a strategy adopted from Affymetrix expression array platforms. More recent versions utilize a less redundant strategy of a limited number of probes per SNP and allele. In short, 25-mer oligonucleotide probes designed to match the target DNA of interest are in-situ synthesized on the array surface. For interrogation of SNPs, specific probes are synthesized for each of the two alleles at separate locations on the array (Fig. 1B). The overall strategy depends on preferential hybridization of perfect complementary target sequence to the probes coupled with the ability to quantify the amount of bound target. By quantification of hybridized targets to separate-allele specific probes their

individual abundance in the sample may be inferred. The Illumina SNP array platforms are more recent and have not undergone the same degree of reformation compared with Affymetrix. It shares the same basic principle of target hybridization to loci specific probes with Affymetrix, and similarly, probe density has increased to comprise millions of markers. However, there are some fundamental differences between the platforms. Illumina utilizes their BeadChip technology that permits probes to be immobilized on silica beads rather than directly onto the array surface. Probe-covered beads are then randomly distributed in microwells covering the array surface, followed by a probe location decoding procedure. Each SNP is interrogated with a single bead type covered by one unique 50-mer probe designed to target the sequence adjacent to the SNP of interest. After target hybridization, alleles are differentiated by a subsequent enzymatic single-base extension of the probe using the hybridized target as template. Base extension results in the incorporation of differentially labeled nucleotides depending on the captured allele. The abundance of the respective SNP alleles in the sample may then be inferred by dual color quantification of signal intensities from each bead type (Fig 1C). It should also be noted that current WGG platforms contain large numbers of probes that are not designed to interrogate SNPs. Rather these probes are solely designed to assay copy number and as such serve as aCGH probes and some are specifically designed to target copy number polymorphic regions, i.e., constitutional copy number variations (CNVs).

B. Principles of Data Extraction and Normalization

Raw data acquisition and processing varies depending on array platform. Arrays are hybridized and labeled according to chemistry-dependent experimental procedures followed by imaging and data extraction. Raw signal measurements for the *A* and *B* alleles are preprocessed, normalized, and summarized over probe replicates or a collection of probes depending on platform. Preprocessing and normalization of probe data is performed to achieve pairs of allele-specific measurements for each SNP locus, and to this end there are various methods described (LaFramboise, 2009). Pairs of normalized allele measurements can subsequently be used to call genotypes and to infer DNA abundance and allele ratio. For calling genotype and calculating allele ratio, observed normalized intensities are related to expected values derived from collections of reference data. Transformation of intensities to relative copy number estimates is essentially also performed by relating values to a collection of normal reference samples (HapMap) or to a matched control.

C. The B Allele Frequency and Relative Copy Number

The B allele frequency (BAF), first presented using Illumina data (Peiffer *et al.*, 2006), is calculated for each SNP individually by transformation of allele intensities and represents the proportion of DNA content for allele *B* as compared to the total DNA content of *A* and *B* alleles together. The proposed transformation involves linear interpolation of allele frequencies from reference data derived from normal samples. Since BAF simply describes the total number

of *B* allele copies divided by the total number of allele copies for that specific locus, a theoretical BAF can be calculated for any given genotype using the following equation:

(1): $BAF = N_B/(N_A+N_B)$

In Eq. (1), N_B is the total number of *B* allele copies and N_A is the total number of *A* allele copies.

Apart from genotyping, SNP arrays provide means for quantification of relative copy numbers at each given loci and current SNP arrays typically contain large numbers of probes that are solely designed to assay copy number and target non-sequence polymorphic loci. These probes can be used for the analysis of CNVs but many are also added to provide increased power and resolution when analyzing acquired copy number aberrations in tumors. Relative copy number ratio values are calculated by comparing observed normalized intensities (sum of *A* and *B*) to the expected, similarly to how BAF is derived, and is typically presented as Log2 relative ratio (LRR). Data from Affymetrix can be converted into BAF and LRR by appropriate normalization and transformation (Wang *et al.*, 2007; Sun *et al.*, 2009). Examples of expected BAF and LRR values for a normal genome and how these values are affected by acquired genetic aberrations is further discussed below.

D. Expected BAF and LRR for a Normal Genome

In a diploid genome, there are only three possible allele combinations for a given locus: homozygosity for the A allele (*AA*), heterozygosity (*AB*) or

homozygosity for the B allele (*BB*). Thus, given Eq. (1), three different BAF values are possible for SNP loci in a normal diploid genome: 0 (*AA*), 0.5 (*AB*), or 1 (*BB*). The genotype status across chromosomes may conveniently be visualized using BAF-plots (Fig. 2). In these plots the BAF values for individual SNPs (y-axis) are plotted with respect to their genomic position (x-axis), similar to a copy number profile. A schematic BAF plot representation of a normal diploid genome is presented in Fig. 2A. As seen in Fig. 2A, any given SNP locus will only have one unique BAF value and this value is defined by its corresponding genotype. The design of current SNP arrays are in practice arbitrary with respect to *A* and *B* alleles, i.e., when considering a large consecutive series of SNP loci, *AA*, *AB*, and *BB* genotypes will ideally appear randomly distributed. Therefore BAF values plotted across a chromosome will give the impression of being "banded" at the macro-level. Fig. 2B displays an experimentally obtained BAF-plot of a chromosome from the analysis of a normal diploid genome. Three seemingly horizontal bands representing *AA*, *AB*, and *BB* genotypes are apparent, closely clustered around the theoretical BAF values of 0, 0.5, and 1, respectively (Fig. 2B). In reality the series of consecutive BAF values across the genome shift continuously between the three states as extensive homozygous genomic segments are normally not observed. The BAF profile of a homozygous genome, e.g., a haploid genome, will consequently present only 2 bands, restricted to theoretical BAF values 0 and 1, whereas a triploid genome will show four bands. The appearance of more than four bands is inevitably the result of mixed samples, e.g., inadvertently mixing DNA from two individuals. However, such chimeric patterns may be observed in clinical samples, for example, when

analyzing recurring leukemias after the patient has undergone bone marrow transplantation (Paulsson *et al.*, 2011). We will in section III discuss BAF values in further detail, and how these values may be used for the identification of regions of genomic alteration.

In section II.C we described how SNP arrays estimate copy numbers for each SNP locus. By definition, a normal diploid genome has two copies of each autosome. Copy numbers are often presented as relative ratios, which are log2 transformed and centered. Therefore the copy number profile of a normal genome is centered on 0, corresponding to 2 copies (Fig. 2B). However, it is worth to mention that constitutional CNVs are quite common (Iafrate *et al.*, 2004; Sebat *et al.*, 2004). Therefore, care should be taken when analyzing tumor samples – to avoid misinterpreting small copy number gains or losses as acquired somatic alterations – especially if matched constitutional blood is unavailable for comparison (Heinrichs *et al.*, 2010).

III. WHOLE GENOME GENOTYPING OF TUMOR SAMPLES

Since the introduction of SNP arrays, a large number of studies have proved these platforms to be important means of analysis of acquired genomic changes. Since SNP arrays can detect chromosomal imbalances at both the copy number level, measured as deviation of LRR, and at the genotype level, measured as deviations of BAF, the combined use of these two measurements can be used for interpretation of underlying genomic imbalances. We will here discuss the basic concept of how copy number, and allelic ratios are affected by common genetic

alterations such as deletions, copy number gains, and copy number neutral events.

A. Changes in BAF and LLR upon Acquired Genomic Alterations

As described above, there are three possible genotypes for a given SNP locus in the normal diploid genome, either heterozygous (*AB*) or homozygous (*AA* or *BB*). Using Eq. (1) for calculating BAF, we also described that these genotypes have BAF values of 0.5 (*AB*), 0 (*AA*), and 1 (*BB*), respectively. In tumors, however, deviations from the normal diploid state is frequently observed, for example, gains of chromosomal regions harboring oncogenes or deletions of regions that harbor tumor suppressor genes. As a consequence, not only the DNA copy number will be affected, but also the balance between *A* and *B* alleles for SNPs that were constitutionally heterozygous (*AB*) within the altered region. Similarly to the normal state, the BAF formula given in Eq. (1) can be used to calculate theoretical BAF, i.e., a representation of the proportion of B alleles to the total number of allele copies. In Table 1 we list a number of possible genotypes and their corresponding BAF values. For example, a region present in three copies can have four possible genotype combinations: *AAA*, *AAB*, *ABB*, and *BBB*, which will have theoretical BAFs of 0/3=0, 1/3=0.33, 2/3=0.67, and 3/3=1, respectively. Similarly, a SNP locus with an *ABBB* genotype will have a BAF value of 0.8. Thus, the simultaneous readout of both BAF and copy number by SNP arrays provides a unique opportunity to extrapolate the actual genotype status of an altered region within a tumor genome. We will in the next paragraphs present a couple of genomic alterations and their effect on BAF and LRR.

With the simple principle of BAF in mind, let us hypothesize a scenario in which a somatic hemizygous deletion has occurred within a diploid tumor genome leading to LOH in the affected region, i.e., the possible genotypes are restricted to either *A* or *B* (Fig. 3A). Thus, BAF values for all germline heterozygous SNPs are shifted from BAF=0.5 to either BAF=0, or BAF=1, depending on which chromosomal homologue that has been lost. The plot will therefore display two horizontal bands of BAF values at 0 and 1. It is important to note that only SNPs heterozygous in the germ line will change their respective BAF value in case of a deletion; constitutively homozygous SNP loci are by definition non-informative for studying acquired allelic imbalances at the genotype level. At the copy number level, however, all SNPs are informative and the deletion will be detected as reduction in LRR for all measured probes within the region (Fig. 3A). In theory, the DNA content for a deletion is reduced to half of that of the normal state, and the theoretical LRR value for affected SNPs would therefore be -1 in Log2 space. However, due to platform limitations the experimentally obtained response on LRR values is typically smaller than the theoretical (Peiffer *et al.*, 2006).

Similarly, a single copy number gain will not only be reflected at the LRR level, but also introduce a shift in BAF at all SNP loci that were germline heterozygous. SNPs within a region affected by a one copy gain caused by duplication of material from one of the homologues will have four possible allele combinations (*AAA, AAB*, *ABB*, and *BBB*), resulting in a four-banded pattern in a BAF plot (Fig. 3A). A region present in three copies may also arise through a two copy gain of material from one homologue in combination with deletion of

material from the second homologue. This will lead to complete homozygosity within the region (only *AAA* or *BBB* genotypes will be present), and the BAF pattern will be indistinguishable from that of LOH caused by deletions. The increase in LRR will however indicate that this region is present in more than two copies.

For more complex alterations involving higher allele copy numbers, multiple paired genotype combinations are possible within the gained region, again depending on which homologues are present and in what proportions. Fig. 3A present two possible scenarios of how a two copy gain can be manifested. In the first example, two imbalanced genotypes are possible for SNPs that were germ line heterozygous (*AAAB* and *ABBB*, BAF=0.25 and BAF=0.75, respectively). Alternatively, a net gain of two copies may arise through duplication of material from both homologues. Germ line heterozygous SNPs will here remain balanced (*AABB*, BAF=0.5) and no shift will be observed at the BAF-level. Similar to the three copy example above, all haplotypes in the four copy region may also be derived from the same homologue leading to complete homozygosity of the segment (*AAAA* and *BBBB*).

Tumors may also display regions of allelic imbalance but without changes in copy number, a state often referred to as copy number neutral imbalance. The combination of genotype and copy number measurements makes SNP arrays ideal for the identification of copy number neutral imbalances. In contrast, such aberrations are undetectable using aCGH. A simple example of a copy number neutral imbalance is when a chromosomal region is deleted and followed by duplication of the remaining allele (Fig. 3A). It must be stressed that definition of

copy number neutral alterations are intimately linked to the ploidy state of the tumor. The ploidy-status of tumors and its implication on BAF and LRR will however be discussed more in detail below. Copy number neutral imbalance is sometimes referred to as uniparental disomy (UPD), which is the terminology used to describe when an individual is constitutionally homozygous for a chromosomal region since both alleles are derived from a single parent. UPDs are observed as the cause of certain recessive genetic disorders and arise through meiotic segregation errors, chromosomal duplications, or mitotic recombination events during early development. Due to its narrow definition – homozygosity caused by two copies from the same parent – and close association with constitutional genetics, we will refrain from using the term UPD when discussing copy number neutral allelic imbalance events.

B. The Mirrored B Allele Frequency (mBAF)

In the examples above we demonstrated how different types of acquired chromosomal alterations influence the BAFs of constitutionally heterozygous SNP loci. A consecutive series of SNP alleles (a haplotype series) on a chromosome homologue is in practice random with respect to its sequence of *As* and *Bs*. If we consider a region affected by a specific genetic alteration we also note that BAF values for the SNPs within this region are symmetrically positioned around the 0.5 axis. A reflection of BAF data along the 0.5 axes can therefore be applied to obtain mirrored BAF (mBAF) values (Assie *et al.*, 2008; Staaf *et al.*, 2008). In Fig. 3B we demonstrate this inherent symmetry for the regions of copy number and/or allelic imbalance presented in Fig. 3A. If non-

informative germ line homozygous SNPs are removed, an mBAF plot will display only one horizontal band reflecting the proportion between the major and minor allele for that specific region. Thus, the use of mBAF will provide similar information as for BAF but requires fewer genotype combinations to describe the genomic state, i.e., the paired genotype combination for the one copy gain presented in Fig. 3A (*AAB*, and *ABB*) have BAF=0.33 and BAF=0.67, but mBAF=0.67: corresponding to the BAF for the genotype that is dominated by B alleles (in this case *ABB*). As exemplified below, mBAF can facilitate identification of segments of allelic imbalance.

C. Delineating Regions of Genomic Imbalance

A number of computational methods have been described for the automated identification of altered regions in tumor genomes analyzed by SNP arrays. As for conventional LOH analysis, at the level of individual SNPs, a matched blood sample is needed as a reference to determine if that specific SNP is subjected to an acquired alteration or not. However, even in case of a matched normal genotype, individual SNPs are generally not sufficient for determining the genotype at a given loci due to possible technical noise. Therefore, one must make use of larger regions of consecutive SNPs to accurately predict genomic imbalances. We previously described that, when considering a larger series of SNPs, a BAF plot will appear as banded and that three bands are seen when analyzing a normal diploid genome. Through the schematic examples of genomic alterations described above (Fig. 3), we also demonstrate that most somatic alterations will introduce shifts in the BAF profile and that these shifts are a

consequence of the particular haplotype combination that constitute that specific alteration. The high resolution of SNP arrays permits inference of allelic imbalance from a continuous stretch of LOH without the need of a matched normal genotype. Initially this possibility was demonstrated using Hidden Markov Model algorithms to infer regions of allelic imbalance (Lin *et al.*, 2004; Beroukhim *et al.*, 2006) but several and more elaborate approaches are currently available for the definition of such genomic segments (Staaf *et al.*, 2008; Li *et al.*, 2011). It should again be stressed that relatively long stretches of homozygosity may be constitutionally present, and therefore care must always be taken when inferring regions of LOH in the absence of a matched normal sample (McQuillan *et al.*, 2008; Heinrichs *et al.*, 2010).

Use of segmentation algorithms, e.g., CBS (Venkatraman and Olshen, 2007), to identify breakpoints delineating regions with a joint underlying genomic state was early adopted for aCGH data and has been repeatedly evaluated (Lai *et al.*, 2005). Segmentation-based approaches can be applied with the overall aim to describe the studied genome as a series of segments ascribed specific BAF and LLR states. Thus, any segment corresponding to a genomic event or alteration is represented by a LRR and BAF that deviate from the normal state: either by imbalanced LRR, BAF, or both. Note that when ascribing a BAF value to a segment, SNPs that are homozygous in the germline are uninformative and are disregarded.

Fig. 4 displays typical BAF and mBAF patterns obtained from a SNP array analysis of a tumor and illustrate how data can be segmented in order to reduce data dimensionality. Thus, instead of describing single SNP loci, we can rather

refer to the alleles and genotype of whole segments, i.e., we refer to haplotypes and genotypes as the collective haplotype/genotype state of one genomic region. A change in the haplotype constitution will result in a change of the banded pattern. It then becomes intuitive that most acquired alterations will introduce a shift in BAF and/LRR, and that changing from one underlying state to another will involve breakpoints in the data delineating genomic alterations (Fig. 4).

D. BAF vs LRR Plots

We have shown that SNP array data provide both genotype and copy number estimates for each SNP that is queried, and that these can be visually represented using mBAF and LRR profile plots. To interpret a specific genetic alteration it is needed to take both mBAF and LRR into account, and their respective relationship can be queried by plotting LRR versus mBAF (Fig. 5). Although values from individual SNPs can be plotted, various segmentation approaches can effectively reduce the complexity of data, i.e., defining regions of genomic balance or imbalance and treating these as individual events assigned representative mBAF and LRR values. When plotting segmented LRR versus mBAF (or BAF) from a tumor with a diploid chromosomal number a characteristic pattern will emerge where genomic regions (segments) with identical allele combinations (genotypes) will appear close to each other within the mBAF/LRR space (Fig. 5).

For example, segments of one copy gain (*BBA*) will appear together as a cluster of values with elevated LRR and mBAF, approaching their theoretical values of mBAF=0.67 and LRR=0.58. Correspondingly, regions representing copy

number loss (*B*) will cluster around their theoretical values of mBAF=1 and LRR=(-1), whereas copy number neutral LOH (*BB* genotypes) will in this example be positioned at the same mBAF as losses (mBAF=1) but at LRR=0 (Fig. 5). All unaltered segments (*AB*) will form a dense cluster at mBAF=0.5 and LRR=0. Hence, each individual tumor will demonstrate a characteristic mBAF/LRR pattern depending on what specific alterations (genotype combinations) have been acquired. Popova and co-authors (Popova *et al.*, 2009) termed this BAF/LRR pattern as the "Genomic Alteration Print" (GAP) of a tumor. Pattern-recognition strategies have been applied on similar representations of SNP array data to resolve tumor cellularity, underlying ploidy of the tumor, as well as intra-tumor heterogeneity (Attiyeh *et al.*, 2009; Popova *et al.*, 2009).

IV. WGG ANALYSES OF COMPLEX AND HETEROGENEOUS CELL POPULATIONS

We have so far discussed relatively simple examples of alterations affecting one homogenous population of tumor cells. In practice however, WGG analyses are often performed on heterogeneous tumor samples that contain more than one distinct population of cells. For example, primary tumor samples are often admixed with cells without somatic alterations. These nonaberrant cells can include cells that define the tumor microenvironment, such as infiltrating lymphocytes and fibroblasts, or vasculature. Normal cells can also be present due to sampling procedures leading to inclusion of varying amounts of tumor-adjacent non-neoplastic tissue. Thus, the proportion of nonaberrant cells will vary from sample to sample. Regardless of the cause and nature of included

nonaberrant cells, the presence of normal diploid cells within a tumor sample can cause problems in downstream analyses and subsequent interpretations of the data. Moreover, cancers may to varying degrees be composed of multiple clones harboring divergent aberrations that are acquired subsequent to the tumor-initiating event. Collectively, the presence of heterogeneity in tumor samples imposes challenges on data analysis and interpretation. By providing the combination of genotype and copy number information, SNP array data can, however, be used for resolving some of this complexity and thereby increase our possibilities to study the mechanisms and actions underlying cancer.

A. Tumor Ploidy

Tumor genomes are often highly aneuploid and may reach near-triploid, tetraploid or even higher ploidy states. Such deviations from the normal diploid state will have direct implications on both BAF and LRR and the underlying aneuploidy has to be taken into account when making assumptions of the genotype status of altered regions. In our examples so far we have only dealt with simple chromosomal alterations on a diploid background. If we instead consider a tumor with a near triploid genome we will expect most chromosomes to be in allelic imbalance. Chromosomal regions without copy number alterations relative to the modal chromosome number can for example have two copies of one allele and one copy of the other allele: resulting in a characteristic 4-banded pattern in the BAF profile (Fig. 6A). Deletions in a triploid *ABB* background will thus be seen as a shift towards either homozygosity, i.e., two

identical alleles are retained (*BB*) (Fig. 6A), or towards allelic balance, i.e., one copy of each allele is retained (*AB*).

A tetraploid tumor, on the other hand, will typically display a BAF profile where most chromosomes have a balanced genotype (*AABB*) (Fig. 6B). Such a scenario could entail tetraploidization from a diploid genome through incomplete cytokinesis, endoreduplication or cell fusion. In theory, BAF cannot be used to discriminate strictly tetraploid cells from diploid cells after duplication of the genome. In reality though, the complexity of the tumor genotype is typically such that tetraploid clones can be discerned since additional alterations are acquired after tetraploidization. A one copy deletion on a tetraploid background is shown in Fig. 6B.

Deviations from the diploid state will also give rise to highly characteristic patterns in the mBAF/LRR plot. This is exemplified in Fig. 7A by a near triploid tumor karyotype and its corresponding mBAF/LRR plot. Note that the majority of chromosomal segments cluster at LRR=0 with mBAF=0.67, indicative of an imbalanced *BBA* genotype. Segments representing deletions, and thus ascribed negative LRR, are seen at either mBAF=0.5 or mBAF=1, depending on which homologue that is lost (AB or BB genotype, respectively). In Fig. 7B, an example of a near tetraploid karyotype and its corresponding mBAF/LRR plot is given. For this karyotype segments without relative copy-number alterations are located at mBAF=0.5 (*AABB*). In contrast to the mBAF/LRR of diploid tumors, a tetraploid background will allow for a variety of possible genotypes for regions subjected to deletions, e.g., *BBA*, *BB*, and *B*.

B. BAF and LRR in an Admixture of Tumor and Normal Cells

The above theoretical examples have focused on situations when there is only one clone present within the sample, i.e., all analyzed cells have identical genotypes. Given that the LRR copy number reflects the DNA content, increasing proportions of cells with a normal karyotype will cause the LRR for a genomic alteration to converge towards that of the normal cells. In much the same way, the BAF patterns of a tumor will be affected by the presence of nonaberrant cells within the sample. However, whereas the change in LRR for imbalanced regions is linearly proportional to the fraction of present diploid cells, the effect on expected BAF is not always linear. Instead the effect on BAF for a given alteration will depend on its' specific genotype. A simple example to illustrate how BAF for a genomic alteration is influenced by the presence of normal cells is given in Fig. 8A. Here, a schematic representation of a sample with eight tumor and two normal cells is displayed, i.e., 20% normal cells are present within the sample. Let us hypothesize that the tumor cells carry a one copy deletion leading to LOH in the affected region. Since BAF simply describes the frequency of B alleles in a given region, Eq. (1) can be used to calculate expected BAF to 0.83 (Fig. 8A). Fig. 8B displays expected BAF and LRR plots for a hemizygous deletion in a sample with 20% normal cell admixture.

Equation (1) can with some minor modifications be used to calculate BAF values for any given locus in case of heterogeneous samples. A general formula

to describe the relationship between BAF and fraction of normal cells is given by the following equation:

(2): $BAF=(x+ N_B (1-x))/(2x+ N_A (1-x) + N_B (1-x))$

In Eq. (2), x is the fraction of cells with a normal karyotype, and N_A and N_B denotes the number of *A* and *B* allele copies for the specific aberration genotype. As previously stated, BAF is not necessarily linearly affected by the proportion of normal cells. In Fig. 8C we use Eq. (2) to plot theoretical mBAF for a number of different chromosomal states as defined by their genotypes and show how these vary with increasing normal cell admixture. For example, mBAF for an aberration with genotype *B*, corresponding to a hemizygous deletion in a diploid tumor, will shift from 1 to 0.56 with increasing fraction of normal cells from 0% to 80%. As shown in Fig. 8C the relationship between mBAF and normal cell admixture is not linear for genotype *B.* Given Eq. (2), a linear relationship is, however, seen for genotype *BB*. The behavior of LRR and BAF in response to normal contamination has been extensively described (Nancarrow *et al.*, 2007), and experimentally corroborated using serial dilution experiments of tumor cell lines and matched normal blood (Assie *et al.*, 2008; Staaf *et al.*, 2008; Van Loo *et al.*, 2010). As a consequence, it is possible to use experimental array SNP data to infer the fraction of nonaberrant cells present within a sample. Several studies have successfully demonstrated this using tumor biopsies by comparing BAF derived estimates with cellularity scores from histological examination (Nancarrow *et al.*, 2007; Assie *et al.*, 2008; Sun *et al.*, 2009). It follows that BAF of an altered region is in fact reflecting the fraction of cells harboring the alteration

and not only the fraction of normal cells. Thus, any deviation from the expected value can be caused by clonal heterogeneity rather that normal contamination. The principles of estimating the fraction of normal cells can be illustrated using a simple example (Fig. 8D). The figure illustrates an expected mBAF/LRR pattern for the example given in Fig. 5, but in this case on a background of 20% normal cells. As demonstrated using well-characterized CLL samples (Staaf *et al.*, 2008), once the cellularity of a sample is resolved it is possible to also estimate the fraction of tumor cells carrying individual alterations. However, the combination of normal contamination and increased clonal heterogeneity can rapidly increase the complexity of the data and thereby reduce the possibility to resolve underlying genotype status.

C. Tumor Subclonality

The presence of genetic variation between different subclones within a tumor mass is a well-known phenomenon. Even though tumor cells generally are clonally related and show identical alterations at some loci, subclonal differences are often observed. Subclonal genetic alterations may readily be identified at the individual cell level by conventional cytogenetics or fluorescence in situ hybridization. Current molecular analyses of bulk samples will however only give an average estimate of all imbalances. For SNP arrays, the effect on BAF and LRR of subclonal alterations will in practice follow the same line of reasoning as discussed in the examples above about nonaberrant cell involvement. If we further expand our example of a sample of 80% tumor cells and 20% normal diploid cells (Figs. 5 and 8D) and hypothesize that 50% of the tumor cells carry

some additional alterations, we can simply calculate expected mBAF for these using Eq. (2). A deletion (*B*) present in 40% of the cells will display mBAF of 0.625. If we instead consider a late copy number gain (*BBA*), mBAF for the altered region will be 0.583. Subclonal events will in this respect behave as regular alterations occurring in a proportion the cells and BAF will be affected as if normal contaminated. When plotting segmented LRR versus mBAF values, subclones are readily discernable as segment values deviating from the expected pattern set by the percentage of normal contamination (Fig. 9, arrows). Subclonal alterations in a region already affected by an earlier alteration will be much harder to detect. Although it is possible to propose plausible models to explain any observed pattern, subsequent validation is necessary to definitively resolve the underlying states. For instance, it is not possible to distinguish between a case including two clones present in a 50:50 relationship that do not share alterations from a homogenous tumor population harboring the union of these alterations but with 50% normal admixture. Likewise, cell populations comprising highly rearranged genomes and mixed ploidy will add complexity beyond examples presented here. Nonetheless, for regions that do not follow expected patterns one can at least assume the presence of subclonal events. Thus, SNP arrays may provide means for the detection of subclonal events and also propose a likely genotype that will explain the observed pattern.

D. Tracing Clonal Relationships Using SNP Arrays

Depiction of copy number gain and loss frequencies across large tumor cohorts highlight recurrent alterations and can be used to classify tumors into

groups with related karyotypes (Russnes *et al.*, 2010). Though providing clues to genetic events important for tumor initiation, progression, and metastasis, studies on non-related individuals will never be able to "recapitulate" clonal evolution and expansion events per se. To be able to discern and model the underlying chronology of events, repeated samples from the same individual has to be studied. Unfortunately such studies are quite rare, most probably since availability of multiple tumor specimens from single individuals is scarce. Certain cancers are, however, more permissive in studying clonal evolution. One example includes urothelial carcinomas where sampling from multifocal and highly recurrent tumors through non-invasive cystoscopies is possible (Höglund, 2007). The limited availability of multiple samples from individual patients can be circumvented by macro or micro dissection (Navin *et al.*, 2010) or cell sorting procedures followed by expansion in animal models (Navin *et al.*, 2011), effectively performing multiple samplings of the same tumor. Interestingly, a number of studies have shown that the bulk of tumor cells at different time points, although sharing some common alterations, differs with respect to their array of genomic alterations. In many cases one must assume an ancestral clone that the tumors are derived from, i.e., there is a clonal relationship but not a strict linear evolution (Höglund, 2007; Mullighan *et al.*, 2008).

As described above, SNP arrays provide opportunities to investigate tumor heterogeneity. The possibility to reveal allelic imbalances and infer genotypes of acquired alterations facilitates elucidation of clonal relationships between tumors. Various models to study cancer evolution by LOH, karyotype, and CGH data have been proposed (Höglund *et al.*, 2005; Letouzé *et al.*, 2010; Navin and

Hicks, 2010). For example TuMult uses a computational approach, tracing breakpoints, for unraveling the succession of genomic alterations that has occurred during the process of carcinogenesis (Letouzé *et al.*, 2010). Given the high resolution of current arrays, the presence of multiple identical breakpoints in tumors is highly indicative of a shared origin. Investigating shared copy number alterations and mapping breakpoints may be supplemented by genotype information provided by SNP array analysis.

We will here present some hypothetical examples of how SNP array data can be used to analyze multiple tumors from the same patient in order to investigate clonal expansion, chronology of events, and divergence in clonal evolution. We first return to our example describing a sample of 80% tumor cells and 20% normal diploid cells in which we demonstrated how intra-tumor heterogeneity could be readily discerned (Figs. 5, 8D, and 9). In our example, the presence of a subclone was indicated by segmented mBAF/LRR values signifying an acquired deletion. The observed BAF for the alteration corresponds to that 50% of the tumor cells carries the deletion, and we can look for the deletion in other tumor samples from the same patient. For instance, if a metastasis or recurrence is available from the same patient we can investigate whether it too carries the identified deletion. The estimated proportion of cells that carry the deletion can yield information on whether the recurrence or metastasis represents an expansion of the specific subclone identified in the primary tumor, e.g., if the deletion is present in the majority of the cells. Recall that the example from Fig. 9 included an additional late alteration: a one copy gain estimated to be present in 50% of the tumor cells. When analyzing the recurrence, the identified gain might

be estimated to be present in the same proportion as the deletion, indicating that both alterations were confined to the same subclone in the primary. Alternatively, one might fail to detect the gain in the recurrence altogether. The latter scenario will suggest that the two alterations were in fact confined to separate subclones in the primary. In line with this simplified example, numerous paired analyses of tumor samples can be imagined that aim to describe plausible relationships between tumors from the same individual.

Apart from discerning possible subclonal expansions, as exemplified above, the inherent properties of SNP arrays provide additional possibilities for tracing clonal hierarchies. It is of importance to stress one obvious, but fundamental, principle in a clonal evolution model; a subsequent clone cannot re-acquire an allele that has been lost, that is, a clone that is heterozygous for a given locus cannot be a direct descendant from a clone that is homozygous at that locus. Such a situation is exemplified in Fig. 10A in which one tumor clone (C1) carries only one homologue of chromosome 9 and the other clone (C2) has retained both homologues. In this example we can conclude that C2 cannot be directly descending from C1, however, the opposite is of course possible. Analogously, homozygous deletions are ideal to discern clonal relationships since complete loss of a locus also represents a state that cannot be reversed.

With this simple principle in mind, we will introduce the concept of "imbalance haplotype" (IH). That is, for any region of allelic imbalance, it is possible to determine the dominating haplotype, i.e., the consecutive series of SNP alleles that are in abundance. We exemplify this for a deletion in which BAF is used to infer the complete haplotype sequences of the parental alleles (Fig.

10B). Importantly, if the actual haplotype is known, it can be used to query an alteration of the same region but in a separate sample from the same patient. By extrapolating the actual haplotype series from BAF values, we therefore can conclude if alterations in tumors from the same patient can be ascribed to the same chromosomal homologue or not.

We will now use this line of reasoning in a hypothetical example aimed to model an underlying hierarchy among tumor clones (Fig. 10C). In the example given we use three tumors (T1, T2, and T3) that are obtained from the same individual, but at separate time points. From the list of alterations, we can identify a focal deletion at 9p21 that is present in all tumors. The IHs for this region are also identical and we can therefore assume that all three tumors stem from a shared cell of origin and that this deletion is an early event. Deletion of the same homologue of chromosome arm 17p is shared by tumors T1 and T2, but not by T3. Thus, T3 cannot be a descendant of either T1 or T2. Neither can T1 and T2 be linearly derived from T3 since this tumor harbors a homozygous deletion at 10q23. Furthermore, the T1 and T2 tumors both display heterozygous deletions of 5q, however, different haplotypes are lost in the respective tumors (incompatible IHs). Thus, neither of these clones can directly have given rise to one another. This simple way of deducing clonal relationships, based on compatible and incompatible events, thus, provide us with an opportunity to connect the tumors hierarchically (Fig. 10C, right). We can conclude that the tumors share a common origin; moreover, we can reject a straightforward linear model of clonal evolution. We can, in addition, infer nodes of lineage deviation representing obligate ancestral clones.

Even though the above example may be overly simplified, it still conveys the basal concept of how SNP arrays may be used to address issues of clonality and tumor evolution. There are, to our knowledge, no reports that use SNP arrays to infer IHs and that take these into consideration when assessing clonal relationships between tumors. Nonetheless, the same conceptual thinking, i.e., demonstrating loss of incompatible genotypes, has been applied in earlier studies using LOH analyses, demonstrating its feasibility (van Tilborg *et al.*, 2000; Lindgren *et al.*, 2006).

V. CONCLUDING REMARKS

Throughout recent years, molecular techniques to study cancer have progressed in terms of resolution and sensitivity, but also with respect to accessibility due to decreased cost. Microarray based platforms have evolved from proof-of-concepts – presented little more than a decade ago – to highly standardized off-the-shelf assays for genome-wide analysis of gene expression, DNA copy number, and genotypes. Undoubtedly, technologies will continue to evolve and much of what is considered at the forefront today will be superseded tomorrow. We have aimed to present some basic concepts pertaining to the analysis of tumor-heterogeneity using genotyping techniques. In doing so, we have also tried to give a brief account of currently available and standardized platforms for genome-wide genotyping. However, we have refrained from discussing in depth any particular analysis methods inherently tied to the mentioned platforms. Much of what has been presented in terms of data interpretation can in theory be applied to genotype and copy number data in

general. Recent developments in high-throughput sequencing techniques lend promise to resolving some of the limitations of current array-based technology in the analysis of tumor-heterogeneity. Mainly, in terms of sensitivity, array-based analysis may fail to detect alterations confined to minor subpopulations. Nonetheless, current techniques have their merits and will undoubtedly continue to contribute to our understanding of tumor heterogeneity, development, and progression.

REFERENCES

Assie, G., LaFramboise, T., Platzer, P., Bertherat, J., Stratakis, C.A., and Eng, C. (2008). SNP arrays in heterogeneous tissue: highly accurate collection of both germline and somatic genetic information from unpaired single tumor samples. *Am J Hum Genet* **82**, 903-915.

Attiyeh, E.F., Diskin, S.J., Attiyeh, M.A., Mosse, Y.P., Hou, C., Jackson, E.M., Kim, C., Glessner, J., Hakonarson, H., Biegel, J.A., and Maris, J.M. (2009). Genomic copy number determination in cancer cells from single nucleotide polymorphism microarrays based on quantitative genotyping corrected for aneuploidy. *Genome Res* **19**, 276-283.

Baumbusch, L.O., Aaroe, J., Johansen, F.E., Hicks, J., Sun, H., Bruhn, L., Gunderson, K., Naume, B., Kristensen, V.N., Liestol, K., Borresen-Dale, A.L., and Lingjaerde, O.C. (2008). Comparison of the Agilent, ROMA/NimbleGen and Illumina platforms for classification of copy number alterations in human breast tumors. *BMC Genomics* **9**, 379.

Beroukhim, R., Lin, M., Park, Y., Hao, K., Zhao, X., Garraway, L.A., Fox, E.A., Hochberg, E.P., Mellinghoff, I.K., Hofer, M.D., Descazeaud, A., Rubin, M.A., et al. (2006). Inferring loss-of-heterozygosity from unpaired tumors using high-density oligonucleotide SNP arrays. *PLoS Comput Biol* **2**, e41.

Chin, K., DeVries, S., Fridlyand, J., Spellman, P.T., Roydasgupta, R., Kuo, W.L., Lapuk, A., Neve, R.M., Qian, Z., Ryder, T., Chen, F., Feiler, H., et al. (2006). Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell* **10**, 529-541.

Curtis, C., Lynch, A.G., Dunning, M.J., Spiteri, I., Marioni, J.C., Hadfield, J., Chin, S.F., Brenton, J.D., Tavare, S., and Caldas, C. (2009). The pitfalls of platform comparison: DNA copy number array technologies assessed. *BMC Genomics* **10**, 588.

Grant, S.F.A., and Hakonarson, H. (2008). Microarray technology and applications in the arena of genome-wide association. *Clin Chem* **54**, 1116-1124.

Gunderson, K.L., Steemers, F.J., Lee, G., Mendoza, L.G., and Chee, M.S. (2005). A genome-wide scalable SNP genotyping assay using microarray technology. *Nat Genet* **37**, 549-554.

Gunnarsson, R., Staaf, J., Jansson, M., Ottesen, A.M., Goransson, H., Liljedahl, U., Ralfkiaer, U., Mansouri, M., Buhl, A.M., Smedby, K.E., Hjalgrim, H., Syvanen, A.C., et al. (2008). Screening for copy-number alterations and loss of heterozygosity in chronic lymphocytic leukemia--a comparative study of four differently designed, high resolution microarray platforms. *Genes Chromosomes Cancer* **47**, 697-711.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. *Cell* **144**, 646-674.

Hehir-Kwa, J.Y., Egmont-Petersen, M., Janssen, I.M., Smeets, D., van Kessel, A.G., and Veltman, J.A. (2007). Genome-wide copy number profiling on high-density bacterial artificial chromosomes, single-nucleotide polymorphisms, and oligonucleotide microarrays: a platform comparison based on statistical power analysis. *DNA Res* **14**, 1-11.

Heinrichs, S., Li, C., and Look, A.T. (2010). SNP array analysis in hematologic malignancies: avoiding false discoveries. *Blood* **115**, 4157-4161.

Höglund, M. (2007). On the origin of syn- and metachronous urothelial carcinomas. *Eur Urol* **51**, 1185-1193; discussion 1193.

Höglund, M., Frigyesi, A., Sall, T., Gisselsson, D., and Mitelman, F. (2005). Statistical behavior of complex cancer karyotypes. *Genes Chromosomes Cancer* **42**, 327-341.

Iafrate, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.K., Qi, Y., Scherer, S.W., and Lee, C. (2004). Detection of large-scale variation in the human genome. *Nat Genet* **36**, 949-951.

Kallioniemi, A., Kallioniemi, O.P., Sudar, D., Rutovitz, D., Gray, J.W., Waldman, F., and Pinkel, D. (1992). Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* **258**, 818-821.

LaFramboise, T. (2009). Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic acids research* **37**, 4181-4193.

LaFramboise, T., Weir, B.A., Zhao, X., Beroukhim, R., Li, C., Harrington, D., Sellers, W.R., and Meyerson, M. (2005). Allele-specific amplification in cancer revealed by SNP array analysis. *PLoS Comput Biol* **1**, e65.

Lai, W.R., Johnson, M.D., Kucherlapati, R., and Park, P.J. (2005). Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics* **21**, 3763-3770.

Letouzé, E., Allory, Y., Bollet, M.A., Radvanyi, F., and Guyon, F. (2010). Analysis of the copy number profiles of several tumor samples from the same patient reveals the successive steps in tumorigenesis. *Genome Biol* **11**, R76.

Li, A., Liu, Z., Lezon-Geyda, K., Sarkar, S., Lannin, D., Schulz, V., Krop, I., Winer, E., Harris, L., and Tuck, D. (2011). GPHMM: an integrated hidden Markov model for identification of copy number alteration and loss of heterozygosity in complex tumor samples using whole genome SNP arrays. *Nucleic acids research*.

Lin, M., Wei, L.-J., Sellers, W.R., Lieberfarb, M., Wong, W.H., and Li, C. (2004). dChipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data. *Bioinformatics* **20**, 1233-1240.

Lindblad-Toh, K., Tanenbaum, D.M., Daly, M.J., Winchester, E., Lui, W.O., Villapakkam, A., Stanton, S.E., Larsson, C., Hudson, T.J., Johnson, B.E., Lander, E.S., and Meyerson, M. (2000). Loss-of-heterozygosity analysis of small-cell lung carcinomas using single-nucleotide polymorphism arrays. *Nat Biotechnol* **18**, 1001-1005.

Lindgren, D., Liedberg, F., Andersson, A., Chebil, G., Gudjonsson, S., Borg, A., Mansson, W., Fioretos, T., and Hoglund, M. (2006). Molecular characterization of early-stage bladder carcinomas by expression profiles, FGFR3 mutation status, and loss of 9q. *Oncogene* **25**, 2685-2696.

Matsuzaki, H., Dong, S., Loi, H., Di, X., Liu, G., Hubbell, E., Law, J., Berntsen, T., Chadha, M., Hui, H., Yang, G., Kennedy, G.C., et al. (2004). Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat Methods* **1**, 109-111.

McQuillan, R., Leutenegger, A.L., Abdel-Rahman, R., Franklin, C.S., Pericic, M., Barac-Lauc, L., Smolej-Narancic, N., Janicijevic, B., Polasek, O., Tenesa, A., Macleod, A.K., Farrington, S.M., et al. (2008). Runs of homozygosity in European populations. *Am J Hum Genet* **83**, 359-372.

Mullighan, C.G., Phillips, L.A., Su, X., Ma, J., Miller, C.B., Shurtleff, S.A., and Downing, J.R. (2008). Genomic analysis of the clonal origins of relapsed acute lymphoblastic leukemia. *Science* **322**, 1377-1380.

Nancarrow, D.J., Handoko, H.Y., Stark, M.S., Whiteman, D.C., and Hayward, N.K. (2007). SiDCoN: a tool to aid scoring of DNA copy number changes in SNP chip data. *PLoS ONE* **2**, e1093.

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., Muthuswamy, L., Krasnitz, A., et al. (2011). Tumour evolution inferred by single-cell sequencing. *Nature*.

Navin, N., Krasnitz, A., Rodgers, L., Cook, K., Meth, J., Kendall, J., Riggs, M., Eberling, Y., Troge, J., Grubor, V., Levy, D., Lundin, P., et al. (2010). Inferring tumor progression from genomic heterogeneity. *Genome Res* **20**, 68-80.

Navin, N.E., and Hicks, J. (2010). Tracing the tumor lineage. *Mol Oncol* **4**, 267-283.

Paulsson, K., Lindgren, D., and Johansson, B. (2011). SNP array analysis of leukemic relapse samples after allogeneic hematopoietic stem cell transplantation with a sibling donor identifies meiotic recombination spots and

reveals possible correlation with the breakpoints of acquired genetic aberrations. *Leukemia*.

Peiffer, D.A., Le, J.M., Steemers, F.J., Chang, W., Jenniges, T., Garcia, F., Haden, K., Li, J., Shaw, C.A., Belmont, J., Cheung, S.W., Shen, R.M., et al. (2006). High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res* **16**, 1136-1148.

Pinkel, D., and Albertson, D.G. (2005). Array comparative genomic hybridization and its applications in cancer. *Nat Genet* **37 Suppl**, S11-17.

Pollack, J.R., Perou, C.M., Alizadeh, A.A., Eisen, M.B., Pergamenschikov, A., Williams, C.F., Jeffrey, S.S., Botstein, D., and Brown, P.O. (1999). Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat Genet* **23**, 41-46.

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

Rajagopalan, H., and Lengauer, C. (2004). Aneuploidy and cancer. *Nature* **432**, 338-341.

Russnes, H.G., Vollan, H.K., Lingjaerde, O.C., Krasnitz, A., Lundin, P., Naume, B., Sorlie, T., Borgen, E., Rye, I.H., Langerod, A., Chin, S.F., Teschendorff, A.E., et al. (2010). Genomic architecture characterizes tumor progression paths and fate in breast cancer patients. *Sci Transl Med* **2**, 38ra47.

Schena, M., Shalon, D., Davis, R.W., and Brown, P.O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467-470.

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Maner, S., Massa, H., Walker, M., Chi, M., Navin, N., Lucito, R., et al. (2004). Large-scale copy number polymorphism in the human genome. *Science* **305**, 525-528.

Staaf, J., Jönsson, G., Ringnér, M., and Vallon-Christersson, J. (2007). Normalization of array-CGH data: influence of copy number imbalances. *BMC Genomics* **8**, 382.

Staaf, J., Lindgren, D., Vallon-Christersson, J., Isaksson, A., Göransson, H., Juliusson, G., Rosenquist, R., Höglund, M., Borg, A., and Ringnér, M. (2008). Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome SNP arrays. *Genome Biol* **9**, R136.

Steemers, F.J., Chang, W., Lee, G., Barker, D.L., Shen, R., and Gunderson, K.L. (2006). Whole-genome genotyping with the single-base extension assay. *Nat Methods* **3**, 31-33.

Sun, W., Wright, F.A., Tang, Z., Nordgard, S.H., Van Loo, P., Yu, T., Kristensen, V.N., and Perou, C.M. (2009). Integrated study of copy number states and genotype calls using high-density SNP arrays. *Nucleic acids research* **37**, 5365-5377.

Tolliver, D., Tsourakakis, C., Subramanian, A., Shackney, S., and Schwartz, R. (2010). Robust unmixing of tumor states in array comparative genomic hybridization data. *Bioinformatics* **26**, i106-114.

Van Loo, P., Nordgard, S.H., Lingjærde, O.C., Russnes, H.G., Rye, I.H., Sun, W., Weigman, V.J., Marynen, P., Zetterberg, A., Naume, B., Perou, C.M., Børresen-Dale, A.-L., et al. (2010). Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci USA* **107**, 16910-16915.

van Tilborg, A.A., de Vries, A., de Bont, M., Groenfeld, L.E., van der Kwast, T.H., and Zwarthoff, E.C. (2000). Molecular evolution of multiple recurrent cancers of the bladder. *Hum Mol Genet* **9**, 2973-2980.

Venkatraman, E.S., and Olshen, A.B. (2007). A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23**, 657-663.

Wang, D.G., Fan, J.B., Siao, C.J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L., et al. (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**, 1077-1082.

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S.F., Hakonarson, H., and Bucan, M. (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.

Wang, Y., and Armstrong, S.A. (2007). Genome-wide SNP analysis in cancer: leukemia shows the way. *Cancer Cell* **11**, 308-309.

Wang, Z.C., Lin, M., Wei, L.-J., Li, C., Miron, A., Lodeiro, G., Harris, L., Ramaswamy, S., Tanenbaum, D.M., Meyerson, M., Iglehart, J.D., and Richardson, A. (2004). Loss of heterozygosity and its correlation with expression profiles in subclasses of invasive breast cancers. *Cancer Res* **64**, 64-71.

Zhao, X., Weir, B.A., LaFramboise, T., Lin, M., Beroukhim, R., Garraway, L., Beheshti, J., Lee, J.C., Naoki, K., Richards, W.G., Sugarbaker, D., Chen, F., et al. (2005). Homozygous deletions and chromosome amplifications in human lung carcinomas revealed by single nucleotide polymorphism array analysis. *Cancer Res* **65**, 5561-5570.

Table I The Association between Genotype and BAF.

Genotype	CN[a]	BAF
-	0 (HD)[b]	-
A	1	0
B	1	1
AA	2	0
AB	2	0.5
BB	2	1
AAA	3	0
AAB	3	0.33
ABB	3	0.67
BBB	3	1
AAAA	4	0
AAAB	4	0.25
AABB	4	0.5
ABBB	4	0.75
BBBB	4	1
AABBB	5	0.6
ABBBB	5	0.8
AAABBB	6	0.5
AABBBB	6	0.67
ABBBBB	6	0.83

[a] CN=Total number of allele copies, [b] HD=Homozygous deletion

Figure Legends

Fig. 1 Schematic illustration of the basic principles of how allele specific intensity values are measured using the Affymetrix and Illumina assays. **A)** Parental homologues comprising one centrally located heterozygous SNP (T/G). **B)** The Affymetrix assay relies on multiple allele specific probes spanning the interrogated SNP and complementary to either T or G. In the illustration only one allele specific probe per allele is depicted. The probes are located in separated features on the array surface and will preferentially hybridize labeled target of perfect complementarity. Relative difference in allele abundance is resolved by comparing the quantified intensities from separate features harboring the respective probes. **C)** The Illumina assay relies on a single loci specific probe complementary to the sequence adjacent to the interrogated SNP. The probe will hybridize both parental homologues in a non-allele specific manner. The hybridized target is used as template in a subsequent enzymatic single-base extension step employing differentially labeled nucleotides. The use of differentially labeled nucleotides permits dual intensity quantification in the same feature and relative difference in allele abundance is resolved by comparing the quantified intensities.

Fig. 2 SNP array analysis of a normal diploid genome. **A)** A schematic illustration of an expected BAF and LRR plot from the analysis of a diploid genome. Each individual SNPs has a BAF value of 0, 0.5, or 1, reflecting the genotype for that specific locus (*AA*, *AB*, and *BB*, respectively). BAF and LRR values for germ line homozygous- and heterozygous SNPs are colored gray and

black, respectively. **B)** Experimentally obtained BAF and LRR plots of a chromosome. In the BAF plot, individual SNPs cluster close to 0, 0.5, or 1, producing three characteristic horizontal bands. At the LRR level, the majority of SNPs cluster around 0, representing the measurement of two DNA copies.

Fig. 3 Schematic examples of common genomic alterations and their expected BAF, LRR, and mBAF. **A)** The constitutional genotype showing the two parental homologues of a diploid genome, each with its own specific haplotype series, is shown at the top. Allele combinations for a tumor with acquired genomic alterations are shown below the parental alleles. The balance between A and B, and the total number of allele copies, will determine the BAF and LRR for each SNP locus. Thus, each alteration causes a shift in the BAF and/or the LRR profile. BAF and LRR values for germ line homozygous- and heterozygous SNPs are colored gray and black, respectively. At the copy number level all SNPs are informative. Expected BAF and LRR plots for the acquired alterations are shown at the bottom. From left to right (possible genotype combinations for germline heterozygous SNPs are given within parenthesis): normal balanced genotype (*AB*), a one copy deletion (*A* and *B* genotypes), a one copy gain (*AAB* and *ABB* genotypes), a two copy gain in which both surplus segments are derived from the same chromosomal homologue (*AAAB*, and *ABBB* genotypes), a balanced two copy gain (*AABB* genotypes), a segment with copy number neutral LOH (*AA* and *BB* genotypes). **B)** Schematic mBAF transformation of the above BAF profile. The mBAF is mirrored at the 0.5-axis yielding only one possible value for SNP loci that were germline heterozygous. For example, in the *AAAB*/*ABBB* two copy gain

segment, the *AAAB* genotypes (BAF=0.2) will be transformed to the mirrored genotype (*BBBA*) with mBAF=0.8. Non-informative homozygous SNPs are excluded from this plot.

Fig. 4 Experimental BAF, mBAF, and LRR plots of two chromosomes obtained from a SNP array analysis of a tumor. Deviations from the expected BAF and LRR patterns for a normal diploid genome are observed, indicative of acquired chromosomal alterations, including an intrachromosomal deletion at 1p21-p31, gain of 5p, and deletion of 5q. The mBAF profiles have been segmented, identifying breakpoints that delineate segments that constitute a discrete genomic state. Identified segments are ascribed specific mBAF and LRR values based on the SNPs between breakpoints. For example, three segments define chromosome 1 whereas chromosome 5 can be described with two separate segments.

Fig. 5 A schematic karyotype of a diploid tumor and its corresponding mBAF/LRR plot. The karyotype illustrates: hemizygous deletions of chromosomes 13, 9p, and 17p, gains of chromosomes 22 (trisomy, *BBA*) and 6p (four copies, *BBBA*), and copy-number neutral allelic imbalance of chromosome 11 (*BB*). The parental chromosomal homologues are colored in yellow or blue, respectively. For the mBAF/LRR plot, each individual circle in the figure represents a continuous chromosomal segment with identical LRR and BAF as given by the tumor karyotype. For example, chromosome 6 is represented by two separate segments: 6p (mBAF=0.75, and LRR=0.65), and the unaffected 6q

(BAF=0.5, LRR=0). Chromosome 11 has no intra-chromosomal breakpoints and is only represented by one segment (BAF=1 and LRR=0).

Fig. 6 Experimental BAF, mBAF, and LRR plots obtained from two tumors with increased ploidy number. **A)** Representation of chromosome 11 for a near triploid tumor. Most of the chromosome displays a three-banded pattern characteristic for a trisomy (mBAF=0.67 and LRR=0). A deletion is however observed at 11q23 (mBAF=0.97, LRR=-0.22), and a gain at 11q24 (mBAF=0.58, mBAF=0.26). The deletion possibly indicates one copy loss resulting in a *BB* genotype. The gain possibly indicates duplication of material from both homologues relative to the trisomic state (*AABBB*). **B)** Chromosome 5 for a near tetraploid tumor. A small segment of the chromosome (5p14-p15) is present in four copies (mBAF=0.51, LRR=0). A net loss of one copy for the remainder of the chromosome results in imbalanced mBAF value and negative LRR value, closely matching an *ABB* genotype.

Fig. 7 Karyotypes of two aneuploid tumors and their corresponding theoretical representations in mBAF/LRR plots**. A)** A near-triploid tumor. The parental homologues are colored in yellow and blue, respectively. Relative copy-number gains are observed for chromosome 22 and 6p. Net losses of material are seen for chromosomes 9p, 9q, 17p, 6q, and chromosome and 13. Copy number neutral LOH is seen for chromosome 11 for which one homologue is present in three copies. In the mBAF/LRR plot the majority of segments are in allelic imbalance (*BBA*) and located at mBAF=0.67 and LRR=0. Deletions or gains are shifted

towards either heterozygosity (*AB* or *AABB*) or towards increased allelic imbalance (*B*, *BB*, or *BBBA*). **B)** A near-tetraploid tumor karyotype and its representation in mBAF/LRR space. Most segments are located at mBAF=0.5 and LRR=0. A number of different genotypes are seen for regions with negative LRR (e.g. *BBA*, *BB*, BA, and *B*). Similarly, a variety of genotypes representing net gain of material are observed (e.g. *BBBAA*, *BBBBA*, and *BBBBAA*), each with its' specific expected LRR and mBAF values.

Fig. 8 BAF and LRR in case of tumor and normal cell admixture. **A)** Schematic example of a sample containing 2 normal diploid cells and 8 tumor cells with an acquired deletion (the lost allele is grayed out). The tumor will contribute with eight *B* alleles, whereas the normal cells contribute with two *A* and two *B* alleles. The expected mBAF for the sample is calculated to 0.83. **B)** Schematic BAF plot illustrating a hemizygous deletion in a sample with 80% tumor cells and 20% normal cells. Germline heterozygous SNPs within the region will not reach their expected BAF values (0 or 1) since the background of normal cells will contribute with both *A* and *B* alleles. **C)** Line-plots of expected mBAF for a number of different genotypes as a function of the fraction of normal diploid cells present within the sample. For example, a tumor segment with an *AAB* genotype has an expected mBAF of 0.6 if intermixed with 50% normal diploid cells. **D)** Schematic mBAF/LRR plot of a sample with 80% tumor and 20% normal cells. The tumor karyotype is identical to the karyotype presented in Fig. 5. Since normal diploid cells (*AB*) are present in the sample, the mBAF and LRR for the respective alterations are shifted towards mBAF=0.5 and LRR=0 along

theoretical lines (gray). Solid gray circles indicate expected LRR and mBAF values in case of no normal cells present.

Fig. 9 Schematic mBAF/LRR plot of a sample with a 20% diploid cell background (80% tumor cells) and two subclonal events within the tumor cell population. The tumor karyotype is identical to the karyotype presented in Fig. 5, although 50% of the tumor cells also have acquired a gain of 5p (*BBA*) and a loss of 5q (*B*). Since only 40% of the cells in the sample carry these alterations, the respective mBAF and LRR for these segments will deviate (arrows) from the pattern observed for the alterations present in all tumor cells.

Fig. 10 Addressing clonal hierarchy using SNP array data. **A)** Schematic example of two tumor clones (C1 and C2). The C2 clone has lost one homologue of chromosome 9. Therefore, subclone C2 cannot be a descendant from C1 (crossed arrow). The reverse is possible (arrow). **B)** Definition of imbalance haplotype (IH). The IH is defined by the series of alleles that are in excess for a given genomic alteration. In this example a loss of a chromosomal segment is illustrated. Each SNP within the IH is called from its respective BAF. Non-informative germ line homozygous SNPs (gray circles) are not considered in the IH sequence. **C)** Schematic example of three tumors derived from the same individual. Identified alterations in each tumor are listed to the left. Incompatible IHs for the 5q deletion is indicated by separate colors. A hierarchical tree describing clonal relationships can be deduced from the given alterations (right).

Two obligate ancestral clones (A1 and A2, respectively) must be assumed as intermediate steps to describe the clonal evolution in this example.

Figure 1

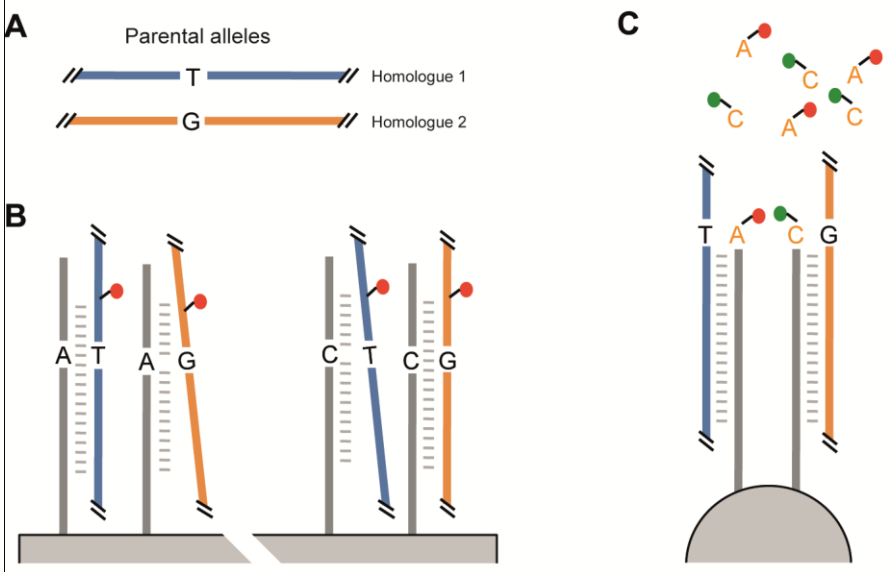

A
Parental alleles
T
Homologue 1
G
Homologue 2
B
A T
A G
C T
C G
C
A
C
A
C
A
C
T A C G

Figure 2

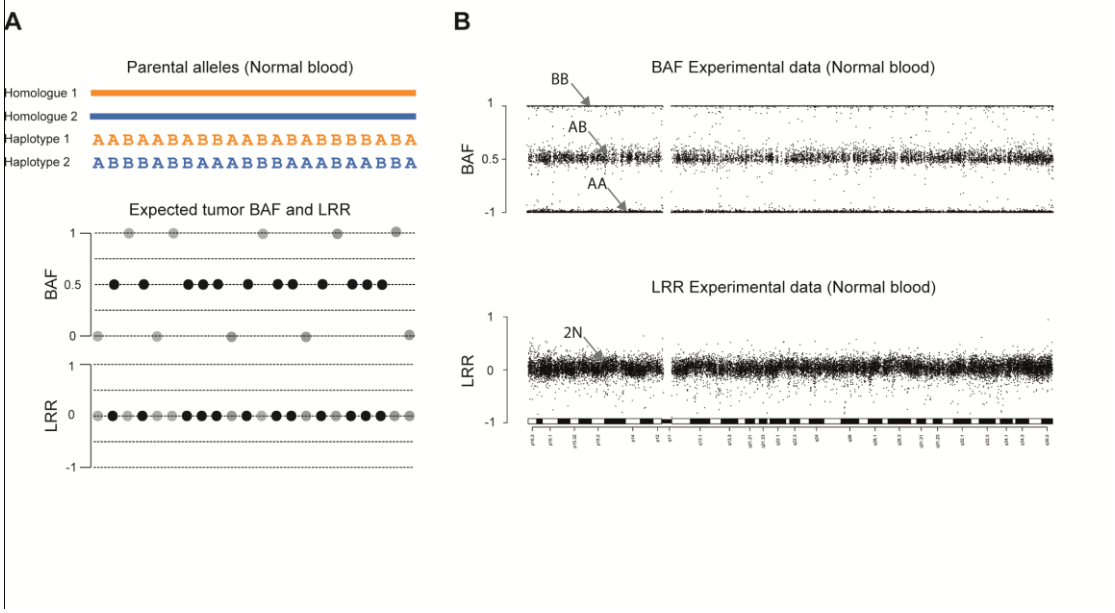

A
Parental alleles (Normal blood)
Homologue 1
Homologue 2
Haplotype 1
Haplotype 2
AABAABABBAABABABBBBABA
ABBBABBBAAABBBAAABAABBA
Expected tumor BAF and LRR
BAF
LRR
1
0.5
0
-1
B
BAF Experimental data (Normal blood)
BB
AB
AA
LRR Experimental data (Normal blood)
2N

Figure 3

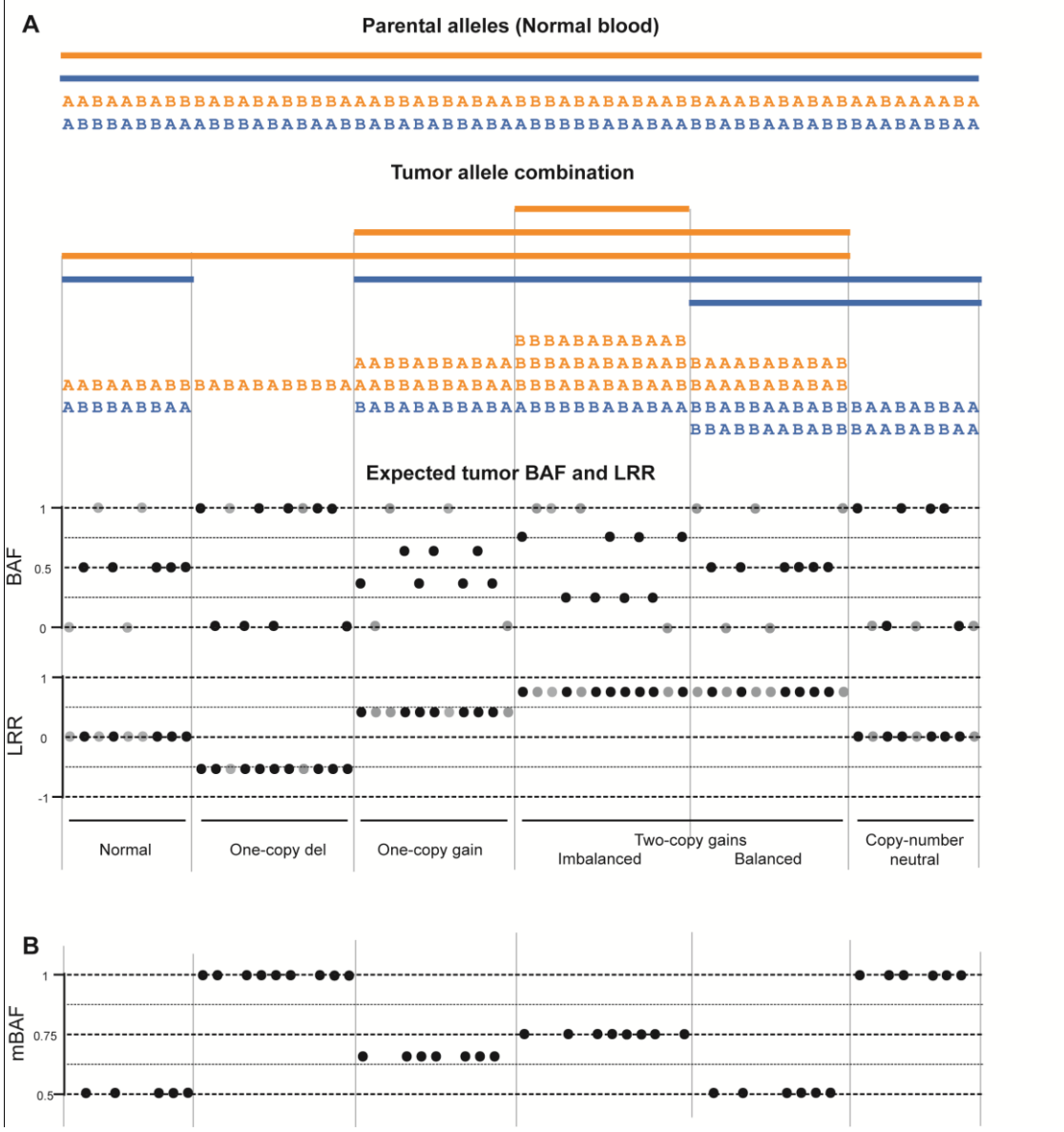

A
Parental alleles (Normal blood)
Tumor allele combination
Expected tumor BAF and LRR
BAF
LRR
Normal
One-copy del
One-copy gain
Two-copy gains
Imbalanced
Balanced
Copy-number neutral
B
mBAF

Figure 4

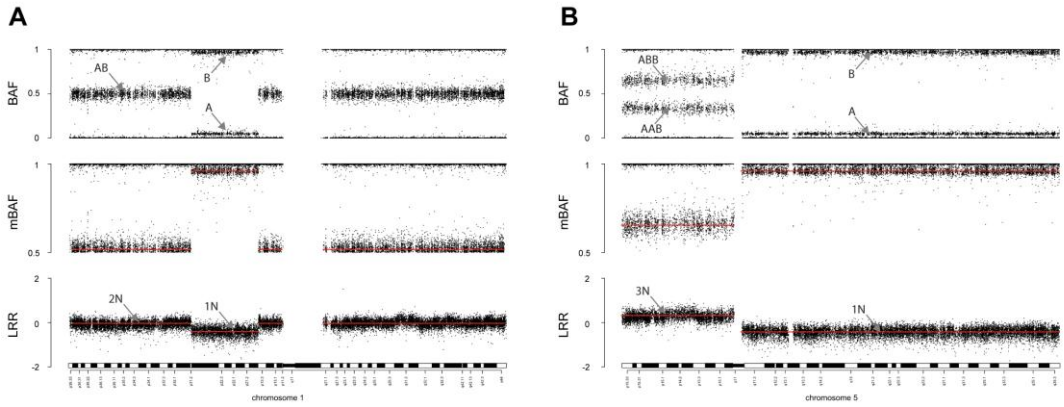

A
BAF
mBAF
LRR
AB
B
A
2N
1N
chromosome 1
B
BAF
mBAF
LRR
ABB
AAB
B
A
3N
1N
chromosome 5

Figure 5

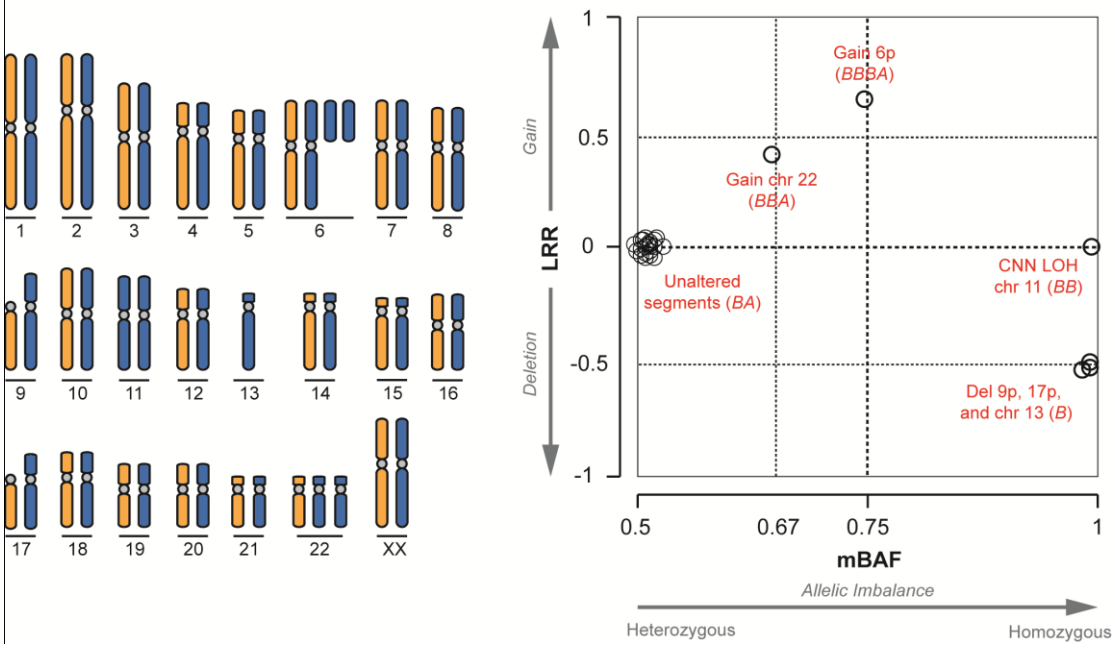

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
XX
Gain
Deletion
LRR
1
0.5
0
-0.5
-1
Gain 6p
(BBBA)
Gain chr 22
(BBA)
Unaltered
segments (BA)
CNN LOH
chr 11 (BB)
Del 9p, 17p,
and chr 13 (B)
0.5
0.67
0.75
1
mBAF
Allelic Imbalance
Heterozygous
Homozygous

Figure 6

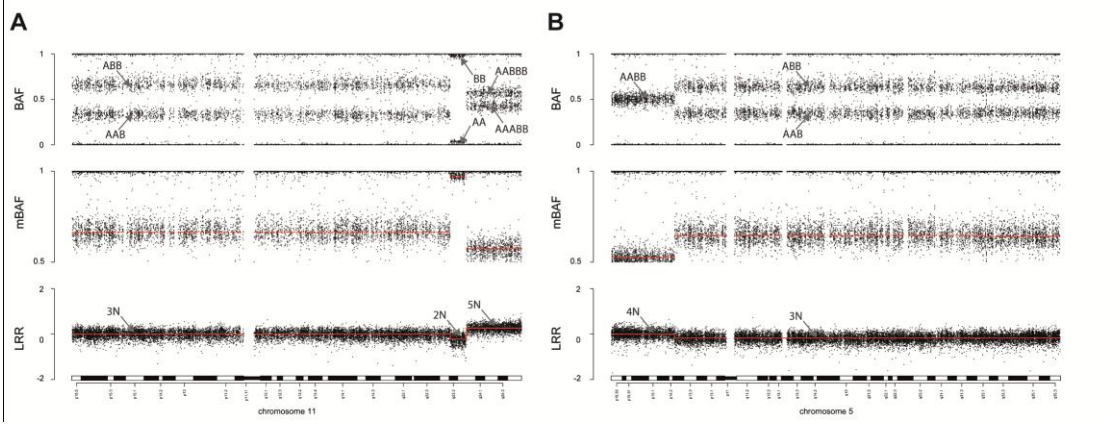

A
B
BAF
mBAF
LRR
ABB
AAB
BB
AABBB
AA
AAABB
3N
2N
5N
AABB
4N
chromosome 11
chromosome 5

Figure 7

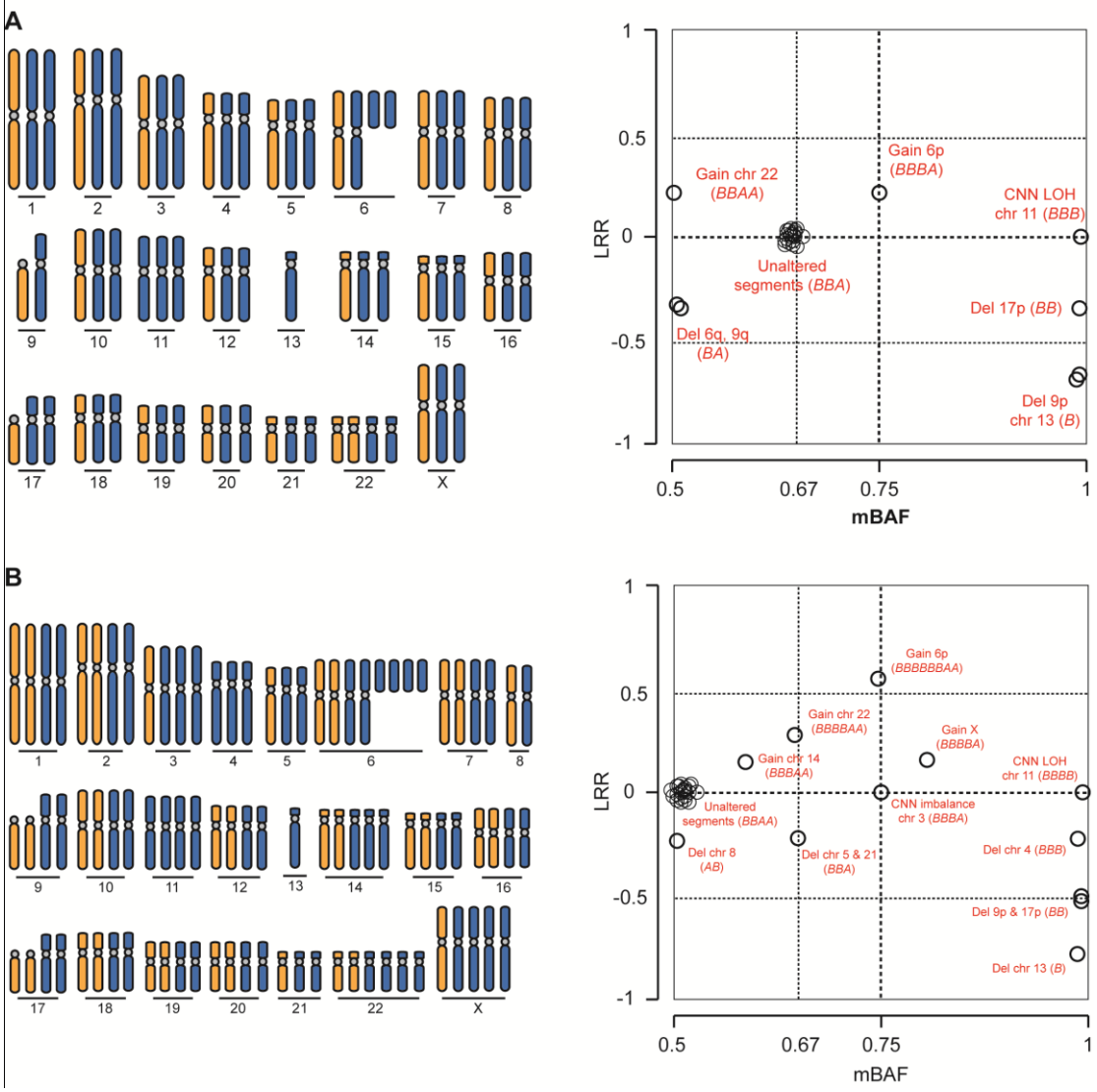

A
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
LRR
mBAF
Gain chr 22 (BBAA)
Gain 6p (BBBA)
CNN LOH chr 11 (BBB)
Unaltered segments (BBA)
Del 17p (BB)
Del 6q, 9q (BA)
Del 9p chr 13 (B)
B
Gain 6p (BBBBBBAA)
Gain chr 22 (BBBBAA)
Gain X (BBBBA)
Gain chr 14 (BBBAA)
CNN LOH chr 11 (BBBB)
Unaltered segments (BBAA)
CNN imbalance chr 3 (BBBA)
Del chr 8 (AB)
Del chr 5 & 21 (BBA)
Del chr 4 (BBB)
Del 9p & 17p (BB)
Del chr 13 (B)

Figure 8

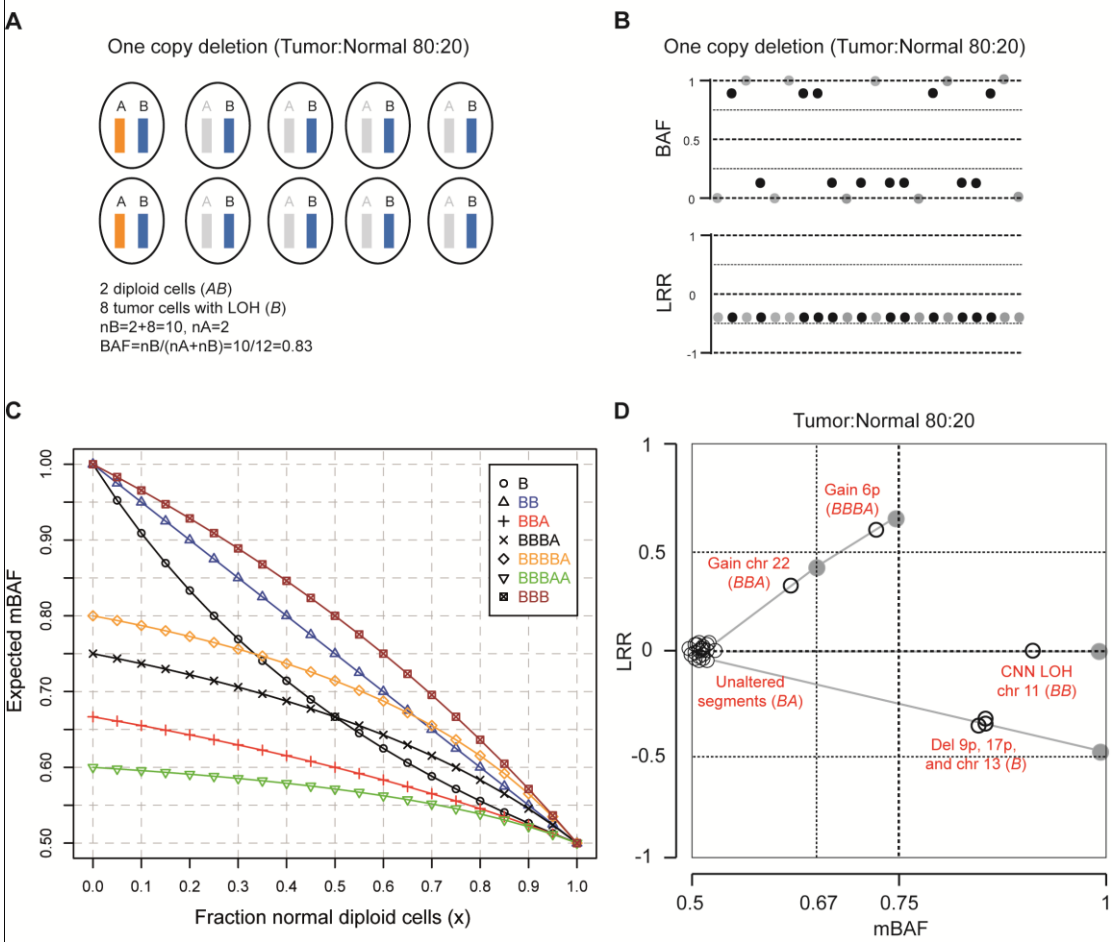

A
One copy deletion (Tumor:Normal 80:20)
A B
2 diploid cells (AB)
8 tumor cells with LOH (B)
nB=2+8=10, nA=2
BAF=nB/(nA+nB)=10/12=0.83
B
One copy deletion (Tumor:Normal 80:20)
BAF
LRR
C
Expected mBAF
Fraction normal diploid cells (x)
B
BB
BBA
BBBA
BBBBA
BBBAA
BBB
D
Tumor:Normal 80:20
LRR
mBAF
Gain 6p (BBBA)
Gain chr 22 (BBA)
Unaltered segments (BA)
CNN LOH chr 11 (BB)
Del 9p, 17p, and chr 13 (B)

Figure 9

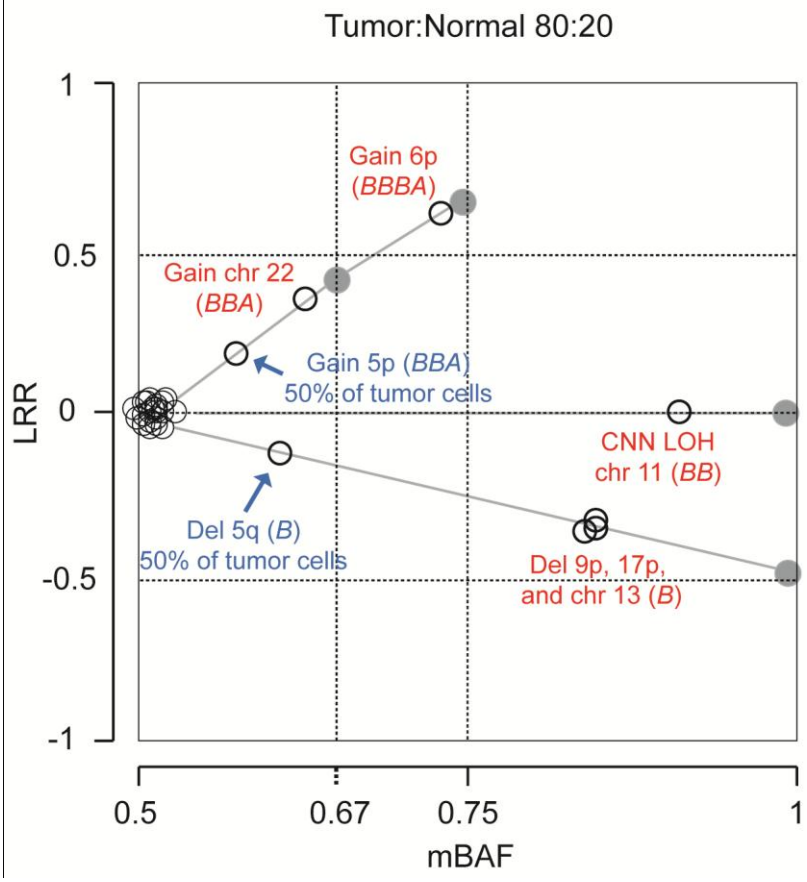
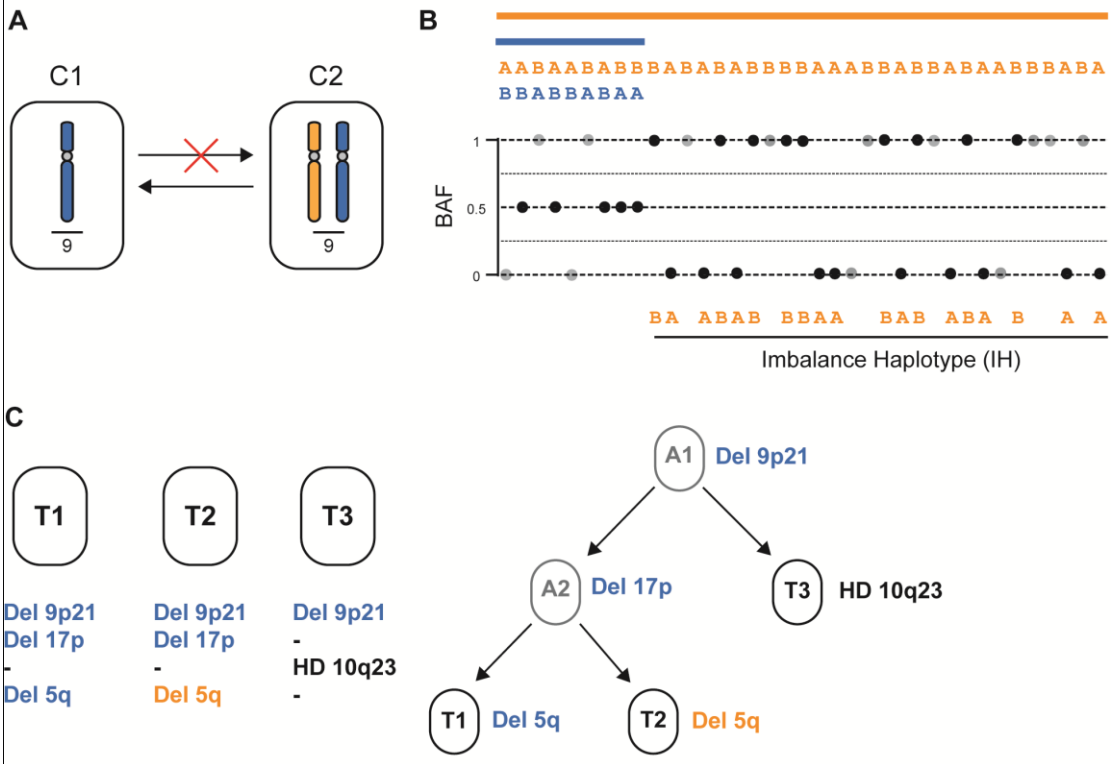

Tumor:Normal 80:20
1
0.5
0
-0.5
-1
LRR
Gain 6p
(BBBA)
Gain chr 22
(BBA)
Gain 5p (BBA)
50% of tumor cells
CNN LOH
chr 11 (BB)
Del 5q (B)
50% of tumor cells
Del 9p, 17p,
and chr 13 (B)
0.5
0.67
0.75
1
mBAF

Figure 10


A
C1
C2
9
9
B
AABAABABBBABABABBBBAAABBABBABAABBBABA
BBABBABAA
1
0.5
0
BAF
BA ABAB BBAA BAB ABA B A A
Imbalance Haplotype (IH)
C
T1
T2
T3
Del 9p21
Del 17p
-
Del 5q
Del 9p21
Del 17p
-
Del 5q
Del 9p21
-
HD 10q23
-
A1
Del 9p21
A2
Del 17p
T3
HD 10q23
T1
Del 5q
T2
Del 5q