Advanced statistical tools for SNP arrays : signal calibration, copy number estimation and single array genotyping

Rippe, R.C.A.

Introduction

1

In this chapter some background to the data used in this thesis is described, in both biological and methodological sense. First, a basic discussion on DNA and the genetics of tumor tissue is given, followed by details on commonly used SNP measurement technology, to close with different applications of the same signals. Most of the descriptions are a strongly simplified version of reality, but this is needed to understand most of the concepts and ideas described in this thesis.

1.1 Human DNA and disease

Recently a book called "The Emperor of all Maladies: A Biography of Cancer" (Mukherjee, 2010) was published. Its message is clear: cancer is a large problem. In general, increasing amounts of evidence have been gathered that each case of cancer or tumor development is related to genetics at least to some extent. More specifically, it has to do with genetic mutations which can be caused by heritable susceptibility or simply by external factors (mutagens) like chemicals or radiation. The human body consists of numerous cells and each of them contains a full copy of our complete DNA. Therefore, there are a lot of opportunities for problems to occur. Small scale changes (mutations) in DNA regularly occur and are not harmful per se, since the structure of DNA has several recovery methods for successful replication. However, if despite the backups a cell with problematic DNA has reproduced, that DNA is also copied. Since cell division is a continuous process in order to replace damaged cells, genetic problems can spread quickly. Mutations occur infrequently, while we see the result of the mutations in the form of so-called polymorphisms, the different DNA variants that can arise from mutations.

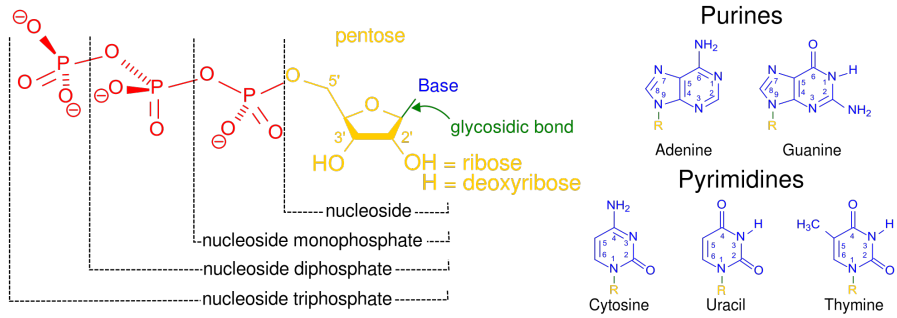

Figure 1.1: Nucleotides and nucleosides: molecules and bases ©Scientific Commons.

Therefore, to understand this problem better, a more detailed description of DNA and polymorphisms is given below.

DNA was first discovered in 1869 by the Swiss biochemist J. F. Miescher. He performed chemical tests on tissue obtained from hospital waste. However, it was not before 1909 that Ph. Levene formulated the first, but incorrect, theory about the chemical structure of DNA. He suggested that it was a large structure of 4 building blocks, the nucleotides. The actual and correct structure was published in 1953 by Watson & Crick in Nature, to be followed later by a paper on DNA replication, by the same authors. They based their publications on X-ray diffraction data (1952) from Rosalind Franklin and colleagues. Their combined efforts taught us some valuable lessons.

Healthy (human) DNA is contained in pairs of chromosomes which are two (long) chains of nucleotides which are referred to using an “alphabet" of four letters, representing molecules (nucleotides) with the bases (nucleosides) adenine, cytosine, guanine and thymine. We distinguish purines and pyrimidines, to which sugar and phosphate groups are attached and together make up the whole nucleotide. See Figure 1.1 for an illustration. A nucleotide with a purine base pairs to one with a pyrimidine base. Human DNA holds about 3 billion of these pairs. Without going into more detail, the chemical and physical properties of all these bonds cause the polymer to coil up into a

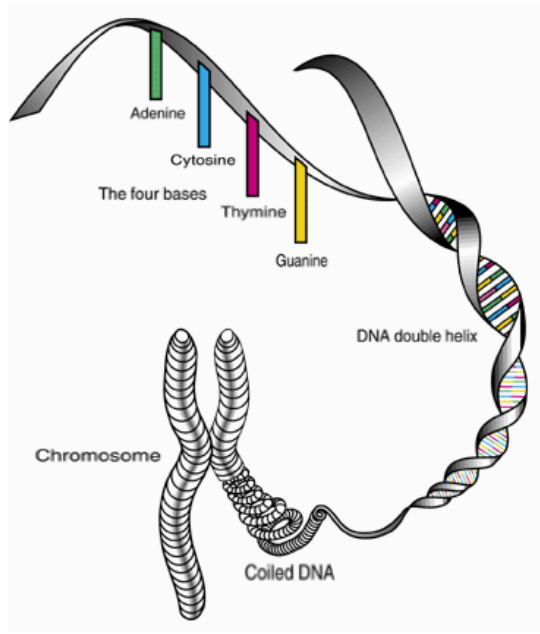

Figure 1.2: DNA: a double helix structure.
Image ©retrieved from https://www.llnl.gov/str/June03/Stubbs.html.

so-called double helix structure, which makes storage of genetic information very compact and safe. This widely known property is described for example in Brown (1999) and Griffiths et al. (2002), and is depicted in Figure 1.2. We can define any (nucleotide) position on a chromosome by counting the number of nucleotides from a unique end.

Genes are specific sets of nucleotides, which can have very different purposes. Some genes code for expression of certain physical features, like eye color, hip width or shape of the nose. Others are involved in regulating the expression of genetic information or serve to simply hold protein structures together. For organisms other than human the same principles hold, but with a few differences. For example, in non-human species the number of chromosome pairs is different from that in humans. Furthermore, where hu-

man DNA holds two chromosome copies - two homologues, called diploid - different numbers of homologues or ploidy also occurs. For example, DNA from some potato species has sets of four homologues - tetraploid - and the Drosophila melanogaster (a type of fruit fly) has 4 pairs of chromosomes (hence is diploid), in which about 75% of (known) human disease genes can be matched to the fruit fly genome (Reiter et al., 2001). Therefore the latter is often used as a genetic model for human disease(s) and to study biological processes such as aging.

The remainder of this introduction and thesis is however restricted to diploid human DNA and a very specific problem therein: nucleotide polymorphisms.

A nucleotide on one of the two chromosome is referred to as an allele (not to be mistaken for a gene allele, which is a large set of nucleotide base pairs). At most positions on our chromosomes we will always find the same nucleotide on both chromosomes. But there are millions of places (about 1 in every 1000 nucleotides) where we see different allele variations (polymorphic occurence) due to a mutation on one of the chromosomes. If the allele for one chromosome is called A and B for the other chromosome and if the chromosomes are indeed identical, then only pairs AA or BB would occur. Small changes can also be AB or BA, which cannot be distinguished. When for a SNP one particular pair of alleles in one person is different from the population, it is called a mutation. If this mutation occurs in more than 1% of the population it is called a variation. These variations are known to occur on specific chromosomal locations, which means that only a small selection of the whole genome sequence (all base pairs in a row), is of interest. They are called Single Nucleotide Polymorphisms (SNPs, pronounced as “snips") and they are being studied extensively in biology and medicine (Adorjan et al., 2002; Altshuler et al., 2005).

These mutations can have (possibly strong) implications, for example breast cancer (Chin et al., 2007). Two general classes of mutations in human DNA are changes in allele copy number (CNV) and imbalance in the allelic ratio. Further mutational distinctions can be made between duplications, inversions and deletions (Conrad et al., 2006). Duplication means that more than two allele copies were found, where two were expected. In

contrast to duplication, deletions of specific sections of chromosomes chromosomal sections can occur, in two ways. In the first, a section is lost in one of the cell reproduction stages. An inversion occurs when a broken chromosome is repaired, but (some of) the parts are reversed before actual repair. Translocations involve 'reordering' of chromosomal sections.

1.2 SNP chips and recent developments in DNA measurement

As described before, each nucleotide with base A, C, G or T on one chromosome is complementary to another. This complementarity is used in a process called hybridization, in which a sequence-specific oligonucleotide (a sequence of 50 bases or less) binds to the DNA strand under treatment. The DNA strand contains the SNP of interest. The SNP measurements are obtained from a photo(n) sensor that captures photons with a given wavelength from one channel, that are emitted after the hybrid DNA is targeted by a laser. By laser illumination the molecules change energetic state and when they fall back to their original state, they emit energy. The result from the photo sensor is a high-resolution image, which is analyzed and translated to numeric values. These numbers are then used in downstream analyses. In the end we have, for one sample, a fluorescence value for each allele in each SNP. Below, two major platforms are discussed, but more exist.

The first major manufacturer is Affymetrix Affmetrix, 2006), known for its trademark 'GeneChip' products. In these single-channel fluorescence chips four probes with oligonucleotides are used (Figure 1.3), to reduce the influence of possible miss-matches. The first and second probe interrogate one direction for the A and B target allele, while the third and fourth interrogate the opposite direction of the the same alleles. Through the years different enzymes were used to split the DNA prior to hybridizing in a particular direction. Measurements are performed on solid surfaces, usually glass or silicon.

The second manufacturer, started in 2001, is Illumina (Fan et al., 2006). One implementation that got Illumina to the front of SNP research was the

Figure 1.3: An Affymetrix SNP chip. ©Affymetrix.

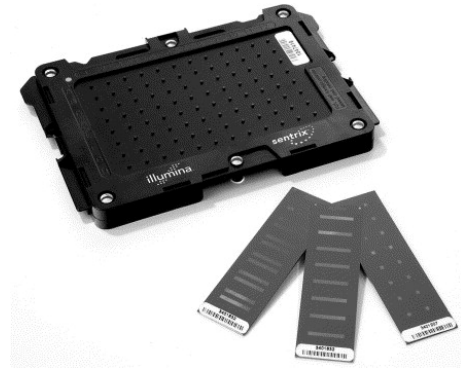

Figure 1.4: An Illumina 8x12 cell well-plate. ©Illumina.

so-called bead array (top device in Figure 1.4, known as the 'GoldenGate Genotyping' technology. One of the trademarks was the use of two-color fluorescence signals (two-photon excitation fluorometry) using bead arrays, which are also used in methylation applications (Bibikova et al., 2006). Common dyes are Cy3 (with a fluorescence emission of 570nm; green light) and Cy5 (670nm: red light). The main plate consisted of 96 (8 by 12) wells for the same number of bead arrays, where in each of the wells a different tissue sample or blood sample was analyzed for about 1600 SNPs. These SNPs were probed on individual beads instead of a solid surface. Later, the SNP resolution increased to 550.000 in more recent technologies like 'Infinium'.

Initial SNP technology for human DNA could target about 1600 SNPs in the whole genome set. In subsequent years the number of targeted SNPs, the SNP resolution, gradually increased through 50.000, 100.000, 500.000 to currently up to 1.200.000 SNPs (full resolution probing) in the most recent platforms.

Apart from the two platforms discussed above, a number of other platforms exist, e.g. Sequenom, Perlegen, Mip, FP-TDI and InVader. However, due to their absence in this thesis, their technical properties are not discussed in further detail.

The latest trend away from SNPs, although not addressed specifically in the following chapters, does not probe at a high number of SNP positions but simply analyzes the whole genomic sequence, including all allele pairs in between currently probed SNP positions. This technique is, unsurprisingly, called 'whole genome sequencing'. Accompanying challenges for this new technique come in terms of data storage (several TBs per individual) and model efficiency (think of memory capacity and computation times).

1.3 Applications of SNP fluorescence signals

The bottom line for the remainder of this thesis, deriving from the above descriptions, is that, per SNP, we have two fluorescence signals that are proportional to the amount of the respective A and B alleles.

One major application of these DNA measurements is to determine which specific combination of alleles is found at individual SNP positions. Assume two channels a and b for the two different alleles, one can measure the double signal for allele A (AA), indicated by double fluorescence strength in channel a and low in b, the double strength signal for B (BB, strong channel b and low a) or equivalent signal strength for each allele (AB, channel a and b in equal proportions) This process is generally referred to as *genotyping* or *genotype calling*. Finding an AA or BB genotype is called homozygous; the alleles are the same for both chromosomes. The AB genotype is called heterozygous. Genotyping of normal (non-tumor) DNA is common practice in e.g. epidemiology (e.g. Huebner et al., 2007).

It comes as no surprise that the genotyping problem was addressed before. For example, companies that manufacture chemical platforms to measure DNA composition (described in section 1.2 in more detail) generally provide their own software to determine genotypes. Third party solutions also exist, like CRLMM (Carvalho et al., 2007) and BirdSeed (Korn et al., 2008). Of course many other methods exist. All of these methods rely one way or another on combined information of multiple reference arrays. Large-scale efforts have been made to catalog known SNP positions for different sets of publicly available arrays, in order to create a 'gold standard' database with genotypes: HapMap (The International HapMap Consortium, 2003; 2007). On the other end, there is also a database that contains annotations for each of these SNPs: BioMart (www.biomart.org).

A second application is to determine deviations from the common composition of two alleles in healthy tissue (e.g. Lips et al., 2005). For example, due to a variety of causes, one (or both) of the alleles may be lost and replaced by either 'empty' DNA, a so-called null-allele, or copied back from the remaining allele. Alternatively, erroneous extra alleles copies may have been created during cell division. Checking for changes in the number of alleles in both chromosome is generally referred to as Copy Number assessment. For example, each chromosome has the same deviation from normal tissue. A more general case is allelic imbalance, where the number of alleles is no longer the same for the two chromosomes. The special case of loss of one allele (out of two) and replacement by the remaining one, hence losing the possibility of finding a heterozygous genotype, is called allelic imbalance (as discussed in McCarroll et al., 2006). From section 1.1 it is immediately clear what the consequences of these problems can be. From this point of view it is a challenge to accurately estimate breakpoints between chromosomal regions with different numbers of alleles. A related challenge is that of finding how many allele copies are found in a specific region. A lot of methods have devised for this specific purpose, like FLasso, CGHseg, DNACopy, VEGA and cumSeg. Several extensive comparisons have been made between subsets of these methods (Lai et al., (2005); Winchester et al., (2009); Tsuang et al., (2010); Muggeo et al., (2011) and Morganella et al. (2010)), all concluding that there is not one procedure that serves all purposes. They collectively

suggest to use multiple methods in conjunction.

For all purposes of SNP signals described above, specialized models exist. One property found in almost all algorithms, is that decisions are made by relating allele intensities between multiple arrays, one SNP at the time (e.g. genotype calls) or by using a set of healthy tissue reference arrays (e.g. CNV profiling). In contrast to this 'school of thinking', the methods described in this thesis use a single array for signal calibration, genotype calling or CNV profiling. This approach can be very useful in situations where new chips have to be designed and a large pool of testing and/or reference material is not available. An example can be found in ALCHEMY by Wright et al. (2010).

1.4 Data format and signal properties

If measurements are performed for an individual, they are performed using a chip or array. Therefore, from here on we use the term "array" for a biological sample, instead of the common use as a multidimensional matrix.

A typical collection of arrays can be seen as a data box, containing only ones and zeros, with three modes: SNPs, arrays and genotypes. Each allele is associated with only one genotype, and hence has a value of 1 in only one of the three genotype layers. This implies that 2/3 of the data box depicted in Figure 1.5 contains no information, or in more statistical terms, is missing. The missingness is a result of this way of structuring the data. Aggregation of the three incomplete genotype layers results in a full matrix with measurements for the alleles on all arrays, without genotype information. This suggests to develop a multi-way model (Kroonenberg, 2008; Smilde et al., 2004) that can cope with large and structural amounts of missing observations. For practical reasons we will not. However, the three-dimensional structure can be used to see and evaluate other interesting properties.

If the technology we described worked perfectly, our story would end here. In practice we observe a number of interesting and relevant patterns in normal DNA:

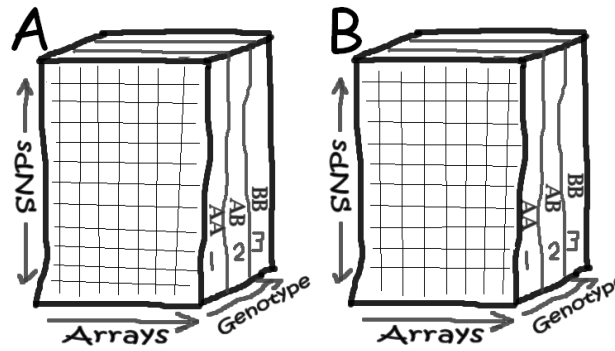

Figure 1.5: SNP data with three modes: SNPs in the rows, samples in the columns and genotypes in the layers.

1. The strength of the fluorescence signals varies systematically between SNPs. If a SNP signal is strong, it is strong in all arrays. A weak SNP signal is weak in all arrays.
2. The strength of the signals varies between biological samples (which is unavoidable; it is caused by differences in the quality of the biological material and the efficiency of DNA extraction) and hybridization.
3. Systematic deviations from the theoretical genotype factors 1 and 2 occur (1 or 2 times the allele).
4. Noise and background signals are present.

The effects described under 1 and 2 are depicted in Figure 1.6. After sorting it is clear that SNPs that have weaker fluorescence signal compared to other SNPs, also are relatively weaker in all arrays. There is a similar effect for arrays, but it is less strong.

1.5 Contributions in this thesis

As valuable as fundamental research is, it is useless without the proper field of application. By coincidence, a prospective user (group) in the Erasmus

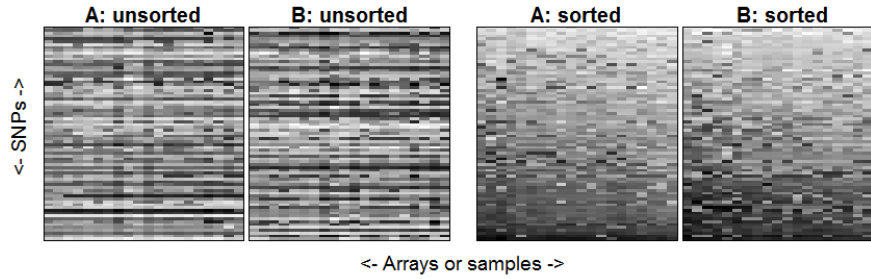

Figure 1.6: Selection of raw SNPs and arrays, based on a small two-color Illumina set. Left: signals unsorted. Right: both signals sorted for rows and columns of A allele intensities.

Medical Center became involved in the development of a tailor-made statistical solution. A short discussion evolved into a more extensive project, from which all parties learned a lot. Detailed communications resulted in more questions as well as theoretical and procedural suggestions.

This thesis introduces a number of new statistical models and algorithms:

- The SCALA model that contains parameters for estimating the systematic effects of SNPs, arrays and genotypes. This model is applied to both alleles.
- The model leads to an extremely large linear regression problem with millions of observations and possibly a million parameters or more. However, it has an extremely sparse structure. A specialized semi-symbolic algorithm allows exact estimation in very short time.
- Once the parameters of the model have been estimated, they are used to eliminate the systematic effects, thereby greatly enhancing the quality of the fluorescence signals. This is called calibration. Further analysis, like genotyping, or copy number estimation is improved.
- It is common practice to perform genotyping SNP by SNP, using relatively large sets of arrays. In this monograph we break with this tradition and perform genotyping for all SNPs on individual arrays. A

semi-parametric mixture model is fitted, with three component densities, one for each of the AA, AB and BB genotypes. Comparison to results of established SNP by SNP algorithms, as found on the HapMap archive shows equal or better performance.

- An interesting part of the SNP practice are so-called waves. When fluorescence signals representing copy numbers are plotted along chromosomes and smoothed, a systematic wavelike pattern becomes visible. A common misunderstanding is that this is a manifestation of a real, slowly changing, wavelike spatial structure. It is shown that the latter is not the case: after calibration the waves disappear.

- Copy number variation (CNV) generally occurs in a segment-wise manner. There is a large literature on smoothing and segmentation of CNV signals, in order to obtain the boundaries of the segments and their levels. A new smoothing algorithm is presented that uses a so-called L_0 penalty on jumps between smoothed values. The result is an extremely sharp segmentation, with extremely smooth segments in between: the segments are constant.

- It is not possible to apply the same smoother to allelic imbalancesignals, because several parallel data bands occur. An existing scatterplot smoother was modified to get sharp segmentation here too.

- The basis of the genotyping method proposed in this monograph is a display of the ratio of the A and B signals versus their sum (on logarithmic scales). Low signals on the sum scale as well as unclear separation between the three genotype groups on the ratio scale indicate low(er) chip quality. The proposed approach to individual arrays is very useful to exploit this knowledge to select only the SNP observations of the highest quality, by a user-defined threshold.

- All models and algorithms are written in R, and combined in a software suite, called SCALA. SCALA also provides both command-line functions (for estimation and calibration, as well as genotpying) and a graphical user interface for interactive (simultaneous) smoothing and plotting of

CNV and allelic imbalance. It can convert DNA array files from different platforms. By user request evaluative and interactive graphs can be created, as well as customized numerical output.

1.6 Thesis outline

After this introduction follow six additional chapters. Chapter 2 to 7 have been written as individual papers, and chapter 8 is a discussion chapter. The software is extensively described in a special section that contains two more individual papers. These address the implementation of most of the concepts and models from chapters 2 to 6.

Chapter 2 describes the linear SCALA models that estimate the calibration parameters. We exploit the structural properties of SNP-specific intensities over arrays in a model that models an overall genotype parameter (over all SNPs in the whole set of arrays) for the three genotypes as well as a model that estimates a set of three genotype parameters for each SNP. Signal calibrations using both types of model are derived and illustrated.

Chapter 3 frames genotyping using a single array against common procedures that call genotypes using a single SNP from multiple arrays. A semi -parametric model is derived that fits three log-concave densities on a two-dimensional histogram.

Chapter 4 discusses the possibilities of removing SNP signals of low quality. Performance for genotyping is assessed in terms of sample call-rate, while performance for CNV calling is illustrated visually, rather than numerically. The proposed low signal filter show improvements for both applications.

Chapter 5 shows that so-called in a copy number signal can be removed effectively using the signal calibration from chapter 2. SNP signals are transformed into aCGH-resembling signals, so that performance by SCALA can be compared with a specialized method for aCGH signals, NoWaves.

Chapter 6 discusses a single-array L_0 penalty signal smoother, ZEN, for CNV signals. Existing methods either need a set of reference arrays, or in single array models, small (random) fluctuations in the signal are modelled,

while ideally only significant signal jumps should be modelled. Performance of the L_0 norm is compared to a multi-array method, VEGA and applications in CNV estimation, estimations of allelic imbalance and scatterplot smoothing are shown.

Chapter 7 showcases the SCALA software suite that was developed based on the ideas and concepts in Chapter 2 to 6. It provides a short theoretical introduction on the implemented models and functions, then discusses details and options for each function, and concludes with an extensive example.

Chapter 8 consists of three parts. It provides a short summary, an overall discussion and suggestions for future research.

Research Articles