Computational methods for large-scale data in medical diagnostics

Piotr Dittwald*

extended abstract of the PhD thesis

Bioinformatic analyses in genetics and proteomics

The bottleneck of the large-scale data processing has made bioinformatic analyses a crucial component in life sciences workflows. The two large fields in biomedical studies, whose rapid development in the recent years has strongly depended on computational methods, are genetics and proteomics. They both are strictly linked to each other, e.g. structural organization of the genome affects the variety of proteins in the organism; on the other hand, proteins are the crucial functional molecules that participate in the process of extracting the information encoded in the genome. In this thesis, we present selected bioinformatic methods and discuss their application in basic research as well as in clinical diagnostics.

Methods and results for genome (in)stability analysis

In the first part of the thesis, we focus on recurrent genomic rearrangements. These structural aberrations (in our studies: deletions, duplications, and inversions of chromosomal fragments) are occurring *de novo* at the same genomic loci in different individuals. A portion of the abnormal number of

*supervisors: dr hab. Anna Gambin, dr hab. Paweł Stankiewicz copies of one or more DNA fragments resulting in an imbalance of DNA is referred to as a Copy-Number Variant (CNV). The main mechanism responsible for recurrent rearrangements is nonallelic homologous recombination (NAHR), wherein recombination breakpoints are located within highly similar DNA sequences, e.g. low-copy repeats (LCRs).

LCRs or segmental duplications (SDs) [Bailey et al., 2002] are defined as pairs of DNA fragments with fraction matching (homology score) over 90% and longer than 1 kb in size. It has been shown [Stankiewicz and Lupski, 2002] that for long LCR elements with high homology (originally the parameters were suggested to be 10-400 kb and 97%), the NAHR events might occur within LCRs causing inversions (for inversely oriented LCRs), deletions or reciprocal duplications (for directly oriented LCRs)[1], or reciprocal translocations.

The commonly used molecular biology experimental method called Microarray-based Comparative Genomic Hybridization (aCGH) allows for high-throughput genome-wide data processing in one experiment [Chial, 2008]. The aCGH method enables detection of CNVs as small as tens of kilobases.

Recurrent deletions and reciprocal duplications

In Dittwald et al. [2013c], based on the literature data, we systematically analyzed the genomic regions of genetic diseases and syndromes associated with NAHR-mediated recurrent deletions and reciprocal duplications. Moreover, we queried and cross-referenced large and unique clinical database of high-resolution genomic analyses performed on patients referred for chromosomal microarray analysis (CMA). The applied algorithms using custom scripts allowed us to filter out cases that refer to NAHR-syndrome regions flanked by directly oriented paralogous (i.e. very similar) LCRs (DP-LCRs). The causative association of the patients' rearrangements with the known genetic syndromes involved manual specification of the selected parameters to tackle the issue of different sensitivity of the CMA arrays. As a result, we were able to determine the prevalence of the known recurrent genomic disorders in the clinical CMA database. We also determined the frequencies of the novel rearrangements. To this aim, we narrowed the study to the cases with genomic breakpoints of the investigated CNVs mapped with a sufficient resolution. We used a statistic model of quasi-Poisson regression,

[1] These rearrangements have often prefix micro referring to their sub-microscopic size.

suitable for count data with missing values, to report genomic features that correlate with the frequency of *de novo* recurrent rearrangements. We also investigated several architectural features of the LCR clusters flanking the interrogated regions.

Furthermore, we constructed a new genome-wide map of the DP-LCR-flanked regions in the human genome, i.e. the genomic regions in which recurrent deletions or reciprocal duplications might occur via LCR-mediated NAHR. We also investigated a concept of LCR cluster (determined by a hierarchical clustering algorithm). The clustering approach enabled us to systematically distinguish between overlapping and adjacent regions, and to combine very similar regions. For example, we were able to identify four novel recurrent NAHR-mediated deletions involving chromosome 2q12.2q13, which were previously referred to as a single region. Selected breakpoints of these novel rearrangements were sequenced using wet-bench experiments, and further clinically characterized. Using annotation of gene location and the OMIM database (`http://www.omim.org/`), we not only identified potentially disrupted genes, but also those of them that might cause disease via NAHR, and might be useful in diagnostics.

The schematic representation of this study is depicted as Figure 1.

Genome-wide analyses of potential recurrent inversions

It should be noted that balanced genomic rearrangements (e.g. paricentric or paracentric inversions) are not detectable by the CMA assays. This limitation may be responsible for underestimation of pathogenic recurrent inversions mediated by NAHR. In Dittwald et al. [2013b], our genome-wide computational approach aimed to investigate human genome instability potentially caused by balanced genomic inversions. We identified a set of inversely oriented, paralogous LCRs (IP-LCRs) that can potentially mediate recurrent inversions via NAHR, by integrating the recent version of human genome build, and the criteria from the literature applied for LCRs than can potentially mediate deletions and duplications. Similarly to the previous section, our algorithms utilized efficient operation on intervals to efficiently analyze the genome. The set of IP-LCRs allowed us to estimate the fraction of the human genome where inversion breakpoints might be located, as well as the fraction of genome potentially unstable due to NAHR mediated by IP-LCRs.

The balanced rearrangements may disrupt the genes harboring the re-

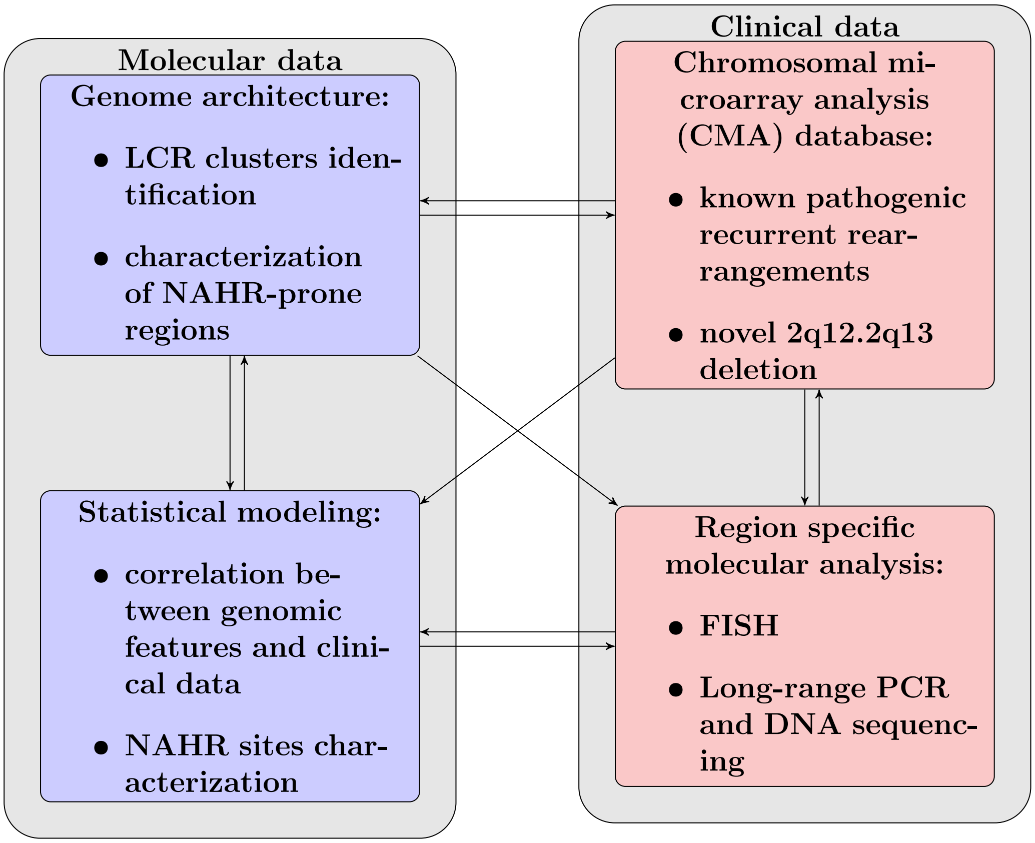
Molecular data
Genome architecture:
LCR clusters identification
characterization of NAHR-prone regions
Statistical modeling:
correlation between genomic features and clinical data
NAHR sites characterization
Clinical data
Chromosomal microarray analysis (CMA) database:
known pathogenic recurrent rearrangements
novel 2q12.2q13 deletion
Region specific molecular analysis:
FISH
Long-range PCR and DNA sequencing

Figure 1: A schematic workflow of the study. The violet and pink colors mark molecular and clinical data, respectively. The arrows indicate the data transfer, which was usually done using automated or semi-automated procedures. Figure courtesy: Dr Anna Gambin.

combination site. Therefore, we reported a set of genes, for which at least one inversion breakpoint is located within the gene, and identified genes that are dosage-sensitive and/or associated with diseases (cf. Figure 2). We also analyzed the X-linked genes, as they have relatively high likelihood of clinically manifesting a disease when disrupted by recurrent inversions in males. Finally, we processed genomic inversions from the Database of Genomic Variants [Zhang et al., 2006] that could be associated with NAHR and estimated the statistical significance of such events.

Methods for proteome analysis

Chemical atoms are built of protons (positively charged), neutrons (not charged), and electrons (negatively charged). Protons and neutrons, also called nucleons, form the nucleus, where the vast majority of the atomic mass is concentrated (therefore the electron mass is omitted in our analyses). Many chemical elements have isotopes[2], i.e. the variants that differ by the amount of neutrons. Here, we will consider only stable isotopes of the five chemical elements building peptides, namely C, H, N, O, and S. The lightest isotope variant is called monoisotopic (in our case these are $^{12}C, ^{1}H, ^{14}N, ^{16}O, ^{32}S$). A mass unit commonly used for chemical molecules is dalton (Da), defined as $\frac{1}{12}$ the mass of carbon ^{12}C, and approximately equal to 1.66×10^{-27} kg. The five considered elements have two (carbons: $^{12}C, ^{13}C$; hydrogens: $^{1}H, ^{2}H$; nitrogens: $^{14}N, ^{15}N$), three (oxygens: $^{16}O, ^{17}O, ^{18}O$), or four (sulphurs: $^{32}S, ^{33}S, ^{34}S, ^{36}S$) isotopic variants. Each of these isotopes has a certain exact mass, denoted as $M_{C_{12}}, \ldots, M_{S_{36}}$, and appears in the nature with a certain probability, denoted as $P_{C_{12}}, \ldots, P_{S_{36}}$.

Mass spectrometry (MS) is one of the most popular analytical method used in proteomics to investigate the content of the chemical mixture, which has already brought a huge portion of insights into the role of biological systems [Cravatt et al., 2007, Chandramouli and Qian, 2009]. The instrumentation used in this method, i.e. mass spectrometer, is composed of the three main parts: (1.) the ionization source – the molecules are charged (i.e. ions are created) and brought to a gas phase; (2.) the mass analyzer – ions are separated by their mass-to-charge (m/z) ratio; (3.) the detector – the

[2] We will consider only stable isotopes, and ignore the radioactive forms which spontaneously undergo the radioactive decay.

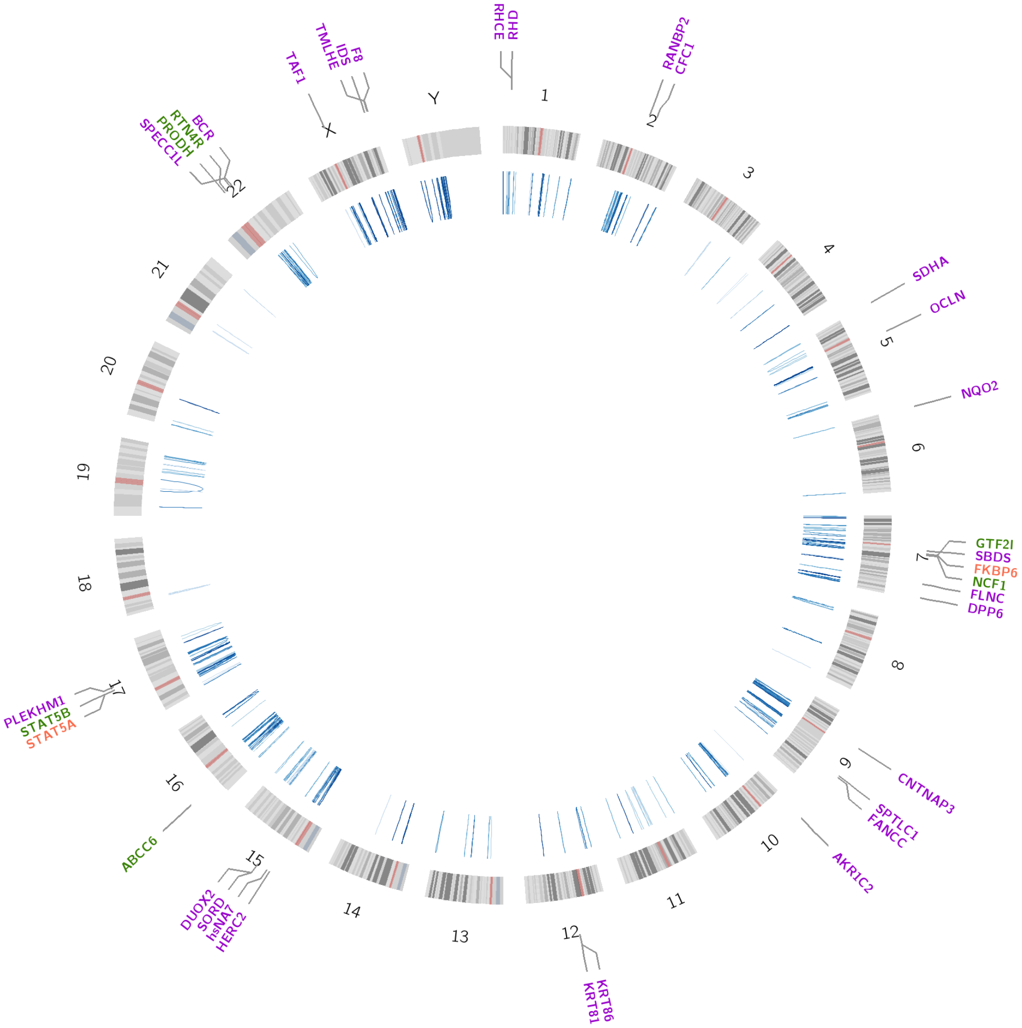

Figure 2: The Circos plot [Krzywinski et al., 2009] depicts the subset of identified genes potentially disrupted by NAHR-mediated inversions genome-wide. We highlighted the genes that are associated with diseases (violet), dosage sensitive (red), and those from both previous groups (green). Figure source: [Dittwald et al., 2013b].

spectrum of signals or peaks is produced, it assigns abundance, i.e. number of ions, for a given m/z.

Isotopic distributions of the molecules

Let us consider the molecule[3] $\xi(v, w, x, y, z)$ of a summaric chemical formula $C_vH_wN_xO_yS_z$, i.e. composed of v carbon, w hydrogen, x nitrogen, y oxygen, and z sulphur atoms. For simplification, we will further omit the parameters v, w, x, y, z, where their presence is obvious from the context.

Analogously to elements, we can also consider isotopic variants of the molecule. Each isotopic variant has its exact mass and a probability, being a sumaric mass and a product of probabilities of occurrence of its atoms, respectively.

The lightest isotopic variant (the one composed purely from the monoisotopic atoms) of the molecule is called a monoisotopic variant. The monoisotopic variant of ξ has an exact mass:

$$M_{mono} = vM_{C_{12}} + wM_{H_1} + xM_{N_{14}} + yM_{O_{16}} + zM_{S_{32}},$$

which is also called a monoisotopic mass of ξ, and a probability:

$$P_{mono} = P_{C_{12}}^{v} \times P_{H_1}^{w} \times P_{N_{14}}^{x} \times P_{O_{16}}^{y} \times P_{S_{32}}^{z}.$$

One can look at the molecule with a different level of accuracy. In a very precise approach, we can consider isotopic fine structure of ξ, where we distinguish between any two isotopic variants as long as they are composed of different number of particular isotopes[4]. However, even for a very small molecules, the number of the fine variants is quite large, and while increasing the number of atoms we can quickly fall into the problem of huge number of configurations that cannot be easily handled. The simplification of the fine approach is to look at the aggregated isotopic variants, where we group together variants with the same number of additional neutrons[5]. Of note, the aggregated variant with zero additional neutrons is always composed of a single fine variant, i.e. the monoisotopic one. The center-mass of aggregated variant is the average mass of all its fine variants.

[3] We will not distinguish between molecules and ions.

[4] We do not distinguish between isoforms, where the order of isotopes matters.

[5] Additional neutrons in comparison to the monoisotopic variant of considered element or molecule.

Results for proteome analysis

Aggregated isotopic variants

Our aim in this part of the analysis is to effectively process the isotopic distribution using the concept of aggregated variant. By q_j we will denote a probability of j-th aggregated isotopic variant of molecule ξ, which can be calculated as:

$$q_j = \sum_k p_{jk} \tag{1}$$

and the center-mass (i.e. expected value) for j-th isotopic variant is defined as:

$$E(m_j) = \bar{m_j} = \frac{\sum_k m_{jk} p_{jk}}{\sum_k p_{jk}}. \tag{2}$$

The m_{jk} and p_{jk} are, respectively, masses and probabilities of the fine variants (indexed by k) with j additional neutrons on comparison to the monoisotopic variant.

In Claesen et al. [2012], we developed the algorithm called BRAIN (**B**affling **R**ecursive **A**lgorithm for **I**sotopic distributio**N** calculations) that is able to compute the aggregated isotope distribution for a given molecule with a formula $C_vH_wN_xO_yS_z$. This algorithm uses two polynomial generating functions. First of these functions, Q, is defined as:

$$\begin{aligned} Q(I; v, w, x, y, z) = \left(P_{C_{12}} I^0 + P_{C_{13}} I^1\right)^v &\quad \times \\ \left(P_{H_1} I^0 + P_{H_2} I^1\right)^w &\quad \times \\ \left(P_{N_{14}} I^0 + P_{N_{15}} I^1\right)^x &\quad \times \\ \left(P_{O_{16}} I^0 + P_{O_{17}} I^1 + P_{O_{18}} I^2\right)^y &\quad \times \\ \left(P_{S_{32}} I^0 + P_{S_{33}} I^1 + P_{S_{34}} I^2 + P_{S_{36}} I^4\right)^z &\quad . \end{aligned}$$

The second function, U, is defined with the usage of the function Q:

$$\begin{aligned} U(I; v, w, x, y, z) = \\ vQ(I; v-1, w, x, y, z) \left(P_{C_{12}} M_{C_{12}} + P_{C_{13}} M_{C_{13}} I^1\right) \\ +wQ(I; v, w-1, x, y, z) \left(P_{H_1} M_{H_1} + P_{H_2} M_{H_2} I^1\right) \\ +xQ(I; v, w, x-1, y, z) \left(P_{N_{14}} M_{N_{14}} + P_{N_{15}} M_{N_{15}} I^1\right) \\ +yQ(I; v, w, x, y-1, z) \left(P_{O_{16}} M_{O_{16}} + P_{O_{17}} M_{O_{17}} I^1 + P_{O_{18}} M_{O_{18}} I^2\right) \\ +zQ(I; v, w, x, y, z-1) \times \\ \left(P_{S_{32}} M_{S_{32}} + P_{S_{33}} M_{S_{33}} I^1 + P_{S_{34}} M_{S_{34}} I^2 + P_{S_{36}} M_{S_{36}} I^4\right) . \end{aligned}$$

The algorithm calculate iteratively the coefficients of both generating functions using the theory of Newton-Girard and Viète's formulae [Séroul, 2000, Vinberg, 2003]. In particular, we obtain the following iterative formula for the probabilities of the aggregated variants:

$$q_j = -\frac{1}{j}\sum_{l=1}^{j} q_{j-l}\psi_l,$$

where ψ_l is a sum of $(-l)$-powers of roots of polynomial $Q(I; v, w, x, y, z)$.

Moreover, in Dittwald et al. [2013a], we implemented BRAIN as a part of R Bioconductor repository [Gentleman et al., 2004] together with a stopping criteria to calculate the substantial part of the isotopic distribution, and applied it in the case study involving batch processing of large protein dataset extracted from the Uniprot database. Namely, we build the linear model predicting the monoisotopic mass based on the corresponding most abundant center-mass. This approach might be potentially useful for experimentalists, who are not able to observe monoisotopic mass for heavy peptides, but would like to use it for molecule identification. We also evaluated the performance of C++ implementation of BRAIN [Hu et al., 2013].

Furthermore, in Dittwald and Valkenborg [2014] we introduced BRAIN 2.0., involving two improvements to decrease both time and memory complexity in obtaining the ratios of the consecutive aggregated isotopic probabilities and a concept to represent the element isotope distribution in a more generic manner than in the original BRAIN.

Finally, we proposed an automatic procedure for discrimination between lipid and peptide signals. The bunch of random forest classifiers is able to distinguish between lipids and peptides based on the features derived from the aggregated isotopic distribution. Moreover, we propose to extend the classification for discriminating between different lipid classes, also using random forest classification. The experiments were tested on lipids and *in silico* digested peptides based on online databases, and a real lipid/peptide mixture analyzed by a mass spectrometer.

Fine isotopic structure

In the next step of the analysis, we tried to characterize the fine structure of aggregated isotopic variants (in practice, we especially looked at the most abundant peaks). We applied a generating function based approach to calculate variance and information theory entropy of mass for the aggregated isotopic variants. More precisely, we first introduced the polynomial:

$$Q^{\perp}(I, J, K; v, w, x, y, z) = \sum_j (\sum_k p_{jk} J^{m_{jk}} K^{m_{jk}}) I^j,$$

and then show that the variance of the aggregated isotopic variants can be obtained by the coefficients of:

$$\frac{\partial^2}{\partial J \partial K} Q^{\perp}(I, J, K; v, w, x, y, z)|_{J=K=1},$$

which is also a polynomial. Also the information theory entropy for the j-th aggregated variant (denoted as $H(j)$) can be calculated using the polynomial generating functions and the following formula:

$$H(j) = \frac{-\sum_k p_{jk} \log(p_{jk})}{\sum_k p_{jk}} + \log(\sum_k p_{jk}).$$

After processing the Uniprot database, we built the linear model for the variance of the most abundant aggregated peak based on its the center-mass. Further, we also estimated the spread of mass distribution of the j-th aggregated variant by:

$$j \cdot (\mu_{2H} - \mu_{15N}).$$

which served for estimating when the overlap between consecutive aggregated peaks occurs.

Articles and manuscripts

The content of the first part of the dissertation, concerning the human instability, is based on the following articles:

- Piotr Dittwald*, Tomasz Gambin* et al (2013). NAHR-mediated copy-number variants in a clinical population: Mechanistic insights into both genomic disorders and Mendelizing traits. (* These authors contributed equally to this work) Genome Research 23, 9: 1395-409,

- Piotr Dittwald*, Tomasz Gambin*, Claudia Gonzaga-Jauregui*, Claudia M.B. Carvalho, James R. Lupski, Pawe Stankiewicz, Anna Gambin (2013). Inverted low-copy repeats and genome instability a genome-wide approach, Human Mutation, 34, 1: 210-20. (*contributed equally).

The content of the second part of the dissertation, concerning the modeling in proteomics, is based on the following articles:

- Jürgen Claesen*, Piotr Dittwald*, Dirk Valkenborg, Tomasz Burzykowski (2012). An efficient method to calculate the aggregated isotopic distribution and exact center-masses, Journal of the American Society for Mass Spectrometry;23(4): 753-63. (* contributed equally)

- Piotr Dittwald, Jürgen Claesen, Tomasz Burzykowski, Dirk Valkenborg, Anna Gambin (2013). BRAIN: a universal tool for high-throughput calculations of the isotopic distribution for mass spectrometry. Analytical Chemistry, 85, 4: 1991-4

- Piotr Dittwald, Dirk Valkenborg (2014). BRAIN 2.0: Time and Memory Complexity Improvements in the Algorithm for Calculating the Isotope Distribution. Journal of the American Society for Mass Spectrometry, 25(4): 588-94,

- Han Hu*, Piotr Dittwald*, Joseph Zaia, Dirk Valkenborg (2013). Comment on "Computation of isotopic peak center-mass distribution by Fourier Transform" (*contributed equally). Analytical Chemistry, 85(24): 12189-92.

and the manuscripts in preparation:

- Piotr Dittwald, Vu Trung Nghia, Glenn A. Harris, Richard M. Caprioli, Raf Van de Plas, Kris Laukens, Anna Gambin, Dirk Valkenborg, Towards automated discrimination of lipids versus peptides from full scan mass spectra,

- Piotr Dittwald, Jürgen Claesen, Dirk Valkenborg, Alan L. Rockwood, Anna Gambin, On isotopic fine structure distribution and limits to resolution in mass spectrometry.

References

J. A. Bailey, Z. Gu, R. A. Clark, K. Reinert, R. V. Samonte, S. Schwartz, M. D. Adams, E. W. Myers, P. W. Li, and E. E. Eichler. Recent segmental duplications in the human genome. *Science*, 297(5583):1003–1007, 2002.

K. Chandramouli and P. Y. Qian. Proteomics: challenges, techniques and possibilities to overcome biological sample complexity. *Human Genomics and Proteomics*, 2009, 2009.

H. Chial. Cytogenetic methods and disease: Flow cytometry, CGH, and FISH. *Nature Education*, 1(1), 2008.

J. Claesen, P. Dittwald, T. Burzykowski, and D. Valkenborg. An efficient method to calculate the aggregated isotopic distribution and exact center-masses. *Journal of the American Society for Mass Spectrometry*, 23:753–763, 2012.

B. F. Cravatt, G. M. Simon, and J. R. Yates. The biological impact of mass-spectrometry-based proteomics. *Nature*, 450(7172):991–1000, 2007.

P. Dittwald and D. Valkenborg. BRAIN 2.0: Time and Memory Complexity Improvements in the Algorithm for Calculating the Isotope Distribution. *Journal of the American Society for Mass Spectrometry*, 25(4):588–594, 2014.

P. Dittwald, J. Claesen, T. Burzykowski, D. Valkenborg, and A. Gambin. BRAIN: a universal tool for high-throughput calculations of the isotopic distribution for mass spectrometry. *Analytical Chemistry*, 85:1991–1994, 2013a.

P. Dittwald, T. Gambin, C. Gonzaga-Jauregui, C. M. Carvalho, J. R. Lupski, P. Stankiewicz, and A. Gambin. Inverted low-copy repeats and genome instability–a genome-wide analysis. *Human Mutation*, 34(1):210–220, 2013b.

P. Dittwald, T. Gambin, P. Szafranski, J. Li, S. Amato, M. Y. Divon, L. X. Rodriguez Rojas, L. E. Elton, D. A. Scott, C. P. Schaaf, W. Torres-Martinez, A. K. Stevens, J. A. Rosenfeld, S. Agadi, D. Francis, S. H. Kang, A. Breman, S. R. Lalani, C. A. Bacino, W. Bi, A. Milosavljevic,

A. L. Beaudet, A. Patel, C. A. Shaw, J. R. Lupski, A. Gambin, S. W. Cheung, and P. Stankiewicz. NAHR-mediated copy-number variants in a clinical population: mechanistic insights into both genomic disorders and Mendelizing traits. *Genome Research*, 23(9):1395–1409, 2013c.

R. C. Gentleman, J. V. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang, and J. Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80, 2004.

H. Hu, P. Dittwald, J. Zaia, and D. Valkenborg. Comment on the computation of isotopic peak center-mass distribution by fourier transform. *Analytical Chemistry*, 85:12189–12192, 2013.

M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. A. Marra. Circos: an information aesthetic for comparative genomics. *Genome Research*, 19(9):1639–1645, 2009.

R. Séroul. *Programming for Mathematicians.* Berlin: Springer-Verlag, 2000.

P. Stankiewicz and J. R. Lupski. Genome architecture, rearrangements and genomic disorders. *Trends in Genetics*, 18(2):74–82, 2002.

E. B. Vinberg. *A course in algebra.* American Mathematical Society, Providence, 2003.

J. Zhang, L. Feuk, G. E. Duggan, R. Khaja, and S. W. Scherer. 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.