Analysis of Somatic Alterations in Cancer Genome: From SNP Arrays to Next Generation Sequencing

Tatiana Popova, Valentina Boeva, Elodie Manié, Yves Rozenholc, Emmanuel Barillot, Marc-Henri Stern

Analysis of Somatic Alterations in Cancer Genome: From SNP Arrays to Next Generation Sequencing

Tatiana Popova
Institut Curie, Centre de Recherche
INSERM U900, INSERM U830
France

Valentina Boeva
Institut Curie, Centre de Recherche
INSERM U900
France

Elodie Manié
Institut Curie, Centre de Recherche
INSERM U830
France

Yves Rozenholc
Université Paris Descartes
MAP5, UMR CNRS 8145
France

Emmanuel Barillot
Institut Curie, Centre de Recherche
INSERM U900, Mines ParisTech
France

Marc-Henri Stern
Institut Curie, Centre de Recherche
INSERM U830
France

1 Introduction

Abnormal genetic content is observed in many tumors and is considered as one of the hallmarks of cancer (Hanahan & Weinberg, 2011). Evolution of cancer is thought to be tightly connected with evolution of cancer genome (Podlaha *et al*., 2012). Studying the cancer genome is important as it may unravel the key genomic events or some particular genomic features, which could shed light on tumor biology and have clinical implications. When considering the genome of an advanced tumor, we observe the "end" point of its evolution often showing numerous acquired rearrangements (Figures 1 & 2). Exact genetic pathways and chronology of events acquisition remain unclear and have only started to be unraveled using whole genome sequencing (Stratton *et al*., 2009). Reconstructed history of several tumor genomes supported the hypothesis of sequential acquisition of genomic events (with possibly variable density in time) and clonal development (Nik-Zainal *et al*., 2012). Although there is still debate whether genomic alterations are causes or consequences of cancer, numerous genomic features have already been linked to initiation and progression in many types of cancer (Tran *et al*., 2012). Some of these genomic features were introduced into clinical practice largely contributing to the diagnostics and treatment choices (Stuart & Sellers, 2009).

In the last decades, several high throughput techniques have been developed to measure genetic alterations in cancer. Spectral Karyotyping (SKY) or other mitotic chromosome imaging gives a general view on chromosomal content in a cancer cell, but the resolution is low, and fine description of genetic alterations based on these images is largely impossible (Figure 2A). Array comparative genomic hybridization (CGH) has been widely in use for the last 20 years showing relative pattern of genetic alterations at rather high resolution (Figure 2B). Single nucleotide polymorphism-based platforms (SNP-arrays) have progressively replaced CGH in tumor characterization, as they allowed estimation of absolute copy numbers and allelic contents (Figure 2C). Next Generation Sequencing (NGS) gives the most complete information about cancer genome, including point mutations and chromosomal translocations at base pair resolution (Figure 2D).

In this chapter we consider basic hypothesis, problem statements and technological and computational solutions for analysis of copy number alterations in tumor genomes. We provide a data mining technique (based on the GAP method described in (Popova *et al*., 2009)) which allows extraction of absolute copy numbers and allelic contents from the whole genome copy number variation and allelic imbalance profiles obtained by SNP arrays or NGS.

2 Basic Hypothesis in Exploratory Analysis of Cancer Genome

What events are essential for tumor initiation and evolution and how they are manifested in the tumor genome? In the simplistic view, there are two major contributors to cancer development: first, tumor promoting events, i.e. "switching on" oncogenes (growth factors, etc); second, elimination of antitumor protection, i.e. "switching off" tumor suppressor genes (genome guardians, checkpoints, etc) (Grander, 1998). Major genomic mechanisms of oncogene activation are gain-of-function mutations, chromosome translocations (which, for example, place the gene under an activating promoter) and amplification of chromosome region containing the gene. Major genomic mechanisms resulting in inactivation of tumor suppressor gene are mutations and deletions of chromosome region containing the gene. The Knudson's two hit hypothesis for recessive tumor suppressor gene implies tumor initiation when two copies of the gene are inactivated (i.e. in the normal diploid human genome, where autosomic genes exist in two copies) (Knudson, 1971).

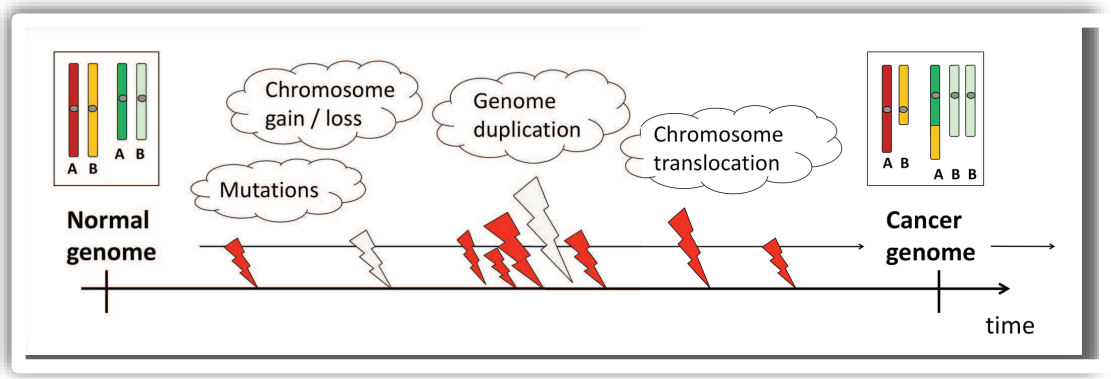


Figure 1: Sequential acquisition of genetic alterations in tumor evolution.

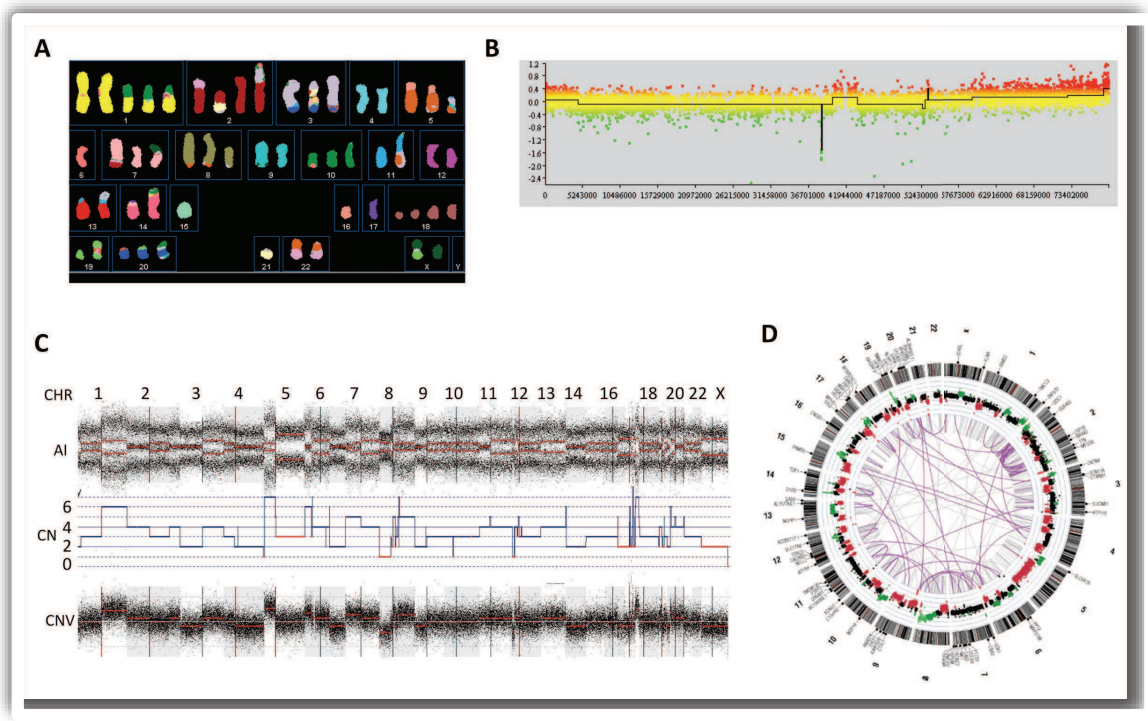


Figure 2: Measuring alterations in a cancer genome. **A.** Whole genome karyotyping image (SKY) of breast cancer cell line MDA-MB-468 (`http://www.path.cam.ac.uk/pawefish/`); **B.** High resolution arrays (BAC, CGH) representing relative copy number variation profile; **C.** High resolution Affymetrix SNP chip 6.0 array profiles of a primary breast tumor: allelic imbalance (*AI*), copy number variation (*CNV*) and recognized absolute copy number (*CN*) profiles; **D.** Circus plot showing chromosomal translocations and copy number variation obtained by NGS for a primary breast tumor (figure from (Natrajan *et al*., 2012)).

Double deletion of a locus or combination of mutation and loss of normal copy (resulting in so called loss of heterozygosity [LOH]) are the genomic indicators of possible tumor suppressor genes. The association between recurrent deletions and tumor suppressors as well as amplifications and oncogenes has been confirmed by the analysis of more than 1300 tumor genomes (Beroukhim *et al*., 2010).

Besides local genomic indications for lost and acquired genes the whole pattern of alterations in a tumor genome, such as chromosomal content (ploidy), number of chromosomal breaks, complexity of intra- and inter-chromosomal rearrangements, may evidence disruption of some pathways, such as DNA repair or chromosome maintenance. For example, *BRCA*$1^{-/-}$ breast tumors display more rearranged genomes compared to non-BRCA1 tumors; alternatively, non-BRCA1 tumors with highly rearranged genomes could be mutated in another gene from the DNA repair pathway, such as *BRCA2* (Popova *et al*., 2012). In this regard analysis of tumor genome profiles could contribute to the tumor classification problem.

A number of studies have shown that although each tumor genome is unique, some alterations observed within the type of tumor/tissue are highly recurrent. For example, more than 90% of triple negative breast tumors carry a mutation in the *TP53* gene and display LOH on the chromosome 17 (Manie *et al*., 2009); almost all clear cell renal cell carcinomas have deletion in the 3p chromosome arm (Hagenkord *et al*., 2011); majority of low grade estrogen receptor positive ductal breast carcinomas have 16q deletion and half of them display 16p gain (Natrajan *et al*., 2009); numerous subtypes of cancers have amplification of 8q; etc. Some of these recurrent alterations were found to target certain tumor related genes as the minimal shared region of alteration clearly identifies a gene locus. Other recurrent alterations comprise the whole chromosome arm. In this case the list of possible functionally related targets is large and uncertain. On the other hand, it was shown that some genomic regions "prefer" to be gained or lost independent of the type of tumor (Beroukhim *et al*., 2010). The current understanding associates some of recurrent alterations with chromosome structure or function (for example, fragile sites, long repeats favoring chromosome recombination, formation of di-centric chromosomes leading to breaks and fusion cycles, interference of transcription and replication in large genes, etc.) (Bignell *et al*., 2010; Cassidy & Venkitaraman, 2012). Genomic alterations "accompanying" causative (driver) mutations or arising as a consequence of functioning due to extensive growth were referred as passenger events (Stratton *et al*., 2009). Thus, the main challenge in tumor genome interpretation is distinguishing alterations associated with the driver and the passenger events.

To conclude, for unraveling cancer genome complexity we need to annotate alterations in tumor genome in order to find recurrent amplifications, deletions, gains, losses, and LOH, which possibly target cancer related genes. We need to be able to characterize tumor ploidy and the level of rearrangements to describe cancer genomes as completely as possible.

3 Measuring Genetic Alterations by SNP Based Technique

The normal human genome contains two copies of each autosome (chromosomes 1-22) and two copies of the sex chromosome X in females (XX) or one copy of each of the sex chromosomes X and Y (XY) in males. These two alleles have paternal and maternal origins and are 99.9% identical in DNA sequence. A difference between two alleles is referred as a genetic variation and includes, in particular, Single Nucleotide Polymorphisms (SNPs) (Figure 3). We say that a SNP is homozygous if paternal and maternal alleles have identical nucleotides; the SNP is heterozygous if paternal and maternal alleles have different nucleotides.

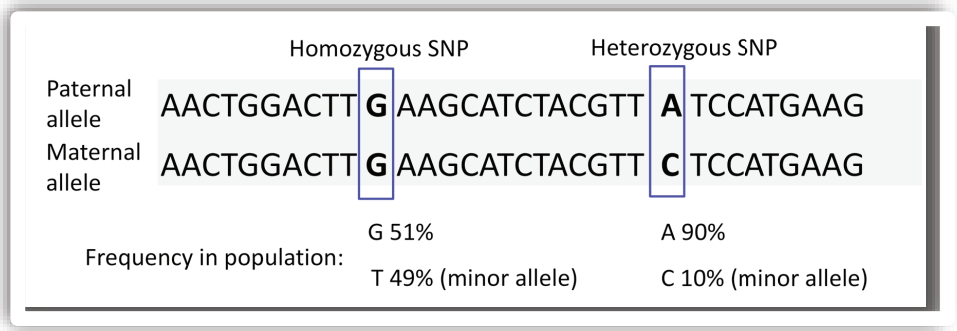


Figure 3: Single Nucleotide Polymorphism (SNPs) in individual genome and in population. Status of SNP in individual genome is defined as homozygous (two identical nucleotides) or heterozygous (distinct nucleotides). Single nucleotide variation is designated as a SNP, if its minor allele frequency exceeds 1% in at least one population; otherwise it is called variant or mutation.

Homozygous SNPs are called non-informative because they do not allow the two alleles to be distinguished. At the moment, near 30 million of SNPs are annotated in the dbSNP database (to be annotated as a SNP, variant frequency at the specific locus should exceed 1% in, at least, one population (Sherry et al., 2001)).

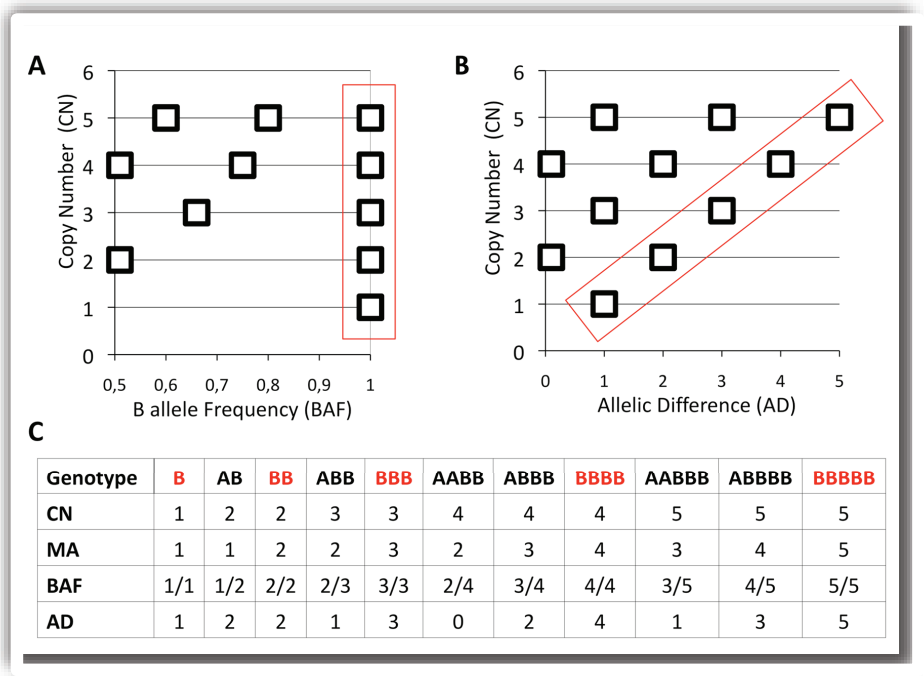
The cancer genome is characterized by abnormal genomic content, meaning, the number of chromosomes and their structure are different from the normal state. Variation in copy numbers is a result of losses of one of the two alleles (or both alleles, so called homozygous deletion, rarely comprising a large genomic region), gains of up to several copies of one or both alleles, and copy neutral LOH (two identical alleles, also called uniparental disomy) (Albertson *et al*., 2003). Which allele is lost or gained in a cancer genome is essentially unknown unless the parents of the patient are also genotyped (that is not a common practice in tumor genomic studies). That is why alleles are designated arbitrarily as A and B, and characterization of allelic content in tumor genome consists of indicating copy number (*CN*) and major allele (*MA*) counts (for example, $CN = 3$ and $MA = 2$ correspond to genotypes AAB and ABB, and could be described as "two identical alleles out of three"). Allelic content is characterized by B Allele Frequency (*BAF*) or Allelic Difference (*AD*), which represent the ratio of the major allele counts and the copy number or the difference between major and minor allele (*MI*) counts $(MI = CN - MA)$, respectively:

$$BAF = MA/CN = n_B^c/CN, \tag{1}$$

$$AD = MA - MI = n_B^c - n_A^c, \tag{2}$$

where *CN* is copy number, n_B^c, n_A^c are numbers of B and A alleles in cancer genome (historically, B allele counts are associated with the major allele counts). Plotting genomic states in two dimensions, namely, allelic content (x axis) and copy number (y axis) gives visual representation of all possible allelic contents and copy numbers, which could be observed in a tumor genome (Figure 4). Monoallelic genomic states (B, BB, BBB, etc., designated by red color in the table and red rectangles in Figure 4) represent genomic states with LOH.

Measuring genetic alterations in a cancer genome with SNP-based technology consists of measuring allelic contents in numerous SNP loci along the genome. If the technology provided perfect measurement



C

Genotype	B	AB	BB	ABB	BBB	AABB	ABBB	BBBB	AABBB	ABBBB	BBBBB
CN	1	2	2	3	3	4	4	4	5	5	5
MA	1	1	2	2	3	2	3	4	3	4	5
BAF	1/1	1/2	2/2	2/3	3/3	2/4	3/4	4/4	3/5	4/5	5/5
AD	1	2	2	1	3	0	2	4	1	3	5

Figure 4: All possible allelic contents for DNA copy number from 1 copy to 5 copies. **A.** Pattern of genomic states if allelic content is characterized by B allele frequency (*BAF*) (Illumina); **B.** Pattern of genomic states if allelic content is characterized by Allelic difference (*AD*) (Affymetrix SNP 6.0, CytoScanHD). **C.** Evaluation of *BAF* and *AD* from *CN* and *MA*. CN: copy number; MA: major allele counts; homozygous genomic states are shown in red; position of each pair (*CN, MA*) is represented on the graphs by the small squares (*MA*: x axis; *CN*: y axis); red rectangles indicate homozygous genomic states (LOH); only $CN \leq 5$ are shown, however, the structure could be easily extended to the higher *CN* levels.

of the number of A and B alleles in each SNP locus, each genomic region would be characterized by the pair of values (*CN, MA*) corresponding to one of the genomic states shown in Figure 4. However, there is no such perfect technology available yet. Whole genome profiles of genetic alterations provided by SNP-arrays (or NGS) represent relative copy number variation (*CNV*) and allelic imbalance (*AI*) profiles (Figures 2C & 5, see Appendix A for SNP array platforms description). Allelic imbalance profile characterizes allelic content by *BAF* or *AD*, depending on the normalization applied to the measured profile. These profiles are affected by technological noise and experimental variation (biological sample preparation); furthermore, the measured tumor sample often has significant admixture of the normal genome (so called normal contamination, as the single cell technology is not yet a common place, the measured tumor sample is usually a mixture of tumor and normal stromal cells). Thus, annotation of measured *relative* variation profiles with *absolute* copy numbers and major allele counts represents a problem statement for further data mining.

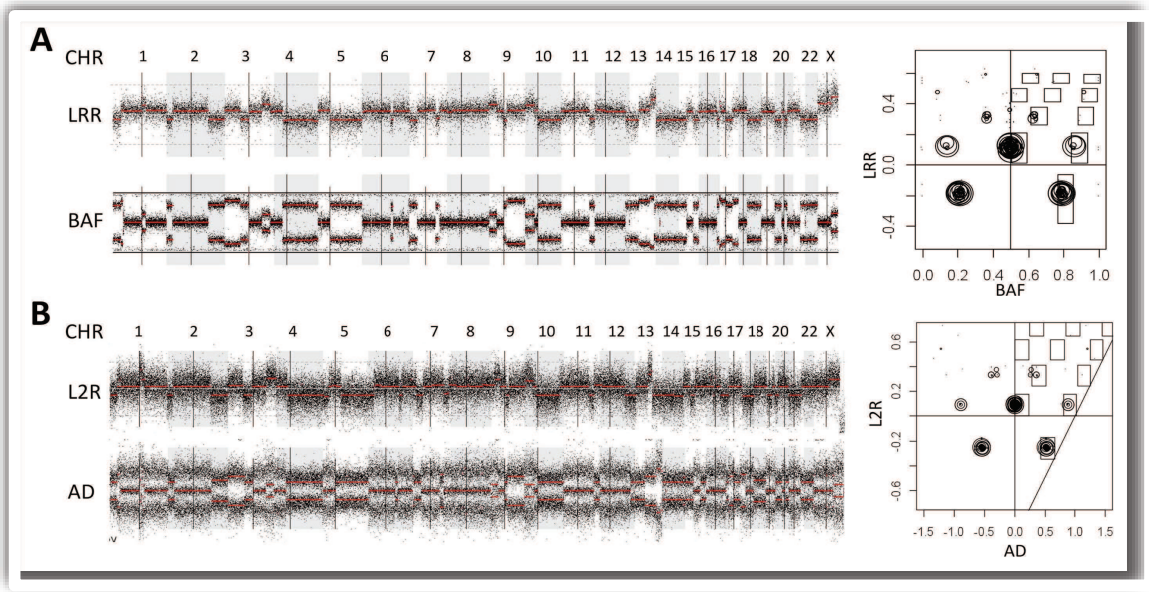


Figure 5: The whole genome profiles of genomic rearrangements for one primary breast tumor measured on two platforms and their GAP plots. **A.** Illumina 300K SNP array profiles (left) and corresponding GAP plot (right): Copy Number Variation (*CNV*) profile is represented by log R ratio (*LRR*), Allelic Imbalances (*AI*) are represented by B allele frequency (*BAF*); **B.** Affymetrix SNP 6.0 profiles (left) and corresponding GAP plot (right): *CNV* profile is represented by Log 2 Ratio (L2R), *AI* are represented by Allelic Differences (*AD*). GAP plots display combined side-view projections of segmented *CNV* and *AI* profiles; each region of genome is represented by two symmetrical circles; position of a circle is defined by median of *CNV* (y axis) and (lower) mode of *AI* (x axis); the size of a circle is proportional to the segment size; rectangles indicate the copy number / allelic content patterns.

4 Genome Alteration Print (GAP) for Mining Genetic Alterations

Here we describe a data mining technique developed for SNP arrays, which allows extraction of absolute copy numbers and allelic contents from the whole genome *CNV* and *AI* profiles. The method is based on the structure denoted by Genome Alteration Print (GAP) (Popova *et al*., 2009). GAP is a two dimensional representation of segmented *CNV* and *AI* profiles of a measured tumor sample and characterizes the spectrum of rearrangements presented in the tumor (Figure 5, right panels). In order to build the GAP of a tumor, *CNV* and *AI* profiles are segmented; the values are smoothed within the segments by median (*CNV*) and mode (*AI*); the segments are plotted on the $AI \times CNV$ plane as the circles of the radius proportional to the segment size (see Appendix B for details).

GAP patterns of the breast tumor sample measured on the Illumina and Affymetrix platforms (Figure 5, right panels) resemble the patterns of copy numbers and allelic contents represented by *BAF* and *AD* respectively (Figure 4). Tumor GAPs differ from the copy number / allelic content plots by (1) uneven distribution of *CN* levels along the vertical axis (due to the original *CNV* profiles represented in the log-

scale: Log R Ratio and Log2Ratio in Illumina and Affymetrix platforms, respectively); and (2) a shift on the horizontal axis toward 0.5 or 0 of *BAF* or AD, respectively (due to the normal contamination present in the tumor).

In order to account for normal contamination in copy number / allelic content template we considered the sample containing proportion p of normal cells and *(1-p)* of tumor cells. Adding proportion p of normal heterozygous (AB) signal to the copy numbers and allelic contents we easily can obtain the model patterns accounting for the normal contamination (Figure 6):

$$CN^p = (1-p)\cdot CN + 2p, \tag{3}$$

$$BAF^p = \frac{(1-p)\cdot n_B^c + pn_B^n}{(1-p)\cdot CN + 2p}, \tag{4}$$

$$AD^p = (n_B^c - n_A^c)\cdot(1-p) + (n_B^n - n_A^n)\cdot p, \tag{5}$$

where CN is copy number; $n_B^c, n_A^c, n_B^n, n_A^n$ are numbers of B and A alleles in (c)ancer and (n)ormal genomes ($n_B^c = [CN/2], \ldots, CN$; $n_A^c = CN - n_B^c$; $n_B^n = 1$, if $n_B^c < CN$; $n_B^n = 1,2$, if $n_B^c = CN$). These two patterns represent the model templates for mining genetic alteration measured by SNP arrays. It is worth noting that LOH are now represented by two sets of genomic states (designated by red quadrangles in the Figure 6) and the distance between the two LOH sets characterizes the level of normal contamination. Because of normal contamination SNPs with acquired homozygosity had addition of normal heterozygous signal, which shifted their *AI* towards heterozygous states (for example, $(1-p)\cdot$ BB + $p\cdot$AB). SNPs homozygous in the germline did not change their position ($(1-p)\cdot$ BB +$p\cdot$BB = BB) (compare Figures 4 & 6). Thus, moderate level of normal contamination could help distinguish segments with germline and acquired homozygosity. However, increasing the normal contamination proportion leads to the pattern shrinkage towards 2 copies, in extremity (when p is close to 1) converging to a normal genome. Acquired and germline homozygosities are not distinguishable in the case of pure tumor sample or cell line.

All the GAPs obtained for a large set of SNP-arrays measuring breast, prostate, lung and some other tumors were found resembling the model patterns of copy numbers and allelic contents defined above. However, the dynamics of change in measured *CNV* profiles was observed to have high degree of variation from sample to sample. To account for this experimental variation we introduced the coefficient of *CNV* contraction q, which is supposed to adjust the difference in scales between model *CN* and measured *CNV* profiles (Figure 6).

The common way of annotating a SNP array profile consisted thus in finding the model template which fits best to the measured GAP. Finding the model template means defining three parameters, namely, proportion of normal contamination (p), coefficient of *CNV* contraction (q), and position of 2 copies on the *CNV* scale (C), such that the model template maximally corresponds to the measured GAP. GAP plane is annotated with *CN*s and *MA*s from the model template and genomic segments are annotated accordingly (Figure 7).

A quality of fitness of the measured GAP to the model template is assessed by the genome coverage in terms of number of SNPs that are explained by the model. However, fitting regular structures might not result in one unique best solution. To avoid ambiguity, we filter redundant templates and choose interpretation with the lowest possible *CN* set and minimal possible normal contamination. All parameters of the model template were set up and extensively tested on a large cohort of tumor genomes and under manual control of recognition quality (visual representation of tumor GAPs and model templates,

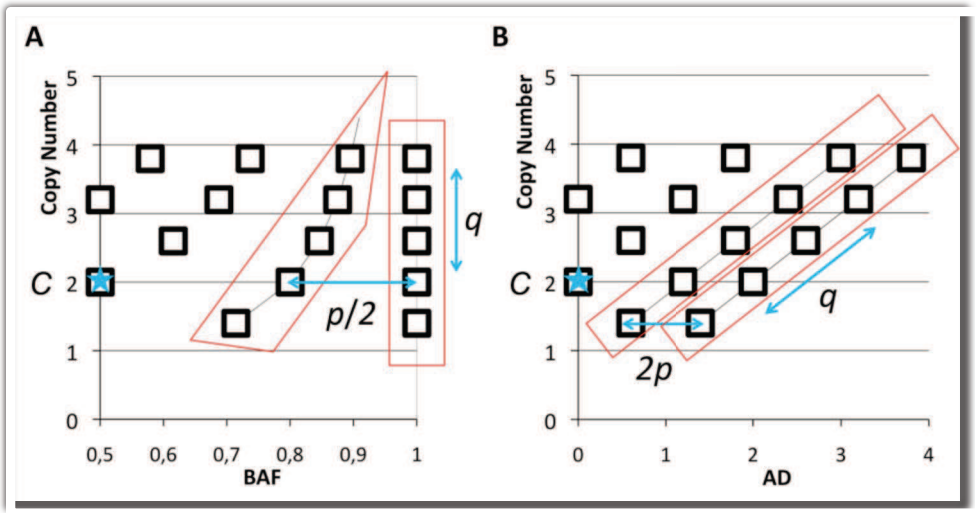


Figure 6: Model templates and three parameters modifying the model (indicated by the blue arrows (p,q) and stars (C)) to fit to a tumor GAP. **A.** Template for Illumina platform. **B.** Template for Affymetrix platform. p is normal contamination proportion; q is coefficient of *CNV* contraction; the stars correspond to 2 copy centering position on *CNV* scale (parameter C). Homozygous states are designated by red quadrants.

as shown in Figure 7, allows judging on correct or incorrect recognition). Current implementation of the GAP method provides near 80% rate of correctly found model templates. Moreover, complicated cases could be annotated manually by introducing three parameters, which (according to a user) provide a proper correspondence between tumor GAP and the model template. More details and algorithm of recognition are presented in Appendix C; R scripts and full details of the application are available at (`http://bioinfo-out.curie.fr/projects/snp_gap/`).

5 Validation of the GAP Method

Our group used the GAP method to process numerous SNP array profiles mainly of breast tumors. As a result of the recognition procedure we obtained the level of normal contamination (p), and the profile of segmental copy numbers and allelic contents (segmental genotypes) for each tumor (one example is shown in Figure 2C). With the current implementation recognition of absolute copy number ranged from 0 to 8 copies (all segments with the copy number variation exceeding 8-copy level were ascribed 8-copy status). Thus, 22 possible segmental genotypes were discriminated (copy number / major allele count): B (1/1); BB (2/2) and AB (2/1); BBB (3/3) and ABB (3/2); BBBB (4/4), ABBB (4/3) and AABB (4/2); etc.

The way we model the proportion of normal contamination (p) for *BAF* evaluation has been already described in a number of publications and confirmed by the dilution series and computer simulations (Nancarrow *et al*., 2007; Staaf *et al*., 2008a). To verify our evaluation of the normal contamination and *AD* we considered the dilution series of the lung cancer cell line (H-1395) available on Affymetrix SNP 6.0 plat-

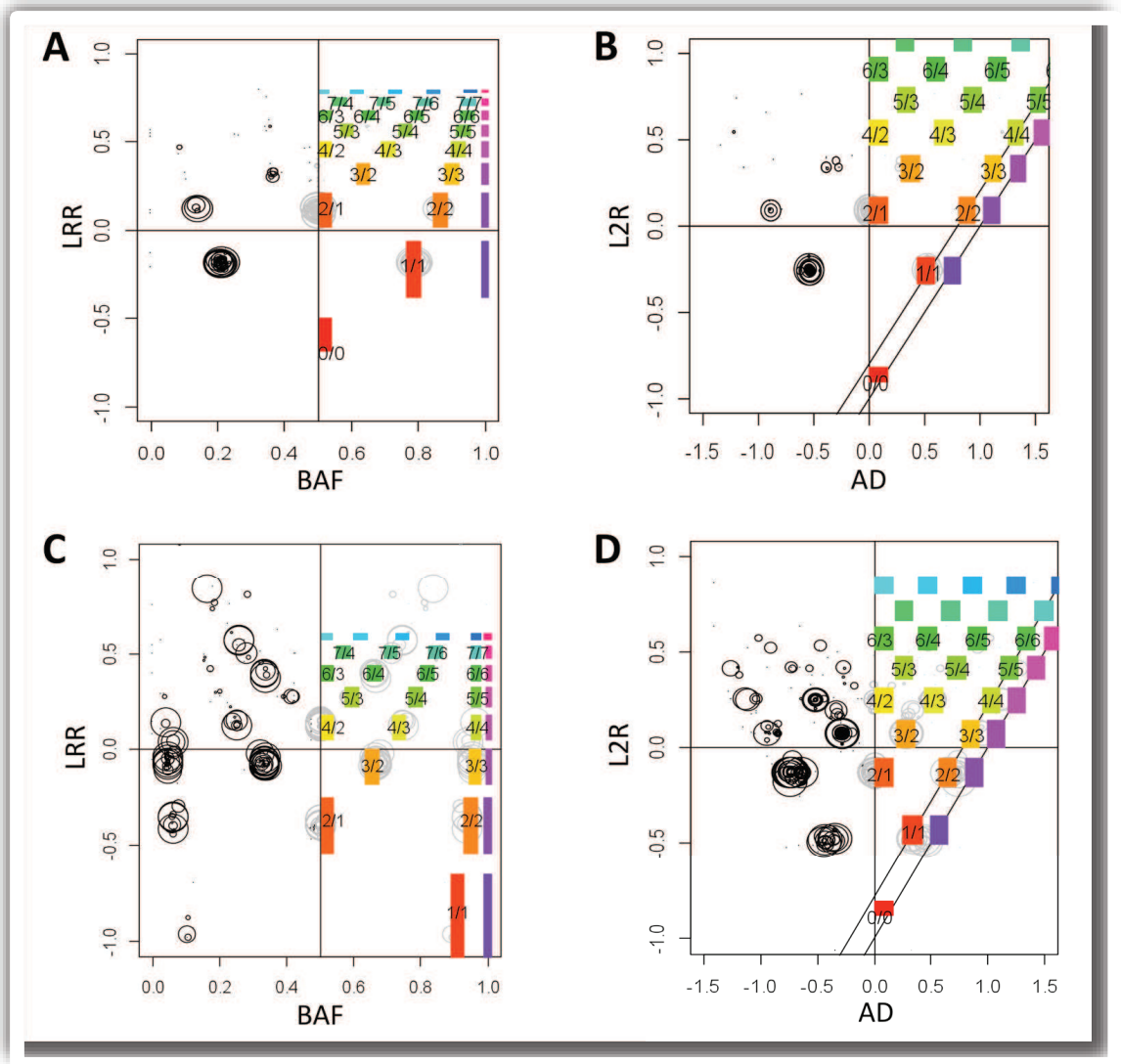


Figure 7: GAP plots and results of automatic recognition procedure of copy number and allelic content in three primary breast tumors. The best fitting model template is shown by colored rectangles, designated by the ratio: *CN/MA*. **A.** and **B.** GAPs and recognition templates for the breast tumor sample shown in Figure 5 on Illumina (**A**) and Affymetrix (**B**) platforms; **C.** and **D.** Examples of model fitting for two over-diploid breast tumors on Illumina (**C**) and Affymetrix (**D**) platforms.

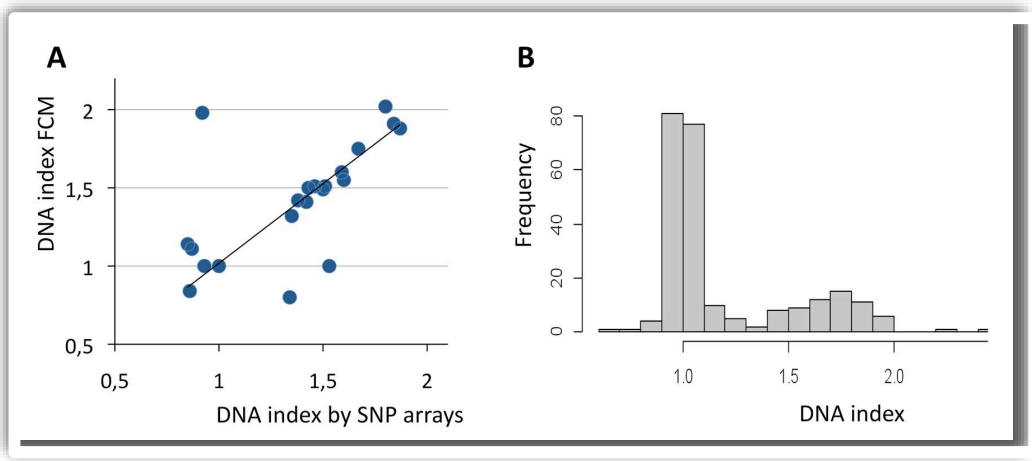


Figure 8: DNA indexes for the breast tumor genomes. **A.** Validation of copy number attribution based on the GAP method: each point represents DNA index of a primary breast tumor measured by SNP array (x-axis) and by flow cytometry (FCM) (y-axis). **B.** Distribution of DNA indexes obtained from the SNP arrays of a big cohort of breast tumors.

form (GEO GSE29172) (Rasmussen *et al.*, 2011). For the proportions 0, 0.3, 0.5, and 0.7 of the normal cells admixture, we obtained the model parameter $2p$ to be 0.1, 0.44, 0.6, and 0.76, respectively (see Figure 6 B). As far as we can conclude from this small series, $2p$ could be considered as a reasonable upper estimation of the normal contamination in Affymetrix SNP 6.0. It is worth noting that (i) the model parameter p significantly underestimated normal contamination; (ii) 70% contaminated tumor displayed almost flat merely recognizable profile.

In order to validate our SNP array interpretation in the total set of copy numbers, we compared DNA contents inferred from recognized copy number profile with those provided by flow cytometry (FCM). DNA index provided by FCM characterizes DNA content of tumor genome relative to normal diploid genome, for which DNA index is set to 1. DNA indexes from SNP arrays were estimated by averaging segmental copy numbers divided by 2. DNA indexes inferred from SNP-arrays are in close correspondence with actual tumor DNA indexes measured by FCM analysis for 20 out of 23 tested samples of breast carcinomas (Figure 8A).

We also validated our recognition procedure by comparing inferred number of chromosomes with known karyotypes or SKY data of 25 breast cancer cell lines (`http://www.lgcstandards-atcc.org/`, `http://www.path.cam.ac.uk/~pawefish/`). The number of chromosomes was estimated by the sum of the copy numbers detected at the pericentric regions. The status of the pericentric region of each chromosome arm was defined by the corresponding juxta-centromeric segment when the latter contained 500 SNPs or more (Affymetrix SNP 6.0). When not measurable, missing values were substituted by the modal copy number of the considered chromosome arm (3.4±2.2 out of 41 chromosome arms per genome were substituted in the series).

Error rate was less than 2 chromosomes per sample (1.7±2.3) in the considered set of breast cancer cell lines (Table 1). To conclude, copy number recognition by the GAP method results in correct predictions

in the majority of tested cases.

6 Other Approaches to Mine SNP Arrays and General Problems

Numerous studies have contributed to better understanding cancer genomics and in particular mining SNP array measurements (Assie *et al*., 2008; Attiyeh *et al*., 2009; Gardina *et al*., 2008; Lamy *et al*., 2007). However, it is the *copy number / allelic content* structure (Figure 4 & 6) standing behind the SNP-based genomic profiling (with linear or non-linear transformation depending on the physical properties of a measuring device and a way of data normalization) that should govern any copy number recognition procedure. Simplicity of copy number / allelic content pattern implies various possible approaches for copy number recognition. Existing integrated approaches includes Hidden Markov Models (PICNIC (Greenman *et al*., 2010), GPHMM (Li *et al*., 2011)) and pattern recognition strategies (GAP (Popova *et al*., 2009), ASCAT (Van Loo *et al*., 2010), TAPS (Rasmussen *et al*., 2011)), Absolut (Carter *et al*., 2012).

Quality of a SNP array mining procedure should be estimated by correspondence between reported tumor characteristics (such as *CN* and *MA* profiles, number of breakpoints, percent of tumor cells) and the actual tumor states in a large set of samples and under various conditions. Problems that could affect recognition potential of all the methods mentioned above could be subdivided into three groups: (i) technical problems associated with quality of measured SNP array profile and physical properties of a measuring device; (ii) tumor heterogeneity; (iii) ambiguity in interpretation of SNP array profiles.

The first class of problems associated with quality of measured SNP array profile includes various effects of unknown origin resulted in non-specific variation in *CNV* profile, such as high percent of outlying values; low-frequency fluctuations, so called waves in *CNV* profile; high-frequency saw-tooth waves, not associated with *CN* change; etc. *AI* profiles appeared to be more stable and not affected by these variations. Thus, data mining techniques more relying on *AI* profiles could be more successful and give more reliable estimation of tumor rearrangements in these cases. However, low quality profiles are more exceptional than regular cases in the latest generation of SNP arrays.

The problems associated with physical properties of measuring devices also include the signal saturation effects, which hamper accurate detection of high copy number levels.

The second class of problems is intra-tumor heterogeneity, which include normal contamination and subclonal structure (Shipitsin *et al*., 2007). Although presumably arisen from a single cell (monoclonal proliferation), cancer progression leads to the sub-populations bearing different genomic alterations (subclones) coexisting in most tumor samples (Navin *et al*., 2010). Variation observed by SNP-array in a tumor genome reflects genomic alterations shared by all tumor cells and subclonal events shared by only subpopulation of tumor. *CN* and *MA* status of an alteration specific to sub-clones is generally indefinable from SNP arrays as the measured signal reflects the sum of unknown subclonal signals in unknown proportions. Intermediate *CNV* levels due to subclonal events could be interpreted as actual *CN* level and overall pattern of copy number set could be thus overestimated.

Another type of problems in mining SNP arrays is associated with classes of tumor genomes not distinguishable by SNP array profiles. This follows from the relative nature of measured profile and regular structure of allelic content / copy number pattern. For example, there is no way to distinguish the genomes displaying AB and B states versus AABB and BB ones, if there are no other events (such as ABB) in the latter case. However, this is probably a rare situation (as we can conclude at least for breast cancer genomes).

Cell line	GAP	ATCC	ATCC min	ATCC max	SKY	Error
HCC2157	63	75	65	79		2
HCC38	75	75	65	79	78	0
Hs-578T	61	59	50	77		0
MDA-MB-157	55	53	52	69	62	0
MDA-MB-231	63	64	52	68		0
MDA-MB-468	66	64	60	67	54	0
MDA-MB-435	58	56	55	62		0
CAMA-1	70	80	68	83		0
BT-483	81	72	46	130		0
MCF-7	84	82	66	87		0
MDA-MB-361	60	56	54	61	56	0
BT-549	83	78	73	80		3
MDA-MB-453	96	90	87	91	90	5
HCC1143	84				84	0
HCC1187	66				64	2
HCC1937	91	100			88	0
MDA-MB-436	40	45				5
HCC1569	99				132	outlier
HCC1599	73				64	9
HCC1954	101				97	4
HCC70	99				95	4
PMC-42	61				64	3
184B5	48	47				1
SKBR.3	85	84			80	1
ZR-75-1	71	72			75	1

Table 1: Number of chromosomes in the cell lines according to GAP estimation, ATCC description and SKY images. Error was calculated as minimal absolute difference between GAP estimation and ATCC and/or SKY numbers.

Obviously, all aforementioned methods performed well on the high quality SNP array profiles with moderate level of normal contamination, low complexity and a good contrast. The differences between the methods start to arise when a tumor sample or its measured profile have compromised quality. PICNIC has problems with normal contamination as it was designed for cell lines, GPHMM tolerated up to 90% of normal contamination well but might be sensitive to sub-clones, GAP usually overestimates the number of breakpoints (an additional smoothing procedure is needed) and looses recognition quality when normal contamination exceeds 70%.

Accurate evaluation of performance of several SNP array mining approaches with respect to the normal contamination and profile complexity (using simulated data) has been done recently by an independent group (Mosen-Ansorena *et al*. 2012). The GAP method showed the best performance among all approaches tested. We consider that the major advantage of the GAP method is its visualization based strategy which allowed setting up an automatic recognition procedure and fitting a number of hidden parameters based on a large set of tumor genomes and under clear-cut manual control.

7 Next Generation Sequencing (NGS)

The recently developed Next generation sequencing (NGS) technology – through whole-genome, whole-exome and whole-transcriptome approaches – provides the most complete view on the cancer genome and has a potential to uncover all spectrum of somatic variation, deduce tumor history and sub-clonal structure (Greenman et al., 2012; Meyerson *et al*., 2010). Extracting absolute copy numbers belongs to the "must have" package for future NGS processing pipelines providing cancer genome characterization from point mutations to large-scale chromosomal rearrangements. Here, we want to outline current methodological and technological issues related to the large-scale absolute copy number and allelic content recognition based on NGS data.

NGS technology consists of shearing tumor genomic DNA into small fragments (200-4000bp), amplifying, and reading fragments using Illumina, SOLiD, PacBio or IonTorrent Analyzers (Metzker, 2010). The single-end or paired-end reads are mapped back to the reference genome, providing coverage of each genomic position by a number of short fragments (reads). Window-averaged number of reads along the genome characterizes *CNV* profile; reads covering known annotated SNP positions could be used to obtain *BAF* profile (Nielsen *et al*., 2011). As soon as *CNV* and *BAF* profiles are obtained, the same strategy of tumor genome annotation that is described for SNP arrays could be applied (Boeva *et al*., 2012). Due to potentially linear relationship between the number of reads and actual DNA copy number, segmented *CNV* and *BAF* profiles should follow the pattern shown in Figure 6A, where coefficient q is no more necessary and dynamics of copy number layers itself could provide the level of normal contamination (see equation (3)) (Boeva *et al*., 2011; Gusnanto *et al*., 2012).

However, at the moment a number of difficulties exist in extracting tractable *CNV* and *BAF* profiles from NGS data. Firstly, the number of reads along the genome depends heavily on regional GC content and read mappability, which result in large unspecific variations in *CNV* profile. Secondly, the reference allele usually has more reads due to asymmetric mapping procedure and *BAF* calculation results in a very noisy profile. Several normalization strategies using a matched normal sample or correcting for GC-content and/or mappability have been suggested (Boeva *et al*., 2011). In the case of targeted sequencing (for example, whole exome sequencing), in addition to mappability and GC-content, a bias resulting from uneven capture in different targeted regions should also be corrected (Li *et al*., 2012; Sathirapongsasuti *et al*., 2011). Mapping reads on the genome indexed with SNPs and exclusion of low quality reads before *BAF* calculation could improve genotype prediction.

Alternatively, huge numbers of reads provided by NGS could help annotating alterations by sequencing-specific approaches. In the case of paired-end sequencing, predicted alterations can be supported and clarified by the fragments encompassing the junction, which could be detected using abnormal reads mapping (Medvedev *et al*., 2010). Moreover, "split-read" mapping (i.e. partial mapping of reads encompassing a junction) can provide base-pair resolution of alterations (Wang *et al*., 2011).

To conclude, accurate evaluation of copy numbers and allelic contents directly from the read counts is still not entirely implemented into powerful bioinformatics tools; mostly because of the short time frame since NGS data are available for the scientific community. Bioinformatics developments and NGS technology itself advance at a staggering rate that will undoubtedly result in new computational solutions in the future.

8 Detection of Tumor Ploidy and Attribution of Gains and Losses

Extracting biologically relevant information from a recognized tumor genome profile represents an important issue. Here we address the detection of tumor ploidy and annotation of gains and losses with the example of breast tumor genomes.

According to the genomic content of a tumor cell, one could distinguish diploid, tetraploid, near-diploid, over-diploid, near-tetraploid or simply aneuploid (highly differing from diploid or tetraploid) status. We have suggested a way to ascribe tumor ploidy status based on the actual genomic content distribution in a large cohort of breast tumors. Inferred DNA index in the cohort of 250 breast tumors showed a bimodal distribution (Figure 8B) similar to those demonstrated for the genomes of various types of cancers (Storchova & Kuffer, 2008). This observation supports the hypothesis of frequent duplication of the whole genome during cancer progression, which explains bimodality (Carter *et al*., 2012; Storchova & Kuffer, 2008). Two modes of the DNA index in the breast cancer cohort were found to be 1 and 1.7 units with the cut-off distinguishing two modes approximately equal to 1.3 (1 corresponds to the normal diploid genome). Thus, tumor genomes with DNA index less than 1.3 were considered to have a ploidy of two and are called "near-diploid genomes"; tumor genomes with DNA index equal or higher than 1.3 were considered to have a ploidy of four and are called "near-tetraploid genomes". Distribution of genomic states in near-diploid and near-tetraploid genomes was generally consistent with attributed ploidy. Near-diploid genomes had genomic DNA mainly presented in 1, 2, 3 copies (B, AB, ABB segmental genotypes) while near-tetraploid genomes had the well represented 2, 3, 4, 5-copy layers (BB, ABB, AABB, AABBB genotypes). Moreover, a significant bias observed in distribution of genomic states (such as high frequency of BB compared to AB states) in the small series of near-tetraploid basal-like breast tumors evidences the late (after a number of alteration events) whole genome duplication (Popova *et al*., 2009). The same conclusion for other types of tumors has been published recently (Carter *et al*., 2012).

Relative genomic alterations such as gains, losses, amplifications should be called according to the ploidy of a tumor sample. We suggested the following procedure: for near-diploid tumors copy loss, gain, amplification is called for the segments with ≤ 1, ≥ 3 and ≥ 4 copies, respectively. For near-tetraploid tumors copy loss, gain, amplification is called for the segments with ≤ 2, ≥ 6 and ≥ 8 copies, respectively. Thus, genomic segments in either of 3, 4, 5-copy states are considered to be not altered in a near-tetraploid genome. Indeed, according to the hypothesis of the late whole genome duplication, 3 and 5 copies represent "secondary" alteration events. Moreover, small relative changes in copy numbers are unlikely to strongly affect gene expression. An additional evidence to support our approach is the similar rate of gains and losses obtained for near-diploid and near-tetraploid basal-like breast tumors. With this approach the most of known recurrent alterations in total set of basal-like breast tumors displayed 80–100% recurrence rate (as compared to 50–60% reported based on the array CGH).

It is worth noting that genetic pathways in other types of tumors could be different from that described for breast cancers. For example, neuroblastoma is known to frequently display genomic content close to the near-triploid state (DNA index of 1.5), implying a high rate of whole chromosome gains or some other mechanisms of genome transformation (Brodeur, 2003). For this reason, DNA index cut-off for ploidy attribution (as well as annotation of gain and losses) obtained for the breast tumors might not be suitable for the other types of tumors. However, we believe that the way of annotating genetic alteration should reflect (as far as possible) tumor genome evolution.

9 Conclusions

In this chapter we present an approach for mining genetic alterations measured by the SNP-based techniques. The main advantages of the GAP method are its visual and contextual simplicity and natural interpretation. Moreover, results of automatic recognition could be easily monitored manually, which increases the confidence in evaluating important cases. Systematic investigation, description and comparative analysis of cancer genomes by SNP arrays and NGS are currently ongoing. Clinical trials have been designed to translate some of the genomic findings into clinic to improve cancer diagnostics and treatment. We hope to see encouraging results in the near future.

Acknowledgements

The authors would like to thank Dr. Amaury Dumont, Jaydutt Bhalshankar and reviewers for productive comments and accurate proofreading of the manuscript.

Appendix A SNP-arrays Platforms and Normalization (Technical Note)

Two major platforms are present in the market: Illumina and Affymetrix, providing SNP arrays containing up to million(s) of SNPs. Having different technological basis, raw data normalization of Illumina and Affymetrix platforms is performed based on the calibrating set of HapMap individuals. Both platforms provide the software for primary data normalization (BeadStudio Illumina, Genotyping Console Affymetrix), producing *CNV* and *AI*.

The latest evaluation of the SNP-arrays on the Illumina platform is essentially similar to all previous versions (only the number of SNPs was increased and significant number of *CN* probes were included; *CN* probes are intended to measure germline *CNV s*, which we are not considering here). All conclusions presented here for Illumina platform equally concern all versions of Illumina SNP arrays. In contrast, the latest versions of Affymetrix platform, Affymetrix SNP 6.0 and CytoScan HD, producing higher quality data, compared with the previous version of SNP-arrays (Affymetrix 100K, 250K, 500K) were subjected to specific normalization. All conclusions presented here were evaluated for Affymetrix SNP 6.0 platform and successfully tested on CytoScan HD with some minor modifications. Normalization of previous versions of Affymetrix (except SNP chip 5.0) could be obtained by Aroma package `http://www.aroma-project.org` and should be treated within the Illumina framework.

For the Illumina platform *CNV* profile is represented by the Log R ratios (*LRR*), which are the log-transformed ratios of experimental and normal reference SNP intensities, centered at zero for each sample. Allelic imbalances are represented by *BAF*, which are the normalized proportions of the B allele signal. Normalization procedure designed by Illumina produces not symmetrical *BAF* profiles, thus, we recommend to symmetrize *BAF* profile by applying the tQN algorithm for quantile normalization (Staaf *et al*., 2008a).

For the Affymetrix platform *CNV* profile is represented by the Log 2 ratios (*L2R*), which are the log-transformed ratios of experimental and normal reference SNP intensities. Allelic imbalances are represented by the *AD*, which represents the difference of A signal and B signal each standardized with respect to their median values in the reference.

Appendix B *CNV* and *AI* Profiles Segmentation

The circular binary segmentation (CBS) algorithm (DNAcopy package, Bioconductor) (Venkatraman & Olshen, 2007) was used for segmentation of all profiles presented here. There exist numerous methods of optimal *CNV* profile segmentation, and any of them could be used instead of CBS. However, the GAP method tolerated "false positive" breakpoints in either profile well, while "false negative" breakpoints could be misleading for pattern recognition. Thus, segmentation method should be as sensitive as possible but could tolerate less specificity.

BAF and *AD* profiles are supposed to be symmetric, thus both profiles were transformed in order to reduce redundancy: $BAF = 0.5 + abs(BAF - 0.5)$ and $AD = abs(AD)$.

BAF and *AD* profiles represent two mode distributions in any genomic segment with heterozygous genotype (in tumor or germline): lower mode corresponds to heterozygous SNPs (informative) and higher mode corresponds to homozygous SNPs (non-informative). Because segmentation of a two mode profile is less trivial than that of a one mode profile, we attempted to reduce *BAF* and *AD* profiles to one mode. In the case of *BAF*, non-informative homozygous SNPs do not depend on copy number and could be filtered out based on the threshold ($BAF > 0.97$, Illumina 300K, as suggested, for example, in (Staaf *et al*., 2008a)).

In the case of *AD*, non-informative homozygous SNPs depend on copy number and simple threshold is not applicable. However, due to the fact that *AD* values for homozygous genomic states equal to the corresponding copy numbers, we applied *AD* filtering based on the *L2R* profile. To filter out homozygous mode, we defined a profile, representing a joint neighborhood of 2^{L2R}. Those points in *AD* profile which fall into the joint neighborhood of 2^{L2R} were filtered out.

Filtered *AI* (*BAF* or *AD*) profiles were segmented by CBS algorithm. *CNV* and *AI* breakpoints were united, providing segmentation of the genomic profile into alteration units with presumably stable copy number and allelic content state.

Appendix C Algorithm of Automatic Recognition of the GAP Pattern

1. Consider GAP of a tumor sample to be a set of genomic segments (alteration units, obtained after segmentation of *CNV* and *AI* profiles) each characterized by the three values: the median of *CNV*, the (lower) mode of *AI*, and the length in *SNP* counts:

$$\mathrm{GAP} = \{CNV_i, AI_i, L_i\}_{i=1,\ldots,n}.$$

2. For each p (proportion of normal contamination), q (coefficient of experimental variation), and model centering C (corresponding to the position of 2 copies on the *CNV* axis) we define a model $\mathrm{GAP}^{\mathrm{T}}$, where each CN/MA pair ($CN = 1,\ldots,5$; $MA = [CN/2],\ldots,CN$) is characterized by the model (CNV_{CN}, $AI_{CN/MA}$) values:

$$\mathrm{GAP}^{\mathrm{T}}_{\mathrm{Illumina}} : LRR_{CN} = C + q \cdot (\log(CN) - 1),$$
$$BAF_{CN/MA} = \frac{(1-p)\cdot n_B^c + p \cdot n_B^n}{(1-p)\cdot CN + 2p},$$

$$\mathrm{GAP}^{\mathrm{T}}_{\mathrm{Affymetrix}}: L2R_{CN} = C + q\cdot\left((CN-1)^{3/4}-1\right),$$
$$AD_{CN/MA} = \left(L2R_{CN} + 1 - \left(2 - n_B^n\right)\cdot 2p\right)\frac{2n_B^c - CN}{CN},$$

where $CN = 1,\ldots,5$; $MA = n_B^c = [CN/2],\ldots,CN$; $n_B^n = 1$, if $n_B^c < CN$; $n_B^n = 1,2$, if $n_B^c = CN$; approximation for *L2R* values was found experimentally, by considering distributions in a large set of Affymetrix SNP 6.0 arrays.

Each $(CNV_{CN}, AI_{CN/MA})$ pair is included into the $\mathrm{GAP}^{\mathrm{T}}$ with a certain rectangle neighborhood, providing a model recognition template (similar to ones in Figure 6).

3. We define a grid in the parameter space: $p \in [0;0.7]$ ($2p \in [0;0.7]$ for $\mathrm{GAP}^{\mathrm{T}}_{\mathrm{Affymetrix}}$), $q \in [0.1;1]$, $C \in [-1;1]$, with the step = 0.05.

4. For each set of parameters (p,q,C) we calculate goodness-of-fit criterion: $K_{p,q,C} = \sum_{i,(CNV_i,AI_i)\in\mathrm{GAP}^{\mathrm{T}}} L_i$, $(CNV_i, AI_i) \in \mathrm{GAP}^{\mathrm{T}}$ if (CNV_i, AI_i) falls into any model rectangle defined by $(CNV_{CN}, AI_{CN/MA})$ pair from $\mathrm{GAP}^{\mathrm{T}}$.

5. We order parameter sets by the criterion value and choose the interpretation with maximal goodness-of-fit: $(p^*, q^*, C^*) = \arg_{\max}(K_{p,q,C})$, after filtering out "redundant" interpretations (in particular, such that superimposition of $\mathrm{GAP}^{\mathrm{T}}$ and tumor GAP results in empty acquired homozygosity states, etc). This procedure results in approximately 80% rate of correctly recognized models.

6. We annotate the segments from the tumor GAP by the closest *CN* and *MA* from the model template (for tumor annotation we enlarge the model $\mathrm{GAP}^{\mathrm{T}}$ to 8 copies).

7. We smooth resulting *CN*/*MA* profiles and exclude redundant breakpoints. R scripts are available at `http://bioinfo-out.curie.fr/projects/snp_gap`. Due to regular updating, some details or constants in implemented code may be slightly different from those described above.

References

Albertson, D. G., Collins, C., McCormick, F. & Gray, J. W. (2003). Chromosome aberrations in solid tumors. Nat Genet 34, 369-76.

Assie, G., LaFramboise, T., Platzer, P., Bertherat, J., Stratakis, C. A. & 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, 903-15.

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. & 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, 276-83.

Beroukhim, R., Mermel, C. H., Porter, D., Wei, G. *et al.* (2010). The landscape of somatic copy-number alteration across human cancers. Nature 463, 899-905.

Bignell, G. R., Greenman, C. D., Davies, H., Butler, A. P. *et al.* (2010). Signatures of mutation and selection in the cancer genome. Nature 463, 893-8.

Boeva, V., Popova, T., Bleakley, K., Chiche, P., Cappo, J., Schleiermacher, G., Janoueix-Lerosey, I., Delattre, O. & Barillot, E. (2012). Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics 28, 423-5.

Boeva, V., Zinovyev, A., Bleakley, K., Vert, J. P., Janoueix-Lerosey, I., Delattre, O. & Barillot, E. (2011). Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. Bioinformatics 27, 268-9.

Brodeur, G. M. (2003). Neuroblastoma: biological insights into a clinical enigma. Nat Rev Cancer 3, 203-16.

Carter, S. L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P. W., Onofrio, R. C., Winckler, W., Weir, B. A., Beroukhim, R., Pellman, D., Levine, D. A., Lander, E. S., Meyerson, M. & Getz, G. (2012). Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol 30, 413-21.

Cassidy, L. D. & Venkitaraman, A. R. (2012). Genome instability mechanisms and the structure of cancer genomes. Curr Opin Genet Dev 22, 10-3.

Gardina, P. J., Lo, K. C., Lee, W., Cowell, J. K. & Turpaz, Y. (2008). Ploidy status and copy number aberrations in primary glioblastomas defined by integrated analysis of allelic ratios, signal ratios and loss of heterozygosity using 500K SNP Mapping Arrays. BMC Genomics 9, 489.

Grander, D. (1998). How do mutated oncogenes and tumor suppressor genes cause cancer? Med Oncol 15, 20-6.

Greenman, C. D., Bignell, G., Butler, A., Edkins, S., Hinton, J., Beare, D., Swamy, S., Santarius, T., Chen, L., Widaa, S., Futreal, P. A. & Stratton, M. R. (2010). PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data. Biostatistics 11, 164-75.

Greenman, C. D., Pleasance, E. D., Newman, S., Yang, F., Fu, B., Nik-Zainal, S., Jones, D., Lau, K. W., Carter, N., Edwards, P. A., Futreal, P. A., Stratton, M. R. & Campbell, P. J. (2012). Estimation of rearrangement phylogeny for cancer genomes. Genome Res 22, 346-61.

Gusnanto, A., Wood, H. M., Pawitan, Y., Rabbitts, P. & Berri, S. (2012). Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data. Bioinformatics 28, 40-7.

Hagenkord, J. M., Gatalica, Z., Jonasch, E. & Monzon, F. A. (2011). Clinical genomics of renal epithelial tumors. Cancer Genet 204, 285-97.

Hanahan, D. & Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. Cell, 144, 646-74.

Knudson, A. G., Jr. (1971). Mutation and cancer: statistical study of retinoblastoma. Proc Natl Acad Sci U S A 68, 820-3.

Lamy, P., Andersen, C. L., Dyrskjot, L., Torring, N. & Wiuf, C. (2007). A Hidden Markov Model to estimate population mixture and allelic copy-numbers in cancers using Affymetrix SNP arrays. BMC Bioinformatics 8, 434.

Li, A., Liu, Z., Lezon-Geyda, K., Sarkar, S., Lannin, D., Schulz, V., Krop, I., Winer, E., Harris, L. & 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 Res 39, 4928-41.

Li, J., Lupat, R., Amarasinghe, K. C., Thompson, E. R., Doyle, M. A., Ryland, G. L., Tothill, R. W., Halgamuge, S. K., Campbell, I. G. & Gorringe, K. L. (2012). CONTRA: copy number analysis for targeted resequencing. Bioinformatics 28, 1307-13.

Manie, E., Vincent-Salomon, A., Lehmann-Che, J., Pierron, G., Turpin, E., Warcoin, M., Gruel, N., Lebigot, I., Sastre-Garau, X., Lidereau, R., Remenieras, A., Feunteun, J., Delattre, O., de The, H., Stoppa-Lyonnet, D. & Stern, M. H. (2009). High frequency of TP53 mutation in BRCA1 and sporadic basal-like carcinomas but not in BRCA1 luminal breast tumors. Cancer Res 69, 663-71.

Medvedev, P., Fiume, M., Dzamba, M., Smith, T. & Brudno, M. (2010). Detecting copy number variation with mated short reads. Genome Res 20, 1613-22.

Metzker, M. L. (2010). Sequencing technologies - the next generation. Nat Rev Genet, 11, 31-46.

Meyerson, M., Gabriel, S. & Getz, G. (2010). Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet 11, 685-96.

Mosen-Ansorena, D., Aransay, A. M. & Rodriguez-Ezpeleta, N. (2012). Comparison of methods to detect copy number alterations in cancer using simulated and real genotyping data. BMC Bioinformatics 13, 192.

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

Natrajan, R., Lambros, M. B., Geyer, F. C., Marchio, C., Tan, D. S., Vatcheva, R., Shiu, K. K., Hungermann, D., Rodriguez-Pinilla, S. M., Palacios, J., Ashworth, A., Buerger, H. & Reis-Filho, J. S. (2009). Loss of 16q in high grade breast cancer is associated with estrogen receptor status: Evidence for progression in tumors with a luminal phenotype? Genes Chromosomes Cancer 48, 351-65.

Natrajan, R., Mackay, A., Lambros, M. B., Weigelt, B. *et al.* (2012). A whole-genome massively parallel sequencing analysis of BRCA1 mutant oestrogen receptor-negative and -positive breast cancers. J Pathol 227, 29-41.

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

Nielsen, R., Paul, J. S., Albrechtsen, A. & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet 12, 443-51.

Nik-Zainal, S., Van Loo, P., Wedge, D. C., Alexandrov, L. B. *et al.* (2012). The life history of 21 breast cancers. Cell 149, 994-1007.

Podlaha, O., Riester, M., De, S. & Michor, F. (2012). Evolution of the cancer genome. Trends Genet 28, 155-63.

Popova, T., Manie, E., Rieunier, G., Caux-Moncoutier, V., Tirapo, C., Dubois, T., Delattre, O., Sigal-Zafrani, B., Bollet, M. A., Longy, M., Houdayer, C., Sastre-Garau, X., Vincent-Salomon, A., Stoppa-Lyonnet, D. & Stern, M. H. (2012). Ploidy and large-scale genomic instability consistently identify basal-like breast carcinomas with BRCA1/2 inactivation. Cancer Res 72, 1-9.

Popova, T., Manie, E., Stoppa-Lyonnet, D., Rigaill, G., Barillot, E. & 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.

Rasmussen, M., Sundstrom, M., Goransson Kultima, H., Botling, J., Micke, P., Birgisson, H., Glimelius, B. & Isaksson, A. (2011). Allele-specific copy number analysis of tumor samples with aneuploidy and tumor heterogeneity. Genome Biol 12, R108.

Sathirapongsasuti, J. F., Lee, H., Horst, B. A., Brunner, G., Cochran, A. J., Binder, S., Quackenbush, J. & Nelson, S. F. (2011). Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. Bioinformatics 27, 2648-54.

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M. & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29, 308-11.

Shipitsin, M., Campbell, L. L., Argani, P., Weremowicz, S. *et al.* (2007). Molecular definition of breast tumor heterogeneity. Cancer Cell 11, 259-73.

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

Staaf, J., Vallon-Christersson, J., Lindgren, D., Juliusson, G., Rosenquist, R., Hoglund, M., Borg, A. & Ringner, M. (2008b). Normalization of Illumina Infinium whole-genome SNP data improves copy number estimates and allelic intensity ratios. BMC Bioinformatics 9, 409.

Storchova, Z. & Kuffer, C. (2008). The consequences of tetraploidy and aneuploidy. J Cell Sci 121, 3859-66.

Stratton, M. R., Campbell, P. J. & Futreal, P. A. (2009). The cancer genome. Nature 458, 719-24.

Stuart, D. & Sellers, W. R. (2009). Linking somatic genetic alterations in cancer to therapeutics. Curr Opin Cell Biol 21, 304-10.

Tran, B., Dancey, J. E., Kamel-Reid, S., McPherson, J. D., Bedard, P. L., Brown, A. M., Zhang, T., Shaw, P., Onetto, N., Stein, L., Hudson, T. J., Neel, B. G. & Siu, L. L. (2012). Cancer genomics: technology, discovery, and translation. J Clin Oncol 30, 647-60.

Van Loo, P., Nordgard, S. H., Lingjaerde, O. C., Russnes, H. G., Rye, I. H., Sun, W., Weigman, V. J., Marynen, P., Zetterberg, A., Naume, B., Perou, C. M., Borresen-Dale, A. L. & Kristensen, V. N. (2010). Allele-specific copy number analysis of tumors. Proc Natl Acad Sci U S A 107, 16910-5.

Venkatraman, E. S. & Olshen, A. B. (2007). A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics 23, 657-63.

Wang, J., Mullighan, C. G., Easton, J., Roberts, S. *et al.* (2011). CREST maps somatic structural variation in cancer genomes with base-pair resolution. Nat Methods 8, 652-4.