University of Warsaw
Faculty of Mathematics, Informatics and Mechanics

Maciej Sykulski

Data analysis and modeling in human genomics

PhD dissertation

Supervisor

dr hab. Anna Gambin

Institute of Informatics
University of Warsaw

November 2014

Author's declaration:
aware of legal responsibility I hereby declare that I have written this dissertation myself and all the contents of the dissertation have been obtained by legal means.

November 7, 2014 .....

date *Maciej Sykulski*

Supervisor's declaration:
the dissertation is ready to be reviewed

November 7, 2014 .....

date *dr hab. Anna Gambin*

Thesis supervisor: dr hab. Anna Gambin Maciej Sykulski

Data analysis and modeling in human genomics

Abstract

Human genomics undergoes fast development thanks to new genome research technologies, such as aCGH (array Comparative Genomic Hybridization), or Next Generation Sequencing. Large amounts of data from patients potentially contain a lot of information about potential diseases, or susceptibility thereto, about mechanism of emergence of genetic diseases, and about mechanisms of evolution. Extracting such information requires careful data processing, constructing of proper statistical models, in order to analyze emerging hypotheses and discriminate between them. The basis for the used methods are algorithms which explore the space of possible models, with the use of Monte Carlo Markov Chain, or using optimizing schemes such as Expectation Maximization, as well as statistical tests and adequate heuristics. This dissertation concerns about issues of aCGH microarray design, comparison of quality of thereof. We propose a method of spatial denoising based on Markov Random Fields, which at the same time allows to recover from aCGH data CNVs (Copy Number Variants). Presented is the database and computational webservice, as well as biomedical results obtained with the use of it.

Keywords: DNA microarrays, aCGH, CNV, genomic disorders, segmentation, Gaussian Markov Random Fields, Bayesian Graphical model, Expectation Maximization, outlier detection, semantic web
ACM Classification: J.3

Promotor rozprawy: dr hab. Anna Gambin Maciej Sykulski

Analiza i modelowanie danych w genomice człowieka

Streszczenie

Genomika człowieka rozwija się szybko dzięki nowym dostępnym technologiom badania genomu, takim jak macierze aCGH (array Comparative Genomic Hybridization) lub nowoczesne metody sekwencjonowania genomu. Duże ilości danych pochodzące od pacjentów zawierają potencjalnie mnóstwo informacji o możliwych chorobach, czy skłonnościach ku takowym, o mechanizmach powstawania chorób genetycznych, także o mechanizmach ewolucji. Wydobycie takich informacji wymaga uważnego przetwarzania danych, konstrukcji właściwych modeli statystycznych w celu analizy i dyskryminacji pomiędzy stawianymi hipotezami. Podstawą używanych metod są algorytmy eksplorujące przestrzeń dopuszczonych modeli przy użyciu metod Monte Carlo Markov Chain, schematów optymalizacyjnych typu Expectation Maximization, testów statystycznych, jak i również adekwatnych heurystyk. W rozprawie zajmujemy się problematyką projektowania płytek aCGH, porównywania jakości tychże, oraz analizy wyników eksperymentów aCGH. Proponujemy opartą na Losowych Polach Markowowskich metodę odszumiania przestrzennego, która jednocześnie pozwala na wykrycie z danych aCGH CNVs (Copy Number Variants), segmentów w genomie pacjenta odpowiadających rearanżacjom. Przedstawiamy serwis bazodanowo-obliczeniowy, który powstał dla potrzeb diagnostyki, oraz wyniki biomedyczne otrzymane przy jego pomocy.

Słowa kluczowe: mikromacierze DNA, aCGH, CNV, choroby genetyczne, segmentacja, gaussowskie losowe pola markowowskie, bayesowskie modele graficzne, algorytm Expectation Maximization, wykrywanie przypadków odstających, sieć semantyczna
Klasyfikacja tematyczna ACM: J.3

Contents

1 Introduction **1**

1.1 Main Results 6

1.2 Scientific publications and other published resources 9

2 aCGH experiment data characteristics and microarray design **15**

2.1 DNA microarray technology 15

2.1.1 Types of DNA microarrays 16

2.1.2 aCGH data characteristics 17

2.2 Signal and noise in DNA microarray data 19

2.2.1 Heteroscedasticity in $\log_2$ratio 21

2.2.2 Signal-to-noise ratio in $\log_2$ratio data 24

2.2.3 Quality Control 25

2.2.4 aCGH data segmentation task 27

2.3 The iterative process of targeted microarray design 30

3 Markov Random Fields for aCGH segmentation and spatial denoising **35**

3.1 Outline of the proposal 37

3.2 Background and Segments Markov random fields model declaration 38

3.2.1 BSMF log-posterior likelihood as a quadratic problem 44

3.2.2 Posterior conditional distributions of the model . . 50

3.2.3 Categorical mixture prior distribution for the segment field 51

3.3 Double linkage in the segment field . . . 55
3.4 Optimizing BSMF parameters with Expectation Maximization . . . 61
3.5 Results: application on aCGH microarrays data . . . 62
3.5.1 Comparison with CBS results . . . 63
3.5.2 Sensitivity to priors . . . 63

4 Functional performance of aCGH design **71**
4.1 Methods . . . 73
4.1.1 Synthetic Array Design . . . 73
4.1.2 Exon CGH Array Design . . . 74
4.1.3 Enhancement of DNACopy . . . 75
4.1.4 Robustness Measure . . . 75
4.1.5 Optimizing Exon CGH Array Design via Noise-induced Discrepancy . . . 79
4.2 Results and discussion . . . 81
4.2.1 Synthetic Data . . . 81
4.2.2 Noise-induced Discrepancy of Optimized Designs . . . 82
4.3 Conclusions . . . 85

5 Rare CNV detection **87**
5.1 Methods . . . 91
5.1.1 Datasets . . . 91
5.1.2 Outstanding CNVs detection . . . 91
5.1.3 Polymorphic regions filtering . . . 96
5.1.4 Validation . . . 96
5.2 Results and discussion . . . 99
5.2.1 Discovery and validation of rare CNVs . . . 100
5.3 Conclusions . . . 101

6 Semantic Web technologies for molecular medicine **103**
6.1 IMID2py – a web application tool for aCGH analysis . . . 104
6.2 IMID2py semantic extension . . . 106
6.3 Biomedical results . . . 109

References **113**

Listing of figures

1.1.1 Plotted is the outline of this dissertation. We aspire to follow the main steps in the process of aCGH design, research, and clinical validation. A more detailed diagram of aCGH design process can be found in Chapter 2 Figure 2.3.2 7

2.2.1 We sampled 50K segments, obtained with CBS algorithm having analyzed $\log_2$ratio from 474 patients aCGH arrays (IMID). We computed variances around segments' $\log_2$ratio means. Two plots were produced: *a)* $\log_2$ratio were exponentiated back to intensities ratios to compute variances, *b)* variances of $\log_2$ratio were produced. LOESS parabola fit visually confirms Agilent's quadratic model which applies to intensity data (a). Quadratic dependence persists in $\log_2$ratio data (b). Signal-to-noise $\log_2$ratio statistics for significant segments (left and right subsets) are as following: Min. 0.9309, 1st Qu. 2.2820, Median 3.545, Mean 13.610, 3rd Qu. 6.564 22

2.2.2 Plots from the stage of quality control of gene expression arrays form patients with bladder cancer (MD Anderson Cancer Center, manuscript in preparation). 25

2.2.3 The grid alignment on Agilent DNA microarray. An image from Feature Extraction software. Misalignment, rotation, wrong fit to the grid may cause several, or many, probes to report biased data. 28

2.2.4 $\log_2$ratio data from aCGH experiment aligned along a chromosome. Aberration at the right end of the chromosome, i.e. shorter deletion and longer duplication is yet to be detected by a segmentation algorithm. Since the aberrated segments have substantial length this wouldn't be a problem for an algorithm, however shorter segments, or segments with lower mean are more difficult to detect. 29

2.3.1 Designing microarray for coverage of specific regions.In version 8 there are 180 000 probes, each 60bp in length. 31

2.3.2 The iterative process of aCGH microarray design. Solid lines represent the main process and data flows. Dotted lines represent feedback of results to subsequent, or same, iterations. Type of information by colors; *orange*: genomic copy number information (CNVs); *green*: regions in the genome either to be covered, or covered by a set of aCGH DNA probes; *violet*: noise reduction, quality control data. 33

3.0.1 Noise in $\log_2$ratio from aCGH microarrays scans. Top: a microarray scan containing linear and non-linear spatial trends, bottom: a well hybridized array with no visible artifacts. 36

3.2.1 Zoom to the top left corner of a microarray. Background noise field levels marked in red/blue, breaks in the field marked black, white dots are spots without a probe. . . . 40

3.2.2 $\log_2$ratio data, scanned from aCGH microarray, aligned along chromosome Y. Red dashed line represents hidden genomic signal segments. A deletion on the right end of the chromosome is present (the segment with a significantly lowered $\log_2$ratio mean ~ -0.5). 41

3.2.3 Factor graph for $\mathcal{L}_{\text{BSMF}}(\theta, Z; x, \Omega)$ model. $\log_2$ratio data $\vec{x}$ marked gray. Gaussian Markov random fields $\vec{a}$, $\vec{b}$, defined on graphs G_{spatial}, G_{genome} accordingly, in white $\log_2$ratio data $\vec{x}$ marked gray. Precision latent variables τ, ρ, ν marked blue, their prior Gamma distributions not charted on the graph, Normal prior for $\vec{b}$ not charted. Discrete variables $Z = \{\vec{y}, \vec{z}, \vec{s}\}$ marked green. Background, segment fields break probabilities p, q marked in yellow. The $\vec{s}$ *segment category state* is introduced in Section 3.2.3. The multivariate $\mathcal{N}_{G_{\text{genome}}}$, $\mathcal{N}_{G_{\text{spatial}}}$ correspond to Gaussian Markov Fields on edges of respective graphs. Interestingly, the graph is a tree graph when vector variables and multivariate distributions are treated as typical Bayes network nodes. 53

3.2.4 Histogram of segment $\log_2$ratio means from IMID2py database. Since most segments have no aberration and segments around 0 dominate, data was divided into 2 groups for better visibility. 54

3.3.1 *Left:* The ratio between the second smallest eigenvalues of matrices M_{single} and $M_{\text{double.w}}$ for $w = 1$ and $w = 0.5$. The $\lambda_0 = 0$ for all matrices, while $\lambda_1^{\text{double.1}}$ is 5 times larger, and $\lambda_1^{\text{double.0.5}}$ is 3 times larger, than the second eigenvalue of the single linkage graph matrix for large matrix sizes n. This shows that the spectral gap for the double linked graph is 5, or 3, times larger than the spectral gap of the single linked graph. *Right:* The ratio of optimal slope coefficients in the segment field $\frac{\hat{a}_1^{\text{single}}}{\hat{a}_1^{\text{double}}}$. Double linked segment fields fit for smaller slopes. (ref. eqn. 3.40). . . 58

3.3.2 The effect of double linkage in the segment field. 60

3.4.1 A hexagonal lattice(equivalently a triangular tiling) graph G and its stratification on Agilent aCGH microarray. Each stratum has its neighborhood non-intersecting with other nodes from the same stratum. Strata no. 1 marked in blue. neighborhood of the center node marked in red. We iterate over strata and maximize log-probability on each one separately, while keeping other strata constant. Maximal cliques of G are triangles, each consisting of vertices from strata $1, 2, 3$. 61

3.5.1 $\log_2$ratio together with segment field in red. Middle of yellow bars indicate original $\log_2$ratio value, together with shifted points allow to see the correction by the background field. 63

3.5.3 BSMF Expectation Maximization run convergence. . . . 65

3.5.4 BSMF Expectation Maximization run convergence of 40 runs. 66

3.5.5 Running times of 40 runs of BSMF segmentations. Computations were made in parallel on a server with 24 cores. Maximal number of iterations was set to 400. For histograms of number of iterations until convergence ref. to Figure 3.5.6 67

3.5.6 Comparison between BSMF and Circular Binary Segmentation (CBS), histograms summarizing 40 microarrays. 68

3.5.7 Comparison with Circular Binary Segmentation (CBS). Marked in red are experiments where the difference in results from by CBS and BSMF were largest in the number of probes in it. Scatterplots reveal dependency between posterior segment break probabilities p, and the size of the symmetric difference between BSMF and CBS segmentations. No such dependency is observed between posterior background noise break probabilities q. 69

3.5.8 Sensitivity to the rate of Gamma prior distribution. . . . 70

3.5.9 Sensitivity to the rate of Beta prior distribution. 70

4.1.1 Plots show log2ratio (y-axis) vs. genomic location (x-axis) for synthetic datasets corresponding to four different array designs: (a) original datasets, (b) dataset with simulated poor hybridization effect, (c) dataset with simulated error-prone analysis procedures, (d) dataset with both effects. 74

4.1.2 The resistance of aberrant segments for increasing noise. y-axis correspond to increasing (log-)noise level, different segments are placed along x-axis (genomic location), the $\log_2$ratio are color-coded. 76

4.1.3 The robustness compared for two synthetic designs. The robustness has been calculated for all probes (upper plot) as well as corresponding weights importance (lower plot). The structure of genomic rearrangements mimics the abnormalities in cancer cells. Good design is coded in blue. Red design contains 20% of poorly hybridizing probes and 15% of outliers (probes causing erroneous scanning). 77

4.1.4 The robustness compared for two synthetic designs. The robustness has been calculated for all probes (upper plot) as well as corresponding weights importance (lower plot). The structure of genomic rearrangements mimics the abnormalities in classical genetic disorder (relatively rare long aberrant segments). Good design is coded in blue. Red design contains 15% of outliers (probes causing erroneous scanning). 79

4.2.1 Segmentations performed on original and optimized designs (see description in the main text). A spoiled probe can be visible around 300 mark. 83

4.2.2 Comparison of *relative noise-induced discrepancy* $R^{\mathcal{A}_i|\mathcal{O}}$ for three optimized designs (see description in the main text). 84

5.0.1 **Processing of logratio data** In each subfigure, rows corresponds to samples and columns to probes. On the left: the effect of rank transformation; the same fragment of the genome represented by logratios (a) and their column ranks (b). The wave pattern is eliminated, while true signal (clear deletion) is strengthen. On the right: the polymorphic region in the middle is surrounded by wave patterns and only one significant deletion is visible (c); markers found by our algorithm indicate only deleted segment, all other spurious signals are ignored (d). 90
5.1.1 This figure presents histograms from samples from μ^1 (L_1 distance) null distributions (limit $|S| \to \infty$, number of cases converging to infinity) for various dimensions k. This sampling undertakes the assumption of column (dimensions) independence. 95
5.1.2 **Rare CNVs detected by our method in 366 samples.** Figure shows the chromosmal location of all segments reported by experts (red), segments predicted by our method (yellow) as well as pathogenic CNVs reported in ISCA (purple) and genes from GAD (blue). . 98
5.2.1 **Evaluation of CNVs detection results.** Venn diagram for the predicted rare CNVs (Predicted), confirmed as pathogenic or uncertain (Reported), segments significantly overlapping with DGV (DGV), segments with GAD genes (GAD), and segments selected as polymorphisms according to polymorphic profile (Polymorphisms). 99
6.1.1 IMID2py exports data to display in UCSC genome browser. 105
6.1.2 IMID2py database scheme created with SchemaSpy (Currier, 2005). 106
6.2.1 Webpage presenting the IMID2py semantic extension. . . 108

List of Tables

2.1.1 Dyeing types of material hybridized to DNA microarrays. 17

2.1.2 Applications of DNA microarrays. 18

2.2.1 For the complete description of the process see (Agilent Technologies, 2013, page 212). This thesis concerns with $\log_2$ratio analysis which results from the above process. Characteristics of resulting data is elaborated further in this thesis; observed $\log_2$ratio standard deviation is plotted in fig. 2.2.1. 20

5.2.1 Selected predicted best scored CNVs confirmed later as pathogenic changes. 100

5.2.2 Selected predicted best scored variants of unknown significance 101

Dedicated to Marysia

Acknowledgements

My work would not have been possible without financial support by the Polish National Science Center grant 2011/01/B/NZ2/00864.

I would like to express gratitude and love to my parents whoose support made this work possible. Greatest love I send to Marysia Skoneczna who always was there where needed.

Special thanks to my supervisor Anna Gambin with whom we worked together on many projects.

1

Introduction

Consider experimental science as the art of signal and noise annotation, backed with their assignment to hypotheses, a discipline rooted in theory and practice. Bioinformatics is the kind of science, build atop vast amounts of biotechnological data, with aspirations to model biological systems. How much and which fragments of genomic information can be learned with the use of DNA microarrays technology? How does the technology of Array Comparative Genomic Hybridization allows to rapidly shift its scientific results into clinical application in human medicine? What insights can be gained from aCGH technology, when results from many experiments are analyzed together, and merged with large databases of experiments from laboratories around the world? How useful is the framework of Bayesian Graphical models in the modeling of aCGH data to extract genomic knowledge? These are the topics that I attempt to explore in this thesis.

Sequencing, and publishing of, the full sequence of human genome, achieved in 2003 by the Human Genome Project[1], marks the entering into the new era in medicine, and in life sciences. The gravity of this milestone achievement stands not only in the fact of reading of a majority of nucleotides from DNA sequence of a human, a great technological and aesthetic achievement, but even more in the potential to understand and influence human health. Human genome was the largest genome, with the length of 3 billion nucleotide base pairs, to be sequenced fully at the time, preceded by the pioneering sequencing of bacteriophage $\phi X174$ in 1977, Epstein-Barr virus in 1984, and the sequencing of the first free-living organism, the bacterium *Haenophilus infuenzae*, in 1995. The fruit fly genome, which is 165 million base pairs in length, about one twentieth the size of the human genome, was sequenced in the year 2000.

What enabled these great achievements was a merge of advancements in biotechnology, and in the methods of genomic information retrieval – statistical methods based on mathematical models of fragmentary data from short sequenced fragments. The method of *shotgun sequencing* was introduced by Sanger and Coulson (1975), where structure of a long genomic sequence is derived from sequencing of DNA randomly cut into short fragments (100 to 1000 base pairs). Improvements to master this technology were made during the years, the important step was the introduction of the *pairwise end sequencing* in the 90's, to which Roach et al. (1995) proposed a successful strategy for DNA preparation (i.e. the choice of lengths of cut DNA fragments), and justified it using computer simulations. This addressed the problem of filling the remaining gaps of a long sequence with reasonable resources, and allowed to speed up the sequencing of large genomes including human genome, as predicted by Weber and Myers (1997).

The story of how we've got to the state of the matters as they're today – the sequence if human genome in open access freely available to anyone – is an interesting one, and includes strong competition in late stages between National Institute of Health, USA and JC Venter's private venture Celera Genomics, the latter company to fill patents applications

[1]See Collins et al. (2003); Guttmacher and Collins (2003).

on thousands of human genes, and culminates in the year 2000 with $50 billion Nasdaq crash after the announcement made by President Clinton that genome sequence, as being the "common heritage of mankind", could not be patented,

Translational Medicine today thus seems within the reach of any team of professionals with the access to biotechnology and internet. It is a collaborative effort to bring the latest scientific results from basic research rapidly into applications in medicine. This can be exemplified by a provision of diagnostic tools, a development, and improvement of, procedures. Education, the spread of knowledge, also falls under this category. Polish authors publishing within this trend include Guzik (2010);Bartnik et al. (2012); Derwińska et al. (2012); Wiśniowiecka-Kowalnik et al. (2013), the latter three publications were prepared with the cooperation of the author while working on this thesis. More than 1000 patients were diagnosed by the Institute of Mother and Child (IMID), Warszawa, with the cooperation of Baylor College of Medicine, TX, USA, and with the help of the software developed for that occasion described briefly in Chapter 6 . Many of these patients carry inborn genetic diseases: autism, epilepsy, mental retardation, certain heart defects. This group of genetic diseases is mediated by aberrations in the genomic material of the carrier. These diseases may be inherited, or they may be introduced *de novo* into patients genome during pre-embryonic and embryonic phases.

Pinpointing in a patients genome the exact genetic aberration responsible for the disease requires two things: *i*) a close look into patient's genome sequence which allows to see mutations, e.g. segment deletions, segment duplications, *ii*) the understanding of each aberration's consequences for the phenotype, i.e. patient's health. The former problem is approached by several biotechnological methods, such as karyotyping, Fluorescent in-situ Hybridization (FISH), DNA microarrays, and recently the full genome Next Generation Sequencing (NGS). This thesis is concerned with the analysis of data from Array Comparative Hybridization (aCGH) – a technique based on DNA microarrays, which

is described in Chapter 2 . The latter problem is constantly being solved through gathering of reports on patients, and their experiments results, in databases around the world. In Chapter 5 presented is the analysis connecting data from our IMID2py database of aCGH results from IMID with aforesaid external databases, in this case those are International Standards for Cytogenomic Arrays database (ISCA) (Faucett, 2010), Genetic Association Database (GAD) (Zhang et al., 2010) and Database of Genomic Variants (DGV) (Zhang et al., 2006).

The importance of information infrastructure in translational medicine cannot be overestimated. The publicly available reference sequence of human genome is essential in annotating, relating, and referring data from labs around the world. The arriving technology of Semantic Web, with marvelous tools for storing and referencing graph databases of knowledge, and the Linked Open Data (LOD) publishing method, are slowly transforming the landscape. In this regard in Chapter 6 summarized is an Early Adoption project of Apache Internet Knowledge Stack, where the IMID2py database is extended with LOD semantic data from UniProt and other databases.

The role of computational methods is undeniable either, with the expectation of sophistication constantly growing. The aCGH technology for genomic aberrations detection is based on statistical analysis and transformations of the outputted data. Plethora of software is implemented for this task, a survey of which is provided by Karimpour-Fard et al. (2010). In Chapter 3 of this thesis an integrated solution to the problem of segmentation and, at the same instance, noise separation in aCGH data is proposed; a solution rooted in the Bayesian framework of Graphical models and Markov fields. Furthermore, in Chapter 4 we implement a statistical measure, and a modification of a popular software, the circular binary segmentation (CBS) algorithm, which is used to compare the quality of different aCGH microarray designs.

Detection of DNA copy number changes in patient's genome is crucial in precise diagnosis of genetic diseases, in understanding of thereof, and the aCGH technology was, and still is, pivotal in medicine, one of

the reasons being some of the pathogenic changes are mosaic and not detectable in conventional karyotyping, reports Stankiewicz and Beaudet (2007).

DNA copy number changes, or Copy Number Variations (CNVs), are gains or losses of chromosomal material. They are associated with many types of genomic disorders like mental retardation, congenital malformations, or autism, according to Lupski (2009); Shaw et al. (2004). Genetic aberrations are characteristic of many cancer types and are thought to drive some cancer pathogenesis process, by Lai et al. (2007); O'Hagan et al. (2003); Snijders et al. (2005); Wang et al. (2006).

The aCGH technology is widely used for identification of segmental copy–number alterations in disease genomes, which is corroborated by many publications including Boone et al. (2010); Miller et al. (2010); Perry et al. (2008). In a typical experiment, DNA is extracted from two genomic samples (test vs reference) and labeled differently. Samples are mixed together and then hybridized to a microarray spotted with DNA probes. Signal fluorescent intensities of each spot from both samples are considered to be proportional to the amount of respective genomic sequence present. A more detailed description is found in Chapter 2 .

The aCGH microarrays can be classified into two types. Targeted arrays aim in detection of known, clinically relevant copy number changes and thus provide a better coverage of selected regions, see e. g. Caserta et al. (2008); Thomas et al. (2005). On the other hand, the whole-genome arrays, provide a coverage of the entire genome Barrett et al. (2004). Each design is constrained by the number of DNA probes on the microarray – hundreds of thousands to more than a million probes on a microarray in 2014. Nevertheless, in many applications, especially clinical, the design of the array should combine these two goals resulting with the exploration of the whole genome, with the special focus on certain specific regions (e.g. containing genes related to the disease under study). An exon array CGH approach proposed recently accurately measures copy-number changes of individual exons in the human genome. Chapter 2 contains description and a diagram of an iterative process of aCGH microarray design, and in Chapter 4 a method to compare designs is outlined and tested.

Assigning significance to signals found in aCGH data is a challenging task, a combination of statistical analysis and of human verification by geneticists. Methods proposed in Chapter 5 aim at improving this task. It's an ongoing effort: improving, automating, and verifying protocols for detection of rare CNVs which underlie diverse spectrum of diseases in human, a perfect example from translational medicine.

Collaboration and cooperation catalyzes translational medicine to happen. This thesis stems from the collaboration between the groups of Ania Gambin from the University of Warsaw, who lead bioinformaticians Tomek Gambin, and Maciek Sykulski (the author), and the group of Paweł Stankiewicz, who lead the Cytogenetic Lab at the Institute of Mother and Child (IMID), Warsaw, where the team was Barbara Wiśniowiecka-Kowalnik, Katarzyna Derwińska, Magdalena Bartnik, and others, under supervision of Ewa Bocian. Moreover, Paweł Stankiewicz connected Polish efforts with the efforts of Baylor College of Medicine (BCM), Houston, USA where leading research on aCGH microarrays and their clinical application takes place. The Polish teams at UW and IMID took part in the design and testing of V8.x OLIGO aCGH chip – a custom-designed array with approximately 180,000 selected "best-performing" DNA oligonucleotide probes on it – which is used as a research and diagnostic tool at IMID and BCM, of which reports in print were made by Bartnik et al. (2012, 2014); Boone et al. (2010); Derwińska et al. (2012); Wiśniowiecka-Kowalnik et al. (2013).

1.1 Main Results

We present the process of aCGH technology from particular to general and back. By that understood is the cycle of *i*) signal vs noise considerations of incoming aCGH microarray data, on which the design of microarrays is dependent, *ii*) processing of data stream in order to derive its segmentation: a structured signal, aligned along human genome used as reference coordinate system, this step is often termed Copy Number Variants (CNVs) calling, *iii*) collecting, and referencing collections of, CNVs, either benign or pathogenic, in a suitable data infrastructure,

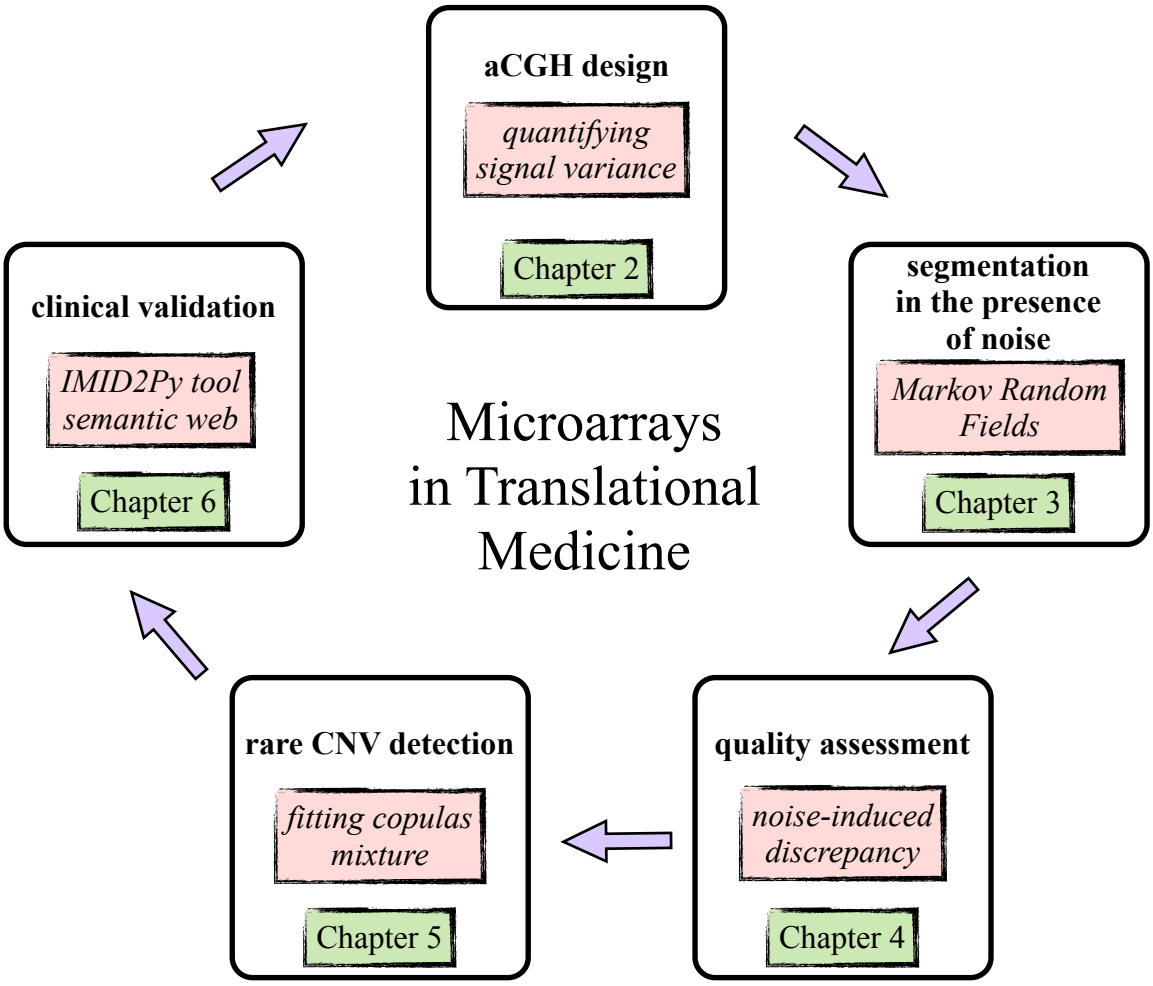


Figure 1.1.1: Plotted is the outline of this dissertation. We aspire to follow the main steps in the process of aCGH design, research, and clinical validation. A more detailed diagram of aCGH design process can be found in Chapter 2 Figure 2.3.2 .

iv) assigning either medical, or evolutionary significance to CNVs, a step involving aggregation of knowledge from various sources on human genetics, where the reference system is either genome sequence coordinates (e.g. genome annotation databases such as UCSC), or names of knowledge network nodes such as genes, proteins, RNA transcripts, transcription factors and their binding sites, *v*) mapping significant knowledge bits back to the genome sequence, to influence the design of aCGH microarrays, with DNA probes printed on which are sensitive to selected genomic aberrations, and which have good signal-to-noise characteristics. The outline of this dissertation aligned with the aforementioned cycle is

sketched on Figure 1.1.1.

In **Chapter 2** we acquaint the reader with the technology of DNA microarrays, and later focus on aCGH method. We analyze characteristics of $\log_2$ratio data from aCGH microarrays, quantify its heteroscedasticity and signal-to-noise ratio on real data, to later introduce $\log_2$ratio segmentation problem, and the most popular Circular Binary Segmentation approach. The section is closed with description and a diagram of the iterative process of targeted microarray design.

The Background and Segments Markov random fields model (BSMF) for segmentation and spatial denoising is declared in **Chapter 3** . This Bayesian Graphical model with conjugate priors, which is a Markov Random Field defined on two graphs: spatial grid, and genomic line, is framed as a partially Quadratic Programming problem, its posterior conditional distributions are given, Expectation Maximization scheme for its optimization is proposed and implemented. The model is then extended with Hidden Markov Model (HMM) state-like prior mixture for segment field, the double linkage modification of genomic neighborhood graph is analyzed. The BSMF Markov Chain possibility is briefly remarked, then the Expectation Maximization scheme for BSMF implementation is explained. Results of the algorithm on real data from IMID2py database are described and plotted, its performance is compared with CBS results, sensitivity to variability in setting of priors is analyzed.

In **Chapter 4** the problem of array design and comparison thereof is taken on. Synthetic data, and modification of real data, with imposed noise is generated. The measure of robustness to noise is proposed for a single DNA probe, and later extended to a whole microarray design resulting with the measure of *relative noise-induced discrepancy*. Method is parametrized by the segmentation algorithm used to identify aberrations. We implemented the efficient Monte Carlo method for testing noise robustness within CBS procedure. Results on synthetic data and in the optimization of a concrete aCGH design are presented.

In **Chapter 5** we propose a novel multiple sample aCGH analysis methodology aiming in rare CNVs detection. The majority of previous approaches dealt with cancer data sets, while we focus on inborn genomic

abnormalities identified in a diverse spectrum of diseases in human. Our method is tested on exon targeted V8.1 OLIGO aCGH microarray by analyzing 366 patients affected with developmental delay/intellectual disability, epilepsy, or autism. The proposed algorithm can be applied as a post-processing filtering to any given segmentation method. With the additional information obtained from multiple samples we efficiently detect significant segments corresponding to rare CNVs responsible for pathogenic changes. The robust statistical framework based on rank statistics applied in our method eliminates the influence of a technical artifact termed in literature as 'waving'.

In **Chapter 6** described are the design and features of IMID2py database used at IMID to gather and analyze aCGH results. Later, we present a semantic extension to our database, namely the results of our Early Adoption project of Apache Internet Knowledge Stack, which involves using the Apache Stanbol software (Auer et al., 2012) to and annotate records in the database with Linked Open Data, and search within it. Uniprot RDF release with Gene Ontology terms, PubMed abstracts, GeneID references is indexed using Stanbol. A concept of a tree of enhancements is introduced, with a set of modules: enhancers, which facilitate certain specific searches within the semantic graph. At the end of the chapter we summarize results obtained by IMID researches with the use of IMID2py database.

1.2 Scientific publications and other published resources

Results in Chapters 4 , 5 stem from the joint work with Tomasz Gambin who published some of these results in his dissertation Gambin (2012). Results in Chapter 3 were obtained in cooperation with Bogusław Kluge, manuscript in preparation, to whom a more detailed thanks are given at the end of the chapter.

Management, perseverance and faith of Ania Gambin, vision, consequence, and vigilance of Paweł Stankiewicz, the atmosphere and the rendition while working with Tomasz Gambin, insights, and proficiency

of Bogusław Kluge, and the hard work of all people from IMID made this thesis possible.

Publications and resources coauthored by the author while working on this dissertation are listed below.

Publications referred to in Chapters 2 .

- P. M. Boone, C. A. Bacino, C. A. Shaw, P. A. Eng, P. M. Hixson, A. N. Pursley, S.-H. L. Kang, Y. Yang, J. Wiszniewska, B. A. Nowakowska, D. del Gaudio, Z. Xia, G. Simpson-Patel, L. L. Immken, J. B. Gibson, A. C.-H. Tsai, J. A. Bowers, T. E. Reimschisel, C. P. Schaaf, L. Potocki, F. Scaglia, T. Gambin, M. Sykulski, M. Bartnik, K. Derwinska, B. Wisniowiecka-Kowalnik, S. R. Lalani, F. J. Probst, W. Bi, A. L. Beaudet, A. Patel, J. R. Lupski, S. W. Cheung, and P. Stankiewicz. Detection of clinically relevant exonic copy-number changes by array CGH. *Human Mutation*, 31(12):1326–1342, 2010. ISSN 1098-1004. doi: 10.1002/humu.21360. URL `http://onlinelibrary.wiley.com/doi/10.1002/humu.21360/abstract` (Boone et al., 2010)

Publication referred to in Chapter 4 .

- T. Gambin, P. Stankiewicz, M. Sykulski, and A. Gambin. Functional performance of aCGH design for clinical cytogenetics. *Computers in Biology and Medicine*, 43(6):775–785, Jan. 2013. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2013.02.008. URL `http://www.computersinbiologyandmedicine.com/article/S0010482513000528/abstract` (Gambin et al., 2013)

Publication referred to in Chapter 5 .

- M. Sykulski, T. Gambin, M. Bartnik, K. Derwinska, B. Wisniowiecka-Kowalnik, P. Stankiewicz, and A. Gambin. Efficient multiple samples aCGH analysis for rare CNVs detection. In *2011 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 406 –409, Nov. 2011. doi: 10.1109/BIBM.2011.38 (Sykulski et al., 2011)

- M. Sykulski, T. Gambin, M. Bartnik, K. Derwinska, B. Wisniowiecka-Kowalnik, P. Stankiewicz, and A. Gambin. Multiple samples aCGH analysis for rare CNVs detection. *Journal of Clinical Bioinformatics*, 3:12, June 2013. ISSN 2043-9113. doi: 10.1186/2043-9113-3-12. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3691624/ (Sykulski et al., 2013)

Publications referred to in Chapters 6 .

- M. Bartnik, E. Szczepanik, K. Derwińska, B. Wiśniowiecka-Kowalnik, T. Gambin, M. Sykulski, K. Ziemkiewicz, M. Kedzior, M. Gos, D. Hoffman-Zacharska, T. Mazurczak, A. Jeziorek, D. Antczak-Marach, M. Rudzka-Dybala, H. Mazurkiewicz, A. Goszczańska-Ciuchta, Z. Zalewska-Miszkurka, I. Terczyńska, M. Sobierajewicz, C. A. Shaw, A. Gambin, H. Mierzewska, T. Mazurczak, E. Obersztyn, E. Bocian, and P. Stankiewicz. Application of array comparative genomic hybridization in 102 patients with epilepsy and additional neurodevelopmental disorders. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 159B(7):760–771, 2012. ISSN 1552-485X. doi: 10.1002/ajmg.b.32081. URL http://onlinelibrary.wiley.com/doi/10.1002/ajmg.b.32081/abstract (Bartnik et al., 2012)

- M. Bartnik, B. Nowakowska, K. Derwińska, B. Wiśniowiecka-Kowalnik, M. Kędzior, J. Bernaciak, K. Ziemkiewicz, T. Gambin, M. Sykulski, N. Bezniakow, L. Korniszewski, A. Kutkowska-Kaźmierczak, J. Klapecki, K. Szczałuba, C. A. Shaw, T. Mazurczak, A. Gambin, E. Obersztyn, E. Bocian, and P. Stankiewicz. Application of array comparative genomic hybridization in 256 patients with developmental delay or intellectual disability. *Journal of Applied Genetics*, 55(1):125–144, Feb. 2014. ISSN 1234-1983, 2190-3883. doi: 10.1007/s13353-013-0181-x. URL http://link.springer.com/article/10.1007/s13353-013-0181-x (Bartnik et al., 2014)

- K. Derwińska, M. Bartnik, B. Wiśniowiecka-Kowalnik, M. Jagla,

A. Rudziński, J. J. Pietrzyk, W. Kawalec, L. Ziólkowska, A. Kutkowska-Kaźmierczak, T. Gambin, M. Sykulski, C. A. Shaw, A. Gambin, T. Mazurczak, E. Obersztyn, E. Bocian, and P. Stankiewicz. Assessment of the role of copy-number variants in 150 patients with congenital heart defects. *Medycyna wieku rozwojowego*, 16(3):175–182, Sept. 2012. ISSN 1428-345X. PMID: 23378395 (Derwińska et al., 2012)

- B. Wiśniowiecka-Kowalnik, M. Kastory-Bronowska, M. Bartnik, K. Derwińska, W. Dymczak-Domini, D. Szumbarska, E. Ziemka, K. Szczałuba, M. Sykulski, T. Gambin, A. Gambin, C. A. Shaw, T. Mazurczak, E. Obersztyn, E. Bocian, and P. Stankiewicz. Application of custom-designed oligonucleotide array CGH in 145 patients with autistic spectrum disorders. *European Journal of Human Genetics*, 21(6):620–625, June 2013. ISSN 1018-4813. doi: 10.1038/ejhg.2012.219. URL `http://www.nature.com/ejhg/journal/v21/n6/abs/ejhg2012219a.html` (Wiśniowiecka-Kowalnik et al., 2013)

Conferences, and other resources referred to in Chapter 6 .

- M. Sykulski and T. Gambin. IMID2py - a database and tools for collection and analysis of aCGH data. In *III Convention of the Polish Bioinformatics Society* ptb (2010). URL `http://www.ptbi3.polsl.pl/files/Program_PTBi_Convention_and_Workshop_2010_ENG.pdf` (Sykulski and Gambin, 2010)

- M. Sykulski. Website: IMiD2py – explore aCGH data, 2012b. the project webpage: `http://bioputer.mimuw.edu.pl/iks/`, the demo webpage: `http://bioputer.mimuw.edu.pl:9442/welcome/` [Online; accessed 5-November-2014] (Sykulski, 2012b)

- M. Sykulski. Videocast: Cytogenetics Lab Stanbol Early Adoption demo, part 1, 2012c. URL `https://www.youtube.com/watch?v=Ua6zN5b3w-M`. [Online; accessed 5-November-2014] (Sykulski, 2012c)

- M. Sykulski. IKS blog: Using Stanbol to enhance medical data by exploring Linked Data, 2012a. URL http://blog.iks-project.eu/using-stanbol-to-enhance-medical-data-by-exploring-linked-data/. [posted on October 25, 2012] Sykulski (2012a)

Before the time working on this dissertation the author coauthored the following publications.

- M. Startek, S. Lasota, M. Sykulski, A. Bulak, A. Gambin, L. Noé, and G. Kucherov. Efficient alternatives to PSI-BLAST. *bulletin of the polish academy of sciences: technical sciences*, 2012. URL http://hal.inria.fr/hal-00749016 (Startek et al., 2012)

- A. Gambin, S. Lasota, M. Startek, M. Sykulski, L. Noé, and G. Kucherov. Subset seed extension to protein BLAST. In *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms (BIOINFORMATICS 2011)*, page 149–158, 2011. URL http://www2.lifl.fr/~noe/files/pp_BIOSTEC11.pdf (Gambin et al., 2011)

2

aCGH experiment data characteristics and microarray design

2.1 DNA microarray technology

In the year 1995 a group of researchers from Stanford University randomly chose 45 complementary DNA clones from *Arabidopsis thaliana* plus 3 control DNA sequences, averaging $\sim$ 1.0kb (kilo-bases), amplified these with PCR, and, with the use of an arraying machine developed in their lab, printed the products, and their duplicates, with the aim to test reproducibility, onto glass slides 3.5mm by 5.5mm each. (Schena et al., 1995) Three printed microarrays were later hybridized[1] with reverse transcribed mRNA from *Arabidopsis thaliana*. To

[1] "Hybridization is the process of establishing a non-covalent, sequence-specific interaction between two or more complementary strands of nucleic acids into a single complex, which in the case of two strands is referred to as a duplex. Oligonucleotides,

minimize experimental variation, two mRNA sources (i.e. root tissue vs leaf tissue) were reverse transcribed to DNA in the presence of fluorescein[2]- and lisamine-labeled nucleotide analogs, respectively. After hybridizing this two-color mixture to one microarray, it was scanned with a laser, and intensity signals were read into a personal computer.

Data returned contained two channels corresponding to two colors used. Under the assumption that scanned intensities correspond to the amount of hybridized DNA material, this enabled to compare gene expression between sources, at the same time minimizing variation in signal from the hybridization process.

At that time already researchers predicted the process to soon scale up the array printing process to produce arrays containing 20,000 cDNA targets. Today DNA microarrays contain up to 2 millions[3] probes (50-75 base pair in length) per array (Agilent Technologies, 2013). One million probe arrays allow to achieve resolution with 2100 median probe spacing on human genome.

DNA microarrays with probes matching to human genome are used in variety of ways: *i*) measuring gene expression, *ii*) DNA aberration detection in postnatal research (Wiśniowiecka-Kowalnik et al., 2013), *iii*) prenatal research, *iv*) cancer research, *v*) genome wide association studies (GWAS) (Consortium, 2012), *vi*) as well as in clinical setting, such as embryo selection during in vitro fertilization (Liu et al., 2013), or prenatal procedures (Van den Veyver et al., 2009).

2.1.1 Types of DNA microarrays

DNA microarrays are produced with several technologies, however, the main concept is the same: a complementary DNA material of specific (possibly short) sequence is attached (printed) on the array surface at a

DNA, or RNA will bind to their complement under normal conditions, so two perfectly complementary strands will bind to each other readily." (Wikipedia)

[2]The idea for fluorescence hybridization was earlier introduced in the experiment type named FISH (fluorescent in-situ hybridization), preceded by ISH experiments (Pardue and Gall, Probes and labeling 1969; John et al., 1969), developed later in 1989 by DeLong (DeLong et al., 1989). FISH is used to this day as a low-throughput verification scheme for aCGH, and in many others applications in microbiology. (Moter and Göbel, 2000)

[3]Affymetrix 6.0 contains over 1.8 million probes.

designated location. On a microarray there are thousands, up to millions, of such locations, with different sequences printed on them.

What makes DNA microarray (chips) and its applications versatile are the various possibilities for:

- the type and the source of the genomic material to be hybridized to a chip, examples include: mRNA, DNA, DNA bound to a particular protein, bisulfite treated DNA where the unmethylated cytosine is converted to uracil, etc.
- dyeing (coloring) of the genomic material to be hybridized to a chip (ref. table 2.1.1); color mixture with reference material
- the choice of the set of DNA sequences printed on a chip

Dyeing Type	Output	Details
one-color	intensity: $\{R_i\}_I$	Allows to measure only absolute signal level, used in gene expression profiling.
two-color	$\log_2$ratio: $\left\{\log_2\left(\frac{R_i}{G_i}\right)\right\}_I$	Allows to measure signal level relative to reference. Fluorescent dyeing is done with Cyanine 3 and Cyanine 5, hence "Red/Green" intensities.
multicolor	vector intensities	"[...] the capacity to simultaneously hybridize eight samples confers an unprecedented flexibility to array-based analyses, providing a 4-fold increase in throughput over standard two-color assays." (Shepard, 2006)

Table 2.1.1: Dyeing types of material hybridized to DNA microarrays.

2.1.2 aCGH data characteristics

Array Comparative Genomic Hybridization (aCGH) is one of the main procedures based on DNA microarrays (for a list of DNA microarray applications see table 2.1.2). In aCGH two colored DNA samples are hybridized to the same DNA microarray and compared: *i*) a reference

Application	Details
Array Comparative Genomic Hybridization (aCGH)	Red-dyed mixture of a sample DNA cut with restriction enzymes is mixed with green-dyed reference DNA (with known properties, e.g. a healthy patient) and then hybridized to microarray.
(aCGH + SNP) microarray	Similar to aCGH, with probes aligned to known Single Nucleotide Polimorphisms (SNPs) locations, and their variants.
Gene expression profiling	One-channel experiment, where transcribed mRNA from a cell is hybridized to a microarray. Levels of gene expression can be measured under variety of conditions.
GeneID	Combination of microarray and PCR technology used to identify organisms, e.g. in food, identifying pathogens.
Chromatin immunoprecipitation on Chip (ChIP)	Isolation, with the use of antibodies, of proteins with DNA sequences bound to them allows to later hybridize these specific sequences to a microarray.
Alternative splicing detection, Fusion genes microarray	Probes are designed to match to the potential splice sites of predicted exons for, or cancerous mutations of, a gene.
Tiling array	A set of overlapping probes densely covering a selected genomic region allows to detect all known, and discover unknown, possible transcripts from the region.

Table 2.1.2: Applications of DNA microarrays.

DNA sample with known properties, *ii*) an examined DNA sample with suspected genomic aberrations.

In the two-channel DNA microarray aCGH experiment an image of a microarray is taken with a scanner, and spots of brightness from the image are transformed into pairs of *Red, Green* intensities:

$$\{R_i, G_i\}_I \tag{2.1}$$

The index sequence I depends on the use case: e.g. I_{array} maps intensities to a (x_i, y_i) positions in the microarray grid of spots, at the same time another indexing I_{genome} maps to specific locus in the reference human

genome (e.g. HG19 Chr9 132423 ← 132482).

Two-channel intensities are usually transformed to a single $\log_2$ratio sequence[4]:

$$\log_2\text{ratio} = \log_2 \frac{R_i}{G_i} \tag{2.2}$$

In this thesis we focus on the analysis of $\log_2$ratio data from DNA microarrays. To understand it's characteristic we first acquaint the reader with the processes from the pipeline which results in obtaining such data. A general overview of the process is presented in the next Section 2.3. A more specific statistical features of $\log_2$ratio data are reviewed in Section 2.2 Signal and noise in DNA microarray data.

2.2 Signal and noise in DNA microarray data

Signal from a hybridized DNA microarray is analyzed and transformed during the following steps of aCGH process:

1. *scanning process:* image analysis, finding image corners, centering grid (Agilent Technologies, 2013),

2. *noise reduction:* background estimation, microarray spatial noise reduction (ref. Section 3)

3. *normalization:* either the simple $\log_2$ratio transformation, however a more sophisticated transformations are proposed in literature: e.g. Variance Stabilization Transform (VST) (Lin et al., 2008). VST agrees with $\log_2$ratio on most data, but on low-hybridized probes.

output: _____ $\log_2$ratio _____

4. *segmentation:* finding regions of elevated, or degraded, hybridization, which correspond to aberrated regions in the genome (e.g. Circularly Binary Segmentation(CBS) algorithm Olshen et al. (2004))

[4]Some authors, as well as some protocols and software (like Agilent Feature Extraction), use another convention of $\log_{10}$ratio. This is rarely relevant. In this thesis, as long as it is not indicated otherwise, we work with $\log_2$ratio data.

5. *CNV detection:* multiple segments from several genomes may correspond to a region of variation, called Copy Number Variant (CNV, ref. chapter 5)

6. *functional mapping:* modifications in genes, transcription factors, SNPs, etc. are mapped to functional groups, such as diseases, signaling pathways, phenotypic traits. The role of large databases that gather data from labs all around the world is unquestionable. Databases referred to in this thesis include UCSC (Meyer et al., 2013), Ensembl (Flicek et al., 2012), Cytogenomic Arrays database (ISCA) (Faucett, 2010), Genetic Association Database (GAD) (Zhang et al., 2010), and Database of Genomic Variants (DGV) (Zhang et al., 2006)

The last two steps vary in technologies other than aCGH. The first three steps usually are performed by the proprietary software delivered by the manufacturer of the microarray, e.g. Agilent Feature Extraction. The building blocks of Agilent Feature Extraction pipeline are enumerated in table 2.2.1).

Agilent Feature Extraction Software Pipeline:
i) Place Grid *ii*) Optimize Grid Fit *iii*) Find Spots *iv*) eXtended Dynamic Range(XDR) extraction (optional, two scans required) *v*) Flag Outliers *vi*) Compute Background, Bias and Error (spatial detrending with LOESS) *vii*) Correct Dye Biases *viii*) Compute Log Ratios *ix*) Calculate QC Metrics

Table 2.2.1: For the complete description of the process see (Agilent Technologies, 2013, page 212). This thesis concerns with $\log_2$ratio analysis which results from the above process. Characteristics of resulting data is elaborated further in this thesis; observed $\log_2$ratio standard deviation is plotted in fig. 2.2.1.

It's important to acknowledge that, although several measures are taken to extract pure signal from microarrays, the intensity data, and resulting $\log_2$ratio data, is quite noisy. As a quantification of this statement we present signal-to-noise estimations in a following Section 2.2.2. To illustrate this fact below we reproduce, and visually verify, the model

for Estimated Feature Variance used in Agilent Feature Extraction software (Agilent Technologies, 2013, page 236).

$$\sigma^2_{Estimated} = \sigma^2_{\text{Labelling/FeatureSynthesis}} + \sigma^2_{\text{Counting}} + \sigma^2_{\text{Noise}} \tag{2.3}$$

$$\sigma^2_{\text{Labelling/FeatureSynthesis}} \propto x^2$$

$$\sigma^2_{\text{Counting}} \propto x$$

$$\sigma^2_{\text{Noise}} = \text{const.}$$

where x is the net signal of a feature.

The $\sigma^2_{\text{Labelling/FeatureSynthesis}}$ term estimates the effects from microarray manufacturing and wet chemistry. These sources of variance turn out to be intensity dependent, it is proportional to the square of the signal.

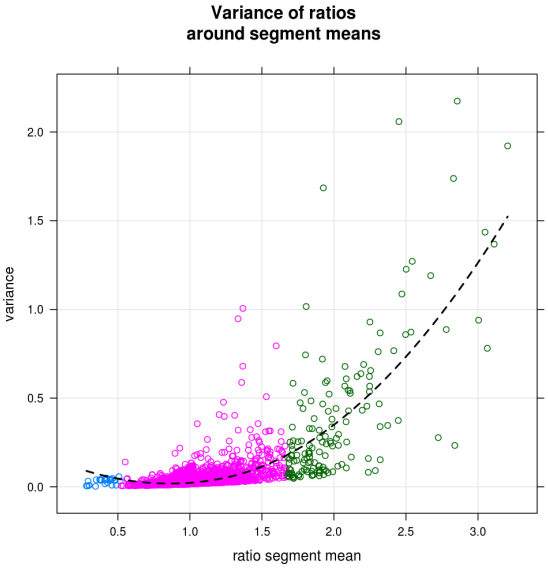
The $\sigma^2_{\text{Counting}}$ term includes noise from scanning process and image analysis (counting pixels and their intensities), it's linearly dependent on intensity.

The σ^2_{Noise} constant term estimates variance from electronic noise in scanner and background level noise in glass.

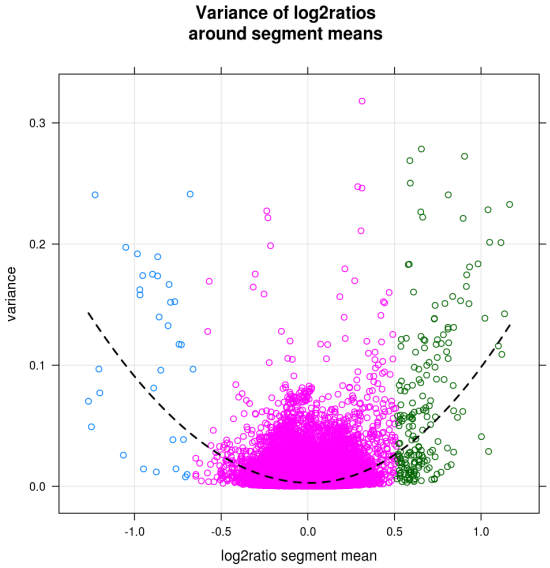
Agilent Feature Extraction software has more than one protocol to determine these terms. Usually, for each analyzed DNA microarray the $\sigma^2_{\text{Labelling/FeatureSynthesis}}$ term is estimated from the coefficient of variation of pixel noise in features. The other two terms are computed by multiplying constants pre-established by Agilent with values depending on the microarray net signal .25 quantile.

2.2.1 Heteroscedasticity in $\log_2$ratio

The model for $\sigma^2_{Estimated}$ is used by Agilent Feature Extraction in background noise estimation protocol, and in quality control for "non-uniformity" outlier calling. Nevertheless, the variance dependence in the outputted $\log_2$ratio signal remains. Figure 2.2.1 purports to visualize this dependence. Here is how this figure was produced. Computing a variance of a signal requires estimating a mean signal intensity (or mean $\log_2$ratio). To do that, we divide data in clusters corresponding to the same signal intensity. This tuns out to be the main task in the analysis



(a) $\sigma^2 \sim$ mean dependence plotted for intensity ratios data (geometric segment means used).



(b) $\sigma^2 \sim$ mean dependence plotted for $\log_2$ratio data.

Figure 2.2.1: We sampled 50K segments, obtained with CBS algorithm having analyzed $\log_2$ratio from 474 patients aCGH arrays (IMID). We computed variances around segments' $\log_2$ratio means. Two plots were produced: *a)* $\log_2$ratio were exponentiated back to intensities ratios to compute variances, *b)* variances of $\log_2$ratio were produced. LOESS parabola fit visually confirms Agilent's quadratic model which applies to intensity data (a). Quadratic dependence persists in $\log_2$ratio data (b). Signal-to-noise $\log_2$ratio statistics for significant segments (left and right subsets) are as following: Min. 0.9309, 1st Qu. 2.2820, Median 3.545, Mean 13.610, 3rd Qu. 6.564

of $\log_2$ratio: segmentation.

Segmentation algorithms, such as Circular Binary Segmentation (Olshen et al., 2004) (detailed in the next Section 2.2.4), cluster features in segments spanning contiguous regions along the genome. To produce Figure 2.2.1 we used 474 results from CBS segmentation algorithm on aCGH arrays. Mean signal intensities were computed, and variances of features around them. We observe that dependence between signal variance and signal intensity remains. Parabolas fitted with local scatterplot smoothing (LOESS) provide visual evidence that quadratic dependency is a valid model for signal intensities ratios, as well as in data transformed to $\log_2$ratio. Verification of this

hypothesis is done with a statistical test similar to Breusch–Pagan test for heteroscedasticity (Breusch and Pagan, 1979). An ordinary least squares model is fitted for squared residuals, and it's F-statistic p-value confirms, or rejects, variance dependence on regressors.

$$\sigma^2_{feature} \sim \beta_0 + \beta_1 \mu_{\text{segment}} + \beta_2 \mu^2_{\text{segment}} \tag{2.4}$$

The summary of this fit in **R** confirms heteroscedasticity and is reprinted below. The significant coefficient, order ~ 0.2 in value, for the quadratic term confirms heteroscedasticity.

```
Call: lm(formula = s2 ~ m + I(m^2))
Residuals:
    Min      1Q  Median      3Q     Max
-0.2476 -0.0044 -0.0034 -0.0006  4.8665
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.629e-03  9.973e-06  464.09   <2e-16 ***
m           -2.153e-02  3.675e-04  -58.59   <2e-16 ***
I(m^2)       1.766e-01  1.574e-03  112.19   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01902 on 3697034 degrees of freedom
Multiple R-squared: 0.003828,Adjusted R-squared: 0.003827
F-statistic:  7103 on 2 and 3697034 DF,  p-value: < 2.2e-16
```

When data is heteroscedastic estimators of variance, confidence intervals, standard errors, coming from linear regression, or based on normality assumption, are biased. This causes errors in statistical tests, if them being not designed specifically for such cases. In the next section the way CBS segmentation algorithm approaches the problem is described. In Chapter 5 we propose a method based on $\log_2$ratio rank statistics, which alleviates the problem of heteroscedasticity.

2.2.2 Signal-to-noise ratio in $\log_2$ratio data

We quantify the amount of noise in the $\log_2$ratio data using signal-to-noise ratio. Another important measure of signal quality used in aCGH setting is Derivative Log2Ratio Spread (DLRS) used in Agilent's Feature Extraction, also used in Leprêtre et al. (2010) to analyze the waving noise phenomenon (described in chapter 5). In literature SNR has several definitions, depending on the nature of a signal source. In this case we use SNR definition as a reciprocal of the *coefficient of variation*.

$$\mathrm{SNR} = \frac{1}{c_v} = \frac{\mu}{\sigma} \tag{2.5}$$

As an indication of signal we use the mean value of $\log_2$ratio in detected significant segment. From 116156 segments in the database (the IMID2py database is described in chapter 6) we've selected all with significant signal[5]:

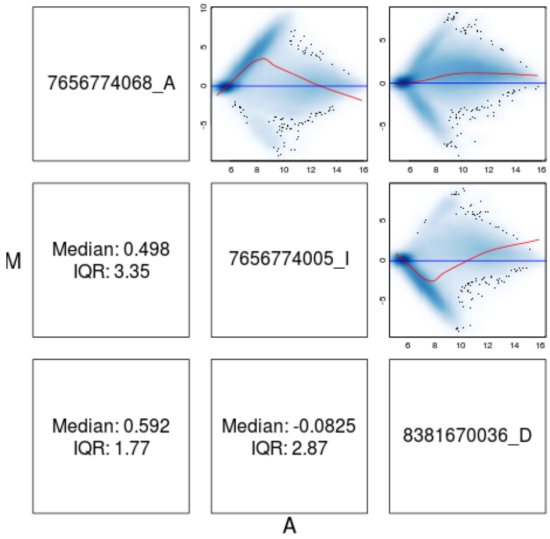
$$\begin{aligned} \left|\mu_{\text{segment}}\right| &> \log_2 \frac{3}{2} \simeq 0.585 \\ \text{length (segment)} &\geq 3 \end{aligned} \tag{2.6}$$

This query returns segments from above 0.9939 quantile of all segments in the database (most segments being no-signal with mean ~ 0), which counts 489 segments, for which we've calculated SNR, taking as an input $\log_2$ratio data.

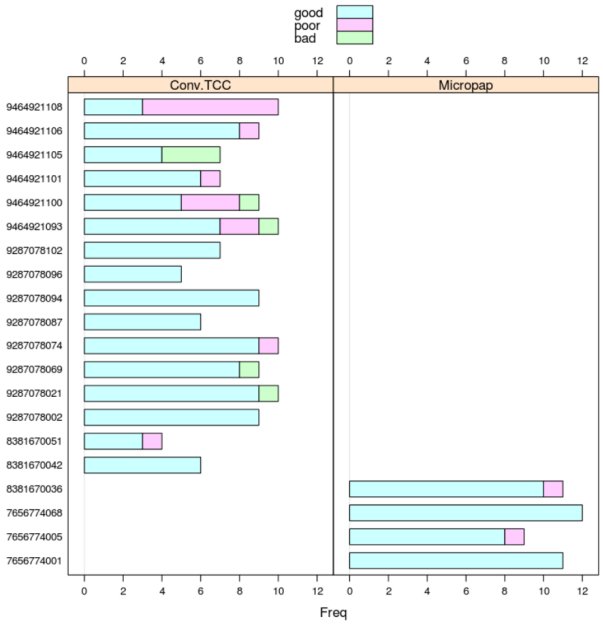
Summary statistics for SNRs based on these segments are printed below:

```
       Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
SNR  0.9102   2.2450   3.2320  11.2800   6.5590 605.4000
```

with more than 10% segments having SNR ≤ 2.



(a) A grid of MA plots of three gene expression arrays vs each other. The middle array is of worse quality, since its intensities are generally lower than other two arrays. (Since an expression array produces signal with only one R channel, data from another array is taken as a second G channel.) Charts produced with **R** package "lumi" (Lin et al., 2008).



(b) Quality of gene expression arrays coming in batches of 12 each. Quality of arrays in the same batch is strongly correlated.

Figure 2.2.2: Plots from the stage of quality control of gene expression arrays form patients with bladder cancer (MD Anderson Cancer Center, manuscript in preparation).

2.2.3 Quality Control

One way to asses a quality of one microarray experiment is MA plot. MA plot is an example of *Bland–Altman plot* (Martin Bland and Altman, 1986), where the vertical axis corresponds to difference of measurements, while the horizontal axis corresponds to average (or sum) of measurements. In the case of

[5]Selecting $\log_2 \frac{3}{2}$ is justified as following. In the case of an organism with a diploid genome, such as human, a duplication of a fragment of a genome (possibly a whole chromosome, as in case of trisomy) results in a duplication of a genomic material, e.g. a gene fragment. If the aforementioned gene fragment is homozygous this results with the change ratio 3/2 of duplicated gene DNA to original DNA

two-color DNA microarray data, a $\log_2$ratio between red-green intensities is plotted vs. its total log-intensity (ref. eqn. 2.7). In the case of one channel DNA microarray, such as a gene expression array, one microarray result is plotted vs. another (ref. fig. 2.2.2a).

$$\begin{aligned} M &= \log\left(R/G\right) &&= \log\left(R\right) - \log\left(G\right) && \text{intensity log-ratio} \\ A &= \frac{1}{2}\log\left(RG\right) &&= \frac{1}{2}\left(\log\left(R\right) + \log\left(G\right)\right) && \text{average log-intensity} \end{aligned} \tag{2.7}$$

MA plot allows to visually inspect microarray results for systematic errors, such as low hybridization of one of the dyes. It is argued in Martin Bland and Altman (1986) that since M is a difference of measurements from the same technology all systematic effects shall cancel, and M shall be approximately Normally distributed, hence 95% of measurements should fall between mean $\pm\, 2\,$sdev. However, in the case of microarrays we shall less be worried about extreme M values, as they very well may indicate genomic signal. On the other hand, a systematic dependence between M and A is an indicator of a problem (ref. fig. 2.2.2a).

Another important characteristic of DNA microarray experiment is that arrays are produced in *batches*: 4–16 arrays of the same type are printed on a larger surface, to be used at the same time. This may be important in analysis, as noise characteristic tends to be more similar between arrays from the same batch (ref. fig. 2.2.2b).

We end this Section Signal and noise in DNA microarray data with a list of noise sources which compose with a genomic signal of a single feature on a microarray.

chemistry, hybridization: probe melting temperature[6], uneven spread of substrates on the microarray introduces spatial trends

[6]Temperature at which half of the DNA duplexes become single stranded due to breaking of the hydrogen bonds between nucleic bases, after which the two strands separate. It's dependent on length and GC content of the sequence, can be approximated with this formula: $T_m \approx 4(\#G + \#C) + 2(\#A + \#T)\ ^\circ C$ (Dieffenbach et al., 1993).

(ref. Section 3)

probe sequence specificity: possibly more than one region similar to a feature DNA sequence is present in the target genome

quality of genomic material: tissue from which genomic material is extracted may contain heterogeneous cell populations, DNA samples may have been stored frozen, etc.

scan image analysis: errors induced in image processing (ref. Figure 2.2.3)

local trends in a genomic sequence: possible explanation include influence of histons on DNA restriction enzymes. (Olshen et al., 2004, end of section 2) (Leprêtre et al., 2010) Chapter 5 validates existence of these trends, and proposes a method to deal with them.

All these together make up a challenge for segmentation algorithms to extract common clusters of genomic material intensities, later to be mapped to functional genomic components (like genes, transcription factors, binding sites, etc.). Only with a successful segmentation and functional mapping a higher level analyses of diseases, genomic disorders, genomic variance in populations, follow.

2.2.4 aCGH data segmentation task

It is clear from previous sections that each aCGH experiment result merges random noise from several sources. However, the interesting part to medical researchers is *genomic signal*. By that phrase we refer to a set of segments aligned on the reference human genome (e.g. HG 18) which correspond to Copy Number Variants, i.e. deletions, duplications, or more complicated rearrangements.

Precisely, since each (non-control) probe on a aCGH microarray matches with its sequence to a specific place in the reference human genome, we imagine an aCGH result as points aligned over a line with x coordinates corresponding to the published reference genome (actually, over a set of disconnected segments corresponding to chromosomes), see

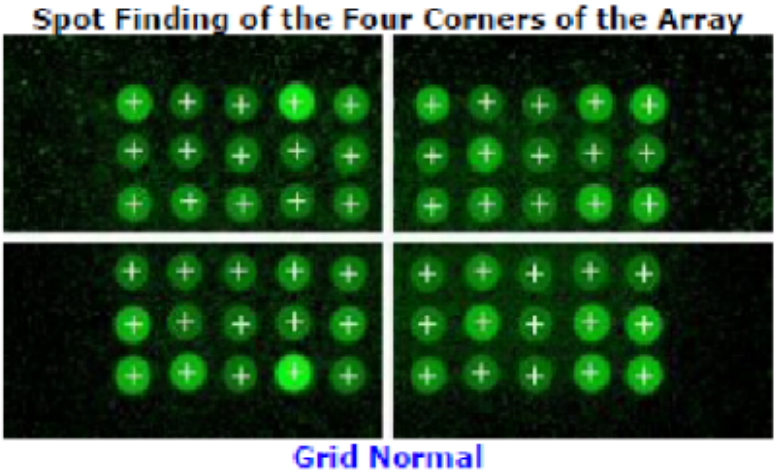


Figure 2.2.3: The grid alignment on Agilent DNA microarray. An image from Feature Extraction software. Misalignment, rotation, wrong fit to the grid may cause several, or many, probes to report biased data.

Figure 2.2.4. Thus, a set of k subsequent DNA probes with high $\log_2$ratio corresponds to a segment on the reference genome, which is duplicated in patient's genome. The task is to discover *breakpoints* between segments of significantly different copy number (most often 1). Breakpoints place between the DNA probes, and they relate with neighboring segments with different CNVs in the following way.

$$\begin{aligned} &\mathrm{Brkp}_n = (i, i+1) \wedge \mathrm{Brkp}_n = (j, j+1) \wedge \mathrm{Brkp}_{n+1} = (k, k+1) \\ &\quad \Leftrightarrow \\ &\mathrm{CNV}\left(\mathrm{Seg}\left[\mathrm{Brkp}_{n-1}[1],\, \mathrm{Brkp}_n[0]\right]\right) \neq \mathrm{CNV}\left(\mathrm{Seg}\left[\mathrm{Brkp}_n[1],\, \mathrm{Brkp}_{n+1}[0]\right]\right) \end{aligned} \tag{2.8}$$

Mean of $\log_2$ratio values in a segment $\mathrm{mean}(\log_2\mathrm{ratio}\left(\mathrm{Seg}\left[a, b\right]\right))$ is what witnesses of segment's copy number.

The theoretical mean values of segments take different values depending on the precise genetics behind, which we leave out of the scope here. An example theoretical value of $\log_2\mathrm{ratio} = 1$ should correspond to a duplication in a haploid genome. Human genome is diploid, however. We suffice to say that $\log_2$ratio mean values outside of $(-0.2, 0.2)$ are of medical interest.

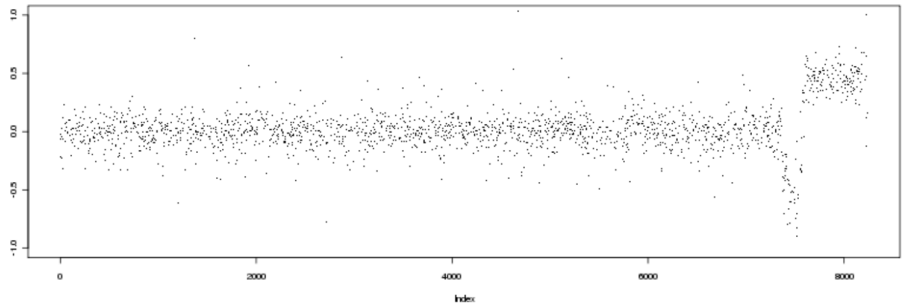


Figure 2.2.4: $\log_2$ratio data from aCGH experiment aligned along a chromosome. Aberration at the right end of the chromosome, i.e. shorter deletion and longer duplication is yet to be detected by a segmentation algorithm. Since the aberrated segments have substantial length this wouldn't be a problem for an algorithm, however shorter segments, or segments with lower mean are more difficult to detect.

Most proposed solutions to recovering CNVs from aCGH data rely on segmentation methods that try to divide the data into segments representing aberrant and normal regions Ben-Yaacov and Eldar (2008a); Cahan et al. (2008); Díaz-Uriarte and Rueda (2007); Lipson et al. (2006).

Comparative studies published so far, e.g. Willenbrock and Fridlyand (2005), nominate Circular Binary Segmentation (CBS), as one of best performing methods for finding copy number segments. The CBS algorithm is implemented e.g. in **R** package DNAcopy (Olshen et al., 2004). For a survey of surveys on aCGH algorithms we refer the reader to Karimpour-Fard et al. (2010).

CBS segmentation algorithm

Segmentation, equivalently a set of breakpoints, can be found recursively. Starting from one large segment for a chromosome, it may be split into two smaller ones. Denote S_n as a sum of $\log_2$ratio values from beginning of a segment up to n-th probe. Then, under assumption of uniform variance $\sigma^2 = 1$, the procedure of *Binary Segmentation*, proposed by Sen and Srivastava (1975), is based on a following t-test-like statistic:

$$Z_i = \frac{\frac{S_i}{i} - \frac{S_n - S_i}{n-i}}{\sqrt{\frac{1}{i} + \frac{1}{n-i}}}, \qquad Z_{BS} = \max_{1 \le i < n} Z_i \tag{2.9}$$

A segment is divided in two smaller segments, if Z_{BS} is above a threshold z-value.

Olshen et al. (2004) propose a slight, but important, modification to binary segmentation, namely they seek not for one but two breakpoints at once. This allows to detect relatively short segments of outstanding intensity, which, in case of binary segmentation, would sink in neighboring noise.

$$Z_{ij} = \frac{\frac{S_j - S_i}{j-i} - \frac{S_n - (S_j - S_i)}{n-j+i}}{\sqrt{\frac{1}{j-i} + \frac{1}{n-j+i}}}, \qquad Z_C = \max_{1 \le i < j \le n} Z_{ij} \tag{2.10}$$

This time, a segment is divided in 3 smaller segments $[1, i-1], [i, j-1], [j+1, n]$ Z_C is above a threshold. The above Z_{ij} statistics is a two-sample t-test statistic (as well as Welch t-test statistics) with both sample variances assumed equal $\sigma_1^2 = \sigma_2^2 = 1$.

In their article Olshen et al. (2004) are aware of heteroscedasticity, non-normality of $\log_2$ratio data. For this reason, instead of extending statistic from Sen and Srivastava (1975) they propose and implement a test based on Z_C cut off values estimation from permuted data. In Chapter 4 we explore and modify **R** package DNAcopy, which implements permutation based CBS, to quantify its, and aCGH design's, robustness to additive noise.

2.3 The iterative process of targeted microarray design

Every design has its constraints. In the case of aCGH microarrays the main constraint is the number of nucleotide probes that can be printed on a chip.

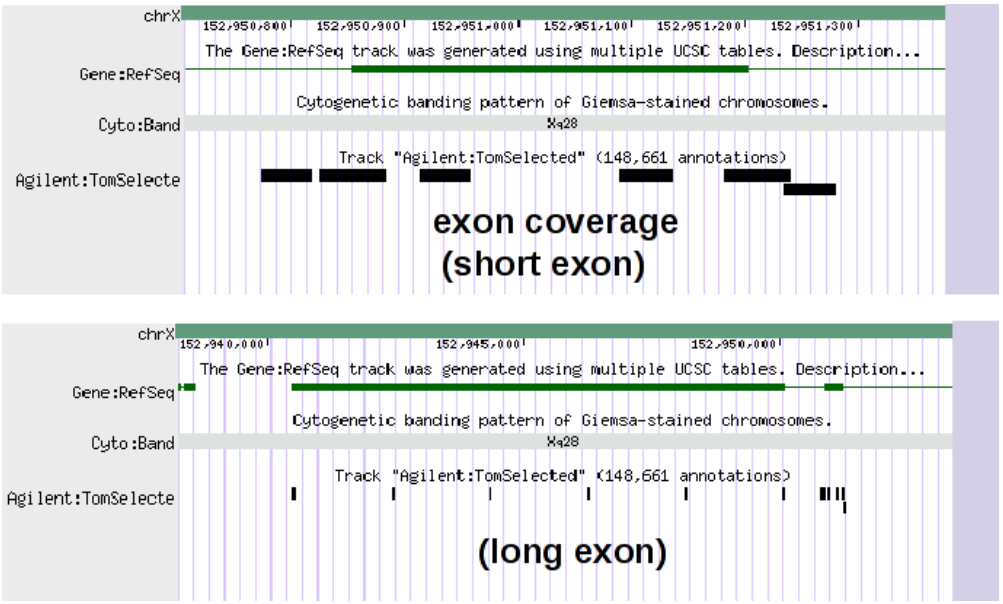


Figure 2.3.1: Designing microarray for coverage of specific regions.In version 8 there are 180 000 probes, each 60bp in length.

The design of the V8 OLIGO chip involved two stages. First, the prototype covering only exonic and microRNA regions was constructed. The main aim at this stage was to develop the array that allows detecting DNA copy number changes of the single exon. Therefore, it was postulated to cover each exon by the same number of oligos. For a given set of 1714 selected genes (including those related to epilepsy, autism, heart defects, mental disorders and other known pathologies) it was decided that each exon would be covered by approximately 6 probes, ref. Figure 2.3.1.

The prototype coverage was two times denser than the desired one in the final version. A set of hybridizations was performed with the prototype version. Performance score of each probe was computed as following: segmentation procedure was applied on data from these experiments. Let us call the empirical cumulative distribution function for distribution of $\log_2$ratio deviations from their segments means $\mathcal{F}$. The distribution $\mathcal{F}$ was estimated from all experiments from the prototype version. For each probe we perform two sided Kolomogorov-Smirnov (K-S) test comparing the $\log_2$ratio deviation from segment mean with distribution $\mathcal{F}$. We assign the p-value obtained in this test as a score of the probe.

Next step involved combining the prototype design with backbone,

i.e., probes putted uniformly across the genome. Densely covered regions, exonic double covered regions were thinned with heuristic approach which considered previously assigned scores and uniformity of nascent coverage (sizes of introduced gaps).

Another step was to cluster MicroRNA probes by distance on the genome to later thin MicroRNA coverage in clustered regions, removing worst performing probes first.

Finally, the design consists of rare (distributed every 10 Kbp) probes in intronic regions and 70 K probes putted uniformly across the genome (backbone).

The iterative process of aCGH microarray design is depicted on Figure 2.3.2.

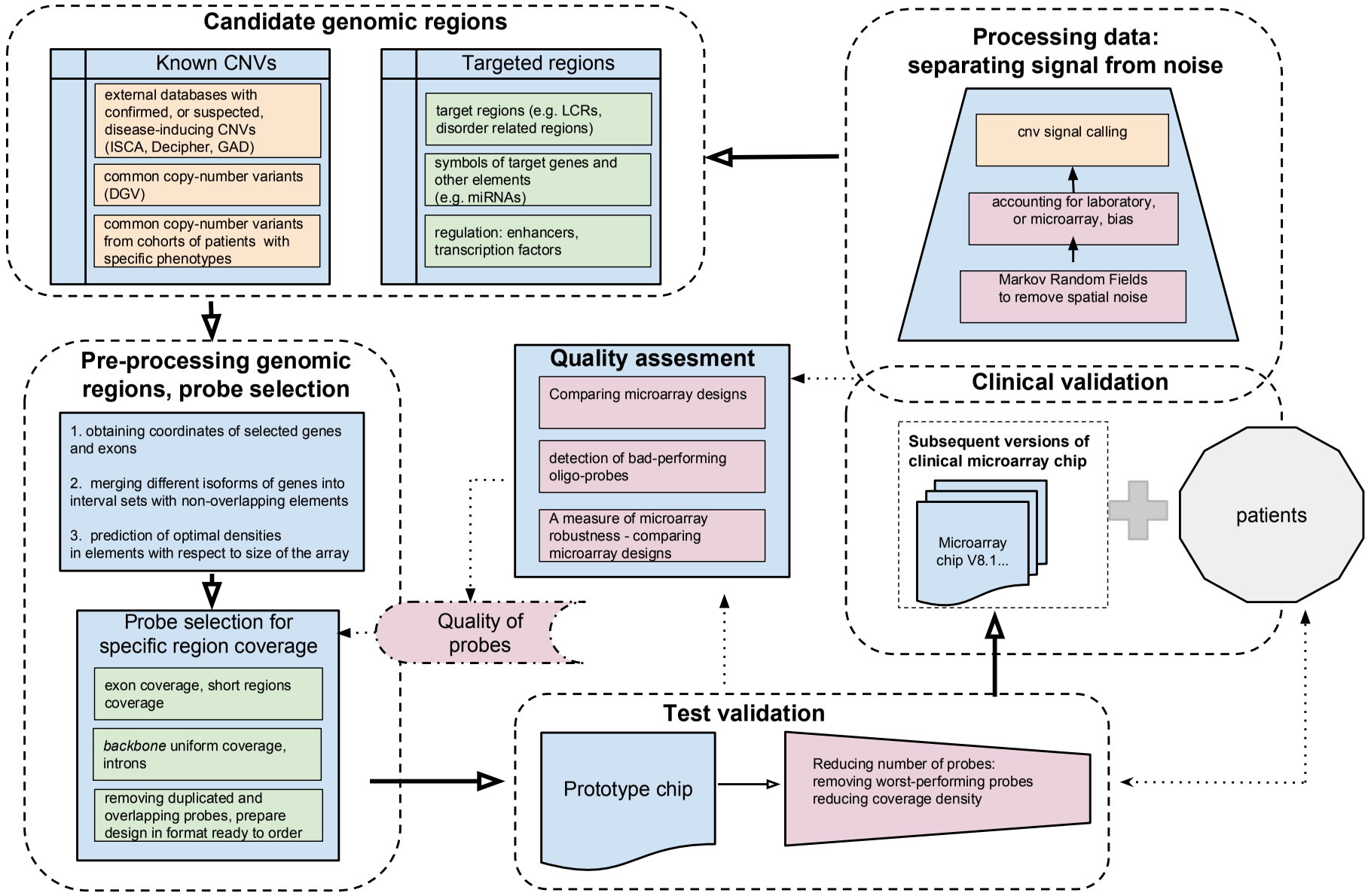


Figure 2.3.2: The iterative process of aCGH microarray design. Solid lines represent the main process and data flows. Dotted lines represent feedback of results to subsequent, or same, iterations. Type of information by colors; *orange*: genomic copy number information (CNVs); *green*: regions in the genome either to be covered, or covered by a set of aCGH DNA probes; *violet*: noise reduction, quality control data.

Control is achieved through learning.
Change is achieved through understanding.

Dalai Lama XIV

3

Markov Random Fields for aCGH segmentation and spatial denoising

The chemistry of the aCGH hybridization process requires spilling and evenly distributing a chemical mixture on a microarray. Without going into the details of the laboratory procedure, we confirm the existence of spatial artifacts on DNA microarray scans in our database in roughly 20%–25% of cases, an example is found on Figure 3.0.1. The Agilent Technologies (2013) Feature Extraction software provides spatial detrend option: a surface fitted with 2D-Loess algorithm on the mean intensities of the negative control probes is treated as a background component and subtracted from the signal. 2D-Loess regression can also be used in a later normalization step. Zahurak et al. (2007) reports that in Agilent FE simple Loess normalization without background subtraction resulted in low variance fold changes that more reliably rank gene expression. In our microarray results database IMID2py, in $\log_2$ratio data, outputted

from Agilent FE, we observe that FE spatial detrending does not fully remove artifacts, the most probable reason being the fact that artifacts are non-linear. Such non-linear artifacts from IMID2py database are depicted on Figure 3.0.1. An important assumption, realized in practice, is of the spatial distribution of DNA probes being independent of mapping of the probes onto the genome sequence.

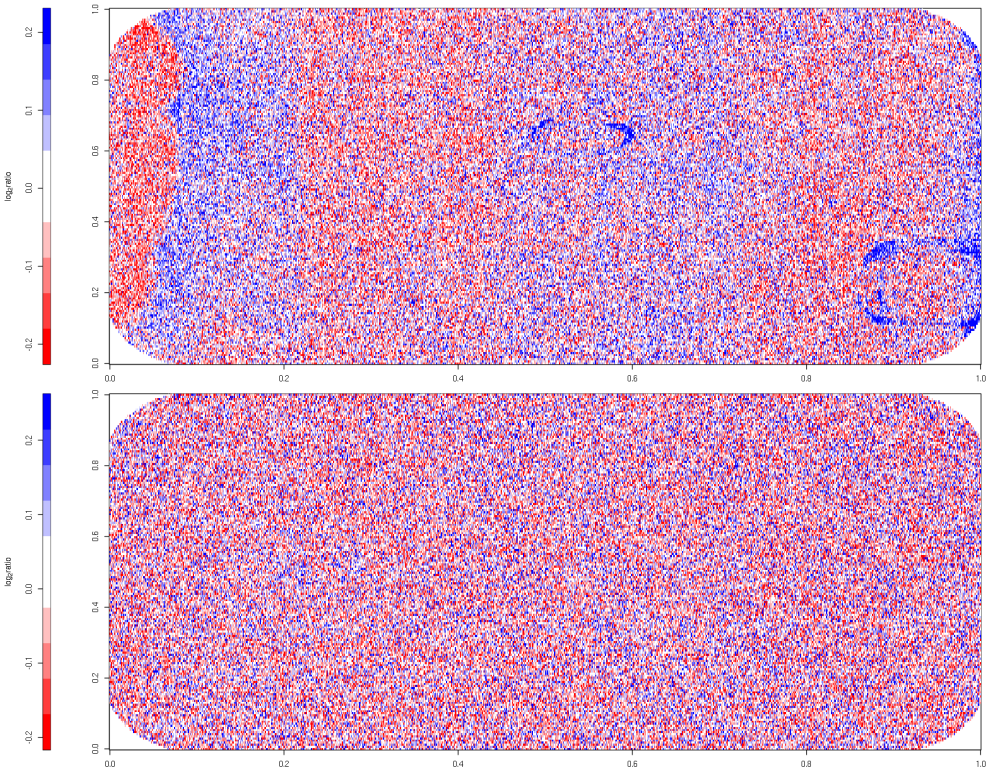


Figure 3.0.1: Noise in $\log_2$ratio from aCGH microarrays scans. Top: a microarray scan containing linear and non-linear spatial trends, bottom: a well hybridized array with no visible artifacts.

The approaches to the problem of microarray spatial detrending involve methods such as: *i*) spatial gradient normalization with two-dimensional Gaussian function (Workman et al., 2002), *ii*) splines with spatial autocorrelation (Baird et al., 2004), *iii*) LOESS fit based on MA plot and residuals (Wilson et al., 2003), *iv*) feed-forward neural network (Tarca et al., 2005), *v*) Hidden Markov Model (Shah et al., 2006). For a survey of these methods we refer to works of Khojasteh et al. (2005), and Neuvial et al. (2006).

Neuvial et al. (2006) in their work propose a method which explicitly deals with strong non-linearity of aberrations, namely, their procedure involves segmentation of the array into spatial areas with similar trend values. Their Neighborhood Expectation Maximization algorithm (NEM, first proposed by Ambroise et al. (1997)) maximizes log-likelihood of a Gaussian mixture with added quadratic "geographic" term dependent on spatial neighborhood of probes:

$$\mathrm{Geo}(c) = \beta \frac{1}{2} \sum_{\{i,j\} \in E_{\mathrm{spatial}}} \sum_{k \in \mathrm{Clusters}} c_{ik} c_{jk} \tag{3.1}$$

where c_{ik} is the expectation of the probe i belonging to the cluster k, and β controls the importance of spatial bounds. Picard et al. (2007) introduce a mixture model with a Gaussian field of segments allowing for breaks in the field, however they do not include spatial noise considerations in their approach. In the next section we introduce a graphical model which merges the two aforementioned methods. Next we describe Expectation Maximization algorithm to estimate parameters of our model, and present its results.

3.1 Outline of the proposal

The following approach to microarray spatial detrending is similar to Neuvial et al. (2006) by the common underlying assumption of the existence of spatial segmentation of a background noise. Our formulation allows to model locally linear trends of noise and non-linear artifacts, as well as genomic signal segmentation in later installments. The proposed model are Gaussian Markov fields on two graphs (equivalently on one graph with two subsets of edges) with Bayesian hierarchical prior probabilities, extended with a Hidden Markov Model in the final installment. We are aware of several modeling systems, and packages, for hierarchical Bayesian interference,to name a few: Stan Modeling Language (Stan Development Team, 2014), flexmix **R** package (Leisch, 2003), BUGS/JAGS (Lunn et al., 2009). We're not convinced that these packages have expressive power to define our model,

or in case they do, to efficiently optimize solution on practical large data, although we leave it as a possibility. Optimization of our model is partially a linear problem, the equivalence between solving linear problems and belief propagation on Gaussian fields is explored extensively by Bickson (2008).

3.2 Background and Segments Markov random fields model declaration

The basic background Gaussian field model. Every microarray is a 2-dimensional surface, with DNA probes printed on it and aligned on a grid. This implies existence of an undirected G_{spatial} neighborhood graph

$$\begin{aligned} G_{\text{spatial}} &= (V, E_{\text{spatial}}) \qquad \text{where} \\ V &= \{i : \text{corresponds to } i\text{-th feature}\} \\ E_{\text{spatial}} &= \left\{\{i,j\} : i,j \text{ are neighbors on the microarray grid graph}\right\}. \end{aligned} \tag{3.2}$$

A scan image of a grid of features from an Agilent DNA microarray can be seen on Figure 2.2.3. We choose to interpret this square lattice, after addition of "top-left-bottom-right" diagonal edges, as a slightly row-shifted hexagonal lattice. The two main reasons for choosing such a lattice are:

- stronger connectedness: after removal of one edge the shortest path between previously neighboring vertices is of length 2 (while for square lattice it's length 4).
- at the same time there exists a regular non-incident vertex coloring with 3 colors, which we call a stratification to 3 independent subsets

A hexagonal lattice graph, and its stratification, is depicted on Figure 3.4.1.

Given the $\log_2$ratio data x_i, we assume the existence of a *background noise field* b_i which is a Gaussian Markov random field with variance $1/\tau$

with neighborhood taken from the graph G_{spatial}.

$$\overset{1}{P}_{\text{background}} \propto \left(\prod_{\{i,j\} \in E_{\text{spatial}}} \sqrt{\frac{\tau}{2\pi}} \right) \exp \left(-\frac{1}{2} \sum_{\{i,j\} \in E_{\text{spatial}}} \tau (b_i - b_j)^2 \right) \tag{3.3}$$

In the simplest model we assume that $\log_2$ratio data x_i is normally distributed around the background field with variance $1/\nu$ (which is simplistic, since we know that x_i also contains signal from genomic deletions or duplications).

$$\overset{1}{P}_{\text{data}} \propto \left(\prod_{i \in V} \sqrt{\frac{\nu}{2\pi}} \right) \exp \left(-\frac{1}{2} \sum_{i \in V} \nu (x_i - b_i)^2 \right) \tag{3.4}$$

Prior distributions for τ, ν are chosen to Gamma distributions, which are conjugate priors to 1/variance (precision) of a Normal distribution. We set $P_{\text{prior b}}$, a prior for b_i, to $\mathcal{N}(0, \tau_b)$, a conjugate prior to mean of a Normal distribution.

$$\begin{aligned} \overset{1}{P}_{\text{prior}} =& \tau^{\Omega_{\tau 0}-1} \exp(-\tau \Omega_{\tau 1}) \cdot \nu^{\Omega_{\nu 0}-1} \exp(-\nu \Omega_{\nu 1}) \\ & \cdot \left(\prod_{i \in V} \sqrt{\frac{\tau_b}{2\pi}} \right) \exp \left(-\frac{1}{2} \sum_{i \in V} \tau_b (0 - b_i)^2 \right) \\ & \cdot \frac{\Omega_{\tau 1}^{\Omega_{\tau 0}}}{\Gamma(\Omega_{\tau 0})} \frac{\Omega_{\nu 1}^{\Omega_{\nu 0}}}{\Gamma(\Omega_{\nu 0})} \qquad \text{/* Gamma normalizing constants */} \end{aligned} \tag{3.5}$$

The required constants for the above set up of priors are $\Omega = \{\Omega_{\tau 0}, \Omega_{\tau 1}, \Omega_{\nu 0}, \Omega_{\nu 1}, \tau_b\}$.

This renders the total log-posterior likelihood for $\theta = \{b, \tau, \nu\}$ to

$$\log \overset{1}{L}(\theta; x, \Omega) = \log \overset{1}{P}_{\text{background}} + \log \overset{1}{P}_{\text{data}} + \log \overset{1}{P}_{\text{prior}}. \tag{3.6}$$

Introducing breaks in the background field. In the second installment breaks in the background field are fixed into the model. A new set of latent $0, 1$ variables is introduced: a priori distributed

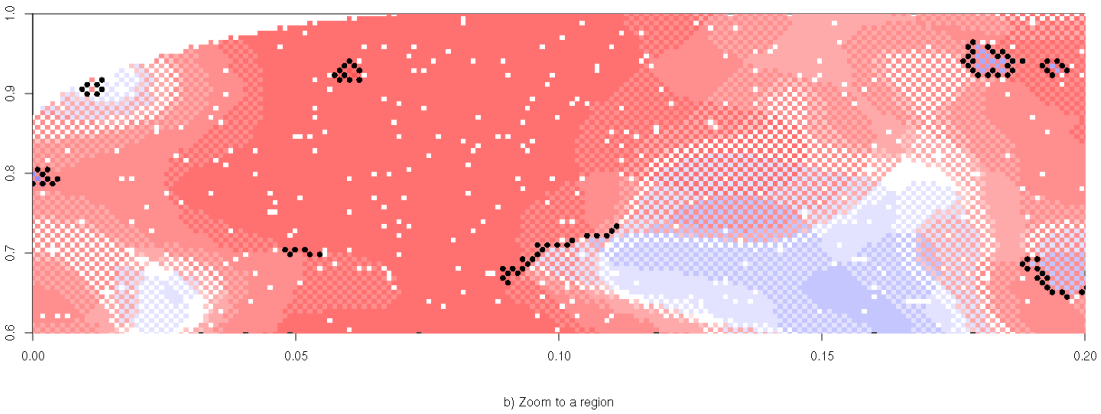


Figure 3.2.1: Zoom to the top left corner of a microarray. Background noise field levels marked in red/blue, breaks in the field marked black, white dots are spots without a probe.

$y_{ij} \sim \text{Bernoulli}(q)$, where q corresponds to the probability of a violent break in the background field (an expected proportion of violent breaks between neighbors in the field). In other words, $y_{ij} == 0$ indicates that two spatial neighboring features i, j differ to much to be described as a Gaussian field (ref. Figure 3.2.1).

$$\overset{2}{P}_{\text{background}} \propto \left(\prod_{\{i,j\} \in E_{\text{spatial}}} \sqrt{\frac{\tau}{2\pi}}^{y_{ij}} \right) \exp\left(-\frac{1}{2} \sum_{\{i,j\} \in E_{\text{spatial}}} y_{ij} \tau (b_i - b_j)^2 \right) \cdot \prod_{\{i,j\} \in E_{\text{spatial}}} (1-q)^{y_{ij}} q^{1-y_{ij}} \tag{3.7}$$

Similarly as before, we introduce a conjugate prior to Bernoulli distribution for q, namely $\text{Beta}(\Omega_{q0}, \Omega_{q1})$ distribution.

$$\begin{aligned} \overset{2}{P}_{\text{prior}} =& \overset{1}{P}_{\text{prior}} \cdot q^{\Omega_{q0}-1}(1-q)^{\Omega_{q1}-1} \\ & \cdot 1/\text{B}\left(\Omega_{q0}, \Omega_{q1}\right) \qquad \text{/* Beta normalizing constant */} \end{aligned} \tag{3.8}$$

The total log-posterior likelihood for $\theta = \{b, y, \tau, \nu, q\}$ and discrete variables $Z = \{y\}$ becomes

$$\log \overset{2}{L}(\theta, Z; x, \Omega) = \log \overset{2}{P}_{\text{background}} + \log \overset{1}{P}_{\text{data}} + \log \overset{2}{P}_{\text{prior}}. \tag{3.9}$$

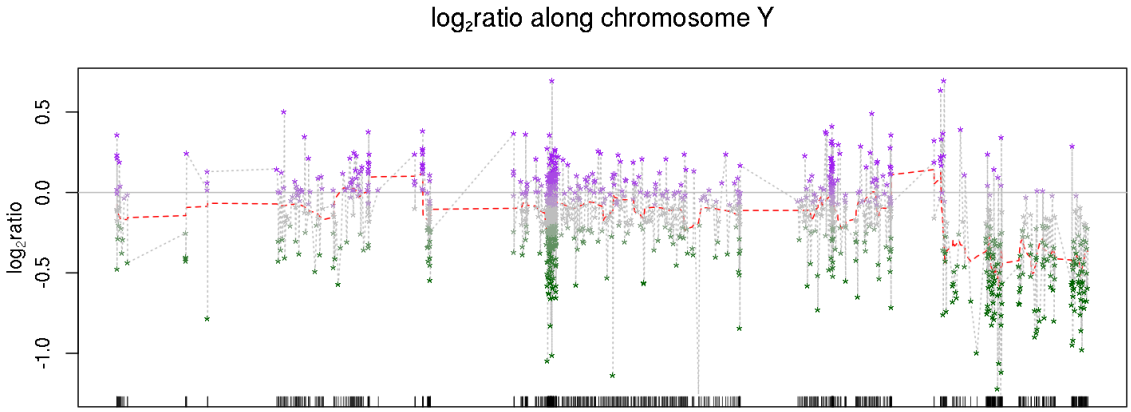


Figure 3.2.2: $\log_2$ratio data, scanned from aCGH microarray, aligned along chromosome Y. Red dashed line represents hidden genomic signal segments. A deletion on the right end of the chromosome is present (the segment with a significantly lowered $\log_2$ratio mean ~ -0.5).

Introducing the segments field along a genome. In the final installment of the Gaussian Markov Field model we address the hidden signal inside x_i $\log_2$ratio data. The natural way to look at $\log_2$ratio data is the undirected line graph G_{genome} along the genome, that is along subsequent chromosomes, as depicted on Figure 3.2.2.

$$\begin{aligned} G_{\text{genome}} &= (V, E_{\text{genome}}) \qquad \text{where} \\ V &= \{i : \text{ corresponds to } i\text{-th feature}\} \\ E_{\text{genome}} &= \big\{\{i, j\} : \ i, j \text{ are probes on the same chromosome} \\ &\qquad\qquad \text{and there is no other probe between them}\big\} \end{aligned} \tag{3.10}$$

If we choose the indexing of the feature set V to agree with the genomic order, then the set of edges E_{genome} is roughly $\big\{\{i, i+1\}\big\}_I$.

Again we assume the existence of *the segment signal field* a_i, which is a Gaussian Markov random field with variance $1/\rho$ with its neighborhood taken from the linear graph G_{genome}. Breaks in the segment field are curated by a set of latent $0, 1$ variables, a priori distributed $z_{ij} \sim \text{Bernoulli}(p)$, where p corresponds to the probability of a violent break in the segment field. Now $z_{ij} == 0$ indicates that signals on two genomic neighboring features i, j differ to much to be described as coming from the same Gaussian field segment.

$$\begin{aligned} P_{\text{segments}} \propto & \left(\prod_{\{i,j\} \in E_{\text{genome}}} \sqrt{\frac{\rho}{2\pi}}^{z_{ij}} \right) \\ & \cdot \exp\left(-\frac{1}{2} \sum_{\{i,j\} \in E_{\text{genome}}} z_{ij} \rho (a_i - a_j)^2 \right) \\ & \cdot \prod_{\{i,j\} \in E_{\text{genome}}} (1-p)^{z_{ij}} p^{1-z_{ij}} \end{aligned} \tag{3.11}$$

This time we assume that $\log_2$ratio data x_i is normally distributed around the background field *and* the segment field.

$$\overset{2}{P}_{\text{data}} \propto \left(\prod_{i \in V} \sqrt{\frac{\nu}{2\pi}} \right) \exp\left(-\frac{1}{2} \sum_{i \in V} \nu (x_i - a_i - b_i)^2 \right) \tag{3.12}$$

The above distribution corresponds to other sources of noise, such as probe specificity, hybridization chemistry. In Section 2.2 heteroscedastic nature of this noise was argued, this is not yet addressed here. Related to this issue is the problem of setting a prior distribution for a_i. One simple possibility is to setup a Normal distribution with a large variance, however, it is false to assume that segment $\log_2$ratio means of genomic signal are normally distributed around 0. In Section 3.2.3 an a_i prior probability is proposed which addresses heteroscedasticity and non-normality. Conjugate prior distributions are assumed for ρ and p,

namely Gamma and Beta.

$$\begin{aligned}\overset{3}{P}_{\text{prior}} =& \overset{2}{P}_{\text{prior}} \cdot \rho^{\Omega_{\rho 0}-1} \exp\left(-\rho\Omega_{\rho 1}\right) \\ & \cdot p^{\Omega_{p0}-1}(1-p)^{\Omega_{p1}-1} \\ & \cdot P_{\text{prior } a} \\ & \cdot \frac{\Omega_{\rho 1}^{\Omega_{\rho 0}}}{\Gamma\left(\Omega_{\rho 0}\right)} \frac{1}{\mathrm{B}\left(\Omega_{p0}, \Omega_{p1}\right)}\end{aligned} \tag{3.13}$$

The complete likelihood for BSMF. The total posterior likelihood for Background-Segment-Markov-Field problem (BSMF), where $\theta = \{a, b, \tau, \nu, \rho, q, p\}$, where $Z = \{y, z\}$, and given a set of constant parameters $\Omega = \{(\Omega_{\tau 0}, \Omega_{\tau 1}), (\Omega_{\nu 0}, \Omega_{\nu 1}), (\Omega_{\rho 0}, \Omega_{\rho 1}), (\Omega_{q0}, \Omega_{q1}), (\Omega_{p0}, \Omega_{p1}), \tau_b, \Omega_a\}$, is

$$\begin{aligned}&\mathcal{L}_{\text{BSMF}}\left(\theta, Z; x, \Omega\right) \propto \\ &\left(\prod_{\{i,j\}\in E_{\text{spatial}}} \sqrt{\frac{\tau}{2\pi}}^{y_{ij}}\right) \exp\left(-\frac{1}{2} \sum_{\{i,j\}\in E_{\text{spatial}}} y_{ij}\tau(b_i - b_j)^2\right) \\ &\cdot \prod_{\{i,j\}\in E_{\text{spatial}}} (1-q)^{y_{ij}} q^{1-y_{ij}} \\ &\cdot \left(\prod_{\{i,j\}\in E_{\text{genome}}} \sqrt{\frac{\rho}{2\pi}}^{z_{ij}}\right) \exp\left(-\frac{1}{2} \sum_{\{i,j\}\in E_{\text{genome}}} z_{ij}\rho(a_i - a_j)^2\right) \\ &\cdot \prod_{\{i,j\}\in E_{\text{genome}}} (1-p)^{z_{ij}} p^{1-z_{ij}} \\ &\cdot \left(\prod_i \sqrt{\frac{\nu}{2\pi}}\right) \exp\left(-\frac{1}{2} \sum_i \nu\left(x_i - a_i - b_i\right)^2\right) \\ &\cdot \tau^{\Omega_{\tau 0}-1} \exp\left(-\tau\Omega_{\tau 1}\right)\ q^{\Omega_{q0}-1}(1-q)^{\Omega_{q1}-1} \\ &\cdot \rho^{\Omega_{\rho 0}-1} \exp\left(-\rho\Omega_{\rho 1}\right)\ p^{\Omega_{p0}-1}(1-p)^{\Omega_{p1}-1} \\ &\cdot \nu^{\Omega_{\nu 0}-1} \exp\left(-\nu\Omega_{\nu 1}\right) \\ &\cdot P_{\text{prior a}} \cdot P_{\text{prior b}}\end{aligned} \tag{3.14}$$

Straightforwardly from the above, taking $P_{\text{prior a}}$ with the same formulation as $P_{\text{prior b}}$, the log-likelihood presents as follows.

$$
\begin{aligned}
&\log \mathcal{L}_{\text{BSMF}}\left(\theta, Z; x, \Omega\right) = \\
&\quad \log \overset{2}{P}_{\text{background}} + \log P_{\text{segments}} + \log \overset{2}{P}_{\text{data}} + \log \overset{3}{P}_{\text{prior}} = \\
&\qquad (\log\tau - \log 2\pi)\frac{1}{2} \sum_{\{i,j\}\in E_{\text{spatial}}} y_{ij} - \tau\frac{1}{2} \sum_{\{i,j\}\in E_{\text{spatial}}} y_{ij}(b_i - b_j)^2 \\
&\qquad\quad + \log(1-q) \sum_{\{i,j\}\in E_{\text{spatial}}} y_{ij} + \log(q) \sum_{\{i,j\}\in E_{\text{spatial}}} (1 - y_{ij}) \\
&\qquad\quad + (\log\rho - \log 2\pi)\frac{1}{2} \sum_{\{i,j\}\in E_{\text{genome}}} z_{ij} - \rho\frac{1}{2} \sum_{\{i,j\}\in E_{\text{genome}}} z_{ij}(a_i - a_j)^2 \\
&\qquad\quad + \log(1-p) \sum_{\{i,j\}\in E_{\text{genome}}} z_{ij} + \log(p) \sum_{\{i,j\}\in E_{\text{genome}}} (1 - z_{ij}) \\
&\qquad\quad + \log(\nu)\frac{1}{2}|V| - \nu\frac{1}{2}\sum_{i\in V}(x_i - a_i - b_i)^2 \\
&\qquad\quad - \tau_b\frac{1}{2}\sum_{i\in V} b_i^2 - \tau_a\frac{1}{2}\sum_{i\in V} a_i^2 \qquad\qquad /^*\ a, b \text{ priors} \sim \mathcal{N}\left(0,\, \tau_\bullet\right)\ ^*/ \\
&\qquad\quad + \log(1-q)\left(\Omega_{\text{q0}} - 1\right) + \log(q)\left(\Omega_{\text{q1}} - 1\right) \\
&\qquad\quad + \log(1-p)\left(\Omega_{\text{p0}} - 1\right) + \log(p)\left(\Omega_{\text{p1}} - 1\right) \\
&\qquad\quad + \log\tau\left(\Omega_{\tau 0} - 1\right) - \tau\Omega_{\tau 1} \\
&\qquad\quad + \log\rho\left(\Omega_{\rho 0} - 1\right) - \rho\Omega_{\rho 1} \\
&\qquad\quad + \log\nu\left(\Omega_{\nu 0} - 1\right) - \nu\Omega_{\nu 1} \\
&\qquad\quad + \text{const.}
\end{aligned}
\tag{3.15}
$$

The BSMF model is a graphical Bayesian model build atop of large graphs $G_{\text{genome}}, G_{\text{spatial}}$. Effectively, the complete Bayes network of the model is large, unless vectors of related variables, as well as their multivariate distributions, are rendered as single vertices. Such factor graph for $\mathcal{L}_{\text{BSMF}}\left(\theta, Z; x, \Omega\right)$ is presented on Figure 3.2.3.

3.2.1 BSMF LOG-POSTERIOR LIKELIHOOD AS A QUADRATIC PROBLEM

Let's immediately point out the few qualitative observations about the total log-likelihood. They partially reveal the landscape of the BSFM, and they come useful later when optimizing for parameters of the model using

Expectation Maximization approach, as well as in the case of marginal distributions estimation with Monte-Carlo Markov Chain.

3.2.1. Claim. *Conditional maximization of the* $\log \mathcal{L}_{BSMF}(\theta, Z; x, \Omega)$ *as a function of vectors* a, b*, while keeping other variables constant, is a Quadratic Programming problem of size* $\mathcal{O}(|V|)$*.*

Proof. The log-likelihood is rewritten as

$$\log \mathcal{L}_{\text{BSMF}}(\theta, Z; x, \Omega) = \frac{1}{2} \begin{bmatrix} a & b \end{bmatrix} Q^{y,z}_{\tau,\rho,\nu} \begin{bmatrix} a \\ b \end{bmatrix} + c^T_{x,\nu,\tau_b} \begin{bmatrix} a \\ b \end{bmatrix} \tag{3.16}$$

$$+ f(\tau, \rho, \nu, p, q, x) \tag{3.17}$$

where $Q^{y,z}_{\tau,\rho,\nu}$ is a symmetric matrix (the precise definition is given in the next Claim 3.2.2) of size $2|V| \times 2|V|$, since $\text{len}(a) = \text{len}(b) = 2|V|$, and c^T_{x,ν,τ_b} is a vector of length $2|V|$. In the case of G_{spatial} being a hexagonal lattice, $\left|E_{\text{spatial}}\right| \leq 3|V|$, and $\left|E_{\text{genome}}\right| < |V|$ since G_{genome} is a line graph. The precise structure depends on G_{spatial}, however, the total number of non-zero elements is less than $\left|E_{\text{spatial}}\right| + \left|E_{\text{genome}}\right| + |V| \leq 5|V|$.

Thus $Q^{y,z}_{\tau,\rho,\nu}$ is a sparse matrix, with $\mathcal{O}(|V|)$ non-zero entries. □

3.2.2. Claim. *Conditional maximization of the* $\log \mathcal{L}_{BSMF}(\theta, Z; x, \Omega)$ *as a function of vectors* a, b*, while keeping other variables constant, is a positive-definite Quadratic Programming problem to which the solution is given by* ${Q^{y,z}_{\tau,\rho,\nu}}^{-1} c_{x,\nu,\tau_b}$

Proof. Expanding the bilinear form formulation of $\log \mathcal{L}_{\text{BSMF}}(\theta, Z; x, \Omega)$ from the previous claim 3.2.1 yields

$$\begin{bmatrix} a & b \end{bmatrix} Q^{y,z}_{\tau,\rho,\nu} \begin{bmatrix} a \\ b \end{bmatrix} = -\tau \sum_{\{i,j\} \in E_{\text{spatial}}} y_{ij}(b_i - b_j)^2 - \rho \sum_{\{i,j\} \in E_{\text{genome}}} z_{ij}(a_i - a_j)^2$$
$$- \nu \sum_{i \in V} (a_i + b_i)^2 - \tau_a \sum_{i \in V} a_i^2 - \tau_b \sum_{i \in V} b_i^2 \tag{3.18}$$

thus, since $\tau > 0 \wedge \rho > 0 \wedge \nu > 0$

$$
\begin{aligned}
& v^T - Q^{y,z}_{\tau,\rho,\nu} v \geq 0 \quad \text{and} \\
& v^T - Q^{y,z}_{\tau,\rho,\nu} v = 0 \Leftrightarrow v = 0 \qquad\qquad \text{thanks to } \tau_a, \tau_b > 0 \\
& \qquad\qquad\qquad\qquad\qquad\qquad \Rightarrow \quad -Q^{y,z}_{\tau,\rho,\nu} \succ 0
\end{aligned}
$$

from which follows that maximization of $\log \mathcal{L}_{\text{BSMF}}(\theta, Z; x, \Omega)$ is equivalent to minimization of a positive definite Quadratic Programming problem. □

3.2.3. Remark. From Claims 3.2.1 , 3.2.2 it follows that the conditional maximization with respect to a, b can be solved, using iterative methods such as Jacobi method, in $\mathcal{O}(|V|)$ time if the $Q^{y,z}_{\tau,\rho,\nu}$ is well conditioned. The lowest eigenvalue of the matrix is bounded from below by $\lambda_0 > \min(\tau_a, \tau_b)$, while the largest eigenvalue can be bounded, from the sparsity of the matrix, by $\mathcal{O}(|V|) \cdot \max(\tau, \rho, \nu, \tau_a \tau_b)$. Unfortunately, this value does become large in practice, since τ, ρ precisions optimize to high values, nevertheless they're bounded from above thanks to prior distributions imposed on them.

3.2.4. Proposition. *[after e.g. (Bickson, 2008)] In a Gaussian Markov random field with a probability function*

$$P(x) = Z^{-1} \exp\left(-\frac{1}{2} x^T E x + c^T x\right)$$

where Z is the partition function, the posterior distribution is multivariate Normal: $P(x) \sim \mathcal{N}\left(\mu,\, E^{-1}\right)$

Proof.

$$\begin{aligned}
\text{let} \qquad \mu &= E^{-1}c \qquad \text{then} \\
P(x) &= Z^{-1} \exp\left(-\frac{1}{2}\mu^T E\mu)\right) \\
&\quad \cdot \exp\left(-\frac{1}{2}x^T Ex + \mu^T Ex - \frac{1}{2}\mu^T E\mu\right) \\
&= Z'^{-1} \exp\left(-\frac{1}{2}\mu^T E\mu\right) = \mathcal{N}\left(\mu,\, E^{-1}\right)
\end{aligned}$$

□

From Proposition 3.2.4, by inverting the final bilinear form matrix, we may obtain posterior confidence intervals, although, this not exactly the same as computing posterior marginal distribution (e.g. with MCMC approach), since here precisions and breaks variables are assumed constant.

3.2.5. Claim. *Conditional maximization of the* $\log \mathcal{L}_{BSMF}(\theta, Z; x, \Omega)$ *as a function of variables* τ, ρ, ν, p, q*, while keeping other variables constant, is a concave problem, and can be optimized globally by maximizing several one-dimensional concave functions, providing that* $\Omega_{\tau 0} > 1 \wedge \Omega_{\rho 0} > 1 \wedge \Omega_{\nu 0} > 1 \wedge \Omega_{q0} > 1 \wedge \Omega_{q1} > 1 \wedge \Omega_{p0} > 1 \wedge \Omega_{p1} > 1$.

Proof. Log-likelihood is a sum of one-dimensional functions with respect to τ, ρ, ν, q, p.

$$\log \mathcal{L}_{\text{BSMF}}(\theta, Z; x, \Omega) = f_{b,y}(\tau) + f_{a,z}(\rho) + f_{a,b}(\nu) + f_y(q) + f_z(p) + f_{b,a,x} \tag{3.19}$$

The posterior conditional distributions are the same as conjugate prior distributions. The conditional log-likelihood with respect to τ is

$$\begin{aligned}
f_{b,y}(\tau) &= \log \text{Gamma}(\alpha, \beta)(\tau) + const. \qquad \text{where} \\
\alpha &= \Omega_{\tau 0} + \frac{\sum_{\{i,j\} \in E_{\text{spatial}}} y_{ij}}{2} \\
\beta &= \Omega_{\tau 1} + \frac{\sum_{E_{\text{spatial}}} \left(b_i - b_j\right)^2}{2}.
\end{aligned} \tag{3.20}$$

Taking the second derivative of log-pdf of Gamma distribution

$$\frac{\mathrm{d}^2 \log \mathrm{Gamma}(\alpha,\beta)(\tau)}{\mathrm{d}\tau^2} = \frac{1-\alpha}{\tau^2} < 0 \quad \Leftrightarrow \quad 1 < \alpha \tag{3.21}$$

and substituting α results in the condition for concavity:

$$\Omega_{\tau 0} > 1 - \frac{\sum_{\{i,j\} \in E_{\text{spatial}}} y_{ij}}{2} \tag{3.22}$$

which is always true if $\Omega_{\tau 0} > 1$, since $y_{ij} > 0$. Similar reasoning applies to ρ, ν. Posterior of p, q is Beta distribution with posterior parameters α', β', which second derivative is

$$\frac{\mathrm{d}^2 \log \mathrm{Beta}(\alpha',\beta')(p)}{\mathrm{d}p^2} = -\frac{\alpha' + p^2(\alpha' + \beta' - 2) - 2(\alpha' - 1)p - 1}{(p-1)^2 p^2} < 0 \quad \Leftrightarrow$$

$$(\beta' - 1)p^2 + (\alpha' - 1)(p-1)^2 > 0 \quad \text{which is true when } \alpha' > 1 \wedge \beta' > 1.$$

The posterior α', β' parameters in Beta distribution are always larger than prior (since $\alpha' = \alpha + \#_{\text{successes}}$, $\beta' = \beta + \#_{\text{failures}}$), so $\Omega_{p0} > 1 \wedge \Omega_{p1} > 1 \Rightarrow \alpha' > 1, \beta' > 1$, and similar reasoning applies to q. □

3.2.6. Claim. *The* $\log \mathcal{L}_{BSMF}(\theta, Z; x, \Omega)$ *as a function of variables* θ *is not everywhere concave on the feasible domain, i.e. its Hessian matrix is not negative-definite, it can be indefinite.*

Proof. As a proof we find a vector h, and point $\hat{\theta}$ for which $h^T H_{\hat{\theta}} h > 0$, where $H_{\hat{\theta}}$ is a hessian matrix of $f(\theta) = \log \mathcal{L}_{\mathrm{BSMF}}(\theta, Z; x, \Omega)$ at point $\hat{\theta}$. Let $h = h_1 \,\mathrm{d}\rho + h_2 \,\mathrm{d}\nu + h_3 \,\mathrm{d}a_i$, and let $\Omega'_{\rho 0} = \Omega_{\rho 0} + \sum_{\{i,j\} \in E_{\text{spatial}}} z_{ij}$ and $\Omega'_{\nu 0} = \Omega_{\nu 0} + |V|$ be the posterior Gamma parameters. Now

$$\begin{aligned}
\mathrm{D}^2_{h,h} f(\theta) &= h_1^2 \tfrac{\mathrm{d}^2 f}{\mathrm{d}\rho^2} + h_2^2 \tfrac{\mathrm{d}^2 f}{\mathrm{d}\nu^2} + h_3^2 \tfrac{\mathrm{d}^2 f}{\mathrm{d}a_i^2} + 2h_1 h_3 \tfrac{\partial^2 f}{\partial \rho \partial a_i} + 2h_2 h_3 \tfrac{\partial^2 f}{\partial \nu \partial a_i} + 2h_1 h_2 \tfrac{\partial^2 f}{\partial \rho \partial \nu} \\
&= h_1^2 \frac{1 - \Omega'_{\rho 0}}{\rho^2} + h_2^2 \frac{1 - \Omega'_{\nu 0}}{\nu^2} - h_3^2 a_i^2 (\rho + \nu)/2 \\
&\qquad - 2h_1 h_3 (z_{i,i+1}(a_i - a_{i+1}) + z_{i,i-1}(a_i - a_{i-1})) \\
&\qquad - 2h_2 h_3 (-x_i + a_i + b_i)\,.
\end{aligned} \tag{3.23}$$

Part of the expression can be rewritten as

$$h_1^2 \frac{1-\Omega'_{\rho 0}}{\rho^2} + h_2^2 \frac{1-\Omega'_{\nu 0}}{\nu^2} - h_3^2 a_i^2 (\rho+\nu)/2 = -h^T C_{\rho,\nu,a_i,z,\Omega} h$$
$$\text{for} \quad C_{\rho,\nu,a_i,z,\Omega} \succeq 0. \tag{3.24}$$

This is true if Ω fulfills conditions from Claim 3.2.5.

The above 2nd derivative can be made positive and arbitrarily large in the direction $h : h_1 = h_2 = h_3 = 1$ by setting a_{i+1}, or a_{i-1},or x_i, to be positive and large enough to impose $h^T H_{\hat{\theta}} h > 0$. This together with $(1\,\mathrm{d}\rho)^T H_{\hat{\theta}} (1\,\mathrm{d}\rho) = \frac{\mathrm{d}^2 f}{\mathrm{d}\rho^2} < 0$ shows indefiniteness of $\log \mathcal{L}_{\text{BSMF}} \left(\hat{\theta}, Z, x, \Omega\right)$ Hessian. □

The above claim suggests that, even with $Z = (y, z)$ held, the $\log \mathcal{L}_{\text{BSMF}} (\theta, Z; x, \Omega)$ may not have a global optimum with respect to θ.

3.2.2 Posterior conditional distributions of the model

The BFSM posterior conditional distributions for θ variables are:

$$
\begin{aligned}
b_i \sim & \mathcal{N}\left(\frac{\tau_0\mu_\tau + \tau \sum\limits_{j:\{i,j\}\in E_{\text{spatial}}} y_{ij}b_j + \nu(x_i - a_i)}{\tau_0 + \tau \sum\limits_{j:\{i,j\}\in E_{\text{spatial}}} y_{ij} + \nu}, \frac{1}{\tau_0 + \tau \sum\limits_{j:\{i,j\}\in E_{\text{spatial}}} y_{ij} + \nu}\right) \\
a_i \sim & \mathcal{N}\left(\frac{\rho_0\mu_\rho + \rho \sum\limits_{j:\{i,j\}\in E_{\text{spatial}}} z_{ij}a_j + \nu(x_i - b_i)}{\rho_0 + \rho \sum\limits_{j:\{i,j\}\in E_{\text{spatial}}} z_{ij} + \nu}, \frac{1}{\rho_0 + \rho \sum\limits_{j:\{i,j\}\in E_{\text{spatial}}} z_{ij} + \nu}\right) \\
\tau \sim & \text{Gamma}\left(\Omega_{\tau 0} + \frac{\sum_{\{i,j\}\in E_{\text{spatial}}} y_{ij}}{2}, \Omega_{\tau 1} + \frac{\sum_{E_{\text{spatial}}} \left(b_i - b_j\right)^2}{2}\right) \\
\rho \sim & \text{Gamma}\left(\Omega_{\rho 0} + \frac{\sum_{\{i,j\}\in E_{\text{genome}}} z_{ij}}{2}, \Omega_{\rho 1} + \frac{\sum_{E_{\text{genome}}} \left(a_i - a_j\right)^2}{2}\right) \\
\nu \sim & \text{Gamma}\left(\Omega_{\nu 0} + \frac{|V|}{2}, \Omega_{\nu 1} + \frac{\sum_V \left(x_i - a_i - b_i\right)^2}{2}\right) \\
q \sim & \text{Beta}\left(\Omega_{q0} + \sum_{E_{\text{spatial}}} y_i, \Omega_{q1} + \left|E_{\text{spatial}}\right| - \sum_{E_{\text{spatial}}} y_i\right) \\
p \sim & \text{Beta}\left(\Omega_{p0} + \sum_{E_{\text{genome}}} z_i, \Omega_{p1} + \left|E_{\text{genome}}\right| - \sum_{E_{\text{genome}}} z_i\right)
\end{aligned}
\tag{3.25}
$$

Conditional distributions of variables from θ and Z are inter-dependent, and their dependency graph is depicted on Figure 3.2.3. The inter-dependency of $\vec{a}$ and $\vec{b}$ constitutes the marginal maximization problem as Quadratic Programming, as it is shown in claims 3.2.1, 3.2.2.

Posterior binomial distributions for the latent $[0, 1]$ field break controlling variables y_{ij}, z_{ij} are derived from Bayes rule for total probability:

$$
P(y_{ij} = 1|\theta, x, \Omega) = \frac{\mathcal{L}(\theta_{[y_{ij}=1]}, Z; x, \Omega)}{\mathcal{L}(\theta_{[y_{ij}=0]}, Z; x, \Omega) + \mathcal{L}(\theta_{[y_{ij}=1]}, Z; x, \Omega)} \tag{3.26}
$$

Precisely, the computation proceeds as follows:

$$P(y_{ij} = 1|\theta, x, \Omega) = \frac{\exp(v)}{\exp(v) + 1} \qquad \text{where}$$
$$v = 0.5(\log \tau - \log 2\pi) - 0.5\tau(b_i - b_j)^2 + \log(1 - q) - \log(q) \quad (3.27)$$

Probabilities for z_{ij} are computed in the same way substituting a_i for b_i and q for p.

3.2.3 Categorical mixture prior distribution for the segment field

Previously we argued that setting $P_{\text{prior a}}$ in the same manner as $P_{\text{prior b}}$, so that $P_{\text{prior a}} \sim \mathcal{N}(0, \tau_a)$, does not match realistic expectation of $\log_2$ratio data. Segment means are not distributed around 0, rather they come form a mixture distribution of (at least) 3 components $\alpha \in \{-1, 0, 1\}$ each corresponding to *deletion, no aberration, duplication* accordingly.[1] Empirical distributions of segment means are plotted on Figure 3.2.4 in form of histograms. These components are assumed to be Normal with means $\mu_{a,-1}, \mu_{a,0}, \mu_{a,1}$ and precisions $\rho_{a,-1}, \rho_{a,0}, \rho_{a,1}$. Setting $\rho_{a,1} > \rho_{a,0}$ and $\rho_{a,-1} > \rho_{a,0}$ partially addresses the heteroscedasticity of $\log_2$ratio.

Prior proportions in the mixture of components are given by a vector of probabilities r_α where $\sum_{\alpha\in\{-1,0,1\}} r_\alpha = 1$. To estimate posterior mixture probabilities, we introduce a vector of categorical variables $\hat{s_i} \in -1, 0, 1$ distributed a priori Categorical(r_α). However, in computations we use a matrix of $0, 1$ variables $\left(s_{i,\alpha}\right)_{|V|\times\{-1,0,1\}}$, which we call *segment category states*, with the property $\forall_i \sum_{\alpha\in\{-1,0,1\}} s_{i,\alpha} = 1$. We end up with the following formula:

$$P_{\text{prior a}} = \prod_{\alpha\in\{-1,0,1\}} \left(r_\alpha \sqrt{\frac{\rho_{a,\alpha}}{2\pi}} \right)^{s_{i,\alpha}} \exp\left(-\frac{1}{2} s_{i,\alpha} \rho_{a,\alpha} (\mu_{a,\alpha} - a_i)^2 \right) \quad (3.28)$$

[1] Actually, due to diploid genome and various possibilities for mutations, there are more than 3 basic levels of signal expected. In literature Hidden Markov Models use 3 to 5 states to model signal categories. (Rueda and Diaz-Uriarte, 2009)

This renders the posterior conditional distribution of a_i to

$$
\begin{aligned}
a_i &\sim \mathcal{N}\left(\frac{m_{\text{post a}}}{\tau_{\text{post a}}}, \frac{1}{\tau_{\text{post a}}}\right) \qquad \text{where} \\
m_{\text{post a}} &= \sum_{\alpha\in\{-1,0,1\}} s_{i,\alpha}\rho_{a,\alpha}\mu_{a,\alpha} + \rho \sum_{j:\{i,j\}\in E_{\text{genome}}} z_{ij}a_j + \nu(x_i - b_i) \\
\tau_{\text{post a}} &= \sum_{\alpha\in\{-1,0,1\}} s_{i,\alpha}\rho_{a,\alpha} + \rho \sum_{j:\{i,j\}\in E_{\text{genome}}} z_{ij} + \nu
\end{aligned} \tag{3.29}
$$

The expectation step for latent $0, 1$ random variables $s_{i,\bullet}$ proceeds as follows, similarly as in the case of y_{ij}.

$$
\begin{aligned}
&P(s_{i,\bullet} = (1,0,0))|\theta, x, \Omega) = \\
&\qquad \frac{\mathcal{L}(\theta, Z_{\left[s_{i,\bullet}=(1,0,0)\right]}; x, \Omega)}{\mathcal{L}(\theta, Z_{\left[s_{i,\bullet}=(1,0,0)\right]}; x, \Omega) + \mathcal{L}(\theta, Z_{\left[s_{i,\bullet}=(0,1,0)\right]}; x, \Omega) + \mathcal{L}(\theta, Z_{\left[s_{i,\bullet}=(0,0,1)\right]}; x, \Omega)}
\end{aligned} \tag{3.30}
$$

Precisely

$$
\begin{aligned}
v_{i,\alpha} &= 0.5(\log\rho_{a,\alpha} - \log 2\pi) - 0.5\rho_{a,\alpha}(\mu_{a,\alpha} - a_i)^2 + \log r_\alpha \\
P(s_{i,\hat{\alpha}}|\theta, x, \Omega) &= \frac{\exp(v_{i,\hat{\alpha}})}{\sum_{\alpha\in\{-1,0,1\}}\exp(v_{i,\alpha})} \quad .
\end{aligned} \tag{3.31}
$$

Hyperpriors for $\mu_{a,\alpha}$, r_α. A natural progressing step is to variate $\mu_{a,\alpha}$ by adding it to the set of optimized variables θ, and complement the log-likelihood with a (hyper)prior Normal distribution (with rather small variance), as below.

$$
P_{\text{prior }\mu_{a,\alpha}} = \sqrt{\frac{\hat{\rho}_{a,\alpha}}{2\pi}}\exp\left(-\frac{1}{2}\hat{\rho}_{a,\alpha}(\hat{\mu}_{a,\alpha} - \mu_{a,\alpha})^2\right) \tag{3.32}
$$

This corresponds to discovering from the data A similar variation step for $\rho_{a,\alpha}$ with Gamma distribution as a hyperprior is feasible.

The vector $\vec{r_\alpha}$ corresponds to the probabilities of discovering genomic deletion, duplications, and neutral segments. It can be varied by setting a prior Dirichlet distribution, which is a conjugate prior to Categorical distribution.

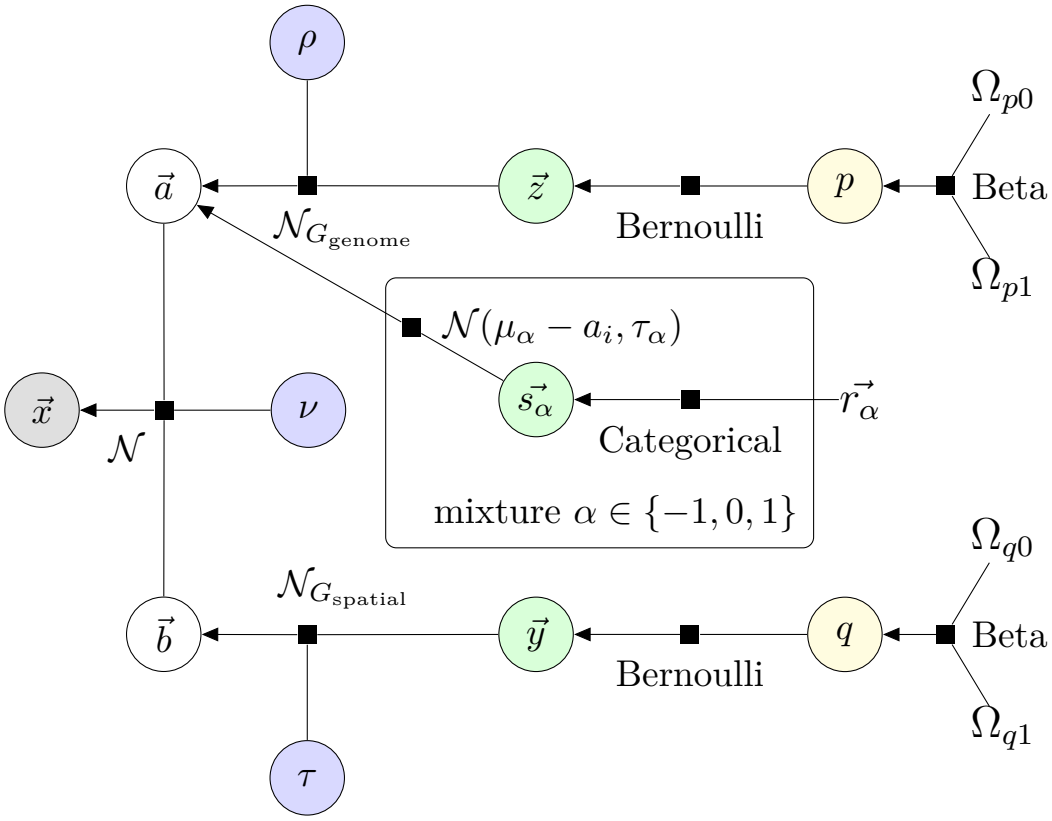


Figure 3.2.3: Factor graph for $\mathcal{L}_{\text{BSMF}}(\theta, Z; x, \Omega)$ model. $\log_2$ratio data $\vec{x}$ marked gray. Gaussian Markov random fields $\vec{a}$, $\vec{b}$, defined on graphs G_{spatial}, G_{genome} accordingly, in white $\log_2$ratio data $\vec{x}$ marked gray. Precision latent variables τ, ρ, ν marked blue, their prior Gamma distributions not charted on the graph, Normal prior for $\vec{b}$ not charted. Discrete variables $Z = \{\vec{y}, \vec{z}, \vec{s}\}$ marked green. Background, segment fields break probabilities p, q marked in yellow. The $\vec{s}$ *segment category state* is introduced in Section 3.2.3. The multivariate $\mathcal{N}_{G_{\text{genome}}}$, $\mathcal{N}_{G_{\text{spatial}}}$ correspond to Gaussian Markov Fields on edges of respective graphs. Interestingly, the graph is a tree graph when vector variables and multivariate distributions are treated as typical Bayes network nodes.

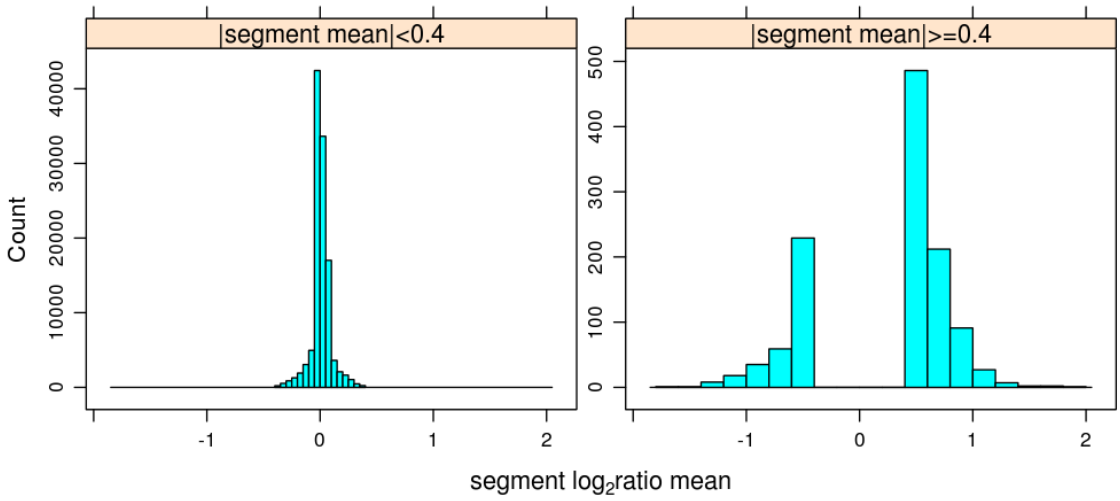


Figure 3.2.4: Histogram of segment $\log_2$ratio means from IMID2py database. Since most segments have no aberration and segments around 0 dominate, data was divided into 2 groups for better visibility.

3.3 Double linkage in the segment field

One characteristic of the line graph G_{genome} and the *segment field* defined on it is the fact that a single break $z_{ij} = 0$ between i-th and j-th neighboring features disconnects the graph, and in such case no information flow is preserved along the genome.

This is especially unwanted in the presence of *spoiled* probes on a microarray. A spoiled probe is a feature which results with $\log_2$ratio being erratic, without correlation to neighboring probes. A probe can turn out to be spoiled in one, a group of, or all experiments, for reasons listed in Section 2.2.3, page 26. We confirm existence of such probes on microarrays from IMID2py database, for a concrete example refer to Figure 4.2.1. Thus, a presence of such a spoiled probe in the segment disrupts dependence structure of probes in the same segment. [2]

This inspires to introduce a simple modification to the single linked segment field, namely *double linked segment field* and its double linked graph

$$G^{\text{double}}_{\text{genome}} = \left(V, E_{\text{genome}} \cup E^{\text{double}}_{\text{genome}}\right) \qquad \text{where}$$
$$E^{\text{double}}_{\text{genome}} = \left\{ \{i,j\} : \begin{array}{r} \text{there is exactly one probe between } i \\ \text{and } j \text{ along the chromosome} \end{array} \right\}. \tag{3.33}$$

Again, if we choose the indexing of the feature set V to agree with the genomic order, then the set of edges $E^{\text{double}}_{\text{genome}}$ is roughly $\left\{\{i, i+2\}\right\}_I$. The modification allows disconnecting single probe outliers from segments without disconnecting the graph, and there is no break in the segment dependency structure.

The relevant part of the $\mathcal{L}_{\text{BSMF}}(\theta, Z; x, \Omega)$ (ref. eqn. 3.11) is modified in

[2] The existence of such probes is one of the reasons for the industry standard to only accept segments with length of at least 3 consecutive probes, see for example Möhlendick et al. (2013).

the following way:

$$\exp\left(-\frac{1}{2}\left(\sum_{\{i,j\}\in E_{\text{genome}}} z_{ij}\rho(a_i-a_j)^2+\sum_{\{i,j\}\in E^{\text{double}}_{\text{genome}}} wz_{ij}\rho(a_i-a_j)^2\right)\right)$$
$$=\exp\left(-\begin{bmatrix}a\end{bmatrix}^T M_{\text{double.w}}\begin{bmatrix}a\end{bmatrix}\right) \tag{3.34}$$

where $w\geq 0$ is the weight of the double link. $w=0$ is the single link, $w=0.5$ is a natural choice from the fact that $a_i-a_{i+2}=(a_i-a_{i+1})+(a_{i+1}-a_{i+2})$ is a sum of two Gaussian variables, hence its variance shall be twice the original. $w=1$ also turns out to be an interesting choice, which is analyzed as follows.

THE CONSEQUENCES OF DOUBLE LINKS are analyzed through the analysis of eigenvalues of the corresponding bilinear form matrices of the single/double linked graph M_{single}, $M_{\text{double.w}}$.

$$M_{\text{single}}=\begin{bmatrix}1&-1&0&\cdots&0&0&0\\-1&2&-1&\ddots&0&0&0\\0&-1&2&\ddots&\ddots&0&0\\\vdots&\ddots&\ddots&\ddots&\ddots&\ddots&\vdots\\0&0&\ddots&\ddots&2&-1&0\\0&0&0&\ddots&-1&2&-1\\0&0&0&\cdots&0&-1&1\end{bmatrix}\qquad M_{\text{double.1}}=\begin{bmatrix}2&-1&-1&\cdots&0&0&0\\-1&3&-1&\ddots&0&0&0\\-1&-1&4&\ddots&\ddots&0&0\\\vdots&\ddots&\ddots&\ddots&\ddots&\ddots&\vdots\\0&0&\ddots&\ddots&4&-1&-1\\0&0&0&\ddots&-1&3&-1\\0&0&0&\cdots&-1&-1&2\end{bmatrix}$$

$$M_{\text{double.w}} = \begin{bmatrix} 1+w & -1 & -w & \cdots & 0 & 0 & 0 \\ -1 & 2+w & -1 & \ddots & 0 & 0 & 0 \\ -w & -1 & 2(w+1) & \ddots & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \ddots & \ddots & 2(w+1) & -1 & -w \\ 0 & 0 & 0 & \ddots & -1 & 2+w & -1 \\ 0 & 0 & 0 & \cdots & -w & -1 & 1+w \end{bmatrix} \tag{3.35}$$

All the above matrices are diagonally weakly dominant [3], hence their eigenvalues are all $\lambda_k \geq 0$. The smallest eigenvalue is $\lambda_0 = 0$, and it corresponds to a constant segment eigenvector $\vec{a} = (m, m, \ldots, m)$. The second eigenvectors are approximately the slope vector $s_{\text{slope}} = \left(-\frac{n-1}{2}, -\frac{n-1}{2}+1, \ldots, \frac{n+1}{2}-1, \frac{n-1}{2}\right)$.

Denote $\tilde{b} = (x - b)$ and let M be any of the linked spatial graph bilinear form matrices. Rewriting the equation 3.16 with appropriate bilinear form matrix Q_{spatial}, assuming $\mathcal{N}(0, \tau_a)$ as a prior for a, and treating

[3] The $M_{\text{double.1}} = D_{\text{genome}} - A_{\text{genome}}$ is the Laplacian of the graph G_{genome}, where D_{genome}, A_{genome} are degree and adjacency matrices accordingly.

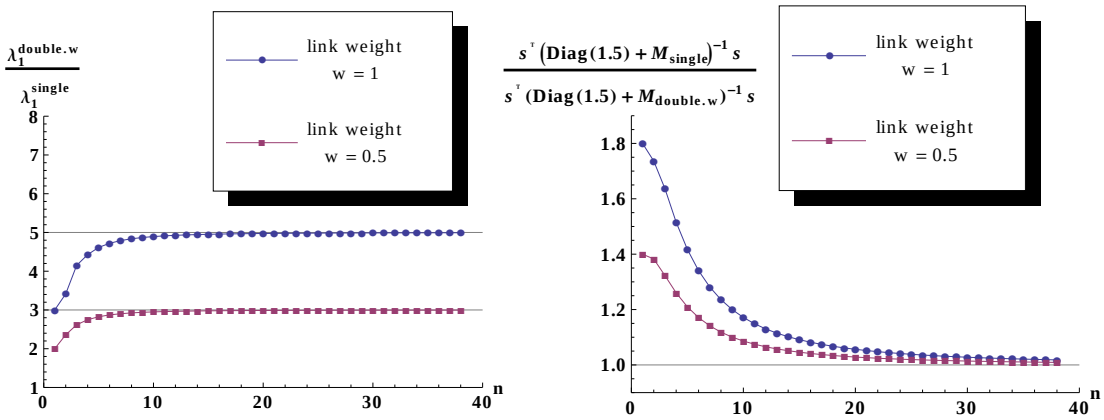


Figure 3.3.1: *Left:* The ratio between the second smallest eigenvalues of matrices M_{single} and $M_{\text{double.w}}$ for $w = 1$ and $w = 0.5$. The $\lambda_0 = 0$ for all matrices, while $\lambda_1^{\text{double.1}}$ is 5 times larger, and $\lambda_1^{\text{double.0.5}}$ is 3 times larger, than the second eigenvalue of the single linkage graph matrix for large matrix sizes n. This shows that the spectral gap for the double linked graph is 5, or 3, times larger than the spectral gap of the single linked graph. *Right:* The ratio of optimal slope coefficients in the segment field $\frac{\hat{a}_1^{\text{single}}}{\hat{a}_1^{\text{double}}}$. Double linked segment fields fit for smaller slopes. (ref. eqn. 3.40).

the $\log \mathcal{L}_{\text{BSMF}}(\theta, Z; x, \Omega)$ as a function of a we obtain

$$\begin{aligned}
\log \mathcal{L}_{\text{BSMF}}(\theta, Z; x, \Omega) &= \frac{1}{2}\begin{bmatrix} a & b \end{bmatrix} Q^{y,z}_{\tau,\rho,\nu} \begin{bmatrix} a \\ b \end{bmatrix} + c^T_{x,\nu,\tau_b} \begin{bmatrix} a \\ b \end{bmatrix} \\
&\quad + f(\tau, \rho, \nu, p, q, x) \\
&= \text{const.} - \langle b|Q_{\text{spatial}}|b\rangle - \left\langle \tilde{b} - a|\tilde{b} - a \right\rangle - \langle a|M|a\rangle - \langle a|\tau_a/2|a\rangle \\
&= \text{const.} - \left\langle \tilde{b}|\tilde{b} \right\rangle + 2\left\langle \tilde{b}|a \right\rangle - \langle a|(I + \tau_a/2)|a\rangle - \langle a|M|a\rangle \\
&= \text{const.} + 2\left\langle \tilde{b}|a \right\rangle - \langle a|M'|a\rangle \\
&\text{where} \quad M' = \text{Diag}(1 + \tau_a/2) + M
\end{aligned} \tag{3.36}$$

Now, the maximization of $\log \mathcal{L}_{\text{BSMF}}(\theta, Z; x, \Omega)$ with respect to a is equivalent to

$$\text{Min} \qquad \langle a|M'|a\rangle - 2\left\langle \tilde{b}|a \right\rangle \tag{3.37}$$

M' is symmetric, strictly positive definite with eigenvalues $1 + \tau_a/2 + \lambda_k$ for $k = 0, \dots, n$. Let $\tilde{b}_k$, a_k denote coefficients in the eigenbasis of M'. The solution to the maximization problem is given by

$$\hat{a} = {M'}^{-1}\tilde{b} \qquad \Leftrightarrow \qquad \hat{a}_k = \tilde{b}_k/(1 + \tau_a/2 + \lambda_k) \tag{3.38}$$

Denote $\hat{a}^{\text{single}} = {M'_{\text{single}}}^{-1}\tilde{b}$, $\hat{a}^{\text{double}} = {M'_{\text{double}}}^{-1}\tilde{b}$. It follows that both cases yield the same segment mean, that is $\hat{a}_0^{\text{single}} = \hat{a}_0^{\text{double}}$, because the λ_0 eigenvalues and corresponding eigenvectors are the same.

It turns out that within a good precision $b_1^{\text{single}} \simeq b_1^{\text{double}}$, which correspond to coefficients of slope component of $\tilde{b}$ and thus

$$\frac{\hat{a}_1^{\text{single}}}{\hat{a}_1^{\text{double}}} \simeq \frac{1 + \tau_a/2 + \lambda_1^{\text{double}}}{1 + \tau_a/2 + \lambda_1^{\text{single}}}\,. \tag{3.39}$$

A reasonable prior setting sets $\tau_a \leq 1$. We hypothetize that

$$\left|\hat{a}_1^{\text{single}}\right| \simeq \frac{1.5 + \lambda_1^{\text{double}}}{1.5 + \lambda_1^{\text{single}}}\left|\hat{a}_1^{\text{double}}\right| \gtrsim 1 \cdot \left|\hat{a}_1^{\text{double}}\right|. \tag{3.40}$$

The hypothesis is verified numerically: the slope vector is chosen for $\tilde{b} \leftarrow s_{\text{slope}}$ and $\frac{\hat{a}_1^{\text{single}}}{\hat{a}_1^{\text{double}}}$ is computed, results on Figure 3.3.1.

The practical effects of the double link introduction are plotted on Figure 3.3.2 where we observe that the optimized segment fields for double links graphs have smaller slopes, to almost no slopes in the case of $w = 1$, which gets balanced with more breaks in segment fields, and which stands in agreement with the above considerations. The slope coefficients ratio from Figure 3.3.1 approaches 1 when $n \to \infty$, however double link causes the fit of segment fields precisions ρ to be larger, hence resulting segment fields have less variance.

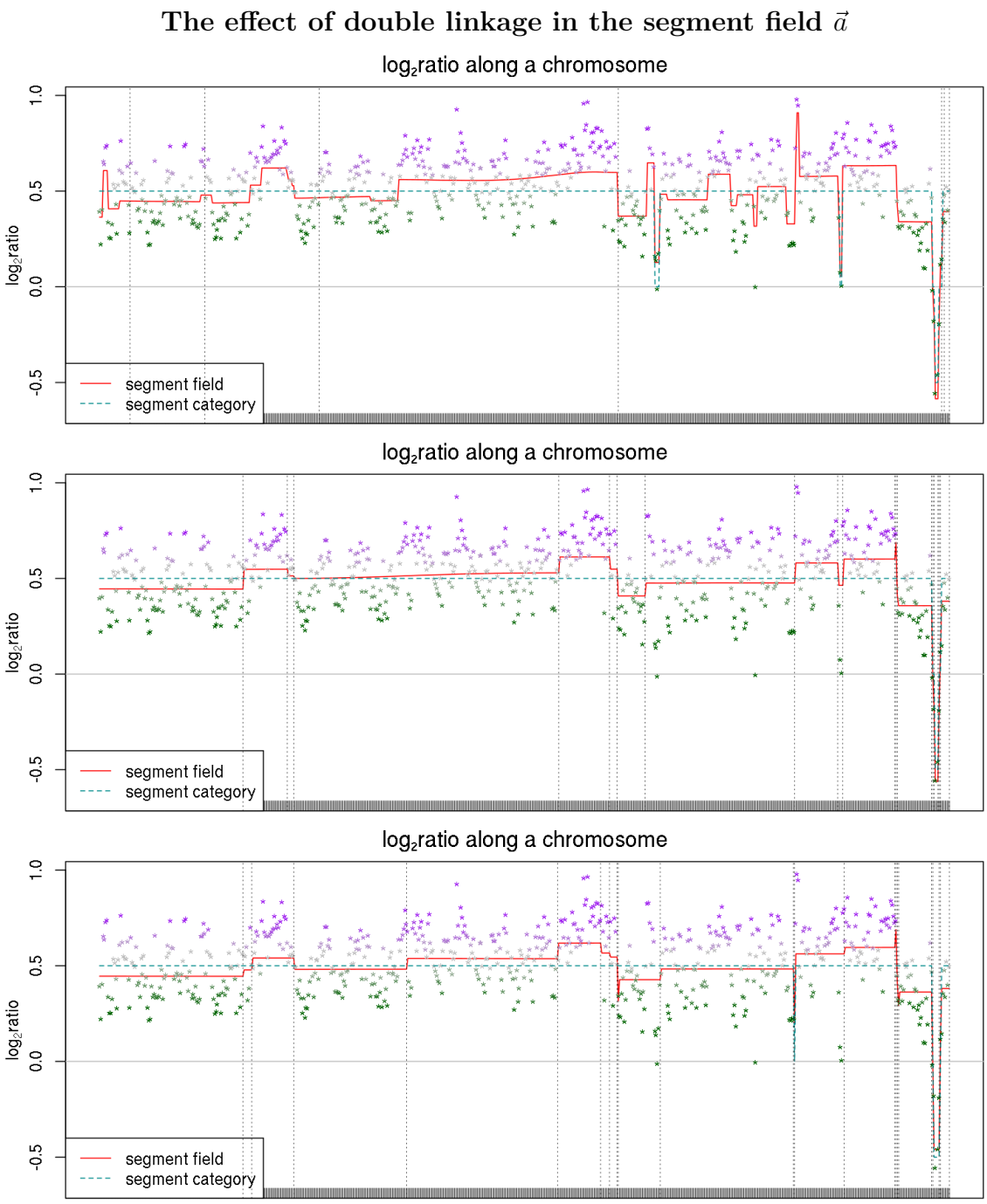


Figure 3.3.2: The effect of double linkage in the segment field (marked in red). *i)* top panel: no additional links, only $\{i, i+1\}$ *ii)* middle panel: $\{i, i+2\}$ with twice the single link variance (weight $w = 0.5$) *iii)* bottom panel: $\{i, i+2\}$ with the same variance (weight $w = 1$). The double linkage stabilizes the segment field (small variance, no slope), as a consequence more breaks in the segment field are imposed. The plotted region indicates duplication in mitochondrial DNA, possibly due to degeneracy of mitochondrial DNA in the reference sample.

3.4 Optimizing BSMF parameters with Expectation Maximization

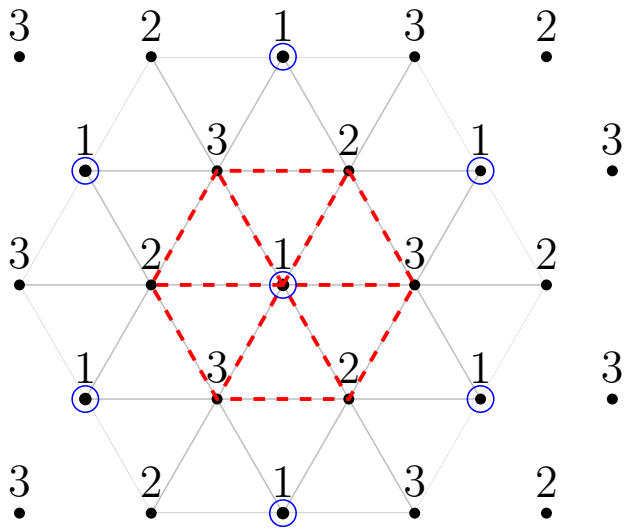


Figure 3.4.1: A hexagonal lattice(equivalently a triangular tiling) graph G and its stratification on Agilent aCGH microarray. Each stratum has its neighborhood non-intersecting with other nodes from the same stratum. Strata no. 1 marked in blue. neighborhood of the center node marked in red. We iterate over strata and maximize log-probability on each one separately, while keeping other strata constant. Maximal cliques of G are triangles, each consisting of vertices from strata $1, 2, 3$.

Estimating posteriors with Gibbs sampler We denote the possibility of a Gibbs sampler for the $\mathcal{L}_{\text{BSMF}}(\theta, Z; x, \Omega)$. The conditional distributions to sample from in each step are listed in Equations 3.25. In the the case of fields a, b one way to proceed is to sample sequentially first over vertices of G_{spatial}, and then over vertices of G_{genome}. The resampling of field values on each vertex is essentially what have to be done, however we may benefit from the structure of graphs and local Markov property of the field. Precisely, we color G_{spatial} graph into 3 strata, as illustrated on Figure 3.4.1. This allows to effectively resample variables in each strata. This concepts also plays role in our Expectation Maximization scheme implementation.

In the EM algorithm background and segment fields are are initialized to averages from neighborhood $\log_2$ratio data in graphs G_{spatial} and G_{genome} respectively.

$$\begin{aligned} a_i &= \frac{x_i + \sum_{j \in N_{\text{genome}}(i)} x_j}{\left|1 + N_{\text{genome}}(i)\right|} & z_i &= 1 \\ b_i &= \frac{x_i + \sum_{j \in N_{\text{spatial}}(i)} x_j}{\left|1 + N_{\text{spatial}}(i)\right|} & y_i &= 1 \end{aligned} \tag{3.41}$$

3.5 Results: application on aCGH microarrays data

To obtain results presented below, we used the following parameters for prior distributions. For Gamma prior distribution for all precision variables we used:

$$\begin{aligned} \Omega_{\tau 0} &= \Omega_{\rho 0} = \Omega_{\nu 0} = 1.1 \\ \Omega_{\tau 1} &= \Omega_{\rho 1} = \Omega_{\nu 1} = 1/5 \end{aligned} \tag{3.42}$$

For Beta prior distribution for probabilities p, q of field breaks we used:

$$\begin{aligned} \Omega_{p0} &= \Omega_{q0} = 2 \\ \Omega_{p1} &= \Omega_{q1} = 50 \end{aligned} \tag{3.43}$$

Normal prior for background field b was set to $\mathcal{N}\left((,\, 0\right), \tau_b = 1)$. Mixture Gaussian prior distribution segment field a initialized with:

$$\begin{aligned} \vec{r_\alpha} &= c(0.01, 0.98, 0.01) \\ (\mu_{a,-1}, \mu_{a,-1}, \mu_{a,-1}) &= (-0.5, 0, 0.5) \\ (\rho_{a,-1}, \rho_{a,-1}, \rho_{a,-1}) &= (\frac{1}{(0.4)^2}, \frac{1}{(0.3)^2}, \frac{1}{(0.4)^2}) \end{aligned} \tag{3.44}$$

.

The convergence properties of EM BSFM optimization are plotted on Figures 3.5.3 and 3.5.4, on run for clear visibility and many for broader overview. It can be observed that, except for field breaks, and field break

probabilities, parameters converge in 30 iterations. Histogram of running times until convergence (or max 400 iterations) is on Figure 3.5.5.

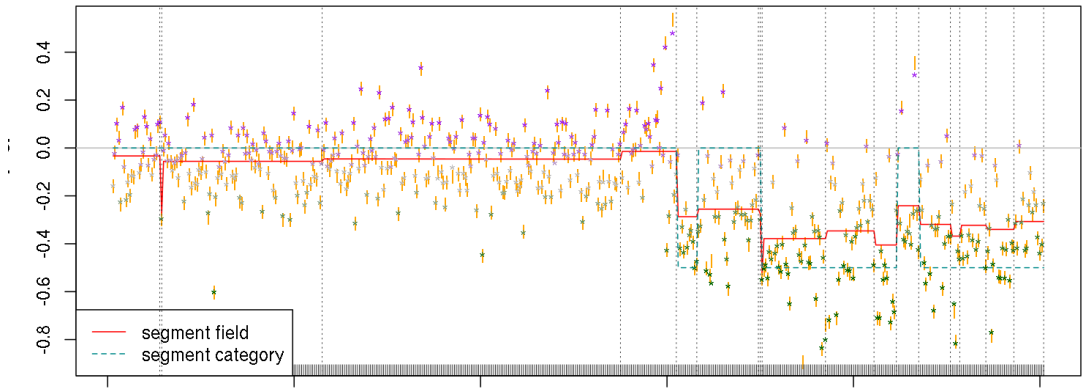


Figure 3.5.1: $\log_2$ratio together with segment field in red. Middle of yellow bars indicate original $\log_2$ratio value, together with shifted points allow to see the correction by the background field.

3.5.1 Comparison with CBS results

To check the validity of segmentations resulting from BSMF we compare 40 BSMF segmentation results with CBS results on the same set of aCGH experiments. Figures 3.5.6 and 3.5.7 summarize this comparison.

3.5.2 Sensitivity to priors

We've run BSMF several times on a selected aCGH experiment with varying prior parameters. On Figures 3.5.8 and 3.5.9 charted is dependence of the final optimal parameters on the rate parameter of the Gamma and Beta distributions.

The author would like to thank dr Bogusław Kluge for conceiving the idea of background and segment fields with breaks, and for the initial EM implementation in **R** with alternating strata on graphs, and for many related and unrelated fruitful discussions. All further modifications to this idea outlined in this chapter, the quadratic problem formulation, their

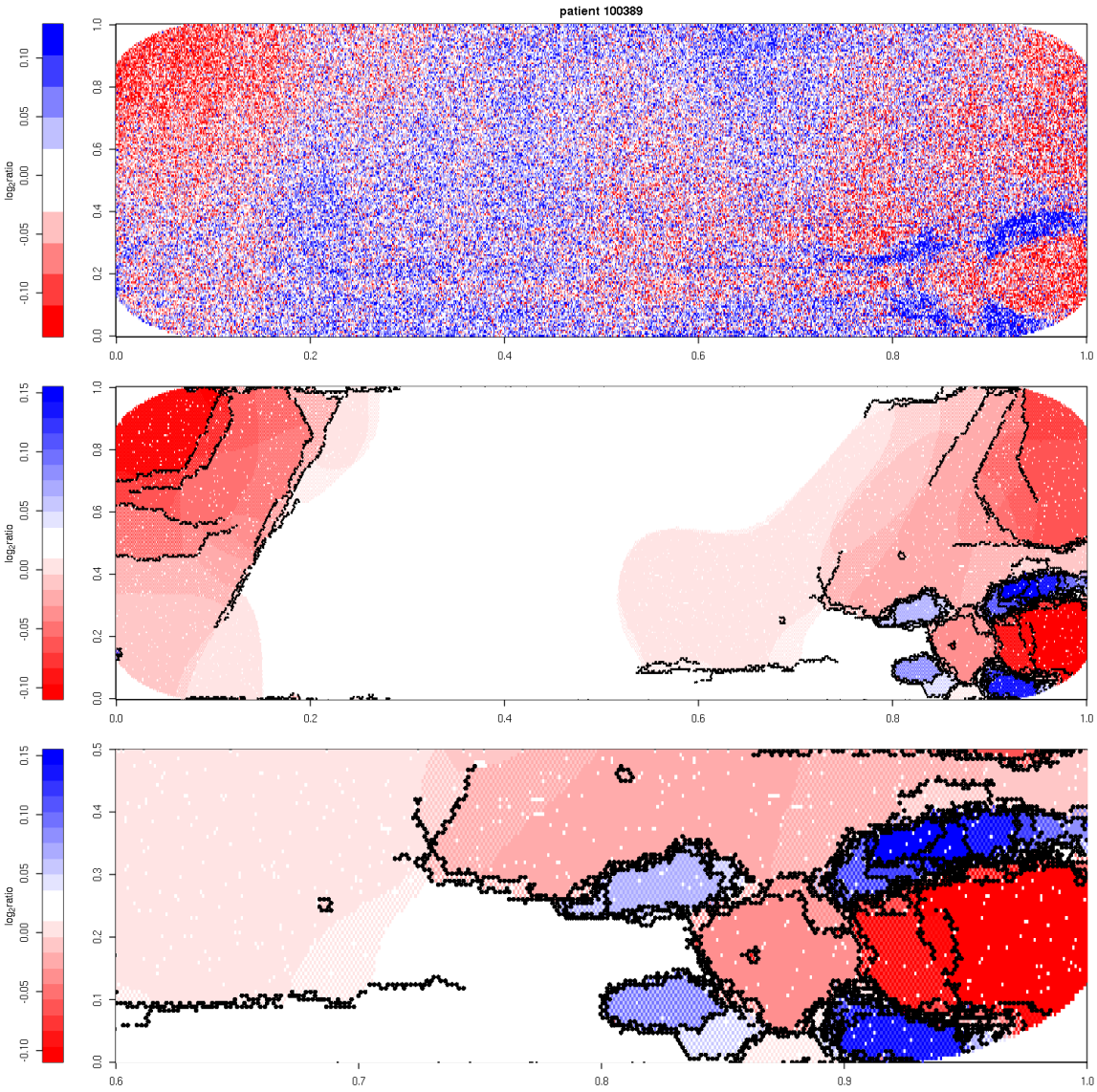


Figure 3.5.2: EM optimized background field of a microarray.

analyses, the experiments and the results elaboration are results of the author.

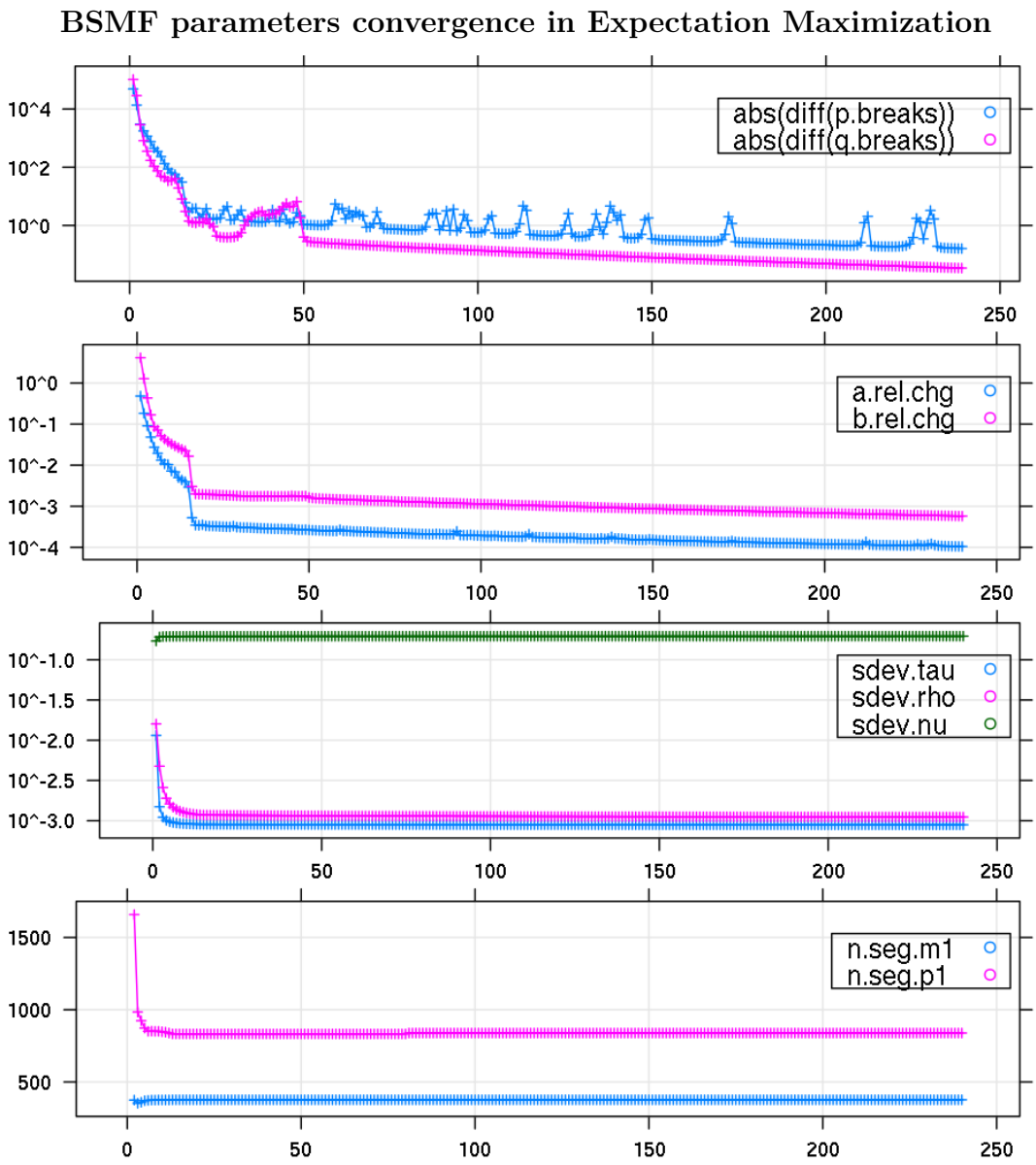


Figure 3.5.3: BSMF convergence in an Expectation Maximization run. Parameters' convergence by iterations (x-axis). Panels from the top: *i)* changes in the number of breaks in segment/background fields(p/q), computed from p, q probabilities *ii)* segment field a and background field b relative change: $\frac{\left\|\vec{a}^{n+1}-\vec{a}^{n}\right\|_2}{\left\|\vec{a}^{n}\right\|_2}$ *iii)* precision parameters τ, ρ, ν converted to standard deviations *iv)* number of features in duplicated segments *n.seg.p1*, in deleted segments *n.seg.s2*, computed from posterior segment categories $s_{i,\alpha}$.

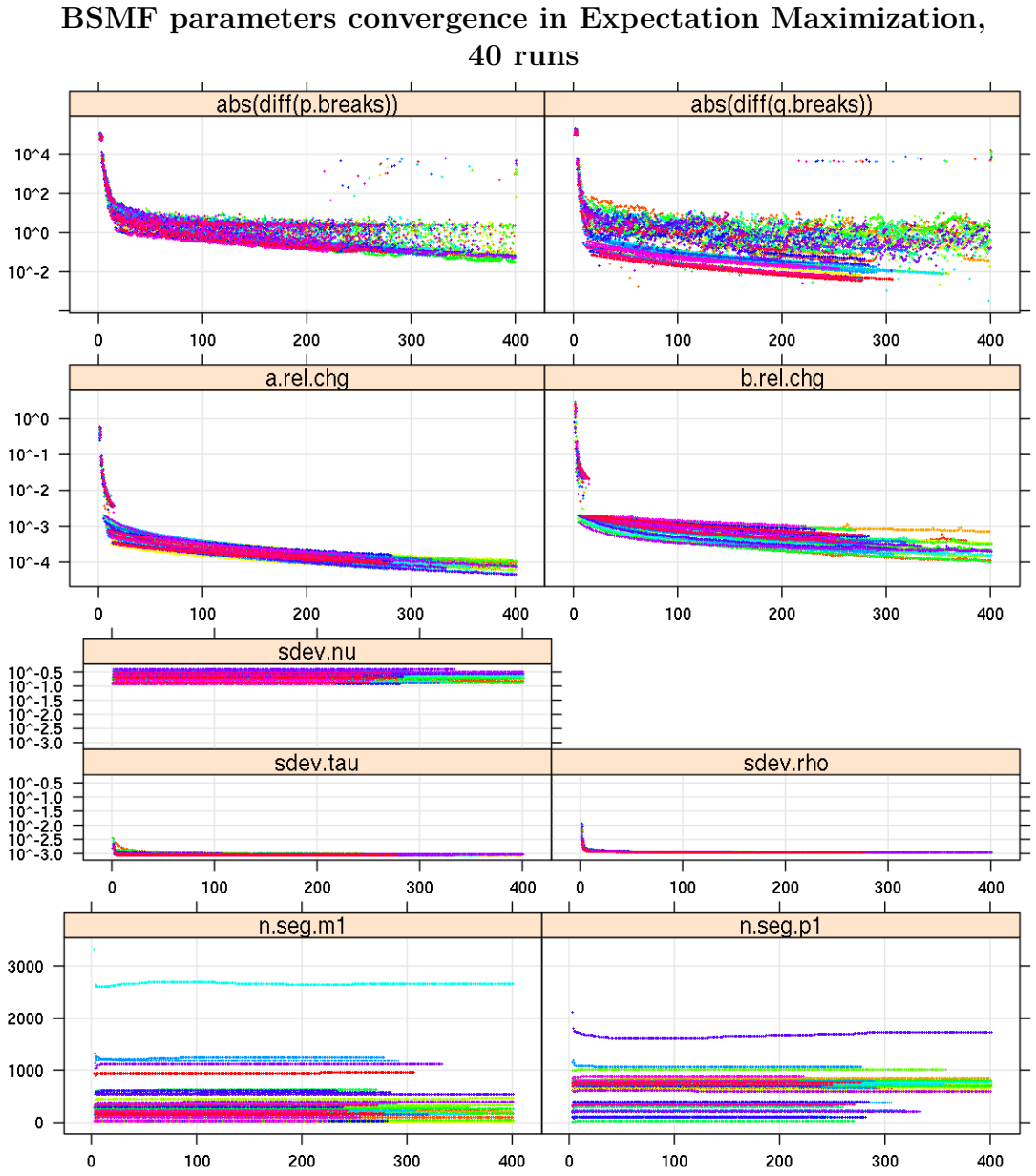


Figure 3.5.4: Data from BSMF EM executions on 40 aCGH microarrays. For the description of charted variables see the caption of Figure 3.5.3.

Running times on Intel(R) Xeon(R) CPU X5690 @ 3.47GHz

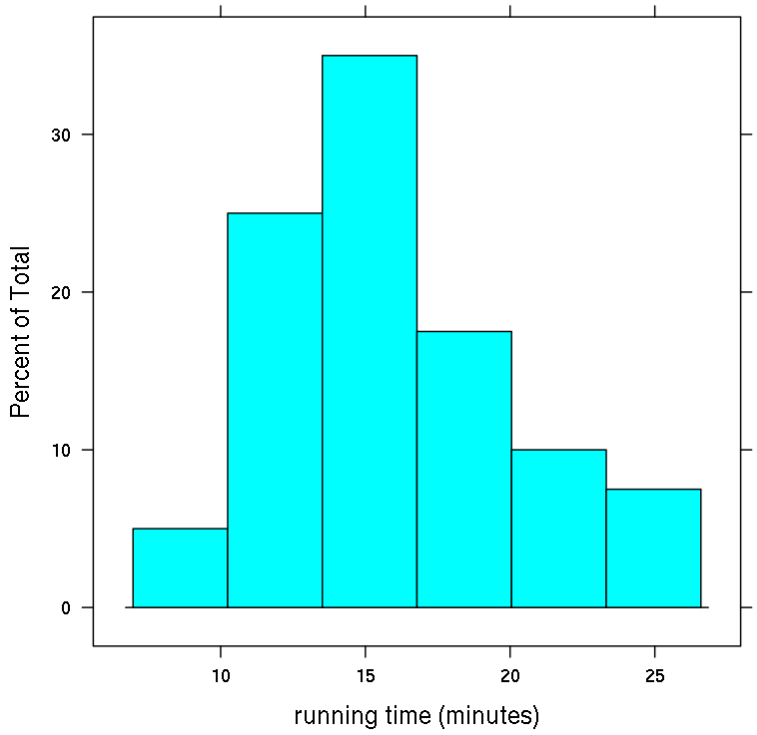


Figure 3.5.5: Running times of 40 runs of BSMF segmentations. Computations were made in parallel on a server with 24 cores. Maximal number of iterations was set to 400. For histograms of number of iterations until convergence ref. to Figure 3.5.6

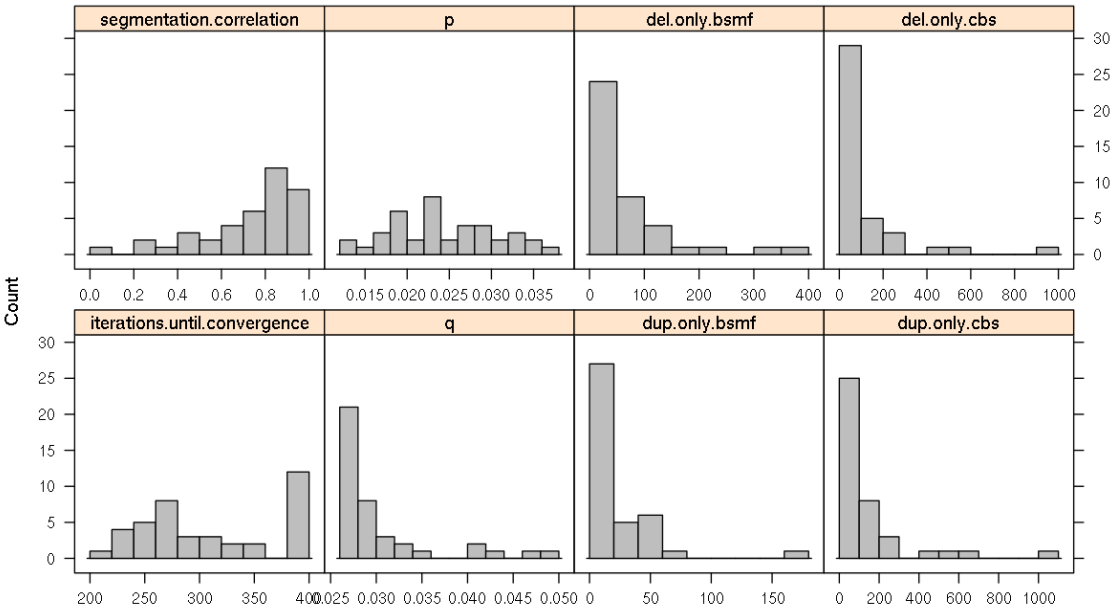


Figure 3.5.6: Comparison between BSMF and Circular Binary Segmentation (CBS), histograms summarizing 40 microarrays. From left to right: *i)* top panel: correlation between $-1, 0, 1$ feature segment indicator vectors (deletion,0,duplication), where high correlation indicates similar segments with |segment mean| > 0.4; bottom panel: BSMF iterations until EM convergence criterion was met *ii)* BSMF posterior probabilities of a break in segment p and background q fields *iii)* feature counts for symmetric difference in detected deletions *iv)* feature counts for symmetric difference in detected duplications.

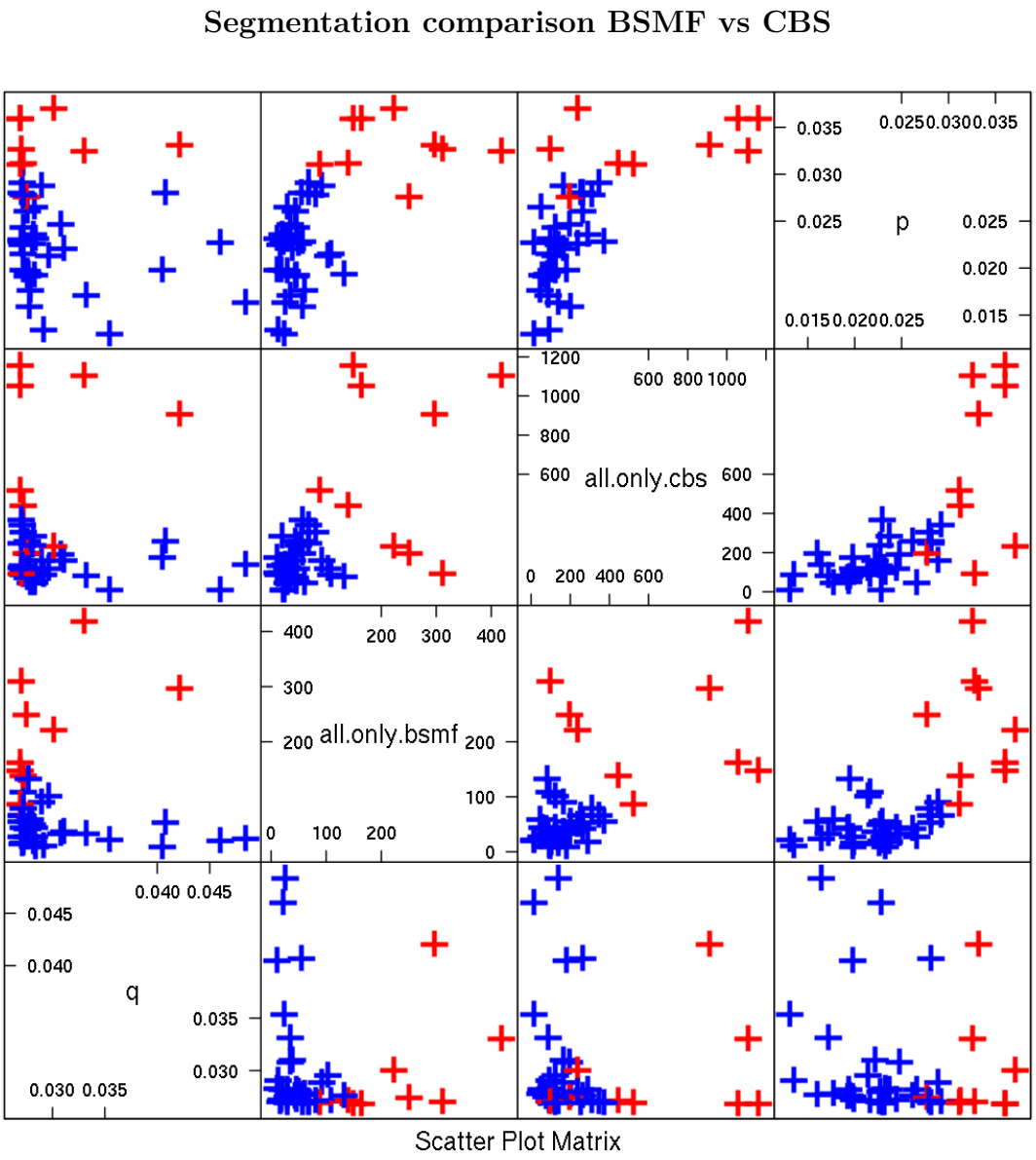
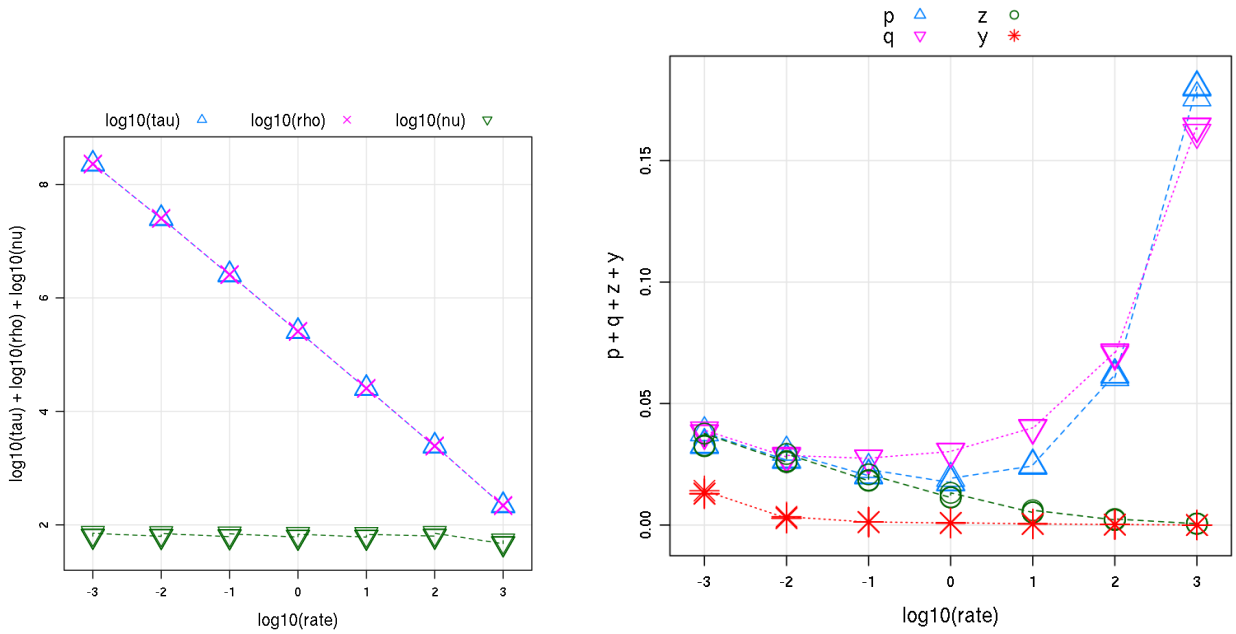
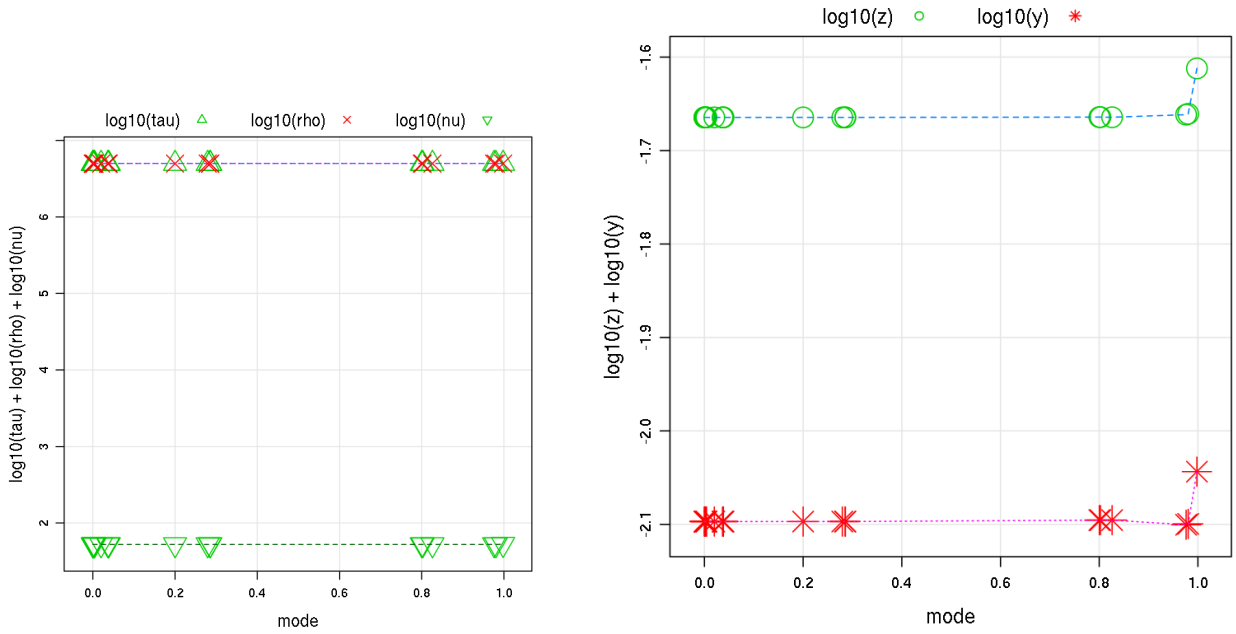


Figure 3.5.7: Comparison with Circular Binary Segmentation (CBS). Marked in red are experiments where the difference in results from by CBS and BSMF were largest in the number of probes in it. Scatterplots reveal dependency between posterior segment break probabilities p, and the size of the symmetric difference between BSMF and CBS segmentations. No such dependency is observed between posterior background noise break probabilities q.



(a) **(b)**

Figure 3.5.8: Sensitivity to the rate of Gamma prior distribution.



(a) **(b)**

Figure 3.5.9: Sensitivity to the rate of Beta prior distribution.

Non, rien de rien
Non, je ne regrette rien
Ni le bien qu'on m'a fait
Ni le mal tout ça m'est bien égal

Michel Vaucaire

4

Functional performance of aCGH design

The array design is the starting point of the study on genomic disorders underlying a given disease, Lemoine et al. (2009) elaborates on the issue. There is a large body of research concerning array design task, see e.g. Lipson et al. (2002, 2007). Similarly, many papers consider the issues of normalization, and detrending of, array CGH data: Chen et al. (2008); Kreil and Russell (2005); Staaf et al. (2007); van Hijum et al. (2008).

It is a reasonable practice, while conducting the large-scale biomedical research projects, to provide several prototype array designs. Very important issue here is the methodology for comparison of the functional quality of different arrays. Often disposing only limited amount of experimental data, researchers develop several array designs and based on their suitable comparison they have to choose the best one for further experiments. The plethora of methods devoted to array design

and normalization studies were proposed in the literature, but only few approaches cope with the problem of comparison between different array designs.

There are some standard statistics calculated for purpose of array comparison. They comprise usually: Signal to Noise Ratio, Derivative Log Ratio Standard Deviation, Background Noise, etc (Carter, 2002). Authors of (Coe et al., 2007) proposed new performance measure called *functional resolution*, which reflect the uniformity of probe spacing on the chip and the sensitivity of the array to single CNVs.

Analogously to other high-throughput technologies (like mass spectrometry or expression microarrays) various sources of technical and biological variation affect the array CGH experiment. The measurement noise comes from the preparation of the microarray slide and the hybridization process, while the biological variability is the result of the heterogeneity of the cells (e.g. mosaicism (Iourov et al., 2008)). However, despite increasing resolution of CGH arrays the variation in signal measurements cannot be eliminated.

Our goal in this Chapter was to develop the framework for performance comparison of different CGH array designs. We decided to explore the concept of robustness. The proposed methodology follows the general concept of robust statistics Hampel et al. (2005), quoting B.D. Ripley *an important area that is used a lot less than it ought to be.*

In our approach we consider the design robust when it is effective in the detection of aberrations in the presence of noise. The segmentation obtained for the given design is treated here as a *robust estimator* of rearrangement regions. Better designs correspond to more robust estimators, i.e., those approximating the aberrations for the data contaminated with the noise.

We decided the built up our method on CBS algorithm for segmentation calling. To test the robustness of a specific design we have to enhance the DNAcopy (package implementing CBS) by incorporating parametrized noise model. Our package named DNAcopyNoise is freely available at `http://bioputer.mimuw.edu.pl/software/DNAcopyNoise`.

Our results are twofold: firstly, using synthetic data we demonstrate

the usefulness of robustness measure for array performance comparison. Secondly, we apply the concept of robustness to select the best one from several optimized designs. The optimization aimed in reducing array size while keeping the same rearrangements detection ability. As a measure of design quality used for optimization purpose we explore so called *noise-induced discrepancy* defined with respect to the performance of the original (full size) array.

4.1 Methods

4.1.1 Synthetic Array Design

Aiming in validation of the robustness approach we generate several datasets using framework from Willenbrock and Fridlyand (2005). Two types of datasets generators are considered: they correspond to different genomic rearrangements structure (high density of relatively short segments, like in cancer tissues versus rare long aberrant segments characteristic to genomic disorders). For each type of data we consider different array designs. E.g., for data of first type, dataset (a) presented in Figure 4.1.1 is the exemplary output of aCGH experiments performed on well designed array. Dataset (b) corresponds to experimental data from the design, in which the inappropriate probe selection resulted in poor hybridization. The generator for dataset (b) is obtained as the following modification of the original generator (a). We choose uniformly at random 20 percent of probes and multiply their signal intensity by the coefficient sampled from Beta distribution with shape parameters $\alpha = 2$ and $\beta = 20$ (unimodal distribution defined on the interval $[0, 1]$).

The generator corresponding to array design giving the dataset (c) mimics the problems arising from erroneous analysis protocol that results in significant background noise. We assume here, that some probes may be erroneously analyzed already during the scanning process and only one from Red (cy5) and Green (cy3) signal is detected. To model such situation we choose uniformly at random 15% of probes and sample their intensities from the beta distribution with parameters $\alpha = 0.7$ and $\beta = 0.7$. Such readouts correspond to the probe signals not well scattered

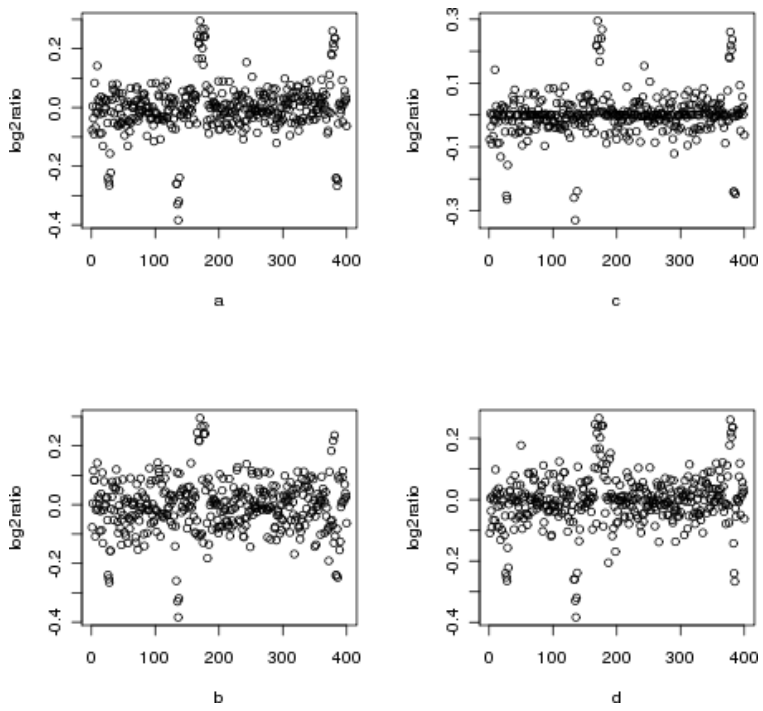


Figure 4.1.1: Plots show log2ratio (y-axis) vs. genomic location (x-axis) for synthetic datasets corresponding to four different array designs: (a) original datasets, (b) dataset with simulated poor hybridization effect, (c) dataset with simulated error-prone analysis procedures, (d) dataset with both effects.

around zero in the typical MA plot (See Section 2.2.3 for definition of MA plot). The design corresponding to dataset (d) suffers from both shortcomings. We generate 40 datasets using each design. One synthetic genome hybridization experiment measure the signal intensities of 10000 probes located on 10 chromosomes.

4.1.2 Exon CGH Array Design

Design quality measures proposed here has been tested on samples obtained in real aCGH experiments. The dataset come from 60 arrays hybridized with DNA from subjects with epilepsy, autism, heart defects and mental disorders. Each experiment was performed on the 180 K exon targeted oligonucleotide array. The construction of exon targeted aCGH microarray is outlined in Section 2.3.

4.1.3 Enhancement of DNACopy

DNAcopy package for R environment implements circular binary segmentation algorithm (Olshen et al., 2004). CBS algorithm finds segmentation by recursively splitting subsequent segments into three, or two smaller ones. Each segment cut is found by maximizing the Z_C statistic, as described in Section 2.2.4.

Segmentation proceeds when the null hypothesis is rejected, that is when Z_C is above upper $\alpha-$quantile of null distribution Z_C^*. CBS algorithm estimates the null distribution with the use of permutation method and tail probability estimation.

We quantify the level of robustness of a segment by introducing a Gaussian noise to the $\log_2$ratio data. The aim is to detect the minimal level of noise that makes the considered segment undetectable with high probability. Finding these values requires extensive sampling as in our model the introduced noise corresponds to highly dimensional random variable. We optimized the implementation by introducing the noise model inside the sampling phase: every permutation in CBS algorithm is sampled with random noise added with mean zero and standard deviation η. This changes the Z_C^* distribution and the sought quantile. This is compared with the previously computed, however scaled accordingly to introduced noise variance, $t_{i+1,j}$ statistic for the analyzed segment. To each aberrant segment k we assign the appropriate noise level η_k by running the original version of CBS segmentation and introducing noise in binary search fashion up to desired precision.

4.1.4 Robustness Measure

It is inevitable that the measurement precision vary considerably between probes depending on the hybridization efficiency. Hence some regions of the genome are analyzed with significantly higher experimental precision than others (Baldocchi et al., 2005). Therefore it is desirable to model the effectiveness of specific array region in detecting aberrations. We propose an approach that allows to evaluate the quality measure for a whole array but also to focus on specific set of probes. In our method we measure the quality of array design using noise robustness of segmentation algorithm

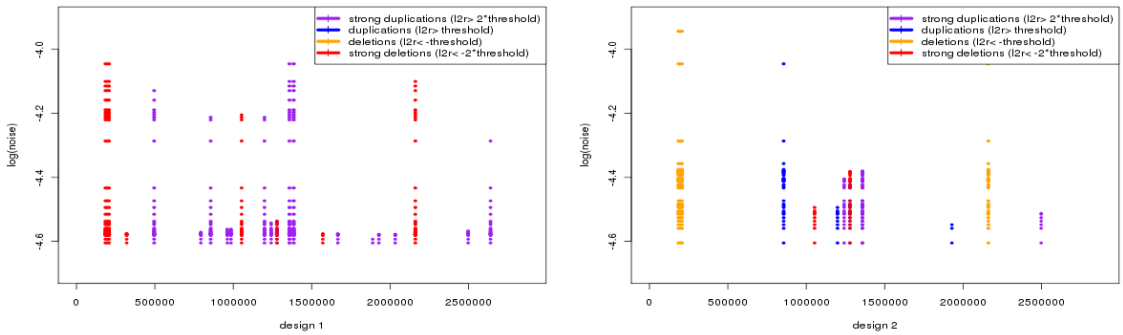


Figure 4.1.2: The resistance of aberrant segments for increasing noise. y-axis correspond to increasing (log-)noise level, different segments are placed along x-axis (genomic location), the $\log_2$ratio are color-coded.

performed for all accessible aCGH experiments.

The intuition behind this approach can be explained in simple terms. Segmentation algorithm provides the information about comparative hybridization experiment. Aberrant segments are easily detectable if they are represented by good quality probes. Good probes should tolerate higher level of measurement noise than poor quality probes. Therefore we conduct segmentation procedure for several increasing noise levels and observe the behavior of aberrant segments. There is certain number of segments found for original experimental data. Then we simulate some measurement noise and repeat segmentation algorithm. Some segments (consisted of poor quality probes) disappear and we continue this process, memorizing for each segment the maximal noise level, for which this segment is still identifiable (for a fixed segment k we denote this value by η_k). The output of several segmentation stages for 2 different (synthetic) designs is presented in Figure 4.1.2. Clearly, the left panel corresponds to more robust design.

Let us fix the aCGH experiment and let η_k denote the noise level of the maximal noise resistance of kth segment defined as above. The level of noise is measured with reference to baseline variation (standard deviation of probes in non aberrant regions). The robustness of probe k is defined

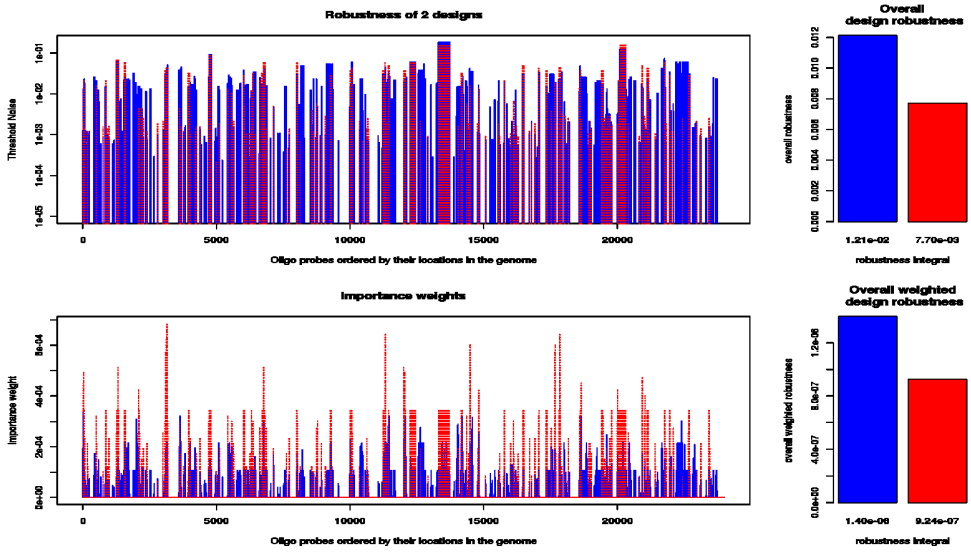


Figure 4.1.3: The robustness compared for two synthetic designs. The robustness has been calculated for all probes (upper plot) as well as corresponding weights importance (lower plot). The structure of genomic rearrangements mimics the abnormalities in cancer cells. Good design is coded in blue. Red design contains 20% of poorly hybridizing probes and 15% of outliers (probes causing erroneous scanning).

as:

$$\theta_k = \frac{\eta_k}{\text{length}(k) \cdot |\text{mean}(k)|} \tag{4.1}$$

where length(k) is the length of kth segment (measured in the number of probes), and $|\text{mean}(k)|$ is the absolute value of mean of signal intensities along the segment. We assign the segment robustness to all the probes it contains.

Now we combine the segmentation robustness of several aCGH experiments into the measure of array design quality. The robustness score for an array is composed from robustness of probes it consists of. Note that, we can estimate the quality only for those probes that are witnesses of some aberration. Consider a single probe k and assume, that it belongs to aberrant segment in some samples (according to segmentation algorithm run for original data). To this probe robustness scores $\theta_k^{i_1}, \theta_k^{i_2}, \ldots \theta_k^{i_m}$ have been assigned in experiments $i_1, \ldots i_m$. Assume,

that there are m accessible experiments in total. As an overall quality of this probe we can take the median of the empirical distribution of robustness scores $\theta_k^{i_1}, \theta_k^{i_2}, \ldots \theta_k^{i_m}$.

However in the case of limited number of accessible experimental data we encounter here the problem of insufficient statistic, because a single probe can be the witness of only few aberrations. To avoid this difficulty we apply the sliding window approach. The empirical distribution of probe robustness is composed for all probes contained in the window of predefined length n (depending on the resolution of an array). The median of this distribution is calculated yielding the smoothed version of the overall probe quality.

The next neighboring window is shifted by the half of the window length. Therefore any single probe contributes to exactly two window statistics (the boundary probes are ignored). Assume that the median (μ_L) from the first window is calculated for i_L events (aCGH experiments in which this probe lies in the aberrant segment) and the second μ_R for i_R events. Then the ith probe robustness for the array $\mathcal{A}$ is defined as:

$$\Theta_i^{\mathcal{A}} = \frac{i_L \mu_L + i_R \mu_R}{i_L + i_R} \tag{4.2}$$

The robustness of array design $\mathcal{A}$ (containing N probes) can be calculated by taking the average robustness of all probes.

However, the important issue here is that the calculation of robustness for some probes relies on many detected aberrations containing this probe, while for others the robustness measure is supported by only few witnesses. Consider once more the probe i and two windows containing it. A support for the ith probe robustness $\Theta_i^{\mathcal{A}}$ is defined as $s_i^{\mathcal{A}} = \frac{i_L + i_R}{nm}$ i.e., the percent of experiments in which this probe or its surrounding probes are witnesses of some aberration.

The support vector is composed of all probe supports $\boldsymbol{s}^{\mathcal{A}} = s_1^{\mathcal{A}}, \ldots, s_i^{\mathcal{A}}, \ldots, s_N^{\mathcal{A}}$. This vector is further transformed into importance weights vector $\boldsymbol{\omega}^{\mathcal{A}} = \omega_1^{\mathcal{A}}, \ldots, \omega_N^{\mathcal{A}}$ by appropriate normalization and scaling (the scaling function flatten out the support vector, as higher support values have roughly the same impact). Finally, the robustness of array

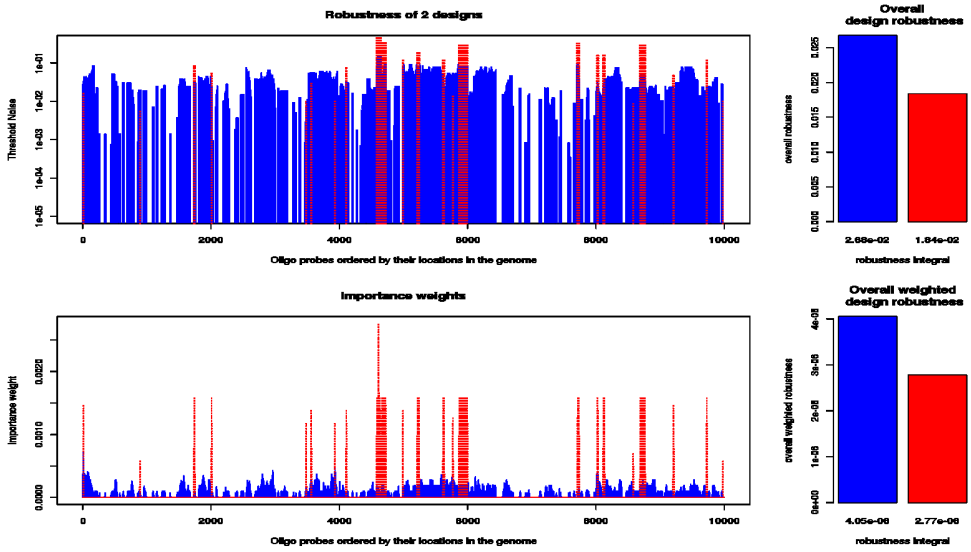


Figure 4.1.4: The robustness compared for two synthetic designs. The robustness has been calculated for all probes (upper plot) as well as corresponding weights importance (lower plot). The structure of genomic rearrangements mimics the abnormalities in classical genetic disorder (relatively rare long aberrant segments). Good design is coded in blue. Red design contains 15% of outliers (probes causing erroneous scanning).

design $\mathcal{A}$ is defined as:

$$\Theta^{\mathcal{A}} = \Sigma_i \omega_i^{\mathcal{A}} \Theta_i^{\mathcal{A}} \tag{4.3}$$

4.1.5 Optimizing Exon CGH Array Design via Noise-induced Discrepancy

The robustness measure $\Theta^{\mathcal{A}}$ defined for a given array design $\mathcal{A}$ allows to estimate the functional performance of $\mathcal{A}$ i.e., the efficiency of rearrangements detection for noisy data. In this section we study the problem of array design optimization. Our goal is to eliminate certain percent of probes to obtain smaller design which has comparable performance.

In the following we make an assumption that the segmentation $\Pi^{\mathcal{O}}$ computed for the original design uncovers the significant genomic signal. The robustness of smaller, filtered designs are measured with respect to

this segmentation. The comparison between the optimized array and the original one is done by means of the gradually changed segmentation for increasing noise level.

Let $\Pi_i^{\mathcal{O}}$ and $\Pi_i^{\mathcal{A}}$ denote segmentations corresponding to sample i analyzed on original $\mathcal{O}$ and optimized design $\mathcal{A}$, respectively. Assuming fixed noise level η and sample number i we define the distance between two segmentations $\sigma_\eta(\Pi_i^{\mathcal{O}}, \Pi_i^{\mathcal{A}})$ similarly to raw distance from (Liu et al., 2006): if both samples express a gain (or loss) at the same loci τ we consider them identical, otherwise this genomic loci contributes to the distance between these segmentation. Below the length of whole genome is denoted by Γ:

$$\sigma_\eta(\Pi_i^{\mathcal{O}}, \Pi_i^{\mathcal{A}}) = \frac{1}{\Gamma} \sum_{\tau:\tau \text{ differs between } \Pi_i^{\mathcal{O}} \text{ and } \Pi_i^{\mathcal{A}}} \text{length}(\tau) \tag{4.4}$$

Above statistic $\sigma_\eta(\Pi_i^{\mathcal{O}}, \Pi_i^{\mathcal{A}})$ corresponds to the fraction of the genome differentiating segmentations $\Pi_i^{\mathcal{A}}$ and $\Pi_i^{\mathcal{O}}$. The total distance σ_η^{tot} is calculated as the average of $\sigma_\eta(\Pi_i^{\mathcal{O}}, \Pi_i^{\mathcal{A}})$ over all m experiments.

$$\sigma_\eta^{\text{tot}(\mathcal{A}|\mathcal{O})} = \frac{1}{m} \sum_{i=1}^{m} \sigma_\eta(\Pi_i^{\mathcal{O}}, \Pi_i^{\mathcal{A}}) \tag{4.5}$$

Finally, the first measure called *noise-induced discrepancy* of smaller array design $\mathcal{A}$ wrt original one $\mathcal{O}$ is defined as follows:

$$\Theta^{\mathcal{A}|\mathcal{O}} = \int_{\eta_{min}}^{\eta_{max}} \sigma_\eta^{\text{tot}(\mathcal{A}|\mathcal{O})} \mathrm{d}\eta \tag{4.6}$$

Proposed measure $\Theta^{\mathcal{A}|\mathcal{O}}$ corresponds to cumulative total distance $\sigma_\eta(\Pi_i^{\mathcal{O}}, \Pi_i^{\mathcal{A}})$ for different noise levels.

Since the *noise-induced discrepancy* includes the average distance $\sigma_\eta(\Pi_i^{\mathcal{O}}, \Pi_i^{\mathcal{A}})$ over all experiments, the result may be influenced by outliers. Moreover, values of $\sigma_\eta(\Pi_i^{\mathcal{O}}, \Pi_i^{\mathcal{A}})$ increase consequently with the noise η, i.e. distances computed for larger η contribute more in overall sum.

To compensate this behavior we propose the notion of *relative noise-induced discrepancy*. Let us fix the sample i and noise level η and

consider statistics $\sigma_\eta(\Pi_i^{\mathcal{O}}, \Pi_i^{\mathcal{A}})$ for all n designs selected for comparison (i.e. $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_n$). Then we make the rank transformation: i.e. instead of distances $\sigma_\eta(\Pi_i^{\mathcal{O}}, \Pi_i^{\mathcal{A}_1})$,$\sigma_\eta(\Pi_i^{\mathcal{O}}, \Pi_i^{\mathcal{A}_2})$,...,$\sigma_\eta(\Pi_i^{\mathcal{O}}, \Pi_i^{\mathcal{A}_n})$ we take their ranks $r_\eta(\Pi_i^{\mathcal{O}}, \Pi_i^{\mathcal{A}_1})$,$r_\eta(\Pi_i^{\mathcal{O}}, \Pi_i^{\mathcal{A}_2})$,...,$r_\eta(\Pi_i^{\mathcal{O}}, \Pi_i^{\mathcal{A}_n})$ [1].

Analogously, to the total distance $\sigma_\eta^{\mathrm{tot}(\mathcal{A}|\mathcal{O})}$, we consider the total ranked distance $r_\eta^{\mathrm{tot}(\mathcal{A}|\mathcal{O})}$, for the array $\mathcal{A}$, as an average $r_\eta(\Pi_i^{\mathcal{O}}, \Pi_i^{\mathcal{A}})$ over all m experiments:

$$r_\eta^{\mathrm{tot}(\mathcal{A}|\mathcal{O})} = \frac{1}{m}\sum_{i=1}^{m} r_\eta(\Pi_i^{\mathcal{O}}, \Pi_i^{\mathcal{A}}) \tag{4.7}$$

Finally, the *relative noise-induced discrepancy* of smaller array design $\mathcal{A}$ with respect to original one $\mathcal{O}$ is defined as follows:

$$R^{\mathcal{A}|\mathcal{O}} = \int_{\eta_{min}}^{\eta_{max}} r_\eta^{\mathrm{tot}(\mathcal{A}|\mathcal{O})} \mathrm{d}\eta \tag{4.8}$$

We would like to emphasize, that the *relative noise-induced discrepancy* $R^{\mathcal{A}|\mathcal{O}}$ is robust to outliers and in contrast to *noise-induced discrepancy* does not favor any noise levels.

4.2 Results and discussion

4.2.1 Synthetic Data

Figure 4.1.3 presents the comparison of two designs evaluated on (synthetic) samples characterized by many relatively short segments (like in cancer tissues). The blue color corresponds to good design. Weaker design (coded in red) contains 20% of poorly hybridizing probes and 15% of outliers. Hence it corresponds to generator (d) from the previous Section.

For all oligo probes we present their robustness $\Theta_i^{\mathcal{A}}$ (upper plot) in logarithmic scale and corresponding importance weights vector $\omega_i^{\mathcal{A}}$(lower plot). It is clearly visible, that the robustness is significantly higher for better (blue) design.

[1] The concept of ranks turns out to be extremely useful in the procedure for rare CNV detection described in the next Chapter.

The evaluation of two other designs tested on typical genomic disorder (not cancer) datasets is illustrated in Figure 4.1.4. Blue color codes the outcome for good design and red color corresponds to design containing 15% of poor probes (yielding $\log_2$ratio readouts classified as outliers), i. e. datasets from this design are obtained from generator of type (c). Analogously as for previous example, the better design yields higher array robustness.

4.2.2 Noise-induced Discrepancy of Optimized Designs

The case study using synthetic datasets justifies that robustness measure can be applied for estimation of the design performance in detecting aberrant regions. The second measure proposed here, i.e. *relative noise-induced discrepancy* turned out to be useful in aCGH design optimization process.

The starting point of the optimization procedure was the calculation of per-oligo quality score. The authors of Mulle et al. (2010) suggest that the high variance of a fixed probe in the $\log_2$ratio across multiple samples originates e.g. from the unreliable binding by target sequences and therefore is related to the poor probe quality. More importantly we observed that also good performing probes may have a high variance in the $\log_2$ratio whenever they are located in polymorphic regions. Therefore we decided to consider the $\log_2$ratio deviations from the signal (as a signal we mean the segment means), denoted by δ_i for $i \in (1, \ldots, N)$, where N states for the number of probes in the array.

Next we observed, that the value of segment mean influences the distribution of these deviations (i.e. higher segment means corresponds to higher deviations). To cope with this problem we replaced a global measure of a probe quality by the simple score that reflects the oligo suitability in the context of surrounding probes. Consider a probe i and its neighborhood (e.g. probes $(i-1), (i+1)$), denoted by i^*. For $\log_2$ratio data from all considered samples we performed the Kolomogorov-Smirnov (K-S) test comparing the distribution of δ_i in all samples to the distribution of δ_{i^*}. From now on, we interpret the p-values form K-S test as a measure of the functional performance of given probe.

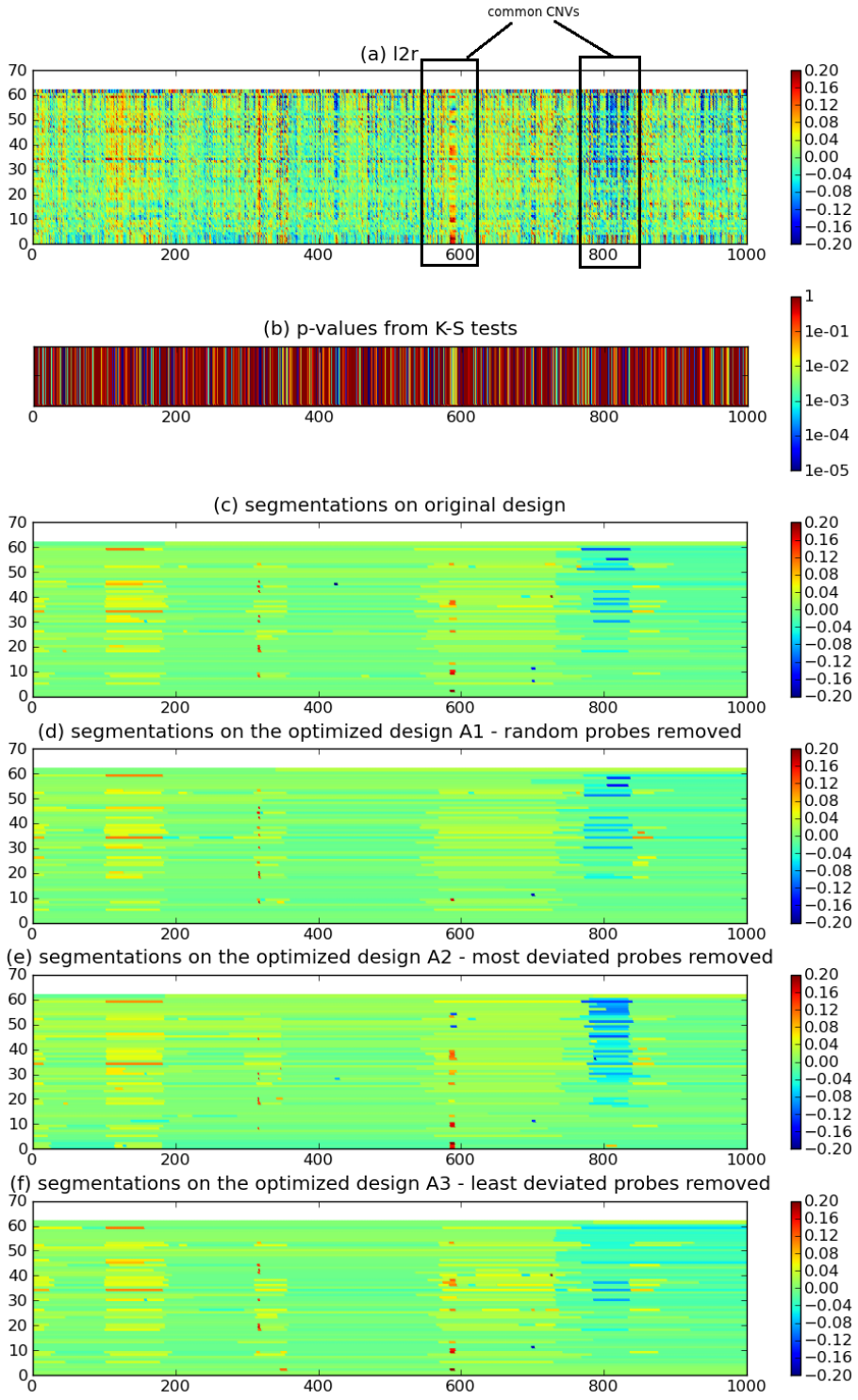


Figure 4.2.1: Segmentations performed on original and optimized designs (see description in the main text). A spoiled probe can be visible around 300 mark.

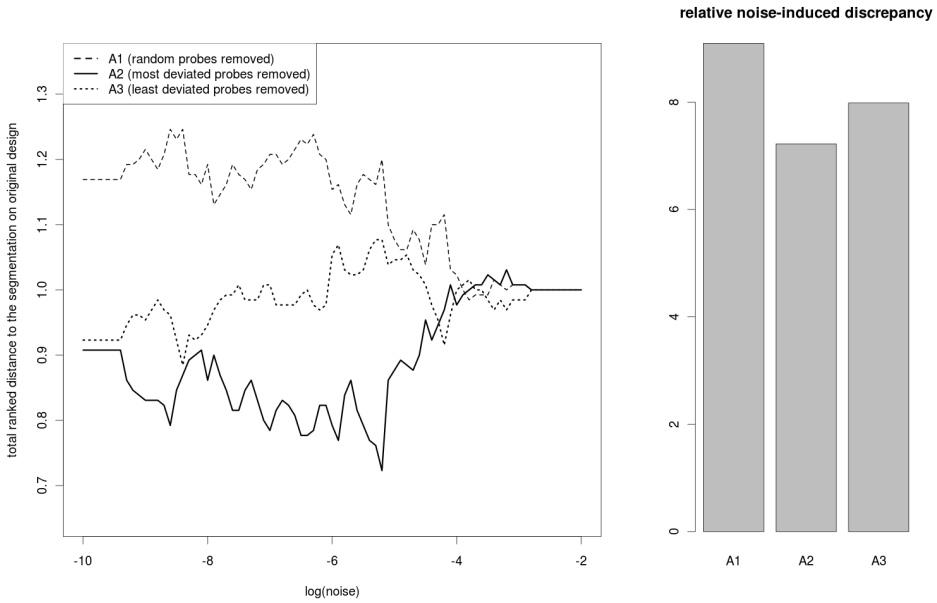


Figure 4.2.2: Comparison of *relative noise-induced discrepancy* $R^{\mathcal{A}_i|\mathcal{O}}$ for three optimized designs (see description in the main text).

Smaller (i.e. optimized) designs were proposed after analysis of the data from 60 aCGH experiments, performed on the 180 K array (Boone et al., 2010) gathered in the software system mentioned in the last Chapter. The analysis aimed in selecting 80% of oligos from original design while keeping the ability to detect all significantly aberrant segments. The influence of design optimization strategy on the robustness of smaller design was analyzed in three case studies. We tested three different approaches of probe selection: (i) uniform sampling (optimized design $\mathcal{A}_1$); (ii) removal of most deviated oligos (optimized design $\mathcal{A}_2$); (iii) removal of least deviated oligos (optimized design $\mathcal{A}_3$). We assumed that the most deviated probes had the lowest p-value in KS-test and the least deviated oligos were defined as probes with the highest K-S p-value. The performance of segmentation method for all described designs is illustrated in Figure 4.2.1. Panel (a) presents the fragment of 1000-oligos distributed along the x-axis; the y-axis corresponds to different samples ($\log_2$ratio value is color-coded). The black boxes mark CNVs common

to almost all considered samples (i.e. duplications near oligo 600-th, and deletions near 800-th.). With high probability these CNVs correspond to benign polymorphisms (not interesting from medical point of view) but in the sequel we use them as a positive controls. Panel (b) visualizes the p-values from K-S tests performed for each oligo on the original design. Panels (c) - (d) present the comparison in segmentation quality for different array designs: segmentations performed on the original design and segmentations performed on the optimized designs $\mathcal{A}_1$, $\mathcal{A}_2$, and $\mathcal{A}_3$, respectively.

Finally Figure 4.2.2 shows the comparison of *relative noise-induced discrepancy*. The left subfigure shows the total ranked distance $r_\eta^{\mathrm{tot}(\mathcal{A}_i|\mathcal{O})}$ versus increasing noise η. The right subfigure presents the values of *relative noise-induced discrepancy* $R^{\mathcal{A}_i|\mathcal{O}}$.

4.3 Conclusions

Several improvements for presented methodology are possible. The challenging problem is whether DNAcopy segmentation method may be replaced by more efficient one (e.g. new segmentation method based on a wavelet decomposition (Ben-Yaacov and Eldar, 2008b)). In our approach we used several simplified assumptions regarding to noise model. In particular, we assumed the Gaussian distribution of signal variation, i.e. we believe that the quality of probe hybridization results in symmetric deviation of signal intensity. However, this assumption may be too optimistic and more profound analysis of the noise behaviour may suggest another type of distribution, e.g. non-symmetric or bi-modal.

Moreover, we treat each probe independently, neglecting possible spatial correlation of the noise. Although this kind of correlation may exists in real data, the standard design protocol places probes on the chip in the random fashion with respect to genome location. This justifies the assumption of non-correlated random perturbation of each probe.

Obviously the limitation of our method is availability of rearrangement data. Proposed measure of design quality does not necessary reflect the global sense of quality, but rather corresponds to

quality of detection of DNA changes in the already analyzed data. Since, our approach is devoted to the analysis of targeted microarrays designed for clinical diagnosis of specific disorders (Bartnik et al., 2012; Wiśniowiecka-Kowalnik et al., 2013), we assume that the coverage of possible rearrangements in previously analyzed DNA is complete enough.

Finally, other possibility of improving array quality is to replace worst performing probes by newly designed oligos or to optimize the spatial rearrangements of probes. However such protocol would require extensive experimental validation.

I have come to believe.

Fox Mulder

5

Rare CNV detection

In recent years there has been an increase in number of probes on the array in aCGH technology – the genomic resolution has improved. Designed high resolution arrays target in detection of changes in single exons, small as several hundred base pairs in size, and facilitate a better detection of CNVs. This helps in clinical interpretation of changes in patients with various clinical phenotypes, especially when a CNV overlaps with a gene known to be causative of the observed clinical phenotype (Boone et al., 2010).

The progress in array resolution increases challenges in aCGH data analysis. The main goal when analyzing aCGH data is to identify genomic regions with rearrangements. The specific challenge in clinical genetic diagnostics is to detect strictly pathogenic CNVs (Koolen et al., 2009).

The primary hallmark of CNV's pathogenicity is its rarity in the population. CNV is considered rare if it is not polymorphic. An aCGH sample signifies a rare CNV if it differs significantly from other samples

in the same genomic region.

Detection of CNVs from aCGH data is a process of separation from noise contiguous blocks of signal along a patient's genome (Figure 5.0.1). This analysis process is called segmentation, and there are plethora of methods and algorithms for detection of CNVs through segmentation.

There are many approaches to identify and describe the structure of the intervals, such as Gaussian models (Picard et al., 2005), hidden Markov models (Cahan et al., 2008), wavelets (Ben-Yaacov and Eldar, 2008b) and quantile regression (Eilers and de Menezes, 2005). Unfortunately, most of the methods suffer from two significant drawbacks: high computational complexity and restriction to single sample analysis.

There exist few methods applicable for the simultaneous analysis of many aCGH samples (Diskin et al., 2006; Mitchell et al., 2007; Nowak et al., 2011), but they are useful only for cancer data analysis, as they assume frequent rearrangement patterns. On the other hand, in the analysis of rare CNVs underlying genomic disorders one has to eliminate non–pathogenic (frequent) polymorphisms. Therefore a different approach should be developed, being the analogue of SCOUT method for rare CNVs in SNP microarrays (Mefford et al., 2009).

An important phenomena, that may affect the aberration calling is waviness. The presence of wave pattern seems be correlated with the GC content or replication timing of the probes, but the underlying mechanism remains unexplained. Waves were observed for different platforms based on DNA hybridization, e.g. ChIP-on-chip DNA methylation studies (Cardoso et al., 2004; Leprêtre et al., 2010).

Almost all segmentation methods detect too many segments (false positives) for dataset containing wave-like noise (the phenomenon occurs especially in tumor samples (van de Wiel et al., 2009)).

Thus, the set of rearrangement regions detected in the segmentation phase needs to be cleared of segments corresponding to non–pathogenic polymorphic changes, wave patterns, and spurious segments resulting from disrupted DNA probes.

In our study, we have focused on *in silico* detection, and supervised

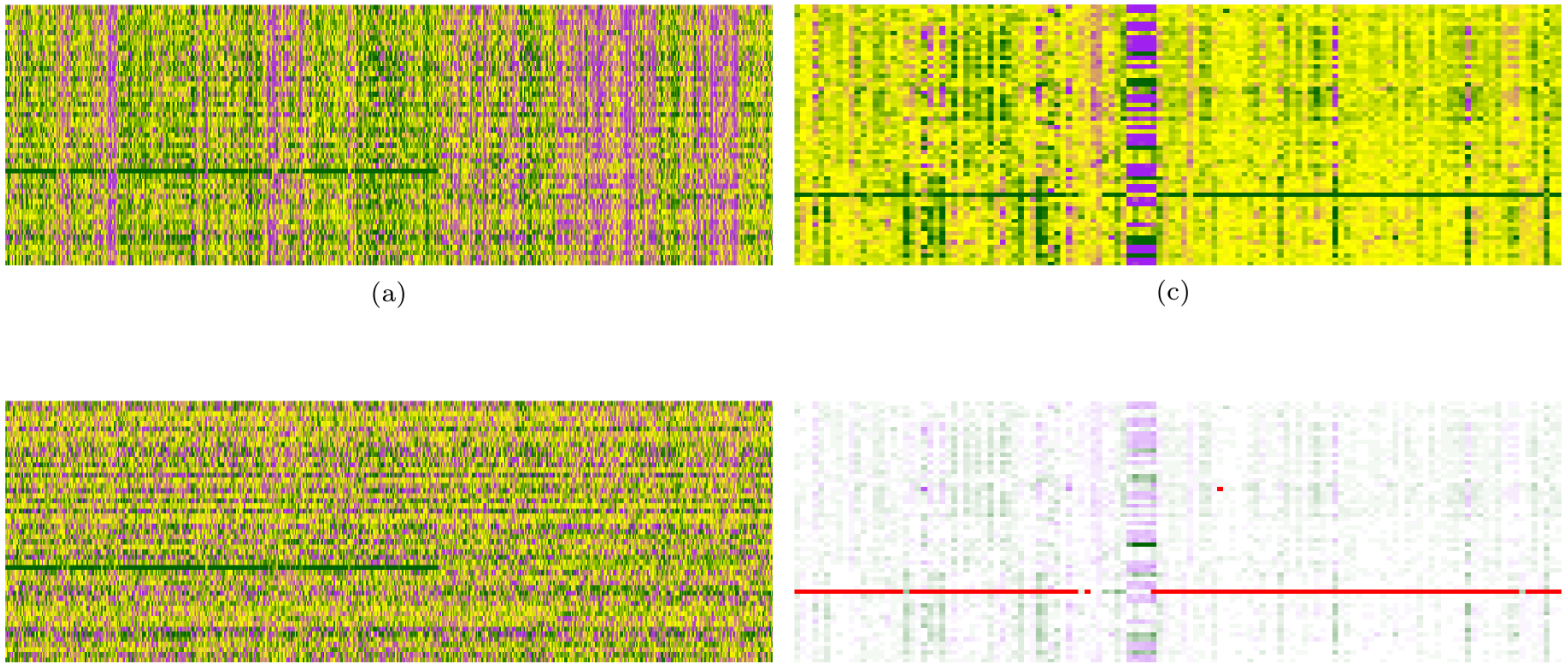
verification, of rare CNVs (i.e. non–polymorphic and outstanding) underlying diverse spectrum of diseases in human. We analyze aCGH samples from a cohort of 366 patients (180 K probes custom exon–targeted CGH array (Boone et al., 2010)) from patients with developmental delay/intellectual disability, epilepsy, or autism. Patients were examined by Institute of Mother and Child, Warsaw, Poland (IMID). We develop and apply novel robust outliers detection procedure to identify aberration associated segments corresponding to the potentially pathogenic changes. We simultaneously process all accessible samples from patients to strengthen information about rearrangements patterns.

To this task we create a procedure which analyzes aCGH data from all samples (a logratio matrix), and detects short fragments of k consecutive probes (k–mers), which are markers of rare CNVs, and which are used to assign statistical, and clinical, significance to detected CNVs.

We augment typical normalizations steps with data transformation to ranks. We propose an outlier statistics, based on ranks, which identifies markers as lying in a 1% tail of the null distribution. This follows the definition of rare pathogenic CNVs, which are nearly absent in control population and present in 1% or less of affected individuals.

From the set of outstanding segments, we sieve out those corresponding to the non–pathogenic polymorphisms, and filter them basing on three main publicly available databases storing the information related to genomic variations and diseases: International Standards for Cytogenomic Arrays database (ISCA) (Faucett, 2010), Genetic Association Database (GAD) (Zhang et al., 2010) and Database of Genomic Variants (DGV) (Zhang et al., 2006).

Our protocol results with a set of medically relevant CNVs. The validation sets the sensitivity of our method for rare CNVs detection to be 96%, and the specificity to be about 94%. We summarize the most interesting 18 CNV segments predicted by our method, that require further analysis (e.g. FISH) in the Validation section. These regions are suspected of being significant to autism, or to mental retardation.



(a) (c)

(b) (d)

Figure 5.0.1: Processing of logratio data In each subfigure, rows corresponds to samples and columns to probes. On the left: the effect of rank transformation; the same fragment of the genome represented by logratios (a) and their column ranks (b). The wave pattern is eliminated, while true signal (clear deletion) is strengthen. On the right: the polymorphic region in the middle is surrounded by wave patterns and only one significant deletion is visible (c); markers found by our algorithm indicate only deleted segment, all other spurious signals are ignored (d).

5.1 Methods

5.1.1 Datasets

The dataset comes from 366 arrays hybridized with DNA of patients suffered from epilepsy, autism, or other neurodevelopmental disorders (developmental delay/intellectual disability) examined at the IMC Cytogenetics Labs. Each experiment was performed on the 180 K custom whole–genome microarray with an exonic–coverage for over 1700 known and candidate genes for neurodevelopmental disorders (Boone et al., 2010).

Microarrays were prepared on Agilent platform, hybridized and scanned by Agilent scanner. We used Agilent Feature Extraction software with default settings, which performs back-ground subtraction, array spatial detrending, dye normalization and logratio calculations from each microarray (Zahurak et al., 2007) (Agilent Technologies).

For further analysis we used outputted logratios – each sample consists of a set of ∼180K logratio intensities mapped to loci in the reference genome hg18 human assembly.

FISH, Multiplex ligation-dependent probe amplification (MLPA), or Polymerase chain reaction (PCR) methods were used for experimental validation.

5.1.2 Outstanding CNVs detection

Although logratio data is already normalized by microarray extraction software, we observe noisy patterns in it: wave bias and experimenter's bias (Figure 5.0.1, also see Discussion). Wave bias has been documented in the literature before (van de Wiel et al., 2009).

Here, we propose a simple and intuitive solution to overcome these two pertaining obstacles, Namely, we replace the logratio signal by its rank, i.e. we analyze the logratio signal relative to other samples.

Figure 5.0.1 justifies the beneficial effect of this approach: at panel (a) the fragment of the genome with hybridization signal is coded by logratios and by their ranks (panel (b)) . After the rank transformation both wave pattern (causing spurious segment calls) and disrupted probes are

eliminated, without affecting the significant segments (one large deletion is visible).

In our approach we analyze aCGH data from all samples simultaneously (the logratio matrix) seeking for markers of rare CNVs (as a marker we mean here short fragment of k consecutive probes – k–mer). Recall that we are interested in rare pathogenic CNVs, which should be nearly absent in control population and present in 1% or less of affected individuals. Therefore our markers correspond to outliers in the set of all k–mers for all samples (presented results were obtained for a parameter $k = 7$). As the outlier detection in high dimensional spaces is a non–trivial task, we decided to use distance-based approach with a suitable choice of metrics (Gogoi et al., 2011).

More precisely, we apply sliding window approach on a rank transformed logratios matrix as follows: for each window spanning the range of k columns, we calculate the distances between the k–mers from all samples. Then, for each k–mer, we compare the average distance to all others in the same window. The crucial step here is to approximate the distribution of average distances between k–mers and classify the k–mer as a outlier (marker) if it lies in a 1% tail of this distribution.

More formally, consider a large $\log_2$ratio matrix L, with dimensions $|S|\times|Q|$, and one of its k-windows L^S_Q, containing $\log_2$ratio data coming from a set of patients $S = \{1, \ldots, n\}$, and from consecutive probes from the set $Q = \{p, \ldots, p + k - 1\}$ (here probe ordering respects probes positions on the reference genome). The transformation of each of k columns into ranks and division of resulting ranks by $|S|+1$ yields *pseudo–ranks* matrix R^S_Q with elements:

$$R^s_q = \frac{\text{rank of } L^s_q \text{ in } L^S_q}{|S|+1}, \quad s \in S, q \in Q \tag{5.1}$$

Let us consider, that S is a patient group sampled from a large group of all patients $\mathcal{S}$, and that rows of R^S contained in $[0, 1]^k$, are in fact pseudo-ranks in columns of $\mathcal{S}$, respectively. Now, R^s_q, taken from a random patient s and probe q, has uniform distribution. Hence, R^S_Q is a sample from distribution $\mathcal{D}_p$ with *c.d.f.* $D_p : [0, 1]^k \to [0, 1]$ with uniform

marginals: $D_p(1,\ldots,u_i,\ldots,1) = u_i \quad \forall i$. However, observe, that if one or more patients in the sample exhibit CNV segment, columns $R^S_{q\in Q}$ are correlated with each other, hence $\mathcal{D}_p$ is not uniform on $[0,1]^k$.

In statistics, distributions with uniform marginals on a hyper-cube $[0,1]^k$ are commonly described using copulas. C is a k-dimensional copula if C is a joint cumulative distribution function of a k-dimensional random vector on the unit cube $[0,1]^k$ with uniform marginals. Several families of copulas (Gaussian copulas, t-copulas, Archimedean copulas), and their properties, were thoroughly studied in literature.

Our method for discriminating outliers is based on a statistics computed for each of n patients: *mean L_q distance to other rank vectors.*

$$\mu^q(s) = \frac{1}{|S|}\sum_{j\in S}\left(\sum_{l=p}^{p+k-1}\left|R^s_l - R^j_l\right|^q\right)^{\frac{1}{q}}, \quad s\in S,\ q\in(0,\inf] \tag{5.2}$$

For the purpose of this work we selected L_1 distance measure, both for simplicity and greater robustness than L_2.

In the case of one dimension $k=1$ and in the continuous limit $|S|\to\inf$, the value of the μ^1 statistics for a patient with pseudo-rank $z\in[0,1]$ is given by:

$$\mu^1(z) = \int_0^1 |t-z|\,dt = z^2 + (1-z)^2 \tag{5.3}$$

$\mu^1(z)$ is monotonous over $z\in[0,\frac{1}{2}]$, and symmetric with respect to $\frac{1}{2}$, z has uniform distribution. Substituting $u = 2|z-\frac{1}{2}|$ we obtain the inverse cumulative distribution function, and further the cdf and the density of the null distribution for $k=1$.

$$\begin{aligned} \mathrm{F}^{-1}_{\mu^1}(u) &= \left(\frac{1+u}{2}\right)^2 + \left(\frac{1-u}{2}\right)^2 = \frac{u^2+1}{2} \quad , \quad u\in[0,1] \\ \mathrm{F}_{\mu^1}(x) &= \sqrt{2x-1} \quad , \quad g_{\mu^1}(x) = \frac{1}{\sqrt{2x-1}} \quad , \quad x\in[\frac{1}{2},1] \end{aligned} \tag{5.4}$$

For $k>1$ the value of the μ^1 statistics for a patient with pseudo-ranks

$z = (z_1, \ldots, z_k) \in [0,1]^k$ is given by:

$$\mu^1(z) = \sum_{i=1}^{k} \int_0^1 |t - z_i| \, dt = \sum_{i=1}^{k} z_i^2 + \sum_{i=1}^{k} (1 - z_i)^2 = ||z||_2^2 + ||1^k - z||_2^2 \tag{5.5}$$

This signifies that the μ^1 statistics converges in limit $|S| \to \infty$ to the sum of squared euclidean distances from two extreme corners of hypercube: 0^k and 1^k (a k-mer in each of these corners has extreme ranks on every probe).

For $k > 1$ if we undertake the independence of pseudo-ranked columns the null distribution $D^k_{\mu^1}$ of μ^1 can be computed as a sum of independent variables. This underlines the adequacy of statistics μ^1 as it converges to the sum of squared euclidean distances from two extreme corners of hypercube: 0^k and 1^k (a k-mer in each of the corners has extreme ranks on every probe). Figure 5.1.1 presents μ^1 limit $|S| \to \infty$ null distributions for various dimensions k for the dimension independence case.

On the other hand, the null hypothesis may assume a certain structure of column correlations, e.g. corresponding to a larger group of patients with CNV segments inside a particular window, and a null distribution may reflect that. First approach we've taken is to fit as a null distribution Beta(α, β) shifted to the appropriate interval $(\min(\mu^1), \max(\mu^1))$. This outlier detection procedure is considered less conservative since Beta has a lighter tail than the $D^k_{\mu^1}$ for small k.

Second approach presupposes that the distribution of k-mers of pseudo-ranks is described by a certain copula C. In case the rank distribution is a certain copula $\mathcal{D}_p = C$, the *c.d.f.* of the null distribution $\mathcal{F}_\mu$ is estimated through approximation of the following integral, by either computing it numerically, or through sampling from the fitted copula C:

$$\begin{aligned} \mathcal{F}_\mu(m) &= \int_{[0,1]^k} \mathbb{1}_{\mathcal{F}_\mu^{-1}(z_1,\ldots,z_k) \leq m} d\mathcal{D}_p(z_1, \ldots, z_k) \\ &= \int_{[0,1]^k} \mathbb{1}_{\sum_{i=1}^k F_\mu^{-1}(z_i) \leq m} dC(z_1, \ldots, z_k) \end{aligned} \tag{5.6}$$

Parameters of copula C are fitted for each window, the null distribution is obtained by integration of the μ^1 statistics over

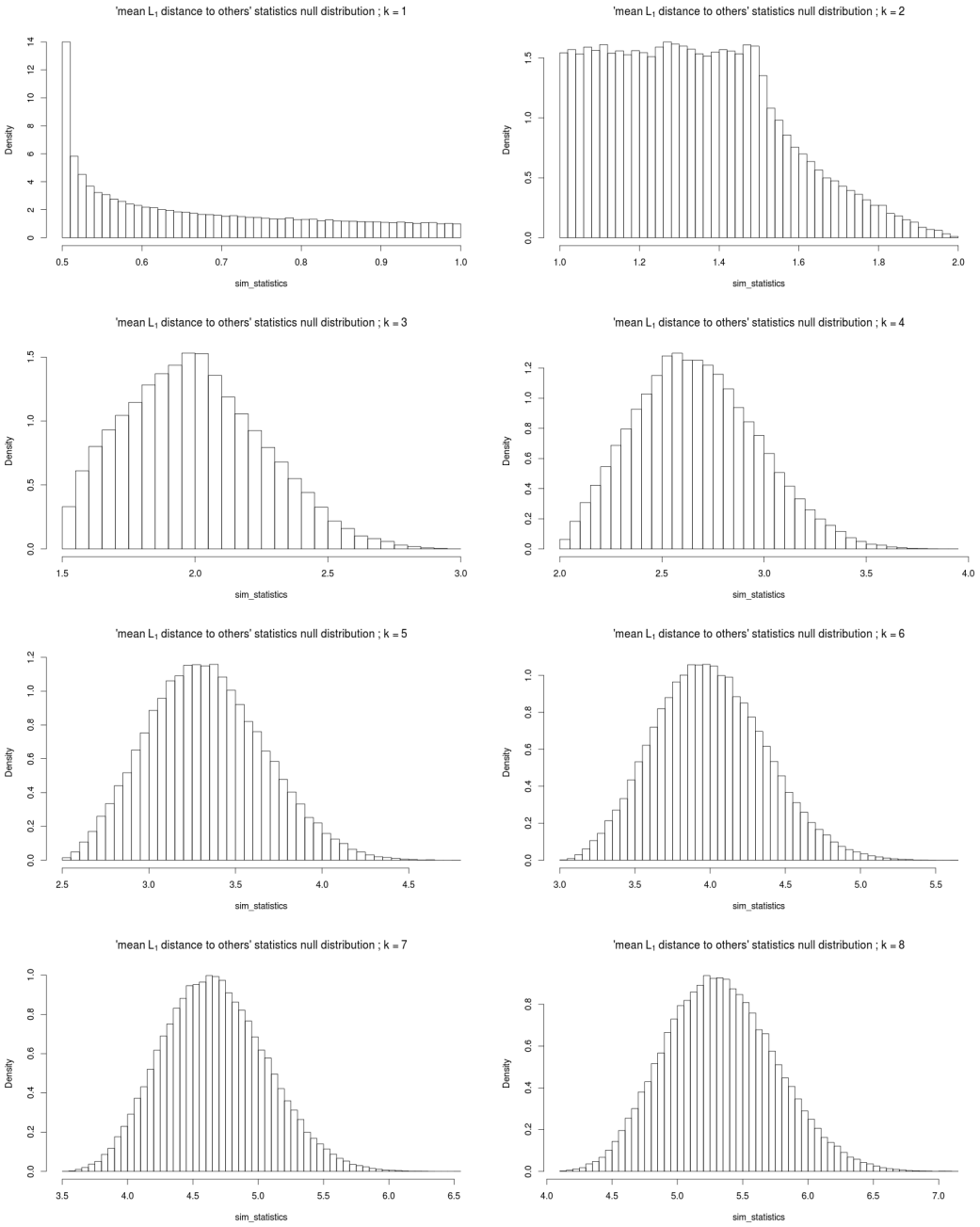


Figure 5.1.1: This figure presents histograms from samples from μ^1 (L_1 distance) null distributions (limit $|S| \to \infty$, number of cases converging to infinity) for various dimensions k. This sampling undertakes the assumption of column (dimensions) independence.

copula C. However, classical families of copulas (Gaussian, t-copula, Archimedean) are not suited to model multidimensional k-mers with

asymmetric dimensional dependencies, a copulas mixture approach is more adequate (Tewari et al., 2011). Then, the mixture approach suffers from huge dimensionality – obtained solutions are only locally optimal, dependent on a mixture fitting starting point. In either approach, k-mers with p-value less than 0.01 (suggested frequency of pathogenic CNVs) are selected as markers.

Markers detected by the outlier detection procedure have to be aligned with the segmentation considered. Then we should filter out segments without any markers inside and order the remaining segments wrt the density of coverage by markers. We assign the *density score* to reported segments (being the percent of the segment covered by markers).

5.1.3 Polymorphic regions filtering

Described method based on outlier detection yield segments corresponding to rare CNVs, but still some segments in highly polymorphic regions could be marked. Therefore the following phase of filtering the non–pathogenic polymorphisms have to be implemented. We map the considered segmentation into probes, i.e. to each probe we assign the value of the mean logratio of the segment containing it. Now, the signal is considered as significant if its absolute value exceeds 0.07 (i.e. log2ratio = 0.24 – the commonly used threshold for aberration).

The number of significant signals in each column i are counted and if there are more than three (1% of 366 samples) we set the i-th coordinate in the vector called *polymorphic profile* to 1, otherwise it is set to 0. In the next step we identify all k-mers of consecutive ones in the polymorphic profile vector. Outlier k-mers that overlap with polymorphic profile are excluded from further analysis.

5.1.4 Validation

To validate segments selected as rare CNVs according to our density score we automate the process of a manual validation of segments based on UCSC (Rhead et al., 2009), i.e. the protocol by which geneticians usually act. Lastly, we compare resulting sets of segments with the set produced manually by geneticists from IMID.

Manual validation by geneticists involves inspecting reported CNVs segments, overlaying them on UCSC tracks. This purposes to filter out known polymorphisms and, by interrogation of all known syndrome regions, to try to narrow down the segment set to only those clinically relevant. This step is followed by FISH or PCR confirmation of the CNVs existence in patient's DNA (Derwińska et al., 2012) (Bartnik et al., 2012). For the automated process we decided to focus on three main databases storing the information related to genomic variations and diseases resulting from it: ISCA (Faucett, 2010), DGV (Zhang et al., 2006) and GAD (Zhang et al., 2010).

During the validation procedure, we correlate the coverage density score for segments with the contents of mentioned databases. We expect, that for medically relevant CNVs the significant intersection with DGV and a non–empty intersection with ISCA and/or GAD should occur.

Figure 5.1.2: Rare CNVs detected by our method in 366 samples. Figure shows the chromosmal location of all segments reported by experts (red), segments predicted by our method (yellow) as well as pathogenic CNVs reported in ISCA (purple) and genes from GAD (blue).

5.2 Results and discussion

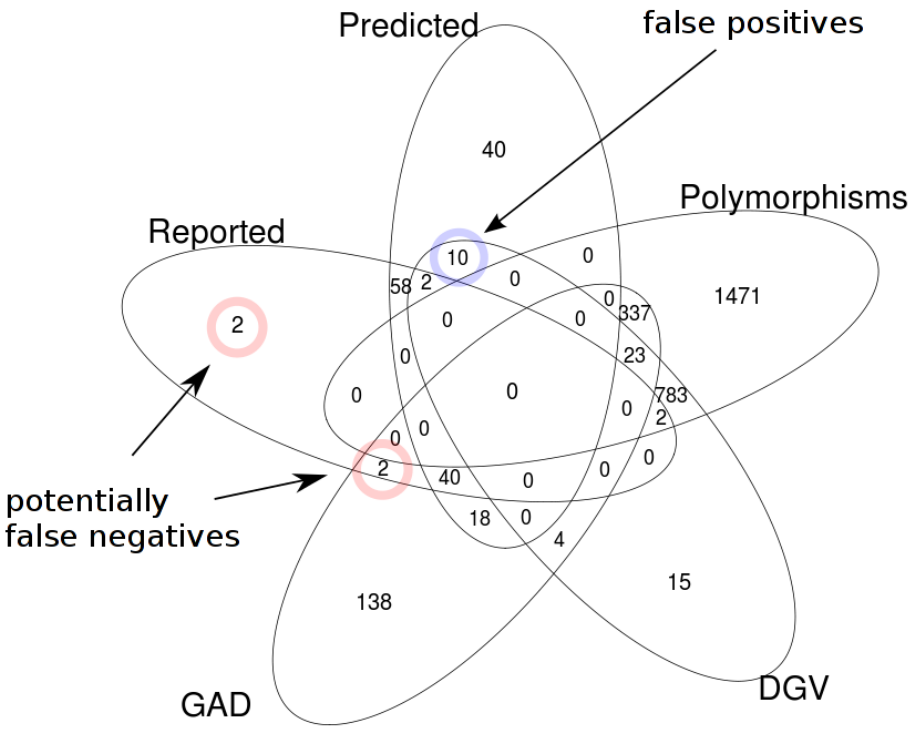


Figure 5.2.1: Evaluation of CNVs detection results. Venn diagram for the predicted rare CNVs (Predicted), confirmed as pathogenic or uncertain (Reported), segments significantly overlapping with DGV (DGV), segments with GAD genes (GAD), and segments selected as polymorphisms according to polymorphic profile (Polymorphisms).

The outcomes of our CNVs detection algorithm are presented in Figure 5.2.1. The Venn diagram depicts five subsets of the set of ca. 3000 segments having absolute mean logratio value greater than 0.1. Considered sets comprise: all reported segments, predicted segments with density score above 50%, segments containing GAD genes, segments with significant intersection with DGV and segments classified as polymorphic.

The procedure based on polymorphic profile eliminated 98% of segments with DGV content (probably benign rearrangements), and 1808 other segments, which are also polymorphic but not reported in DGV. We can estimate the sensitivity of our method to be ca. 96% (only 4 segments reported by experts are missed from all 102 reported but non-DGV segments) and the specificity about 94% (as false positives we classify 10 predicted but non-reported segments having significant DGV intersection).

5.2.1 Discovery and validation of rare CNVs

Proposed method identified 168 potentially pathogenic duplications and deletions (having coverage density score 50%). Importantly, 100 rearrangements have been also confirmed as real pathogenic changes (c.f. Table 5.2.1) or as changes of unknown significance suitable for further analysis (listed in Table 5.2.2).

case	gain/loss	cytoband	size (Mb)	oligo nr.	score	ISCA	GAD	diagnosis
1	del	1q43q44	3.3	166	100%	24	—	mental retardation
2	del	3q13.2q13.31	4.5	154	99%	4	2	autism
3	del	Xp22.12	1.6	100	95%	53	1	mental retardation
4	del	17q21.31	0.3	84	96%	23	3	mental retardation
5	del	5q14.3q15	5.4	596	95%	5	—	mental retardation
6	del	Xq22.1q22.3	5.2	167	91%	51	2	mental retardation
7	del	2q37.2q37.3	6.3	736	88%	19	2	mental retardation
8	del	15q13.3q14	8	873	87%	2	2	mental retardation

Table 5.2.1: Selected predicted best scored CNVs confirmed later as pathogenic changes.

case	gain/loss	cytoband	size (Mb)	oligo nr.	score	ISCA	diagnosis
1	del	8q22.2	0.25	56	100%	5	autism
2	del	5q35.3	0.7	27	93%	6	mental retardation
3	dup	3p26.3	0.33	21	90%	10	mental retardation
4	dup	12q24.32	0.4	9	88%	7	mental retardation
5	dup	4q28.2	0.12	77	87%	6	mental retardation
6	dup	3p22.3	1.2	30	83%	4	autism
7	dup	6q25.3	0.9	17	82%	4	mental retardation
8	del	4q21.23q21.3	0.95	22	81%	4	autism

Table 5.2.2: Selected predicted best scored variants of unknown significance

5.3 Conclusions

Many recent studies have emphasized the role of CNVs in the etiology of many human diseases, with rare variants being particularly important (Mefford and Eichler, 2009). Current methods for detection of CNVs in individual samples are not capable to infer such information, while most approaches for multi sample analysis focus on frequent CNVs present in tumor samples. We propose the efficient solution filling this gap that can be used for accurate detection of rare CNVs and has potential use in clinical diagnostics. Since our procedure produces a set of markers for rare CNVs, it may be efficiently used to filter a segmentation produced by any other segmentation algorithm, and help with identification of segments corresponding to rare pathogenic polymorphisms.

The ongoing study on a group of 366 individuals confirmed large part of our predictions (see previous section, Table 5.2.1 and 5.2.2).

Moreover, the validation of the proposed segments scoring indicates the significant enrichment of high scoring segments in disease genes from GAD database and impoverishment in benign CNVs present in DGV database. Futhermore, the extensive intersection of rearrangements detected by us with data stored in ISCA indicates the potential pathogenic changes in our segments. The presented method is robust in the sense of sensitivity to outliers coming from spurious probes, or any singular outliers of other type, when comparing to segmentation on each sample separately

(DNACopy algorithm was used in this study). Last but not least, it is also resistant to waviness. The DNA–copy segmentation algorithm used in the first stage of our method can be replaced by any other procedure, and more importantly it can be also skipped at all. In that case, we can cluster the significance markers found during the second phase along the genome to obtain longer segments. This idea leads to multi-sample segmentation algorithm that can be highly efficient and we plan to exploit it in the future.

Success consists of going from failure to failure without loss of enthusiasm.

Winston Churchill

6

Semantic Web technologies for molecular medicine

Results from array experiments, either RNA expression arrays, or aCGH DNA arrays, are difficult to interpret, in clinical setting, and in research. The first step involves separating signal from noise with specialized algorithms, and mapping signals to genomic regions. In the second step, data from each sample is still large and contains many signals of various importance. The main task of the analysis is to assess importance, and assign meaning to those signals. For the quality of diagnosis, or research, it's crucial to investigate and validate findings from experiment results, underline phenotype-genotype links, with the use of various external data sources. To facilitate this process many genome browsers were made available in the last decade, most prominent examples being UCSC. However, genomic browsers are not sufficient, and the whole process benefits from specialized software.

6.1 IMID2py – a web application tool for aCGH analysis

Our team in cooperation with the Institute of Mother and Child in Warsaw (IMID) and Baylor College of Medicine developed and deployed an aCGH analysis software in use at the IMID Cytogenetics Lab. IMID2py enables gathering data, removing noise, statistical signal calling on aCGH chip scans, signal segments reviewing, and reporting. The software was used to analyze aCGH experiment results for research and in clinical setting. Our team participated in the process of microarray chip design, and further analysis.

The overall aim of the project was to analyze four groups of patients suffering from: autism, epilepsy, inborn heart diseases, and mental retardation.

The software tool developed by us helps in managing patients, experiments and computation tasks. More precisely, it enables to (i) present segmentation results, charts, compare patients, summarize genome analysis; (ii) gather phenotype data; (iii) run analyses and (iv) generate reports.

IMID2py simplifies significantly the process of sample analysis by providing the concise presentation of preprocessed and filtered data and customized data filtering. It also gives links to appropriate external tools and databases, like **UCSC Genome Browser** (c.f. Fig 6.1.1) or **Decipher** - **D**atabas**e** of **C**hromosomal **I**mbalance.

The most crucial functionality is detecting segments in patients genomes (further labeling with known genes, microRNAs and known CNVs). To assess the significancy of a given aberrant segment we similar cases from gathered patients are reported.

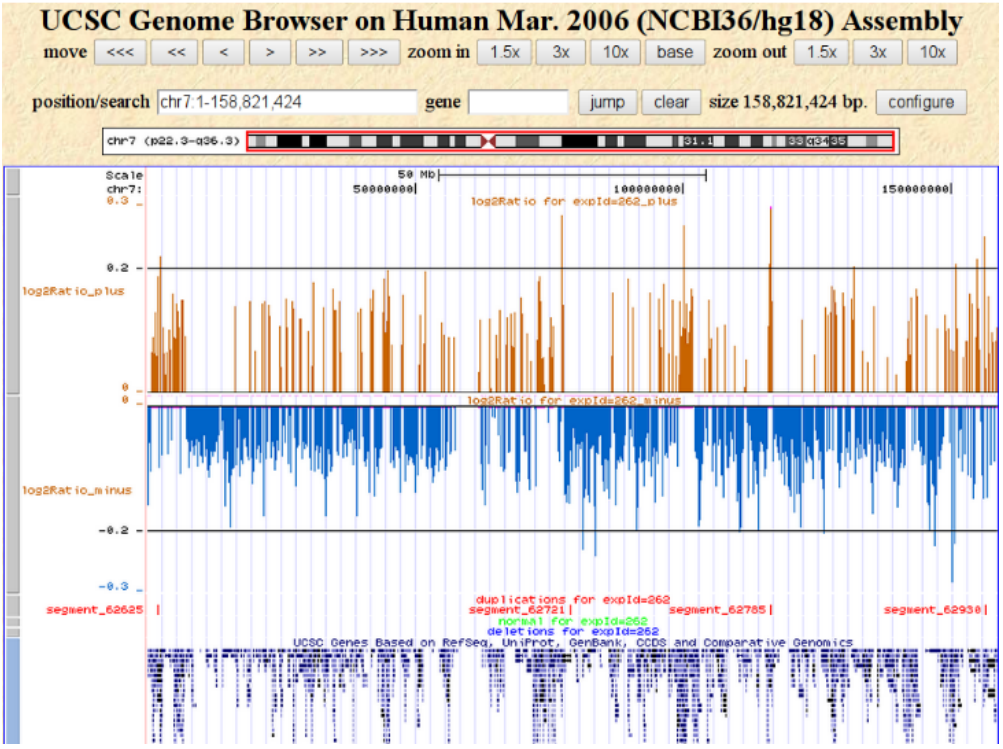


Figure 6.1.1: IMID2py exports data to display in UCSC genome browser.

Tools that are used in IMID2py project include: at data layer – MySQL (main database), SQLite (phenotype form schemes - portability); MVC: Python (web2py web kit); for processing and analysis: R (Bioconductor) and Python (scipy/numpy). All analyses, reports and data are shared in „the cloud". Figure 6.1.2 presents the IMID2py database scheme.

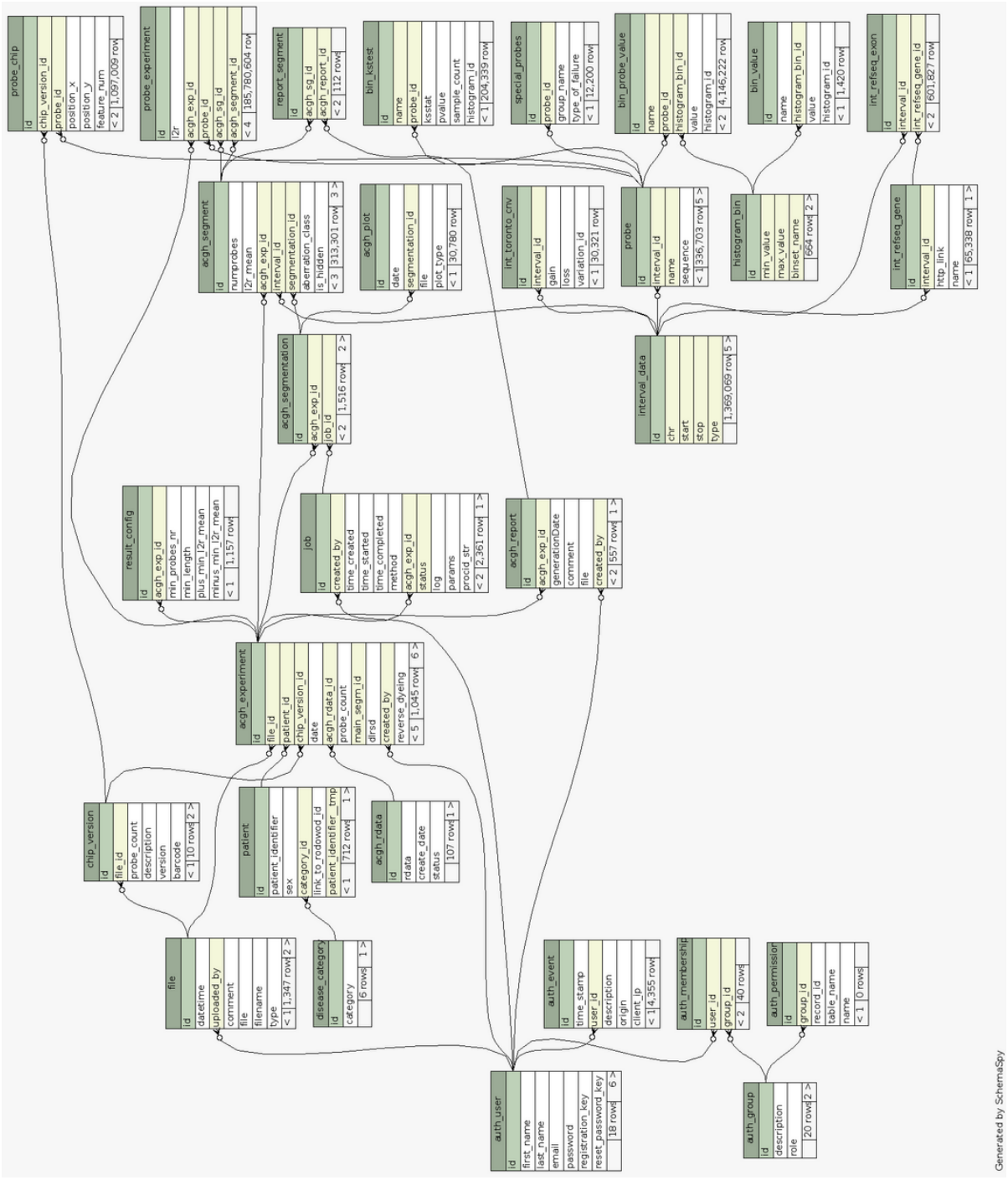


Figure 6.1.2: IMID2py database scheme created with SchemaSpy (Currier, 2005).

6.2 IMID2py semantic extension

Our recent contribution to the process of analysis of genetic data

is an experimental extension of IMID2py with semantic technologies, and with the use of Linked Open Data Sykulski (2012a,b,c). See Figure 6.2.1 shows the screenshot of the webpage presenting the IMID2py semantic extension. Utilizing Apache Stanbol components (Entityhub, Contenthub, Enhancment Engine) enabled us to propose a solution which allows geneticists to: (i) annotate medical content (i.e. a result from a genetic experiment) with relevant data; (ii) formulate various queries to Linked Open Data resources; (iii) use several provided automatic enhancers; (iv) search indexed Linked Open Data entities with VIE auto-complete and (v) search among tagged annotations, with facets provided by Stanbol Contenthub.

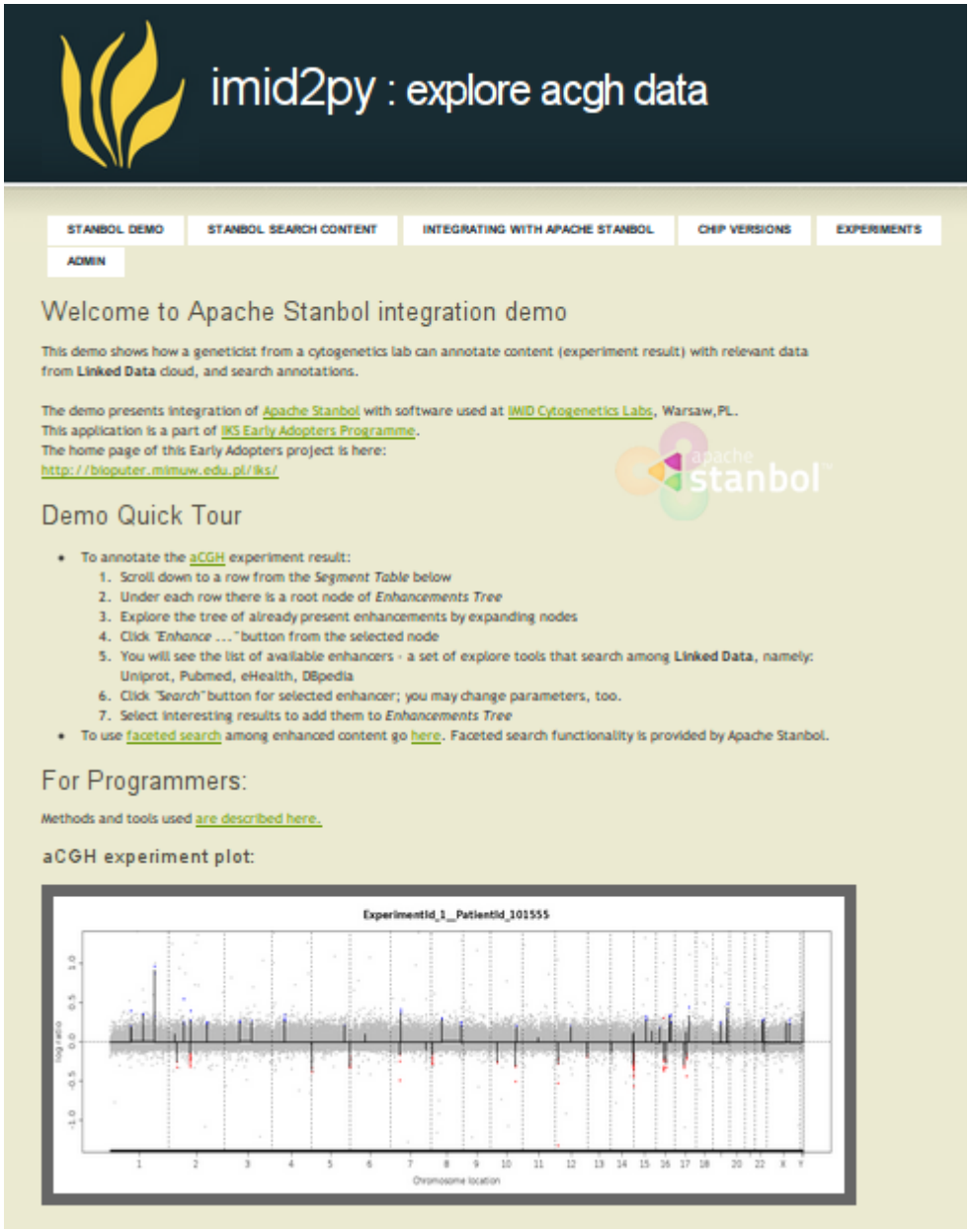


Figure 6.2.1: Webpage presenting the IMID2py semantic extension.

In our solution, user creates a *tree of enhancements* for her/his content: this is a small part of a LOD cloud which users find relevant. A user is able to search for enhancements thanks to Entityhub with pre-indexed linked data from large open databases: UNIPROT, *PubMed, eHealth.*

Users explore Linked Data by asking semi-automatically generated queries, and by reviewing results returned by Stanbol Entityhub. We have

provided several enhancers based on Stanbol Entityhub query language. Finding out from our users what is most useful, and providing them with useful tools is what we were trying to achieve.

IMID2py is used at the cytogenetic lab to review aCGH results from our patients. A geneticists job is to assist doctors in stating clinical diagnosis. However, since genetics is a very rapidly developing field, part of the job is to perform research on difficult, unknown, cases.

By allowing users to easily document their research path in the *tree of enhancements* and later search among, and create reports of their findings, we try to enable reasonable use of constantly-growing Linked Data. Further development will provide more enhancers and Enhancement Chains, abstraction of available enhancers to facilitate more thorough, more automatic research and reporting.

6.3 Biomedical results

As mentioned in the Introduction over 1000 patients were diagnosed with the help of our system. We participated in four projects aiming in the analysis of genetic rearrangements underlying following diseases: epilepsy, developmental delay or intellectual disability, congenital heart defects and autistic spectrum disorder. Below we sketch main results by citing the abstract of articles summarizing the projects.

Epilepsy and additional neurodevelopmental disorders (Bartnik et al., 2012).

Copy-number variants (CNVs) collectively represent an important cause of neurodevelopmental disorders such as developmental delay (DD)/intellectual disability (ID), autism, and epilepsy. In contrast to DD/ID, for which the application of microarray techniques enables detection of pathogenic CNVs in -10-20% of patients, there are only few studies of the role of CNVs in epilepsy and genetic etiology in the vast majority of cases remains unknown. We have applied whole-genome exon-targeted oligonucleotide array comparative genomic hybridization (array CGH) to a cohort of 102 patients with various types of epilepsy with or without additional neurodevelopmental abnormalities. Chromosomal

microarray analysis revealed 24 non-polymorphic CNVs in 23 patients, among which 10 CNVs are known to be clinically relevant. Two rare deletions in 2q24.1q24.3, including KCNJ3 and 9q21.13 are novel pathogenic genetic loci and 12 CNVs are of unknown clinical significance. Our results further support the notion that rare CNVs can cause different types of epilepsy, emphasize the efficiency of detecting novel candidate genes by whole-genome array CGH, and suggest that the clinical application of array CGH should be extended to patients with unexplained epilepsies.

Developmental delay or intellectual disability (Bartnik et al., 2014).

We used whole-genome exon-targeted oligonucleotide array comparative genomic hybridization (array CGH) in a cohort of 256 patients with developmental delay (DD)/intellectual disability (ID) with or without dysmorphic features, additional neurodevelopmental abnormalities, and/or congenital malformations. In 69 patients, we identified 84 non-polymorphic copy-number variants, among which 41 are known to be clinically relevant, including two recently described deletions, 4q21.21q21.22 and 17q24.2. Chromosomal microarray analysis revealed also 15 potentially pathogenic changes, including three rare deletions, 5q35.3, 10q21.3, and 13q12.11. Additionally, we found 28 copy-number variants of unknown clinical significance. Our results further support the notion that copy-number variants significantly contribute to the genetic etiology of DD/ID and emphasize the efficacy of the detection of novel candidate genes for neurodevelopmental disorders by whole-genome array CGH.

Congenital heart defects (Derwińska et al., 2012).

Congenital heart defects are the most common group of major birth anomalies and one of the leading causes of infant deaths. Mendelian and chromosomal syndromes account for about 20% of congenital heart defects and in some cases are associated with other malformations, intellectual disability, and/or dysmorphic features. The remarkable conservation of genetic pathways regulating heart development in animals suggests that

genetic factors can be responsible for a significantly higher percentage of cases. Our aim was to assess of the role of CNVs in the etiology of congenital heart defects using microarray studies. Genome-wide array comparative genomic hybridization, targeting genes known to play an important role in heart development or responsible for abnormal cardiac phenotype was used in the study on 150 patients. In addition, we have used multiplex ligation-dependent probe amplification specific for chromosome 22q11.2 region. We have identified 21 copy-number variants, including 13 known causative recurrent rearrangements (12 deletions 22q11.2 and one deletion 7q11.23), three potentially pathogenic duplications (5q14.2, 15q13.3, and 22q11.2), and five variants likely benign for cardiac anomalies. We suggest that abnormal copy-number of the ARRDC3 and KLF13 genes can be responsible for heart defects. Our study demonstrates that array comparative genomic hybridization enables detection of clinically significant chromosomal imbalances in patients with congenital heart defects.

Autistic spectrum disorders (Wiśniowiecka-Kowalnik et al., 2013).

Autism spectrum disorders (ASDs) are a heterogeneous group of neurodevelopmental disorders, including childhood autism, atypical autism, and Asperger syndrome, with an estimated prevalence of 1.0-2.5% in the general population. ASDs have a complex multifactorial etiology, with genetic causes being recognized in only 10-20% of cases. Recently, copy-number variants (CNVs) have been shown to contribute to over 10% of ASD cases. We have applied a custom-designed oligonucleotide array comparative genomic hybridization with an exonic coverage of over 1700 genes, including 221 genes known to cause autism and autism candidate genes, in a cohort of 145 patients with ASDs. The patients were classified according to ICD-10 standards and the Childhood Autism Rating Scale protocol into three groups consisting of 45 individuals with and 69 individuals without developmental delay/intellectual disability (DD/ID), and 31 patients, in whom DD/ID could not be excluded. In 12 patients, we have identified 16 copy-number changes, eight (5.5%) of which likely contribute to ASDs. In addition to known recurrent CNVs

such as deletions 15q11.2 (BP1-BP2) and 3q13.31 (including DRD3 and ZBTB20), and duplications 15q13.3 and 16p13.11, our analysis revealed two novel genes clinically relevant for ASDs: ARHGAP24 (4q21.23q21.3) and SLC16A7 (12q14.1). Our results further confirm the diagnostic importance of array CGH in detection of CNVs in patients with ASDs and demonstrate that CNVs are an important cause of ASDs as a heterogeneous condition with a variety of contributory genes.

References

III Convention of the Polish Bioinformatics Society, 2010. URL http://www.ptbi3.polsl.pl/files/Program_PTBi_Convention_and_Workshop_2010_ENG.pdf. Cited on pages 12 and 126.

Agilent Technologies. Agilent feature extraction software - user guide. URL http://www.genomics.agilent.com/files/Manual/G4460-90025_FE_User.pdf. Cited on page 91.

Agilent Technologies. Agilent SurePrint g3 human catalog CGH microarrays, 2013. URL http://www.chem.agilent.com/library/brochures/5990-3368en_lo.pdf. Cited on pages xv, 16, 19, 20, 21, and 35.

C. Ambroise, M. Dang, and G. Govaert. Clustering of spatial data by the EM algorithm. In A. Soares, J. Gómez-Hernandez, and R. Froidevaux, editors, *geoENV I — Geostatistics for Environmental Applications*, number 9 in Quantitative Geology and Geostatistics, pages 493–504. Springer Netherlands, Jan. 1997. ISBN 978-90-481-4861-5, 978-94-017-1675-8. URL http://link.springer.com/chapter/10.1007/978-94-017-1675-8_40. Cited on page 37.

S. Auer, L. Bühmann, C. Dirschl, O. Erling, M. Hausenblas, R. Isele, J. Lehmann, M. Martin, P. N. Mendes, B. van Nuffelen, C. Stadler, S. Tramp, and H. Williams. Managing the life-cycle of linked data with the LOD2 stack. In *Proceedings of the 11th International Conference on The Semantic Web - Volume Part II*, ISWC'12, page 1–16, Berlin, Heidelberg, 2012. Springer-Verlag. ISBN 978-3-642-35172-3. doi: 10.1007/978-3-642-35173-0_1. URL http://dx.doi.org/10.1007/978-3-642-35173-0_1. Cited on page 9.

D. Baird, P. Johnstone, and T. Wilson. Normalization of microarray data using a spatial mixed model analysis which includes splines. *Bioinformatics (Oxford, England)*, 20(17):3196–3205, Nov. 2004. ISSN 1367-4803. doi: 10.1093/bioinformatics/bth384. Cited on page 36.

R. A. Baldocchi, R. J. Glynne, K. Chin, D. Kowbel, C. Collins, D. H. Mack, and J. W. Gray. Design considerations for array CGH to oligonucleotide arrays. *Cytometry. Part A: The Journal of the International Society for Analytical Cytology*, 67(2):129–136, Oct. 2005. Cited on page 75.

M. T. Barrett, A. Scheffer, A. Ben-Dor, N. Sampas, D. Lipson, R. Kincaid, P. Tsang, B. Curry, K. Baird, P. S. Meltzer, Z. Yakhini, L. Bruhn, and S. Laderman. Comparative genomic hybridization using oligonucleotide microarrays and total genomic DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 101(51): 17765–17770, 2004. Cited on page 5.

M. Bartnik, E. Szczepanik, K. Derwińska, B. Wiśniowiecka-Kowalnik, T. Gambin, M. Sykulski, K. Ziemkiewicz, M. Kedzior, M. Gos, D. Hoffman-Zacharska, T. Mazurczak, A. Jeziorek, D. Antczak-Marach, M. Rudzka-Dybala, H. Mazurkiewicz, A. Goszczańska-Ciuchta, Z. Zalewska-Miszkurka, I. Terczyńska, M. Sobierajewicz, C. A. Shaw, A. Gambin, H. Mierzewska, T. Mazurczak, E. Obersztyn, E. Bocian, and P. Stankiewicz. Application of array comparative genomic hybridization in 102 patients with epilepsy and additional neurodevelopmental disorders. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 159B(7):760–771, 2012. ISSN 1552-485X. doi: 10.1002/ajmg.b.32081. URL http://onlinelibrary.wiley.com/doi/10.1002/ajmg.b.32081/abstract. Cited on pages 3, 6, 11, 86, 97, and 109.

M. Bartnik, B. Nowakowska, K. Derwińska, B. Wiśniowiecka-Kowalnik, M. Kędzior, J. Bernaciak, K. Ziemkiewicz, T. Gambin, M. Sykulski, N. Bezniakow, L. Korniszewski, A. Kutkowska-Kaźmierczak, J. Klapecki, K. Szczałuba, C. A. Shaw, T. Mazurczak, A. Gambin, E. Obersztyn, E. Bocian, and P. Stankiewicz. Application of array comparative genomic hybridization in 256 patients with developmental delay or intellectual disability. *Journal of Applied Genetics*, 55(1):125–144, Feb. 2014. ISSN 1234-1983, 2190-3883. doi: 10.1007/s13353-013-0181-x. URL http://link.springer.com/article/10.1007/s13353-013-0181-x. Cited on pages 6, 11, and 110.

E. Ben-Yaacov and Y. C. Eldar. A fast and flexible method for the segmentation of aCGH data. *Bioinformatics (Oxford, England)*, 24 (16):i139–145, 2008a. Cited on page 29.

E. Ben-Yaacov and Y. C. Eldar. A fast and flexible method for the

segmentation of aCGH data. *Bioinformatics*, 24(16):i139 –i145, 2008b. Cited on pages 85 and 88.

D. Bickson. Gaussian belief propagation: Theory and aplication. *arXiv:0811.2518 [cs, math]*, Nov. 2008. URL http://arxiv.org/abs/0811.2518. arXiv: 0811.2518. Cited on pages 38 and 46.

P. M. Boone, C. A. Bacino, C. A. Shaw, P. A. Eng, P. M. Hixson, A. N. Pursley, S.-H. L. Kang, Y. Yang, J. Wiszniewska, B. A. Nowakowska, D. del Gaudio, Z. Xia, G. Simpson-Patel, L. L. Immken, J. B. Gibson, A. C.-H. Tsai, J. A. Bowers, T. E. Reimschisel, C. P. Schaaf, L. Potocki, F. Scaglia, T. Gambin, M. Sykulski, M. Bartnik, K. Derwinska, B. Wisniowiecka-Kowalnik, S. R. Lalani, F. J. Probst, W. Bi, A. L. Beaudet, A. Patel, J. R. Lupski, S. W. Cheung, and P. Stankiewicz. Detection of clinically relevant exonic copy-number changes by array CGH. *Human Mutation*, 31(12):1326–1342, 2010. ISSN 1098-1004. doi: 10.1002/humu.21360. URL http://onlinelibrary.wiley.com/doi/10.1002/humu.21360/abstract. Cited on pages 5, 6, 10, 84, 87, 89, and 91.

T. S. Breusch and A. R. Pagan. A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, 47(5):1287–1294, Sept. 1979. ISSN 0012-9682. doi: 10.2307/1911963. URL http://www.jstor.org/stable/1911963. Cited on page 23.

P. Cahan et al. wuHMM: a robust algorithm to detect DNA copy number variation using long oligonucleotide microarray data. *Nucleic Acids Research*, 36(7):e41, 2008. Cited on pages 29 and 88.

J. Cardoso et al. Genomic profiling by DNA amplification of laser capture microdissected tissues and array CGH. *Nucleic Acids Research*, 32(19):e146, Jan. 2004. Cited on page 88.

Carter. Comparative analysis of comparative genomic hybridization micro array technologies: report of a workshop sponsored by the wellcome trust. *Cytometry*, 49(2):43–48, 2002. Cited on page 72.

D. Caserta, M. Benkhalifa, M. Baldi, F. Fiorentino, M. Qumsiyeh, and M. Moscarini. Genome profiling of ovarian adenocarcinomas using pangenomic BACs microarray comparative genomic hybridization. *Molecular Cytogenetics*, 1:10, 2008. Cited on page 5.

H. H. Chen, F. Hsu, Y. Jiang, M. Tsai, P. Yang, P. S. Meltzer, E. Y. Chuang, and Y. Chen. A probe-density-based analysis method for array CGH data: sim ulation, normalization and centralization.

Bioinformatics (Oxford, England), 24(16):1749–1756, 2008. Cited on page 71.

B. P. Coe, B. Ylstra, B. Carvalho, G. A. Meijer, C. Macaulay, and W. L. Lam. Resolving the resolution of array CGH. *Genomics*, 89(5):647–653, 2007. Cited on page 72.

F. S. Collins, M. Morgan, and A. Patrinos. The human genome project: lessons from large-scale biology. *Science*, 300(5617):286–290, 2003. URL http://www.sciencemag.org/content/300/5617/286.short. Cited on page 2.

T. E. P. Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, Sept. 2012. ISSN 0028-0836. doi: 10.1038/nature11247. URL http://www.nature.com/nature/journal/v489/n7414/full/nature11247.html. Cited on page 16.

J. Currier. Schemaspy: Graphical database schema metadata browser. *Source Forge, Aug*, 2005. Cited on pages xiv and 106.

E. F. DeLong, G. S. Wickham, and N. R. Pace. Phylogenetic stains: ribosomal RNA-based probes for the identification of single cells. *Science*, 243(4896):1360–1363, Mar. 1989. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.2466341. URL http://www.sciencemag.org/content/243/4896/1360. Cited on page 16.

K. Derwińska, M. Bartnik, B. Wiśniowiecka-Kowalnik, M. Jagla, A. Rudziński, J. J. Pietrzyk, W. Kawalec, L. Ziółkowska, A. Kutkowska-Kaźmierczak, T. Gambin, M. Sykulski, C. A. Shaw, A. Gambin, T. Mazurczak, E. Obersztyn, E. Bocian, and P. Stankiewicz. Assessment of the role of copy-number variants in 150 patients with congenital heart defects. *Medycyna wieku rozwojowego*, 16(3):175–182, Sept. 2012. ISSN 1428-345X. PMID: 23378395. Cited on pages 3, 6, 12, 97, and 110.

C. W. Dieffenbach, T. M. Lowe, and G. S. Dveksler. General concepts for PCR primer design. *PCR methods and applications*, 3(3):S30–37, Dec. 1993. ISSN 1054-9803. Cited on page 26.

S. J. Diskin et al. STAC: a method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments. *Genome Research*, 16(9):1149–1158, 2006. Cited on page 88.

P. Du, W. A. Kibbe, and S. M. Lin. lumi: a pipeline for processing illumina microarray. *Bioinformatics*, 24(13):1547–1548, 2008. Not cited.

R. Díaz-Uriarte and O. M. Rueda. ADaCGH: a parallelized web-based application and r package for the analysis of aCGH data. *PloS One*, 2 (1):e737, 2007. Cited on page 29.

P. H. C. Eilers and R. X. de Menezes. Quantile smoothing of array CGH data. *Bioinformatics*, 21(7):1146 –1153, Apr. 2005. Cited on page 88.

W. A. Faucett. International standard cytogenomic array consortium. interview by alyson krokosky, sharon f terry. *Genetic Testing and Molecular Biomarkers*, 14(5):585, Oct. 2010. Cited on pages 4, 20, 89, and 97.

P. Flicek, I. Ahmed, M. R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gil, C. García-Girón, L. Gordon, T. Hourlier, S. Hunt, T. Juettemann, A. K. Kähäri, S. Keenan, M. Komorowska, E. Kulesha, I. Longden, T. Maurel, W. M. McLaren, M. Muffato, R. Nag, B. Overduin, M. Pignatelli, B. Pritchard, E. Pritchard, H. S. Riat, G. R. S. Ritchie, M. Ruffier, M. Schuster, D. Sheppard, D. Sobral, K. Taylor, A. Thormann, S. Trevanion, S. White, S. P. Wilder, B. L. Aken, E. Birney, F. Cunningham, I. Dunham, J. Harrow, J. Herrero, T. J. P. Hubbard, N. Johnson, R. Kinsella, A. Parker, G. Spudich, A. Yates, A. Zadissa, and S. M. J. Searle. Ensembl 2013. *Nucleic Acids Research*, page gks1236, Nov. 2012. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gks1236. URL http://nar.oxfordjournals.org/content/early/2012/11/30/nar.gks1236. Cited on page 20.

A. Gambin, S. Lasota, M. Startek, M. Sykulski, L. Noé, and G. Kucherov. Subset seed extension to protein BLAST. In *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms (BIOINFORMATICS 2011)*, page 149–158, 2011. URL http://www2.lifl.fr/~noe/files/pp_BIOSTEC11.pdf. Cited on page 13.

T. Gambin. Design of experiments and genomic data analysis in array-based CGH technology. 2012. URL http://goo.gl/mIofb9. Cited on page 9.

T. Gambin, P. Stankiewicz, M. Sykulski, and A. Gambin. Functional performance of aCGH design for clinical cytogenetics. *Computers in Biology and Medicine*, 43(6):775–785, Jan. 2013. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2013.02.008. URL http://www.computersinbiologyandmedicine.com/article/S0010482513000528/abstract. Cited on page 10.

P. Gogoi et al. A survey of outlier detection methods in network anomaly identification. *The Computer Journal*, 54:570–588, Apr. 2011. Cited on page 92.

A. E. Guttmacher and F. S. Collins. Welcome to the genomic era. *New England Journal of Medicine*, 349(10):996–998, 2003. ISSN 0028-4793. doi: 10.1056/NEJMe038132. URL http://www.nejm.org/doi/full/10.1056/NEJMe038132. Cited on page 2.

T. J. Guzik. Medycyna translacyjna - czyli z laboratorium do łóżka chorego ... i z powrotem. *Kosmos*, 59(1-2):257–262, 2010. URL http://yadda.icm.edu.pl/yadda/element/bwmeta1.element.bwnjournal-article-ksv59p257kz. Cited on page 3.

F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. Wiley Series in Probability and Statistics, 2005. Cited on page 72.

I. Y. Iourov, S. G. Vorsanova, and Y. B. Yurov. Chromosomal mosaicism goes global. *Molecular Cytogenetics*, 1:26, 2008. Cited on page 72.

A. Karimpour-Fard, L. Dumas, T. Phang, J. M. Sikela, and L. E. Hunter. A survey of analysis software for array-comparative genomic hybridisation studies to detect copy number variation. *Human Genomics*, 4(6):421, Aug. 2010. ISSN 1479-7364. doi: 10.1186/1479-7364-4-6-421. URL http://www.humgenomics.com/content/4/6/421/abstract. Cited on pages 4 and 29.

M. Khojasteh, W. L. Lam, R. K. Ward, and C. MacAulay. A stepwise framework for the normalization of array CGH data. *BMC bioinformatics*, 6:274, 2005. ISSN 1471-2105. doi: 10.1186/1471-2105-6-274. Cited on page 36.

D. A. Koolen et al. Genomic microarrays in mental retardation: A practical workflow for diagnostic applications. *Human Mutation*, 30 (3):283–292, 2009. Cited on page 87.

D. P. Kreil and R. R. Russell. There is no silver bullet–a guide to low-level data transforms and normalisation methods for microarray data. *Briefings in Bioinformatics*, 6(1):86–97, 2005. Cited on page 71.

C. Lai et al. SIRAC: supervised identification of regions of aberration in aCGH datasets. *BMC Bioinformatics*, 8:422, 2007. Cited on page 5.

F. Leisch. FlexMix: A general framework for finite mixture models and latent class regression in r, 2003. URL http://epub.wu.ac.at/712/. Cited on page 37.

S. Lemoine, F. Combes, and S. L. Crom. An evaluation of custom microarray applications: the oligonucleotide design challenge. *Nucleic Acids Research*, 37(6):1726–1739, 2009. Cited on page 71.

F. Leprêtre et al. Waved aCGH: to smooth or not to smooth. *Nucleic Acids Research*, 38(7):e94, Apr. 2010. Cited on pages 24, 27, and 88.

S. M. Lin, P. Du, W. Huber, and W. A. Kibbe. Model-based variance-stabilizing transformation for illumina microarray data. *Nucleic Acids Research*, 36(2):e11–e11, Feb. 2008. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkm1075. URL http://nar.oxfordjournals.org/content/36/2/e11. Cited on pages 19 and 25.

D. Lipson, P. Webb, and Z. Yakhini. Designing specific oligonucleotide probes for the entire s. cerevisiae transcriptome. *Algorithms in Bioinformatics*, pages 491–505, 2002. Cited on page 71.

D. Lipson, Y. Aumann, A. Ben-Dor, N. Linial, and Z. Yakhini. Efficient calculation of interval scores for DNA copy number data analysis. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, 13(2):215–228, 2006. Cited on page 29.

D. Lipson, Z. Yakhini, and Y. Aumann. Optimization of probe coverage for high-resolution oligonucleotide acgh. *Bioinformatics*, 23:e77–83, 2007. Cited on page 71.

J. Liu, J. Mohammed, J. Carter, S. Ranka, T. Kahveci, and M. Baudis. Distance-based clustering of CGH data. *Bioinformatics*, 22(16): 1971–1978, 2006. doi: 10.1093/bioinformatics/btl185. Cited on page 80.

J. Liu, S. A. Salem, A. C. Peck, Z. Yang, R. D. Salem, and E. S. Sills. On-site array CGH applications in clinical in vitro fertilization: Reproductive outcomes and impact on cryopreservation of non-transferred human embryos. *Journal of Biomolecular Techniques : JBT*, 24(Suppl):S11, May 2013. ISSN 1524-0215. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3635430/. Cited on page 16.

D. Lunn, D. Spiegelhalter, A. Thomas, and N. Best. The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, 28 (25):3049–3067, Nov. 2009. ISSN 1097-0258. doi: 10.1002/sim.3680. URL http://onlinelibrary.wiley.com/doi/10.1002/sim.3680/abstract. Cited on page 37.

J. R. Lupski. Genomic disorders ten years on. *Genome Medicine*, 1(4): 42, 2009. Cited on page 5.

J. Martin Bland and D. Altman. Statistical methods for assesing agreement between two methods of clinical measurement. *The Lancet*, 327(8476):307–310, Feb. 1986. ISSN 0140-6736. doi: 10.1016/S0140-6736(86)90837-8. URL http://www.sciencedirect.com/science/article/pii/S0140673686908378. Cited on pages 25 and 26.

H. C. Mefford and E. E. Eichler. Duplication hotspots, rare genomic disorders, and common disease. *Current Opinion in Genetics & Development*, 19(3):196–204, June 2009. Cited on page 101.

H. C. Mefford et al. A method for rapid, targeted CNV genotyping identifies rare variants associated with neurocognitive disease. *Genome Research*, (9):1579 –1585, 2009. Cited on page 88.

L. R. Meyer, A. S. Zweig, A. S. Hinrichs, D. Karolchik, R. M. Kuhn, M. Wong, C. A. Sloan, K. R. Rosenbloom, G. Roe, B. Rhead, B. J. Raney, A. Pohl, V. S. Malladi, C. H. Li, B. T. Lee, K. Learned, V. Kirkup, F. Hsu, S. Heitner, R. A. Harte, M. Haeussler, L. Guruvadoo, M. Goldman, B. M. Giardine, P. A. Fujita, T. R. Dreszer, M. Diekhans, M. S. Cline, H. Clawson, G. P. Barber, D. Haussler, and W. J. Kent. The UCSC genome browser database: extensions and updates 2013. *Nucleic Acids Research*, 41(D1):D64–D69, Jan. 2013. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gks1048. URL http://nar.oxfordjournals.org/content/41/D1/D64. Cited on page 20.

D. T. Miller et al. Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *American Journal of Human Genetics*, 86(5):749–764, May 2010. Cited on page 5.

Mitchell et al. Assessing the significance of conserved genomic aberrations using high resolution genomic microarrays. *PLoS Genet*, 3(8):e143, 2007. Cited on page 88.

A. Moter and U. B. Göbel. Fluorescence in situ hybridization (FISH) for direct visualization of microorganisms. *Journal of Microbiological Methods*, 41(2):85–112, July 2000. ISSN 0167-7012. doi: 10.1016/S0167-7012(00)00152-4. URL http://www.sciencedirect.com/science/article/pii/S0167701200001524. Cited on page 16.

J. G. Mulle, V. C. Patel, S. T. Warren, M. R. Hegde, D. J. Cutler, and M. E. Zwick. Empirical evaluation of oligonucleotide probe selection for DNA microarrays. *PloS One*, 5(3):e9921, 2010. ISSN 1932-6203. doi:

10.1371/journal.pone.0009921. URL http://www.ncbi.nlm.nih.gov/pubmed/20360966. Cited on page 82.

B. Möhlendick, C. Bartenhagen, B. Behrens, E. Honisch, K. Raba, W. T. Knoefel, and N. H. Stoecklein. A robust method to analyze copy number alterations of less than 100 kb in single cells using oligonucleotide array CGH. *PLoS ONE*, 8(6):e67031, June 2013. doi: 10.1371/journal.pone.0067031. URL http://dx.doi.org/10.1371/journal.pone.0067031. Cited on page 55.

P. Neuvial, P. Hupé, I. Brito, S. Liva, . Manié, C. Brennetot, F. Radvanyi, A. Aurias, and E. Barillot. Spatial normalization of array-CGH data. *BMC Bioinformatics*, 7(1):264, May 2006. ISSN 1471-2105. doi: 10.1186/1471-2105-7-264. URL http://www.biomedcentral.com/1471-2105/7/264/abstract. Cited on pages 36 and 37.

G. Nowak et al. A fused lasso latent feature model for analyzing multi-sample aCGH data. *Biostatistics*, June 2011. Cited on page 88.

R. C. O'Hagan et al. Array comparative genome hybridization for tumor classification and gene discovery in mouse models of malignant melanoma. *Cancer Res*, 63:5352–5356, 2003. Cited on page 5.

A. B. Olshen et al. Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics*, 5:557–72, 2004. Cited on pages 19, 22, 27, 29, 30, and 75.

G. H. Perry et al. The fine-scale and complex architecture of human copy-number variation. *American journal of human genetics*, 82:685–95, 2008. Cited on page 5.

F. Picard, S. Robin, E. Lebarbier, and J.-J. Daudin. A segmentation/clustering model for the analysis of array CGH data. *Biometrics*, 63(3):758–766, Sept. 2007. ISSN 1541-0420. doi: 10.1111/j.1541-0420.2006.00729.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1541-0420.2006.00729.x/abstract. Cited on page 37.

F. Picard et al. A statistical approach for array CGH data analysis. *BMC Bioinformatics*, 6(1):27, 2005. Cited on page 88.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL http://www.R-project.org/. Not cited.

B. Rhead et al. The UCSC genome browser database: update 2010. *Nucleic Acids Research*, 38(Database):D613–D619, 2009. Cited on page 96.

J. C. Roach, C. Boysen, K. Wang, and L. Hood. Pairwise end sequencing: a unified approach to genomic mapping and sequencing. *Genomics*, 26 (2):345–353, Mar. 1995. ISSN 0888-7543. doi: 10.1016/0888-7543(95) 80219-C. URL http://www.sciencedirect.com/science/article/pii/088875439580219C. Cited on page 2.

O. M. Rueda and R. Diaz-Uriarte. RJaCGH: Bayesian analysis of aCGH arrays for detecting copy number changes and recurrent regions. *Bioinformatics*, 25(15):1959–1960, Aug. 2009. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btp307. URL http://bioinformatics.oxfordjournals.org/content/25/15/1959. Cited on page 51.

F. Sanger and A. R. Coulson. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, 94(3):441–448, May 1975. ISSN 0022-2836. doi: 10.1016/0022-2836(75)90213-2. URL http://www.sciencedirect.com/science/article/pii/0022283675902132. Cited on page 2.

D. Sarkar. *Lattice: Multivariate Data Visualization with R*. Springer, New York, 2008. URL http://lmdvr.r-forge.r-project.org. ISBN 978-0-387-75968-5. Not cited.

M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470, Oct. 1995. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.270.5235.467. URL http://www.sciencemag.org/content/270/5235/467. Cited on page 15.

A. Sen and M. S. Srivastava. On tests for detecting change in mean. *The Annals of Statistics*, 3(1):98–108, Jan. 1975. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1176343001. URL http://projecteuclid.org/euclid.aos/1176343001. Cited on pages 29 and 30.

S. P. Shah, X. Xuan, R. J. DeLeeuw, M. Khojasteh, W. L. Lam, R. Ng, and K. P. Murphy. Integrating copy number polymorphisms into array CGH analysis using a robust HMM. *Bioinformatics*, 22(14):e431–e439, July 2006. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btl238. URL http://bioinformatics.oxfordjournals.org/content/22/14/e431. Cited on page 36.

C. J. Shaw et al. Comparative genomic hybridisation using a proximal 17p bac/pac array detects rearrangements responsible for four genomic disorders. *J Med Genet*, 41:113–119, 2004. Cited on page 5.

J. R. E. Shepard. Polychromatic microarrays: Simultaneous multicolor array hybridization of eight samples. *Analytical Chemistry*, 78(8): 2478–2486, Apr. 2006. ISSN 0003-2700, 1520-6882. doi: 10.1021/ac060011w. URL http://europepmc.org/abstract/MED/16615753. Cited on page 17.

G. K. Smyth. Limma: linear models for microarray data. In R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, and W. Huber, editors, *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, pages 397–420. Springer, New York, 2005. Not cited.

A. M. Snijders et al. Rare amplicons implicate frequent deregulation of cell fate specification pathways in oral squamous cell carcinoma. *Oncogene*, 24:4232–42, 2005. Cited on page 5.

J. Staaf, G. Jonsson, M. Ringner, and J. Vallon-Christersson. Normalization of array-cgh data: influence of copy number imbalances. *BMC Genomics*, 8:382, 2007. Cited on page 71.

Stan Development Team. *Stan Modeling Language Users Guide and Reference Manual, Version 2.4*, 2014. URL http://mc-stan.org/. Cited on page 37.

P. Stankiewicz and A. L. Beaudet. Use of array CGH in the evaluation of dysmorphology, malformations, developmental delay, and idiopathic mental retardation. *Current Opinion in Genetics & Development*, 17 (3):182–192, June 2007. ISSN 0959-437X. doi: 10.1016/j.gde.2007.04.009. URL http://www.sciencedirect.com/science/article/pii/S0959437X07000743. Cited on page 5.

M. Startek, S. Lasota, M. Sykulski, A. Bulak, A. Gambin, L. Noé, and G. Kucherov. Efficient alternatives to PSI-BLAST. *bulletin of the polish academy of sciences: technical sciences*, 2012. URL http://hal.inria.fr/hal-00749016. Cited on page 13.

W. Stein et al. *Sage Mathematics Software (Version 6.1.1)*. The Sage Development Team. http://www.sagemath.org. Not cited.

M. Sykulski. IKS blog: Using Stanbol to enhance medical data by exploring Linked Data, 2012a. URL http://blog.iks-project.eu/using-stanbol-to-enhance-medical-data-by-exploring-linked-data/. [posted on October 25, 2012]. Cited on pages 13 and 107.

M. Sykulski. Website: IMiD2py – explore aCGH data, 2012b. the project webpage: http://bioputer.mimuw.edu.pl/iks/, the demo webpage: http://bioputer.mimuw.edu.pl:9442/welcome/ [Online; accessed 5-November-2014]. Cited on pages 12 and 107.

M. Sykulski. Videocast: Cytogenetics Lab Stanbol Early Adoption demo, part 1, 2012c. URL https://www.youtube.com/watch?v=Ua6zN5b3w-M. [Online; accessed 5-November-2014]. Cited on pages 12 and 107.

M. Sykulski and T. Gambin. IMID2py - a database and tools for collection and analysis of aCGH data. In *III Convention of the Polish Bioinformatics Society* ptb (2010). URL http://www.ptbi3.polsl.pl/files/Program_PTBi_Convention_and_Workshop_2010_ENG.pdf. Cited on page 12.

M. Sykulski, T. Gambin, M. Bartnik, K. Derwinska, B. Wisniowiecka-Kowalnik, P. Stankiewicz, and A. Gambin. Efficient multiple samples aCGH analysis for rare CNVs detection. In *2011 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 406 –409, Nov. 2011. doi: 10.1109/BIBM.2011.38. Cited on page 10.

M. Sykulski, T. Gambin, M. Bartnik, K. Derwinska, B. Wisniowiecka-Kowalnik, P. Stankiewicz, and A. Gambin. Multiple samples aCGH analysis for rare CNVs detection. *Journal of Clinical Bioinformatics*, 3:12, June 2013. ISSN 2043-9113. doi: 10.1186/2043-9113-3-12. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3691624/. Cited on page 11.

A. L. Tarca, J. E. K. Cooke, and J. Mackay. A robust neural networks approach for spatial and intensity-dependent normalization of cDNA microarray data. *Bioinformatics (Oxford, England)*, 21(11):2674–2683, June 2005. ISSN 1367-4803. doi: 10.1093/bioinformatics/bti397. Cited on page 36.

A. Tewari, M. J. Giering, and A. Raghunathan. Parametric characterization of multimodal distributions with non-gaussian modes. In *2012 IEEE 12th International Conference on Data Mining Workshops*, volume 0, pages 286–292, Los Alamitos, CA, USA, 2011. IEEE Computer Society. ISBN 978-0-7695-4409-0. doi: 10.1109/ICDMW.2011.135. Cited on page 96.

R. Thomas, A. Scott, C. F. Langford, S. P. Fosmire, C. M. Jubala, T. D. Lorentzen, C. Hitte, E. K. Karlsson, E. Kirkness, E. A.

Ostrander, F. Galibert, K. Lindblad-Toh, J. F. Modiano, and M. Breen. Construction of a 2-Mb resolution BAC microarray for CGH analysis of canine tumors. *Genome Research*, 15(12):1831–1837, 2005. Cited on page 5.

M. A. van de Wiel et al. Smoothing waves in array CGH tumor profiles. *Bioinformatics*, 25(9):1099 –1104, May 2009. Cited on pages 88 and 91.

I. B. Van den Veyver, A. Patel, C. A. Shaw, A. N. Pursley, S.-H. L. Kang, M. J. Simovich, P. A. Ward, S. Darilek, A. Johnson, S. E. Neill, W. Bi, L. D. White, C. M. Eng, J. R. Lupski, S. W. Cheung, and A. L. Beaudet. Clinical use of array comparative genomic hybridization (aCGH) for prenatal diagnosis in 300 cases. *Prenatal diagnosis*, 29 (1):29–39, Jan. 2009. ISSN 0197-3851. doi: 10.1002/pd.2127. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3665952/. Cited on page 16.

S. A. F. T. van Hijum, R. J. S. Baerends, A. L. Zomer, H. A. Karsens, V. Martin-Requena, O. Trelles, J. Kok, and O. P. Kuipers. Supervised lowess normalization of comparative genome hybridization data–application to lactococcal strain comparisons. *BMC Bioinformatics*, 9:93, 2008. Cited on page 71.

Y. Wang, F. Makedon, and J. Pearlman. Tumor classification based on dna copy number aberrations determined using snp arrays. *Oncology reports*, 15 Spec no.:1057–9, 2006. Cited on page 5.

J. L. Weber and E. W. Myers. Human whole-genome shotgun sequencing. *Genome Research*, 7(5):401–409, May 1997. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.7.5.401. URL http://genome.cshlp.org/content/7/5/401. Cited on page 2.

Wikipedia. Nucleic acid thermodynamics — wikipedia, the free encyclopedia. URL http://en.wikipedia.org/wiki/Nucleic_acid_thermodynamics. [Online; accessed 1-October-2014]. Cited on page 16.

H. Willenbrock and J. Fridlyand. A comparison study: applying segmentation to array cgh data for downstream analyses. *Bioinformatics*, 21:4084–4091, 2005. Cited on pages 29 and 73.

D. L. Wilson, M. J. Buckley, C. A. Helliwell, and I. W. Wilson. New normalization methods for cDNA microarray data. *Bioinformatics (Oxford, England)*, 19(11):1325–1332, July 2003. ISSN 1367-4803. Cited on page 36.

B. Wiśniowiecka-Kowalnik, M. Kastory-Bronowska, M. Bartnik, K. Derwińska, W. Dymczak-Domini, D. Szumbarska, E. Ziemka, K. Szczałuba, M. Sykulski, T. Gambin, A. Gambin, C. A. Shaw, T. Mazurczak, E. Obersztyn, E. Bocian, and P. Stankiewicz. Application of custom-designed oligonucleotide array CGH in 145 patients with autistic spectrum disorders. *European Journal of Human Genetics*, 21(6):620–625, June 2013. ISSN 1018-4813. doi: 10.1038/ejhg.2012.219. URL http://www.nature.com/ejhg/journal/v21/n6/abs/ejhg2012219a.html. Cited on pages 3, 6, 12, 16, 86, and 111.

Wolfram Research, Inc. Mathematica, Version 7.0, 2008. Not cited.

C. Workman, L. J. Jensen, H. Jarmer, R. Berka, L. Gautier, H. B. Nielser, H.-H. Saxild, C. Nielsen, S. Brunak, and S. Knudsen. A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biology*, 3(9):research0048, Aug. 2002. ISSN 1465-6906. doi: 10.1186/gb-2002-3-9-research0048. URL http://genomebiology.com/2002/3/9/research/0048/abstract. Cited on page 36.

M. Zahurak, G. Parmigiani, W. Yu, R. B. Scharpf, D. Berman, E. Schaeffer, S. Shabbeer, and L. Cope. Pre-processing agilent microarray data. *BMC Bioinformatics*, 8(1):142, May 2007. ISSN 1471-2105. doi: 10.1186/1471-2105-8-142. URL http://www.biomedcentral.com/1471-2105/8/142/abstract. PMID: 17472750. Cited on pages 35 and 91.

J. Zhang et al. Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. *Cytogenetic and Genome Research*, 115(3-4):205–214, 2006. Cited on pages 4, 20, 89, and 97.

Y. Zhang et al. Systematic analysis, comparison, and integration of disease based human genetic association data and mouse genetic phenotypic information. *BMC Medical Genomics*, 3:1, 2010. Cited on pages 4, 20, 89, and 97.

Colophon

This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 11 point Arno Pro, designed by Robert Slimbach in the style of book types from the Aldine Press in Venice, and issued by Adobe in 2007. A template, which can be used to format a PhD thesis with this look and feel, has been released under the permissive MIT (X11) license, and can be found online at github.com/suchow/ or from the author at suchow@post.harvard.edu.

Compilation of this thesis was done with XeLaTeX. Below are listed files and packages used. Newer versions of *mathspec* package are known to cause problems with large formulas, they conflict with package *breqn*.