TU Delft

Delft University of Technology

Computational analysis of copy number profiles of tumors

Van Dyk, Ewald

DOI
10.4233/uuid:6562195d-ae4c-4150-bfdd-c5282c21954b

Publication date
2019

Document Version
Final published version

Citation (APA)
Van Dyk, E. (2019). *Computational analysis of copy number profiles of tumors*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:6562195d-ae4c-4150-bfdd-c5282c21954b

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

COMPUTATIONAL ANALYSIS OF COPY NUMBER PROFILES OF TUMORS

COMPUTATIONAL ANALYSIS OF COPY NUMBER PROFILES OF TUMORS

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus, Prof. dr. ir. T.H.J.J. van der Hagen,
chair of the Board for Doctorates
to be defended publicly on
Wednesday 9 January 2019 at 12:30 o'clock

by

Hendrik Oostewald VAN DYK

Master of Engineering in Computer Engineering, North-West University, South Africa,
born in Pretoria, South Africa

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus,	chairperson
Prof.dr. L.F.A. Wessels	Delft University of Technology, promotor
Prof.dr.ir. M.J.T. Reinders	Delft University of Technology, promotor

Independent members:

Prof.dr. J.J. Goeman,	Leiden University Medical Center
Prof.dr. R.C.H.J. van Ham	Delft University of Technology
Prof.dr. J. Jonkers	Leiden University
Prof.dr. M. van de Wiel	Free University Amsterdam
Dr.ir. P.D. Moerland	University of Amsterdam
Prof.dr. A. Hanjalic	Delft University of Technology, reserve member

Keywords: copy number profile, segmentation, recurrent aberrations, recurrent copy number breaks, oncogene, tumor suppressor, driver gene, scale space, Euler characteristic

Printed by: Ipskamp Printing

Cover by: Ewald van Dyk

Copyright © 2018 by Hendrik Oostewald van Dyk

ISBN 978-94-6384-003-3

An electronic copy of this dissertation is available at
http://repository.tudelft.nl/

CONTENTS

1

INTRODUCTION

Cancer is a collection of diseases that occurs when cells are altered in such a way as to stimulate undesirable cell-proliferation. Altered pathways affect factors such as the evasion of growth suppressors, avoidance of apoptosis and stimulation of growth factors [1]. Important DNA repair mechanisms are also deactivated leading to genomic instability [2]. In this way, a heterogeneous collection of replicating cells undergo an evolutionary process, all trying to outcompete each other in a hostile environment comprised of, amongst other the immune system and other extracellular factors [3]. Tumors that develop are often heterogeneous where no single clone fully outcompetes others and through it's interaction with the tumor microenvironment orchestrates events that eventually leads to metastasis [4]. With that said, recent data does suggest that metastatic dissemination can occur early in the formation of tumors [5]. Nevertheless, the heterogeneous nature of tumors partially explains why cancer is so difficult to cure [6]. An effective single drug treatment, for example, might kill off 99% of cancer cells only to leave 1% resistant that eventually replicate and render the drug useless.

Oncogenes are key players in the development of cancer. Oncogenes are genes which, when activated, are causally linked to the development and progression of cancer [7]. There are many types of oncogenes. They can, for example, be mutated and over expressed growth factor receptors such as ERBB2 in breast cancer and EGFR in glioblastoma [8, 9]. Other oncogenes such as MYC code transcription factors that have widespread consequences for the expression of downstream genes related to cell proliferation and growth [10]. The RAS family of genes are examples of oncogenes whose overexpression directly effects signalling pathways related to proliferation, growth and survival [11]. Yet another example is hTERT which codes for a catalytic subunit of the enzyme telomerase that maintains telomere stability in almost all cancer cells [12].

Another class of cancer genes are tumor suppressors. These are genes that usually protect the cell from becoming cancerous and can be grouped into three categories: caretaker, gatekeeper and landscaper genes [13–15]. In contrast to oncogenes, these genes are usually inactivated by, for example, point mutations or copy number deletions, typically resulting in lower protein expression in cancer cells. A class of caretaker genes

code for products that are crucial for maintaining mutational and chromosomal stability. BRCA1 is an example typically associated with breast cancer and plays a key role in error free repair of DNA double-strand breaks [16]. In contrast, gatekeeper genes produce products that inhibit cell proliferation. Examples include RB1 [17] and CDKN1B [18, 19] that both inhibit the cell cycle. Finally, the landscaper genes do not directly control cellular growth, but encode for gene products that inhibit the local microenvironments from becoming conductive to unregulated cell proliferation. Examples include PTEN [20] and SMAD5 [21]. Tumor suppressor genes can belong to more than one of these categories. Probably the most important tumor suppressor gene (in cancer) is TP53 [22], who's product, p53, is known as 'the guardian of the genome' and is crucial for preventing germline mutations. On the other hand, p53 also halts the cell cycle when DNA is damaged and triggers apoptosis in severe cases [23]. With these considerations, it is clear that TP53 behaves as both a caretaker and gatekeeper.

We refer to oncogenes and tumor suppressors collectively as cancer genes. Multiple cancer genes need to be activated (inactivated) before cancer can develop, but that does not imply that they are necessarily altered on the genome . There are many epigenetic factors that can influence the expression of cancer genes. BRCA1, for example, is often silenced in breast cancer due to hypermethylation of its promoter region [24]. It is important to note that cancer genes are, by definition, causally linked to cancer. Just because a gene's expression level is significantly correlated with a given cancer type does not automatically imply it is a cancer gene. Over expression of the oncogenic MYC transcription factor will, for example, modulate the expression of genes that are consequently correlated, but not causally linked, to cancer. We refer to genes that are 'along for the ride' as passenger genes. It is therefore very difficult to distinguish cancer genes from passengers when considering gene expression data alone (even across multiple samples).

In contrast to gene expression data, much more can be said about the causal links between genes and cancer when considering genomic alterations. One can for example do a genome wide analysis to see if germline single nucleotide polymorphisms (SNPs) significantly associate with a cancer type relative to a normal population. This type of analysis is called a genome-wide association study (GWAS) [25] and due to the large number of SNP positions considered on the genome and the weak effects associated with germline variants, these methods typically have low statistical power (reveal low true positive rates) in complex diseases such as cancer. One therefore requires a large number of samples (cancer samples and normal samples). With that said, genes containing significant SNP associations are still not necessarily causally related to the cancer type at hand due to a phenomena called linkage disequilibrium [26]. This is the phenomena where SNPs on the genome are non-randomly associated with one another (across individuals irrespective of disease type). Nevertheless, with conditional analysis [27] it is possible to identify candidate genomic loci that likely contain causal hereditary cancer genes.

It is also possible to limit such an analysis to somatic point mutations (acquired during the patient's lifetime) in a tumor when searching for cancer genes. The number of somatic mutations vary greatly between different tumors and cancer types, ranging from less than 300 mutations across the genome in cancer types such as pilocytic astrocytoma and more than 30000 per genome in lung cancers and melanomas [28]. Since these mu-

tations are acquired independently for different patients, linkage disequilibrium is not an issue and it is possible to identify cancer genes that are significantly enriched for mutations across a patient cohort. Furthermore, we expect to find more cancer genes compared to a germline GWAS approach since we directly measure the somatic mutations from the tumor. On the other hand, the sparsity of somatic mutations will necessarily lead to a loss in statistical power. To counter this power loss, it is useful to smooth mutation profiles in a scale space methodology [29]. However this, once again, comes at the expense of pinpointing cancer genes and only identifying (often broad) loci.

Another source of genomic alterations that can be used to detect cancer genes are somatic copy number alterations/aberrations (CNAs). Unlike point mutations, these represent sections of DNA that are deleted or copied mulitple times across the genome. They can vary from focal indels (less than a kilo-base pair long) to chromosome wide gains and losses (also referred to as amplifications and deletions, respectively). Copy number gains of a gene can increase its overall expression and it is therefore possible to identify oncogenes by searching for genes that show frequent copy number gains across different tumor samples. Similarly, tumor suppressors can be identified by detecting genes that are frequently deleted across multiple tumors.

In this thesis we focus only on DNA copy number data and developed computational approaches that pinpoint cancer genes in recurrently aberrated regions.

1.1. MEASURING DNA COPY NUMBER ALTERATIONS

Due to the diploid nature of human DNA, any given stretch of DNA usually has two copies in a normal cell. There will always be germline copy number variations, but these insertions and deletions are typically focal and sparse across the genome of a normal cell. In contrast, depending on the cancer type, highly unstable cancerous cells are riddled with copy number alterations. These events can be broad, with whole chromosome arms being deleted or gained, or focal where only a small section of less than 1 kilo base pair (kbp) is deleted or amplified. In copy number analysis the goal is to determine the number of copies available for any given stretch of DNA in a cancerous cell. This is typically achieved by measuring the quantity of a specific genomic sequence across the genome using short sequences (also called molecular probes) that are synthesised to match specific sequences on the genome. There are various platforms available for performing these tests including fluorescent in situ hybridization, comparative genomic hybridization and high resolution array-based tests on array comparative genomic hybridization (aCGH) [30, 31]. The SNP array technology is an example of a high resolution platform that uses both copy number probes and SNP probes which make it possible to determine the copy number state of both alleles on the genome [32]. Copy number measurements and allele counts can also be derived from DNA whole genome sequencing (WGS) and whole exome sequencing (WES) data. Probe measurements are usually presented as an intensity log ratio between the cancer sample and a normal reference (often determined on the blood of the same patient). By placing the probe measurements at their appropriate positions on the genome, one obtains a copy number profile that can be regarded as the sum of a piecewise constant function (a continues stretch with constant copy number) and additive measurement noise (Fig. 1.1).

A large number of segmentation algorithms exist that are specifically designed to

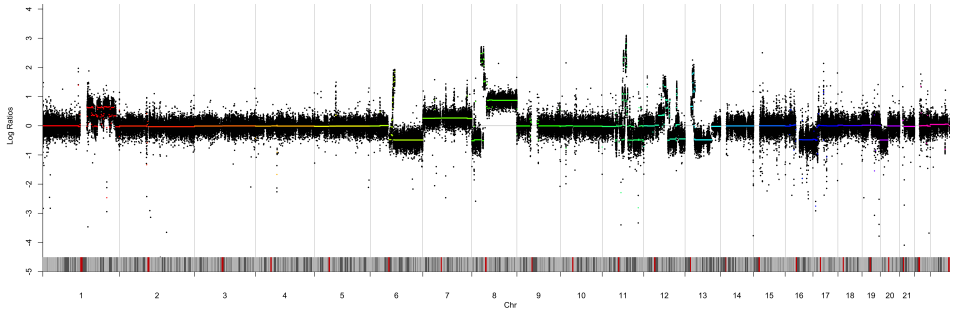

Figure 1.1: Illustration of a DNA copy copy number profile for a breast cancer sample. Black dots indicate raw unsegmented copy number intensity measures. The (coloured) piecewise constant profile is obtained after applying a segmentation algorithm.

filter out measurement noise from the copy number profile. This is achieved by identifying breakpoints in the piecewise constant copy number signal and estimating the copy number between adjacent breakpoints. One might expect the segment values to represent integer values as pieces of DNA are gained and lost in whole integer values. However, this is often not the case since measurements usually represent an averaged profile across a large pool of heterogeneous cells. Nevertheless it is often assumed that there are two signal components mixed in the population, one representing a dominant clone in the tumor and one representing stromal (normal) DNA. As the stromal component is assumed to have a normal diploid genome the signal originating from the tumor cells will be attenuated. In order to obtain high quality profiles, tumor purities (percentage of cells that represent the tumor) above 70% are typically required. Another complicating factor that makes it hard to derive integer copy number values is the fact that tumor cells could be characterized by abnormal ploidy. The ploidy of a tumor cell indicates the average number of chromosome sets in a cell. This might be significantly different than that observed in the (diploid) normal reference cell and the actual DNA content in the tumor is unknown [33–36]. Advanced segmentation algorithms such as ASCAT [37] and ABSOLUTE [38] estimate tumor ploidy and purities simultaneously and use this to compute integer copy number values.

1.2. IDENTIFYING ONCOGENES AND TUMOR SUPPRESSORS FROM DNA COPY NUMBER PROFILES

Sometimes an amplified or deleted piece of DNA can activate or deactivate an oncogene or tumor suppressor, respectively. It is believed that the majority of amplifications and deletions, however, are passenger events, i.e. they result from genomic instability but do not incur a selective advantage for the tumor cell. For simplicity, we only consider amplifications for detecting oncogenes in this discussion. Similar arguments hold for deletions and tumor suppressors. It is possible to define a null model for somatic passenger aberrations and identify significantly recurrent somatic copy number amplifications across multiple tumors in any given cancer cohort. Somatic passenger aber-

rations occur independently between tumors and if amplifications in a locus occur at a rate much higher than expected, it is reasonable to assume that at least a subset of these aberrations are drivers, i.e. contain an oncogene. There exist a diverse set of algorithms that are specifically designed for detecting recurrent aberrations at specified amplification widths [39–50]. The goal is to call regions that are likely to contain oncogenes with an associated p-value. This has proven to be a very difficult task, since aberrations contributing to a recurrence event can vary greatly in width across tumor samples and might be strongly entangled with passenger events.

A simple approach for detecting regions containing oncogenes is to aggregate (sum) segmented copy number profiles across a cohort of tumors and call local maximum peaks that exceed a specified significance threshold. Many algorithms are too strict in the way they employ these thresholds. They typically perform multiple test correction across the genome to avoid over-calling, but fail to account for the high correlation observed between copy number measurement in close proximity. This correlation occurs naturally due to the segmented nature of copy number data and much statistical power can be gained by accounting for it. The second problem with this approach is that driver aberrations are likely to overlap with passenger amplifications (especially with large sample sizes) resulting in a peak that does not overlap with the oncogene. To be more specific, even though a local maximum peak in the aggregate profile might be above the significance threshold, it might be called too focal (i.e. miss the oncogene) due to one tumor sample containing a focal passenger amplification that happens to overlap with highly recurrent driver aberrations. Therefore one needs to call wider 'peaks' to ensure that recurrent regions contain the oncogene (Fig. 4.1a,g). This is at the expense of pinpointing the oncogene, since wider regions might overlap multiple passenger genes. A good example of an algorithm that attempt to solve this problem by widening peaks is RegBounder in the GISTIC2 pipeline [42]. Yet another challenge to overcome is the fact that multiple peaks might be associated with a broad recurrent region and it is not clear which peak is associated with an oncogene. There are a number of algorithms including JISTIC [51], GISTIC2 [42] and RAIG [50] that identify independent peaks with what are called peel-off algorithms (Fig. 4.1b-f). They iteratively peel away amplified segments from single tumor profiles that overlap with a peak and if a secondary peak remains significant, it is called as well.

Although it is clear that the problem at hand is complex, the overall goal is clear. A powerful algorithm is required to 1) call as many recurrent regions as possible that contain cancer genes, 2) call recurrent regions as focal as possible without missing the cancer genes and 3) call as few false positive regions (without cancer genes) as possible.

1.3. OUTLINE OF THESIS

In this thesis we are mainly concerned with detecting cancer genes from DNA copy number data with high specificity and statistical power. Throughout we employ a powerful statistic called the expected Euler characteristic that is often used in random field theory [52–57]. This statistic helps us to effectively correct for multiple testing where tests in close genomic proximity are highly correlated. Although this test is often employed to estimate accurate family wise error rates (FWER), we extend the theory to control for false discovery rates (FDR), which is often more powerful in exploratory algorithms [58]. We

show that this technique naturally lends itself to our application domain and often provides us with analytical solutions that would otherwise lead to time consuming permutation tests. In Chapter 2 we start with a real world diagnostic application of copy number profiles. Here we use DNA copy number profiles to distinguish between BRCA1-like and sporadic breast cancer tumors. It has been shown that patients with BRCA1-like profiles benefit from platinum-based chemotherapies. Unfortunately, measurement platforms change over time and it is therefore important to ensure that our BRCA1-like classifiers are robust across different platforms. Here we employ a strategy for mapping any platform to a feature space that can be used by a single classifier. We show that prediction performance is concordant between platforms. In Chapter 3, we describe an algorithm called ADMIRE, which employs a scale space methodology for detecting recurrent aberrations at different widths (from focal to chromosome-wide recurrent aberrations). In Chapter 4, we introduce an algorithm called RUBIC which, instead of detecting recurrent aberrations, detects recurrent breakpoints across tumor samples. This method naturally avoids many of the complications discussed earlier and is much better at pinpointing cancer genes in focal loci than ADMIRE or other state-of-the-art algorithms. Nevertheless it places less emphasis than ADMIRE on broad recurrent structures. The theoretical framework for ADMIRE and RUBIC constitute a large part of the thesis and are added as supplements at the end of the respective chapters. In Chapter 5, we apply and extend the theory on the expected Euler characteristic for two dimensional (two channel) data and apply it as a single sample segmentation algorithm on SNP platforms and compare it to state of the art algorithms.

REFERENCES

[1] D. Hanahan and R. A. Weinberg, *Hallmarks of cancer: the next generation,* cell **144**, 646 (2011).

[2] T. Abbas, M. A. Keaton, and A. Dutta, *Genomic instability in cancer,* Cold Spring Harbor perspectives in biology **5**, a012914 (2013).

[3] M. Greaves and C. C. Maley, *Clonal evolution in cancer,* Nature **481**, 306 (2012).

[4] G. Lorusso and C. Rüegg, *The tumor microenvironment and its contribution to tumor evolution toward metastasis,* Histochemistry and cell biology **130**, 1091 (2008).

[5] H. Hosseini, M. M. Obradović, M. Hoffmann, K. L. Harper, M. S. Sosa, M. Werner-Klein, L. K. Nanduri, C. Werno, C. Ehrl, M. Maneck, *et al.*, *Early dissemination seeds metastasis in breast cancer,* Nature (2016).

[6] C. Swanton, *Cancer evolution: the final frontier of precision medicine?* (2014).

[7] C. M. Croce, *Oncogenes and cancer,* New England Journal of Medicine **358**, 502 (2008).

[8] Z. Mitri, T. Constantine, and R. O'Regan, *The her2 receptor in breast cancer: pathophysiology, clinical use, and new advances in therapy,* Chemotherapy research and practice **2012** (2012).

[9] H. Zhang, A. Berezov, Q. Wang, G. Zhang, J. Drebin, R. Murali, and M. I. Greene, *Erbb receptors: from oncogenes to targeted cancer therapies,* The Journal of clinical investigation **117**, 2051 (2007).

[10] C. V. Dang, L. M. Resar, E. Emison, S. Kim, Q. Li, J. E. Prescott, D. Wonsey, and K. Zeller, *Function of the c-myc oncogenic transcription factor,* Experimental cell research **253**, 63 (1999).

[11] J. Downward, *Targeting ras signalling pathways in cancer therapy,* Nature Reviews Cancer **3**, 11 (2003).

[12] T. Sundin and P. Hentosh, *Intertesting association between telomerase, mtor and phytochemicals,* Expert reviews in molecular medicine **14**, e8 (2012).

[13] N. C. Levitt and I. D. Hickson, *Caretaker tumour suppressor genes that defend genome integrity,* Trends in molecular medicine **8**, 179 (2002).

[14] B. Vogelstein and K. W. Kinzler, *Cancer genes and the pathways they control,* Nature medicine **10**, 789 (2004).

[15] F. Michor, Y. Iwasa, and M. A. Nowak, *Dynamics of cancer progression,* Nature reviews cancer **4**, 197 (2004).

[16] K. Yoshida and Y. Miki, *Role of brca1 and brca2 as regulators of dna repair, transcription, and cell cycle in response to dna damage,* Cancer science **95**, 866 (2004).

[17] L. Murphree and W. F. Benedict, *Retinoblastoma: clues to human oncogenesis,* Science **223**, 1028 (1984).

[18] L. Hengst and S. I. Reed, *Translational control of p27kip1 accumulation during the cell cycle,* Science **271**, 1861 (1996).

[19] S. S. Millard, J. S. Yan, H. Nguyen, M. Pagano, H. Kiyokawa, and A. Koff, *Enhanced ribosomal association of p27kip1 mrna is a mechanism contributing to accumulation during growth arrest,* Journal of Biological Chemistry **272**, 7093 (1997).

[20] A. Bronisz, J. Godlewski, J. A. Wallace, A. S. Merchant, M. O. Nowicki, H. Mathsyaraja, R. Srinivasan, A. J. Trimboli, C. K. Martin, F. Li, *et al.*, *Reprogramming of the tumour microenvironment by stromal pten-regulated mir-320,* Nature cell biology **14**, 159 (2012).

[21] B. Bierie and H. L. Moses, *Tumour microenvironment: Tgfβ: the molecular jekyll and hyde of cancer,* Nature Reviews Cancer **6**, 506 (2006).

[22] S. Surget, M. P. Khoury, and J.-C. Bourdon, *Uncovering the role of p53 splice variants in human malignancy: a clinical perspective,* OncoTargets and therapy **7**, 57 (2014).

[23] P. H. Shaw, *The role of p53 in cell cycle regulation,* Pathology-Research and Practice **192**, 669 (1996).

[24] L. Zhang and X. Long, *Association of brca1 promoter methylation with sporadic breast cancers: Evidence from 40 studies,* Scientific reports **5**, 17869 (2015).

[25] T. A. Manolio, *Genomewide association studies and assessment of the risk of disease,* New England Journal of Medicine **363**, 166 (2010).

[26] M. Slatkin, *Linkage disequilibrium - understanding the evolutionary past and mapping the medical future,* Nature Reviews Genetics **9**, 477 (2008).

[27] J. Yang, T. Ferreira, A. P. Morris, S. E. Medland, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. N. Weedon, R. J. Loos, *et al.*, *Conditional and joint multiple-snp analysis of gwas summary statistics identifies additional variants influencing complex traits,* Nature genetics **44**, 369 (2012).

[28] L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, S. A. Aparicio, S. Behjati, A. V. Biankin, G. R. Bignell, N. Bolli, A. Borg, A.-L. Børresen-Dale, *et al.*, *Signatures of mutational processes in human cancer,* Nature **500**, 415 (2013).

[29] J. De Ridder, A. Uren, J. Kool, M. Reinders, and L. Wessels, *Detecting statistically significant common insertion sites in retroviral insertional mutagenesis screens,* PLoS computational biology **2**, e166 (2006).

[30] C. Price, *Fluorescence in situ hybridization,* Blood reviews **7**, 127 (1993).

[31] D. Pinkel and D. G. Albertson, *Comparative genomic hybridization,* Annu. Rev. Genomics Hum. Genet. **6**, 331 (2005).

[32] T. LaFramboise, *Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances,* Nucleic acids research , gkp552 (2009).

[33] F. Mitelman, *Recurrent chromosome aberrations in cancer,* Mutation Research/Reviews in mutation research **462**, 247 (2000).

[34] D. G. Albertson, C. Collins, F. McCormick, and J. W. Gray, *Chromosome aberrations in solid tumors,* Nature genetics **34**, 369 (2003).

[35] Z. Storchova and D. Pellman, *From polyploidy to aneuploidy, genome instability and cancer,* Nature reviews Molecular cell biology **5**, 45 (2004).

[36] Z. Storchova and C. Kuffer, *The consequences of tetraploidy and aneuploidy,* Journal of cell science **121**, 3859 (2008).

[37] P. Van Loo, S. H. Nordgard, O. C. Lingjærde, H. G. Russnes, I. H. Rye, W. Sun, V. J. Weigman, P. Marynen, A. Zetterberg, B. Naume, *et al.*, *Allele-specific copy number analysis of tumors,* Proceedings of the National Academy of Sciences **107**, 16910 (2010).

[38] S. L. Carter, K. Cibulskis, E. Helman, A. McKenna, H. Shen, T. Zack, P. W. Laird, R. C. Onofrio, W. Winckler, B. A. Weir, *et al.*, *Absolute quantification of somatic dna alterations in human cancer,* Nature biotechnology **30**, 413 (2012).

[39] O. M. Rueda and R. Diaz-Uriarte, *Finding recurrent copy number alteration regions: a review of methods,* Current Bioinformatics **5**, 1 (2010).

[40] A. Ben-Dor, D. Lipson, A. Tsalenko, M. Reimers, L. O. Baumbusch, M. T. Barrett, J. N. Weinstein, A.-L. Børresen-Dale, and Z. Yakhini, *Framework for identifying common aberrations in dna copy number data,* in *Research in Computational Molecular Biology* (Springer, 2007) pp. 122–136.

[41] R. Beroukhim, G. Getz, L. Nghiemphu, J. Barretina, T. Hsueh, D. Linhart, I. Vivanco, J. C. Lee, J. H. Huang, S. Alexander, *et al., Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma,* Proceedings of the National Academy of Sciences **104**, 20007 (2007).

[42] C. H. Mermel, S. E. Schumacher, B. Hill, M. L. Meyerson, R. Beroukhim, and G. Getz, *Gistic2. 0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers,* Genome Biol **12**, R41 (2011).

[43] S. J. Diskin, T. Eck, J. Greshock, Y. P. Mosse, T. Naylor, C. J. Stoeckert, B. L. Weber, J. M. Maris, and G. R. Grant, *Stac: A method for testing the significance of dna copy number aberrations across multiple array-cgh experiments,* Genome research **16**, 1149 (2006).

[44] A. Magi, M. Benelli, S. Yoon, F. Roviello, and F. Torricelli, *Detecting common copy number variants in high-throughput sequencing data by using jointslm algorithm,* Nucleic acids research , gkr068 (2011).

[45] S. Morganella, S. M. Pagnotta, and M. Ceccarelli, *Finding recurrent copy number alterations preserving within-sample homogeneity,* Bioinformatics **27**, 2949 (2011).

[46] A. Niida, S. Imoto, T. Shimamura, and S. Miyano, *Statistical model-based testing to evaluate the recurrence of genomic aberrations,* Bioinformatics **28**, i115 (2012).

[47] F. Sanchez-Garcia, P. Villagrasa, J. Matsui, D. Kotliar, V. Castro, U.-D. Akavia, B.-J. Chen, L. Saucedo-Cuevas, R. R. Barrueco, D. Llobet-Navas, *et al., Integration of genomic data enables selective discovery of breast cancer drivers,* Cell **159**, 1461 (2014).

[48] E. van Dyk, M. J. Reinders, and L. F. Wessels, *A scale-space method for detecting recurrent dna copy number changes with analytical false discovery rate control,* Nucleic acids research **41**, e100 (2013).

[49] V. Walter, A. B. Nobel, and F. A. Wright, *Dinamic: a method to identify recurrent dna copy number aberrations in tumors,* Bioinformatics **27**, 678 (2011).

[50] H.-T. Wu, I. Hajirasouliha, and B. J. Raphael, *Detecting independent and recurrent copy number aberrations using interval graphs,* Bioinformatics **30**, i195 (2014).

[51] F. Sanchez-Garcia, U. D. Akavia, E. Mozes, and D. Pe'er, *Jistic: identification of significant targets in cancer,* BMC bioinformatics **11**, 189 (2010).

[52] R. J. Adler, *On generalising the notion of upcrossings to random fields,* Advances in Applied Probability **8**, 789 (1976).

[53] R. J. Adler and A. M. Hasofer, *Level crossings for random fields,* The Annals of Probability **4**, 1 (1976).

[54] K. J. Worsley, *Estimating the number of peaks in a random field using the hadwiger characteristic of excursion sets, with applications to medical images,* The Annals of Statistics **23**, 640 (1995).

[55] K. J. Worsley, S. Marrett, P. Neelin, A. C. Vandal, K. J. Friston, A. C. Evans, *et al., A unified statistical approach for determining significant signals in images of cerebral activation,* Human brain mapping **4**, 58 (1996).

[56] K. Worsley, M. Andermann, T. Koulis, D. MacDonald, and A. Evans, *Detecting changes in nonisotropic images,* Human brain mapping **8**, 98 (1999).

[57] T. Nichols and S. Hayasaka, *Controlling the familywise error rate in functional neuroimaging: a comparative review,* Statistical methods in medical research **12**, 419 (2003).

[58] Y. Benjamini and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing,* J R Stat Soc B **57**, 289 (1995).

2

PLATFORM COMPARISONS FOR IDENTIFICATION OF BREAST CANCERS WITH A BRCA-LIKE COPY NUMBER PROFILE

Ewald van Dyk, Philip C. Schouten, Linde M. Braaf, Lennart Mulder, Esther H. Lips, Jorma J. de Ronde, Laura Holtman, Jelle Wesseling, Michael Hauptmann, Lodewyk F.A. Wessels, Sabine C. Linn & Petra M. Nederlof

This chapter is published in Breast Cancer Research and Treatment. Volume 139. Pages 317–327 [1].

Previously we employed bacterial artificial chromosome (BAC) array Comparative Genomic Hybridization (aCGH) profiles from BRCA1 and -2 mutation carriers and sporadic tumours to construct classifiers that identify tumour samples most likely to harbour BRCA1 and -2 mutations, designated 'BRCA1 and -2 like' tumours, respectively. The classifiers are used in clinical genetics to evaluate unclassified variants and patients for which no good quality germline DNA is available. Furthermore, we have shown that breast cancer patients with BRCA-like tumour aCGH profiles benefit substantially from platinum-based chemotherapy, potentially due to their inability to repair DNA double strand breaks, providing a further important clinical application for the classifiers.

The BAC array technology has been replaced with oligonucleotide arrays. To continue clinical use of existing classifiers we mapped oligonucleotide aCGH data to the BAC domain, such that the oligonucleotide profiles can be employed as in the BAC classifier.

We demonstrate that segmented profiles derived from oligonucleotide aCGH show high correlation with BAC aCGH profiles. Furthermore, we trained a support vector machine (SVM) score to objectify aCGH profile quality. Using the mapped oligonucleotide aCGH data we show equivalence in classification of biologically relevant cases between BAC and oligoucleotide data. Furthermore, the predicted benefit of double strand break inducing chemotherapy due to a homologous recombination defect is retained.

We conclude that oligonucleotide aCGH data can be mapped to and used in the previously developed and validated BAC aCGH classifiers. Our findings suggest it is possible to map copy number data from any other technology in a similar way.

2.1. INTRODUCTION

Inactivating mutations in the BRCA1 and -2 genes cause a large increase in breast and ovarian cancer risk [2]. The proteins encoded by these genes are important in the homologous recombination DNA repair pathway. Homologous recombination (HR) is the only known error-free repair pathway for DNA double strand breaks (DSB). Inactivation of components of this pathway results in genomic instability [3]. Array Comparative Genomic Hybridization (aCGH) can be used to analyze this genomic instability. It was previously reported that BRCA1 and -2 mutation carriers reveal specific tumour aCGH signatures that differ from non-BRCA-mutated tumours [4, 5].

Classifiers were trained on these patterns in BAC aCGH data to identify potential BRCA1/2 mutated patients or evaluate BRCA1/2 variants of unknown significance in the clinical genetics setting. These classifiers can also be used when germline DNA is not available [6, 7]. Not all cases that are classified as BRCA-mutated possess a pathogenic mutation that is currently reported by Dutch clinical genetics centers; therefore we term these profiles BRCA-like. Unidentified mutations or other mechanisms of down-regulation of these genes, such as promoter hypermethylation may also be responsible for BRCA-like aCGH profiles and be a sign of HR deficiency [8]. Accordingly, we found that patients with BRCA-like tumours selected from a phase III clinical trial that randomized patients between a conventional chemotherapy regimen and a DSB-inducing regimen highly benefitted from the latter [9, 10]. Due to technological advances which allow analysis of FFPE samples on oligonucleotide based platforms, bacterial artificial chromosome (BAC) based aCGH profiling has been replaced in many laboratories [11]. De-

spite the variety of manufacturers and platform designs, aCGH platforms are concordant and robust in detecting large aberrations. None of the platforms seems superior [12–16]. However, none of these studies were performed in the setting with existing genomic classifiers.

The aim of this study is to validate the application of previously developed and validated BRCA1 and -2 BAC aCGH classifiers using oligonucleotide aCGH data to allow other labs to use the classifiers. In the clinical setting, classification robustness is imperative to ensure the diagnosis remains the same regardless of the copy number platform employed. We specifically investigate if and how using copy number data generated by a different platform influences this classification.

2.2. METHODS

2.2.1. PATIENTS

Patients in this study have been described in earlier reports [6, 7, 9, 10]. Briefly, we hybridized samples that were previously hybridized to BAC arrays when sufficient DNA was available. Samples were divided into 3 cohorts, BRCA1 (n=20 re-hybridizations) and BRCA2 (n=39 re-hybridizations) mutation carriers, controls which possess no known BRCA1/2 mutations that are routinely tested in the clinical genetics setting (83 re-hybridizations). An independent set consisted of a selection of patients that have been treated in a randomized controlled trial of high dose versus conventional chemotherapy. The conventional regimen consisted of 5 courses of 500 mg/m2 5-fluorouracil, 90 mg/m2 epirubicin and 500 mg/m2 cyclophosphamide (FEC). Patients in the high dose treatment cohort were administered 4 courses of FEC, followed by stem cell harvesting and 1 course of 1600 mg/m2 carboplatin, 480 mg/m2 thiotepa and 6000 mg/m2 cyclophosphamide. We matched 77 BRCA1 or -2-like cases to 77 non-BRCA-like cases (112 re-hybridizations of patients with sufficient DNA available) on age, pathological T stage, number of positive lymph nodes and systemic and surgical treatment. Due to strong correlations it was not possible to match a group of patients that did not significantly differ in ER positivity and Bloom-Richardson grade.

All patients have given written informed consent to be included in the study [17], which was approved by the institutional review committee. According to Dutch law, this implied consent allows for the analysis of residual tissue specimens obtained for diagnostic purposes and anonymized publication of the results. (http://www.federa.org/code-goed-gebruik-van-lichaamsmateriaal-2011).

2.2.2. DNA ISOLATION AND aCGH PROFILES

BAC aCGH profiles were available from previous studies [6, 7, 9, 10]. For the BRCA1/2 and sporadic breast cancer datasets we generated new aCGH profiles on the Nimblegen 135k human aCGH platform for cases that had sufficient DNA available. DNA isolation was performed as reported before [6, 7, 9, 10].

Tumour DNA was labelled with Cy3 and female pooled reference DNA (G1521, Promega) was labelled with Cy5 using the ENZO labelling kit for BAC arrays (ENZ-42670, ENZO Life Sciences). Unincorporated nucleotides were removed with the Qiagen MinElute PCR Purification Kit (28004, Qiagen). Subsequently, tumour and reference DNA

were pooled and pelleted using an Eppendorf Concentrator (5301, Eppendorf). The pellets were resuspended in hybridisation mix (NimbleGen Hybridization Kit, Roche Nimblegen) and the sample loaded on the array. Hybridisation was at 42°C for 40-72 hours (Maui Hybridization System, BioMicro Systems). Slides were washed three times (Roche NimbleGen Wash Buffer Kit) and scanned at 2um double pass using an Agilent High Resolution Microarray Scanner (Scanner model: G2505C, Agilent). The resulting image files were further analyzed using NimbleScan software (Roche Nimblegen). Grids were aligned on the picture manually and per channel pair files generated. The Nimblescan DNACopy algorithm was applied at default settings and the unaveraged DNAcopy text files used for further analyses.

2.2.3. COMPUTATIONAL ANALYSIS

PREPROCESSING

Raw aCGH profiles as generated by Nimblegen software were processed in Matlab or R 2.15.0 [18]. Outliers were removed using the algorithm proposed in Circular Binary Segmentation package employing default settings [19]. Segmentation was performed with the cghseg package [20] using a modified Bayes information criterion to select the number of breakpoints [21].

QUALITY ASSESSMENT

We trained a binary support vector machine (SVM) to score the quality of oligonucleotide aCGH profiles. Training was performed on 138 profiles that were assigned to four quality categories by an expert, based on impression of overall noise and dynamic range of the aberrations. The four categories are “very poor”, “poor”, “good” and “very good”. These labels were later converted to two classes by combining the “very poor” and “poor”, and “good” and “very good” labels, to reach sufficient numbers to train the SVM. The SVM was trained using the PRTools toolbox in Matlab [22]. We modified the classifier to report probabilistic scores ranging from zero (poor) to one (good) with a 0.5 boundary threshold. Features employed include noise variance (we define noise as the difference between the raw aCGH profile and the segmented profile), the dynamic range of segments, the signal (segment) to noise power ratio, skewness of the noise and the mean and median segment sizes. We employed a Gaussian radial basis kernel function and optimized the kernel width and Hinge loss penalty factors using 10-fold cross-validation. An outer loop 10-fold cross-validation was performed to assess the predictive performance of the classifier and an independent test set with 265 samples was employed as a complete independent validation.

MAPPING

We mapped oligonucleotide probes to BAC clones by assigning oligonucleotide probes that overlap with a specific BAC clone to that BAC clone. All log-ratio intensities of oligonucleotide probes that were assigned to a specific BAC were averaged to determine the oligonucleotide derived copy number value for the associated BAC probe. 430 BAC clones that were not covered by any oligo probes were ignored. After performing these steps for all BAC clones, we then treated the resulting oligo-derived aCGH profiles as if they were generated on a BAC platform and applied the BAC classifiers.

OPTIMIZING CLASSIFIERS

We retrained shrunken centroid classifiers [23] optimized on oligonucleotide using the class labels 'BRCA1 mutated or methylated' and 'non-BRCA1 mutated or methylated' [3–5]. We chose the classifier with the lowest cross-validation error rate for further analyses.

ASSESSMENT OF PLATFORM SIMILARITY

We used the R packages ClassDiscovery [24], KCsmart [25], to analyse differences between the original BAC profiles and mapped aCGH profiles. We used the BootstrapClusterTest, with Pearson correlation, complete linkage on 5 clusters, sampled 200 times to assess clustering of BAC and mapped oligonucleotide data.

CLINICAL DATA ANALYSIS

Survival analysis and classification concordance were performed with the R packages survival [26] and epiR [27]. Cohen's Kappa values were calculated for concordance between classification on BAC and oligonucleotide aCGH data. The analyses were performed with the marker by treatment interaction design [28] and reported with REMARK criteria if applicable (Table 2.4), and with recommendations specific for predictive marker studies [29–31]. Univariate Kaplan-Meier curves were generated and compared with log rank statistic. Multivariate Cox models were generated to assess the interaction between double strand break inducing chemotherapy and BRCA1 or -2-like CGH status. The hazard rates of patients with a BRCA-like tumour were compared with patients with a non-BRCA-like tumour in a separate model. Both Cox models were corrected for ER status, tumour size, number of positive lymph nodes, histological grade, and prognostic value of BRCA-like status.

2.3. RESULTS

2.3.1. MAPPING OF OLIGONUCLEOTIDE DATA TO BAC PLATFORM POSITIONS

After preprocessing, we mapped oligonucleotide probes to BAC clones and averaged the probes overlapping with the genomic footprint of one clone (mapped profile). 2847 out of 3277 original BAC clones were covered by oligonucleotide probes.

For the following analyses we use two different profiles, the original BAC aCGH (BAC profile) and the oligo-nucleotide aCGH derived mapped profile (oligo-mapped profile). To test the similarity between the oligo-mapped and original BAC profiles we applied several methodologies. First, we used hierarchical clustering (Pearson correlation, complete linkage) on 112 matched BAC and oligo-mapped profiles (Fig. 2.1). When we cluster unsegmented, log-ratio profiles, partial clustering occurs due to platform differences rather than biological replicates (red and green dendrogram bars). However, this platform bias is reduced considerably if we segment both the oligo-mapped and BAC profiles. We confirmed that we obtain similar clustering results using three additional methods. First, clustering only the log-ratios of the BAC and mapped profile employed in the BRAC1/2 classifiers. Second, by principal components analysis and third, by demonstrating that the fraction of the genome aberrated by the KCsmart algorithm is smaller in matched pair analysis compared to non-matched pair analysis (Fig. 2.6).

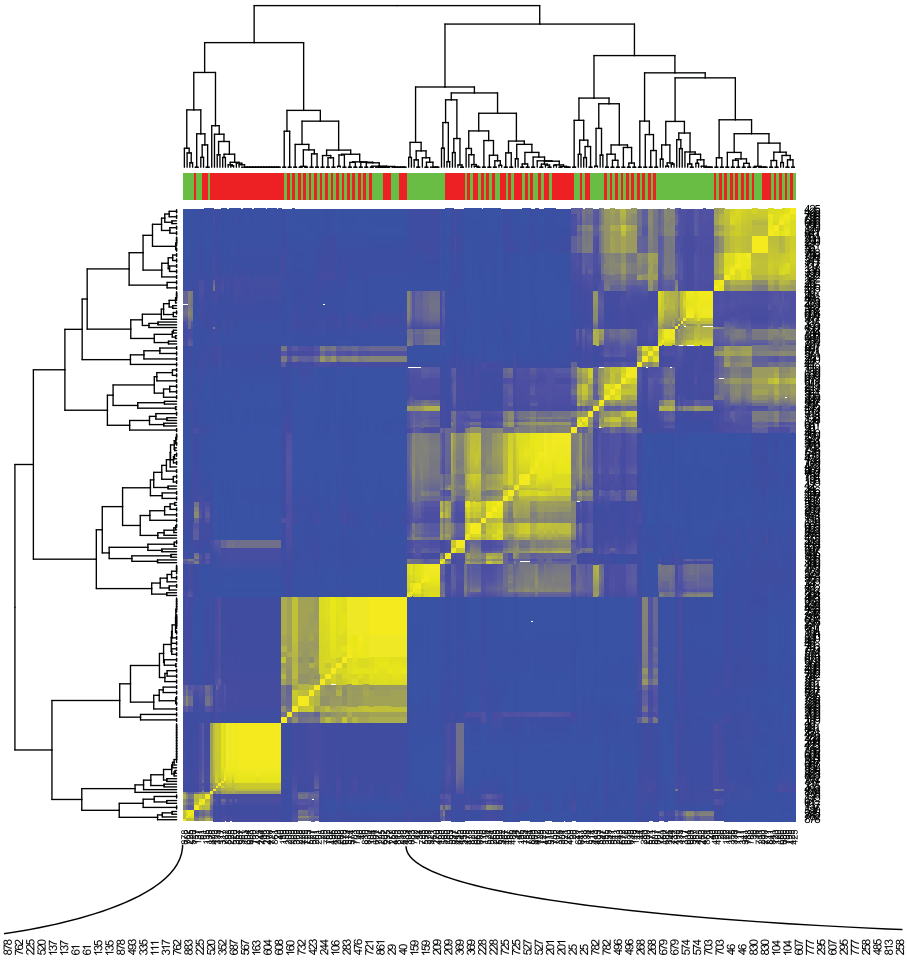

Figure 2.1: Bootstrapped hierarchical clustering of log ratios CGH profiles (**a**). *Red* and *green* column colors represent BAC and oligo-mapped oligonucleotide CGH profiles. Heatmap colors range from 0 (never clustered in same cluster, *blue*) to 1 (always clustered in same cluster, *yellow*). The blowout shows sample names to demonstrate clustering of biological pairs. Clustering segmented data shows less clustering by platform, indicated by *blocks* of *red* and *green* in the columns.

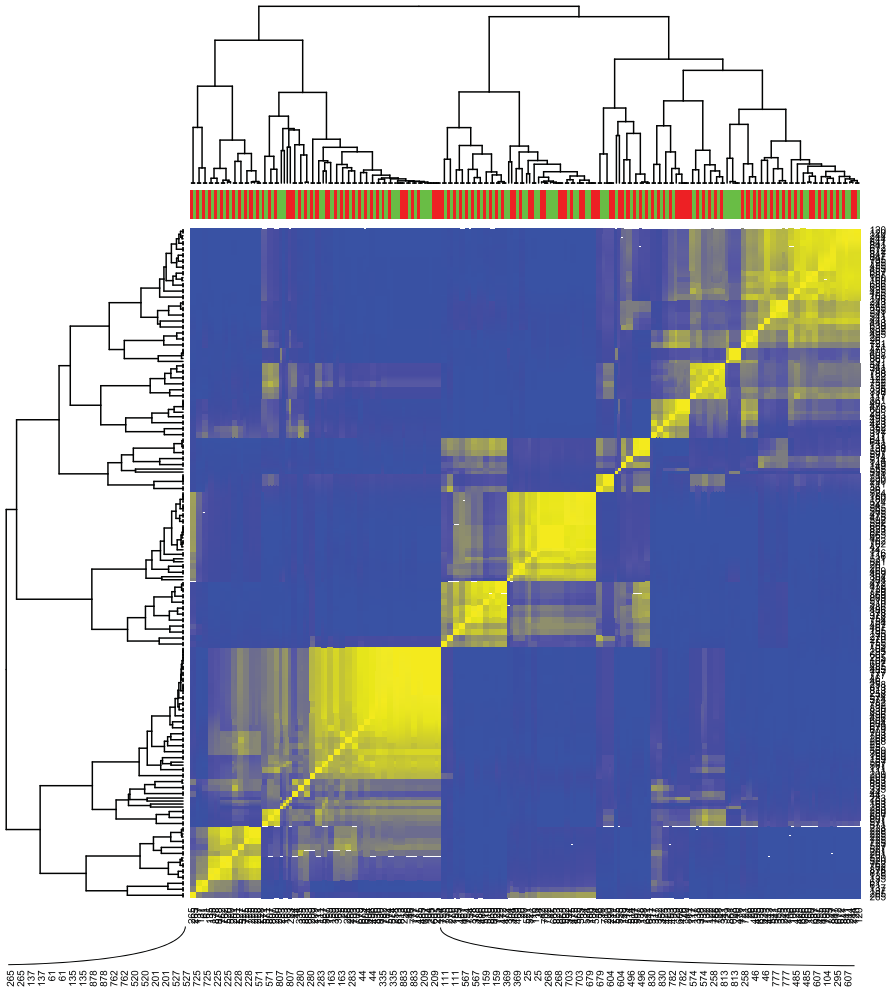

Figure 2.1: Bootstrapped hierarchical clustering of segmented CGH profiles (**b**). *Red* and *green* column colors represent BAC and oligo-mapped oligonucleotide CGH profiles. Heatmap colors range from 0 (never clustered in same cluster, *blue*) to 1 (always clustered in same cluster, *yellow*). The blowout shows sample names to demonstrate clustering of biological pairs. Clustering segmented data shows less clustering by platform, indicated by *blocks* of *red* and *green* in the columns.

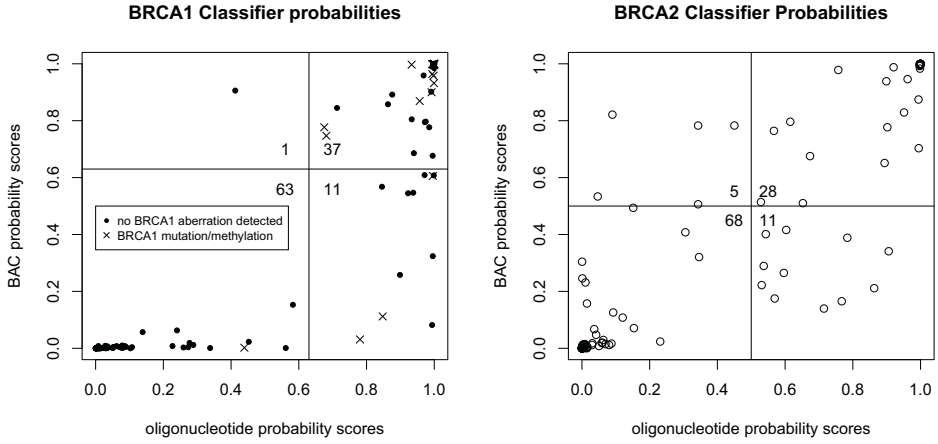

Figure 2.2: Classifier probability scores for the BRCA1 (*left*) and -2 (*right*) classifier. On the x-axis the scores assigned on oligonucleotide aCGH data, on the y-axis the score based on BAC aCGH data. The *horizontal* and *vertical* line represent the cutoff for BRCA-like. *Numbers* represent the number of patients in the quadrant. For BRCA1 mutated and methylated tumours are presented by an ×.

2.3.2. CONCORDANCE OF BAC CLASSIFIER WITH OLIGO-MAPPED DATA AND WITH OPTIMIZED CLASSIFIERS

Probability scores of samples belonging to the BRCA1-like or non-BRCA1-like group can be obtained [23]. This score ranges from 0 (non-BRCA1-like) to 1 (BRCA1-like). The cutoff for assigning to the BRCA1-like group is 0.63 and 0.5 for the BRCA2-like group [7, 9]. We investigated these classifier probability scores assigned to the BAC aCGH data and oligonucleotide aCGH data (Fig. 2.2). The kappa value for inter-observer agreement was 0.78 (95% CI: 0.59-0.96) for the BRCA1 classifier and 0.67 (95% CI: 0.49-0.86) for the BRCA2 classifier. Partly discordant classification is expected, due to the fact that the classifiers were optimized on BAC aCGH data. This stems from the fact that some features had to be removed due to non-overlap between BAC and oligonucleotide aCGH platform in rehybridized samples.

This problem is potentially overcome by training optimized shrunken centroid classifiers on the oligo-nulceotide data to see whether we could improve performance. We did this on unsegmented, segmented and mapped data and compared to classification using BAC data. The results are similar within the confidence interval, although we observe a slight loss in specificity when employing mapped data (Fig. 2.3). As we observe similar performances of the classifiers we will use the mapped data approach for further analysis, since it does not result in a different classifier, only different input data (Fig. 2.2 and 2.7).

2.3.3. DISCORDANT CLASSIFICATION

We investigated whether we could find an obvious bias between BAC and oligo-mapped profiles. Some of the classifier BAC probes could not be mapped as there were no oligonu-

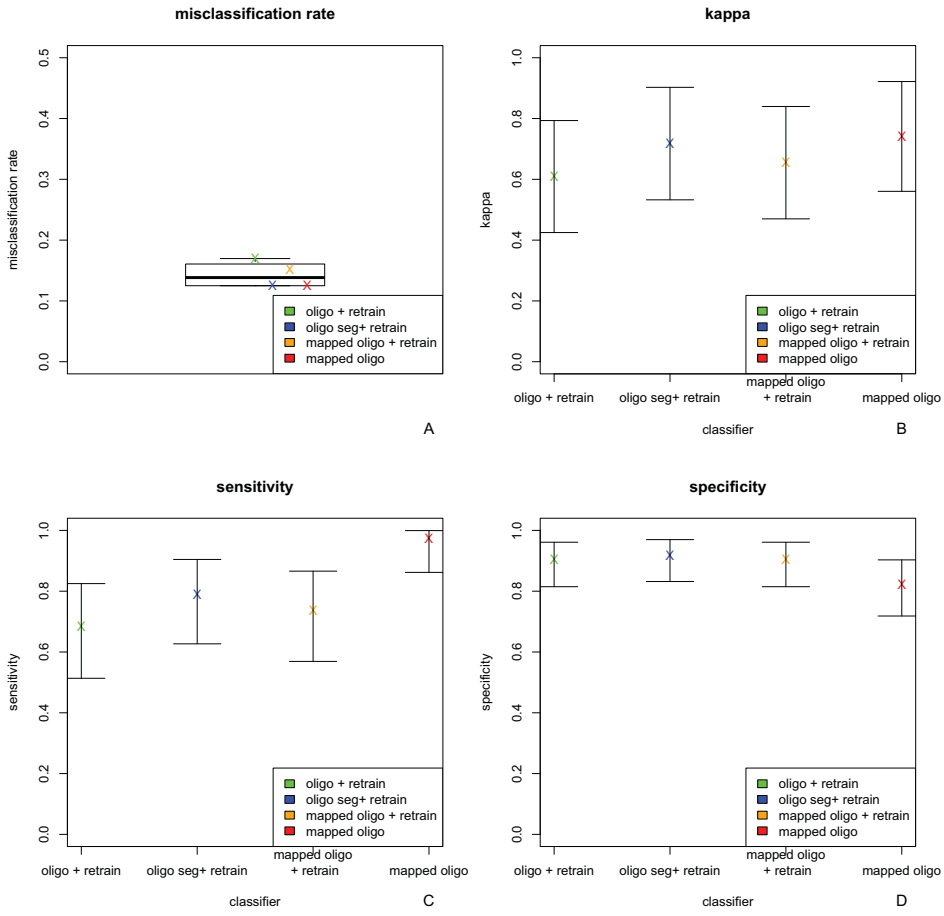

Figure 2.3: Misclassification rates (**a**), kappa (**b**), sensitivity (**c**) and specificity (**d**) with 95% confidence intervals of an optimized classifier trained on unsegmented (oligo + retrain), segmented (segmented + retrain) and mapped (mapped + retrain) oligonucleotide data, and of using the BAC classifier with mapped data (mapped).

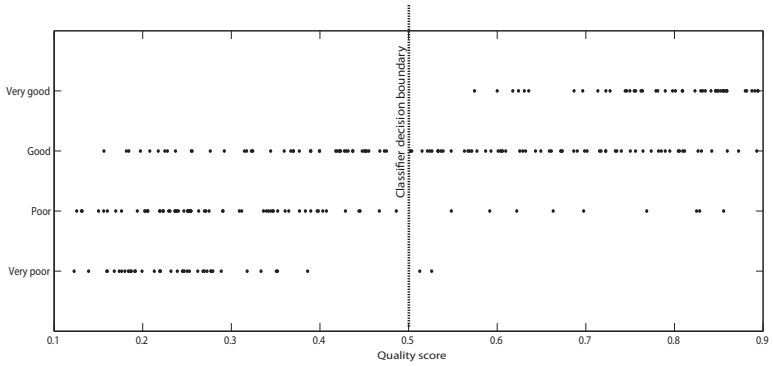

Figure 2.4: Quality scores predicted by the support vector machine for an independent test set consisting of 265 aCGH profiles. The profiles were visually labeled into four quality groups as shown on the y-axis and classified by an SVM classifier depicted on the x-axis.

cleotide probes on those locations, and they were therefore removed from the classifier. When we removed the same probes in the BAC aCGH classifiers, the performance did not differ for the BAC vs oligonucleotide aCGH comparison (Table 2.1). However, some loss of information was observed for the BRCA2 classifier when we compared the original classifier with the classifier with non-mappable probes removed.

Table 2.1: Kappa value of classification with BAC versus oligonucleotide data compared with BAC without non-mappable probes vs. oligonucleotide data.

Classifier	**BAC non-mappable removed vs. mapped oligonucleotide**	**BAC vs. mapped oligonucleotide**	**BAC vs. non-mappable removed BAC**
BRCA1	0.80 (95% CI: 0.61-0.98)	0.78 (95% CI: 0.59-0.96)	1 (95% CI: 0.80-1)
BRCA2	0.64 (95% CI: 0.46-0.83)	0.67 (95% CI: 0.49-0.86)	0.80 (95% CI: 0.62-0.98)

Although we observed high concordance in aCGH profiles, it could be that discordant classification occurs only in the profiles with lower quality. Therefore, we trained a support vector machine classifier on quality labels (independent of BRCA classification) assigned to each profile by an experienced molecular geneticist (PMN). The classifier yields a score between 0 and 1 that represents the profile quality of the experimental data. Cross-validation on the training set yielded an accuracy of 89.9% and an independent test set consisting of 265 profiles (40 very poor, 73 poor, 111 good and 50 very good) yielded an accuracy of 79.6%. The trend between the score predictions and visually assigned labels on the test set is shown in Fig. 2.4.

Classification was not related to aCGH quality score (Fig. 2.8 and 2.9). Combining this result and the high concordance between BAC and oligo-mapped profiles we conclude that despite the quality of FFPE samples in general, both the wet lab experimental method and the classification method are robust. We therefore continue analyses with-

out removing samples.

Although the BAC and oligonucleotide aCGH profiles are very similar we observed that some regions have slightly different amplitudes (Fig. 2.10 and 2.11). These differences are not necessarily restricted to discordantly classified cases. However, discordantly classified cases already have very similar distances to both classifier centroids in the BAC aCGH data. In other words, these samples are not convincingly assigned to a specific class or conversely, the posterior probability of belonging to a specific class is around 0.5. (Fig. 2.12). This and a mix of experimental noise and removed classifier features is likely the cause of discordant classification.

2.3.4. PREDICTION OF CHEMOTHERAPY BENEFIT

We previously reported that patients with a BRCA1 or -2-like profile, as determined by aCGH, derive benefit from double strand break inducing chemotherapy [9, 10]. We tested whether this benefit is still observed when mapped oligonucleotide data is used.

We matched the group of BRCA1 or -2-like tumours to a non-BRCA1 or -2-like group. For 112 out of 154 patients enough DNA was available to generate an oligonucleotide CGH profile. For the analysis of the BRCA1 classifier we excluded patients that scored BRCA2-like to increase the power of the analysis.

Table 2.2 summarizes the patients for whom enough DNA was present and the patients in the analysis of BRCA1 classifier performance.

Kaplan-Meier curves of BRCA1 classification of 112 patients on BAC aCGH profiles and oligonucleotide profiles are shown in Fig. 2.5. Patients that have a non-BRCA-like tumour do not derive significant benefit from a DNA double strand break inducing regimen. Patients with either a BRCA1- or -2-like CGH tumour have an improved survival on DNA double strand break inducing chemotherapy (log rank p = 0.007).

In the multivariate analysis we corrected for ER status, stage, number of positive lymph nodes and tumour grade. Table 2.3 shows the hazard rates for recurrence corrected for these prognostic factors for the BRCA1classifier. The interaction test for a differential effect between patients with the marker and without the marker receiving double strand break inducing chemotherapy was significant for the BRCA1 BAC classifier. On oligonucleotide data the interaction test is 0.227 for the BRCA1 classifier. Although some discordant classification occurred, the classifiers perform similar within the confidence interval. Results for the BRCA2 classifier are similar (data not shown).

2.4. DISCUSSION

We investigated whether we could robustly map data from an oligonucleotide aCGH platform to a BAC platform, such that classification of tumours as BRCA1 or -2-like remains largely unchanged [6, 7, 23].

We chose this approach for several reasons. First, aCGH data is very robust and data generated with different platforms is very comparable [12–16]. However, no data is available for cross-platform shrunken centroid classification on aCGH data. Secondly, we observed that classifiers trained on mutation status predict loss of gene function due to other mechanisms, such as promoter hypermethylation [9]. We lack a gold standard for these BRCA-like aCGH cases without routine diagnostic BRCA1 or -2 mutations. For

Table 2.2: Clinical characteristics of the samples that had DNA available from 230 aCGH profiles in [9]. Column BRCA1 analysis describes the patients in the validation of prediction of double strand break inducing chemotherapy with the BRCA1 classifier, which excludes BRCA2-like aCGH patients. ER: Estrogen receptor, PR: progesterone receptor; HER2 : Human Epidermal growth factor Receptor 2; FEC: 5-fluorouracil, epirubicin, cyclophosphamide; CTC: carboplatin, thiotepa, cyclophosphamide.

	All Patients		BRCA1 analysis	
	BAC non-BRCA	BAC BRCA	oligo non-BRCA	oligo BRCA
ER status				
negative	9	38	6	41
positive	42	23	36	7
PR status				
negative	16	43	12	45
positive	35	16	30	1
missing	0	2	0	2
HER2 status				
negative	51	61	42	48
Pathological T stage				
stage I/II	47	49	38	40
stage III	4	11	4	8
missing	0	1	0	0
No. positive lymph nodes				
< 10	31	38	28	29
>= 10	20	23	14	19
Bloom-Richardson grade				
grade I/II	34	17	29	10
grade III	14	40	10	35
missing	3	4	3	3
BAC BRCA1 status				
non-BRCA	51	23	41	11
BRCA	0	38	1	37
BAC BRCA2 status				
non-BRCA	51	28	40	33
BRCA	0	33	2	15
Chemotherapy regimen				
conventional (5*FEC)	23	31	18	25
DSB inducing (4*FEC+1*CTC)	28	30	24	23
Recurrence, 2nd primary, death				
No	28	27	22	21
Yes	23	34	20	27
Oligo BRCA1 status				
non-BRCA	45	19	42	0
BRCA	6	42	0	48
Oligo BRCA2 status				
non-BRCA	44	29	42	31
BRCA	7	32	0	17

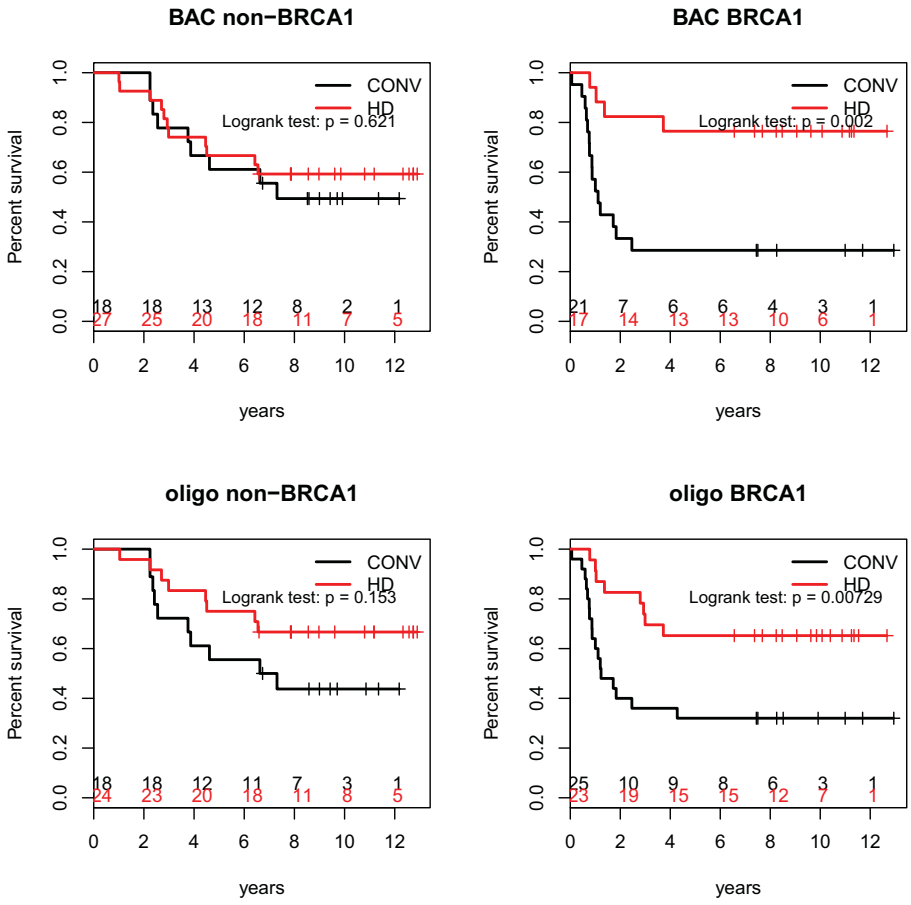

Figure 2.5: Kaplan-Meier curves of recurrence-free survival of non-BRCA1-like CGH patients and BRCA1-like CGH patients classified on BAC data (*top*) and mapped oligonucleotide (*bottom*) data. Patients that are not BRCA1-like do not derive significant benefit from double strand break inducing chemotherapy. Patients that have a BRCA-like CGH tumour derive substantial benefit of the double strand break inducing chemotherapy regimen compared to a conventional FEC regimen.

Table 2.3: Multivariate analysis of recurrence free survival of BRCA1-like CGH tumours. HR=hazard rate, 95% CI = 95% confidence interval, HD=high dose chemotherapy, ~ interaction test: p=0.015; # interaction test: p=0.227.

		BAC			Mapped oligonucleotide		
		HR	95% CI	P	HR	95% CI	P
ER	+ vs. -	0.869	0.337-2.238	0.771	0.941	0.362-2.448	0.901
pT	3 vs. 1 or 2	3.673	1.552-8.695	0.003	3.755	1.604-8.789	0.002
Grade	3 vs. 1 or 2	1.423	0.539-3.755	0.476	1.301	0.502-3.367	0.588
No. positive lymph nodes	>= 10 vs. < 10	2.56	1.274-5.146	0.008	2.576	1.302-5.098	0.007
BRCA1 CGH	+ vs. -	2.675	0.798-8.971	0.111	1.826	0.579-5.760	0.304
BRCA1 CGH & DSBi chemo		0.13	0.037-0.463	0.002 ~	0.311	0.117-0.831	0.02 #
BRCA1 CGH & Conv chemo		1			1		
non-BRCA1 CGH & DSBi chemo		0.874	0.352-2.169	0.772	0.51	0.205-1.270	0.148
non-BRCA1 CGH & Conv chemo		1			1		

clinical applications this group is interesting as these patients benefit from double strand break inducing therapy. Retraining the classifier may result in loss of this group of patients. Third, an important limiting factor is the availability of rare clinical samples. We encountered loss of patients for whom enough DNA was available, demonstrating the need to cherish and to reserve them for new questions rather than repeated quality and technology development steps. Fourth, if it is feasible to use data from other platforms than were used to train the classifiers, many existing datasets can be used to validate our findings.

On the first point, we have demonstrated that it is possible to approximate original data to an extent that unsupervised methods cluster biological replicates rather than platforms on which the data was generated. Robustness of cross-platform comparisons of copy number data has been reported before, provided that the resolution and coverage of the platforms are sufficient [12–16]. In concordance with these reports is our observation that segmented profiles correlate better than non-segmented data. Making use of the high correlation between neighbouring probes and therefore obtaining robust estimates of copy number changes at a certain locus in training classifiers could therefore improve existing shrunken centroid classifiers that were trained on non-segmented data.

Only one study proposes methods for preprocessing aCGH data for cross-platform comparison class discovery. This study was conducted in the setting of new class discov-

ery based on unsupervised analysis [32]. These methods do not assume a pre-existing classifier, however the strategy is similar to ours: mapping to a single format combined with noise reduction, which we do by averaging the noisier oligonucleotide probe measurements located within a BAC clone. We did not apply scaling as the proposed method involves changing log 2 ratios to Z scores. Neither did we apply interpolation to a common genomic grid as the features of the classifier (ie. the chromosomal positions of the BAC clones) are fixed.

On the second point, the lack of gold standard for BRCA-like cases, and optimizing the existing classifiers we refer to the mRNA expression field, in which the use of shrunken centroid classifiers is common. Several methods have been proposed and compared for cross-platform comparisons of gene expression data. However, none of the cross-platform comparison methods allow for normalization without intrinsically changing the measurements or distributions of measurements. Applying such a method would require adjustment of the original classifiers to prevent differential classification due to different preprocessing methodologies. Adjusting the classifiers would lead to similar problems as are currently observed in the gene expression field, where discordance between platforms, classifier centroids or gene list and inter-observer variation prevent robust classification of single tumours [33–35].

In our translation we encountered occurrences of discordant classification. This is expected for several reasons: experimental noise, optimization on BAC aCGH data and loss of information due to missing classifier features. We observed high kappa scores between classification of BAC derived and oligo-mapped aCGH profiles as well as with classification based on optimized classifiers. We demonstrated that optimization results in different classifiers with similar performance. Higher resolution platforms, such as copy number profiling of FFPE samples by next generation sequencing could counter the loss by non-mappable probes [36] . However, the pattern that we capture is already present on lower resolution platforms, as the BAC aCGH [6, 7]. The estimates that we present are conservative, in the sense that they will likely improve when more dense platforms are used to estimate copy number at a certain genomic location. Furthermore, we assessed various sources of biases that could lead to discordant classification but could not discover an obvious bias that would always lead to discordant classification. Discordant classification seems to occur due to experimental noise for aCGH profiles that are approximately equally distanced to the BRCA aCGH and sporadic aCGH signature. Given the high kappa values and the retention of performance in predicting clinical outcome we deem the classifiers robust for predicting outcome of patients that benefit from double strand break inducing chemotherapy likely due to a molecular defect in homologous recombination DNA repair. Furthermore, these classifiers may be used for prediction of benefit of PARP1 inhibitors. Validation in independent cohorts is required for further clinical implementation.

In our attempt to find a source of discordant classification we developed a support vector machine to capture aCGH quality as visually assessed by experienced molecular geneticists. This classifier had good classification performance in an independent validation set and can thus be used to determine aCGH profile quality. Previously, either single variable measures or visual inspection has been used to describe aCGH quality [16, 30]. Single variables do not adequately describe aCGH profiles and various mea-

sures to describe quality, and the SVM can thus be viewed as an addition to objectify aCGH quality assessment.

To conclude, we present a method for mapping between DNA copy number platforms. This method results in highly correlated classification of BRCA1 and -2-like class in a large cohort of patients for which both BAC and oligonucleotide aCGH-based classification data are available. For prediction of chemotherapy benefit the mapped data predicts only slightly worse but with large overlap in the confidence interval. We argue that this method can be used to map data generated on other platforms such as SNP arrays and the likely new platform of choice, next generation sequencing data as well.

2.5. ACKNOWLEDGMENTS

We thank Marieke Vollebergh for discussion and Tesa Severson for critically reading the manuscript.

This study was carried out within the framework of CTMM, the Center for Translational Molecular Medicine (www.ctmm.nl), project Breast CARE grant 030-104, and Life Sciences Center Amsterdam (LSCA) Validation fund.

2.6. CONFLICTS OF INTEREST

SC Linn and PM Nederlof are named inventors on a patent application for the BRCA1 and -2 like array Comparative Genomics Hybridization classifiers.

The other authors do not disclose any conflict of interest.

2.7. SUPPLEMENTAL METHODS

We used the prcomp function from the R base package stats to plot samples mapped to the first two principal components to assess clustering of BAC and mapped oligonucleotide CGH data. We used KCsmart to compare 1000 times 50 paired BAC and mapped oligonucleotide CGH profiles with 50 non-paired profiles at 200 permutations with an FDR of 5%, kernel size of 1mb and sample density of 50kb. We averaged the size of the genome called significantly different between both groups and plotted these.

Table 2.4: Supplementary tables.

REporting recommendations for tumour MARKer prognostic studies (REMARK)	
Introduction	
1. State the marker examined, the study objectives, and any pre-specified hypotheses.	BRCA1 and -2 like classification employing oligonucleotide array CGH data instead of BAC array CGH data on which original classifiers were trained results in similar classification
Materials and Methods	
Patients	
2. Describe the characteristics (e.g. disease stage or comorbidities) of the study patients, including their source and inclusion and exclusion criteria.	Cohorts of BRCA1/2 and non-BRCA1 and -2 mutated patients described in [6, 7]. Patients from Randomized Controlled Trial by [17]. BAC array CGH profiles have been described in [9].
3. Describe treatments received and how chosen (e.g. randomised or rule-based).	Randomized between: 5×FEC vs. 4×FEC+1×CTC
Specimen characteristics	
4. Describe type of biological material used (including control samples), and methods of preservation and storage.	DNA isolated from FFPE archival tissue
Assay methods	
5. Specify the assay method used and provide (or reference) a detailed protocol, including specific reagents or kits used, quality control procedures, reproducibility assessments, quantitation methods, and scoring and reporting protocols. Specify whether and how assays were performed blinded to the study end point.	Array CGH, see methods for precise experimental procedures and quality control
Study design	

6. State the method of case selection, including whether prospective or retrospective and whether stratification or matching (e.g. by stage of disease or age) was employed. Specify the time period from which cases were taken, the end of the follow-up period, and the median follow-up time.	Rehybridization of the SA Joosse cohorts [6, 7], for which enough DNA was available to do array CGH. Rehybridization of matched sample (BRCA like vs non-BRCA like) of the Vollebergh cohort [9]. Matching was done on age, T stage, number of positive lymph nodes, systemic and surgical treatment received. Due to strong correlations between BRCA status and ER status and Bloom Richardson grade these factors could not be matched for, they have been corrected for in multivariate analysis
7. Precisely define all clinical end points examined.	Reported disease free survival; OS results are similar. Control was analysis based on BAC data instead of oligonucleotide data
8. List all candidate variables initially examined or considered for inclusion in models.	Standard prognostic variables: ER, T stage, number of lymph nodes, Bloom Richardson grade
9. Give rationale for sample size; if the study was designed to detect a specified effect size, give the target power and effect size.	All patients with a BRCA like BAC aCGH tumor were included and matched to a control
Statistical analysis methods	
10. Specify all statistical methods, including details of any variable selection procedures and other model-building issues, how model assumptions were verified, and how missing data were handled.	Compare results based on BAC data with results based on oligonucleotide data
11. Clarify how marker values were handled in the analyses; if relevant, describe methods used for cut-point determination.	Existing classifiers were used, with different input data. Cutoff determination has been described in Joosse et al [7] and Vollebergh et al [9]
Results	
Data	
12. Describe the flow of patients through the study, including the number of patients included in each stage of the analysis (a diagram may be helpful) and reasons for dropout. Specifically, both overall and for each subgroup extensively examined report the numbers of patients and the number of events.	154 patients were matched, for 112 patients sufficient amounts of DNA were available for rehybridization (see Table 2.2) 90 patients (exclude patients with missing data and BRCA2 like aCGH tumors) for the comparison between BAC based classification and oligonucleotide based classification
13. Report distributions of basic demographic characteristics (at least age and sex), standard (disease-specific) prognostic variables, and tumour marker, including numbers of missing values.	Table 2.2

Analysis and presentation	
14. Show the relation of the marker to standard prognostic variables.	Table 2.1 (patients were matched, no test performed) and [6, 7, 9]
15. Present univariate analyses showing the relation between the marker and outcome, with the estimated effect (e.g. hazard ratio and survival probability). Preferably provide similar analyses for all other variables being analysed. For the effect of a tumour marker on a time-to-event outcome, a Kaplan-Meier plot is recommended.	Fig. 2.5
16. For key multivariable analyses, report estimated effects (e.g. hazard ratio) with confidence intervals for the marker and, at least for the final model, all other variables in the model.	Table 2.3
17. Among reported results, provide estimated effects with confidence intervals from an analysis in which the marker and standard prognostic variables are included, regardless of their significance.	Table 2.3
18. If done, report results of further investigations, such as checking assumptions, sensitivity analyses, internal validation.	Interaction test between marker and treatment with confidence intervals to assess predictive value. Investigations of biases that could lead to discordant classification (Fig. 2.2, Fig 2.6-2.11), due to data quality issues, specific experimental platform biases and data preprocessing.
Discussion	
19. Interpret the results in the context of the pre-specified hypotheses and other relevant studies; include a discussion of limitations of the study.	Within the confidence interval classification retains predictive value for mutation status and benefit of therapy. Lack of a golden standard for BRCAness leads to conclusion that classifiers act very similar and as expected a bit worse due to optimalization on BAC data. See further discussion
20. Discuss implications for future research and clinical value.	See discussion

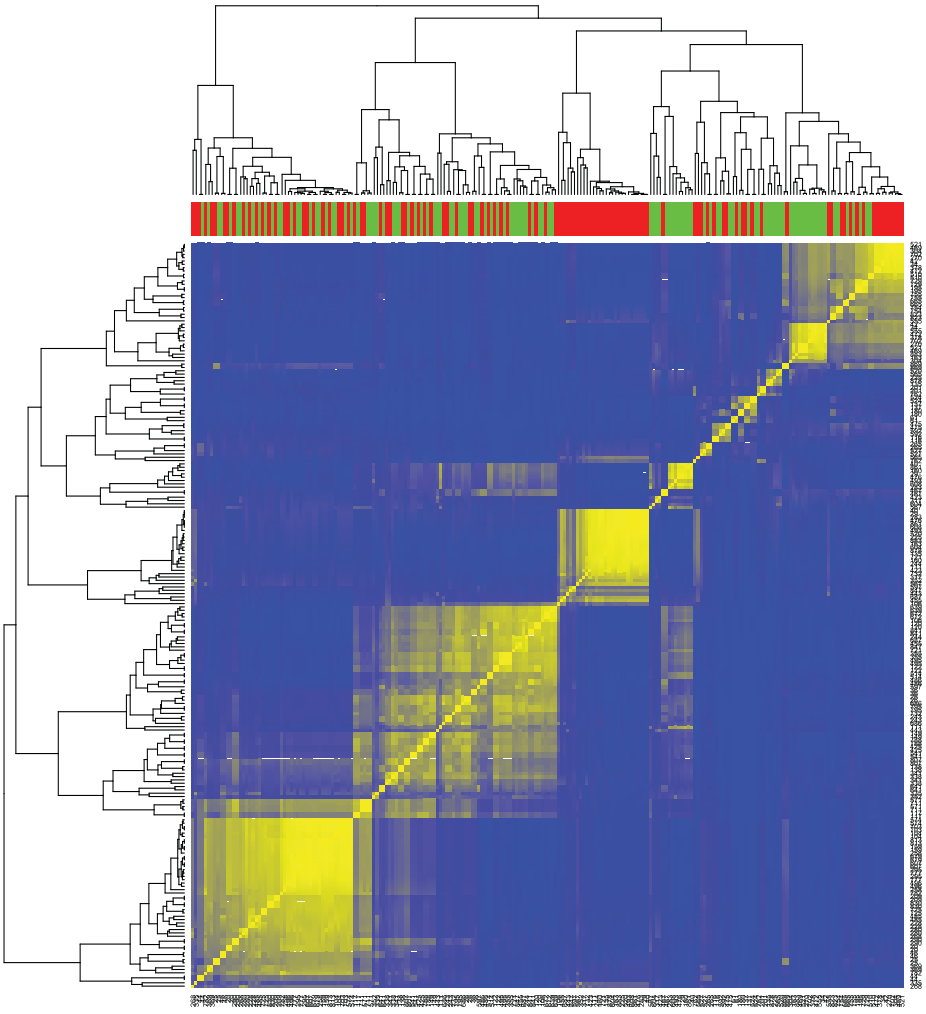

Figure 2.6: (**a**) Bootstrapped hierarchical clustering of log ratios CGH profiles. Only ratios present in the BRCA1 classifier are used. *Red* and *green* column colors represent BAC and oligo-mapped oligonucleotide CGH profiles. Heatmap colors range from 0 (never clustered in same cluster, *blue*) to 1 (always clustered in same cluster, *yellow*). The blowout shows sample names to demonstrate clustering of biological pairs. Clustering segmented data shows less clustering by platform, indicated by blocks of *red* and *green* in the columns.

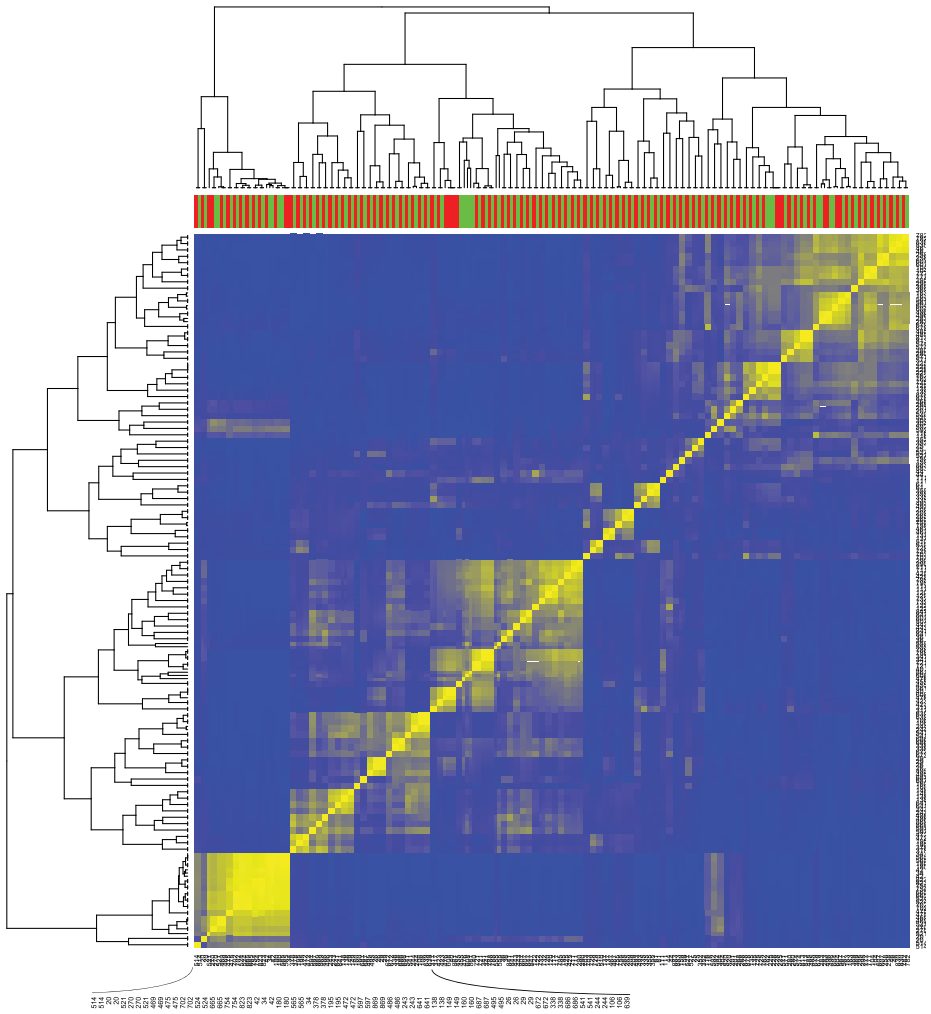

Figure 2.6: (**b**) Bootstrapped hierarchical clustering of segmented log ratios CGH profiles. Only ratios present in the BRCA2 classifier are used. *Red* and *green* column colors represent BAC and oligo-mapped oligonucleotide CGH profiles. Heatmap colors range from 0 (never clustered in same cluster, *blue*) to 1 (always clustered in same cluster, *yellow*). The blowout shows sample names to demonstrate clustering of biological pairs. Clustering segmented data shows less clustering by platform, indicated by blocks of *red* and *green* in the columns.

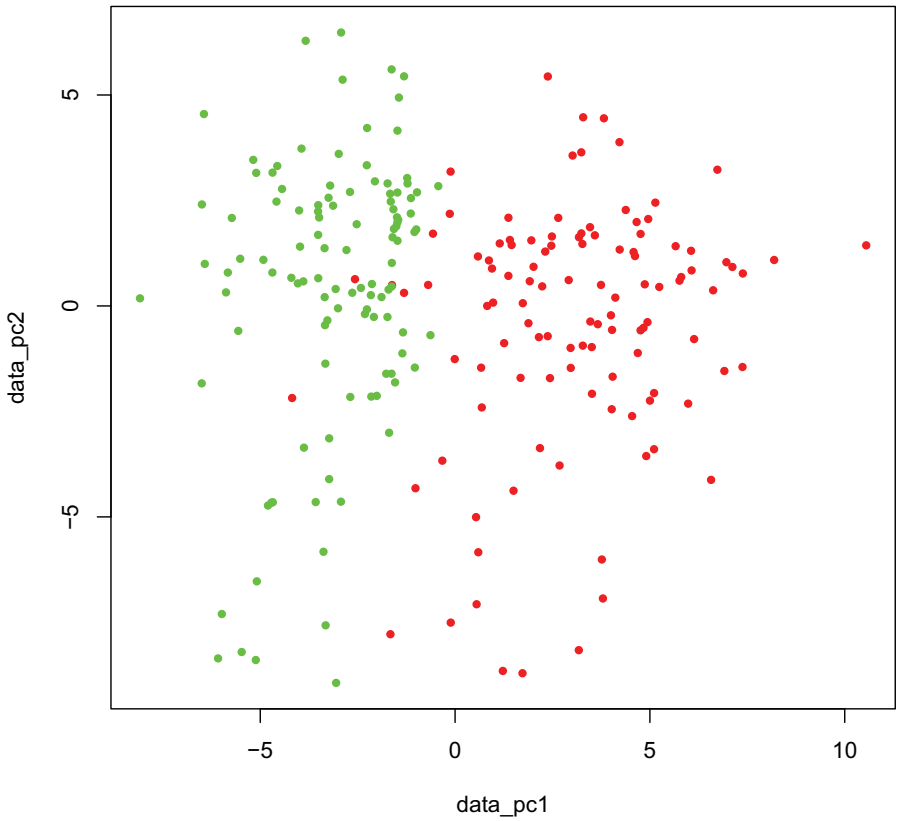

Figure 2.6: (**c**) Principal component analysis of log ratios and BAC (*red*) and mapped (*green*) profiles.

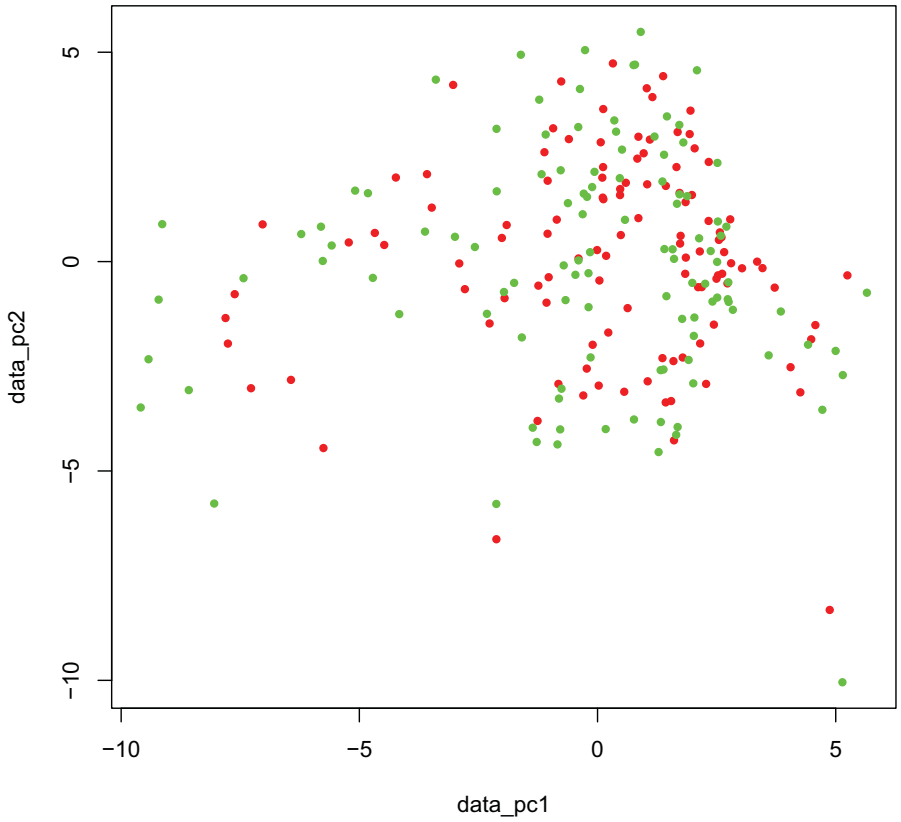

Figure 2.6: (**d**) Principal component analysis of segmented log ratios and BAC (*red*) and mapped (*green*) profiles.

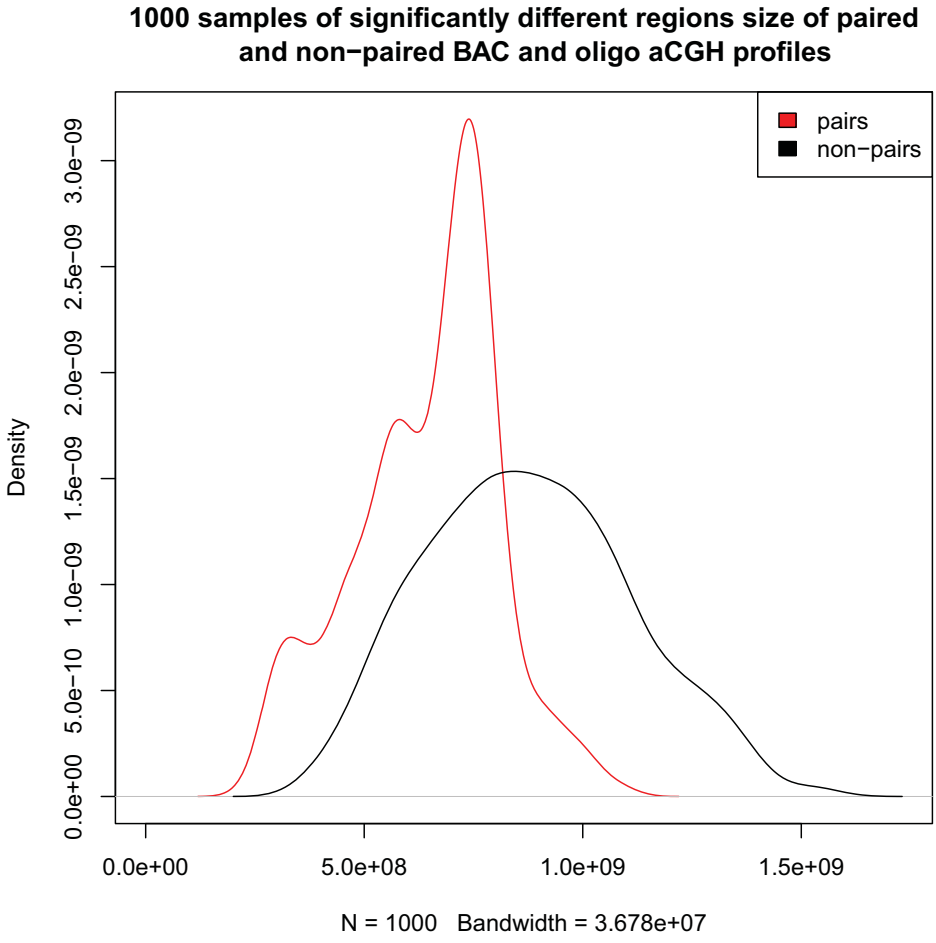

Figure 2.6: (**e**) Density plot of KCsmart analysis of 50 paired BAC-mapped and non-paired log ratio samples sampled 1000 times. Size of genome called aberrantly was summed for each sample.

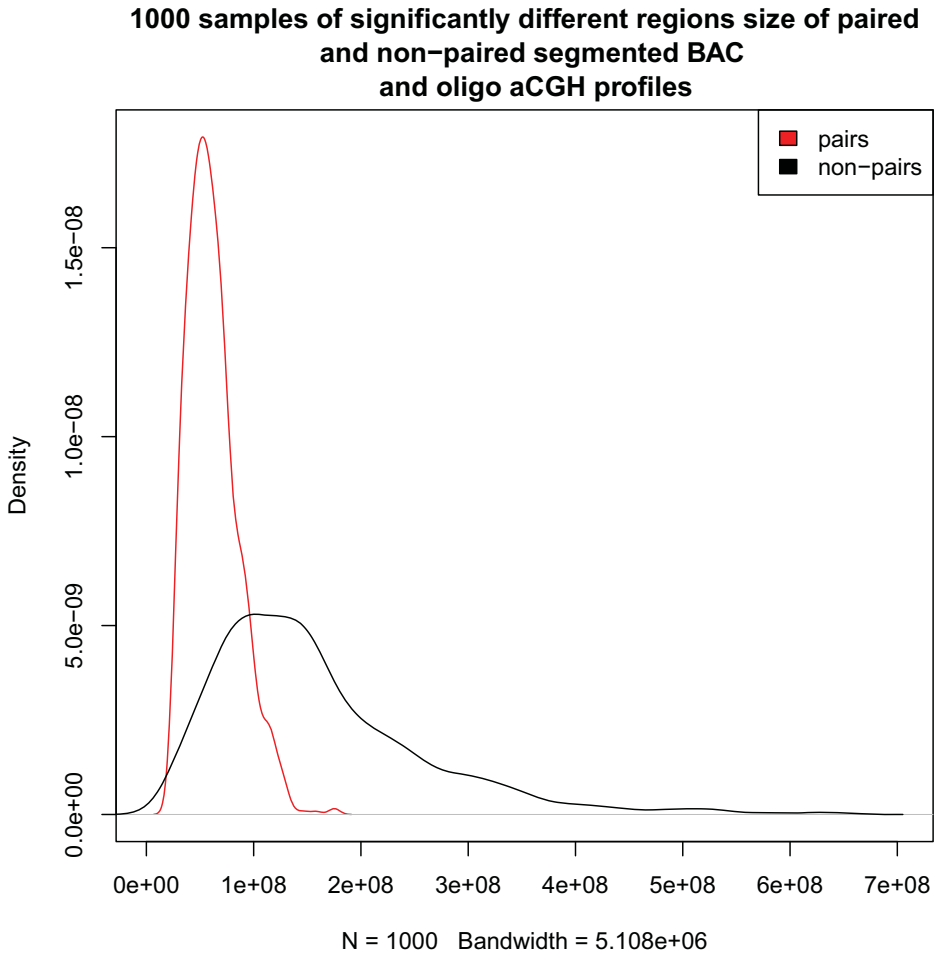

Figure 2.6: (**f**) Density plot of KCsmart analysis of 50 paired BAC-mapped and non-paired segmented log ratio samples sampled 1000 times. Size of genome called aberrantly was summed for each sample.

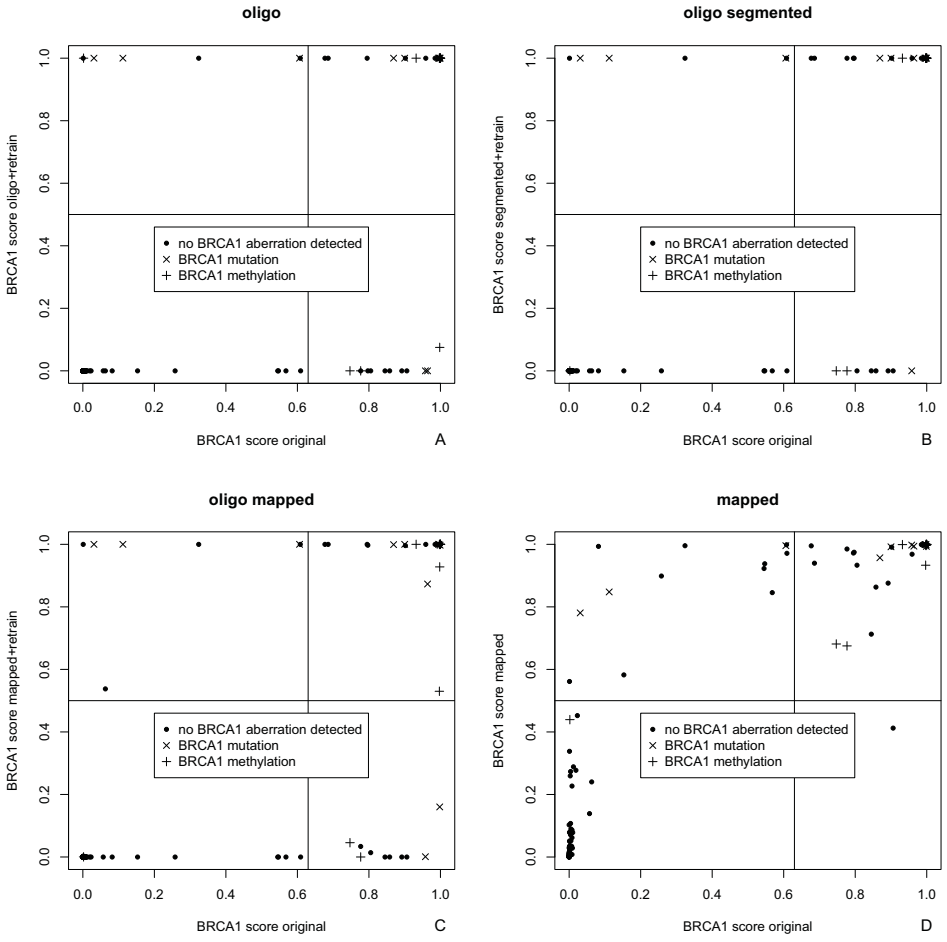

Figure 2.7: **a**) BRCA1 probability scores of a classifier optimized on oligonucleotide data plotted against the scores of the original BAC classifier. (**b**) BRCA1 probability scores of a classifier optimized on segmented oligonucleotide data plotted against the scores of the original BAC classifier. (**c**) BRCA1 probability scores of a classifier optimized on oligo-mapped data plotted against scores of the original BAC classifier. (**d**) BRCA1 probability scores of the oligo-mapped data classified by the BAC classifier plotted against the original BAC classifier.

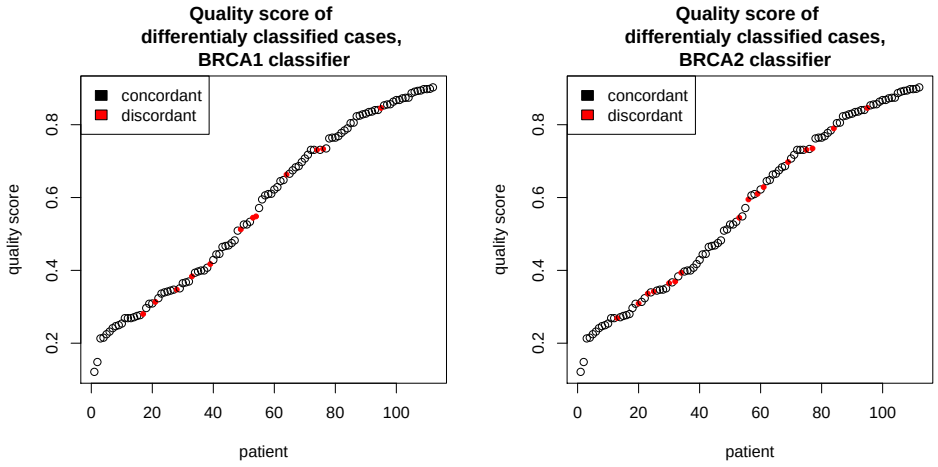

Figure 2.8: Quality scores of samples that were discordantly assigned to the non-BRCA or BRCA class per classifier.

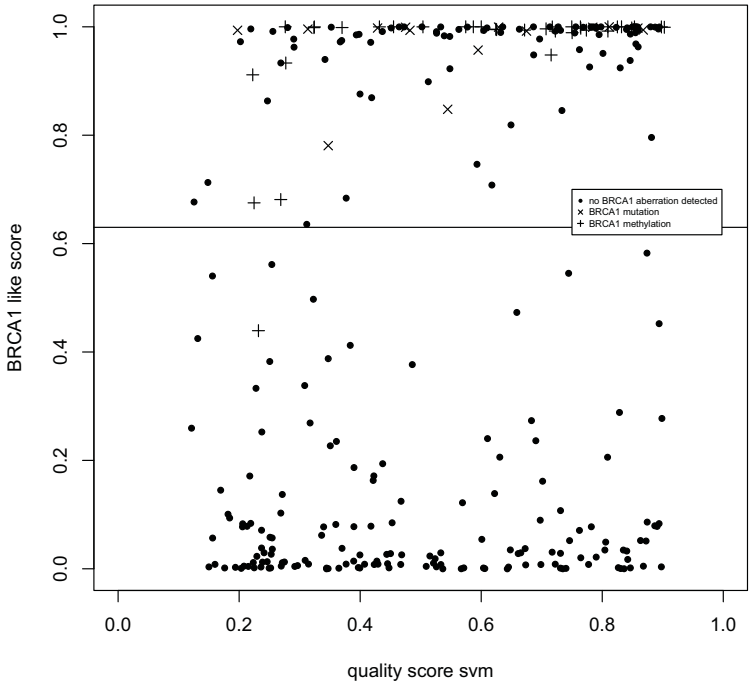

Figure 2.9: Quality scores assigned by the SVM quality classifier plotted against BRCA1 probability scores. + are samples with BRCA1 methylation, × are samples with a pathogenic BRCA1 mutation, • are samples without BRCA1 methylation or a known pathogenic BRCA1 mutation.

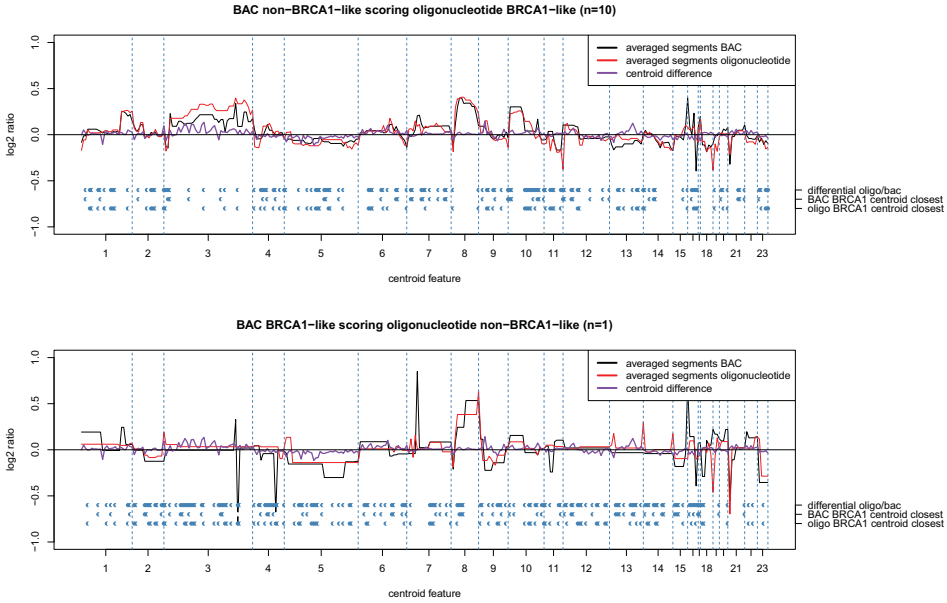

Figure 2.10: Top panel: Averaged BAC profiles (*red*) of segmented log ratios employed in the BRCA1 classifier for cases that switch from non-BRCA1 like to BRCA1 like. Averaged mapped profiles (*red*) and difference of non-BRCA1 like and BRCA1 like centroids (*purple*). Blue spots represent log ratios for which a switch to closest centroid occurs and whether the BAC log ratio is closest or the mapped log ratio is closest to the BRCA1 centroid. Bottom panel: Averaged BAC profiles (*red*) of segmented log ratios employed in the BRCA1 classifier for cases that switch from BRCA1 like to non-BRCA1 like. Averaged mapped profiles (*red*) and difference of non-BRCA1 like and BRCA1 like centroids (*purple*). *Blue* spots represent log ratios for which a switch to closest centroid occurs and whether the BAC log ratio is closest or the mapped log ratio is closest to the BRCA1 centroid.

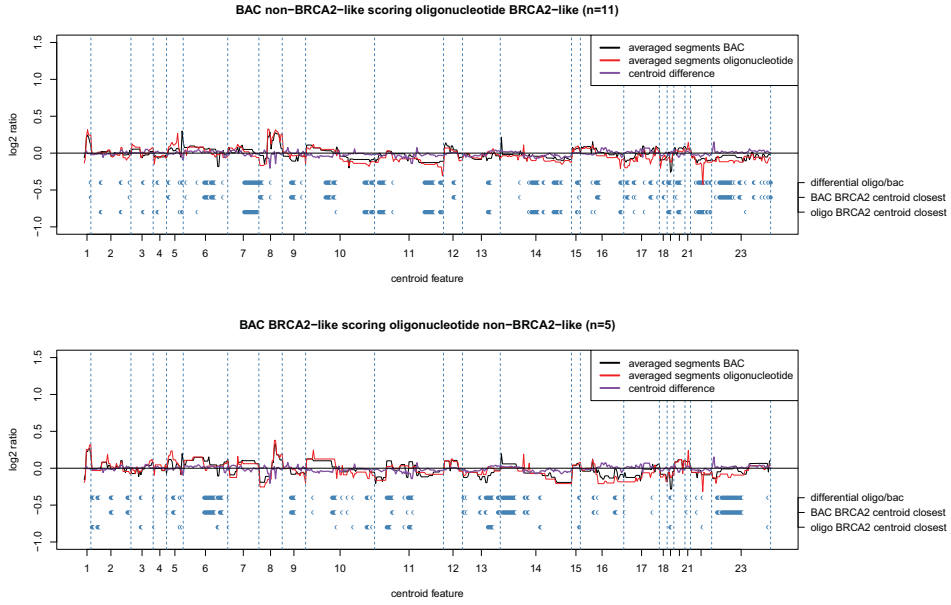

Figure 2.11: Top panel: Averaged BAC profiles (*red*) of segmented log ratios employed in the BRCA1 classifier for cases that switch from non-BRCA2 like to BRCA2 like. Averaged mapped profiles (*red*) and difference of non-BRCA2 like and BRCA2 like centroids (*purple*). *Blue* spots represent log ratios for which a switch to closest centroid occurs and whether the BAC log ratio is closest or the mapped log ratio is closest to the BRCA2 centroid. Bottom panel: Averaged BAC profiles (*red*) of segmented log ratios employed in the BRCA2 classifier for cases that switch from BRCA2 like to non-BRCA2 like. Averaged mapped profiles (*red*) and difference of non-BRCA2 like and BRCA2 like centroids (*purple*). *Blue* spots represent log ratios for which a switch to closest centroid occurs and whether the BAC log ratio is closest or the mapped log ratio is closest to the BRCA1 centroid.

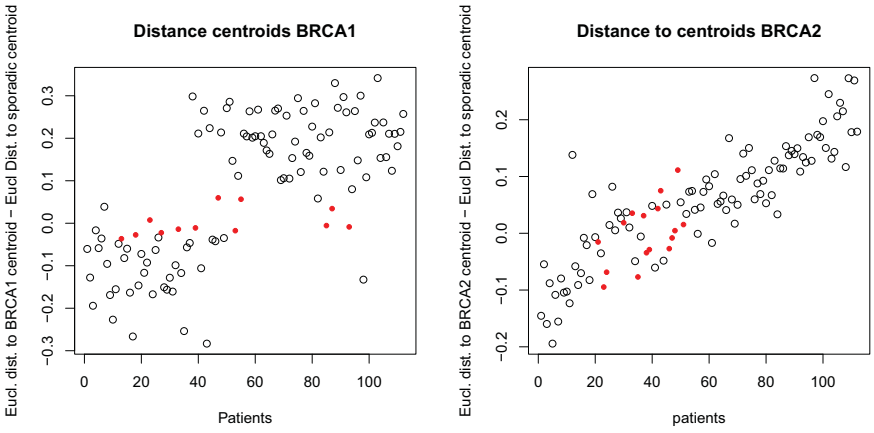

Figure 2.12: For each patient the difference between the Euclidean distance of the BAC data to BRCA centroid and the non-BRCA centroid. Discordant cases are marked in *red*.

REFERENCES

[1] P. C. Schouten, E. van Dyk, L. M. Braaf, L. Mulder, E. H. Lips, J. J. de Ronde, L. Holtman, J. Wesseling, M. Hauptmann, L. F. Wessels, *et al.*, *Platform comparisons for identification of breast cancers with a brca-like copy number profile,* Breast cancer research and treatment **139**, 317 (2013).

[2] D. Thompson and D. Easton, *The genetic epidemiology of breast cancer genes,* Journal of mammary gland biology and neoplasia **9**, 221 (2004).

[3] A. R. Venkitaraman, *Linking the cellular functions of brca genes to cancer pathogenesis and treatment,* Annual Review of Pathological Mechanical Disease **4**, 461 (2009).

[4] L. F. Wessels, T. Van Welsem, A. A. Hart, L. J. Van't Veer, M. J. Reinders, and P. M. Nederlof, *Molecular classification of breast carcinomas by comparative genomic hybridization: a specific somatic genetic profile for brca1 tumors,* Cancer research **62**, 7110 (2002).

[5] E. H. van Beers, T. van Welsem, L. F. Wessels, Y. Li, R. A. Oldenburg, P. Devilee, C. J. Cornelisse, S. Verhoef, F. B. Hogervorst, L. J. van't Veer, *et al.*, *Comparative genomic hybridization profiles in human brca1 and brca2 breast tumors highlight differential sets of genomic aberrations,* Cancer Research **65**, 822 (2005).

[6] S. A. Joosse, E. H. van Beers, I. H. Tielen, H. Horlings, J. L. Peterse, N. Hoogerbrugge, M. J. Ligtenberg, L. F. Wessels, P. Axwijk, S. Verhoef, *et al.*, *Prediction of brca1-association in hereditary non-brca1/2 breast carcinomas with array-cgh,* Breast cancer research and treatment **116**, 479 (2009).

[7] S. A. Joosse, K. I. Brandwijk, P. Devilee, J. Wesseling, F. B. Hogervorst, S. Verhoef, and P. M. Nederlof, *Prediction of brca2-association in hereditary breast carcinomas using array-cgh,* Breast cancer research and treatment **132**, 379 (2012).

[8] S. A. Joosse, K. I. Brandwijk, L. Mulder, J. Wesseling, J. Hannemann, and P. M. Nederlof, *Genomic signature of brca1 deficiency in sporadic basal-like breast tumors,* Genes, chromosomes and cancer **50**, 71 (2011).

[9] M. Vollebergh, E. Lips, P. Nederlof, L. Wessels, M. Schmidt, E. Van Beers, S. Cornelissen, M. Holtkamp, F. Froklage, E. de Vries, *et al.*, *An acgh classifier derived from brca1-mutated breast cancer and benefit of high-dose platinum-based chemotherapy in her2-negative breast cancer patients,* Annals of Oncology **22**, 1561 (2011).

[10] S. Linn, E. Lips, P. Nederlof, L. Wessels, J. Wesseling, M. van de Vijver, E. De Vries, H. van Tinteren, J. Jonkers, M. Hauptmann, *et al.*, *Genomic patterns resembling brca-mutated breast cancers and benefit of intensified carboplatin-based chemotherapy in her2-negative breast cancer.* Journal of Clinical Oncology **29**, 10505 (2011).

[11] J. L. Costa, G. Meijer, B. Ylstra, and C. Caldas, *Array comparative genomic hybridization copy number profiling: a new tool for translational research in solid malignancies,* in *Seminars in radiation oncology*, Vol. 18 (2008) pp. 98–104.

[12] L. O. Baumbusch, J. Aarøe, F.-E. Johansen, J. Hicks, H. Sun, L. Bruhn, K. Gunderson, B. Naume, V. Kristensen, K. Liestøl, *et al.*, *Comparison of the agilent, roma/nimblegen and illumina platforms for classification of copy number alterations in human breast tumors,* BMC genomics **9**, 379 (2008).

[13] C. Curtis, A. G. Lynch, M. J. Dunning, I. Spiteri, J. C. Marioni, J. Hadfield, S.-F. Chin, J. D. Brenton, S. Tavaré, and C. Caldas, *The pitfalls of platform comparison: Dna copy number array technologies assessed,* BMC genomics **10**, 588 (2009).

[14] S. D. Hester, L. Reid, N. Nowak, W. D. Jones, J. S. Parker, K. Knudtson, W. Ward, J. Tiesman, and N. D. Denslow, *Comparison of comparative genomic hybridization technologies across microarray platforms,* Journal of biomolecular techniques **20**, 135 (2009).

[15] N. Wicker, A. Carles, I. G. Mills, M. Wolf, A. Veerakumarasivam, H. Edgren, F. Boileau, B. Wasylyk, J. A. Schalken, D. E. Neal, *et al.*, *A new look towards bac-based array cgh through a comprehensive comparison with oligo-based array cgh,* BMC Genomics **8**, 84 (2007).

[16] O. Krijgsman, D. Israeli, J. C. Haan, H. F. van Essen, S. J. Smeets, P. P. Eijk, M. Steenbergen, D. Renske, K. Kok, S. Tejpar, *et al.*, *Cgh arrays compared for dna isolated from formalin-fixed, paraffin-embedded material,* Genes, Chromosomes and Cancer **51**, 344 (2012).

[17] S. Rodenhuis, M. Bontenbal, L. V. Beex, J. Wagstaff, D. J. Richel, M. A. Nooij, E. E. Voest, P. Hupperets, H. van Tinteren, H. L. Peterse, *et al.*, *High-dose chemotherapy with hematopoietic stem-cell rescue for high-risk breast cancer,* New England Journal of Medicine **349**, 7 (2003).

[18] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria (2014).

[19] E. Venkatraman and A. B. Olshen, *A faster circular binary segmentation algorithm for the analysis of array cgh data,* Bioinformatics **23**, 657 (2007).

[20] F. Picard, E. Lebarbier, M. Hoebeke, G. Rigaill, B. Thiam, and S. Robin, *Joint segmentation, calling, and normalization of multiple cgh profiles,* Biostatistics **12**, 413 (2011).

[21] N. R. Zhang and D. O. Siegmund, *A modified bayes information criterion with applications to the analysis of comparative genomic hybridization data,* Biometrics **63**, 22 (2007).

[22] R. Duin, P. Juszczak, P. Paclik, P. Pakalska, D. De Ridder, D. Tax, and S. Verzakov, *A matlab toolbox for pattern recognition, version 4,* Delft Pattern Recognition Research (2007).

[23] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, *Diagnosis of multiple cancer types by shrunken centroids of gene expression,* Proceedings of the National Academy of Sciences **99**, 6567 (2002).

[24] K. Coombes, *ClassDiscovery: Classes and methods for "class discovery" with microarrays or proteomics [Internet]* (2012).

[25] J. J. de Ronde, C. Klijn, A. Velds, H. Holstege, M. J. Reinders, J. Jonkers, and L. F. Wessels, *KC-SMARTR: An R package for detection of statistically significant aberrations in multi-experiment aCGH data,* BMC research notes **3**, 298 (2010).

[26] T. Therneau, *A Package for Survival Analysis in S. R package version 2.37–2* (2012).

[27] M. Stevenson, T. Nunes, J. Sanchez, R. Thornton, J. Reiczigel, J. Robison-Cox, and P. Sebastiani, *epiR: An R package for the analysis of epidemiological data,* R package version 0.9–43 (2012).

[28] D. J. Sargent, B. A. Conley, C. Allegra, and L. Collette, *Clinical trial designs for predictive marker validation in cancer treatment trials,* Journal of Clinical Oncology **23**, 2020 (2005).

[29] M. A. Vollebergh, J. Jonkers, and S. C. Linn, *Genomic instability in breast and ovarian cancers: translation into clinical predictive biomarkers,* Cellular and molecular life sciences **69**, 223 (2012).

[30] K. Beelen, W. Zwart, and S. C. Linn, *Can predictive biomarkers in breast cancer guide adjuvant endocrine therapy?* Nature reviews Clinical oncology **9**, 529 (2012).

[31] L. M. McShane, D. G. Altman, W. Sauerbrei, S. E. Taube, M. Gion, and G. M. Clark, *Reporting recommendations for tumor marker prognostic studies (REMARK),* Journal of the National Cancer Institute **97**, 1180 (2005).

[32] K. d. Jong, E. Marchiori, A. Van der Vaart, S. Chin, B. Carvalho, M. Tijssen, P. Eijk, P. Van den Ijssel, H. Grabsch, P. Quirke, *et al., Cross-platform array comparative genomic hybridization meta-analysis separates hematopoietic and mesenchymal from epithelial tumors,* Oncogene **26**, 1499 (2007).

[33] B. Weigelt, A. Mackay, R. A'hern, R. Natrajan, D. S. Tan, M. Dowsett, A. Ashworth, and J. S. Reis-Filho, *Breast cancer molecular profiling with single sample predictors: a retrospective analysis,* The lancet oncology **11**, 339 (2010).

[34] A. Mackay, B. Weigelt, A. Grigoriadis, B. Kreike, R. Natrajan, R. A'Hern, D. S. Tan, M. Dowsett, A. Ashworth, and J. S. Reis-Filho, *Microarray-based class discovery for molecular classification of breast cancer: analysis of interobserver agreement,* JNCI: Journal of the National Cancer Institute **103**, 662 (2011).

[35] P.-E. Colombo, F. Milanezi, B. Weigelt, and J. S. Reis-Filho, *Microarrays in the 2010s: the contribution of microarray-based gene expression profiling to breast cancer classification, prognostication and prediction,* Breast Cancer Research **13**, 212 (2011).

[36] H. M. Wood, O. Belvedere, C. Conway, C. Daly, R. Chalkley, M. Bickerdike, C. McKinley, P. Egan, L. Ross, B. Hayward, *et al., Using next-generation sequencing for high resolution multiplex analysis of copy number variation from nanogram quantities of DNA from formalin-fixed paraffin-embedded specimens,* Nucleic acids research **38**, e151 (2010).

3

A SCALE-SPACE METHOD FOR DETECTING RECURRENT DNA COPY NUMBER CHANGES WITH ANALYTICAL FALSE DISCOVERY RATE CONTROL

Ewald van Dyk, Marcel J.T. Reinders & Lodewyk F.A. Wessels

This chapter is published in Nucleic Acids Research 2013. Volume 41, Issue 9. Page e100 [1].

Tumor formation is partially driven by DNA copy number changes, which are typically measured using array comparative genomic hybridization, SNP arrays and DNA sequencing platforms. Many techniques are available for detecting recurring aberrations across multiple tumor samples, including CMAR, STAC, GISTIC and KC-SMART. GISTIC is widely used and detects both broad and focal (potentially overlapping) recurring events. However, GISTIC performs false discovery rate control on probes instead of events. Here we propose Analytical Multi-scale Identification of Recurrent Events, a multi-scale Gaussian smoothing approach, for the detection of both broad and focal (potentially overlapping) recurring copy number alterations. Importantly, false discovery rate control is performed analytically (no need for permutations) on events rather than probes. The method does not require segmentation or calling on the input dataset and therefore reduces the potential loss of information due to discretization. An important characteristic of the approach is that the error rate is controlled across all scales and that the algorithm outputs a single profile of significant events selected from the appropriate scales. We perform extensive simulations and showcase its utility on a glioblastoma SNP array dataset. Importantly, ADMIRE detects focal events that are missed by GISTIC, including two events involving known glioma tumor-suppressor genes: CDKN2C and NF1.

3.1. INTRODUCTION

DNA copy number alterations in cancer, typically recorded by array comparative genomic hybridization (aCGH), single nucleotide polymorphism (SNP) arrays and (more recently) sequencing, can reveal interesting genes that are important for diagnosis, prognosis and targeted therapeutics. However, genomic instability typically introduces random or passenger alterations that make it hard to distinguish recurring alterations (possibly harboring driver genes) from the rest in single sample (tumor) measurements.

A number of statistical methods have been developed to detect aberrations that recur at high frequencies across multiple samples. These methods include CMAR [2], Significance Testing for Aberrant Copy numbers (STAC) [3], Hierarchical Hidden Markov model (H-HMM) [4], Genomic Identification of Significant Targets in Cancer (GISTIC) [5], GISTIC2.0 [6], JISTIC [7] and Kernel Convolution: a Statistical Method for Aberrant Region deTection (KC-SMART) [8].

CMAR and STAC require discretized copy number alteration profiles where genomic regions take on one of three discrete states: a loss, no-aberration or a gain. Although this is partially justified because copy number changes in DNA are discrete in nature, measurements are typically performed on DNA extracted from a heterogeneous pool of cell populations, which could cause deviations from the expected discrete values. Therefore, CMAR and STAC disregard valuable information by ignoring the amplitude of gains or losses in single samples. H-HMM does not require discretized profiles but uses three hidden states to model losses, absence of aberrations and gains.

GISTIC2.0 requires non-discretized, but segmented, profiles. Segmentation (typically performed on single sample profiles) reduces measurement noise, but approximates a signal that varies across the genome with a piecewise constant signal, requiring selection of segment boundaries (breakpoints). Breakpoints can be missed (in noisy profiles), and therefore, segmentation also introduces a form of discretization.

All methods used to detect recurring aberrations, in one way or another, aggregate (sum) all the sample profiles either in raw, segmented or discretized form. This results in a significant reduction in biological noise (passenger events) with respect to signal (recurring events). In addition, aggregation also reduces measurement noise, justifying an approach followed by, e.g. KC-SMART, that avoids segmentation all together and performs smoothing on the aggregated profile.

In particular, GISTIC2.0 and KC-SMART use a statistical framework that weighs both the amplitude and frequency of recurrence in its detection procedure. JISTIC is an adaptation of GISTIC, and all arguments used for GISTIC2.0 in this article also apply to GISTIC and JISTIC.

Possibly the single most desirable property of GISTIC2.0 is its ability to detect focal recurring events embedded in broader events (such as whole chromosome arms being deleted) through a peel-off algorithm requiring knowledge of segment boundaries provided by a segmentation algorithm. However, to the best of our knowledge, there are no approaches that analytically (without resorting to permutation tests) characterize the significance of recurring events and, at the same time, use a principled approach for automatic scale selection (required level of smoothing) while guaranteeing a specified error rate (average number of falsely detected recurrent events).

For an extensive review on (many more) methods, see [9].

Here we present ADMIRE (Analytical Multi-scale Identification of Recurring Events), a smoothing methodology, with the following features:

- Segmentation and/or calling are not required for the genomic profiles. Instead, reduction of measurement noise is achieved by performing smoothing on the aggregated profile;
- Automatic scale selection, or selection of the level of smoothing, is performed on the aggregated profile to increase the power for detecting recurrent events. For example, broad recurrent events are detected with higher significance if we allow for a higher level of smoothing. An important characteristic of the approach is that the error rate is controlled across all scales and that the algorithm outputs a single profile of significant events selected from the appropriate scales;
- A recursive procedure to detect statistically significant focal recurrent events that are embedded in broader events;
- An analytical method that controls the expected number of detected false-positive recurrent events (and therefore helps avoid time-consuming permutation tests)

3.2. METHODS

The ADMIRE methodology is summarized in Fig. 3.1 and described in subsequent subsections. In this example, and subsequent simulations, we simulate aCGH profiles, but any technique, such as SNP arrays (see Results) or sequencing, might be used in principle. In Fig. 3.1, the left column (Column I) illustrates the methodology on measured profiles, whereas Column II illustrates the construction of the null distribution (the expected behavior of the aggregated profile if none of the copy number alterations are recurrent). Multiple aCGH samples are summed [Fig. 3.1B.I (Fig. 3.1, Row B, Column I)] to obtain a single aggregated profile in which recurrent aberrations reveal high peaks compared with passenger events. This indicates that in our model, we consider both the frequency and amplitude of events, similar to the approach followed by GISTIC2.0 and KC-SMART. Next we perform kernel smoothing at different scales (Fig. 3.1C.I) to reduce measurement noise. Fig. 3.1A.II illustrates how we can simulate profiles that share no recurrent events by performing cyclic permutations on each profile individually, Fig. 3.1B.II shows the summation of the resulting profiles to obtain a representative null hypothesis that closely resembles a stationary Gaussian random process and Fig. 3.1C.II shows the kernel convolution per scale. In Fig. 3.1 (Column II), these steps (permutation, summation and smoothing) are repeated 1000 times to obtain an empirical approximation of the null distribution per scale. These distributions are used to derive a threshold per scale corresponding to the desired false discovery rate (FDR) or family-wise error rate (FWER) of passenger events. The permutation test is shown for illustration purposes. ADMIRE avoids permutations altogether by exploiting an analytical relationship between the desired threshold and FDR or FWER. We apply the constant thresholds derived at each scale (kernel width) to obtain recurrent segments for each scale separately (Fig. 3.1D.I). In Fig. 3.1D.I and II, we regard only detected recurrent segments that are of sufficient resolution (the detected event is large compared with the kernel width) and take the union of all significant segments across all scales. The

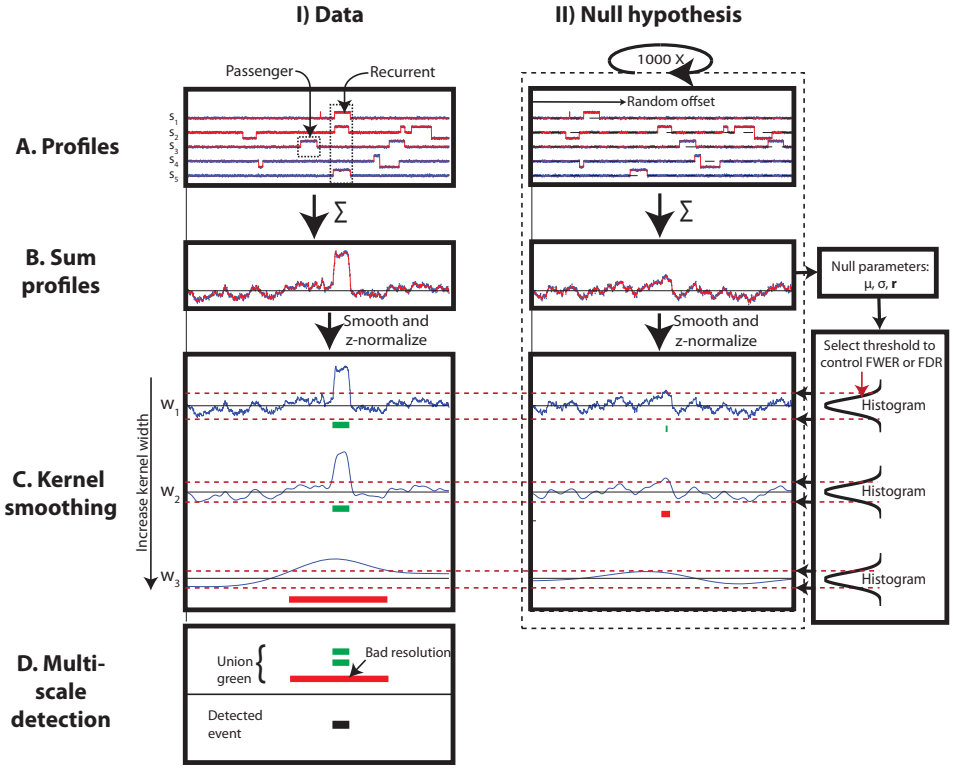

Figure 3.1: Illustrating the steps involved for detecting recurring aberration in multiple copy number alteration profiles with the multi-scale ADMIRE approach. All plots in the left column, Column I, represent data with recurrent events, and Column II shows the exact same procedure when permuting the data to construct a cyclic shift null hypothesis. Column I: (**A**) Illustration of five (of 100) simulated aCGH profiles with recurring events and a number of passenger (random) aberrations. (**B**) The first step in detecting recurring events is to sum all profiles (100 samples) to a single aggregated profile. (**C**) A Gaussian kernel is convolved with the aggregated profile and z-normalized, as described in the text. This is done with many different kernel widths so that focal events can be detected with small kernels and broad events with larger kernels. Ultimately, constant thresholds (derived from the empirical null as outlined in Column II) will be applied on the smoothed signal (both upper and lower tail), as illustrated by the red dashed lines. (**D**) Illustration of how we combine all the events found on multiple scales. Basically, we take the union of all events found on all scales; however, for all kernels (except the smallest), we perform a filtering procedure to ensure the proper resolution. The procedure is simple in that we only keep those events that are substantially (20 times) larger then the kernel width (more on this in the text). Column II: Illustration of the permutation of profiles where each profile's probes are cyclically shifted with a random offset (Panel A) and the summation of the resulting profiles (Panel B) to obtain a representative null hypothesis that closely resembles a stationary Gaussian random process with parameters μ, σ and the auto-correlation **r**. Panel C shows the kernel convolution per scale. In this illustration, we propose to repeat the steps in Panels A, B and C one thousand times to obtain an empirical approximation of the null distribution and use these distributions to derive a threshold per scale corresponding to the desired control of FDR and FWER. However, in this article, we derive an analytical relationship between the thresholds and FWER or FDR.

final step (not shown in Fig. 3.1) involves a recursive procedure to detect focal recurrent events embedded in broad events. In the following sections, we will run through all these steps in more detail.

3.2.1. AGGREGATION

Consider an ordered set of small genomic sequences $\langle \text{seq}_i \rangle$ ($\langle\rangle$ means the set is ordered) that are centered at genomic positions $\langle p_i \rangle$ on a normal reference genome. Each such sequence has an average copy number $\langle t_{s,i} \rangle$ across all cells in a specified tumor sample s. Furthermore, for a normal cell we have a reference copy number for each sequence $\langle n_i \rangle$ (typically $n_i = 2$ for a diploid sequence). From now on we assume that we have an unbiased probe measurement of the log ratio (the base of the log is irrelevant for the subsequent analysis) $a_{s,i} = \log(t_{s,i}/n_i)$, where a positive (negative) value indicates a gain (loss) in the tumor sample.

To find recurring losses or gains, we simply add all sample profiles into one aggregated profile (Fig. 3.1B.I). The aggregated probe values are given by:

$$a_i = \sum_{s=0}^{S-1} a_{s,i}, \tag{3.1}$$

where $s \in \{0,1,...,S-1\}$ and $i \in \{0,1,...,P-1\}$ are the sample and probe indices respectively.

This process is the same as that proposed by KC-SMART and GISTIC2.0, with the fundamental exception that we do not split gains and losses. Little power is lost by doing this, except for clear cases where a region (of the same size) is recurrently lost and gained. The major advantage of not splitting gains and losses is that relevant statistics (such as FDR control) become analytically tractable.

3.2.2. THE NULL-HYPOTHESES

We propose to model the null distribution by performing random cyclic permutations. This implies that for genomic profile s, we push all probes by a random number U_s positions to the right. The U_s probes that are pushed out of the genomic profile on the right are cycled around and fill the U_s empty positions that are created on the left of the profile. This process is performed for each sample independently (Fig. 3.1A.II). We prefer this over random permutation of the probes in a sample profile because it destroys the recurrence structure but retains the auto-correlation between probes. After every sample has undergone a random cyclic shift, all the profiles are aggregated (Fig. 3.1B.II). More specifically,

$$\begin{aligned} A^0_i &= \sum_{s=0}^{S-1} A^0_{s,i} \\ A^0_{s,i} &= a_{s,R_{s,i}} \\ R_{s,i} &= i + U_s (\text{mod} P), \end{aligned} \tag{3.2}$$

where U_s is a uniform random variable covering $\{0,1,...,P-1\}$. Note that each individual probe is identically distributed and identical to the distribution obtained from a permuting null hypothesis, as we randomly select one of the log ratios in each sample. It is also clear that the cyclic auto-correlation remains unchanged for each sample. Furthermore, $A^0 = \{A^0_0, A^0_1, ..., A^0_{P-1}\}$ is a homogeneous random process since the correlation between probes is independent of the probe labels and depends only on their relative ordering on the genome.

We can easily obtain analytical expressions for the mean, variance and auto-correlation of A^0:

$$\begin{aligned}
\mu &= \sum_{s=0}^{S-1} \mu_s, & \mu_s &= \frac{1}{P} \sum_{i=0}^{P-1} a_{s,i} \\
\sigma^2 &= \sum_{s=0}^{S-1} \sigma_s^2, & \sigma_s^2 &= \frac{1}{P} \sum_{i=0}^{P-1} (a_{s,i} - \mu_s)^2 \\
r(\Delta i) &= \sum_{s=0}^{S-1} r_s(\Delta i), & r_s(\Delta i) &= \frac{1}{\sigma^2 P} \sum_{i=0}^{P-1} \left[(a_{s,i} - \mu_s) \times \right. \\
& & & \left. (a_{s,i+\Delta i (\mathrm{mod} P)} - \mu_s) \right]
\end{aligned} \tag{3.3}$$

Alternatively, we can represent the auto-correlation function with a $P \times P$ diagonal-constant correlation matrix $\mathbf{r}$ with $r_{i,j} = r(i-j)$.

Because we are summing multiple profiles, the random process will become multivariate Gaussian (a consequence of the central limit theorem), and the parameters in Eq. 3.3 fully describe the random process.

Technically, it is more desirable to calculate a homogeneous auto-correlation measure based on genomic distance instead of probe index, as probes are not equally spaced. Nonetheless, the proposed scheme provides a good approximation.

3.2.3. SMOOTHING WITH A FIXED KERNEL WIDTH

As we do not assume that the input samples are segmented, and therefore contain substantial measurement noise, it is desirable to smooth the aggregated profile (Fig. 3.1C). We describe an optimal kernel smoothing methodology based on the assumption that the null hypothesis is a random Gaussian process. The idea is that if we fix the kernel type (e.g. Gaussian) and the kernel width (i.e. the number of nearby probes to average, in our case controlled by the standard deviation of the Gaussian kernel), we can normalize the smoothed (continuous) profile so that each point on the genome has exactly the same normal distribution (mean zero and variance one) in the null process. This way we can apply a constant threshold across the whole genome when detecting recurring aberrations.

Since we do not assume that the input samples are segmented and therefore contain substantial measurement noise, it is desirable to smooth the aggregated profile (Fig. 3.1C). We describe an optimal kernel smoothing methodology based on the assumption that the null-hypotheses is a random Gaussian process. The idea is that if we fix the kernel type (for example Gaussian) and the kernel width (i.e. the number of nearby probes to average, in our case controlled by the standard deviation of the Guassian kernel) we can normalize the smoothed (continuous) profile so that each point on the genome has exactly the same normal distribution (mean zero and variance one) in the null process. This way we can apply a constant threshold across the whole genome when detecting recurring aberrations.

The first step is to smooth the signal by convolving the aggregated profile with a kernel.

$$F_w^0(g) \quad = \quad k_w(g) * \sum_{i=0}^{P-1} A_i^0 \delta(g - p_i), \tag{3.4}$$

where g is the position on the genome, $F_w^0(g)$ is the smoothed random process, $k_w(g)$ is the kernel of width w ($\exp(-0.5g^2/w^2)$ for a Gaussian kernel) and p_i is the genomic location of probe i. $*$ represents the convolution operator and δ is the Dirac delta function.

The smoothed function $F_w^0(g)$ is a linear combination of $\{A_0^0, A_1^0, ..., A_{P-1}^0\}$ with coefficients $\{k_w(g-p_0), k_w(g-p_1), ..., k_w(g-p_{P-1})\}$ for any given point in space g.

We can calculate the exact mean and variance of $F_w^0(g)$ as follows:

$$\begin{aligned} \mu_w(g) &= \mu\, m_w(g) \\ \sigma_w^2(g) &= \sigma^2 s_w^2(g), \end{aligned} \tag{3.5}$$

where

$$\begin{aligned} m_w(g) &= k_w(g) * \sum_{i=0}^{P-1} \delta(g - p_i) \\ s_w^2(g) &= \vec{k}_w^T(g) \mathbf{r} \vec{k}_w(g) \end{aligned} \tag{3.6}$$

$\vec{k}_w^T(g)$ is a $1 \times P$ column vector equal to the kernel coefficients $[k_w(g-p_0), k_w(g-p_1), ..., k_w(g-p_{P-1})]^T$ and $\mathbf{r}$ is the auto-correlation matrix.

We choose a threshold function such that $F_w^0(g)$ has the same (single tale) P-value ρ for any given g. Therefore:

$$P\big(F_w^0(g) \geq t(g)\big) = \rho \tag{3.7}$$

As $F_w^0(g)$ is Gaussian we get:

$$t(g) = \mu_w(g) + \sigma_w(g)\, t_\rho, \tag{3.8}$$

where t_p is a constant threshold that controls $F_w^0(g)$ at a P-value ρ.

Equivalently, we can z-normalize $F_w^0(g)$ to apply a constant threshold t_ρ represented by the z-normalized smoothed random process $H_w^0(g)$:

$$H_w^0(g) = \frac{F_w^0(g) - \mu\, m_w(g)}{\sigma\, s_w(g)} \tag{3.9}$$

It is worth mentioning that H_w^0 is a differentiable (smooth) normal random process (with mean zero and variance one for all g), but is non-homogeneous (unlike the discrete random process $\langle A_i^0 \rangle$) due to unequal probe spacings.

3.2.4. COUNTING SIGNIFICANT EVENTS

We ultimately seek to provide a list of genomic regions (broad or focal events) that are significantly recurring and therefore likely to be relevant in cancer development. In providing such a list, we are interested in controlling the expected proportion of regions that are in error (passenger events). We call this the event-based FDR. Before we can do this, it is important to first define what we mean by an event.

For a fixed threshold and kernel width, we define positive and negative excursion sets as follows:

$$\begin{aligned} a^+(h_w, t) &= \{g \in a | h_w(g) \geq t\} \\ a^-(h_w, t) &= \{g \in a | h_w(g) \leq -t\}, \end{aligned} \tag{3.10}$$

where h_w is the smoothed (and z-normalized) aggregate profile (see Eq. 3.9) and a is the set of all g considered (the genome). a^+ and a^- represent all genomic regions that are deemed recurrently gained and lost, respectively (relative to the threshold t). Due to the Gaussian null hypothesis, we will focus all attention on a^+ and realize that symmetric arguments exist for a^-.

We define positive recurring events to be the connected components of a^+. For a smoothed aggregated profile $h_w(g)$ and fixed threshold t, we represent the total number of detected events with $\chi(h_w, t)$. Note that counting the number of events is equivalent to counting the number of up-crossing on the threshold and adding one if the left boundary point is above the specified threshold.

3.2.5. ANALYTICAL RELATIONSHIP BETWEEN THE THRESHOLD AND THE EXPECTED NUMBER OF EVENTS FOUND IN THE NULL-HYPOTHESIS

For H_w^0 (smoothed non-homogenous Gaussian process), we can find an exact analytical expression that relates any given t to the expected number of events found ($E[\chi(H_w^0, t)]$), with the only restriction being that the kernel selected must be differentiable up to the second order. A large amount of work has been done on finding $E[\chi(H_w^0, t)]$ for homogeneous fields [10–13] and little on non-homogeneous fields [14]. Therefore, we extend the theory for non-homogeneous (one-dimensional) processes in Section 3.8.1 and show the final result here. More specifically, for a non-homogeneous process, the expected number of events is given by:

$$E[\chi(H_w^0, t)] = \frac{1}{2}\text{erfc}(\frac{t}{\sqrt{2}}) + \frac{e^{-t^2/2}}{2\pi}\int_{g\in a}\sqrt{v_w(g)}dg, \tag{3.11}$$

where

$$\begin{aligned} v_w(g) &= \text{Var}[\frac{d}{dg}H_w^0(g)] \\ &= \frac{\frac{d}{dg}\bar{k}_w^T(g)\mathbf{r}\frac{d}{dg}\bar{k}_w(g)}{\bar{k}_w^T(g)\mathbf{r}\bar{k}_w(g)} - (\frac{\bar{k}_w^T(g)\mathbf{r}\frac{d}{dg}\bar{k}_w(g)}{\bar{k}_w^T(g)\mathbf{r}\bar{k}_w(g)})^2 \end{aligned} \tag{3.12}$$

$v_w(g)$ is a function that represents the roughness of the random process (naturally the variance in the derivative) and depends entirely on the probe locations, smoothing and auto-correlation $\mathbf{r}$ (and is independent of parameters μ and σ because we z-normalized). For a rough random process (when we perform little smoothing), the integral in Eq. 3.11 will be large and reflects the severity of multiple testing.

Note that we do not concern ourselves with estimating the full distribution of $\chi(H_w^0, t)$ but only the mean. $E[\chi(H_w^0, t)]$ is a sufficient statistic for calculating the FDR (explained later). $E[\chi(H_w^0, t)]$ is also an upper-bound for the FWER and becomes tight for practical FWERs (< 0.1) [15].

We specifically used Gaussian kernels in this work, but Eq. 3.11 hold for all kernels that are twice differentiable. For the application of detecting recurrent events, it is desirable to use a symmetric kernel that drops to zero, such as Gaussian, Student t, Cauchy or wavelet kernels. As the kernel is implemented in a discrete setting, it is also important to ensure that the kernel has a limited frequency bandwidth so that the smoothed aggregated profile can be sampled at a reasonable (Nyquist) rate.

3.2.6. MULTI-SCALE DETECTION

Previous sections indicate how we can control $E[\chi(H_w^0, t)]$ for a fixed kernel width. GISTIC2.0 performs no smoothing on the aggregated profile (or effectively smooth with a small kernel width) and relies on noise reduction through segmentation on single profiles. Fig. 3.2 shows that for unsegmented profiles, we can gain power by considering many kernel widths in parallel. For example if we try to detect broad recurring events, we gain power when increasing the kernel width. On the

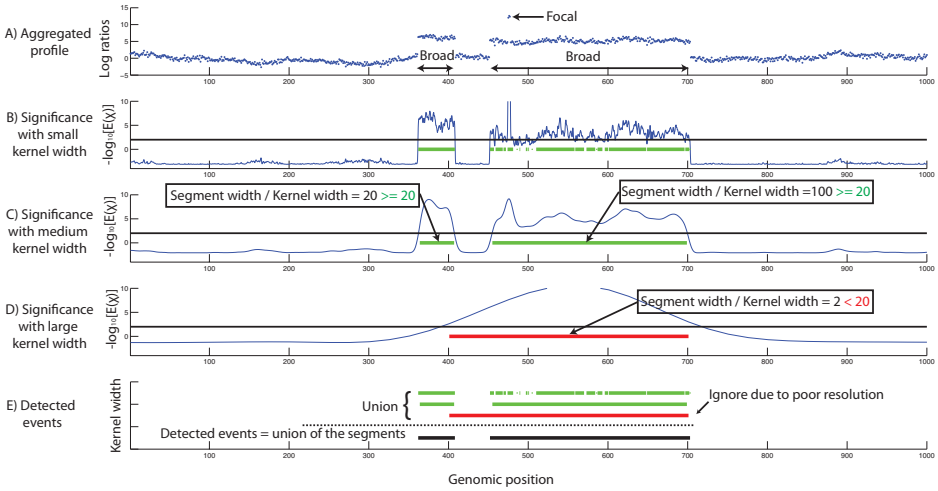

Figure 3.2: Illustration showing how power can be gained by considering multiple scales (levels of smoothing). (**A**) A simulated aggregated profile with two broad recurring gains and one focal gain embedded in a broad event. (**B**) Significance level of the aggregated profile for little smoothing (small kernel width). Owing to the small kernel width, the resolution is high and the boundaries on the detected regions are fairly accurate. This is at the expense of power and results in hundreds of significant segments instead of two broad events. (**C**) Significant power is gained for intermediate kernel widths and the two broad events are found as desired. Furthermore, the resolution is high enough (the segment size is much greater than the kernel width) and therefore the boundaries of the significant events are sufficiently accurate (compared with the aberration size). (**D**) High power is observed for large kernel widths (significance level exceeds the threshold by far) but the resolution is so low that two events are merged into one and boundary estimates are poor. (**E**) We obtain the final estimate of recurring segments by taking the union of all detected events on all scales that reveal sufficient resolution. Note that the focal events embedded in broad events are completely missed. Furthermore, significance in these figures is represented by the expected number of events $E[\chi]$ found across the whole genome (as predicted by the null hypothesis). The threshold is selected at $E[\chi] = 0.01$, a close upper-bound for the FWER of 0.01.

other hand, large kernel widths will reduce the resolution of profiles and estimated recurrent region boundaries will be inaccurate and focal events lost. This is illustrated in Fig. 3.2. In Panel B, the resolution is high, resulting in accurate boundaries but low power causing the broad event to be shattered in many small events. In Panel D, the power is high, but the boundaries are inaccurate. Panel C shows a good compromise between boundary precision and power. Therefore, it is desirable to restrict the size of allowed kernels based on the size of detected events. To be more specific, at any given scale (except the smallest kernel width considered, as the resolution is assumed to be high), all detected events that have a detected width smaller than $\alpha = 20$ times the kernel width will be ignored because they result in a poor resolution. Rather, these events are detected at a smaller scale to ensure a proper accuracy of the event boundaries. In fact, for $\alpha = 20$, at least 70% of any detected event will overlap with a real recurrent event (see Section 3.8.2). α can be set by the user, and in Section 3.9.1 we illustrate how different settings of α influence results (see Fig. 3.9).

Finally, the union of all the remaining significant regions across all scales represents the recurring events in the data. This multi-scale procedure will more likely merge events that appear on the smallest scale than create new ones on a larger scale. This enables us to keep control over the number of detected events (see Section 3.8.2 and Fig. 3.5).

3.2.7. UPDATING THE NULL-PARAMETERS BASED ON KNOWN RECURRENT EVENTS

Parameters μ, σ and $\mathbf{r}$ will, in general, be conservative estimates for the non-recurrent null hypothesis if estimated on all probes, especially if a large proportion of these probes are recurrent. Therefore, it is desirable to ignore all probes that are known to be recurrent when estimating the null parameters. This is done iteratively by first calculating conservative parameter estimates (with all probes considered) and then removing all the probes that are deemed recurrent through the multi-scale detection procedure. If we re-calculate the null parameters (which will be less conservative) with the remaining probes only, more recurring events will we found. This process is repeated until no more new recurring events can be found (see Section 3.8.3 commenting on the convergence behavior). Although this method will drastically increase power, the null parameters will either be slightly optimistic or remain conservative if some recurrent events remain undetected.

3.2.8. RECURSIVE MULTI-LEVEL DETECTION OF RECURRING ABERRATIONS

The events detected by the procedure as described thus far include focal and broad events, but we are not yet able to detect focal events that are embedded in broad events. To find those, we propose a recursive scheme that finds new events that are embedded in earlier detected events. For example, lets say that we find (among other) one broad recurrent gain that starts (ends) at genomic location g_s (g_e). We re-estimate the null parameters μ, σ and $\mathbf{r}$ from all probes between g_s and g_e and perform the multi-scale analysis to find smaller events embedded within this broad event. This procedure for finding a focal event within a broad event is illustrated in Fig. 3.3. Again we iteratively update the null parameters until the null region converges (a new null region inside the broad event). Note that the boundaries of the detected broad event (g_s and g_e) might be inaccurate and therefore embedded focal events might be detected at the border of the initial broad event. As these are a result of the boundary inaccuracy, we simply ignore them (unless, e.g. it is a focal gain within a gain). We repeat this recursive procedure until no more events can be found and represent the results in recursive levels.

On a final note, not only does the recursive multi-level detection procedure allow us to detect recurring events embedded in broad recurring events, but also helps to improve our estimate on $E[\chi]$, as explained in Section 3.8.4.

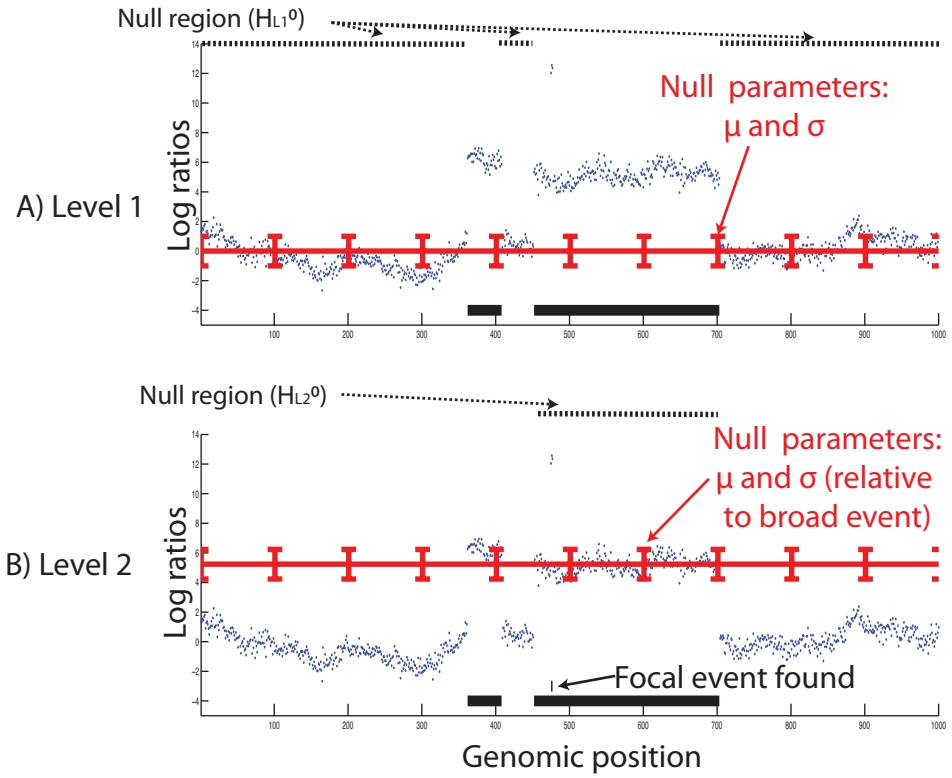

Figure 3.3: Illustrating the recursive multi-level detection methodology. (**A**) On recursive level 1, we detect recurrent aberrations with the proposed multi-scale methodology. Note that the region in which we finally estimate the null parameters (μ, σ and $\mathbf{r}$) is restricted to H^0_{L1}, as illustrated by the dotted line at the top of the figure. (**B**) On recursive level 2, we follow the exact same procedure, except this time, estimate the null parameters in the broad event H^0_{L2}. This allows us to detect embedded focal events inside broader events.

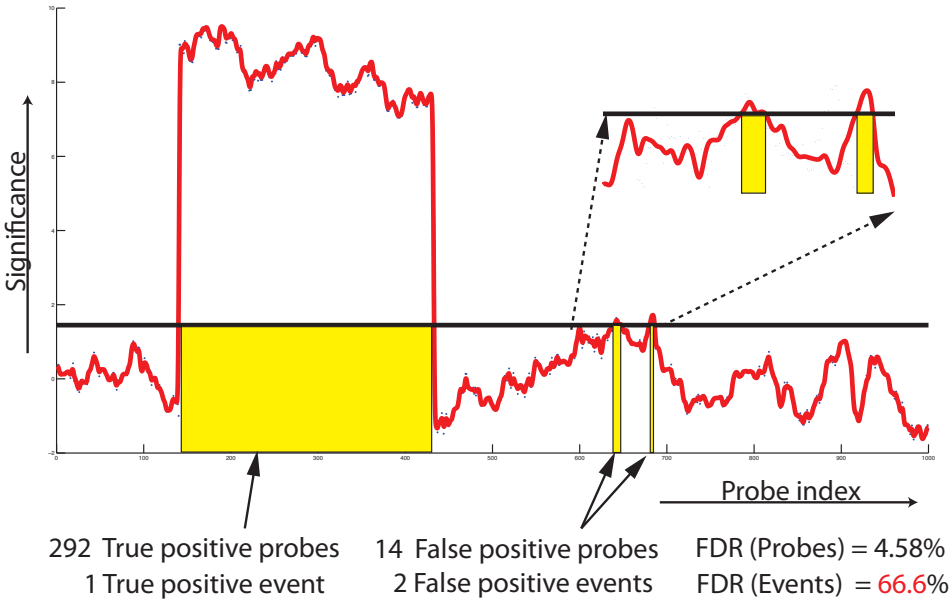

Figure 3.4: Probe-based versus event-based FDR control. Illustration on how controlling the probe-based FDR (expected proportion of detected probes that are false-positives) can introduce an unexpected proportion of focal events simply due to the presence of broad chromosomal recurring aberrations.

3.2.9. FDR CONTROL

As we are able to predict the expected number of events found in the null hypothesis, we can also control the event-based FDR, the expected proportion of detected events that are false discoveries.

To see this, consider the Benjamini-Hochberg procedure [16] that controls the FDR at level q for m independent or positive dependent tests: Let $p_{(1)} \leq p_{(2)} \leq ... \leq p_{(m)}$ be the ordered observed P-values and m_0 the number of true null hypotheses. If we reject the null hypotheses for tests $p_{(1)} \leq p_{(2)} \leq ... \leq p_{(k)}$, where

$$k = \max\{i : m_0 p_{(i)} \leq i q\}, \tag{3.13}$$

then FDR $\leq q$. If we reject all tests with a P-value lower than ρ, then the expected number of false-positive tests $E_\rho(\#FP) = m_0\rho$ (irrespective of the correlation that might exist between tests). Therefore, Eq. 3.13 can be rewritten:

$$k = \max\{i : E_{p_{(i)}}(\#FP) \leq i q\} \tag{3.14}$$

For our application, Eq. 3.14 is intuitive. For the i'th detected event, if the ratio between the expected number of false-positive events ($E(\chi)$) and the number of events i detected is smaller than the FDR (q), then the FDR will be in control. We can lower the detection threshold until the inequality in Eq. 3.14 is violated.

We propose the following procedure to find an appropriate value for $E[\chi]$ to control the FDR at level q:

- Set $n = 1$, $E_1[\chi] = q$ and $i_1 = 0$;
- REPEAT:

- Detect recurrent events using ADMIRE with thresholds corresponding to $E_n[\chi]$. Count the number of detected events i_{n+1};
- IF $i_{n+1} \leq i_n$: BREAK;
- Set $E_{n+1}[\chi] = (i_{n+1}+1)q$;
- Set $n = n+1$

This methodology is different from that performed in GISTIC2.0. GISTIC2.0 regards each probe as an independent test (owing to the random permutation scheme) and uses the methodology proposed by Benjamini and Hochberg [17] to control the probe-based FDR (i.e. the proportion of false-positive probes). In contrast, ADMIRE performs event-based FDR, and this subtle, yet profound, difference is illustrated in Fig. 3.4.

3.3. RESULTS

This section starts with an artificial, simulated dataset to illustrate several properties of ADMIRE. We start off by demonstrating that the theoretical estimate of the expected number of events, $E[\chi]$, is indeed a good approximation of the empirically observed number of events under a wide range of experimental conditions. Then we move on to show that $E[\chi]$ is a close upper-bound of the FWER and that the ADMIRE algorithm does control event-based FDR at the desired level. Finally, we demonstrate the properties of ADMIRE on a real-world glioma dataset.

3.3.1. DATASETS

SIMULATED DATASETS

We simulate aCGH profiles on a genome consisting of 2.4×10^8 bps and randomly select 12 000 probe positions for measurements. For each profile, we select 159 random breakpoints (160 segments) on the genome, of which a random selection of 50% of the segments take on log ratios of 0 (all probes in these segments). The remaining segments randomly take on log ratios of $+1$ and -1, representing passenger gains and deletions, respectively. We also add random Gaussian (measurement) noise to each profile (with variance σ_n^2) for a specified signal to noise ratio (SNR) defined as $\text{SNR} = 1/\sigma_n^2$. For example, an SNR of 10^{99} implies negligible measurement noise.

When recurrent events are added, we typically specify a width, location and frequency of recurrence across samples. For simplicity, probes covered by recurrent events take on values of $+1$ or -1 to represent recurrent gains or losses, respectively. For example, we might decide to add a recurrent event centered at 1.2×10^8 bps, 1×10^6 bps wide with a 30% frequency of occurrence across all samples.

In total, there are three global parameters that will be varied across experiments: (i) the number of samples to aggregate (S), (ii) the SNR and (iii) the number of recurrent aberrations. For every recurrent aberration, we also specify the width, genomic location and frequency of occurrence.

For a detailed description on how we typically generate such a dataset, see Section 3.9.2.

THE GLIOMA DATASET

To demonstrate the properties of ADMIRE on real data, we used the dataset described by Beroukhim *et al.* [5] consisting of 141 high-quality glioma samples (107 primary Glioblastoma multiforme (GBM), 15 secondary GBMs and 19 lower-grade gliomas) to aggregate. DNA was hybridized on a Affymetrix $\approx 100{,}000$ SNP array platform. Batch effects and systematic errors were removed using the exact methodology described by Beroukhim *et al.* [5] (see their Supporting information). All samples were segmented using Gain and Loss Analysis of DNA (GLAD) [18] to reduce measurement noise (this was done for both GISTIC2.0 and ADMIRE), and all known copy number variation probes were removed from the analysis.

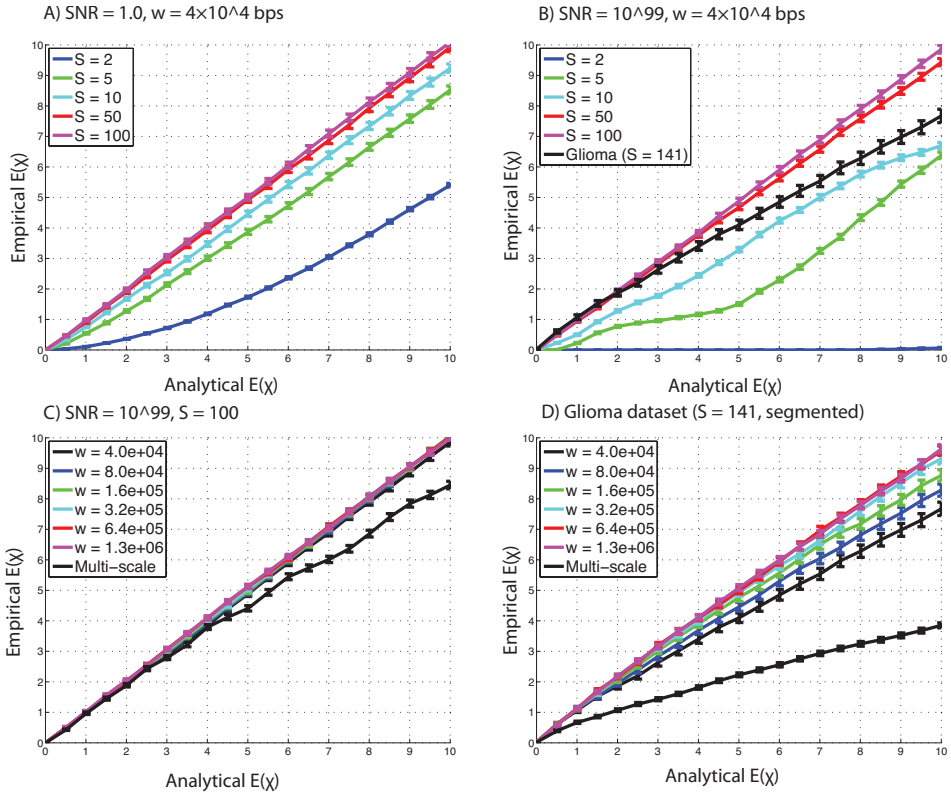

Figure 3.5: Illustration of the relationship between the analytical estimates of $E[\chi]$ (x-axis) and that measured across 1000 simulations (y-axis) of aCGH profiles containing only passenger events. (**A**) We fix the kernel width to be small (40 kb) and the SNR at 1 to represent measurement noise. We vary the number of samples to aggregate for each simulation experiment. (**B**) A similar experiment on simulated aCGH profiles where we added no measurement noise (SNR = 1×10^{99}) and therefore effectively work with segmented samples. The black line depicts the result obtained when using cyclic permutation to create a null hypothesis on the glioma dataset. (**C**) The number of simulated samples to aggregate is fixed at 100 and the kernel width is varied, showing good theoretical predictions for all kernels. The black line indicates the mean number of events detected when we apply multi-scale selection. (**D**) Similar results are depicted when using cyclic permutations to create a null hypothesis on the glioma dataset. The genome size for the simulated data is only 12×10^8 bps, whereas the glioma dataset consists of all probes stretching from chromosome 1 to 22. Error bars indicate the standard error of the empirical $E[\chi]$.

3.3.2. $E[\chi]$ SIMULATIONS

We simulate aCGH profiles using the methodology proposed earlier; however, we do not add any recurrent aberrations.

To investigate whether our theoretical model of the expected number of detected events ($E[\chi]$) is accurate for different thresholds, noise levels and kernel widths, we performed the following experiments. We varied the number of samples to aggregate, S, such that $S \in \{2, 5, 10, 50, 100\}$; the SNR assumed two values, SNR $\in \{1, 1 \times 10^{99}\}$ and the Gaussian kernel width was set to $w \in \{4 \times 10^4, 8 \times 10^4, 1.6 \times 10^5, 3.2 \times 10^5, 6.4 \times 10^5, 1.3 \times 10^6\}$. For combinations of these variables, we simulated 1000 artificial datasets.

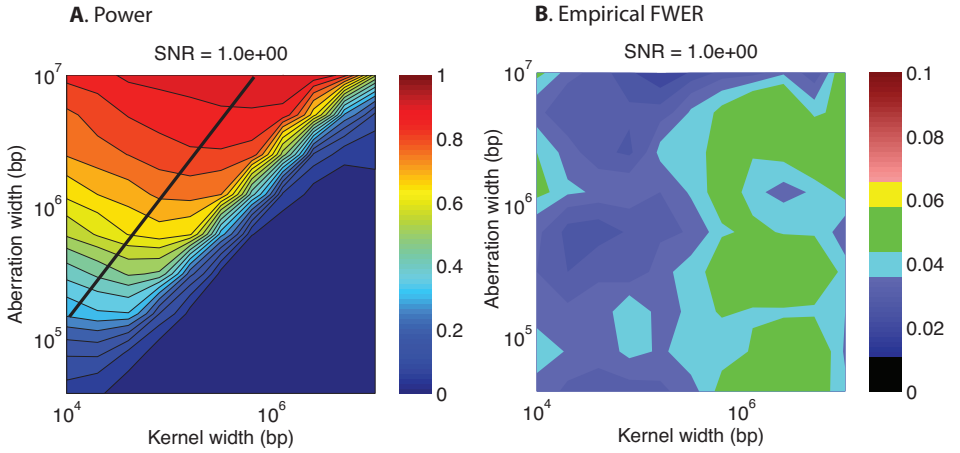

Figure 3.6: (**A**) A representative plot of the power for detecting a recurring aberration as a function of the aberration size and kernel width for the SNR fixed at 1. In this experiment, we added only a single recurring aberration per experiment and fixed $E[\chi]$ at 5%. The black line indicates the maximum allowed kernel width at which an aberration can be detected if we apply filtering with $\alpha = 20$ in the multi-scale methodology. See Fig. 3.11 for similar plots at different SNRs. (**B**) The empirical FWER. The green regions indicate that the measured FWER is within 1 standard deviation of the expected 5% FWER.

In Fig. 3.5A, we show the relationship between the analytical and empirical $E[\chi]$ as the detection threshold is varied for a fixed kernel width of 4×10^4 bps (two probes per kernel width, on average) and an SNR of 1 ($\sigma_n^2 = 1$ per sample). We show this result for all values of S.

Fig. 3.5B is similar to Fig. 3.5A, except that we do not add measurement noise. This serves to illustrate that our approach can also be applied to segmented data.

The main conclusion drawn from Fig. 3.5A and B is that the analytically predicted $E[\chi]$ becomes more accurate as we increase the number of aggregated samples due to the central limit theorem. For smaller sample sizes, the theoretical estimate is conservative.

In Fig. 3.5C, we fix S to 100 and the SNR to 1×10^{99} and vary the kernel widths to show that the analytical estimate of $E[\chi]$ remains accurate for all kernel widths. We also show that the empirical $E[\chi]$ is smaller than the analytical $E[\chi]$ if we perform the multi-scale detection.

Next we investigated the relationship between the empirical and theoretical estimate of $E[\chi]$ on the glioma dataset. To obtain an empirical estimate of $E[\chi]$, we constructed a null hypothesis by repeating the cyclic permutation procedure, aggregation and kernel smoothing as outlined in Fig. 3.1II, one thousand times on the glioma dataset. The results for $S = 141$ and all kernel widths including the multi-scale analysis are depicted in Fig. 3.5D. Overall, the theoretical prediction serves as a relatively tight upper-bound for the empirical estimate, but depends on the kernel width. More specifically, the estimate of $E[\chi]$ becomes more accurate for larger kernels owing to adjacent probes being averaged (and again the central limit theorem suggests better convergence).

Overall, this experiment shows that the analytical $E[\chi]$ is sufficiently accurate and that the multi-scale procedure produces conservative results.

3.3.3. FWER SIMULATIONS

We observed earlier that $E[\chi]$ is a close upper-bound for the FWER [15], and in this section, we perform simulations to verify this fact. We simulated aCGH profiles using the same methodology proposed earlier.

We fix the number of samples to aggregate to 100 and only add one recurrent event centered at 120 Mbps with a given width, w_a, and a 30% chance of occurrence per sample.

In every simulation, we also fix the kernel width and therefore do not perform a multi-scale analysis. Neither do we search for embedded events through recursion. However, we do update the null parameters iteratively based on known recurrences. See Section 3.9.3 for a detailed description of the experiment.

Fig. 3.6A depicts a typical power plot as a function of aberration size and kernel width - for an elaborate collection of these plots for different SNRs, see Fig. 3.11. This plot shows how the power changes (for the analytical FWER fixed at 5%) for detecting recurring aberrations of different sizes (one event per simulation) while varying the kernel width. We can observe that for a fixed kernel width, the power decreases as the aberration size decreases. In fact, there is an abrupt drop in power when the aberration size equals the kernel width, as indicated by the diagonal ridge in the panel. In general, we can conclude that as long as the aberration is larger than the kernel width (region above the diagonal line), we have more power to detect the aberration. Fig. 3.6B shows that the measured FWER (the chance of detecting one or more false-positives) is close to that predicted by $E[\chi]$, as expected. From these simulations it is clear that for any recurrent aberration of a fixed width, a fixed kernel width can be selected to gain optimal power. If the kernel width becomes too large, we observe a drastic loss in power, as indicated by the lower right corner in Fig. 3.6A. Note that in contrast, Fig. 3.2 suggests that larger kernels increase the power, but if we extend Fig. 3.2 to show even larger kernels, the significance levels will drop drastically.

3.3.4. FDR SIMULATIONS

For the FDR experiments, we expanded the simulated dataset described previously to include recurrent events of different sizes and to have overlapping recurrent events. This will allow the possibility to estimate the capacity of the *complete* ADMIRE algorithm to control the FDR. More specifically, we expanded the simulated dataset by adding $N_b = 2$ broad (20×10^6 bps, 1000 probes, on average) and $N_m = 5$ medium-size (2×10^6 bps, 100 probes, on average) non-overlapping recurrent events at random locations (albeit consistent between samples) on the genome. Furthermore, we added a varying number (N_f) of recurrent focal events (100 kb, five probes, on average) across the genome (potentially overlapping with the broad- and medium-size events). For each recurrent event, we select a random frequency (between 0 and 1) of occurrence across samples.

The complete ADMIRE algorithm has been applied with a specified analytical FDR. The number of samples to aggregate (S) is varied, as well as the SNR and the number of focal recurring events (N_f). An event is considered a true-positive if at least 70% of the detected region overlaps with a true recurrent region (the multi-scale detection procedure with $\alpha = 20$ filtering guarantees an overlap of at least 70%). The number of true-positive events is then the sum of the number of true recurrent broad (maximum two), medium (maximum five) and focal events found. The empirical FDR is calculated by averaging the proportion of falsely detected events across 1000 simulation experiments. Likewise, the empirical power is the average proportion of true recurring events that are detected. For example, when we add only one recurring focal event, we hope to detect eight true events (two broad, five medium and one focal). If, for example, we detect four of the eight recurrent events and one extra false event in one simulation, the measured FDR would be 20% and the power 50%.

In Fig. 3.7A, we fix the number of samples to aggregate to 200 and the SNR is set high (zero measurement noise and profiles are segmented). We vary the number of focal events and the

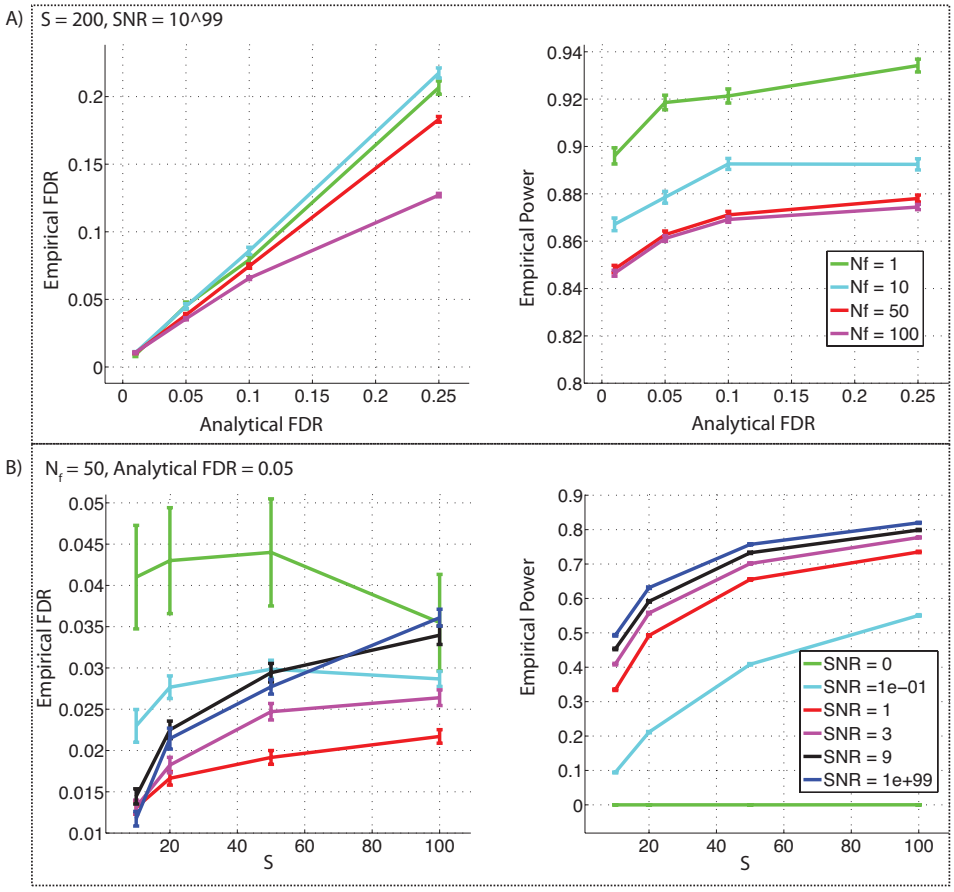

Figure 3.7: The relationship between the theoretically predicted analytical FDR and empirical FDR and power for a simulated dataset. (**A**) The empirical FDR (left panel) and power (right panel) as a function of the analytical FDR (varied between 1 and 25%) for the number of true focal recurrent events assuming the following values, $N_f \in \{1, 10, 50, 100\}$, while keeping the number of samples to aggregate per simulation fixed at 200, i.e. $S = 200$. Furthermore, we do not add any noise, as the SNR = 10^{99}, implying that all samples are segmented. (**B**) The empirical FDR (left panel) and power (right panel) as a function of the number of samples to aggregate S for the SNR assuming the following values, SNR $\in \{0, 0.1, 1, 3, 9, 10^{99}\}$, while keeping the number of focal recurrent events and FDR fixed at 50 ($N_f = 50$) and 5%, respectively.

analytical FDR and represent the measured FDR and power. In Fig. 3.7B, we fix the number of focal recurrent events to 50 and the analytical FDR to 5%, while varying the number of samples that are aggregated and the SNR.

From Fig. 3.7A and B, it is clear that the empirical FDR is smaller than that predicted analytically. The three main reasons for this are the following:

- Inaccurate estimation of the null random process parameters μ, σ and $\mathbf{r}$. The higher the number of true positives missed, the more conservative the null parameter estimates are and the true FDR will be smaller than predicted. Ultimately, this estimate will be most conservative if we estimate null parameters across the whole genome. In Fig. 3.7A, we can clearly see that the FDR decreases when we increase the number of recurrent events. This is because for a fixed threshold, the expected number of undetected events is proportional to the total number of events (this is a simple consequence of how we generated the data). Therefore, for a larger number of recurrent events, the expected number of events that go undetected will be large and therefore the null parameters will be more conservative. This is especially noticeable in Fig. 3.7B, where the FDR is fixed at 5% and we vary the SNR. For an SNR of zero, the null parameters will be accurate (as recurrent events do not exist) and we expect the FDR estimate to be close to the predicted value, whereas for an SNR of one, the null parameters will include a significant proportion of the recurrent signal. This situation improves again for higher SNRs owing to an increase in power;
- The multi-scale procedure in Fig. 3.2 also ensures a conservative estimate on $E[\chi]$, as illustrated in Fig. 3.5C (and D);
- If the number of samples to aggregate (S) is small, the Gaussian model becomes inaccurate for the null hypothesis. This explains the reduced FDR for small values of S in Fig. 3.7B. Note that for SNR = 0, the Gaussian model is accurate no matter how many profiles we aggregate.

The power curves in Fig. 3.7A (right panel) counterintuitively suggest that we lose power when increasing the number of focal events. However, if we consider that medium-size ($N_m = 5$) and broad-size ($N_b = 2$) events are detected with much higher power, it becomes obvious. For example, if we add only one focal event, then 7/8 of all recurrent events are of medium or broad size, whereas for 100 focal events, this ratio is only 7/107.

3.3.5. APPLICATION ON GLIOMA DATA

We compare the recurring events found by both ADMIRE and the latest version of GISTIC2.0 at 25% FDR on the glioma dataset described earlier. The results in Fig. 3.8 reveal that ADMIRE finds many more events (in total 223 focal and broad events) than GISTIC2.0 (50 focal and broad events). All the known glioma tumor suppressors and oncogenes found by GISTIC2.0 are also recovered by ADMIRE. Although GISTIC2.0 performs probe-based FDR, and is therefore expected to be optimistic (see Fig. 3.4), there are many sources of power loss that are overcome by ADMIRE as follows:

- Substantial power is gained, as regions that are known to be significantly recurrent are ignored when estimating the null parameters;
- We account for the auto-correlation in the genomic profiles (in the null hypothesis), and as nearby probes reveal high positive correlations, the severity of multiple testing is reduced;
- By considering multiple scales (levels of smoothing), we gain substantial power for detecting broader events.

We give a multi-level representation of the events found by both ADMIRE and GISTIC2.0 in Fig. 3.8. GISTIC2.0 dichotomizes events into focal and broad (chromosome arm-length) recurrences. All events found by GISTIC2.0 on a chromosome-arm level are indicated on the first level

(+1 for gains or -1 for losses). After removing aberrations that stretch across whole chromosome arms, GISTIC2.0 also finds probes that are significantly recurrent with q-values below the 25% probe-based FDR level. All regions defined by these probes are represented on a second level (+2 for gains and -2 for losses). GISTIC2.0 uses an arbitrated peel-off algorithm to identify multiple potential target regions inside each significant region below the q-value threshold. The boundaries of these regions are then fine-tuned using an algorithm called RegBounder [6]. These regions are then represented on the third level. In contrast, ADMIRE makes no such distinction and simply adds levels until convergence. ADMIRE only adds more focal regions on a higher level if it can be proved significantly recurrent (below 25% event-based FDR) with respect to its immediate background (the level below).

Visually it is clear that ADMIRE shows an increase in power for detecting broad events (due to the multi-scale approach), as can be seen, for example, when looking at the third level of recurrent deletions in chromosome 1p in Fig. 3.8B.III (containing CHD5). In contrast, GISTIC2.0 only finds a focal recurrent aberration (close to CHD5). The aggregated profile in Fig. 3.8B.II reveals that indeed the broad event (third recursive level) detected by ADMIRE is likely a real event (of the same width), but it is difficult to prove significance of the focal event found by GISTIC2.0 relative to this background. It is possible to look for maximal peaks inside the broad event to help guide us towards genes that are likely relevant, but cannot be significantly distinguished from neighboring genes. In this sense, ADMIRE is more conservative at detecting focal events than GISTIC2.0.

One can argue that it is important to detect broad events with high power (justifying the multi-scale methodology). To see why, consider a single scale analysis (with little or no smoothing). One might not have the necessary power to detect some broad events; however, random (passenger/measurement noise) focal events that surpass the threshold (in combination with the broad event) will lead to shattered positives. In contrast, the multi-scale procedure will likely detect the broad event, and if not, we regard the overlapping focal events (that surpass the threshold) to be non-random (with respect to its immediate background).

ADMIRE detects a number of focal events that are missed by GISTIC2.0, including two events involving known glioma tumor suppressor genes: CDKN2C and NF1. The focal recurrent event overlapping with CDKN2C is showcased in Fig. 3.8D. NF1 is showcased in Fig. 3.12.

3.4. DISCUSSION

ADMIRE is an algorithm designed to assist in the discovery of broad and focal (potentially overlapping) recurring events. It does not require segmentation of single sample genomic profiles and therefore admits heterogeneous samples that do not display clear breakpoints in copy number. ADMIRE performs a kernel smoothing methodology on the aggregated profile that optimizes the power for detecting recurring events if the null hypothesis closely resembles a Gaussian random process. Our previous algorithm, KC-SMART, is an example of another kernel smoothing methodology. Compared with KC-SMART, ADMIRE shows a drastic increase in power, especially for focal aberrations, when we fix the FWER at 5% (see Fig. 3.10).

Furthermore, ADMIRE performs analytical event-based FDR control instead of probe-based FDR. The user thus receives a list of recurrent regions for which the expected proportion of false regions is lower than that specified by the FDR.

From a technical perspective, ADMIRE gains power in detecting recurring events by accounting for the auto-correlation between probes (reduces the severity of multiple testing), performing a multi-scale smoothing methodology (especially helps for detecting broad events) and perhaps most importantly by estimating the behavior of passenger events (the null hypothesis) in regions that do not contain known recurrent events. Although it might be regarded as unimportant to detect broad events with high power (as focal events are expected to be of greater importance when

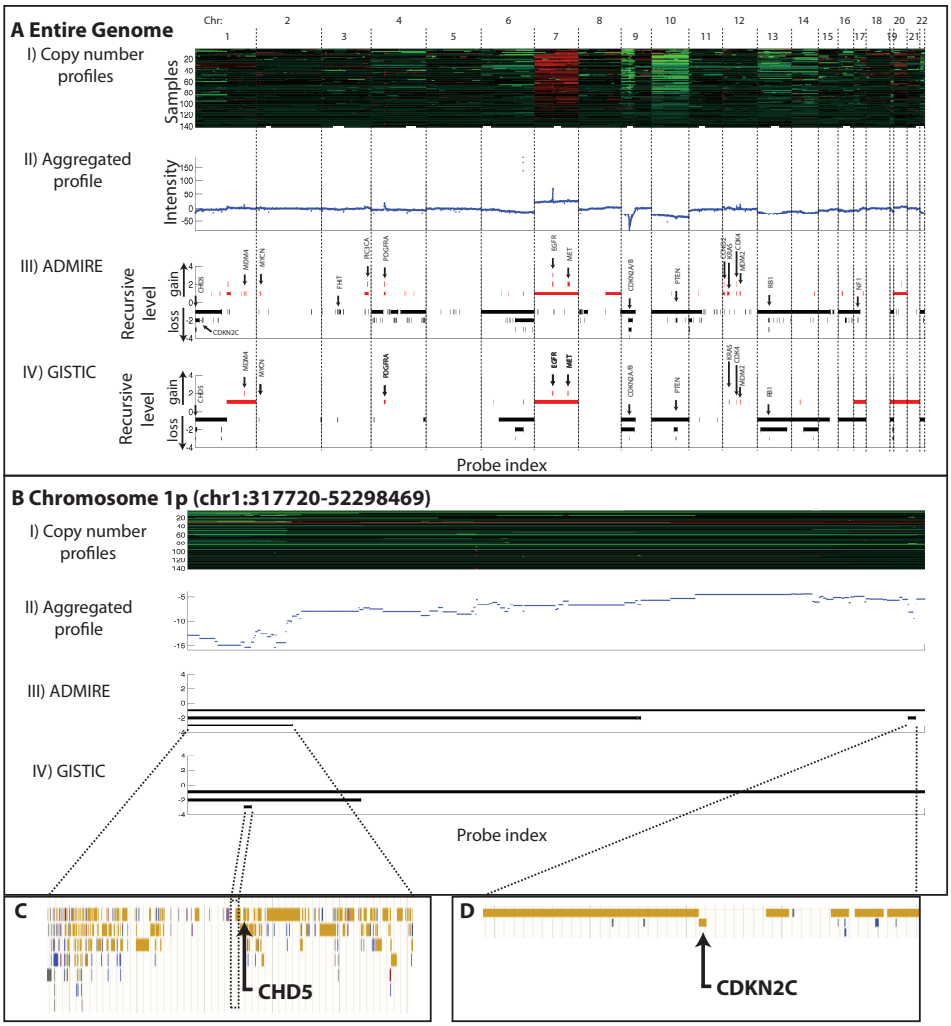

Figure 3.8: Comparison of detected recurring events detected by ADMIRE and GISTIC2.0 on the glioma dataset. (**A**) Summary of the recurrent aberrations found by both ADMIRE and GISTIC2.0 on the entire genome. (**A.I**) The SNP array profiles for 141 glioma samples. Red (green) represents amplifications (deletions). (**A.II**) The sum of all the SNP array profiles. (**A.III**) A multi-level representation of the recurring events found by ADMIRE at 25% event-based FDR. The first recursive level shows all the broad and focal events that are not embedded in broad events. The second level shows more focal (or less broad) events embedded in broad first-level events, etc.(**A.IV**) Results found by GISTIC2.0 at 25% probe-based FDR. The first level (+1/-1 for gains or losses, respectively) represents all the broad recurrent events found at the chromosome arm level. After removing segments that stretch across whole chromosome arms, all segments with q-values below 0.25 are represented on the second level. Finally, focal regions are detected using the RegBounder algorithm and represented on the third level. Therefore, red events (positive levels) represent recurring gains (levels move upwards) and black (negative levels) represents deletions (with levels moving downwards). (**B**) A zoom of the result in Panel A, showing the first part of chromosome 1p. (**C**) The top recursive level (most focal) event found by ADMIRE containing the CHD5 gene. It is interesting to note that GISTIC2.0 finds a much more focal area close to CHD5; however, with careful observation of the aggregated profile in (**B.II**) it is obvious that no focal event can be called with high significance by ADMIRE at this point. (**D**) Shows the recurring region found by ADMIRE containing the known glioma tumor suppressor gene CDKN2C that was missed by GISTIC2.0.

searching for relevant genes), we argue that this is of central importance, as one might expect that for every broad event missed, a number of potentially false focal events might be detected in this region simply due to passenger events revealing peaks in an elevated region (shattered events) in the aggregated profile.

We introduced an analytical expression for the expected Euler characteristic, which simply counts up-crossings and not explicitly how long the signal remains above the amplitude threshold (the so-called sojourn time). Intuitively this could present a problem, but ADMIRE solves the problem by using the scale space to automatically tune the power to match the aberration width.

We also introduced a method that allows us to control the FDR (based on the expected Euler characteristic) without resorting to time-consuming permutation tests. We are therefore able to perform complex procedures, such as updating the null-process parameters in the recursive multi-level detection scheme, within a realistic time frame.

The methodology is justified from a theoretical perspective and justified with empirical simulations. Also when we test the method on a glioblastoma dataset, we find many more potentially interesting recurrent events (including two known glioma tumor suppressors CDKN2C and NF1) that approximately form a superset of those found by GISTIC2.0. Note that ADMIRE does not make a binary distinction between broad and focal events since multiple levels of increasingly focal events are derived from the data.

On a final note, the amount of primary memory used by ADMIRE depends on the probe locations and the minimum kernel width specified. If the whole human genome is covered with probes (say 3 million or more probes) and the minimum kernel width specified is 1 kb, the maximum memory usage will be 2 GB, which might be smaller than the dataset itself. Computation time is largely influenced by the number of recurrent aberrations detected, which might take up to 8 h on an Intel Core i7-950 processor for a dataset consisting of 3 million probes, 200 samples and 200 recurrent events.

3.5. AVAILABILITY

ADMIRE can be downloaded at http://bioinformatics.nki.nl/admire/ . This includes a zipped file with the required Matlab code and glioma dataset.

3.6. FUNDING

This work was part of the BioRange programme of the Netherlands Bioinformatics Centre (NBIC), which is supported by the Netherlands Genomics Initiative (NGI). Funding for open access charge: The Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands.

Conflict of interest statement . None declared.

3.7. ACKNOWLEDGEMENTS

The authors thank Jeroen de Ridder and Theo Knijnenburg for valuable discussions regarding different aspects of the work. Special thanks go to Guillem Rigaill for checking and providing invaluable insight in many theoretical parts

3.8. SUPPLEMENTARY METHODS

3.8.1. ANALYTICAL EXPRESSION FOR THE EULER CHARACTERISTIC

In this section we derive an analytical expression for the expected number of events ($E[\chi]$) found above a fixed threshold and kernel width (see Eq. 3.11 in the main text) for a suitably regular (as defined later) non-homogenous Gaussian (multi-variate) random process $H(g)$ in a compact domain $a = \{g : b_l \leq g \leq b_r\}$, where b_l and b_r are constant boundaries (for example, the boundaries of a chromosome arm). Furthermore, we assume that $H(g)$ has mean zero and variance one at any given location g. Therefore the random process is uniquely defined by the non-homogenous correlation function $r(g_1, g_2)$ or alternatively $r_g(\Delta g) = r(g, g + \Delta g)$. From now on, any realization of the random process $H(g)$ will be represented with a lowercase $h(g)$.

The null process derived in Eq. 3.9 meets the required criterion. However, as will be shown later we do not need an explicit solution for $r_g(\Delta g)$. Instead, the same information is captured by the expected variation in the derivative of $H(g)$ (see Eq. 3.12).

Much work has been done on finding the expected number of events above a fixed threshold for homogeneous Gaussian random fields of any dimensionality [11, 12]. Little work has been done on treating non-homogenous fields except for work done in neuro-imaging [14]. Unfortunately, this work is not directly compatible with our application and therefore we derive the necessary equation to relate $E[\chi]$ with the desired threshold.

To start, we defined the excursion set a^+ of any realization $h(g)$ as the region (a subset of a) where $h(g)$ is higher than or equal to the fixed threshold t:

$$a^+(h,t) \quad = \quad \{g \in a : h(g) \geq t\} \tag{3.15}$$

We define the 'number of events' in a to be the number of maximally connected subsets (ordered by inclusion) of a^+ and denote it with the one dimensional Euler characteristic $\chi(h,t,a)$. It is clear that $\chi(h,t,a)$ is related to the number of up-crossings on t if $h(g)$ is continuous. If, for any sample function $h(g)$, $h(b_l)$ is smaller than t (on the left boundary), the Euler characteristic is equal to the number of up-crossings. On the other hand if $h(b_l)$ is higher than or equal to the threshold the Euler characteristic will be the number of up-crossings plus one. We denote the number of up-crossings with $\chi_{\mathrm{DT}}(h,t,a)$ since this is exactly the differential topology (DT) characteristic for a one dimensional function [12]. $\chi(h,t,a)$ and $\chi_{\mathrm{DT}}(h,t,a)$ are related as follows:

$$\chi(h,t,a) \quad = \quad I(h(b_l) - t) + \chi_{\mathrm{DT}}(h,t,a), \tag{3.16}$$

where

$$I(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \tag{3.17}$$

We are interested in evaluating the expected Euler characteristic:

$$\begin{aligned} E[\chi(H,t,a)] &= P[H(b_l) \geq t] + E[\chi_{\mathrm{DT}}(H,t,a)] \\ &= \frac{1}{2}\mathrm{erfc}(\frac{t}{\sqrt{2}}) + E[\chi_{\mathrm{DT}}(H,t,a)] \end{aligned} \tag{3.18}$$

SUITABLY REGULAR PROCESSES

We denote the first and second order derivatives (if they exist) of the process $H(g)$ by $H'(g)$ and $H''(g)$ respectively. We define a random process to be suitably regular on a (for a fixed threshold t) if the following conditions hold as described by [11] and adapted for our one dimensional application:

1. A sample function $h(g)$ has a.s. (almost surely) continuous derivatives up to second order with finite variance in an open neigbourhood of a. This condition is satisfied if we use kernels for smoothing with the same requirements (such as a Gaussian kernel).
2. The set $\{g \in a : h(g) = t \wedge h'(g) = 0\}$ is a.s. empty. In other words the probability density function $p[H'(g)|H(g) = t]$ should be finite at zero and the set $\{g \in a : h(g) = t\}$ a.s. finite.
3. The set $\{g \in \{b_l, b_r\} : h(g) = t\}$ is a.s. empty.

A random process is therefore considered to be suitably regular if any realization h is a.s. sufficiently smooth (ensured by the kernel convolution) and consist of a finite number of points where t are crossed, the derivative being non zero at each point.

First we observe that a suitably regular Gaussian random process has a differentiable correlation function $r_g(\Delta g)$ with respect to Δg up to second order. We also observe that:

$$\begin{aligned}
\mathrm{Var}[\frac{d}{dg}H(g)] &= \lim_{\Delta g \to 0} \mathrm{Var}[\frac{H(g+\Delta g) - H(g)}{\Delta g}] \\
&= \lim_{\Delta g \to 0} \frac{1}{(\Delta g)^2} E\big[\big(H(g+\Delta g) - H(g)\big)^2\big] \\
&= \lim_{\Delta g \to 0} \frac{2\big(1 - r_g(\Delta g)\big)}{(\Delta g)^2} \\
&= -\lim_{\Delta g \to 0} \frac{r'_g(\Delta g)}{\Delta g} \\
&= -\lim_{\Delta g \to 0} r''_g(\Delta g), \qquad (3.19)
\end{aligned}$$

where r'_g and r''_g represent the first and second order derivative r_g with respect to Δg. In the last two steps we consecutively applied L'Hôpital's rule.

Therefore, r_g comply with the following properties:

$$\begin{aligned}
r_g(0) &= 1 \\
r'_g(0) &= 0 \\
r''_g(0) &= -\mathrm{Var}[\frac{d}{dg}H(g)] \qquad (3.20)
\end{aligned}$$

THE EXPECTED NUMBER OF UP-CROSSINGS

Before we derive an analytical expression for the expected number of up-crossing, lets first derive a few useful lemmas.

Let us define a function $q : \mathbb{F} \to \mathbb{R}$:

$$q(r) = \lim_{\Delta g \to 0^+} \frac{(k \circ r)(\Delta g)}{\Delta g}, \qquad (3.21)$$

where

$$\begin{aligned}
k(r) &= \frac{1}{2\pi|\Sigma(r)|^{1/2}} \times \\
&\quad \int_t^\infty \int_{-\infty}^t e^{-\frac{1}{2}[x,y]\Sigma^{-1}(r)[x,y]^T} dxdy \\
\Sigma(r) &= \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix} \qquad (3.22)
\end{aligned}$$

and $\mathbb{F}$ is the set of all functions r for which the limit is finite.

Lemma 1. *Consider a suitably regular random Gaussian process H with $\forall g \in \mathbb{R}\ H(g) \sim N(0,1)$ and correlation function $r_g(\Delta g) = r(g, g+\Delta g)$. If $\forall g \in a$, $r_g \in \mathbb{F}$, then*

$$E[\chi_{DT}(H,t,a)] = \int_{b_l}^{b_r} q(r_g)dg \tag{3.23}$$

Proof. The DT characteristic is simply the number of points $g \in a$ satisfying the following conditions:

1. $h(g) = t$
2. $h'(g) > 0$

Note that the number of points satisfying $h(g) = t$ for $g \in a$ representing up and down crossings (say $c_1 < c_2 < ... < c_k$) are a.s. finite and distinct due to the regularity conditions. Let us define grid points via sequences $a^n : 2^n \to a$ such that:

$$\begin{aligned} a^n(0) &= b_l \\ a^n(i+1) &= a^n(i) + \Delta g^n, \end{aligned} \tag{3.24}$$

where $\Delta g^n = \dfrac{b_r - b_l}{2^n}$ and denote the set of grid points by $\{a^n\} = \{a^n(i) | i < 2^n\}$. It is easy to prove by induction that for any $k, l \in \mathbb{N}, k < l$ implies that $\{a^k\} \subset \{a^l\}$.

Next, let us define the infinite sequence $\langle \chi_n | n < \mathbb{N} \rangle$, with elements

$$\chi_n = \sum_{i=0}^{n-1} f^n\big(h,t,a^n(i)\big)\Delta g^n, \tag{3.25}$$

where

$$f^n(h,t,g) = \begin{cases} \dfrac{1}{\Delta g^n} & \text{if } h(g) < t \le h(g+\Delta g^n) \\ 0 & \text{otherwise} \end{cases} \tag{3.26}$$

It is rather easy to show that the sequence $\langle \chi_n | n < \mathbb{N} \rangle$ is:

- increasing;
- bounded from above by $\chi_{\mathrm{DT}}(h,t,a)$;
- there a.s. exist an $n \in \mathbb{N}$ for which $\chi_n = \chi_{\mathrm{DT}}(h,t,a)$, since there exist a.s. only a finite number of crossings;

Therefore we conclude that:

$$\begin{aligned} \lim_{n\to\infty} \chi_n &= \lim_{n\to\infty} \sum_{i=0}^{n-1} f^n\big(h,t,a^n(i)\big)\Delta g^n \\ &= \chi_{\mathrm{DT}}(h,t,a) \end{aligned} \tag{3.27}$$

Now let us consider the expected value of and given χ_n:

$$E[\chi_n] = \sum_{i=0}^{n-1} E[f^n\big(H,t,a^n(i)\big)]\Delta g^n, \tag{3.28}$$

It is clear at this point that the sequence $\langle E[\chi_n] | n < \mathbb{N} \rangle$ will converge to $E[\chi_{\mathrm{DT}}(h,t,a)]$ since, for any realization of the process, $\langle \chi_n | n < \mathbb{N} \rangle$ converges to $\chi_{\mathrm{DT}}(h,t,a)$. Therefore we set out to compute

$$\begin{aligned} E[\chi_{\mathrm{DT}}(h,t,a)] &= \lim_{n\to\infty} E[\chi_n] \\ &= \lim_{n\to\infty} \sum_{i=0}^{n-1} E[f^n\big(H,t,a^n(i)\big)]\Delta g^n \\ &= \int_{b_l}^{b_r} \lim_{n\to\infty} E[f^n\big(H,t,g\big)]dg \end{aligned} \tag{3.29}$$

if the limit inside the integral is finite for all $g \in a$. Therefore the problem is reduced to finding the expectation of $f^n\big(H,t,g\big)$, which is fully determined by the joint probability density function of $x = h(g)$ and $y = h(g+\Delta g^n)$ with correlation $r = r_g(\Delta g^n)$ (see Eq. 3.26):

$$p(x,y) = \frac{1}{2\pi|\Sigma(r)|^{1/2}} e^{-\frac{1}{2}[x,y]\Sigma(r)^{-1}[x,y]^T}, \tag{3.30}$$

where

$$\Sigma(r) = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix} \tag{3.31}$$

From Eq. 3.26 we get:

$$\begin{aligned} E[f^n(H,t,g)] &= \frac{1}{\Delta g^n} P[x < t, y \geq t] \\ &= \frac{(k \circ r_g)(\Delta g^n)}{\Delta g^n}, \end{aligned} \tag{3.32}$$

We are interested in finding an expression for (if it exists):

$$q(r_g) = \lim_{n\to\infty} E[f^n(H,t,g)] = \lim_{\Delta g \to 0^+} \frac{(k \circ r_g)(\Delta g)}{\Delta g} \tag{3.33}$$

This concludes the proof of lemma 1. □

Lemma 2. *Consider two functions r_1 and r_2 that are differentiable up to the second order in open intervals $(0,\delta_1)$ and $(0,\delta_2)$ respectively, where $\delta_1, \delta_2 > 0$. Let us assume that $r_1 \in \mathbb{F}$ and that :*

$$\begin{aligned} \lim_{g\to 0^+} r_1(g) &= \lim_{g\to 0^+} r_2(g) = 1 \\ \lim_{g\to 0^+} r_1'(g) &= \lim_{g\to 0^+} r_2'(g) = 0 \\ \lim_{g\to 0^+} r_1''(g) &= \lim_{g\to 0^+} r_2''(g) < 0 \end{aligned} \tag{3.34}$$

If these conditions hold, then $r_2 \in \mathbb{F}$ and $q(r_1) = q(r_2)$.

Proof. Note that there exist δ_3 and δ_4 such that $0 < \delta_3, \delta_4 < \min(\delta_1, \delta_2)$ such that r_1 and r_2 have negative first and second order derivatives in the intervals $(0,\delta_3)$ and $(0,\delta_4)$ respectively and furthermore the images are equal $r_1[(0,\delta_3)] = r_2[(0,\delta_4)] = (\epsilon, 1)$ for some $\epsilon \in (0,1)$. Since r_2 is decreasing in the interval $(0,\delta_4)$, its inverse exists and we define a function with domain $(0,\delta_3)$ and range $(0,\delta_4)$:

$$w(g) = (r_2^{-1} \circ r_1)(g) \tag{3.35}$$

First note that $w(g)$ is an increasing function, since both r_1 and r_2^{-1} are decreasing. Also, as expected:

$$\begin{aligned}\lim_{g\to 0^+} w(g) &= \lim_{g\to 0^+} r_2^{-1}(r_1(g)) \\ &= \lim_{r_1\to 1^-} r_2^{-1}(r_1) \\ &= 0 \end{aligned} \tag{3.36}$$

Furthermore, for $g \in (0,\delta_3)$:

$$\frac{dw}{dg} = \frac{\frac{dr_1}{dg}}{\frac{dr_2}{dw}} \tag{3.37}$$

Note that the denominator ($\frac{dr_2}{dw}$) and its derivative (with respect to g) is strictly non-zero in the interval $g \in (0,\delta_3)$ since r_2 have negative first and second order derivatives. Also note that $\lim_{g\to 0^+} \frac{dr_1}{dg} = \lim_{g\to 0^+} \frac{dr_2}{dw} = \lim_{w\to 0^+} \frac{dr_2}{dw} = 0$ (See Eq. 3.34). Therefore we can apply L'Hôpital:

$$\begin{aligned}\lim_{g\to 0^+} \frac{dw}{dg} &= \lim_{g\to 0^+} \frac{\frac{d^2r_1}{dg^2}}{\frac{d^2r_2}{dw^2}\frac{dw}{dg}} \\ &= 1/\lim_{g\to 0^+} \frac{dw}{dg} \\ \therefore \lim_{g\to 0^+} \frac{dw}{dg} &= 1 \end{aligned} \tag{3.38}$$

Where we noticed that $\lim_{g\to 0^+} \frac{d^2r_1}{dg^2} = \lim_{g\to 0^+} \frac{d^2r_2}{dw^2} = \lim_{w\to 0^+} \frac{d^2r_2}{dw^2} < 0$ (see Eq. 3.34).

To summarize,

- w has domain $(0,\delta_3)$ and range $(0,\delta_4)$;
- w is an increasing function with first and second order derivatives;
- $\lim_{g\to 0^+} w(g) = 0$;
- $\lim_{g\to 0^+} w'(g) = 1$;

On a different note, k is clearly differentiable in in the domain $(\epsilon, 1)$ since Σ is positive definite.

Now we have everything we need to prove the lemma. For any function r_2 (including r_1) satisfying Eq. 3.34:

$$\lim_{g\to 0^+} (k \circ r_2)(g) = 0 \tag{3.39}$$

since $r_2 \to 1^-$ and Σ becomes singular. Therefore by L'Hôpital:

$$q(r_2) = \lim_{g\to 0^+} \frac{dk}{dr_2}\frac{dr_2}{dg} \tag{3.40}$$

if the limit exist.

We know that $q(r_1)$ is finite (since $r_1 \in \mathbb{F}$) and again by L'Hôpital:

$$\begin{aligned} q(r_1) &= \lim_{g\to 0^+} \frac{(k\circ r_2 \circ w)(g)}{g} \\ &= \lim_{g\to 0^+} \frac{dk}{dr_2}\frac{dr_2}{dw}\frac{dw}{dg} \\ &= \lim_{g\to 0^+} (\frac{dk}{dr_2}\frac{dr_2}{dw}) \lim_{g\to 0^+} \frac{dw}{dg} \\ &= \lim_{w\to 0^+} (\frac{dk}{dr_2}\frac{dr_2}{dw}) \\ &= q(r_2) \end{aligned} \tag{3.41}$$

which concludes the the proof for lemma 2. □

Lemma 3. *For a stationary correlation function r:*

$$q(r) = \frac{e^{-t^2/2}}{2\pi}\sqrt{-r''(0)} \tag{3.42}$$

Proof. Consider a suitably regular stationary random Gaussian process H with correlation function r and $\forall g \in \mathbb{R}\ H(g) \sim N(0,1)$. Lemma 1 states that if $q(r)$ is finite then

$$\begin{aligned} E[\chi_{DT}(H,t,[0,1])] &= \int_0^1 q(r)dg \\ &= q(r) \end{aligned} \tag{3.43}$$

In the second step we used the fact that r is constant for all g in a stationary process.

Also, by Eq. 3.2 of Theorem 3.1 in [11] when applied to a one dimensional field:

$$E[\chi_{DT}(H,t,[0,1])] = \frac{e^{-t^2/2}}{2\pi}\sqrt{-r''(0)} \tag{3.44}$$

□

Theorem 1. *Consider a suitably regular (non-stationary) random Gaussian process H with* $\forall g \in \mathbb{R}\ H(g) \sim N(0,1)$ *and correlation function* $r_g(\Delta g) = r(g, g+\Delta g)$. *Then*

$$E[\chi_{DT}(H,t,a)] = \frac{e^{-t^2/2}}{2\pi}\int_{b_l}^{b_r} \sqrt{Var[\frac{d}{dg}H(g)]}dg \tag{3.45}$$

Proof. According to Lemma 1, we simply need to find an expression for $q(r_g)$ for all g. The properties specified in Eq. 3.20 hold for all r_g. Certainly for each such r_g there exist a stationary r_g^s with the same properties. By Lemma 2 and 3:

$$\begin{aligned} q(r_g) &= q(r_g^s) \\ &= \frac{e^{-t^2/2}}{2\pi}\sqrt{-r_g^{s\prime\prime}(0)} \\ &= \frac{e^{-t^2/2}}{2\pi}\sqrt{-r_g{}''(0)} \\ &= \frac{e^{-t^2/2}}{2\pi}\sqrt{\mathrm{Var}[\frac{d}{dg}H(g)]} \end{aligned} \tag{3.46}$$

□

3.8.2. DETAILS ON MULTI-SCALE DETECTION

RESOLUTION PARAMETER α:

The filter parameter α is related to the minimum possible spatial overlap between a detected event D and a real recurrent event R, which we indicate with the similarity coefficient J_1:

$$J_1(D,R) = \frac{|D \cap R|}{|D|} \tag{3.47}$$

To see this, consider for example a Gaussian kernel truncated at $\pm 3w$ (the kernel weights after $\pm 3w$ are negligible and can safely be ignored for computational purposes) and a true recurrent event R with unknown boundaries g_s and g_e. Clearly at a fixed scale w the smoothed signal is influenced by the recurrent event on the interval $[g_s - 3w, g_e + 3w]$. Therefore if we detect a recurrent event larger than αw we are guaranteed that $J_1(D,R) \geq \frac{\alpha-6}{\alpha} 100\%$ unless a false positive occurs. For a symmetric kernel that is decreasing with distance from the center, the minimum overlap is restricted to $\frac{\alpha-2\kappa}{\alpha} 100\%$, where we choose κ such that:

$$\frac{\int_{\kappa w}^{\infty} k_w(g)dg}{\int_{-\infty}^{\infty} k_w(g)dg} \ll 1 \tag{3.48}$$

The idea is simply to keep the kernel weights low at a distance κw ($\kappa = 3$ for a Gaussian kernel) so that the contribution from a significant recurrence at that distance is negligible in the null region.

Small α means that the aberration widths are allowed to be comparable in size to the kernel width. Fig. 3.6 in the main text shows that the power increases when we allow for the kernel width to be similar in size of the aberration. However, the resolution (J_1) decreases.

If we let $\alpha \to \infty$ we gain good resolution, but the power would be equivalent to that in a single scale analysis (at the smallest scale). As a compromise we choose $\alpha = 20$ that guarantees a minimum overlap J_1 of 70% (a parameter that the user can vary). In the final step of the multi-scale selection we take the union of all recurrent events surviving the α filtering and as a consequence all maximally connected regions resulting from the union also have a minimum overlap of 70% (or whatever specified by the user) with true events. If the smallest kernel width considered is sufficiently small, we do not particularly care whether this overlap is high since we expect the detected event to be sufficiently zoomed into potentially interesting driver genes.

Usually, it is not a good idea to sacrifice power for a high resolution by letting $\alpha \to \infty$ (single scale analysis), because large events might be shattered into smaller pieces as illustrated in Fig. 3.2B. Although shattered events are strictly not false positives, we might employ a more strict overlap condition such as the Jaccard similarity coefficient to call true positives:

$$J_2(D,R) = \frac{|D \cap R|}{|D \cup R|} \tag{3.49}$$

For a small shattered event D, $J_1 = 1.0$ and J_2 would be small. We therefore define two different types of undesirable events for an overlap threshold o (e.g. 70%):

- False positives: Events for which $J_1 < o$;
- Shattered positives: Events for which $J_1 \geq o$ and $J_2 < o$;

For a fixed threshold o the shattered positive rate will decrease for a decreasing value of α, whereas the FDR will remain unchanged (for false positives) as long as α remains large enough to ensure a high J_1.

Simulation results where we vary the α parameter is illustrated in Section 3.9.1. Here we also observe the effect that α has on the shattered positive rate.

THE EXPECTED NUMBER OF EVENTS

It is important to ask what effect the proposed multi-scale detection has on the expected number of significant regions (from now on referred to as $E[\chi]$) found in the null-hypothesis if we keep $E[\chi_{\text{theory}}] = E[\chi(H^0_w, t)]$ constant and the same on each scale. First note that if we consider a very small kernel (say 1 kbp) then we are performing almost no smoothing and any region above a fixed threshold is considered significant with a high resolution. In this case our estimate on $E[\chi(H^0_{1\text{kbp}}, t)]$ will be accurate (since we don't remove any segments that don't survive the $20w$ filter rule). On the other hand, if for larger kernels, we only retain regions that are at least $20w$ in size, our estimate $E[\chi(H^0_w, t)]$ will be much higher then the actual expectation on that particular scale. For any region $a = [g_s, g_e]$ let us define χ_a to be the total number of events found in $[g_s, g_e]$ on the smallest kernel considered. Furthermore, let $\mathbf{G}(a)$ be the property 'a is a detected event for some $H_w(g)$ and survives the $20w$ filtering rule'. We make the following assumption on the null-hypothesis:

$$P[\chi_a > 1 | \mathbf{G}(a)] \quad > \quad P[\chi_a = 0 | \mathbf{G}(a)] \tag{3.50}$$

In effect we assume that if there exist a large region that is considered significantly elevated, then we expect the chance to be higher for this region to contain multiple significant 'hits' on a small scale then no 'hits' at all.

If this assumption holds, then clearly the multi-scale procedure will more likely merge events on the smallest possible scale than create new ones on a larger scale, and since we control the expected number of events on the smallest scale at $E[\chi_{\text{theory}}]$, we expect:

$$E[\chi] \quad < \quad E[\chi_{\text{theory}}] \tag{3.51}$$

and therefore have $E[\chi]$ under control.

3.8.3. DETAILS ON UPDATING THE NULL-PARAMETERS

Consider a smoothed aggregated profile with no recurrent events which can be modeled with a Gaussian random process with parameters μ, σ and $\mathbf{r}$. In the main text we proposed a methodology to iteratively update these null-parameters by detecting peaks that surpass a given threshold t and re-estimating the parameters by ignoring these significant regions. Even with no recurrent events there is always a possibility of detecting false positives and consequently an updated estimate on σ will be biased (and generate optimistic results) after one or more iteration. We set out to prove that this bias (assuming there are no recurrent events) will converge after infinite iterations and that the effect on the estimated number of false events ($E[\chi]$) will be negligible for reasonable thresholds t. If recurrent events are present, σ will be a conservative estimate and therefore we consider only the worst case scenario with no recurrences.

Consider the variance of the random Gaussian process σ and the respective threshold t to control $E[\chi]$. Furthermore, assume that t_i is the expected threshold on iteration i. If there are no recurrent events (unknown for real data), $t_1 = t$, the desired threshold. On the second iteration, all measurements larger than t_1 will be ignored when calculating a new variance estimate and will therefore be biased. In fact, we can predict how t_i changes across each iteration:

$$t_{i+1} \quad = \quad t\sqrt{1 - \text{erfc}(\frac{t_i}{\sqrt{2}\sigma}) - \sqrt{\frac{2}{\pi}}\frac{t_i}{\sigma}e^{-\frac{1}{2}(\frac{t_i}{\sigma})^2}} \tag{3.52}$$

Using Eq. 3.11 we can also predict how the estimate on $E[\chi]$ increase with each iteration. Therefore, for a fixed starting threshold t we can predict the ratio between the predicted $E[\chi]$ after infi-

nite iterations and the $E[\chi]$ for the desired threshold t:

$$R(t/\sigma) \quad = \quad e^{\frac{1}{2}[(\frac{t}{\sigma})^2-(\frac{t_\infty}{\sigma})^2]} \tag{3.53}$$

R is a decreasing function of t strictly larger than one that only converge for values $t/\sigma > 2.1617$. In practice we are typically interested in thresholds $t/\sigma > 3$ and therefore $R < 1.1632$ (for a typical value $t/\sigma = 4, R = 1.0092$). This illustrates that the iterative procedure only leads to a slight overestimation of $E[\chi]$ by a factor of maximally 1.1632 (if no recurrent events are present). In real data, where many recurrent events are present, our estimate on $E[\chi]$ will likely remain conservative.

3.8.4. DETAILS ON RECURSIVE MULTI-LEVEL DETECTION

Not only does the recursive multi-level detection procedure described in the text allow us to detect recurring events embedded in broad recurring events, but also helps to improve our estimate on $E[\chi]$ (the expected number of false recurring events found) when not all parts of the genome (the recurring events) can be described by the null-hypothesis. We illustrate this concept in Fig. 3.3 in the main text. Note that the region in which we estimate the null-parameters μ, σ and $\mathbf{r}$ is restricted to H^0_{L1} in Fig. 3.3A as illustrated by the dotted line at the top of the figure. However in Eq. 3.11 (in the main text) we integrate across the whole genome and therefore the expected number of detected events in H^0_{L1} will be lower than $E[\chi]$ (the expected number of events if the null-model held across the whole genome) by a factor $\int_{H^0_{L1}} / \int_G$ (the erfc term is small). In the second recursion step (Fig. 3.3B) we follow the exact same procedure except this time estimate the null-parameters in the broad event H^0_{L2}. This allows us to detect embedded focal events inside broader events. Again we make sure the test is genome wide (although we only estimate and detect events in H^0_{L2} we still integrate across the whole genome in Eq. 3.11 in the main text) and therefore the expected number of null-events found inside H^0_{L2} will differ from $E[\chi]$ (the expected number of events if the null-model held across the whole genome for the new parameters) by a factor $\int_{H^0_{L2}} / \int_G$. The total number of expected random events found in both recursion steps will be $\frac{\int_{H^0_{L1}} + \int_{H^0_{L2}}}{\int_G} E[\chi]$. In fact, if we consider all recursion steps the expected number of falsely detected events (focal, broad and embedded) will approach $E[\chi]$ and we effectively avoid the need for step up/down multiple testing procedures. The only added assumption here is that the expected number of false focal events discovered inside false broad events is insignificant.

3.9. SUPPLEMENTARY RESULTS

3.9.1. RESOLUTION PARAMETER α ON SIMULATED DATA

SUMMARY

We investigated the effect that the resolution parameter α (which is related to the accuracy of event boundaries) has on the FDR and the power for detecting recurrent aberrations of different genomic lengths. We generated simulated data using the methodology proposed in Section 3.3.4. Specifically, we added $N_b = 2$ broad (20×10^6 bps), $N_m = 5$ (2×10^6 bps) medium and $N_f = 50$ (100×10^3 bps, with an average of five probes) focal recurrent events. For each simulation, the recurrent events were placed at random locations (potentially overlapping) with a recurrence frequency (across the samples) randomly selected between zero and one. Passenger events and noise were simulated using the procedure described in Section 3.3.1. We fixed the number of aCGH samples to aggregate to $S = 200$ and the analytical FDR level to 5%. We vary the SNR and α parameter. Finally, we distinguish between true and false positives based on an overlap threshold $o = 70\%$

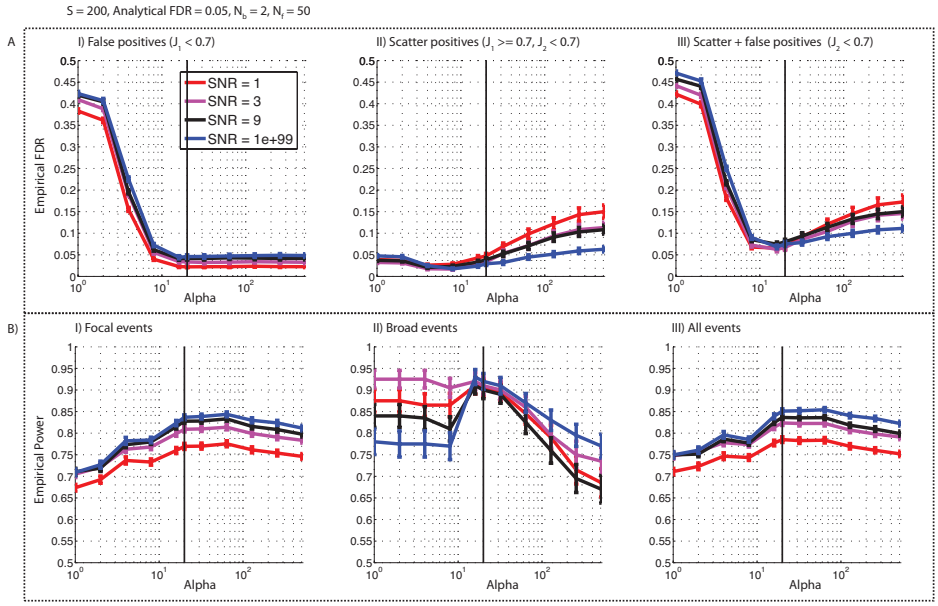

Figure 3.9: Simulation results illustrating the effect that the resolution parameter α (on x-axis of each plot) has on FDR and power for a simulated case study. Simulations were performed for SNR $\in \{1, 3, 9, 1e99\}$ and repeated 1000 times. In each experiment we add two broad, five medium and 50 focal recurrent events. In each plot, the vertical black line indicates $\alpha = 20$. (**A.I**) All detected events that are covered by less than 70% ($J_1 < 0.7$) of all true recurrent regions is considered a false positive. The y-axis represents the average ratio of detected events that are false positives (measured FDR). (**A.II**) All detected events that are covered by more than 70% of a true recurrence ($J_1 \geq 0.7$), but with a Jaccard similarity coefficient below 70% is considered a shattered positive. The y-axis represents the average ratio of detected events that are shattered positives. (**A.III**) All detected events that have a Jaccard similarity coefficient below 70% for all recurrent regions are considered a strict false positive. The y-axis represents the average ratio of detected regions that are strict false positives (**B.I**) The y-axis represents the power for detecting focal events with a Jaccard similarity coefficient above 70%. B.II) This is the same as (**B.I**) except that we show the power for detecting broad events. (**B.III**) Here we show the power for detecting all events (focal, medium and broad).

Results

Fig. 3.9 illustrates the empirical FDR and power for detecting recurrent aberrations (on 1000 simulations) of different sizes, for different SNRs and α parameters.True and false positives are discriminated based on an overlap threshold $o = 70\%$. Before we continue it is important to realize that for large α values we are effectively ignoring all scales except the smallest (i.e. we perform little smoothing), since none of the events detected at a larger scales will survive the α filtering. This is in contrast to $\alpha = 0$, where all detected events on all scales are combined.

In Fig. 3.9A.I we measure the expected proportion of detected events for which $J_1 < 70\%$ (false positives). As predicted in Section 3.8.2, for α value above 20, this proportion is below 5% (the FDR). For α values smaller than 20 the accuracy on event boundaries become poor and the measured FDR grows high.

In Fig. 3.9A.II we measure the expected proportion of detected events that are true positives, but $J_2 < 70\%$ (shattered positives). It is important to observe that these errors are reduced when we allow for α to grow small (and effectively use the full scale space). This effect is observed even for segmented samples (SNR = $1e99$). This illustrates that the scale space methodology is not only useful for reducing measurement noise, but also biological noise (i.e. passenger events).

In Fig. 3.9A.III we illustrate the measured FDR if we consider shattered positives to be false. That is, we regard all detected events with a Jaccard similarity index below 70% as false positives. It is interesting to note that for $\alpha = 20$, the FDR is close to minimal, but still higher than the desired FDR level of 5%. This is because the shattered positive rate cannot be controlled, but reduced due to the multi-scale analysis - an aspect ignored by single-scale methods. Furthermore, it is not surprising that $\alpha = 20$ is close to the minimum FDR, since we chose a corresponding overlap threshold of $o = 70\%$.

In Fig. 3.9B.I we illustrate the average proportion of recurrent focal events (of which there are 50 per simulation) that are detected (Jaccard similarity coefficient above 70%) while varying α. Note that for small α values we lose power. This is mainly due to low precision on event boundaries. For large α values we see a minimal reduction in power. This is because little value is added by the scale-space for such small events, since they are detected at small scales only.

In Fig. 3.9B.II we illustrate the power for detecting broad events. Again for low α values we lose considerable power due to low precision on event boundaries. However, for very large α values we also see a drastic decrease in power. This is mainly due to scattering and the multi-scale procedure becomes invaluable.

Finally, in Fig. 3.9B.III we illustrate the power for detecting all events (focal and broad included). The power observed is similar to that in Fig. 3.9B.I since most events are focal per simulation (50 as apposed to two).

Discussion

Not only do we reduce the shattered positive rate with the multi-scale procedure if we select α properly ($\alpha = 20$ for an overlap of 70%), but we also increase the power for detecting true positives (especially broad events). This is true even for segmented samples, which illustrates that smoothing not only reduces measurement noise but also biological noise (passenger events).

Furthermore, in Fig. 3.5C and D, we see that the multi-scale procedure reduces our estimate on $E[\chi]$ which will result in a conservative FDR estimate. However, this does not imply that the power for detecting recurrent events are reduced compared to a single-scale analysis (since we join events on all scales, you cannot lose events found at the smallest scale). This simulation study serves to illustrate that the multi-scale procedure (with proper selection of α) can drastically increase the power for detecting events (especially broad events), which makes a conservative estimate on the FDR well worth it.

3.9.2. KC-SMART VS. ADMIRE SMOOTHING METHODOLOGIES

SUMMARY

Both KC-SMART [8] and the newly proposed method performs kernel smoothing on aggregated profiles and applies a constant threshold on the smoothed signal to decide which locations represent recurring aberrations or not. To make these methods directly comparable, we do not split positive and negative aberrations into two problems for KC-SMART.

The new method performs smoothing with Eq. 3.9 (in the main text) and relies on data dependent parameters μ, σ and $\mathbf{r}$ (defined in Eq. 3.3 in the main text) and the platform dependent probe locations, whereas KC-SMART uses only probe locations to normalize the smoothed signal. KC-SMART uses the following normalization scheme:

$$h_{KC}(g) = \frac{f_w(g)}{k_w(g) * \sum_{i=0}^{P-1} \delta(g - p_i)} \tag{3.54}$$

The purpose of this experiment is to show that the new smoothing method increases the power of detecting recurring aberrations if we apply a constant threshold (theoretically justified by Eq. 3.7 in the main text). We do this by simulating 1000 aggregated profiles (with only one recurring aberration) and calculate the proportion of these tests that reveal the recurring aberration with both smoothing methods (new and KC-SMART) when controlling the (two-tailed) FWER at 5%. Since we know the location of the recurring aberration, we do not need to rely on the cyclic-shift null hypotheses and estimate the required threshold from these 1000 simulations to control the FWER. Therefore the power of KC-SMART can be compared directly to the newly proposed method.

SIMULATED DATA

- Genome size: 240 Mbps, $a = \{g | 0 \leq g \leq 120\text{Mbps}\}$;
- Number of probes: 12000;
- Probe positions: Random positions on the genome (each probe has only 1 bp width);
- Number of samples to aggregate: 100;
- Segmentation: Each sample is considered to consist of 160 segments with 159 random breakpoints;
- Segment amplitudes: All probes within a segment takes on a value of either -1, 0 or $+1$ with probabilities 25%, 50% and 25% respectively. Therefore, for each sample we expect 80 aberrated segments (with amplitude -1 or $+1$) that are not recurring;
- Recurring segment center: $c_a = 120$ Mbps;
- Recurring segment widths: We perform parallel experiments with widths (w_a) 40 kbps, 80 kbps, 160 kbps, 320 kbps, 640 kbps, 1.28 Mbps, 2.56 Mbps, 5.12 Mbps and 10.24 Mbps;
- Recurring segment amplitude: $+1$;
- Probability of recurring segment per sample: Each sample (of the 100 samples to aggregate) has a 30% chance of containing a recurring aberration;
- Signal to noise ratio (SNR): For a given SNR, Gaussian noise is added to each sample with mean zero and variance 1/SNR. Parallel experiments are performed at SNRs of 0 (no random or recurring aberrations), 0.1, 1.0, 3.0, 9.0 and 10^{99} (no noise);

EXPERIMENTAL PROCEDURE

For each parallel experiment (for data sets with a given recurring aberration width and SNR) we perform smoothing across multiple scales. We do this for kernel widths w_k = 10 kbps, 20 kbps, 40 kbps, 80 kbps, 160 kbps, 320 kbps, 640 kbps, 1.28 Mbps, 2.56 Mbps, 5.12 Mbps and 10.24 Mbps;

Define the recurring aggregated region to be $a_r = [c_a - w_a/2 - 3w_k, c_a + w_a/2 + 3w_k]$. Define the non-recurring aggregated region to be $a_n = a - a_r$. We perform the following steps.

- STEP 1: Generate an aggregated profile using the methodology proposed earlier;
- STEP 2: Smooth the aggregated profile using a specified kernel width (estimate null parameters from a_n);
- STEP 3: Calculate and store the maximum peak of the smoothed profile in the region a_n and a_r;
- STEP 4: Repeat steps one to three 1000 times. Therefore we will have 1000 maximum peak values for both a_n and a_r;
- STEP 5: Define t to be the 97.5th highest maximum peak in a_n. This will be the thresholds that approximately controls the FWER (two tail) at 5%;
- STEP 6: Define p (the power) to be the proportion of maximum peaks in a_r above t;
- STEP 7: Repeat steps one to six for the two different smoothing methods (KC-SMART and the new method);

RESULTS

Fig. 3.10 shows the power obtained for all combinations of SNR, aberration width and kernel width for KC-SMART and the new smoothing method. The main conclusion from these results are that Eq. 3.9 (in the main text) improves the power for detecting recurring aberrations when compared to KC-SMART for the proposed simulation data. This is especially true for small kernel widths if we observe data with low signal to noise ratios. From these figures we can also observe that changing the kernel width allows us to focus on an aberration with the appropriate size for maximal power.

3.9.3. FWER CONTROL FOR SIMULATED DATA

SUMMARY

This test is very similar to the one proposed in the KC-SMART comparison. The big difference is that we do not use our knowledge of the recurring aberration locations in order to control the FWER. In other words we can only find estimates on the recurrent (a_r) and null (a_n) regions. We treat the method as a black box which outputs a smoothed aggregated profile and a threshold that is analytically determined to control the FWER at 5%.

EXPERIMENTAL PROCEDURE

- STEP 1: Generate an aggregated profile using the methodology proposed in Section 3.9.2;
- STEP 2: Set the null-region estimate equal to the whole genome $\hat{a}_n = a$;
- STEP 3: Estimate null-parameters μ, σ^2 and $\mathbf{r}$ using only probes in $\hat{a}_n$;
- STEP 4: Smooth the aggregated profile using a specified kernel width (w_k) and calculate the analytical threshold t that controls $E[\chi]$ at 2.5% (remember we are performing a two tailed test);
- STEP 5: Set $\hat{a}_r = \{g \in a | h_{w_k}(g) \geq t\}$. Set $\hat{a}_n = a - \hat{a}_r$;
- STEP 6: Repeat steps 3 to 5 until $\hat{a}_n$ converges. Now we have a final estimate for t;

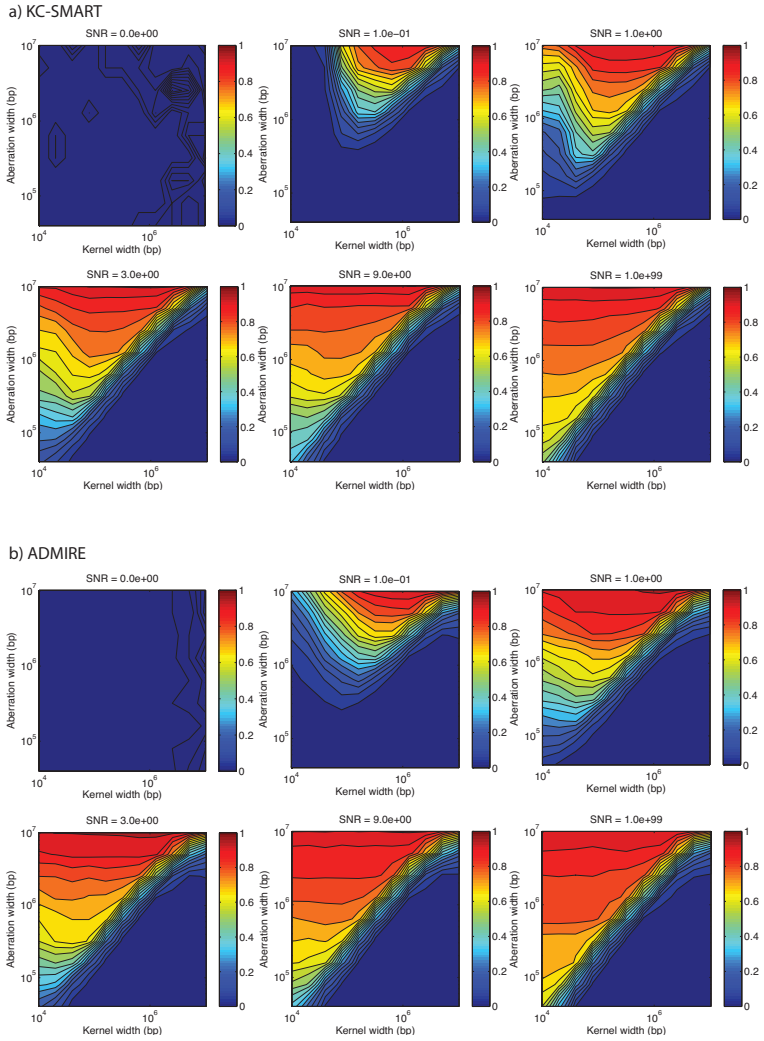

Figure 3.10: KC-SMART and the new method (ADMIRE) are similar in the sense that both perform kernel smoothing on an aggregated profile and then apply a constant threshold to decide whether an aberration is recurring or not. We simulated aCGH profiles with a know recurring aberration as described earlier and test the power of detecting this aberration with a) KC-SMART and b) the new method (ADMIRE) when fixing the FWER at 5%. Each plot represents the power (proportion of recurring aberration detected in 1000 simulations) while varying the aberration width (w_a) and kernel width (k_w). Inspection reveals an increase in power with ADMIRE (especially for small kernels).

- STEP 7: Determine whether any peaks are above t in the non-recurring region a_n and the recurring region a_r;
- STEP 8: Repeat steps one to seven 1000 times;
- STEP 9: Define p (the power) to be the proportion of simulations that reveal one or more peaks in a_r above t;
- STEP 10: Define FWER (empirical) to be the proportion of simulations that reveal one or more peaks in a_n above t;

RESULTS

Fig. 3.11 shows the power obtained for all combinations of SNR, aberration width and kernel width from the proposed smoothing method and the measured FWER (which was controlled analytically).

It is interesting to note that for segmented data (no Gaussian noise on any segments) the analytical FWER control becomes conservative. This is mainly due to the fact that the expected Euler characteristic is an upper bound on the FWER and nearby peaks are highly correlated (since the measurement noise is small). To see why, consider:

$$\begin{aligned} E[\chi] &= P[\chi = 1] + 2P[\chi = 2] + 3P[\chi = 3] + \ldots \\ \text{FWER} &= P[\chi = 1] + P[\chi = 2] + P[\chi = 3] + \ldots \end{aligned} \tag{3.55}$$

If the correlation between probes extend much further than the kernel width (for example, smoothing with a small kernel on segmented samples), we get situations like:

$$P[\chi = 1] \approx P[\chi = 2] \approx P[\chi = 3], \text{etc} \tag{3.56}$$

and therefore the expectation becomes conservative.

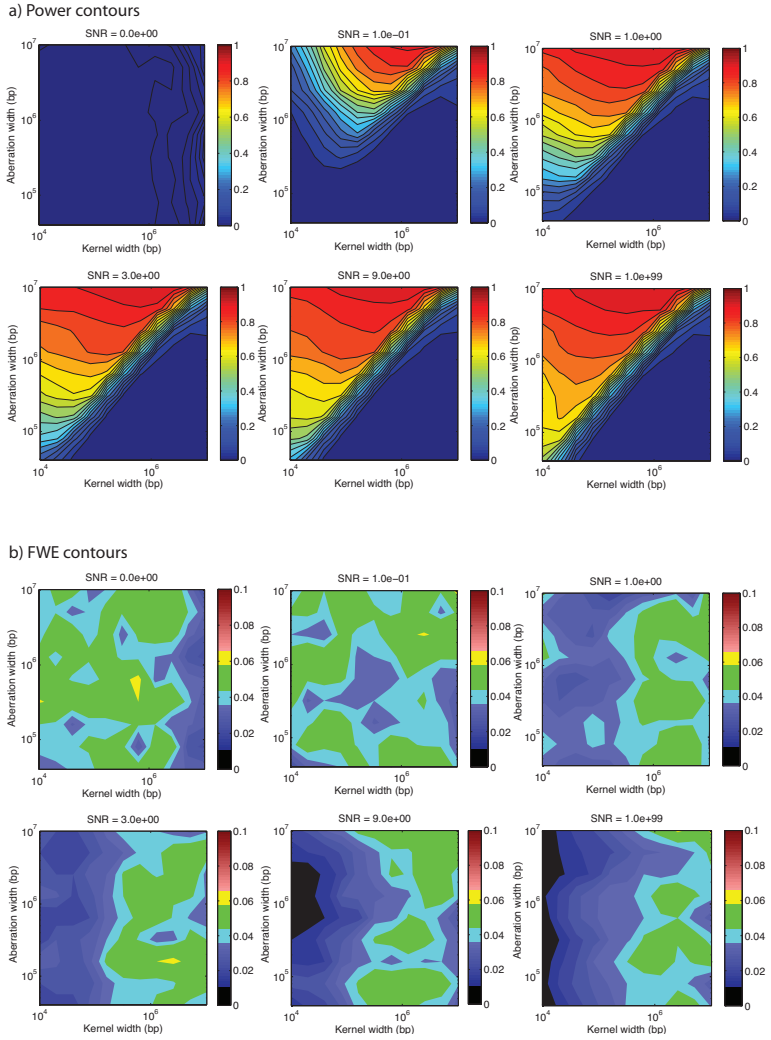

Figure 3.11: In this experiment we again simulate 1000 aggregated profiles. Then for different recurring aberration widths, kernel widths and SNRs we calculate a) the power and b) the estimated FWER. The color bar in the FWER plots should be interpreted as follows: green represents one standard deviation from 5% FWER, cyan and yellow represent FWERs within two standard deviations, blue and red represent extremes below or above two standard deviations and black represents a FWER below 1%.

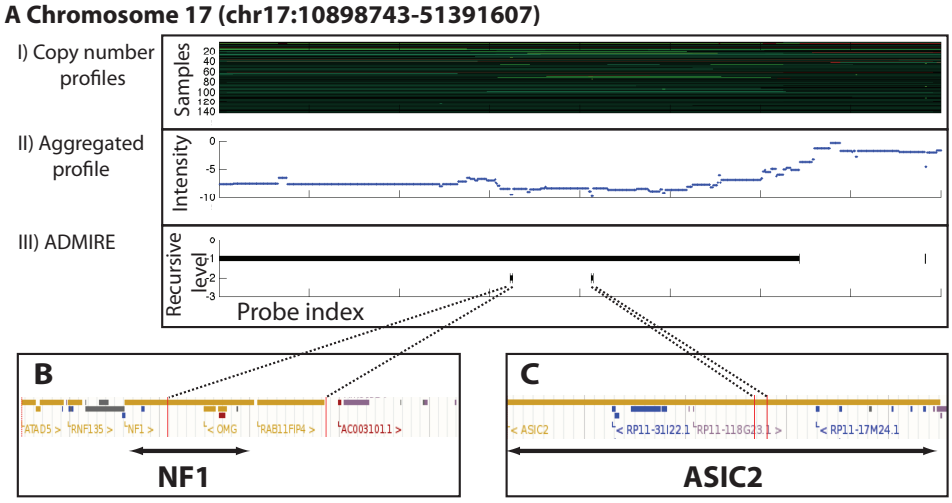

Figure 3.12: Example of ADMIRE on a part of Chromosome 17 where no recurrent aberrations are found by GISTIC2.0. (**A.I**) The SNP-array profiles for 141 Glioma samples. Red (green) represents amplifications (deletions). (**A.II**) The aggregated (sum) of all the samples. (**A.III**) A multi-level representation of the recurring events found by ADMIRE at 25% event-based FDR (across the whole genome). The first recursive level shows a broad deletion. If we re-estimate the null-parameters in this broad event we find two extra focal events. (**B**) Shows that the left most focal event overlaps with the NF1 Glioma tumor suppressor gene. (**C**) The second focal recurrent loss is located in the ASIC2 gene. High-grade glioma tumors are associated with the low functional expression of ASIC2 [19, 20].

REFERENCES

[1] E. van Dyk, M. J. Reinders, and L. F. Wessels, *A scale-space method for detecting recurrent dna copy number changes with analytical false discovery rate control,* Nucleic acids research **41**, e100 (2013).

[2] C. Rouveirol, N. Stransky, P. Hupé, P. La Rosa, E. Viara, E. Barillot, and F. Radvanyi, *Computation of recurrent minimal genomic alterations from array-cgh data,* Bioinformatics **22**, 849 (2006).

[3] B. Taylor, J. Barretina, N. Socci, P. DeCarolis, M. Ladanyi, M. Meyerson, S. Singer, and C. Sander, *Functional copy-number alterations in cancer,* PLoS One **3**, e3179 (2008).

[4] S. Shah, W. Lam, R. Ng, and K. Murphy, *Modeling recurrent dna copy number alterations in array cgh data,* Bioinformatics **23**, i450 (2007).

[5] R. Beroukhim, G. Getz, L. Nghiemphu, J. Barretina, T. Hsueh, D. Linhart, I. Vivanco, J. Lee, J. Huang, S. Alexander, *et al.*, *Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma,* Proceedings of the National Academy of Sciences **104**, 20007 (2007).

[6] C. H. Mermel, S. E. Schumacher, B. Hill, M. L. Meyerson, R. Beroukhim, and G. Getz, *Gistic2. 0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers,* Genome biology **12**, R41 (2011).

[7] F. Sanchez-Garcia, U. Akavia, E. Mozes, and D. Pe'er, *Jistic: identification of significant targets in cancer,* BMC bioinformatics **11**, 189 (2010).

[8] C. Klijn, H. Holstege, J. de Ridder, X. Liu, M. Reinders, J. Jonkers, and L. Wessels, *Identification of cancer genes using a statistical framework for multiexperiment analysis of nondiscretized array cgh data,* Nucleic Acids Research **36**, e13 (2008).

[9] O. M. Rueda and R. Diaz-Uriarte, *Finding recurrent copy number alteration regions: a review of methods,* Current Bioinformatics **5**, 1 (2010).

[10] R. J. Adler and A. M. Hasofer, *Level crossings for random fields,* The Annals of Probability **4**, 1 (1976).

[11] R. J. Adler, *On generalising the notion of upcrossings to random fields,* Advances in Applied Probability **8**, 789 (1976).

[12] K. J. Worsley, *Estimating the number of peaks in a random field using the hadwiger characteristic of excursion sets, with applications to medical images,* The Annals of Statistics , 640 (1995).

[13] K. J. Worsley, S. Marrett, P. Neelin, A. C. Vandal, K. J. Friston, A. C. Evans, *et al.*, *A unified statistical approach for determining significant signals in images of cerebral activation,* Human brain mapping **4**, 58 (1996).

[14] K. Worsley, M. Andermann, T. Koulis, D. MacDonald, and A. Evans, *Detecting changes in nonisotropic images,* Human brain mapping **8**, 98 (1999).

[15] T. Nichols and S. Hayasaka, *Controlling the familywise error rate in functional neuroimaging: a comparative review,* Statistical methods in medical research **12**, 419 (2003).

[16] Y. Benjamini and D. Yekutieli, *The control of the false discovery rate in multiple testing under dependency,* Annals of statistics , 1165 (2001).

[17] Y. Benjamini and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing,* J R Stat Soc B **57**, 289 (1995).

[18] P. Hupé, N. Stransky, J. Thiery, F. Radvanyi, and E. Barillot, *Analysis of array cgh data: from signal ratio to gain and loss of dna regions,* Bioinformatics **20**, 3413 (2004).

[19] B. Berdiev, J. Xia, L. McLean, J. Markert, G. Gillespie, T. Mapstone, A. Naren, B. Jovov, J. Bubien, H. Ji, *et al.*, *Acid-sensing ion channels in malignant gliomas,* Journal of Biological Chemistry **278**, 15023 (2003).

[20] W. Vila-Carriles, Z. Zhou, J. Bubien, C. Fuller, and D. Benos, *Participation of the chaperone hsc70 in the trafficking and functional expression of asic2 in glioma cells,* Journal of Biological Chemistry **282**, 34381 (2007).

4

RUBIC IDENTIFIES DRIVER GENES BY DETECTING RECURRENT DNA COPY NUMBER BREAKS

Ewald van Dyk, Marlous Hoogstraat, Jelle ten Hoeve, Marcel J.T. Reinders & Lodewyk F.A. Wessels

This chapter is published in Nature communications 2016. Volume 7. Page 12159 [1].

The frequent recurrence of copy number aberrations across tumor samples is a reliable hallmark of certain cancer driver genes. However, state-of-the-art algorithms for detecting recurrent aberrations fail to detect several known drivers. In this study, we propose RUBIC, an approach that detects recurrent copy number breaks, rather than recurrently amplified or deleted regions. This change of perspective allows for a simplified approach as recursive peak splitting procedures and repeated re-estimation of the background model are avoided. Furthermore, we control the false discovery rate on the level of called regions, rather than at the probe level, as in competing algorithms. We benchmark RUBIC against GISTIC2 (a state-of-the-art approach) and RAIG (a recently proposed approach) on simulated copy number data and on three SNP6 and NGS copy number data sets from TCGA. We show that RUBIC calls more focal recurrent regions and identifies a much larger fraction of known cancer genes.

4.1. INTRODUCTION

Due to genomic instability, cancer cells often exhibit a large number of somatic copy number aberrations (SCNA) many of which are believed to play a pivotal role in tumor development or progression. Specifically, SCNAs represent one of the mechanisms to activate oncogenes and inactivate tumor suppressors[2, 3].

Given a large collection of somatic copy number profiles of tumors, an important challenge is to distinguish driver from passenger aberrations. The exact genomic locations of somatic passenger aberrations are expected to be variable across different tumor samples. In contrast, driver aberrations often recur on the same locus across tumor samples, which allows them to be identified in a properly defined statistical framework. Identification of driver aberrations is important as it allows us to identify (new) oncogenes and tumor suppressors.

Many algorithms have been developed for detecting recurrent copy number aberrations [4–15], highlighting the relevance of discovering novel oncogenes and tumor suppressors. However, this problem is still far from being solved as state-of-the-art approaches fail to identify known oncogenes and tumor suppressors in large sample sets. For example, while *EGFR* is one of the most frequently amplified oncogenes in Glioblastoma [16], neither RAIG and especially not GISTIC2 detects the complete recurrently amplified region harboring *EGFR*.

One of the main difficulties in detecting recurrent copy number aberrations arises from the heterogeneous nature of driver aberrations across samples, ranging from focal aberrations covering a single gene to broad aberrations spanning a whole chromosome arm. Algorithms should call recurrent regions as focally as possible to pinpoint the driver genes and hence maximizing specificity. Conversely, too much emphasis on focality could result in driver genes being confused with passengers in close proximity, simply due to off target focal passenger aberrations overlapping with a broader recurrent locus. This results in reduced sensitivity. Therefore, a proper approach should strike a good balance between sensitivity and specificity.

The great majority of algorithms, including the algorithm we propose, start by splitting copy number gains and losses into separate data sets and therefore detect oncogenes and tumor suppressors separately. Throughout, we will only consider the copy number gains - deletions are treated in a symmetric fashion. When considering gains, the first step of existing algorithms is to detect broad loci that are amplified at a significant frequency. Subsequently, heuristics are applied to identify separate focal recurrences within these loci (Fig. 4.1a-g). This so-called peak splitting is achieved in two possible ways. In the first approach, the null model is adapted [13, 17] based on the local background to determine whether a smaller locus is recurrently amplified in an already recurrent locus. This requires the re-estimation of many parameters on smaller loci, resulting in a loss of statistical power. The second approach employs greedy peel-off algorithms [6, 7, 12] that call local maximum peaks in recurrent loci (Fig. 4.1b,c) and then remove all aberrated

segments that overlap with the identified maximum peak (Fig. 4.1d). Subsequently, new maximal peaks are identified (Fig. 4.1e) based on a reduced data set, and this loss in power can result in potentially missing important driver genes in close proximity to the original maximum peak. After iterating these steps, a list of independent peaks are generated (Fig. 4.1f). The boundaries of these peaks are sensitive to passenger aberrations and a post processing step is employed (e.g. the RegBounder algorithm [7]) to broaden the peaks and improve the probability of including the correct driver genes (Fig. 4.1g).

With RUBIC (Recurrent Unidirectional Break Identification by Clustering) we follow a completely different approach. Specifically, RUBIC detects recurrent copy number breaks instead of recurrent amplifications or deletions. A recurrent break marks a region where a significant portion of the samples show transitions in copy number from neutral to gain (positive break) or from gain to neutral (negative break) (Fig. 4.1i). RUBIC is based on a simple idea: if we can prove significant recurrence of breaks that occur in close proximity of each other, a subset of these breaks are most likely associated with driver aberrations. Regions enclosed between recurrent positive breaks on the left and recurrent negative breaks on the right will most likely harbor a putative oncogene. This new approach has several advantages. First, it simplifies the identification of recurrent regions significantly: there is no need for complicated peak splitting or peel-off algorithms. Second, power is maximized as the recurrent breaks are identified based on all samples and by employing a null model based on the behavior of passenger aberrations on the complete genome. This is in contrast to peak splitting approaches that require recursive re-estimation of the null model on an ever-decreasing locus width or recursive identification of maximal peaks on an ever-decreasing number of samples in peel-off algorithms.

In summary, by focusing on recurrent breaks, RUBIC becomes independent of the regions between the breaks. Specifically, RUBIC circumvents the difficulties of current algorithms outlined above which stem from aiming to call regions at the right size. RUBIC is simple, computationally efficient and outperforms existing methods on both simulated and real data sets. It calls more true positive regions (between 1.4 and 3.6 times more than GISTIC2) at more (appropriate) focal widths thus pinpointing the responsible driver genes. Finally, the algorithm only requires a single parameter, controlling the false discovery rate of called regions.

4.2. RESULTS

4.2.1. OVERVIEW

RUBIC detects significantly recurrent breaks in the aggregate copy number profile of a collection of tumor samples (Fig. 4.1h-k). Essentially, RUBIC performs hierarchical clustering on the aggregate profile (Fig. 4.1j, red line). It starts with segments spanning a single measurement probe and iteratively joins neighboring segments until a significant break between segments in the aggregate profile is encountered. As only neigboring segments can be joined, the complexity of the clustering problem is significantly reduced. Due to the nature of hierarchical clustering, this implies that all remaining breaks between segments in the aggregate profile are significant. All significant breaks in the aggregate profile represent segment boundaries, and the average aggregate copy number profile between breaks represents the segment amplitude (Fig. 4.1j, black line). To determine the significance of a break, we require a break recurrence measure and a significance test. The break recurrence measure, which scores a break between two adjacent segments, is equal to the difference in segment amplitudes. Intuitively, this makes sense, since a high frequency in breaks (across samples) result in a large jump in the aggregate (Methods). Significance of the recurrence measure is represented by the expected Euler characteristic (Methods), and we employ a null model obtained through cyclic permutation of the tumor profiles (Methods). During the hierarchical clustering, RUBIC employs the expected Euler characteristic as similarity measure, thus

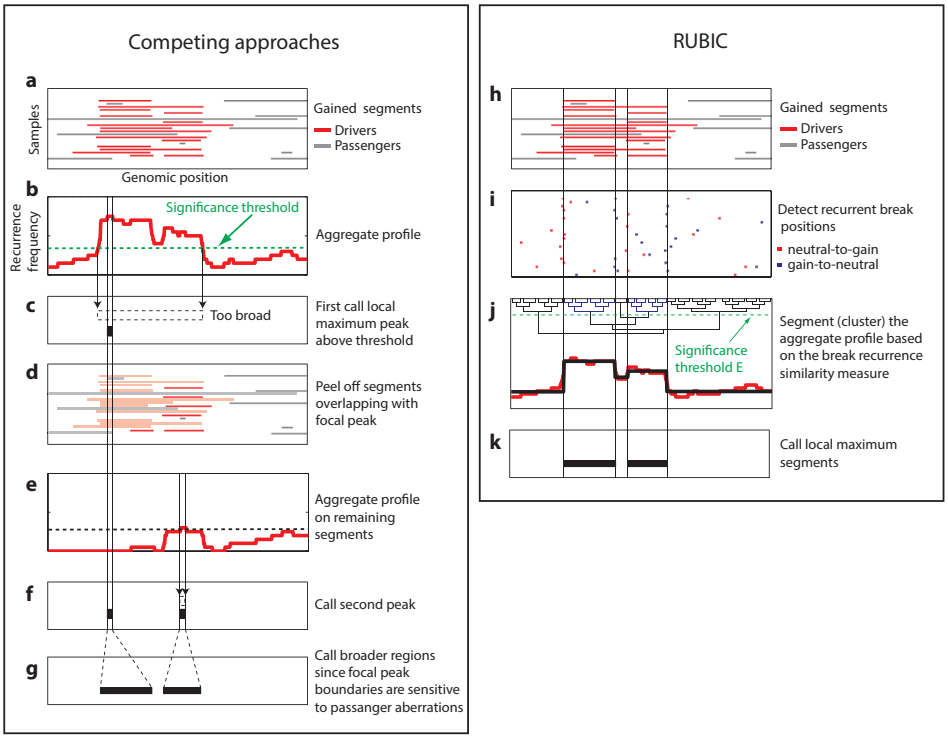

Figure 4.1: Algorithmic steps of competing approaches and RUBIC. (**a**) A heat map of the gains in simulated copy number profiles for 20 samples. Segments that activate oncogenes (driver aberrations) are shown in red and passenger aberrations in grey. (**b**) The copy number profiles in **a** are aggregated (summed) to produce the aggregate gain profile. The dashed line represents a significance threshold based on a null model, obtained by, for example, permutation of the probe indices in **a**. (**c**) The calling of the maximum peak in the aggregate profile within the genomic region where the aggregate profile exceeds the significance threshold. (**d**) Copy number segments overlapping with the maximal peak are removed from the data set. (**e**) Based on the adapted data set, a new aggregate profile and significance threshold are computed. (**f**) As in **c**, a maximum peak is called in the adjusted aggregate profile. (**g**) Finally, a post processing step is employed to broaden the peaks and improve the probability of including the correct driver genes. (**h**) The same input data set depicted in **a**. (**i**) Positions of recurrent breaks in the copy number profiles. Neutral-to-gain breaks are depicted in red and gain-to-neutral breaks in blue. (**j**) The segmented profile (in black) resulting from performing hierarchical clustering on the aggregate profile (in red). During clustering, RUBIC employs the expected Euler characteristic as similarity measure, thus allowing termination of the clustering when all segments are separated by significant breaks with similarity measures below a fixed, predetermined threshold (green dashed line). The dendrogram resulting from clustering the aggregate profile is also depicted, with the significance threshold used as cutoff to produce the depicted segmentation. (**k**) Local maximum segments are called and such segments are expected to contain putative oncogenes.

allowing termination of the clustering when all segments are separated by significant breaks with similarity measures below a fixed, predetermined threshold, E. We choose to use the expected Euler characteristic as a significance measure because it directly links the global threshold, E (used to terminate clustering) to the expected number of false positive regions called in Fig. 4.1k (Methods). This results in error control at the segment level, rather than the probe level, as in competing approaches. The clustering produces a segmented aggregate profile, where the positions of the breaks in the aggregate profile indicate regions of significantly recurrent breaks in the sample profiles (Fig. 4.1j). Finally local maximal segments are called (Fig. 4.1k). Such segments are expected to contain putative oncogenes as only gains were employed in this example. Our implementation of RUBIC can be downloaded at http://ccb.nki.nl/software/.

4.2.2. BENCHMARKING ON SIMULATED DATA SETS

In order to benchmark RUBIC and competing approaches we generated a simulated data set of copy number profiles. In contrast to most available simulation approaches that artificially insert recurrent copy number aberrations of fixed widths at any given locus, we employed a preselected set of 100 driver genes as starting point. We generated a copy number profile for each sample based on an idealized evolutionary model. Briefly, we simulate genomic instability by inserting random amplifications and deletions across the genome for many individual cells. In some cells, amplifications activate oncogenes and deletions inactivate tumor suppressors. Such driver aberrations modulate the proliferation rate of an individual cell. The cell with the highest score is then regarded as the dominant clone which we use to represent the sample. This process is repeated for each sample in our analysis. Simulated copy number profiles exhibit complex recurrence patterns developing on both focal and broad scales. For more information on the model and the simulated profiles see the Methods.

We systematically compared RUBIC to GISTIC2 (a state-of-the-art approach) and RAIG (a recently proposed approach) on simulated data sets generated using our evolutionary model. We employed all three algorithms to separately detect recurrent amplifications and deletions.

For GISTIC2 and RAIG we used exactly the same parameter settings as for the real tumor data sets (Supplementary Methods in [1]). RUBIC requires only a single parameter to be set: the FDR. For all algorithms, results were generated at an FDR level of 25%. Each algorithm reports a list of regions and genes (partially) overlapping with these regions. We removed all called regions that did not overlap with any genes. Such regions were never reported by RUBIC or RAIG. Only GISTIC2 reported four such regions in all simulations performed, and suggested nearby genes in brackets, none of which were drivers. We also removed regions larger than 10 Mega base pairs (Mbp), since they usually contain many genes and that makes it difficult to pinpoint the drivers. Although rare, such broad regions are sometimes called by GISTIC2 and RUBIC, but not RAIG.

We evaluated the performance based on three measures: 1) the proportion of driver genes that overlapped with called recurrent regions (true positives); 2) the proportion of called regions that do not overlap with any of the driver genes (false positives) and 3) the average driver density in called regions. The third measure scores the ability of algorithms to call regions as focally as possible, i.e. the capacity to pinpoint drivers.

We varied the number of samples from 10 to 1000, and for each number of samples we generated five simulated data sets from which we extracted recurrent regions. RUBIC outperforms both GISTIC2 and RAIG in terms of the number of drivers detected as well as the driver density while controlling the FDR (Fig. 4.2). Both RUBIC and GISTIC2 achieve an FDR well below the set rate of 25%. For RUBIC, the measured false discovery rate is stable at 5% across sample sizes, but much lower than the 25% FDR selected. This is due to the fact that the cyclic shift null model is conservative. Even though RAIG performs fairly well on previously reported simulation studies, it performs

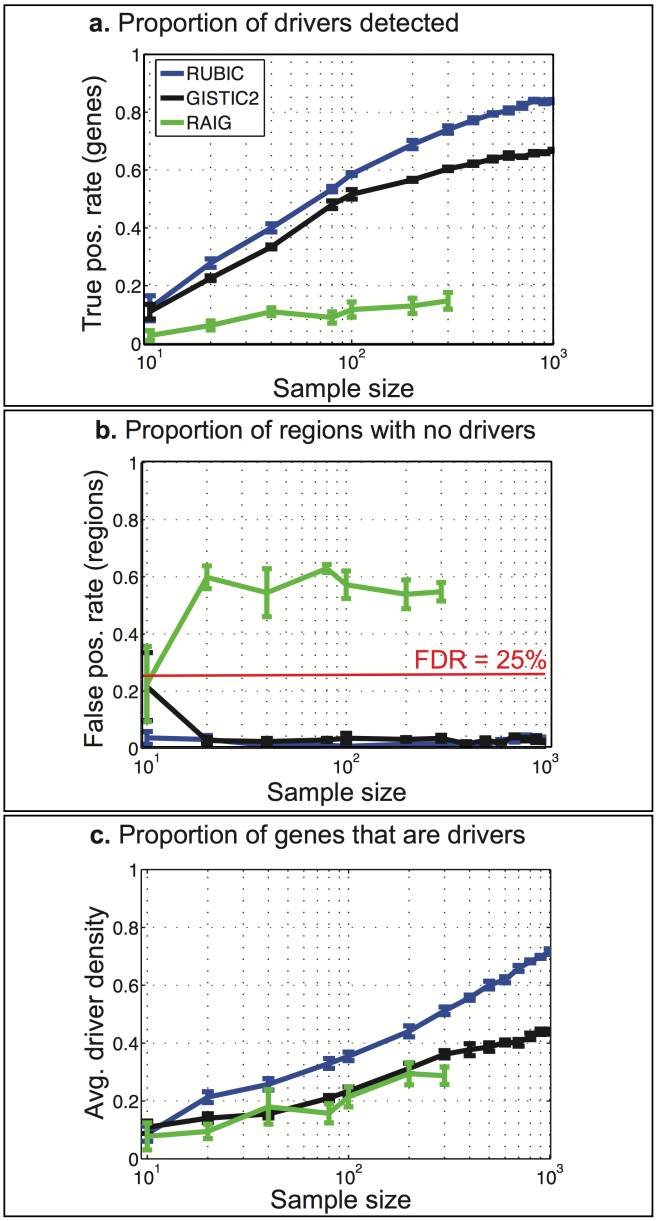

Figure 4.2: Performance on simulated data. The performance of RUBIC, GISTIC2 and RAIG at recovering driver genes in simulated data for different sample sizes (represented on the x-axes). (**a**) Proportion of (known) driver genes that overlapped with called recurrent regions. (**b**) Proportion of called recurrent regions that overlapped with none of the driver genes. (**c**) Average (across called regions) proportion of genes that are drivers within each called recurrent region. Error bars represent the standard error based on five samplings.

significantly worse than RUBIC and GISTIC2 on all measures for this evolutionary model. In addition, RAIG does not scale well computationally with regard to the number of samples. In Fig. 4.2 we only show RAIG results for up to 300 samples as we only depict results for processes that took less than two weeks to complete.

4.2.3. COMPARISON ON THREE TCGA SNP6 DATA SETS

We compared RUBIC, RAIG and GISTIC2 on SNP6 copy number profiles of three cancer data sets from TCGA: 1080 Breast Invasive Carcinoma (BRCA) samples, 577 Glioblastoma Multiforme (GBM) samples and 450 Colon Adenocarcinoma (COAD) samples. We used optimized parameters for GISTIC2 and RAIG as specified in the Supplementary Methods in [1]. We set the FDR at 25% and extracted both recurrent amplifications and deletions with all three algorithms. As in the simulation experiments, we only report regions that overlapped with at least one gene and excluded all regions larger than 10 Mbp.

Unlike the simulation study, we do not know which genes are drivers and therefore we selected 463 genes (Supplementary Data 1 in [1]) as positive controls from the Sanger Institute Cancer Gene Census (referred to as 'Census genes') [18]. We score each algorithm based on four measures: 1) the total number of focal recurrent regions detected ('No. regions'); 2) the number of regions that overlap with Census genes ('No. Census regions'); 3) the total number of Census genes detected ('No. Census genes') and 4) the average driver density in the called regions. The driver density of a region is defined as one divided by the number of genes overlapping the region and is therefore a measure of how good the respective algorithms are at identifying drivers. While this (conservative) measure is optimal when every region contains a single driver, we do not rule out the possibility of multiple weak drivers in a region. If the data supports the presence of multiple (weak) drivers, as suggested in the literature [19], RUBIC will detect these. Table 4.1 summarizes the results obtained for the three algorithms on all three data sets. Each entry has two values (separated with a slash) representing recurrent gains and losses, respectively.

RUBIC calls more recurrent regions than both GISTIC2 and RAIG on all three data sets. Given that the FDR was set at 25% the majority of these regions are expected to contain true driver genes. In fact, the larger number of recurrent regions called by RUBIC also results in a larger yield of Census genes. RAIG calls fewer amplified regions than GISTIC2 on the BRCA data set, but none of these regions contain a Census gene. On the GBM and COAD data sets, RAIG calls more regions than GISTIC2, however, the called regions contain fewer Census genes. These results suggest that the RAIG error rate is high, which is consistent with our observations in the simulation study. The superior ability of RUBIC and GISTIC2 to recover Census genes was also confirmed by a global analysis. Specifically, by employing a cyclic permutation test, we found an enrichment for Census genes ($p < 0.05$, permutation test) in all data sets for both RUBIC and GISTIC2, but not for RAIG. In fact, only the amplified regions called by RAIG on the GBM data set showed significant enrichment for Census genes.

While the average driver density estimates for RUBIC are smaller than those obtained by GISTIC2 for the gains, these values are not strictly comparable since RUBIC calls many more regions. This is because recurrent regions that are only detected by RUBIC do not recur as frequently as those that were detected by both algorithms. Regions of lower recurrence will necessarily be called broader and therefore result in a lower average driver density for RUBIC. Specifically, if we only look at the 27 amplified regions where RUBIC and GISTIC2 overlap in BRCA, the average driver densities are comparable, with 0.29 and 0.35 for RUBIC and GISTIC2 respectively. Of these 27 amplified regions, 16 were called (slightly) more focally by GISTIC2. However, 6 of these regions called by RUBIC included extra Census genes. In contrast, none of the 11 regions that were called more broadly by GISTIC2 included any extra Census genes. These results suggest that GISTIC2 indeed tends to call amplified regions too focally. Perhaps the best example illustrating that GISTIC2

tends to call amplifications too focally is *EGFR* in glioblastoma. The skeptical reader might suspect that we ran GISTIC2 with sub-optimal parameters, but in fact we downloaded the GISTIC2 results (with optimized parameters) from http://firebrowse.org/. Counter-intuitively, if we run GISTIC2 on smaller subsets (< 577) of the Glioblastoma data set, we do actually detect *EGFR*. The reason is that GISTIC2 calls regions wider for smaller sample sizes (Fig. 4.2c), but ironically falls prey to passenger aberrations that 'distract' from the true driver aberration at larger sample sizes.

Deletions called by RUBIC are more focal than those called by GISTIC2 (higher average driver density in Table 4.1), while the opposite is true for amplifications. The asymmetry between achieved average driver densities (gains vs. losses) in the RUBIC results makes sense from a biological perspective: while tumor suppressors can be inactivated by deletions of sub-genic size, aberrations resulting in over-expression of oncogenes typically cover the whole gene and are therefore expected to be wider. RUBIC only called 8%, 15% and 9% of the amplified regions based on a break inside a gene for the BRCA, GBM and COAD datasets, respectively. In contrast, 34%, 65% and 53% of all called deletions were based on a break inside a gene for the same respective datasets.

When considering the overlap in the Census genes retrieved by the three approaches (Fig. 4.3) we notice that RUBIC returns the largest number of Census genes and that the majority of the Census genes retrieved by GISTIC2 and RAIG are a subset of the Census genes retrieved by RUBIC. In the breast cancer data set, GISTIC2 was able to call a single unique broad amplified region that was not detected by RUBIC. This region resides at the end of Chromosome 1q and contains a single Census gene. For the deletions, GISTIC2 called 11 unique regions that were not detected by RUBIC. Three of these regions overlapped with Census genes. One of these regions on Chromosome 9q is very broad (9 Mbp) and contains seven Census genes. This single region explains most of the disparity between the results of RUBIC and GISTIC2 in Fig. 4.3a. RUBIC did call this region, but it was filtered out as it just exceeded 10 Mbp.

Some known oncogenes and tumor suppressors are only captured by RUBIC, such as *MDM4* in breast, *APC* in colon and *EGFR* in Glioblastoma (Fig. 4.3). *EGFR* is the most frequently amplified gene in Glioblastoma, yet neither GISTIC2 nor RAIG detects it. GISTIC2 missed *EGFR* because it called a false focal peak (containing no overlapping genes) near *EGFR* and peeled away most of the segments overlapping with the false peak that also overlap with *EGFR* (Fig. 4.4). Interestingly, RUBIC calls two regions, consistent with the observation that 24%-67% of all glioblastoma's are Type III deletion mutants where Exons 2-7 are deleted [20]. This result also suggests that *EGFR-AS1* might be an oncogene in its own right.

4.2.4. Focused analysis of the breast cancer data set

We analyzed the BRCA data set more closely and show a genomewide overview of called regions by all three algorithms in Fig. 4.5. Here we also highlight (in red) a small subset of bona fide and/or recently validated oncogenes (52 in total) and tumor suppressors (12 in total) specifically associated with breast cancer. The list is constructed based on strong evidence for the involvement of each of the genes in breast cancer, and is largely based on two published lists [17, 21]. See Supplementary Table 1 in [1]. A subset of the oncogenes in this list were only recently validated [17]. RAIG, GISTIC2 and RUBIC were able to recover, respectively, 0, 13 and 34 of these bona fide oncogenes and 2, 5 and 5 of these tumor suppressors. All five regions containing the tumor suppressors were called more focally by RUBIC as compared to GISTIC2. A global enrichment test with a cyclic permutation scheme shows that all RUBIC and GISTIC2 regions are highly enriched for bona fide oncogenes and tumor suppressors (all p-values $< 10^{-3}$, permutation test).

Zooming in on some loci, we illustrate examples in which RUBIC outperforms both GISTIC2 and RAIG. First, we find that RUBIC is able to recover four validated oncogenes missed by both GISTIC2 and RAIG (Fig. 4.5b). In the second example, GISTIC2 called an amplification peak too focally and missed *MIR21* (Fig. 4.5c). Finally, we show an example where GISTIC2 called a too

Table 4.1: Summary of detected regions on SNP6 data set. Recurrent copy number regions predicted by RUBIC, GISTIC2 and RAIG on BRCA, GBM and COAD. For each subtable containing the results of a specific cancer type, the rows represent the following: the first row (labeled 'No. regions') represents the total number of focal recurrent regions detected by each algorithm. The second row shows the number of regions that overlap with Census genes. The third row represents the total number of Census genes detected. The last row shows the average driver density in the called regions. Each entry has two values (separated with a slash) representing recurrent gains and losses, respectively.

Breast cancer (BRCA) ($n = 1080$)			
Methods	RUBIC	GISTIC2	RAIG
No. regions (gains/losses)	100/58	28/31	11/41
No. Census regions (gains/losses)	48/16	15/17	0/5
No. Census genes (gains/losses)	63/26	16/33	0/5
Avg. driver density (gains/losses)	0.21/0.41	0.34/0.10	0.80/0.57
Glioblastoma (GBM) ($n = 577$)			
Methods	RUBIC	GISTIC2	RAIG
No. regions (gains/losses)	40/152	22/36	25/58
No. Census regions (gains/losses)	23/26	14/13	7/6
No. Census genes (gains/losses)	33/34	15/15	7/6
Avg. driver density (gains/losses)	0.29/0.71	0.39/0.19	0.59/0.56
Colon adenocarcinoma (COAD) ($n = 450$)			
Methods	RUBIC	GISTIC2	RAIG
No. regions (gains/losses)	23/72	17/31	27/50
No. Census regions (gains/losses)	11/12	8/9	6/5
No. Census genes (gains/losses)	16/14	10/10	6/7
Avg. driver density (gains/losses)	0.14/0.58	0.21/0.20	0.36/0.46

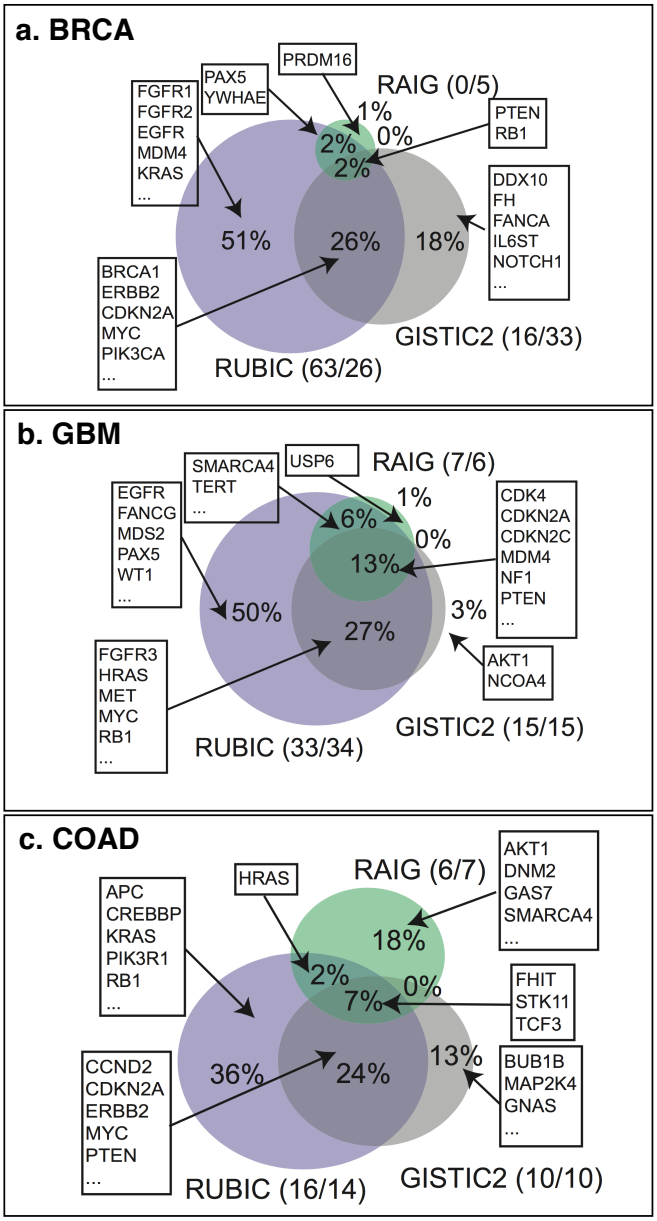

Figure 4.3: Detected Census genes. Venn diagram of Census genes that overlapped with called recurrent regions in RUBIC, GISTIC2 and RAIG. (**a**,**b** and **c**) illustrates this for the breast (BRCA), glioblastoma (GBM) and colon (COAD) cancer data sets, respectively. The numbers separated by a slash (in brackets) represent Census gene counts for gains and losses, separately.

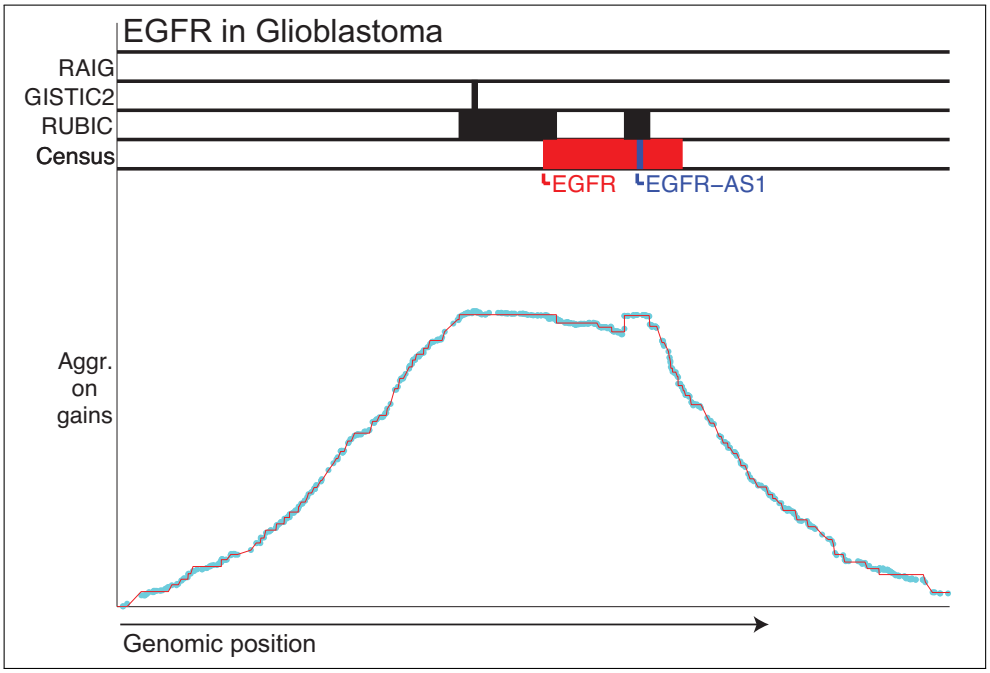

Figure 4.4: Recurrence at the EGFR locus. Genomic representation of *EGFR* and called recurrent regions in its proximity by RUBIC, GISTIC2 and RAIG on the Glioblastoma data set. The cyan profile represents the aggregate copy number profile. The RUBIC segmented aggregate is depicted in red. The rows with labels RUBIC, GISTIC2 and RAIG show the genomic locations of regions called by each of these algorithms. The row with label 'Census' shows the location of *EGFR* (in red) and *EGFR-AS1* (in blue).

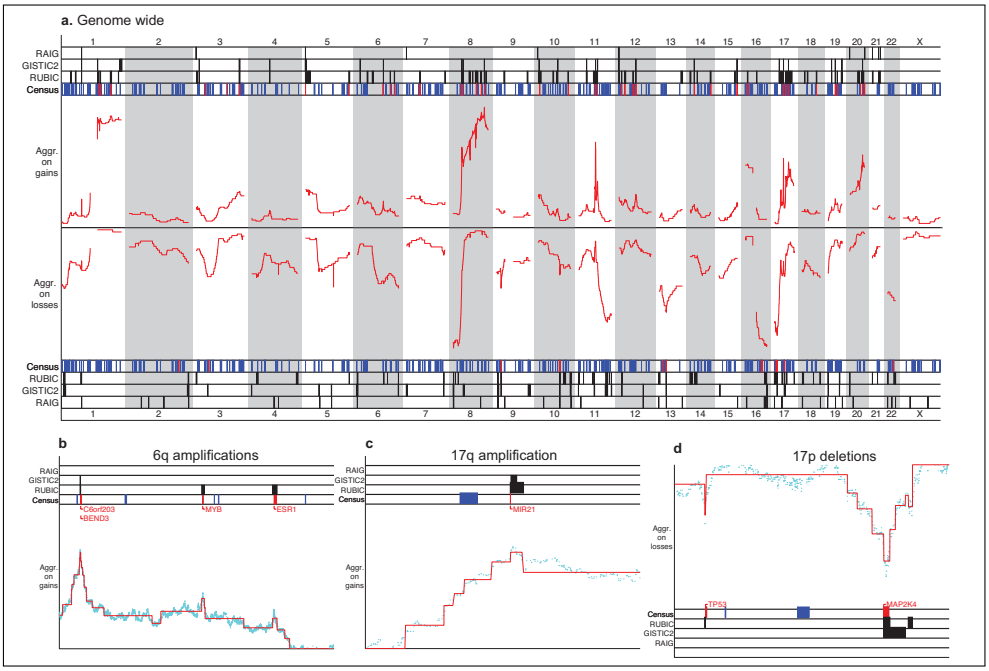

Figure 4.5: Genome-wide overview of detected regions in breast cancer. Genomic representation of recurrent regions found by RUBIC, GISTIC2 and RAIG in the breast cancer data set. (**a**) RUBIC segmented aggregate profiles (in red) across the whole genome for gains and losses in the top and bottom halves respectively. The rows with labels RUBIC, GISTIC2 and RAIG show the genomic locations of called recurrent regions. The row with label 'Census' shows the locations of Census genes in blue. Validated breast cancer genes are represented in red (in the Census row). (**b** - **d**) Example (zoomed in) loci with validated gene names added in red. The cyan profiles represent the aggregate of all samples before RUBIC segmentation depicted in red.

broad deletion in *MAP2K4* (Fig. 4.5d). More generally, for regions overlapping with RUBIC, GISTIC2 consistently calls broader deletions without introducing any extra known tumor suppressors or Census genes, while the broader amplifications called by RUBIC do include more oncogenes and Census genes as compared to those called by GISTIC2. This suggests that the smaller deletions and larger amplifications called by RUBIC improve driver detection.

4.2.5. COMPARISON ON NEXT GENERATION SEQUENCING

In order to investigate the applicability of RUBIC to copy number profiles derived with Next Generation Sequencing (NGS) technology we compared RUBIC and GISTIC2 on two additional NGS data sets. The first data set consists of copy number profiles of 90 breast cancer samples (not overlapping with TCGA samples) derived from low coverage (< 1x average) whole genome sequencing (lcWGS) [22]. The second set contains 383 TCGA breast cancer copy number profiles derived from whole exome sequencing (WES). Since all comparisons indicated that RAIG is not a competitive approach, we only benchmarked RUBIC against GISTIC2 on the NGS data sets. We used optimized parameters for GISTIC2 as before, set the FDR at 25% and extracted both recurrent amplifications and deletions with both algorithms. As before, we only report regions that overlapped with at least one gene and excluded all regions larger than 10 Mbp. The results indicate that both RUBIC and GISTIC2 can be successfully applied to NGS data, as we recover recurrent regions containing known drivers, albeit at a lower average driver density. The lower density is a direct consequence of the fact that the sample size of the NGS data is lower compared to the SNP6 data, resulting in larger called regions and hence a lower average density. The results also indicate that the observations we made based on the SNP6 data regarding the relative performance of RUBIC and GISTIC2 can be extrapolated to NGS data (Table 4.2). Specifically, we show that RUBIC consistently identifies more recurrent regions, more Census genes and more bona fide breast cancer genes at comparable or higher driver densities. On the lcWGS data set, RUBIC detects a larger number of Census genes, in spite of the fact that the sample size is much lower than the SNP6 breast cancer data set. This is most likely caused by two factors. First, the lcWGS set contains many BRCA-like samples, characterized by BRCA1/2 specific but highly unstable copy number profiles, increasing the likelihood of detecting recurrent aberrations. Second, due to the smaller sample size, the regions called by RUBIC are larger, hence increasing the likelihood of detecting more Census genes. While the overlap of amplified regions identified on lcWGS profiles with the SNP6 recurrent regions is around 50%, it is consistent at that level for both RUBIC and GISTIC2. There are two reasons why we expect this overlap to be low. First, the collection of samples used for lcWGS is highly enriched for the BRCA-like samples compared to the TCGA SNP6 dataset. Second, the collection of patient samples used for the lcWGS does not overlap with the TCGA data set and the obtained overlap is therefore consistent with an FDR of 25%. In contrast, 87% of the amplifications detected by RUBIC in the WES dataset overlap with those found in the SNP6 dataset. The patient samples in the WES dataset are a subset of those comprising the SNP6 dataset and there is no enrichment for any particular subtype (the 383 samples were selected randomly). This suggests that RUBIC is robust against the platform differences, in contrast to GISTIC2 that obtains only 61% overlap.

4.2.6. FRAGILE SITE ANALYSIS

Since RUBIC calls recurrent breakpoints, it is reasonable to ask whether we are not simply calling breakpoints at fragile sites. To answer this, we would have to test whether the recurrent regions called by RUBIC are enriched for fragile sites. We employed a published list [23] of fragile sites and combined that with an unpublished list obtained from the Sanger Institute to construct a list of 127 rare and common fragile sites and performed a permutation-based enrichment test (Supplementary Methods and Supplementary Data 2 in [1]). We could not find any enrichment for fragile sites in recurrent regions called by RUBIC for either the SNP6, lcWGS or WES profiles in

Table 4.2: Summary of detected regions on NGS data sets. Recurrent copy number regions predicted by RUBIC and GISTIC2 for BRCA datasets derived from low coverage whole genome sequencing (lcWGS) and TCGA whole exome sequencing (WES). For each subtable containing the results of a specific sequencing platform the rows represent the following: the first row (labeled 'No. regions') represents the total number of focal recurrent regions detected by each algorithm. The second row shows the number of regions that overlap with Census genes. The third row represents the total number of Census genes detected. The fourth row shows the number of BRCA bona fide oncogenes / tumor suppressors detected. The fifth row shows the enrichment p-values for bona-fide drivers in regions based on a cyclic permutation test. The sixth row shows the average driver density in the called regions. The final row shows the proportion of regions detected in NGS data that overlap with regions found for the SNP6 TCGA data set. Each entry has two values (separated with a slash) representing recurrent gains and losses, respectively.

BRCA (lcWGS) ($n = 90$)			
	Methods	RUBIC	GISTIC2
1	No. regions (gains/losses)	80/43	26/29
2	No. Census regions (gains/losses)	47/10	17/7
3	No. Census genes (gains/losses)	90/21	25/20
4	No. bona fide genes	32/3	10/2
	(52 oncogenes/12 tumor suppressors)		
5	Enrichment p-values for bona fide genes	2×10^{-4}/0.022	$< 1 \times 10^{-4}$/0.083
6	Avg. driver density (gains/losses)	0.12/0.24	0.09/0.13
7	Region overlap with SNP6	0.50/0.42	0.42/0.33

BRCA (WES) ($n = 383$)			
	Methods	RUBIC	GISTIC2
1	No. regions (gains/losses)	46/9	13/3
2	No. Census regions (gains/losses)	32/4	10/1
3	No. Census genes (gains/losses)	58/ 14	16/1
4	No. bona fide genes	28/2	10/1
	(52 oncogenes/12 tumor suppressors)		
5	Enrichment p-values for bona fide genes	$< 1 \times 10^{-4}$/0.018	$< 1 \times 10^{-4}$/0.021
6	Avg. driver density (gains/losses)	0.07/0.25	0.08/0.21
7	Region overlap with SNP6	0.87/0.56	0.61/0.33

Table 4.3: Percentage of called regions overlapping with fragile sites and the associated enrichment p-values computed with permutation tests.

RUBIC regions on SNP6 profiles			
Recurrence type	BRCA (n = 1080)	GBM (n = 577)	COAD (n = 450)
Gains: % overlap (p-value)	20% (0.91)	38% (0.10)	17% (0.95)
Losses: % overlap (p-value)	36% (0.10)	19% (0.68)	29% (0.52)

RUBIC regions on lcWGS and WES profiles		
Recurrence type	lcWGS (n = 90)	WES (n = 383)
Gains: % overlap (p-value)	34% (0.43)	33% (0.61)
Losses: % overlap (p-value)	35% (0.52)	33% (0.59)

any of the cancer types considered, as indicated in Table 4.3.

4.3. DISCUSSION

To identify cancer genes residing in recurrently aberrated genomic regions, we follow a completely different approach from current state of the art approaches. Rather than focusing on the recurrence of regions, we introduced RUBIC, an approach that considers the recurrence of breaks. This results in a significant simplification of the algorithm as there is no need for recursive identification of smaller recurrent regions in broader regions via complicated peak splitting approaches. An added advantage of the fact that RUBIC focuses on breaks reflecting the relative change in copy number between segments, rather than the cumulative strength of an aberration across samples, is that the need for an arbitrary reference state is diminished. RUBIC requires only a single input parameter (the FDR) and controls the FDR at the level of regions, rather than probes as most competing approaches. Although users are discouraged from inputting raw unsegmented copy number data into RUBIC, we do expect RUBIC to be less sensitive to the choice of a segmentation algorithm since our theoretical approach does not explicitly require piecewise constant segments. In contrast, algorithms like GISTIC2 that directly peel-off segments when calling peaks will be sensitive to the specific choice of segmentation algorithm. In a comparison with GISTIC2 and RAIG, we show that RUBIC calls significantly more recurrent regions and identifies a much larger fraction of regions containing known cancer genes (from the Cancer Gene Census).

We developed a gene centric simulation model to employ in our benchmarking studies. In this model, we define hypothetical driver genes, simulate genomically unstable copy number profiles and apply evolutionary pressure which results in driver genes being selectively aberrated. We believe this is an improvement over existing simulation approaches as 1) it focuses on genes rather than aberrations; 2) it is an approximation (albeit quite rough) of the evolutionary processes going on in real tumors and 3) it produces recurrence patterns that closely resemble patterns in real data sets. It is therefore suited for revealing shortcomings in existing approaches. For example, RAIG reports a very high recall and precision rate in a previous simulation study [15] which simulated driver aberrations, rather than driver genes. However, it commits many false positives when calling driver genes in data generated with our simulation model. On simulated data from this model, RUBIC outperforms both GISTIC2 and RAIG on all measures: it finds more driver genes, calls very few regions that don't overlap with driver genes and calls these regions more focally. Nonetheless, RUBIC does tend to call fewer false positive regions than expected based on the set FDR due to its conservative null-model. This is because many copy number breaks belong to driver aberrations that are not provably recurrent. Yet we include these breaks in our null-model which should ideally only contain breaks associated with passenger aberrations.

On the three TCGA data sets we employed, we selected 463 genes from the Cancer Gene Census to employ as positive controls. It should be noted that one should not expect all of these genes to be involved in tumor development and maintenance specifically in breast cancer, glioblastoma and colon cancer, since they have been found to be somatically mutated in a much broader variety of cancer types. Nor should we expect all of them to be activated or inactivated through copy number aberrations, as most of these genes were identified based on the occurrence of other aberrations, such as point mutations. In fact, in some cases we do not even know the status (oncogene or tumor suppressor) of the genes. We therefore also considered a much smaller subset of bona fide or validated breast cancer oncogenes ($n = 52$) and tumor suppressors ($n = 12$).

As stated in the introduction, algorithms should strive to accurately pinpoint drivers by calling recurrent regions as focally as possible. On the other hand, as we have shown, too much emphasis on focality results in calling passengers (only) in close proximity to drivers (*EGFR* in Glioblastoma being a good example). This problem is threefold. First, it results in the driver genes being missed, reducing the true positive rate. Second, passenger genes are called erroneously, increasing the false positive rate. Finally, these erroneously called passengers are often reported as highly significant, since they do occur in highly recurrent regions.

Finally, we have demonstrated that RUBIC is not only applicable to SNP6-derived copy number profiles, but can also successfully be applied to copy number profiles derived from NGS data. We showed that the results obtained in the comparison with GISTIC2 on the SNP6 data also hold for NGS data, both in the setting of copy number profiles derived from low coverage whole genome sequencing as well as whole exome sequencing.

While it is beyond the scope of this work, the methodology of RUBIC can be applied to other application domains. For example, large-scale projects such as The Encyclopedia of DNA Elements (ENCODE) are generating large amounts of ChipSeq data. Typically these profiles are subjected to peak calling to identify, for example, binding sites of transcription factors or domains characterized by a specific chromatin mark. The segmentation approach we proposed here can be employed to segment ChipSeq profiles in order to identify binding peaks and domains. Note that this will amount to the application of the segmentation to a single sample. However, as ChipSeq is also being applied in tumor material on a more regular basis, we foresee that RUBIC will also be applied to patient cohorts to detect recurrently occurring peaks or domains.

4.4. METHODS

4.4.1. THE BREAK RECURRENCE MEASURE

Suppose we have a genomic region R of width w and cut it at position g_0 into two regions: R_L and R_R. Let the widths of these regions be denoted by w_L and w_R, respectively. Let the average of the aggregate profile for regions R_L and R_R be denoted by $\hat{\mu}_L$ and $\hat{\mu}_R$, respectively. Positive breaks near g_0, that significantly recur across samples, will result in $\hat{\mu}_R > \hat{\mu}_L$. In contrast, under the null model (which models passengers) positive and negative breaks are equally likely to occur anywhere in R. From this it follows that, under the null, the expected means will be equal, i.e. $E[\hat{\mu}_R] = E[\hat{\mu}_L]$. It is important to note that this equality holds even if R is fully contained within a recurrent region. It is this observation that removes the need to employ the peak splitting algorithms mentioned in the introduction.

It can be shown that a recurrent break occurs at g_0 in R by showing that the value of $t(g_0) = \hat{\mu}_R - \hat{\mu}_L$ is significant according to the null model. This is similar to performing a two-sample t-test where the two samplings are represented by the aggregate log ratio measurements in R_L and R_R respectively. Formally, $t(g)$, is also a function of $w = (w_L, w_R)$, and will be denoted by $t_w(g)$. The larger w_L and w_R, the more statistical power one attains. However, if these regions are too large and extend beyond loci in which recurrent breaks of the opposite sign occur, the power will decrease considerably.

In Section 4.4.4 we show how hierarchical clustering based on the significance of t_w can be employed to simultaneously find appropriate values for w_L and w_R and identify recurrent breaks based on the break recurrence measure, t.

4.4.2. THE NULL MODEL

We employ a null model to describe passenger breaks and hence identify recurrent breaks by evaluating the significance of the break recurrence measure. We use a cyclic shift permutation scheme described in detail in the literature [13, 14] to define a null model. To sample from the null distribution, we shift probe indices by a random offset for each copy number profile independently. In this scheme, all break locations become independent across samples, while the inherent genomic dependencies within each sample, for example chromothripsis patterns, are retained. As with the real data, we sum all cyclically shifted sample profiles per probe (locus) to form one realization of the aggregate profile under the null. (For notational convenience, a specific realization will be denoted by the index i.)

By repeatedly permuting profiles one can estimate the probability that breaks recur at observed frequencies by chance alone. This null model is conservative, since we would ideally only model passenger breaks, whereas many of the breaks in our data contribute to driver events. To reduce this bias, we first detect driver breaks with RUBIC and then update the null model after deleting these breaks. We repeat these two steps iteratively (see Section 4.7.2).

4.4.3. MEASURING THE SIGNIFICANCE OF BREAK RECURRENCE

For a fixed w, each t_w can be associated with a (two-tailed) p-value derived from the null model. We will, instead, use a different measure of significance called the expected Euler characteristic [13, 24, 25]. This measure is more natural in our application and will allow us to directly control the false discovery rate on called recurrent regions rather than probes, as explained later. The idea is as follows: for any fixed realization of the null model (indexed with i), a fixed w and a fixed non-negative threshold t, we define positive and negative excursion sets: $A^+_{w,i} = \{g : t_{w,i}(g) \geq t\}$ and $A^-_{w,i} = \{g : t_{w,i}(g) \leq -t\}$, respectively. We count the number of disjoint regions in each and denote these with $\chi^+_{w,i}$ and $\chi^-_{w,i}$, respectively. The sum of these counts, $\chi_{w,i}(t) = \chi^+_{w,i} + \chi^-_{w,i}$, is known as the Euler characteristic. We can then compute the expected Euler characteristic across

realizations: $\bar{\chi}_w(t) = \sum_{i \in I} \chi_{w,i}(t)/|I|$, where I represents the set of all possible permutations (see Section 4.7.1 and Fig. 4.6).

On actual data, for a fixed scale w_0 and position g_0, we can compute a value $t_0 = t_{w_0}(g_0)$. $\bar{\chi}_{w_0}(|t_0|)$ can be interpreted as a measure of significance (small values being significant). In fact, it is an upper bound for the familywise error rate (FWER) if we regard each locus g as a separate test and it is a tight bound for small values: $\bar{\chi}_w < 0.1$ [24]. It is important to note that the Euler characteristic allows us to link the value of the break recurrence measure at a specific locus and a fixed scale, $t_0 = t_{w_0}(g_0)$, to the significance of the number of called recurrent regions in the aggregate profile.

There are two major advantages of using $\bar{\chi}_w$ as a significance measure. First, there exists an analytical approximation that relates $t = t_w(g)$ to $\bar{\chi}_w(|t|)$ that is highly accurate for the majority of scales (see Eq. 4.3 in Section 4.7.1). This means that we can avoid time consuming permutation tests for many choices of w (see the subsection entitled 'RUBIC segmentation based on both the permutation and analytically derived expected Euler characteristic' in Section 4.7.1). The second, and more important reason, is that we can directly compute the false discovery rate on called recurrent regions (not breaks) using $\bar{\chi}_w$. We clarify this in Section 4.4.5.

4.4.4. SEGMENTATION

Ultimately, RUBIC is a segmentation algorithm on the aggregate profile. We essentially approximate the aggregate profile with a piecewise constant function with jump discontinuities at significantly recurrent breaks. The jump discontinuities represent significant breaks in the aggregate profile. The jump height at position g is exactly equal to $t_w(g)$, where $w = (w_L, w_R)$ represents adjacent segment widths. We regard breaks in the aggregate profile as significant if $\bar{\chi}_w$ is small.

RUBIC segmentation is an agglomerative hierarchical clustering algorithm that starts with the most fine-grained segmentation, where each probe is a unique segment, and iteratively merges adjacent segments. As a measure of the similarity of two segments, we use $\bar{\chi}_{w_s}$, where w_s corresponds to the widths of the segments under consideration. In each iteration, we merge segments with the highest (least significant) $\bar{\chi}_{w_s}$ score. We continue merging segments until all remaining $\bar{\chi}_{w_s}$ scores are less than or equal to a fixed global threshold, E. This implies that the jump discontinuities separating the remaining segments are all significant ($< E$) and hence represent recurrent breaks. In the segmented profile, segments residing between recurrent breaks are represented by a single value, the average of the aggregate profile in that segment. Since all segments are naturally sorted on the genome, and we only need to consider adjacent segments for merging, we can efficiently perform the clustering in $P\log(P)$ time, where P is the number of probes on the genome. Fig. 4.1j shows the resulting segmentation when we perform this procedure for a fixed significance threshold, E.

4.4.5. CALLING

The final step in the algorithm is to simply call all the local maximum segments in Fig. 4.1j producing the result illustrated in Fig. 4.1k. A segment is defined as a local maximum when it is bordered by positive and negative jump discontinuities on its left and right, respectively. One can then expect to find oncogenes inside these called segments since positive (negative) jump discontinuities correspond to significantly recurrent ($< E$) positive (negative) breaks, i.e. recurrent amplifications.

The remaining question is: how to choose the global threshold E? The benefit of using the Euler characteristic as similarity measure is that $E/2$ is an upper bound on the expected number of false positive local maximum segments (called regions) that result in the data (see Section 4.7.5). Since there is a direct correspondence between the number of false positive regions and the threshold E, we can directly apply the Benjamini-Hochberg procedure [26] to control the false discovery rate on recurrently amplified called segments [13]. We illustrate the Benjamini-

Hochberg procedure with an example. Suppose we specify the FDR level at 25%. We then start by setting $E = 2 \times 0.25$, knowing that the expected number of false positive called regions will be below 0.25. We then count the number of called regions after clustering, say there are 70. At this point we choose $E = 70 \times 2 \times 0.25$. We continue adapting E until the number of called regions remain unchanged. Say, for example, we end up with 100 called regions. At this point $E = 100 \times 2 \times 0.25 = 50$. The expected number of false positives will be below $E/2 = 25$, which is 25% of the 100 called regions.

4.4.6. SIMULATING COPY NUMBER EVOLUTION WITH KNOWN DRIVER GENES

Given the lack of a real copy number data sets for which all the oncogenes and tumor suppressors are known, it is very hard to compare algorithms in terms of specificity and sensitivity on real data. This type of analysis can only be achieved through simulation. The majority of simulation studies are performed by artificially inserting numerous recurrent copy number aberrations of fixed widths for any given locus [10]. Such simulations are not designed to give a direct answer to how good algorithms are at pinpointing driver genes, since they define driver aberrations rather than driver genes. In fact, it is questionable whether amplifications of a fixed width recur across multiple samples in real data sets, except for events occurring on the chromosome arm level.

Since real copy number aberrations are subject to selective pressure, we expect oncogenes (tumor suppressors) to be found in recurrently amplified (deleted) regions without the need for fixed recurrent segment widths. For example, an oncogene can be frequently amplified across samples even though the associated aberration widths vary considerably. A small subset of these amplifications might be focal enough so as to cover only this one gene. With enough samples, it is likely that there is a sufficient number of these focal aberrations to unambiguously call the oncogene as being recurrently aberrated. Consequently, finding driver genes is not the same as finding recurrent aberrations of a fixed width.

For that reason, we simulated copy number profiles based on an idealized evolutionary model in which a fixed number of oncogenes and tumor suppressors are known. We assigned a proliferation coefficient to each gene which indicates the influence of that gene on cell proliferation. More specifically, the contribution of a gene to cell proliferation is the product of the coefficient and the average copy number fold change of that gene with respect to the normal diploid state (see Section 4.7.3). In our performance study, we selected 100 random genes from the human genome as drivers and assigned to each a proliferation score drawn from a normal distribution. Positive (negative) coefficients represent oncogenes (tumor suppressors). During the simulated evolutionary process, copy number changes were introduced in the profile, resulting in copy number changes in several genes, including oncogenes and tumor suppressors.

We started the simulation by creating a copy number neutral (diploid) dominant clone, with a genome of the same size as the human genome. We then evolved the copy number profile of a single sample by repeating the following randomization and selection steps 20 times:

- Randomization: derive 100 descendants from the dominant clone by adding 10 random copy number aberrations at random locations on the genome. The width and copy number log ratios of the aberrations are extracted from the TCGA breast cancer data set (see Section 4.7.3).
- Selection: based on the proliferation coefficients and copy number values of the 100 selected driver genes, compute the overall proliferation of each descendent, select the descendent with the highest proliferation score and define it as the new dominant clone.

The final dominant clone represents the final copy number profile of a single sample. This process is repeated for every sample. Simulated copy number profiles resemble what we observe in real data, with complex recurrence patterns developing on both focal and broad scales.

4.5. DATA AVAILABILITY

The lcWGS and simulated DNA copy number data that support the findings of this study are available in GitHub, https://github.com/ewaldvandyk/RUBIC-datasets.git. The TCGA SNP6 and WES data that support the findings of this study are available from TCGA but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. We provide full details on the TCGA data that we employed as well as the processing steps that were applied to these data to obtain the input profiles employed in our analyses. Hence, after obtaining the data from TCGA under licence our results can be reproduced. All of the remaining data are available within the Article and Supplementary Information files or available from the authors upon request.

4.6. ACKNOWLEDGEMENTS

We would like to thank Gergana Bounova, Sander Canisius, Magali Michaut, Daniel Vis and Guillem Rigaill for many discussions and critical reading of the manuscript. We would like to thank Tycho Bismeijer and Nicola Bonzanni for assistance with the software.

4.7. SUPPLEMENTARY METHODS

4.7.1. COMPUTING THE EXPECTED EULER CHARACTERISTIC IN RUBIC SEGMENTATION

OVERVIEW

RUBIC segments the aggregate copy number profile using agglomerative clustering. It starts by considering each probe to be a unique segment and continues merging adjacent segments until a stopping criteria is met. For each pair of adjacent segments we compute a similarity measure called the expected Euler characteristic and in each step we merge adjacent segments with the highest similarity measure. This process continues until all similarity measures are below a fixed threshold E. In this section we describe in detail how to estimate the expected Euler characteristic between adjacent segments (which generally depends on the widths of the segments). First, we describe a method involving a permutation scheme that is accurate but is computationally inefficient. Second, we describe an analytical approximation that is often (but not always) accurate and otherwise conservative. Finally, when clustering it is important to decide which approximation to use. We discuss this in the final part of this section. A derivation of the analytical approximation and the reason why the expected number of local maximum segments in the null model will be below the threshold E is discussed in the last two sections of the chapter (Sections 4.7.4 and 4.7.5, respectively).

ESTIMATING THE EXPECTED EULER CHARACTERISTIC WITH PERMUTATIONS

In Fig. 4.6a, we give a detailed breakdown on exactly how to estimate the expected Euler characteristic with a permutation scheme for a fixed scale $w = (w_L, w_R)$ and fixed non-negative threshold t.

In one realization of the null (illustrated by the dashed box in Fig. 4.6a), we shift probe indices by a random offset for each sample independently. Probes that are shifted beyond chromosome boundaries are shifted into adjacent chromosomes' start positions and probes shifted beyond the end of the last chromosome are shifted into the start positions of the first chromosome. In this scheme, all break locations are independent between samples, while the inherent genomic dependancies (for example chromothripsis) between breaks are retained within each sample. The next step is to simply sum all the cycled samples' profiles.

To compute the Euler characteristic for a fixed realization, we need to compute $t_w(g)$ at every genomic position g. To efficiently do this we define a wavelet kernel as follows:

$$k_w(g) = \begin{cases} +1/w_R & -w_R \le g < 0 \\ -1/w_L & 0 \le g \le w_L \\ 0 & \text{elsewhere} \end{cases} \tag{4.1}$$

$t_w(g)$ is computed using this kernel with a computationally efficient operation called convolution (indicated by the symbol $*$ in Fig. 4.6). Essentially, for each locus g_0, it computes $t_w(g_0)$ by 1) reversing and shifting the kernel to location g_0, $k_w(g_0 - g)$, and 2) multiplying it with the aggregate profile (using the scalar product). The resulting convolved profile is shown in the lower right of Fig. 4.6a. This is equivalent to first convolving the individual samples' profiles with the kernel and then summing (because convolution is distributive over addition).

For the fixed realization i, we can count the number of disjoint regions of $\chi^+_{w,i}$ ($\chi^-_{w,i}$) above the threshold t (below $-t$) as illustrated by the red (green) numbers in Fig. 4.6a. The Euler characteristic for i is then simply the sum of these counts.

If we repeat this permutation scheme many (N) times, we can estimate the expectation across realizations as indicated by the formula at the bottom of Fig. 4.6a.

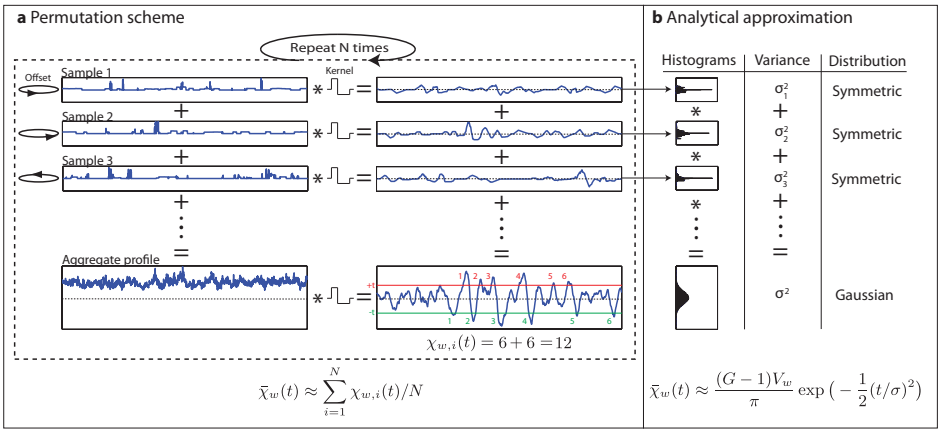

Figure 4.6: **Illustration on how to compute the expected Euler characteristic.** We assume a fixed scale $w = (w_L, w_R)$ and fixed positive threshold t in the null-model. Throughout, $*$ represents the convolution operator. (**a**) Illustration on how to approximate the expected Euler characteristic with the permutation scheme. In one realization of the null, we cyclicly shift each sample's probes with a random offset (independently for each sample). Each sample is convolved with the wavelet kernel corresponding to $w = (w_L, w_R)$ and the results are shown in the curves in the column on the right in **a**. We sum the resulting convolved profiles to obtain the curve in the lower right corner in **a**. Equivalently, we can first sum all the cyclicly shifted samples to produce a realization of the aggregate in the lower left corner and then convolve it with the wavelet. The Euler characteristic is computed by counting disjoint regions above (below) the threshold t ($-t$). This procedure is repeated N times and estimates the expected Euler characteristic using the equation at the bottom of **a**. (**b**) We can accurately approximate the expected Euler characteristic analytically without the permutation scheme if w_L and w_R are large enough. We compute the histogram of each sample's convolved profile. We compute the variance in each histogram and note that they are symmetric. Since all samples are cyclicly shifted independently, we can compute the distribution (variance) on the aggregate by convolving (adding) the histograms (single sample variances). Finally, we approximate the null as a Gaussian process with an analytically derived expected Euler characteristic as shown by the bottom equation in **b**.

AN ANALYTICAL APPROXIMATION OF THE EXPECTED EULER CHARACTERISTIC

For many scales w (when w_L and w_R are large), we can accurately approximate $\bar{\chi}_w(t)$ analytically as illustrated in Fig. 4.6b. We present our copy number data as a $S \times G$ matrix of copy number log ratios $[c_{s,g}]$, where S is the number of samples and G the number of probe measurements on the genome. Each row in the matrix represents a single sample profile c_s, where probes are sorted on the genome. Each sample profile can be convolved with the kernel to produce $t_{w,s} = k_w * c_s$.

First we compute the histogram for each sample's convolved profile ($t_{w,s}$). Note that these histograms remain unchanged for cyclically shifted profiles. These histograms have mean zero (the kernel integrates to zero) and are somewhat symmetric (positive breaks are as likely to occur as negative breaks in every sample). Due to independence between samples in the null, we can compute the histogram on the aggregate convolved profile by convolving all the sample histograms. It is then a consequence of the central limit theorem that allows us to approximate the convolved aggregate profile as a multivariate Gaussian random process with zero mean. Furthermore, this process will be stationary due to the cyclic shift hypothesis (all probes behave in the same way everywhere, since we randomly offset them for each realization). For a stationary zero-mean multivariate Gaussian process there are only two parameters that need to be estimated: The variance σ_w^2 and auto-correlation r_w. The auto-correlation is a function of Δg with $r_w(\Delta g)$ equal to the Pearson correlation between probe measurements separated by Δg probes. It turns out that, in order to relate $\bar{\chi}_w$ to a threshold t, we only need to compute $\rho_w = r_w(1)$, i.e. the Pearson correlation between adjacent probe measurements. We explicitly compute the variance σ_w^2 and ρ_w as follows:

$$\begin{aligned} \sigma_w^2 &= \sum_{s=1}^{S} \sigma_{w,s}^2, \qquad \sigma_{w,s}^2 = \frac{1}{G} \sum_{g=1}^{G} t_{w,s}^2(g) \\ \rho_w &= \frac{1}{\sigma_w^2} \sum_{s=1}^{S} \rho_{w,s}, \qquad \rho_{w,s} = \frac{1}{G} \sum_{g=1}^{G} t_{w,s}(g)\, t_{w,s}\big((g+1)/G\mathbb{Z}\big) \end{aligned} \tag{4.2}$$

We can then accurately relate the expected Euler characteristic to a positive threshold t for a stationary Gaussian process as follows:

$$\begin{aligned} \bar{\chi}_w(t) &= \frac{(G-1)V_w}{\pi} \exp\big(-\frac{1}{2}(t/\sigma_w)^2\big), \quad \text{where} \\ V_w &= \arccos(\rho_w) \end{aligned} \tag{4.3}$$

RUBIC SEGMENTATION BASED ON BOTH THE PERMUTATION AND ANALYTICALLY DERIVED EXPECTED EULER CHARACTERISTIC

In RUBIC segmentation, we need to compute $\bar{\chi}_w$ for many different scales. To do so purely based on the permutation scheme is computationally prohibitive and it is desirable to use the analytical estimate instead.

The Gaussian assumption does hold for the majority of kernel choices w. However, when w_L and w_R are small, the approximation becomes inaccurate. For example, suppose we choose $w = (1,1)$. In this case, $t_{w,s}$ will be zero everywhere except at locations where copy number breaks occur. Due to the sparsity of $t_{w,s}$, the Gaussian assumption fails and the analytical prediction will be liberal.

Due to these considerations we need to perform segmentation using a hybrid between these estimates. The methodology is simple: We cluster segments based purely on the analytical model at first. After segmenting with this methodology, there will only be a small number of jump discontinuities in the segmented profile. All of these jump discontinuities will be significant according to the analytical estimate. It is only at this point where we recompute significance values (on the small set of jump discontinuities remaining) based on the permutation scheme. After we did so,

we continue segmenting with the analytical estimates. We iteratively continue until all jump discontinuities are significant based on the permutation estimate.

With this procedure, we are always ensured that the expected number local maximum segments (that we end up calling) in the aggregate segmented profile is below or equal to $E/2$, where E is the global threshold used to stop clustering. This is true no matter what analytical approximation we use for $\bar{\chi}_w$. If this analytical approximation is conservative, we end up merging segments with jump discontinuities that should have been called significant and therefore the number of local maximum segments will be lower than $E/2$. On the other hand, if the analytical approximation is liberal, we do not lose power or incur more false positives. Instead, we end up testing many jump discontinuities with the permutation scheme. The analytical approximation we use tend to be liberal for small kernel sizes (and we should not expect to lose power because of it). Nevertheless, by the time we stop merging segments, the majority of jump discontinuities remaining define larger kernels, where the analytical approximations hold well.

Due to the overall accuracy of our analytical estimate, we rarely need to iterate these steps. Usually after segmenting the aggregate profile with the analytical approximations, the permutation scheme also calls the jump discontinuities significant and we stop.

4.7.2. ITERATIVELY UPDATING THE NULL-MODEL

Generally, we consider a copy number aberration to be a driver if it provides a selective advantage in tumor initiation and progression. Driver aberrations are likely to effect (and therefore overlap) with oncogenes and tumor suppressor genes (driver genes). Passenger aberrations on the other hand are believed to provide no significant selective advantage and occur due to genomic instability. It is not necessarily true that all aberrations that overlap with driver genes are driver aberrations. For example, it could happen that an aberration amplifies an inactive allele of an oncogene. Nevertheless, if we can prove that neutral-to-gain (or gain-to-neutral) breaks recur significantly across samples, then we have good reason to believe that at least a subset of them belong to driver aberrations.

In Fig. 4.7 we show an example of 20 DNA copy number samples (only the gains). In the top panel we show all the driver aberrations in red, whereas passenger aberrations are shown in grey. Note that we generally don't have prior knowledge on which aberrations are drivers and we show it here only for illustrative purposes. The important point here is that not all driver aberrations (or breaks) will necessarily recur significantly. Therefore it is not possible to fully discriminate drivers from passengers in the data.

The cyclic shift null-model describes the behavior of passenger aberrations and is based on the assumption that they occur randomly on the genome. Unfortunately, this scheme will be conservative, since the overall break density will be higher in the cyclic null model than the true passenger break density. Although the break locations are random in the null, a large portion of the breaks that are scatter across the genome originate from drivers that are concentrated in fixed genomic loci in the data (see the lower panel in Fig. 4.7). Therefore, the cyclic shift null-model will over estimate the background break density. Nevertheless, we can significantly reduce this bias by iteratively detecting recurrent break points with RUBIC and then remove the concentrated breaks from our null model. Notably, this will also automatically remove the apparent breaks between adjacent chromosome boundaries.

At each iteration we perform RUBIC segmentation on the aggregate profile with family wise error (FWER) control (it is not particularly natural to do FDR control for updating the null). Family wise error control is straight forward and achieved by setting the clustering threshold at E = FWER without applying the Benjamini-Hochberg procedure. We set the family wise error rate equal to the FDR level. We segment each sample using the break locations detected by RUBIC (with segment amplitudes equal to the mean in that particular sample's copy number ratio). From each sample,

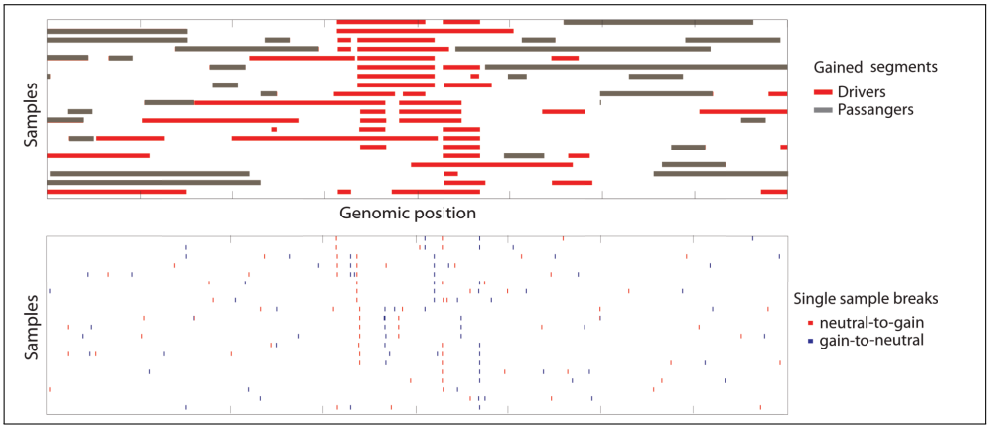

Figure 4.7: **Illustrating an example with 20 DNA copy number profiles.** The top panel shows copy number gains as they occur across the genome (x-axis) for each sample (y-axis). Driver aberrations are shown in red and passengers in grey. The bottom panel shows the copy number break points for each sample. Neutral-to-gain breaks are shown in red, while gain-to-neutral breaks are shown in blue.

we subtract this segmented profile and update the null accordingly (i.e. cyclically permute based on the new sample profiles). As a result, all frequently recurrent breaks are canceled. Note that the 'cancellation' of breaks will not necessarily align perfectly with the correct driver breaks. This leads to a slight power loss (slightly more conservative than need be) since we are technically introducing new breaks locally (and therefore increase the break density in the null). Nevertheless they will occur in close proximity (and be of opposite sign) to the true driver breaks and greatly improves statistical power (especially for larger kernels).

At each iteration, we re-segment the aggregate of the original dataset (not the profiles with the cancelation breaks) based on the updated null-model. We continue iterating until no new breaks can be detected at the specified FWER level.

We use the resulting null-model (after convergence) to finally segment the aggregate with FDR control.

4.7.3. EVOLUTIONARY MODEL FOR SIMULATING COPY NUMBER PROFILES

SIMULATING PASSENGER ABERRATIONS

When we simulate tumor instability in our simulation model, we randomly add copy number aberrations that were extracted from the TCGA breast cancer datasets. Each copy number sample in the Level 3 TCGA data is represented as a list of segments with start and end positions and the average copy number log ratio of each segment. A copy number aberration does not typically correspond to these segments directly. For example, suppose there are only two aberrations in chromosome 1q, one focal amplification and one broad gain of the whole chromosome arm. We want to extract these two events separately, but in the level 3 data at our disposal there will be three segments corresponding to 1q. To extract aberrations we follow a simple strategy. In each sample and in each chromosome, we extract the segment with the highest (positive or negative) copy number log ratio. We then merge adjacent copy number segments that border the chosen segment into one large segment. We take care to associate the merged segment with a log ratio that is the weighted average of the two segments contributing (the weights depend on the sizes of the segments). We iteratively extract segments in this fashion until no more segments remain.

This procedure is performed for each sample and each chromosome in our TCGA breast cancer dataset. Since we use these segments to simulate passenger aberrations that are located at random positions and samples, we need not record their locations nor which samples they are from, but only their genomic widths and log ratio values. We do add a flag indicating whether the segment is as wide as the chromosome, i.e. was the last segment extracted from a complete chromosome. Therefore we end up with a large aberration list with three fields: the width, log ratio and chromosome wide flag.

When we add a passenger aberration to an existing profile, we select a random chromosome (the probability of each chromosome is weighted by its length) and a random aberration from the list described above. If the chosen segment is flagged as chromosome wide, we change its size to the chromosome width and center it to cover the whole chromosome. Otherwise, we center the aberration at a random position in the chromosome and clip the segment at chromosome boundaries if necessary. Finally, we add the segment to the existing profile.

COMPUTING THE PROLIFERATION SCORE

We defined a list of oncogenes and tumor suppressors d_i $(1 \leq i \leq D)$. For each we assigned a proliferation coefficient α_i where positive (negative) values are associated with oncogenes (tumor suppressors). From this we can compute a proliferation score P for any fixed copy number profile by assuming that cell proliferation is linearly related to the average copy number dosage (f_i) (of gene d_i) and α_i:

$$P = \sum_{i \in \{1,\ldots,D\}} \alpha_i f_i \tag{4.4}$$

The dosage f_i of gene d_i is based on the average copy number log ratio $\bar{c_i}$ of the gene and is a measure of the fold change relative to a normal reference:

$$f_i = \text{sign}(\bar{c_i})(A^{|\bar{c_i}|} - 1), \tag{4.5}$$

where $A = 2$ is the log base.

4.7.4. DERIVATION OF THE ANALYTICAL APPROXIMATION OF THE EXPECTED EULER CHARACTERISTIC

We will now derive an accurate analytical expression relating the expected Euler characteristic to a fixed non-negative threshold t for a discrete Gaussian random process with constant variance σ^2, zero mean and non-stationary correlation function $r(g)$. In our application we have a discrete stationary Gaussian random process defined at a finite number of probes $\left(T_i : i \in \{0,1,\ldots,G\}\right)$, but the theory applies to non-stationary processes too. Much work has been done on estimating the expected Euler characteristic in stationary and discrete processes [27–29]. These estimates are typically not very accurate when adjacent probes are weakly (or negatively) correlated. On the other had, for smooth stationary Gaussian processes, there exist exact expressions for the expected Euler characteristics [25]. Our strategy will be to interpolate discrete processes with a smooth (in the sense that it is continuously differentiable up to any order) Gaussian random process $H(g)$ (Fig. 4.8). From the smoothed process we can derive an extremely simple and exact expression for the expected Euler characteristic.

Without loss of generality, we will assume that the discrete process has a variance equal to one for all probes. For simplicity, we assume that the Euler characteristic is equal to the sum of the up-crossings (the red circles in Fig. 4.8) and down-crossings (the green circle in Fig. 4.8) at the thresholds t and $-t$ respectively. We refer to these up-crossings and down-crossings collectively as the level-crossings of t. Strictly, if $|t_0| > t$, where t_0 is the left most probe measurement, the

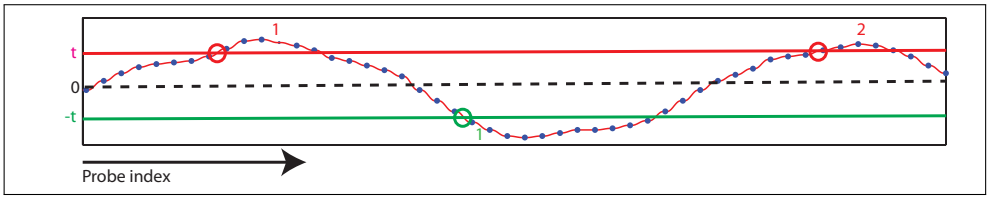

Figure 4.8: **Smooth approximation of a discrete Gaussian process.** We interpolate a discrete stationary Gaussian process (indicated with blue dots) with a smooth and non-stationary Gaussian process (in red). Counting regions above (below) a fixed threshold t ($-t$) can be accomplished by counting up-crossing (down-crossings) in the smoothed profile as indicated by the red (green) circles.

Euler characteristic will be equal to the number of crossing points plus one. Although it is easy to correct for this boundary effect, it is usually negligible and we will not consider it any further in this section.

Traditionally, the expected Euler characteristic is only computed for up-crossing above a positive threshold. In our application, we also count the down-crossings. Since a Gaussian random process is symmetric with respect to the zero line, the expected number of level crossings will be double the number of up-crossings.

The reason we prefer to work with a smooth random process is due to Theorem 1 in Section 3.8.1. We restate the theorem here for convenience:

Theorem 2. *Consider a non-negative threshold t, a closed interval $R = [g_L, g_R]$ and a suitably regular (non-stationary) random Gaussian process H with $\forall g \in \mathbb{R}\ H(g) \sim N(0,1)$. The expected number of level crossings for a threshold t in R is equal to:*

$$\bar{\chi}(t) = \frac{e^{-t^2/2}}{\pi} \int_R \sqrt{Var[\frac{d}{dg} H(g)]\, dg} \tag{4.6}$$

For the definition of a suitably regular process, see the subsection entitled 'Suitably regular processes' in Section 3.8.1.

To smooth the discrete process, we convolve it with a particular bump function f satisfying the following properties:

- It is smooth (continuously differentiable up to any order).
- $\{g : f(g) = 0\} = (-\infty, 1] \cup [1, \infty)$
- $\{g : f(g) > 0\} = (-1, 1)$
- $f(0) = 1$
- $\forall 0 \le g \le 1$, f(g) + f(g - 1) = 1

The exact choice of this function is not important, however a good example is illustrated in Fig. 4.9 and is defined as follows:

$$\begin{aligned} f(g) &= \Gamma(g)/\Gamma(0), \text{ where} \\ \Gamma(g) &= \int_{-1}^{g} \beta(2u+1) - \beta(2u-1)\, du \\ \beta(g) &= \gamma(1-g)\gamma(1+g) \\ \gamma(g) &= \begin{cases} e^{-1/g} & g > 0 \\ 0 & g \le 0 \end{cases} \end{aligned} \tag{4.7}$$

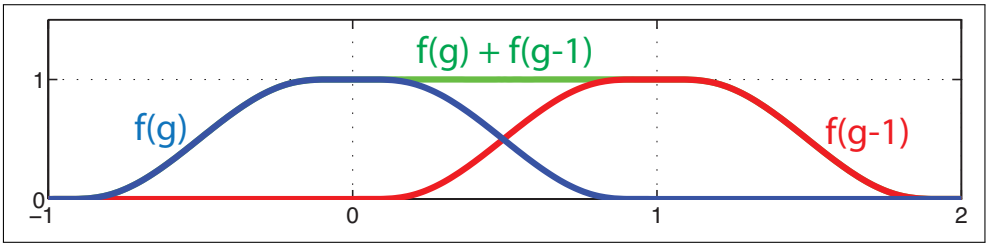

Figure 4.9: **Illustrating the properties of the Bump function used for smoothing the discrete Gaussian process.**

We convolve a realization of the discrete process $\left(t_i : i \in \{0,1,...,G-1\}\right)$ with f to produce the smoothed profile:

$$s(g) \quad = \quad f(g - \lfloor g \rfloor) t_{\lfloor g \rfloor} + \left(1 - f(g - \lfloor g \rfloor)\right) t_{\lfloor g \rfloor + 1} \tag{4.8}$$

Note that at any particular position g, s only depends on the two adjacent random variables $t_{\lfloor g \rfloor}$ and $t_{\lfloor g \rfloor + 1}$. The covariance matrix of two variables t_i and t_{i+1} is presented as follows:

$$\Sigma_i \quad = \quad \begin{bmatrix} 1 & \rho_i \\ \rho_i & 1 \end{bmatrix}, \tag{4.9}$$

where $\rho_i = \mathrm{Cov}(t_i, t_{i+1})$. Using this, we can easily compute the variance of S at any position g:

$$\begin{aligned} \mathrm{Var}[S(g)] \quad &= \quad [f_g, 1 - f_g] \Sigma_{\lfloor g \rfloor} [f_g, 1 - f_g]^T, \text{ where} \\ f_g \quad &= \quad f(g - \lfloor g \rfloor) \end{aligned} \tag{4.10}$$

Generally, $\mathrm{Var}[S(g)]$ is equal to one at integer values, but strictly smaller at intermediate values. Therefore, our final step in smoothing the discrete process is to z-normalize. The final smoothed interpolation function of a realization of the discrete random process is therefore defined as follows:

$$h(g) \quad = \quad \frac{s(g)}{\sqrt{\mathrm{Var}[S(g)]}} \tag{4.11}$$

In Fig. 4.8 we show an example realization of a discrete process (with blue dots) and the resulting smoothed function $h(g)$ in red.

Theorem 3. *Consider a non-negative threshold t, a non-stationary discrete random Gaussian process $(T_i : i \in \{0,1,...,G-1\})$ with $\forall i \in \{0,1,...,G-1\}$ $T_i \sim N(0,1)$. The Expected number of level crossings of the interpolation process $H(g)$ in the region $[0, G-1]$ is equal to:*

$$\bar{\chi}(t) \quad = \quad \frac{e^{-t^2/2}}{\pi} \sum_{i=0}^{G-2} \arccos(\rho_i) \tag{4.12}$$

Proof. Since f is smooth and h is an algebraic expression of f, h is itself smooth and $\forall g \in [0, G-1], \mathrm{Var}[H(g)] = 1$. As a consequence, all the conditions in Theorem 2 are satisfied. The only part that needs a proof is:

$$\int_0^1 \sqrt{\mathrm{Var}[\frac{d}{dg} H(g)]}\, dg \quad = \quad \arccos(\rho_0), \tag{4.13}$$

For any $g \in [\lfloor g \rfloor, \lfloor g \rfloor + 1]$, only the variables $T_{\lfloor g \rfloor}$ and $T_{\lfloor g \rfloor + 1}$ are involved in an analogous manner as T_0 and T_1 are involved for $g \in [0,1]$. From this observation and Eq. 4.13, Theorem 3 immediately follows.

In all subsequent steps, we only consider $g \in [0,1]$. First we can simplify Eq. 4.10:

$$\begin{aligned} C = \mathrm{Var}[S(g)] &= [f, 1-f]\Sigma_0[f, 1-f]^T \\ &= 2\alpha f^2 - 2\alpha f + 1, \end{aligned} \tag{4.14}$$

where $\alpha = 1 - \rho_0$. Next we rewrite Eq. 4.11 as follows:

$$h = \frac{f t_0 + (1-f) t_1}{\sqrt{C}} = \frac{f}{\sqrt{C}}(t_0 - t_1) + \frac{1}{\sqrt{C}} t_1 \tag{4.15}$$

Taking the derivative yields

$$h' = (p + qf)(t_0 - t_1) + q t_1, \tag{4.16}$$

where

$$p = f' C^{-1/2}, \;\; q = -\frac{1}{2} C^{-3/2} C' \tag{4.17}$$

It is easy to show that:

$$\mathrm{Cov}(t_0 - t_1, t_1) = \begin{bmatrix} 2\alpha & -\alpha \\ -\alpha & 1 \end{bmatrix}, \tag{4.18}$$

As a consequence,

$$\mathrm{Var}[H'] = [p + qf, q] \begin{bmatrix} 2\alpha & -\alpha \\ -\alpha & 1 \end{bmatrix} [p + qf, q]^T, \tag{4.19}$$

After a long, but straightforward, derivation we obtain the following:

$$\begin{aligned} \sqrt{\mathrm{Var}[H']} &= \frac{r'}{r^2 + 1} \\ r &= \sqrt{\frac{1-\rho_0}{1+\rho_0}}(2f - 1) \end{aligned} \tag{4.20}$$

If we set $r_0 = -\sqrt{\frac{1-\rho_0}{1+\rho_0}}$ and $r_1 = \sqrt{\frac{1-\rho_0}{1+\rho_0}}$,

$$\begin{aligned} \int_0^1 \sqrt{\mathrm{Var}[H']}\, dg &= \int_{r_0}^{r_1} \frac{r'}{r^2+1}\, dr \\ &= \arctan(r_1) - \arctan(r_0) \\ &= \arccos \rho_0 \end{aligned} \tag{4.21}$$

□

4.7.5. THE CLUSTERING THRESHOLD E CONTROLS THE EXPECTED NUMBER OF FALSE POSITIVES

OVERVIEW

If we perform hierarchical clustering on a realization of the null model and stop merging segments when all similarity measures ($\bar{\chi}$) between adjacent segments are below a global threshold E, then there will be a certain number of local maximum segments. The purpose of this section is to show that the expected number of local maximum segments resulting across realizations will be less than or equal to $E/2$. In this section it is not necessary to assume that the null process (describing the aggregate) is Gaussian. However, there are a number of properties that are required for the null model. Unfortunately we need to introduce some details and therefore we introduce these properties gradually. All these properties hold in the cyclic permutation scheme that we use. We start with the first two properties

- Property 1: The distribution of the aggregate of a probe (across realizations) is independent of the probe index, i.e. all probes have the same distribution.
- Property 2: The aggregate process is stationary, i.e. the covariance between probe measurements in the aggregate depends only on the index distance between them and not the actual probe indices

THE EULER CHARACTERISTIC AT A FIXED SCALE $w = (w_L, w_R)$ AND FIXED POSITIVE THRESHOLD t IN THE NULL MODEL

For a fixed scale we convolve the aggregate profile of a fixed realization of the null with a kernel k_w which results in values $(t_g : 0 \le g \le G-1)$, where G is the number of probes. Strictly, $t_g = t_w(g)$ is a function of the scale $w = (w_L, w_R)$ and we drop this for convenience. The corresponding null process $(T_g : 0 \le g \le G-1)$ is stationary and have mean zero everywhere, since the kernel integrates to zero. Next, we introduce the third property of the null:

- Property 3: The distribution of each T_g is symmetric with respect to zero.

This assumption holds, since our null model describes the behavior of passenger aberrations on the genome. Positive (neutral-gain) breaks are as likely to occur as negative (gain-neutral) breaks in each sample and their locations are also random on the genome.

For a fixed threshold t we can compute the Euler characteristic by counting crossings as indicated in Fig. 4.10. There are two possible ways in which $(t_g : 0 \le g \le G-1)$ can cross the positive threshold t:

- Up-crossings: probe g is considered an up-crossing when $t_g \ge t$ and $t_{g-1} < t$. In Fig. 4.10 we mark these events with green circles.
- Down-crossings: $t_g \le t$ and $t_{g-1} > t$. These are the blue circles in Fig. 4.10.

By convention we never count a crossing at probe index 0, since there is no probe to the left for comparison.

We count the number of up-crossings and down-crossings and denote it with $\chi^{\uparrow}_w(t)$ and $\chi^{\downarrow}_w(t)$ respectively. Similarly we can count the number of crossings at threshold $-t$ and denote it with $\chi^{\uparrow}_w(-t)$ (red circles) and $\chi^{\downarrow}_w(-t)$ (orange circles). In this notation, the expected Euler characteristic that we use as a similarity measure is equal to $\bar{\chi}_w(|t|) = E\left[\chi^{\uparrow}_w(t) + \chi^{\downarrow}_w(-t)\right]$, where the expectation is across realizations. As a consequence of Property 3 and the fact that expectation is a linear operator, we can see that:

$$E\left[\chi^{\uparrow}_w(t)\right] = E\left[\chi^{\downarrow}_w(t)\right] = E\left[\chi^{\uparrow}_w(-t)\right] = E\left[\chi^{\downarrow}_w(-t)\right] = \frac{1}{2}\bar{\chi}_w(|t|) \tag{4.22}$$

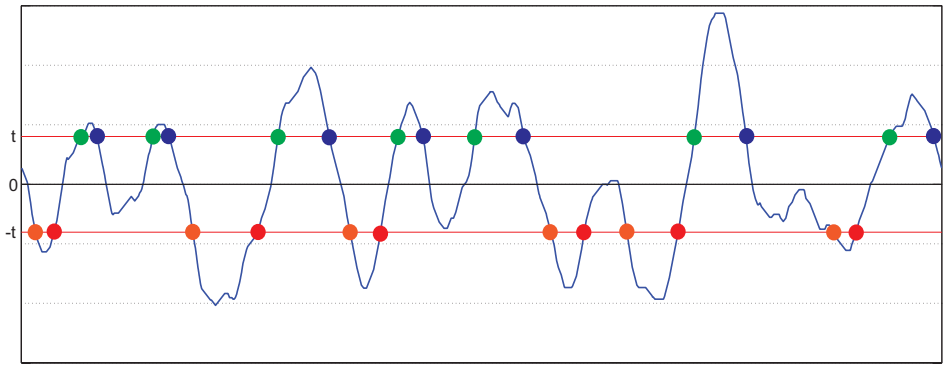

Figure 4.10: **Illustrating the type of t-level crossings that exist.** t up-crossings and down-crossings are represented by green and blue circles respectively. $-t$ up-crossings and down-crossings are represented by red and orange circles respectively

Counting up-crossing at t across the whole genome is equivalent to summing up-crossings in sub-intervals as follows:

$$\chi^{\uparrow}_{w}(t) = \chi^{\uparrow}_{w,\{0,\ldots,g_0\}}(t) + \chi^{\uparrow}_{w,\{g_0,\ldots,g_1\}}(t) + \ldots + \chi^{\uparrow}_{w,\{g_n,\ldots,G-1\}}(t) \tag{4.23}$$

Note that we include g_0 in both the intervals $\{0,\ldots,g_0\}$ and $\{g_0,\ldots,g_1\}$. This is because we never count the crossing in the left most probe of an interval (to stay in accordance with our definition).

From this observation, properties 1, 2 and, again, linearity of expectation, we see that

$$\bar{\chi}^{\uparrow}_{w,\{g_1,\ldots,g_2\}}(t) = E\left[\chi^{\uparrow}_{w,\{g_1,\ldots,g_2\}}(t)\right] = \frac{g_2 - g_1}{2(G-1)}\bar{\chi}_w(|t|) \tag{4.24}$$

We can define $\bar{\chi}^{\downarrow}_{w,\{g_1,\ldots,g_2\}}(t)$, $\bar{\chi}^{\downarrow}_{w,\{g_1,\ldots,g_2\}}(-t)$ and $\bar{\chi}^{\uparrow}_{w,\{g_1,\ldots,g_2\}}(-t)$ in a similar fashion and note that these expectations are all equal.

It is convenient to work with a density measure on the expected Euler characteristic. This density is exactly equal to the expected number of crossings at one probe (or the expected Euler characteristic at one probe). A single probe cannot be more than one type of crossing and we are therefore computing the expectation of a Bernoulli trail which is therefore also exactly equal to the probability of a probe to be a crossing. Due to the previous equation and Eq. 4.22 we can compute these densities as follows (setting $g_2 = g_1 + 1$):

$$\begin{aligned} \Delta\bar{\chi}_w(|t|) &= \frac{1}{G-1}\bar{\chi}_w(|t|) \\ \Delta\bar{\chi}^{\uparrow}_w(t) = \Delta\bar{\chi}^{\downarrow}_w(t) = \Delta\bar{\chi}^{\uparrow}_w(-t) = \Delta\bar{\chi}^{\downarrow}_w(-t) &= \frac{1}{2}\Delta\bar{\chi}_w(|t|) \end{aligned} \tag{4.25}$$

As an application, say we wish to compute the expected number of up-crossings at the negative threshold -t over P probes. Then we simply compute $\frac{P-1}{2}\Delta\bar{\chi}_w(|t|)$.

It is also convenient to transform the vector $(t_g : 0 \leq g \leq G-1)$ into a new vector $(\tau_g : 0 \leq g \leq G-1)$, where each τ_g represents a significance measure in terms of the expected Euler characteristic. For each t_g we can compute $\bar{\chi}_w(|t_g|)$, which is always positive and is uninformative with respect to the sign of t_g. Furthermore, a large value t_g will correspond to a low value for $\bar{\chi}_w(|t_g|)$.

For the sake of convenience, we would like τ_g to be an increasing function of t_g, i.e. large values in τ_g represent highly significant. With these considerations in mind, we define τ_g as follows:

$$\tau_g = \tau_w(g) = -\text{sign}(t_g)\log(\frac{\bar{\chi}_w(|t_g|)}{G-1}) \tag{4.26}$$

We should note that for any scale w and value t_g, $\bar{\chi}_w(|t_g|) \leq G-1$, since the number of crossing points can never exceed the number of probes minus one (probe zero is never a crossing). Therefore the log will always be non-positive. τ_g will always have the same sign and be an increasing function of t_g. By convention, we also set $\text{sign}(0) = 0$.

Counting crossing in $(t_g : 0 \leq g \leq G-1)$ with respect to a positive threshold t is equivalent to counting crossings in $(\tau_g : 0 \leq g \leq G-1)$ with respect to the threshold $T = -\log(\frac{\bar{\chi}_w(t)}{G-1})$. We can also express the expected Euler characteristic density in Eq. 4.25 in terms of the threshold T:

$$\begin{aligned} \Delta\bar{\chi}(T) &= \Delta\bar{\chi}_w(|t|) = e^{-T} \\ \Delta\bar{\chi}^{\uparrow}(T) = \Delta\bar{\chi}^{\downarrow}(T) &= \Delta\bar{\chi}^{\uparrow}(-T) = \Delta\bar{\chi}^{\downarrow}(-T) = \frac{1}{2}\Delta\bar{\chi}(T) \end{aligned} \tag{4.27}$$

This form is extremely convenient, since $\Delta\bar{\chi}$ only depends on T and not the scale w.

As before, we can also count the number of crossing in a restricted interval and denote them by $\chi^{\uparrow}_{\{g_1,\ldots g_2\}}(T)$, etc. We can compute the expected number of up-crossings at T in a restricted interval $\{g_1,\ldots g_2\}$ as follows: $\bar{\chi}^{\uparrow}_{\{g_1,\ldots g_2\}}(T) = \frac{g_2-g_1}{2}\Delta\bar{\chi}(T)$.

Counting local extrema in the segmented aggregate

In agglomerative clustering, we continue merging segments until all similarity measures between adjacent segments are smaller than a global threshold E. A hypothetical realization of the null after segmenting the aggregate is illustrated in Fig. 4.11a. The blue dots represent the aggregate of each probe, while the piecewise constant red graph represent the aggregate profile after clustering. Each segment has a height equal to the mean aggregate. By convention, we define a local extremum as a segment that is supported by two jump discontinuities of opposite sign. As a consequence, we do not count segments on the boundaries as local extrema. In Fig. 4.11a, there are exactly three local extrema.

In Fig. 4.11b we computed the difference between adjacent probes in the segmented profile. This results in a sparse function that is equal to zero everywhere except perhaps at the jump discontinuities. The height of a jump discontinuity at location g_n is exactly equal to $t_{g_n} = t_{(\omega_n,\omega_{n+1})}(g_n)$. In Fig. 4.11c, we map each t_{g_n} to its corresponding expected Euler characteristic $\bar{\chi}_{(\omega_n,\omega_{n+1})}(|t_{g_n}|)$. These are the similarity measures that we used for clustering and are all below the threshold E since this is the point at which we stopped merging. In Fig. 4.11d we show the τ profile, where each t_{g_n} is converted to its respective τ_{g_n} value as explained earlier. Note that the threshold E in Fig. 4.11d corresponds exactly to the threshold $T = -\log(\frac{E}{G-1})$.

The green ellipses in Fig. 4.11d illustrates a different way in which we count local extrema. Counting local extrema is performed in exactly the same way we compute the Euler characteristic for a fixed scale, except that we:

- use the τ profile in Fig. 4.11d,
- use the threshold T, and
- ignore all the probes where there are no jump discontinuities, i.e. we collapse the profile with new indices corresponding to the jump discontinuities.

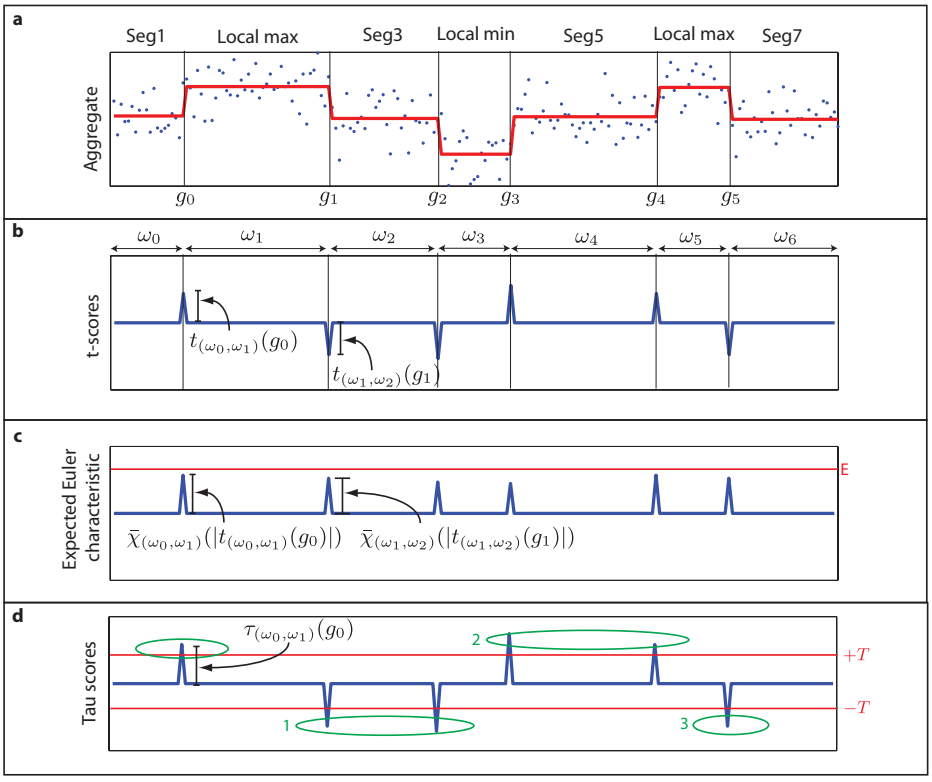

Figure 4.11: **Illustrating how we count local maximum segments.** (**a**) Illustration of the aggregate profile (in blue) and the segmented profile (in red). (**b**) The size of the jump discontinuities in the segmented profile is represented with t-scores. (**c**) Each t-score is transformed into a significance score known as the expected Euler characteristic. (**d**) Each significance score is finally converted to a signed tau-score)

We represent this number with $\chi(T)$. Note that $\chi(T)$ does not depend on any of the scale parameters, since we are counting on the segmented profile and not just any particular scale. Also note that for all τ values at the jump discontinuities it follows that $|\tau| \geq T$ (again because this is where we stop clustering). The above definition of $\chi(T)$ is only valid in this case. In the next section we show how to compute $\chi(T)$ in cases where $|\tau| < T$.

At the end of clustering, $\chi(T)$ will always be equal to the number of local extrema in the segmented profile (which equals three in Fig. 4.11a). Generally, the number of local maximum segments will be approximately equal to half of the number of local extrema. To be exact:

- $\chi(T)/2$ if $\chi(T)$ is even,
- $(\chi(T)+1)/2$ if $\chi(T)$ is odd and the left most jump discontinuity $t_{(\omega_0,\omega_1)}(g_0)$ has a positive sign (this is the case in Fig. 4.11), and
- $(\chi(T)-1)/2$ if $\chi(T)$ is odd and the left most jump discontinuity has a negative sign.

When segmenting a realization of the null, the chance that the left most break is positive is exactly 50%. The expected number of local maximum segments across realizations will therefore be exactly half of $E[\chi(t)]$.

The goal of this section is therefore to show that $E[\chi(T)] \leq E$.

Computing $\chi(T)$ when some τ values are below T

We just showed how to compute $\chi(T)$ when all jump discontinuities $|\tau_{g_n}| \geq T$. However, this is only the case at the terminal stage of clustering. For intermediate stages, there will be discontinuities where $|\tau_{g_n}| < T$. In general, the count $\chi(T)$ is computed in a slightly different manner than for a fixed scale.

We count crossings on the vector $(\tau_{g_n} : 0 \leq n \leq S-1)$, where S is the number of breaks in the segmentation. We define a discontinuity at index n as a T up-crossing if $\tau_{g_n} \geq T$ and $\tau_{g_{n-1}} < T$. By convention, the first discontinuity at index 0 will never be an up-crossing. We denote the set of indices that correspond to T up-crossings with C^+. Similarly, we define a discontinuity at index n as a $-T$ down-crossings if $\tau_{g_n} \leq -T$ and $\tau_{g_{n-1}} > -T$. We denote the set of indices that correspond to $-T$ down-crossings with C^-. Note that C^+ and C^- are disjoint.

We now introduce new types of crossings that we collectively refer to as switch crossings. We define a discontinuity at index n as a T switch-crossing if:

$$n \in C^+ \quad \text{and} \quad \max\{i \in C^- : i < n\} \geq \max\{i \in C^+ : i < n\} \tag{4.28}$$

Similarly, n is a $-T$ switch crossing if:

$$n \in C^- \quad \text{and} \quad \max\{i \in C^+ : i < n\} \geq \max\{i \in C^- : i < n\} \tag{4.29}$$

By convention, we define the max of an empty set to be equal to 0, i.e. $\max\{\} = 0$. We denote the sets of T and $-T$ switch-crossings with S^+ and S^-. Notably, $S^+ \subseteq C^+$, $S^- \subseteq C^-$. Also note that: $\min(C^+ \cup C^-) \in S^+ \cup S^-$, i.e. the left most crossing (at T or $-T$) is automatically a switch crossing. This is because the max will be 0 in both cases and the '$\geq$' statement becomes true.

The Euler characteristic is defined as the cardinality of $C^+ \cup C^-$. However, we define $\chi(T)$ as the cardinality of $S^+ \cup S^-$. Note that if $\forall n \in \{0,...,S-1\}, |\tau_{g_n}| \geq T$, $\chi(T)$ will be equal to the Euler characteristic, otherwise it will be less than or equal.

We can also compute $\chi(T)$ when segmenting on a restricted number of probes $\{m, m+1, ..., n\}$ and we denote it with $\chi_{\{m,...,n\}}(T)$. We should be careful to note that we only segment on this restricted set of probes. Generally, the resulting segmentation will not be the same as when performed on $\{0,...,G-1\}$, since there is no guarantee that jump discontinuities will result at m and n. Later on, however, we will only compute $\chi_{\{m,...,n\}}(T)$ when the global segmentation have jump discontinuities at m and n.

MERGING ADJACENT SEGMENTS IN ONE ITERATION OF CLUSTERING

Fig. 4.12 illustrates the step in agglomerative clustering where two segments are merged with the lowest $|\tau|$ separating them at location g_n. The red graph in Fig. 4.12a shows the segmented aggregate before merging segments of widths ω_n and ω_{n+1} at iteration k. The black graph shows the segmented aggregate after merging the segments at iteration $k+1$. In Fig. 4.12b we show the lag-one difference profile of the segmented aggregate (similar to Fig. 4.11b) before merging and in 4.12c we show the corresponding tau plot (similar to Fig. 4.11d) before merging. In Fig. 4.12b we show the thresholds in the diff graph that corresponds to $T = -\log(\frac{E}{G-1})$ in the τ profile. Note that these thresholds differ between discontinuities, since they are all at different scales. Fig. 4.12d and 4.12e are the same as 4.12b and 4.12c respectively, except that these relate to the segmented aggregate after merging segments (iteration $k+1$). The only differences between Fig. 4.12c and 4.12e are:

- We remove the jump discontinuity corresponding to $\tau^k_{g_n} = \tau_{(\omega_n,\omega_{n+1})}(g_n)$
- We replace $\tau^k_{g_{n-1}} = \tau_{(\omega_{n-1},\omega_n)}(g_{n-1})$ with $\tau^{k+1}_{g_{n-1}} = \tau_{(\omega_{n-1},\omega_n+\omega_{n+1})}(g_{n-1})$
- We replace $\tau^k_{g_{n+1}} = \tau_{(\omega_n,\omega_{n+1})}(g_{n+1})$ with $\tau^{k+1}_{g_{n+1}} = \tau_{(\omega_n+\omega_{n+1},\omega_{n+2})}(g_{n+1})$

In this specific example, the new value $\tau^{k+1}_{g_{n+1}}$ is above the threshold T. If $\tau^{k+1}_{g_{n-1}}$ was also significant, we would stop merging at this point.

JUMP DISCONTINUITIES CAN BECOME SIGNIFICANT AFTER MERGING. CASE 1: BORDERING A SIGNIFICANT DISCONTINUITY OF THE OPPOSITE SIGN

In Fig. 4.12d we show $t_{(\omega_n+\omega_{n+1},\omega_{n+2})}(g)$ evaluated at every g (not just g_{n+1}). The important point to note here is that although $t_{(\omega_n+\omega_{n+1},\omega_{n+2})}(g_{n+1})$ is above the threshold $\tilde{\chi}^{-1}_{(\omega_n+\omega_{n+1},\omega_{n+2})}(E)$, it is short lived with a crossing in close proximity. It is extremely likely that such a crossing will occur in the region $\{g_{n+1},...,g_{n+1}+\omega_n\}$. This is illustrated in Fig. 4.12f. Here we show the kernel (in black) $k_{(\omega_n+\omega_{n+1},\omega_{n+2})}$ when shifted to $g_{n+1}+\omega_n$, i.e. $k_{(\omega_n+\omega_{n+1},\omega_{n+2})}(g_{n+2}+\omega_n-g)$. $t_{(\omega_n+\omega_{n+1},\omega_{n+2})}(g_{n+2}+\omega_n)$ is evaluated by computing the average aggregate in the right lobe of the kernel and subtracting the average in the left lobe. These averages are shown by the dotted lines in Fig. 4.12f.

As a consequence of this observation, we specify yet another property of the null model. Given a segmentation with jump discontinuities
$\{\tau_{(\omega_n,\omega_{n+1})}(g_n), \tau_{(\omega_{n+1},\omega_{n+2})}(g_{n+1}), \tau_{(\omega_{n+2},\omega_{n+3})}(g_{n+2})\}$ and an arbitrary positive thresholds T.

- Property 4: If $\tau_{(\omega_n,\omega_{n+1})}(g_n) < T$, $\tau_{(\omega_{n+1},\omega_{n+2})}(g_{n+1}) < T$ and $\tau_{(\omega_{n+2},\omega_{n+3})}(g_{n+2}) \leq -T$, then:

$$P[\forall g \in \{g_{n+1},...,g_{n+1}+\omega_n\}, \tau_{(\omega_n+\omega_{n+1},\omega_{n+2})}(g) \geq T] \ll P[\tau_{(\omega_n+\omega_{n+1},\omega_{n+2})}(g_{n+1}) \geq T], \tag{4.30}$$

where $P[]$ represents the probability.

Although it is possible to think up pathological examples in which
$\forall g \in \{g_{n+1},...,g_{n+1}+\omega_n\}, \tau_{(\omega_n+\omega_{n+1},\omega_{n+2})}(g) \geq T$, we have never observed it in any realization of the null-model or the real data for that matter. In contrast, it often happens that
$\tau_{(\omega_n+\omega_{n+1},\omega_{n+2})}(g_{n+1}) \geq T$.

JUMP DISCONTINUITIES CAN BECOME SIGNIFICANT AFTER MERGING. CASE 2: BORDERING A DISCONTINUITY THAT IS NOT SIGNIFICANT

Next we introduce a property of the null model similar to Property 4.

Given a segmentation with jump discontinuities
$\{\tau_{(\omega_n,\omega_{n+1})}(g_n), \tau_{(\omega_{n+1},\omega_{n+2})}(g_{n+1}), \tau_{(\omega_{n+2},\omega_{n+3})}(g_{n+2})\}$ and an arbitrary positive thresholds T.

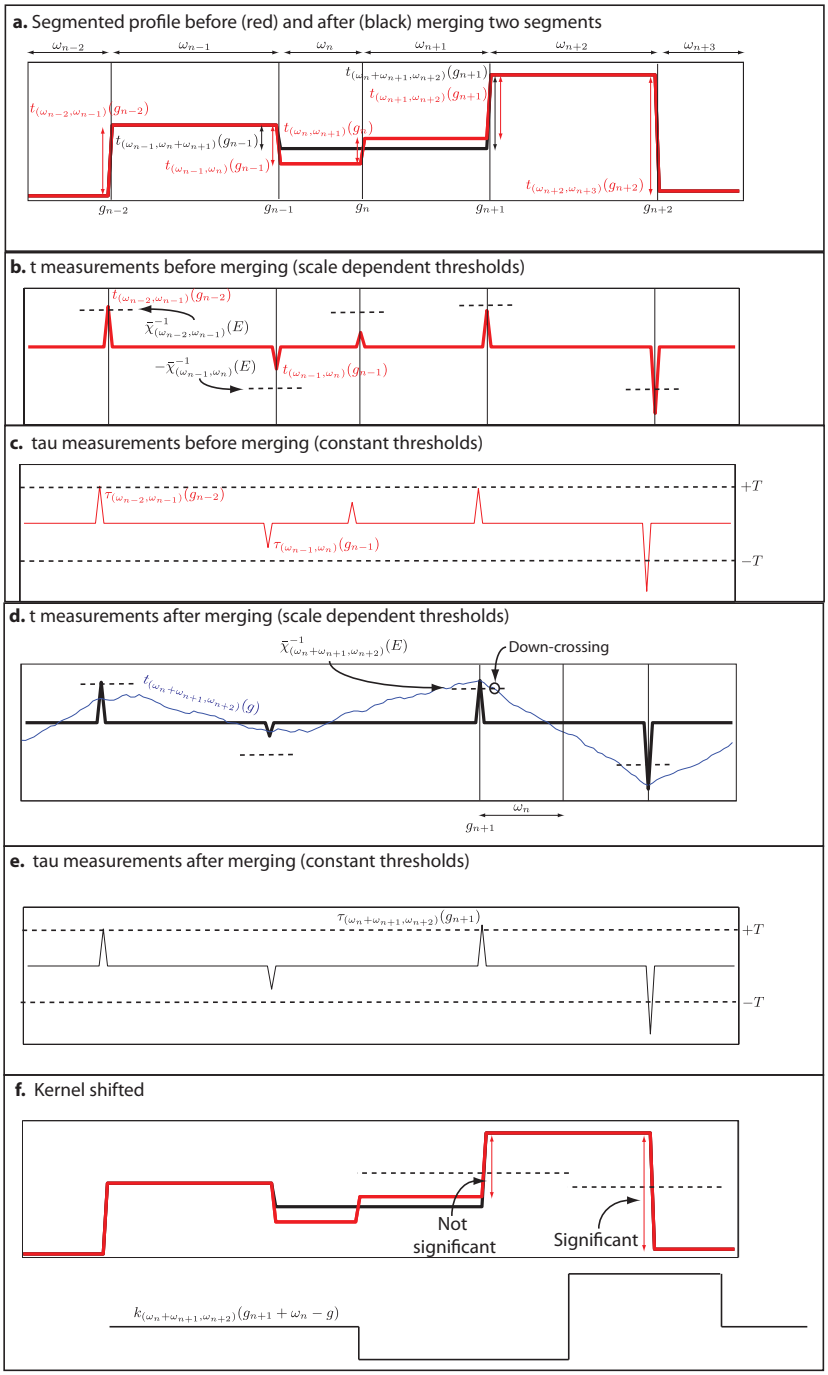

Figure 4.12: **Merging segments in one iteration of clustering.** (**a**) Illustrating the segmented profile before and after merging two segments. (**b**) and (**c**) illustrates the t- (lag-one difference) and tau-scores respectively before merging. (**d**) and (**e**) are the same as (**b**) and (**c**) respectively after merging. In (**d**), we also shows what the aggregate profile looks like after convolving with a fixed kernel. (**f**) We show why the t-score drops below the significance threshold when we shift the kernel.

- Property 5: If $\tau_{(\omega_n,\omega_{n+1})}(g_n) < T$, $\tau_{(\omega_{n+1},\omega_{n+2})}(g_{n+1}) < T$ and $\tau_{(\omega_{n+2},\omega_{n+3})}(g_{n+2}) < T$, then

$$P[\forall g \in \{g_{n+1},...,g_{n+1}+\min(\omega_n+\omega_{n+1},\omega_{n+2})\}, \tau_{(\omega_n+\omega_{n+1},\omega_{n+2})}(g) \geq T] \\ \ll P[\tau_{(\omega_n+\omega_{n+1},\omega_{n+2})}(g_{n+1}) \geq T], \quad (4.31)$$

Note that this scenario is quite similar to the one in Property 4. The major difference here is that we don't require $\tau_{(\omega_{n+2},\omega_{n+3})}(g_{n+2})$ to be below the threshold $-T$. Due to the relaxed constraint, we need to extend the domain in which we search for a down-crossing from the width w_n to $\min(\omega_n+\omega_{n+1},\omega_{n+2})$. The threshold T is large in the sense that segmentation at iteration k only resulted in jump discontinuities lower than T in the neighborhood of g_{n+1}. $\min(\omega_n+\omega_{n+1},\omega_{n+2})$ is equal to the width of one of the lobes in the kernel corresponding to the scale $w = (\omega_L,\omega_R)$, where $\omega_L = \omega_n+\omega_{n+1}$ and $\omega_R = \omega_{n+2}$. Without loss of generality, let us assume it is w_L. Property 5 is justified by the fact that t_{g_0} and $t_{g_0+w_L}$ is weakly correlated in the null and because T is high in the neighborhood. The weak correlation stems from the fact that the kernel is anti-correlated with itself when shifted by w_L probes:

$$\int_{-\infty}^{\infty} k_w(g_0-g)k_w(g_0+w_L-g)dg \quad < \quad 0 \quad (4.32)$$

As with Property 4, it is possible to think up pathological examples in which $\forall g \in \{g_{n+1},...,g_{n+1}+\min(\omega_n+\omega_{n+1},\omega_{n+2})\}, \tau_{(\omega_n+\omega_{n+1},\omega_{n+2})}(g) \geq T$. However we never observed this in realizations of the null-model.

INVARIANCE IN THE SIGNS OF τ WHEN MERGING SEGMENTS

It is easy to prove that:

$$t_{(\omega_{n-1},\omega_n+\omega_{n+1})}(g_n) \quad = \quad t_{(\omega_{n-1},\omega_n)}(g_{n-1}) + \alpha\, t_{(\omega_n,\omega_{n+1})}(g_n), \quad (4.33)$$

where $\alpha = \frac{\omega_{n+1}}{\omega_n+\omega_{n+1}} < 1$. From this it follows that:

- Property 6: If $|\tau_{(\omega_n,\omega_{n+1})}(g_n)| < |\tau_{(\omega_{n-1},\omega_n)}(g_{n-1})|$ then either:

$$\begin{aligned} \text{sign}\big(\tau_{(\omega_{n-1},\omega_n+\omega_{n+1})}(g_{n-1})\big) &= \text{sign}\big(\tau_{(\omega_{n-1},\omega_n)}(g_{n-1})\big), \text{ or} \\ |\tau_{(\omega_{n-1},\omega_n+\omega_{n+1})}(g_{n-1})| &\leq |\tau_{(\omega_n,\omega_{n+1})}(g_n)| \end{aligned} \quad (4.34)$$

What this means is that if we merge two segments that differ with the least significance (remove the discontinuity at g_n), then for the new jump discontinuity evaluated at g_{n-1}, $\tau_{(\omega_{n-1},\omega_n+\omega_{n+1})}(g_{n-1})$ will have the same sign as the old one $\tau_{(\omega_{n-1},\omega_n)}(g_{n-1})$ or else will be even less significant than the discontinuity that was removed.

CONTROL ON THE EXPECTED NUMBER OF LOCAL MAXIMUM SEGMENTS WHEN CLUSTERING

We are now finally in a position to prove that the expected number of local maximum segments will be less than $E/2$ when clustering until all jump discontinuities have a similarity measure below E.

We start with iteration $k = 0$, where every probe is a unique cluster, i.e. the segmented aggregate is exactly the same as the aggregate. At each successive iteration, we merge segments that are separated with the lowest $|\tau| < T$ score. For any threshold T, we can compute $\chi(T)$ at any iteration k and denote it with $\chi^k(T)$. We can also segment on an arbitrary interval of probes $\{q,...,r\}$ and denote $\chi_{\{q,...r\}}(T)$ at iteration k with $\chi^k_{\{q,...r\}}(T)$

We need to prove that:

$$\forall q \forall r \forall k E[\chi^k_{\{q,...,r\}}(T)] \leq (q-r)\Delta\bar{\chi}(T), \quad (4.35)$$

where $T = -\log(\frac{E}{G-1})$.

For the special case where $q = 0, r = G-1$:

$$\begin{aligned} E[\chi^k_{\{0,\ldots,G-1\}}(T)] &\leq (G-1)\Delta\bar{\chi}(T) && \text{Eq. 4.35} \\ &= (G-1)e^{-T} && \text{Eq. 4.27} \\ &= (G-1)e^{\log(E/(G-1))} \\ &= E \end{aligned} \tag{4.36}$$

If we then choose the smallest k for which all τ values at the jump discontinuities are significant, we know that the expected number of local extrema is less than or equal to E. As we observed, the expected number of local maximum segments will be less than or equal to $E/2$.

To prove Eq. 4.35 we use double induction. First on the number of probes $P = r - q + 1$:

Induction hypothesis 1:

$$r - q + 1 < P \implies \forall k E[\chi^k_{\{q,\ldots,r\}}(T)] \leq (q-r)\Delta\bar{\chi}(T) \tag{4.37}$$

From this we need to prove that it also holds for $r - q + 1 = P$. We do so by induction on k.

Suppose $k = 0$. This is the case where we have not yet started merging segments. Every probe is a unique segment. At this point all τ's are at the same scale $w = (1,1)$. It follows directly from the definition of the expected Euler characteristic at a fixed scale that:

$$E[\chi^0_{\{q,\ldots,r\}}(T)] \leq (q-r)\Delta\bar{\chi}(T) \tag{4.38}$$

Now suppose the statement is true for a fixed k.

Induction hypothesis 2:

$$r - q + 1 = P \implies E[\chi^k_{\{q,\ldots,r\}}(T)] \leq (q-r)\Delta\bar{\chi}(T) \tag{4.39}$$

We need to prove that it also holds for $k+1$. For this we will consider three unique scenarios in which segments are merged from iteration k to $k+1$. All other possible scenarios that can be conceived are either straight forward or strictly symmetric. Without loss of generality we will assume $q = 0$ and $r = P-1$ (we simply relabel the indices and this is legal due to Property 1 and 2), that the discontinuity locations at iteration k will be at $\{g_0, g_1, \ldots g_{n-1}, g_n, g_{n+1}, \ldots\}$ and that it is discontinuity g_n that will be merged in iteration $k+1$.

Scenario 1

This scenario is depicted in Fig. 4.13. This scenario is based on two criterion. First, we assume that the right most significant discontinuity g_L before g_{n-1} ($L = \max\{i < n-1 : |\tau_{(\omega_i,\omega_{i+1})}(g_i)| \geq T\}$) has $\tau_{(\omega_L,\omega_{L+1})}(g_L) \geq T$. Second, we assume that $\tau_{(\omega_{n+2},\omega_{n+3})}(g_{n+2}) \leq -T$.

If $|\tau_{(\omega_{n-1},\omega_n)}(g_{n-1})| \geq T$ and $|\tau_{(\omega_{n+1},\omega_{n+2})}(g_{n+1})| \geq T$, $\chi^{k+1}_{\{0,\ldots,P-1\}}(T) \leq \chi^k_{\{0,\ldots,P-1\}}(T)$ (property 6) and by induction hypothesis 2, $E[\chi^{k+1}_{\{0,\ldots,P-1\}}(T)] \leq (P-1)\Delta\bar{\chi}(T)$. Therefore, without loss of generality, let us assume $|\tau_{(\omega_{n+1},\omega_{n+2})}(g_{n+1})| < T$. In this case:

$$\begin{aligned} \chi^{k+1}_{\{0,\ldots,P-1\}}(T) \leq\ & \chi^s_{\{0,\ldots,g_{n-1}\}}(T) + \\ & \chi^t_{\{g_n,\ldots,P-1\}}(T) + \\ & 2U(\tau_{(\omega_n+\omega_{n+1},\omega_{n+2})}(g_{n+1}) \geq T) \end{aligned} \tag{4.40}$$

Here, s and t are the number of iterations required to segment on the intervals $\{0,\ldots,g_{n-1}\}$ and $\{g_n,\ldots,P-1\}$ respectively when considered as separate problems. $U()$ is the indicator function

with range $\{0,1\}$. Due to Property 4, we know that $\tau_{(\omega_n+\omega_{n+1},\omega_{n+2})}(g_{n+1}) \geq T$ implies (with high certainty) that there will be a down-crossing in the interval $\{g_{n+1},...,g_{n+1}+\omega_n\}$. As a consequence,

$$\begin{aligned}\chi^{k+1}_{\{0,...,P-1\}}(T) \quad \leq \quad & \chi^{s}_{\{0,...,g_{n-1}\}}(T) + \\ & \chi^{t}_{\{g_n,...,P-1\}}(T) + \\ & 2\chi^{\downarrow}_{\{g_{n+1},...,g_{n+1}+\omega_n\}}(T),\end{aligned} \tag{4.41}$$

Finally, we can compute the expectation:

$$\begin{aligned}E[\chi^{k+1}_{\{0,...,P-1\}}(T)] \quad \leq \quad & E[\chi^{s}_{\{0,...,g_{n-1}\}}(T)] + \\ & E[\chi^{t}_{\{g_n,...,P-1\}}(T)] + \\ & 2E[\chi^{\downarrow}_{\{g_{n+1},...,g_{n+1}+\omega_n\}}(T)] \\ \leq \quad & (g_{n-1})\Delta\bar{\chi}(T) + \quad \text{(Induction hypothesis 1)} \\ & (P-1-g_n)\Delta\bar{\chi}(T) + \quad \text{(Induction hypothesis 1)} \\ & 2(g_n-g_{n-1})\big(\frac{1}{2}\Delta\bar{\chi}(T)\big) \quad \text{(Eq. 4.27)} \\ = \quad & (P-1)\Delta\bar{\chi}(T)\end{aligned} \tag{4.42}$$

SCENARIO 2

This scenario is depicted in Fig. 4.14. This scenario is base on three criterion. First, we assume that the right most significant discontinuity g_L before g_{n-1} ($L = \max\{i < n-1 : |\tau_{(\omega_i,\omega_{i+1})}(g_i)| \geq T\}$) has $\tau_{(\omega_L,\omega_{L+1})}(g_L) \geq T$. Second, we assume that the left most significant discontinuity g_R after g_{n+2} ($R = \min\{i > n+2 : |\tau_{(\omega_i,\omega_{i+1})}(g_i)| \geq T\}$) has $\tau_{(\omega_R,\omega_{R+1})}(g_R) \leq -T$. The final criteria is $|\tau_{(\omega_{n+2},\omega_{n+3})}(g_{n+2})| < T$.

If $|\tau_{(\omega_{n-1},\omega_n)}(g_{n-1})| \geq T$ and $|\tau_{(\omega_{n+1},\omega_{n+2})}(g_{n+1})| \geq T$, then $\chi^{k+1}_{\{0,...,P-1\}}(T) \leq \chi^{k}_{\{0,...,P-1\}}(T)$ (property 6) and by induction hypothesis 2, $E[\chi^{k+1}_{\{0,...,P-1\}}(T)] \leq (P-1)\Delta\bar{\chi}(T)$. Therefore, without loss of generality, let us assume $|\tau_{(\omega_{n+1},\omega_{n+2})}(g_{n+1})| < T$. In this case:

$$\begin{aligned}\chi^{k+1}_{\{0,...,P-1\}}(T) \quad \leq \quad & \chi^{s}_{\{0,...,g_{n-1}\}}(T) + \\ & \chi^{t}_{\{g_{n+1},...,P-1\}}(T) + \\ & 2U(\tau_{(\omega_n+\omega_{n+1},\omega_{n+2})}(g_{n+1}) \geq T)\end{aligned} \tag{4.43}$$

Note that the only difference between Eq. 4.40 and 4.43 is that we consider the interval $\{g_{n+1},...,P-1\}$ instead of $\{g_n,...,P-1\}$ in the second term. Due to Property 5, we can derive $E[\chi^{k+1}_{\{0,...,P-1\}}(T)] \leq (P-1)\Delta\bar{\chi}(T)$ following the same strategy proposed in scenario 1.

SCENARIO 3

The final scenario that we will consider is depicted in Fig. 4.15. This is the scenario where $\tau_{(\omega_{n-2},\omega_{n-1})}(g_{n-2}) \leq -T$ and $\tau_{(\omega_{n+2},\omega_{n+3})}(g_{n+2}) \leq -T$.

If either $\tau_{(\omega_{n-1},\omega_n)}(g_{n-1}) \geq T$ or $\tau_{(\omega_{n+1},\omega_{n+2})}(g_{n+1}) \geq T$, then $\chi^{k+1}_{\{0,...,P-1\}}(T) \leq \chi^{k}_{\{0,...,P-1\}}(T)$ and by induction hypothesis 2, $E[\chi^{k+1}_{\{0,...,P-1\}}(T)] \leq (P-1)\Delta\bar{\chi}(T)$. Therefore we will assume that

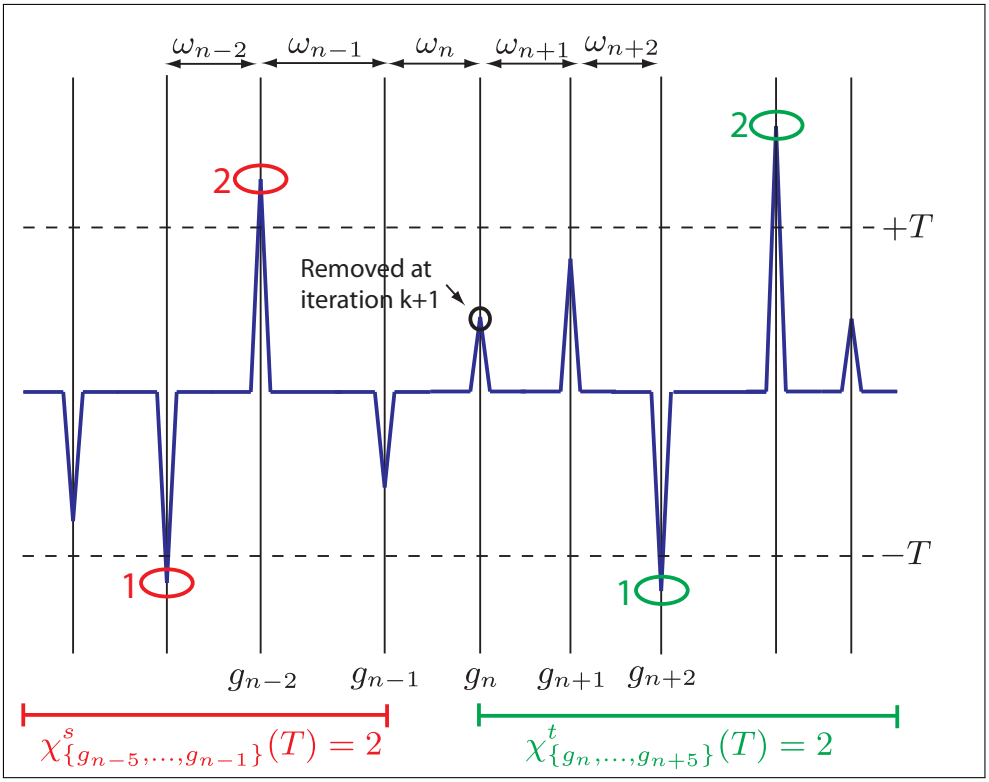

Figure 4.13: **Illustrating Scenario one when merging adjacent segments.** When two segments are merged the encircled break at g_n is removed. This depicts the scenario when the tau-score at break g_{n-2} is above the threshold T and at break g_{n+2} is below $-T$.

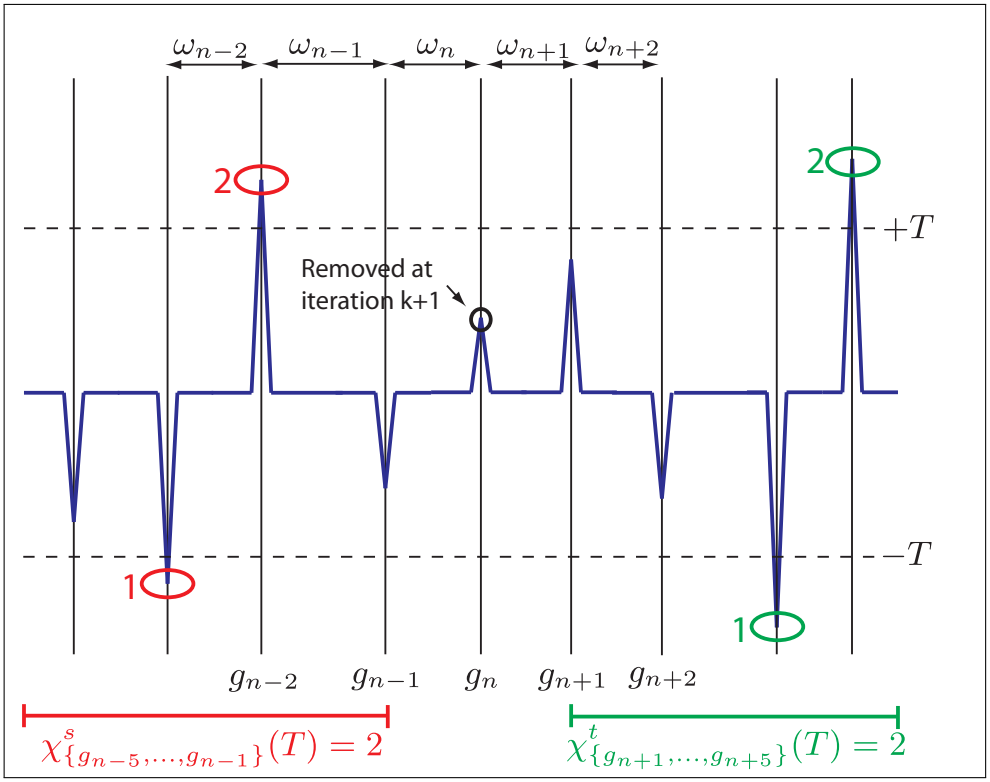

Figure 4.14: **Illustrating Scenario two when merging adjacent segments.** When two segments are merged the encircled break at g_n is removed. This depicts the scenario when the tau-score at break g_{n-2} is above the threshold T, insignificant at g_{n+2} and below $-T$ at break g_{n+3}.

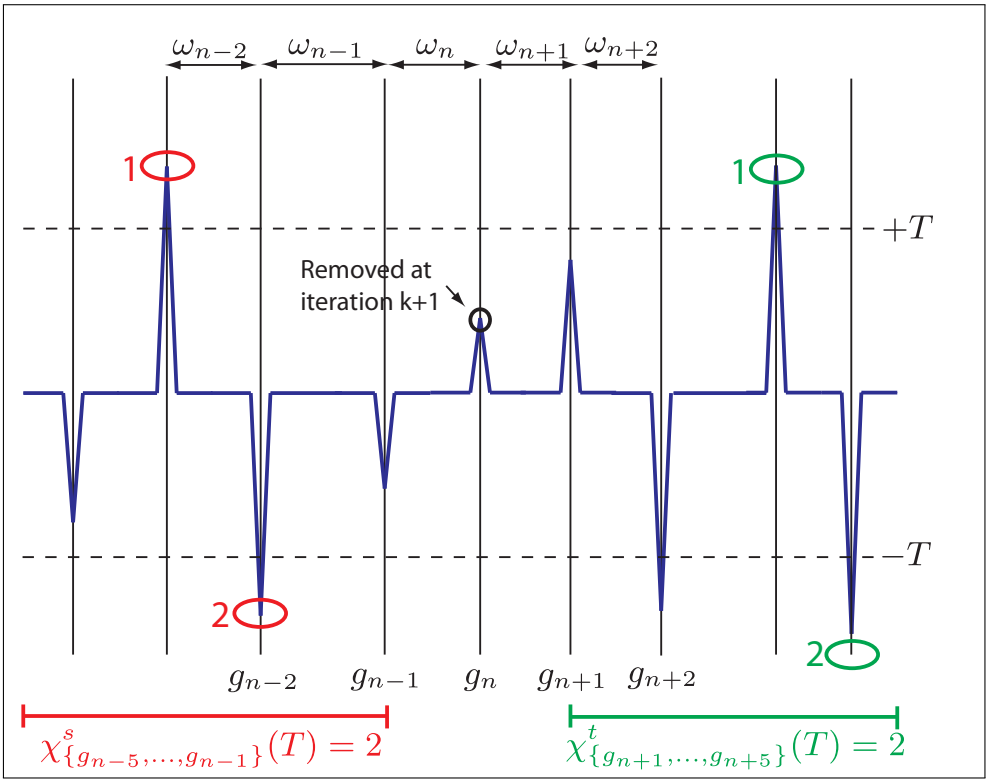

Figure 4.15: **Illustrating Scenario three when merging adjacent segments.** When two segments are merged the encircled break at g_n is removed. This depicts the scenario when the tau-scores at breaks g_{n-2} and g_{n+2} are below $-T$.

neither are above T. In this case:

$$\begin{aligned}\chi^{k+1}_{\{0,\ldots,P-1\}}(T) \quad \leq \quad & \chi^{s}_{\{0,\ldots,g_{n-1}\}}(T) + \\ & \chi^{t}_{\{g_{n+1},\ldots,P-1\}}(T) + \\ & 2U(\tau_{(\omega_{n-1},\omega_n+\omega_{n+1})}(g_{n-1}) \geq T) + \\ & 2U(\tau_{(\omega_n+\omega_{n+1},\omega_{n+2})}(g_{n+1}) \geq T) \end{aligned} \tag{4.44}$$

Due to Property 4, we have:

$$\begin{aligned}\chi^{k+1}_{\{0,\ldots,P-1\}}(T) \quad \leq \quad & \chi^{s}_{\{0,\ldots,g_{n-1}\}}(T) + \\ & \chi^{t}_{\{g_{n+1},\ldots,P-1\}}(T) + \\ & 2\chi^{\downarrow}_{\{g_{n+1},\ldots,g_{n+1}+\omega_n\}}(T) + \\ & 2\chi^{\uparrow}_{\{g_{n-1}-\omega_{n+1},\ldots,g_{n-1}\}}(T) \end{aligned} \tag{4.45}$$

And the result follows when taking expectations.

REFERENCES

[1] E. Van Dyk, M. Hoogstraat, J. Ten Hoeve, M. J. Reinders, and L. F. Wessels, *Rubic identifies driver genes by detecting recurrent dna copy number breaks,* Nature communications **7**, 12159 (2016).

[2] R. McLendon, A. Friedman, D. Bigner, E. G. Van Meir, D. J. Brat, G. M. Mastrogianakis, J. J. Olson, T. Mikkelsen, N. Lehman, K. Aldape, *et al.*, *Comprehensive genomic characterization defines human glioblastoma genes and core pathways,* Nature **455**, 1061 (2008).

[3] Cancer Genome Atlas Research Network, *Comprehensive molecular characterization of clear cell renal cell carcinoma,* Nature **499**, 43 (2013).

[4] O. M. Rueda and R. Diaz-Uriarte, *Finding recurrent copy number alteration regions: a review of methods,* Current Bioinformatics **5**, 1 (2010).

[5] A. Ben-Dor, D. Lipson, A. Tsalenko, M. Reimers, L. O. Baumbusch, M. T. Barrett, J. N. Weinstein, A.-L. Børresen-Dale, and Z. Yakhini, *Framework for identifying common aberrations in dna copy number data,* in *Research in Computational Molecular Biology* (Springer, 2007) pp. 122–136.

[6] R. Beroukhim, G. Getz, L. Nghiemphu, J. Barretina, T. Hsueh, D. Linhart, I. Vivanco, J. C. Lee, J. H. Huang, S. Alexander, *et al.*, *Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma,* Proceedings of the National Academy of Sciences **104**, 20007 (2007).

[7] C. H. Mermel, S. E. Schumacher, B. Hill, M. L. Meyerson, R. Beroukhim, and G. Getz, *Gistic2. 0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers,* Genome Biol **12**, R41 (2011).

[8] S. J. Diskin, T. Eck, J. Greshock, Y. P. Mosse, T. Naylor, C. J. Stoeckert, B. L. Weber, J. M. Maris, and G. R. Grant, *Stac: A method for testing the significance of dna copy number aberrations across multiple array-cgh experiments,* Genome research **16**, 1149 (2006).

[9] A. Magi, M. Benelli, S. Yoon, F. Roviello, and F. Torricelli, *Detecting common copy number variants in high-throughput sequencing data by using jointslm algorithm,* Nucleic acids research , gkr068 (2011).

[10] S. Morganella, S. M. Pagnotta, and M. Ceccarelli, *Finding recurrent copy number alterations preserving within-sample homogeneity,* Bioinformatics **27**, 2949 (2011).

[11] A. Niida, S. Imoto, T. Shimamura, and S. Miyano, *Statistical model-based testing to evaluate the recurrence of genomic aberrations,* Bioinformatics **28**, i115 (2012).

[12] F. Sanchez-Garcia, U. D. Akavia, E. Mozes, and D. Pe'er, *Jistic: identification of significant targets in cancer,* BMC bioinformatics **11**, 189 (2010).

[13] E. van Dyk, M. J. Reinders, and L. F. Wessels, *A scale-space method for detecting recurrent dna copy number changes with analytical false discovery rate control,* Nucleic acids research **41**, e100 (2013).

[14] V. Walter, A. B. Nobel, and F. A. Wright, *Dinamic: a method to identify recurrent dna copy number aberrations in tumors,* Bioinformatics **27**, 678 (2011).

[15] H.-T. Wu, I. Hajirasouliha, and B. J. Raphael, *Detecting independent and recurrent copy number aberrations using interval graphs,* Bioinformatics **30**, i195 (2014).

[16] S. K. Rao, J. Edwards, A. D. Joshi, I.-M. Siu, and G. J. Riggins, *A survey of glioblastoma genomic amplifications and deletions,* Journal of neuro-oncology **96**, 169 (2010).

[17] F. Sanchez-Garcia, P. Villagrasa, J. Matsui, D. Kotliar, V. Castro, U.-D. Akavia, B.-J. Chen, L. Saucedo-Cuevas, R. R. Barrueco, D. Llobet-Navas, *et al.*, *Integration of genomic data enables selective discovery of breast cancer drivers,* Cell **159**, 1461 (2014).

[18] S. A. Forbes, N. Bindal, S. Bamford, C. Cole, C. Y. Kok, D. Beare, M. Jia, R. Shepherd, K. Leung, A. Menzies, *et al.*, *Cosmic: mining complete cancer genomes in the catalogue of somatic mutations in cancer,* Nucleic acids research **39**, D945 (2010).

[19] N. L. Solimini, Q. Xu, C. H. Mermel, A. C. Liang, M. R. Schlabach, J. Luo, A. E. Burrows, A. N. Anselmo, A. L. Bredemeyer, M. Z. Li, *et al.*, *Recurrent hemizygous deletions in cancers may optimize proliferative potential,* Science **337**, 104 (2012).

[20] A. B. Heimberger, D. Suki, D. Yang, W. Shi, and K. Aldape, *The natural history of egfr and egfrviii in glioblastoma patients,* Journal of translational medicine **3**, 38 (2005).

[21] T. Santarius, J. Shipley, D. Brewer, M. Stratton, and C. Cooper, *A census of amplified and overexpressed human cancer genes,* Nature Reviews Cancer , 59 (2010).

[22] P. Schouten, A. Grigoriadis, T. Kuilman, H. Mirza, J. Watkins, S. Cooke, E. van Dyk, T. Severson, O. Rueda, M. Hoogstraat, *et al.*, *Robust brca1-like classification of copy number profiles of samples repeated across different datasets and platforms,* Molecular Oncology , 1274 (2015).

[23] A. Fungtammasan, E. Walsh, F. Chiaromonte, K. Eckert, and M. KD, *A genome-wide analysis of common fragile sites: what features determine chromosomal instability in the human genome?* Genome Research , 993 (2012).

[24] T. Nichols and S. Hayasaka, *Controlling the familywise error rate in functional neuroimaging: a comparative review,* Statistical methods in medical research **12**, 419 (2003).

[25] R. J. Adler and A. M. Hasofer, *Level crossings for random fields,* The Annals of Probability **4**, 1 (1976).

[26] Y. Benjamini and D. Yekutieli, *The control of the false discovery rate in multiple testing under dependency,* Annals of statistics , 1165 (2001).

[27] K. J. Worsley, A. C. Evans, S. Marrett, and P. Neelin, *A three-dimensional statistical analysis for cbf activation studies in human brain,* Journal of Cerebral Blood Flow & Metabolism **12**, 900 (1992).

[28] K. J. Worsley, *Estimating the number of peaks in a random field using the hadwiger characteristic of excursion sets, with applications to medical images,* The Annals of Statistics , 640 (1995).

[29] K. J. Worsley, S. Marrett, P. Neelin, A. C. Vandal, K. J. Friston, A. C. Evans, *et al.*, *A unified statistical approach for determining significant signals in images of cerebral activation,* Human brain mapping **4**, 58 (1996).

5

DNA COPY NUMBER SEGMENTATION WITH CLUSTERING ON THE EXPECTED EULER CHARACTERISTIC

Ewald van Dyk, Marcel J.T. Reinders & Lodewyk F.A. Wessels

Manuscript in preparation.

A large number of methods exist for segmenting DNA copy number profiles measured on microarray, SNP and sequencing platforms. Many of these algorithms perform very well on idealized simulation models. However, on real data sets, the performance varies for different algorithms. This makes it particularly hard for practitioners to select the appropriate algorithm for their particular needs. Recently, a comparison pipeline called Jointseg *was introduced that allows for objective comparison of algorithms on realistic simulations based on data from specific platforms. An important problem is that most of these algorithms employ different parameters that need to be tuned to ensure comparable performance.* Jointseg *resolves this issue to allow for objective comparison and subsequently suggests that there is no single best segmentation algorithm when comparing the performance on data from the SNP6 Affymetrix and Illumina platforms or when varying the tumor fraction. We introduce* RUBICseg*, a new segmentation algorithm that employs a single intuitive parameter: the false discovery rate of selected break locations. To perform the segmentation,* RUBICseg *not only employs the copy number profile itself, but also the decrease in heterozygosity profile derived from B-allele frequencies. We illustrate within the* Jointseg *framework that* RUBICseg *is consistently one of the top performing algorithms for samples with tumor fractions above* 70%*. These results hold true for both the Affymetrix and Illumina platforms.*

5.1. INTRODUCTION

DNA copy number alterations are a hallmark of cancer[1] and many of these aberrations play a pivotal role in tumor development and progression. As a consequence, it is important to 1) accurately identify the genomic locations of copy number breakpoints and 2) determine the type of aberrations that occur. These include, but are not limited to homozygous deletions, hemizygous deletions, copy-neutral loss of heterozygosity and copy number gains. Many different technological platforms exist for measuring copy number profiles including array-comparative genomic hybridization (aCGA)[2, 3], single nucleotide polymorphism (SNP) arrays[4] and next generation sequencing.

Typically, DNA copy number states in a tumor sample are measured at fixed genomic locations (probes) and the changes are typically expressed as a log ratio relative to a normal cell reference. The top (c-channel) panel in Fig. 5.1 illustrates a typical copy number profile when probe measurements (in log ratio form) are sorted on a reference genome. Normal human cells are diploid, which means that each cell contains two copies of DNA inherited from two different biological parents. In tumor cells, DNA segments of varying sizes (from a few kilobases to whole chromosome arms) are often deleted or copied multiple times[5], which implies that the true biological signal in the copy number profile should be piecewise constant, where positive (negative) log ratios indicate segments with a net gain (loss) in alleles relative to a normal diploid cell. Many platforms also have SNP probes that allows one to see exactly which allele gets deleted or amplified. This extra dimension is usually represented in terms of the B allele frequency (BAF)[6] as indicated in the middle (b-channel) panel of Fig. 5.1. Note that for any particular copy number segment there will be two or more modes in the b-channel. This happens because, for any given patient, it is unknown which chromosome in the diploid pair the B allele belongs to. In any copy number state there will always be two modes centered close to zero and one, representing the homozygous AA and BB germline alleles, respectively. BAF measurements associated with germline homozygous alleles are uninformative and need to be removed from the analysis. We refer to the remaining SNPs as informative. Probes contributing to other modes can only arise when the germline is heterozygous (AB). For example, if a single copy is gained (from only one parental chromosome) in a heterozygous germline region, there will be two modes in the BAF signal representing AAB and ABB respectively representing BAF frequencies of 1/3 and 2/3. It should be noted that in such cases BAF frequency measurements will usually be biased towards 0.5 due to contamination of

the tumor sample with normal cells[6] (usually represented as the tumor fraction). The contamination of the tumor tissue with normal cells will also result in non-integer copy number measurements. The b-channel should not be directly modelled as a piecewise constant signal due to its multimodal distribution. Instead, we first transform the informative BAF measurement into what is called the 'decrease in heterozygosity' $d = 2|b - \frac{1}{2}|$ for all SNPs that are heterozygous[7] in the germline (lower panel in Fig. 5.1).

From Fig. 5.1 it is clear that both the c- and d-channels are piecewise constant. The break locations in both channels are the same since they belong to the same copy number events. One should be able to predict break locations better when considering these channels jointly for two reasons. First, segmenting on the d-channel alone will lead to a loss in statistical power since not all probes are informative (e.g. those that are germline homozygous) and therefore leads to a reduction in probe resolution. Nevertheless, the d-channel should not be ignored since it is sometimes easier to detect sudden jumps in the d-channel than in the c-channel (compare the top and bottom panel in Fig. 5.1) relative to their respective noise levels. Second, even if it were possible to attain full resolution for the d-channel, one has more statistical power when jointly considering two dimensions, instead of either dimension alone.

Various statistical methods (called segmentation algorithms) have been developed in the last decade to estimate the locations of copy number break points based on the c-channel (overall copy number), d-channel (derived from BAF) or both[8–16]. Most of these algorithms can be classified into four categories as explained by Pierre-Jean *et. al.*[17]:

- Hidden Markov model (HMM) methods. These algorithms are based on the assumption that there are a limited number of hidden discrete copy number states[16, 18–20].
- Multiple change-point methods where it is assumed that the observed signal is effected by abrupt changes and that the signal between the changes is homogenous. *Cghseg*[9] is an example of such a method.
- Fused lasso methods. These methods assume that successive probe measurements should mostly have the same estimate and this is achieved by regularizing with an L_1 penalty on successive differences. One example of such an algorithm is *GFLars*[14, 21].
- Recursive segmentation algorithms. These methods also assume that the copy number profile is a piecewise constant signal and that the segments can be identified by recursively splitting larger segments into smaller ones. Algorithms in this category include *CBS*[22], *PSCBS*[12] and *RBS*[13].

Since there are many segmentation algorithms, it is not always clear which one to use for any given dataset. Recently, Pierre-Jean *et. al.*[17] developed *Jointseg*, a comparison pipeline that 1) accurately simulates copy number profiles from various technological platforms and also 2) introduced a way in which different segmentation algorithms can be compared in a supervised manner. This provides a useful tool for practitioners to choose the correct segmentation algorithm for their specific application and also allows for rapid testing of new segmentation methods. Notably, with *JointSeg* it was illustrated that there is no single segmentation algorithm that outperforms all others for all different technological platforms, but that methods that use both the c- and d-channel frequently outperform others.

Here we introduce a new multiple change-point segmentation algorithm that jointly segments on both the c and d channel and relies on only one intuitive parameter: the false discover rate (FDR) of detected breakpoints. This algorithm employs a powerful statistic called the expected Euler characteristic that allows us to accurately recover true DNA break locations. We compared our approach to state of the art algorithms using *Jointseg* and show that it often outperforms existing methods for high (but realistic) tumor fractions (> 70%) for both the Affymetrix and Illumina SNP platforms.

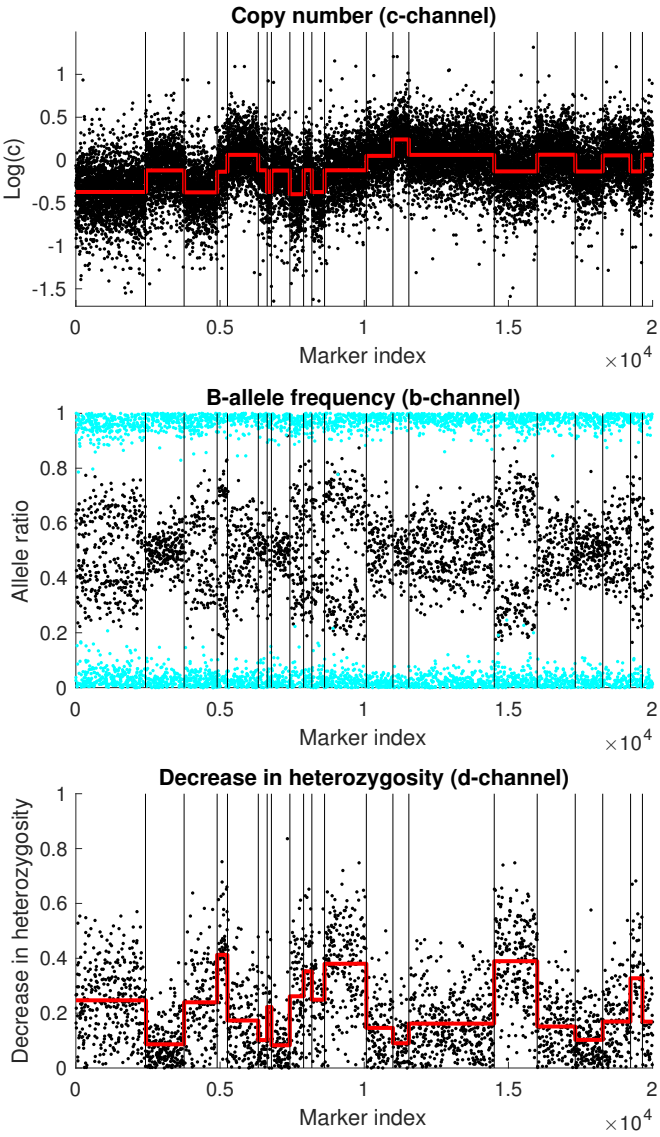

Figure 5.1: Illustration of a DNA copy copy number profile derived from the Affymetrix SNP6 platform. The top panel shows the log ratio (with respect to normal) of the copy number measurements for markers sorted across the genome in black (c-channel). A *RUBICseg* segmentation is depicted by the piece-wise constant function in red. The true break locations are indicated by the vertical black lines. The middle panel shows the B-allele frequency for SNP markers (approximately 1/2 of the markers). Uninformative markers due to germline homozygosity are shown in cyan. The bottom panel shows the loss in heterozygosity (d-channel) for the remaining informative markers with *RUBICseg* segmentation in red.

5.2. METHODS

5.2.1. *Jointseg* DATASETS.

We used *Jointseg* to simulate realistic copy number profiles derived from two copy number platforms:

- Affymetrix SNP6.0 microarray. Simulations were derived from lung cancer data accessible through the NCBI GEO database[23], accession GSE29172. Lung cancer samples were mixed with matched blood to simulate tumor fractions of 50, 70 and 100%.
- Illumina HumanCNV-Duo v1. Simulations were derived from breast cancer data accessible through the NCBI GEO database[23], accession GSE11976. Breast cancer samples were mixed with matched blood to simulate tumor fractions of 50, 79 and 100%.

Approximately half of the probes in the Affymetrix platform are SNP probes (from which b-allele frequencies can be derived). Since approximately 1/3 of all SNPs are heterozygous in a given human's germline, we can derive decrease in heterozygosity scores (d-channel) for only 1/6 of the probes. In contrast, the Illumina platform has b-allele frequencies for each probe, and therefore has informative measurements for 1/3 of the probes. For a detailed description of these datasets and how DNA copy number datasets are simulated, see Pierre-Jean *et. al.*[17].

5.2.2. PERFORMANCE MEASURE OF *Jointseq* EMPLOYED IN COMPARISON.

For each generated copy number profile, segmentation algorithm tested and tolerance threshold employed, one can compute the true positive and false positive break rates. Unfortunately these value pairs cannot be compared for different algorithms since they represent different points on their receiver operating (ROC) curves. Ideally, one would like to compare the area under these curves (AUC). To tune each algorithm's parameters by sweeping across the parameter space and using the ROC as readout is unrealistic. For this reason, *Jointseg* sweeps across the different number of called breaks in each algorithm by successively removing breaks in such a way as to minimise the mean square error. For example, if *CBS* called 30 breaks, *Jointseg* will first select only one break that minimises the mean square error and record the true and false positive counts. Next, joinseg selects the two best breaks, etc. This process is repeated until 10 errors are detected. In this way a partial ROC curve is reconstructed for each algorithm (up to 10 false positives) with false positive counts on the x-axis and true positive counts on the y-axis. Finally, as a performance measure, the partial AUCs computed (the area under the partial ROC divided by $10 \times B$, where B is the number of true positive breaks). In this scheme a pAUC of 0 means that the first 10 best breaks are false positives, whereas a perfect pAUC of 1.0 means that the top B breaks are true positives.

5.2.3. GAUSSIAN DATASETS.

In addition to the *Jointseg* datasets, we also simulated artificial copy number profiles with Gaussian measurement noise at different signal to noise ratio (SNR) levels. For each SNR level, we simulate 10 copy number profiles with 20000 probes and 20 breakpoints uniformly distributed across probe measurements. Each segment in the c-channel takes on mean log copy number segment values derived from the Affymetrix SNP6.0 platform for tumor fractions of 100%. Note that we use tumor fractions of 100% as the Gaussian noise is added to this clean profile. Decrease in heterozygosity (d-channel) levels were chosen similarly. Specifically, these segment levels were randomly chosen from the following copy number states: normal, gain of one copy, hemizygous deletion, homozygous deletion and copy-neutral LOH. This results in noise-free bi-variate piecewise constant profiles of the c- and d-channels. We add bi-variate Gaussian noise with zero mean

and covariance

$$\begin{aligned} \Sigma &= \begin{bmatrix} \sigma_c^2 & r\sigma_c\sigma_d \\ r\sigma_c\sigma_d & \sigma_d^2 \end{bmatrix}, \text{where} \\ \sigma_c^2 &= \frac{\text{Var}[c]}{\text{SNR}} \\ \sigma_d^2 &= \frac{\text{Var}[d]}{\text{SNR}} \\ r &= 0.3 \end{aligned} \tag{5.1}$$

Var[.] represents the variance in the clean c and d profiles. r was arbitrarily chosen and represents the Pearson correlation between the noise in the c and d channels. Finally, in each profile, we randomly set 83.3% (5/6) of the probe values in the d-channel to NA to simulate uninformative measures.

5.2.4. MODEL ASSUMPTIONS.

Many segmentation algorithms assume that a given DNA copy number (c-channel) profile is the sum of a piecewise constant signal with additive, constant variance, zero mean Gaussian noise and that the noise is uncorrelated between different probes. Together, the c- and d-channel can also be regarded as a two dimensional piecewise constant signal. We will also assume that the noise is additive and a zero mean stationary bi-variate Gaussian process with constant covariance matrix (between the c- and d-measurements) and is uncorrelated between different probes (i.e. white noise). We will show that the covariance matrix of the noise can be estimated accurately from the data and can therefore be regarded as known (i.e. we will not model the uncertainty in the covariance estimate). The goal of the segmentation algorithm is therefore to recover the piecewise constant signal. This is equivalent to detecting the location of breakpoints. The piecewise constant segments are then simply the regions between breakpoints and we estimate the c- and d-channel amplitudes in each segment by averaging the measurements between the breakoints.

5.2.5. APPROACH.

We perform an iterative clustering approach in which neighboring segments are merged if there is no indication of a breakpoint between them, starting with every probe as a separate segment. Segments that are least likely to be separated by a breakpoint are merged first. Hence, to decide whether two segments need to be merged we need an estimate of the likelihood that they are separated by a breakpoint. We base this on the difference between the means of the c- and d-measurements in the two neighboring segments.

The significance of the difference in segment means depends on the noise in the profile and the length of the segments. As there are many possible break locations to choose from, we need to correct for multiple testing to calculate the appropriate significance level. For example, for fixed (adjacent) segment lengths (denoted by w_L and w_R probes for the left and right segments respectively), there are theoretically $G-1$ possible break locations, where G is the number of probes in the profile (we ignore boundary effects since the noise can be modelled beyond the boundaries). The differences in segment means between adjacent break locations are generally correlated (especially for large values of w_L and w_R) so that a simple family wise error rate (FWER) correction like the Bonferroni procedure will lead to significant power loss. Therefore, we propose to use a different correction measure called the expected Euler characteristic that is a tight upper bound for the FWER, while accounting for the correlation between adjacent break measurements.

To summarize, iterative clustering is performed based on the expected Euler characteristic, where higher values (least significant) are merged first. Clustering continues until all remaining expected Euler characteristic measures (associated with the remaining break locations) are

sufficiently low. Finally, the exact locations of the boundaries of the segments are fine-tuned by minimizing the L^2 distance between a fitted piecewise-constant function and the actual measurements. This represents the final segmentation of the profile.

In the following sections, we discuss the different steps in detail (see Fig. 5.2): 1) estimating the covariance of the measurement noise; 2) defining a break measure based on the difference in adjacent segment means; 3) computing the dependency between break measures at adjacent break locations (needed for multiple test correction); 4) choosing a significance measure used for iterative clustering; 5) computing the expected number of false positive breaks at any given iterative step in clustering 6) introducing the Euler characteristic; 7) computing the expected Euler characteristic; and 8) fine tuning break locations.

Note that, as stated earlier, we do not have d-measurements for all probes, therefore we can regard these as missing values. For now, we will only consider the special (albeit unrealistic) case where all measurements in the d-channel are known. In the supplementary methods, we extend the method to account for missing data points.

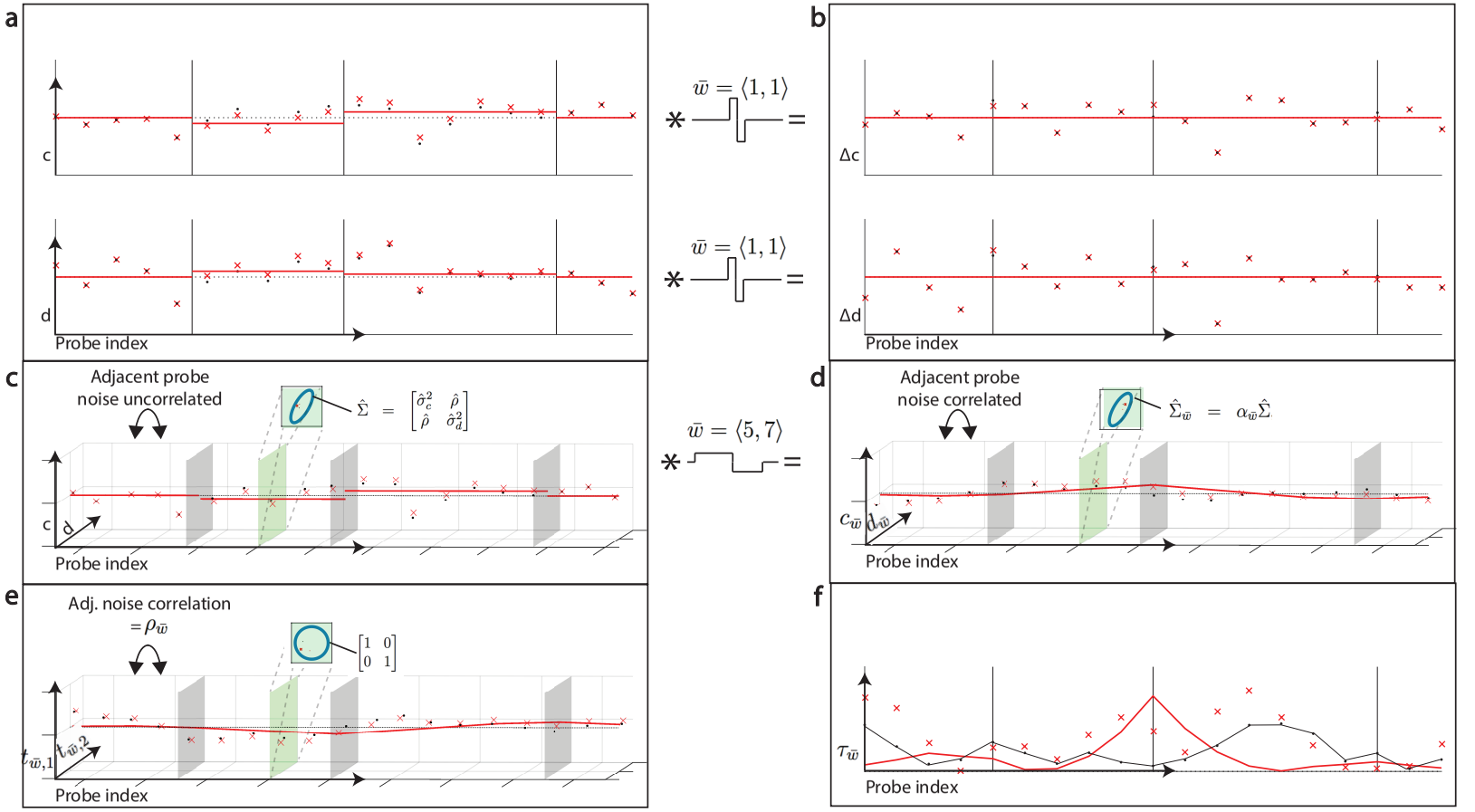
a
c
d
Probe index
$\bar{w} = \langle 1, 1 \rangle$
$\bar{w} = \langle 1, 1 \rangle$
$\bar{w} = \langle 5, 7 \rangle$
b
Δc
Δd
Probe index
c
Adjacent probe
noise uncorrelated
$\hat{\Sigma} = \begin{bmatrix} \hat{\sigma}_c^2 & \hat{\rho} \\ \hat{\rho} & \hat{\sigma}_d^2 \end{bmatrix}$
Probe index
d
Adjacent probe
noise correlated
$\hat{\Sigma}_{\bar{w}} = \alpha_{\bar{w}} \hat{\Sigma}$
$c_{\bar{w}}$
$d_{\bar{w}}$
Probe index
e
Adj. noise correlation
$= \rho_{\bar{w}}$
$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$
$t_{\bar{w},1}$
$t_{\bar{w},2}$
Probe index
f
$\tau_{\bar{w}}$
Probe index

Figure 5.2: Illustration of the steps involved in computing the break measure $\tau_{\bar{w}}$ for the fixed scale $\bar{w} = \langle 5, 7 \rangle$. In (**a**), we illustrate an example piecewise constant profile with Gaussian noise added to both the c- and d-channel. The (piecewise constant) signal is illustrated in solid red, the noise profile is shown as black dots and the sum of the signal and noise (the measured profile) is shown with red crosses. (Note that this color coding holds for all panels). In (**b**), we show the lag-one difference profile where noise and observed measures only differ at the break locations which allows for accurate estimation of the noise covariance. Note that computing the lag one difference is equivalent to convolving the profile with a kernel function as indicated between **a** and **b** (the asterisk represent the convolution operator). In (**c**), we represent (**a**) in three dimensions. It is important to note that the c- and d-channel noise measurements are correlated (see the green plane), but noise measurements between different probes are uncorrelated. In (**d**), we convolve the example profile with a kernel representing the scale $\bar{w} = \langle 5, 7 \rangle$. Due to the smoothing, adjacent probes will be correlated. In (**e**), we rotate and scale the c- and d-channels to ensure that the resulting noise covariance in each probe is the identity matrix. The Pearson correlation between adjacent probes is $\rho_{\bar{w}}$ as computed in Eq. 5.10. In (**f**), we show the one dimensional $\tau_{\bar{w}}$ profile that we use as a break measure.

5.2.6. ESTIMATING THE NOISE COVARIANCE.

The first step in recovering the piecewise constant signal is to model the noise. We regard the measurement noise as a zero mean, stationary and bi-variate Gaussian process. We also assume that the noise is uncorrelated between different probes. Therefore we only have to estimate the joint covariance matrix of noise measurements on the c- and d-channel. If we assume that the number of breakpoints (B) in the profile are small compared to the total number of probes, we can do this accurately by considering the lag-one difference profile (Fig. 5.2a-b). Note that under the assumption of a small number of breakpoints, the lag-one difference operation removes most of the signal, retaining mostly noise. Also note that since the noise in adjacent probes is uncorrelated, the variance in the lag-one difference noise profile will be twice the variance of that in a single probe. As a consequence, we can accurately (albeit conservatively) estimate the c-channel noise variance as half the noise variance in the lag-one profile[24]. Similar arguments apply for the d-channel and the covariance between the c- and d-channel. We estimate the noise covariance $\hat{\Sigma}$ as follows (Fig. 5.2c):

$$\begin{aligned}
\hat{\Sigma} &= \begin{bmatrix} \hat{\sigma}_c^2 & \hat{\rho} \\ \hat{\rho} & \hat{\sigma}_d^2 \end{bmatrix}, \text{where} \\
\hat{\sigma}_c^2 &= \frac{1}{2(G-1)} \sum_{i=1}^{G-1} (c_{i+1} - c_i)^2 \\
\hat{\sigma}_d^2 &= \frac{1}{2(G-1)} \sum_{i=1}^{G-1} (d_{i+1} - d_i)^2 \\
\hat{\rho} &= \frac{1}{2(G-1)} \sum_{i=1}^{G-1} (c_{i+1} - c_i)(d_{i+1} - d_i),
\end{aligned} \tag{5.2}$$

where c_i and d_i are probe measurements in the c- and d-channel, respectively, at probe index i. G is the total number of probes in the profile.

5.2.7. DEFINING A BREAK MEASURE BETWEEN ADJACENT SEGMENTS.

In each iteration of the clustering process we need to decide whether two adjacent segments will be joined or not. We will base this decision on the break measure of adjacent segments. Specifically, to see if there is a break at position g, we need to subtract segment means from the left and right of g. Therefore, a mean difference measurement for any given profile depends on three parameters: the break index g and the two adjacent segment lengths denoted by w_L and w_R. We represent the width parameters as a single vector $\bar{w} = \langle w_L, w_R \rangle$ and refer to this vector as the 'scale' at which breaks are detected.

First, for both channels we estimate the mean difference between the left and right segments:

$$\begin{aligned}
c_{\bar{w}}(g) &= \frac{1}{w_R} \sum_{i \in R_{\bar{w}}(g)} c_i - \frac{1}{w_L} \sum_{i \in L_{\bar{w}}(g)} c_i \\
d_{\bar{w}}(g) &= \frac{1}{w_R} \sum_{i \in R_{\bar{w}}(g)} d_i - \frac{1}{w_L} \sum_{i \in L_{\bar{w}}(g)} d_i,
\end{aligned} \tag{5.3}$$

where

$$\begin{aligned}
L_{\bar{w}}(g) &= \{g - w_L + 1, ..., g\} \\
R_{\bar{w}}(g) &= \{g + 1, ..., g + w_R\}
\end{aligned} \tag{5.4}$$

Second, we estimate the covariance matrix of $c_{\bar{w}}(g)$ and $d_{\bar{w}}(g)$ under the null hypothesis, i.e. when there is no break in $L_{\bar{w}}(g) \cup R_{\bar{w}}(g)$. In this case, $c_{\bar{w}}(g)$ and $d_{\bar{w}}(g)$ will be zero mean jointly Gaussian. The covariance matrix of $c_{\bar{w}}(g)$ and $d_{\bar{w}}(g)$ is then a constant multiple of $\hat{\Sigma}$ (Fig. 5.2d):

$$\begin{aligned} \hat{\Sigma}_{\bar{w}} &= \alpha_{\bar{w}} \hat{\Sigma}, \text{ where} \\ \alpha_{\bar{w}} &= \frac{1}{w_L} + \frac{1}{w_R} \end{aligned} \tag{5.5}$$

Next, we decorrelate and scale $c_{\bar{w}}(g)$ and $d_{\bar{w}}(g)$ to obtain two new variables $t_{\bar{w},1}(g)$ and $t_{\bar{w},2}(g)$ (Fig. 5.2e):

$$\bar{t}_{\bar{w}}^T(g) = [c_{\bar{w}}(g), d_{\bar{w}}(g)] V D^{1/2}, \tag{5.6}$$

where $\bar{t}_{\bar{w}}(g)$ is a 2×1 column vector with components $t_{\bar{w},1}(g)$ and $t_{\bar{w},2}(g)$, V contains the eigenvectors in column format of $\hat{\Sigma}_{\bar{w}}^{-1}$, D is the diagonal matrix with eigenvalue entries and $()^T$ represents the transposition operation. The advantage of using $t_{\bar{w},1}(g)$ and $t_{\bar{w},2}(g)$ instead of $c_{\bar{w}}(g)$ and $d_{\bar{w}}(g)$ is that they are both normalized (variance one) and identically and independently distributed under the null.

Finally, we define our break measure as the square of the Euclidean distance of $\bar{t}_{\bar{w}}(g)$ from the origin (Fig. 5.2f):

$$\tau_{\bar{w}}(g) = \bar{t}_{\bar{w}}^T(g) \bar{t}_{\bar{w}}(g) \tag{5.7}$$

Note that $\tau_{\bar{w}}(g)$ follows a chi-square distribution with two degrees of freedom under the null, i.e. when there are no breaks in $L_{\bar{w}}(g) \cup R_{\bar{w}}(g)$).

5.2.8. ESTIMATING DEPENDANCY BETWEEN ADJACENT BREAK MEASURES.

Let us assume that $\bar{w}$ is fixed and we perform a test at every position g for a potential break employing our break measure. Then we need to correct for multiple testing across g. If the break measures $\tau_{\bar{w}}(g)$ were independent for different g's, a Bonferroni correction would suffice. However, this is not the case due to overlapping windows and in this section we set out to compute the dependency between $\tau_{\bar{w}}(g)$ and $\tau_{\bar{w}}(g+1)$. In the section describing the expected Euler characteristic, we show how to correct for multiple testing using this information.

For a fixed $\bar{w}$ and under the null, let 1) $\tau_{\bar{w}}(g)$ be a realization of $T^2_{\bar{w}}$, a stationary, discrete, bivariate, two degree of freedom, chi-square random process and 2) $t_{\bar{w},1}(g)$ and $t_{\bar{w},2}(g)$ be realizations of $T_{\bar{w},1}$ and $T_{\bar{w},2}$, two independent and identical Gaussian processes with unit variance. Then, it follows from the definition of $\tau_{\bar{w}}(g)$ that $T^2_{\bar{w}}$ is given by

$$T^2_{\bar{w}} = \left(T_{\bar{w},1}\right)^2 + \left(T_{\bar{w},2}\right)^2 \tag{5.8}$$

Since $T_{\bar{w},1}$ and $T_{\bar{w},2}$ are identical and independent processes and $T^2_{\bar{w}}$ is fully determined by these, the dependency between $T^2_{\bar{w}}(g)$ and $T^2_{\bar{w}}(g+1)$ is fully characterised by either $T_{\bar{w},1}$ or $T_{\bar{w},2}$ and we drop the subscripts and work with a single Gaussian process denoted by $T_{\bar{w}}$. In fact, when computing the expected Euler characteristic we only need to compute the correlation between $T_{\bar{w}}(g)$ and $T_{\bar{w}}(g+1)$, denoted by $\rho_{\bar{w}}$.

Under the null, $T_{\bar{w}}(g)$ and $T_{\bar{w}}(g+1)$ can be written as a linear combination of i.i.d. Gaussian

variables Z_i, each with zero mean:

$$\begin{aligned} T_{\bar{w}}(g) &= \frac{1}{w_R}\sum_{i\in R_{\bar{w}}(g)} Z_i - \frac{1}{w_L}\sum_{i\in L_{\bar{w}}(g)} Z_i \\ T_{\bar{w}}(g+1) &= \frac{1}{w_R}\sum_{i\in R_{\bar{w}}(g+1)} Z_i - \frac{1}{w_L}\sum_{i\in L_{\bar{w}}(g+1)} Z_i \\ \rho_{\bar{w}} &= E[T_{\bar{w}}(g)T_{\bar{w}}(g+1)], \end{aligned} \tag{5.9}$$

where $E[.]$ is expectation. Note that $E[Z_iZ_j] = 0$ for every $i \neq j$ and that the variance $E[Z_iZ_i]$ equals $\frac{1}{1/w_R + 1/w_L}$ (since the variance of $T_{\bar{w}}(g)$ is unity). From these facts, we can show that:

$$\rho_{\bar{w}} = \frac{\frac{w_R-1}{w_R^2} + \frac{w_L-1}{w_L^2} - \frac{1}{w_L w_R}}{\frac{1}{w_R}+\frac{1}{w_L}} \tag{5.10}$$

5.2.9. A SIGNIFICANCE MEASURE FOR ITERATIVE CLUSTERING.

Note that at a given iteration of the clustering, we would like to compare the significance of breaks at different scales in order to decide which breaks to merge. As the break measure is not comparable between different scales, we define a p-value, which is comparable between different scales. Specifically, we define scale dependent decreasing functions $p_{\bar{w}} : \mathbb{R} \to \mathbb{R}$ which take our break measure as input and outputs a significance value, with values close to zero representing high significance (like a p-value). In the supplementary section, we show that a natural choice for these functions is:

$$\begin{aligned} p_{\bar{w}}(\tau) &= P[M_{\bar{w}} \geq \tau], \text{where} \\ M_{\bar{w}} &= \max_{g\in\{1,\ldots,G\}} T^2_{\bar{w}}(g) \end{aligned} \tag{5.11}$$

$M_{\bar{w}}$ is a scale dependent random variable that tracks the maximum value of the process $T^2_{\bar{w}}$ and $P[\cdot]$ represents probability.

5.2.10. COMPUTING THE EXPECTED NUMBER OF FALSE POSITIVE BREAKS.

Suppose we performed iterative clustering until there are b breakpoints (i.e. $b+1$ clusters). Label the break locations $g_1, g_2, \ldots, g_b$ and their respective scales with $\bar{w}_1, \bar{w}_2, \ldots, \bar{w}_b$, where $\bar{w}_i = \langle w_{L,i}, w_{R,i}\rangle$ for all $i \in \{1,2,\ldots,b\}$. Note that the scales are constrained so that for every $i < b$, $w_{R,i} = w_{L,i+1}$. Furthermore, breaks are associated with break measures $\tau_1, \tau_2, \ldots, \tau_b$, where for all i, $\tau_i = \tau_{\bar{w}_i}(g_i)$. Also, for each break measure, we compute a significance $p_{\bar{w}_1}(\tau_1), \ldots, p_{\bar{w}_b}(\tau_b)$. We define a break at index i to be positive if $p_{\bar{w}_i}(\tau_i) \leq e$, for some constant significance threshold e and denote the number of positive breaks by b^+ ($\leq b$). In the next iteration, we merge clusters associated with the least significant break of index $\text{argmax}_i\, p_{\bar{w}_i}(\tau_i)$.

In the supplementary methods we show that an upper bound for the expected number of false positive breaks under the null for any level of clustering (any value of b) is directly related to e:

$$E[B^+] \leq be \tag{5.12}$$

B^+ is the random variable representing the number of called positive breaks (below significance level e) under the null (i.e. all called breaks are false positive breaks). Eq. 5.12 holds under the null, but for actual data it is necessary to perform the Benjamini-Hochberg procedure to select the appropriate number of breaks for a fixed FDR[25–27]. For any level of clustering b, this amounts to the following procedure:

- Sort the values $p_{\bar{w}_1}(\tau_1), ..., p_{\bar{w}_b}(\tau_b)$ in ascending order and re-label them $p_1, p_2, ..., p_b$.
- Compute q-values: $q_i = \frac{b}{i} p_i$.
- Set $k = \min\{i : q_i > \text{FDR}\}$ and call all breaks associated with $\{p_1, p_2, ..., p_{k-1}\}$ as positives, i.e. $b^+ = k - 1$.

To select the appropriate level of clustering, we start with $b = 1$, compute b^+ (which is zero or one). Then we continue to $b = 2$ and recompute b^+. This process is repeated until b^+ starts to decrease. We denote the final value b^+ with #breaks. Generally, there is no closed form solution for the functions $p_{\bar{w}}$, however we will use the expected Euler characteristic as a tight upper-bound for effective control on $E[B^+]$.

5.2.11. THE EULER CHARACTERISTIC OF A CHI-SQUARE RANDOM PROCESS.

For a fixed scale, Eq. 5.11 is equivalent to the FWER at a given level of τ. The Bonferroni corrected p-value at level τ is defined to be the expected number of measurements in the null process above the constant threshold τ, which is simply equal to the single test p-value for τ times the number of tests (G). If G is large and all tests are uncorrelated, the Bonferroni corrected p-value is a tight upper bound for the FWER (if FWER < 0.1).

Since $T^2_{\bar{w}}$ might be highly correlated between nearby g (see Eq. 5.10) the Bonferroni correction might be too conservative and result in significant power loss. For example, if we apply a fixed high threshold τ on the null chi-square process $T^2_{\bar{w}}$, it might happen that for nine out of ten realizations none of the measurements cross the threshold, and that for the tenth realization, ten measurements for *adjacent* g's cross the threshold. This means that the average number of single measurement tests across the ten realizations that exceed the threshold (the Bonferroni FWER estimate) will be much higher (1.0 in this case) than the probability of one or more measurements to exceed the threshold per realization (the true FWER is 0.1 in this example). For this reason it is more powerful to compute the expected number of *regions* above the threshold i.e. count adjacent values for g simultaneously crossing the threshold as one event. In the previous example, the ten adjacent measurements crossing the threshold would only be counted as one event.

To formalize this, consider the i^{th} realization of the null process $\tau^{(i)}_{\bar{w}}$. Assume also that g is defined on all real numbers $[1, G]$ (not just the integers $\{1, 2, ..., G\}$). One can linearly interpolate $\tau^{(i)}_{\bar{w}}$ for all real g and define a positive excursion set based on a positive threshold τ:

$$a^+(\tau^{(i)}_{\bar{w}}, \tau) \quad = \quad \{g \in [1, G] : \tau^{(i)}_{\bar{w}}(g) \geq \tau\} \tag{5.13}$$

We define the Euler characteristic $\chi(\tau^{(i)}_{\bar{w}}, \tau)$ to be the number of connected components in $a^+(\tau^{(i)}_{\bar{w}}, \tau)$. Since $\tau^{(i)}_{\bar{w}}$ is a realization of the chi-square process $T^2_{\bar{w}}$, we can also regard the Euler characteristic $\chi(\tau^{(i)}_{\bar{w}}, \tau)$ as a realization of the random variable $\chi(T^2_{\bar{w}}, \tau)$ that takes on integer values. For short we denote this random variable with $\chi_{\bar{w}}(\tau)$.

5.2.12. THE EXPECTED EULER CHARACTERISTIC.

There exists an exact analytical expression that relates a non-negative threshold τ to the expected Euler characteristic for the chi-square random process[26, 28]:

$$\begin{aligned} \tilde{\chi}_{\bar{w}}(\tau) \quad &= \quad E[\chi_{\bar{w}}(\tau)] \\ &= \quad e^{-\tau/2} + \frac{\tau^{1/2} e^{-\tau/2}}{\sqrt{2\pi}} R_{\bar{w}} \end{aligned} \tag{5.14}$$

with

$$R_{\tilde{w}} = \int_{[1,G]} \sqrt{\mathrm{Var}[\frac{d}{dg} T_{\tilde{w}}(g)]} dg \tag{5.15}$$

We showed in the Section 4.7.4[27] that $R_{\tilde{w}}$ can be estimated accurately as:

$$R_{\tilde{w}} = (G-1)\arccos(\rho_{\tilde{w}}), \tag{5.16}$$

where $\rho_{\tilde{w}}$ is as defined in Eq. 5.10.

5.2.13. THE EXPECTED EULER CHARACTERISTIC IS AN UPPER BOUND FOR THE MAXIMUM STATISTIC.

Next, we show that $\tilde{\chi}_{\tilde{w}}(\tau)$ is an upper-bound for $p_{\tilde{w}}(\tau)$. We know that for a realisation $\tau^{(i)}_{\tilde{w}}$ of $T^2_{\tilde{w}}$, a g for which $\tau^{(i)}_{\tilde{w}}(g) \geq \tau$ exist if and only if the Euler characteristic $\chi(\tau^{(i)}_{\tilde{w}}, \tau)$ is greater than zero. Therefore:

$$\begin{aligned} p_{\tilde{w}}(\tau) &= P[\chi(T^2_{\tilde{w}}, \tau) = 1] + P[\chi(T^2_{\tilde{w}}, \tau) = 2] + ... \\ &\leq P[\chi(T^2_{\tilde{w}}, \tau) = 1] + 2P[\chi(T^2_{\tilde{w}}, \tau) = 2] + ... \\ &= E[\chi(T^2_{\tilde{w}}, \tau)] \\ &= \tilde{\chi}_{\tilde{w}}(\tau) \end{aligned} \tag{5.17}$$

Throughout iterative clustering, we use the upper bound $\chi_w(\tau)$ instead of $p_w(\tau)$.

5.2.14. ESTIMATING BREAK LOCATIONS.

Although the clustering procedure is effective at selecting the appropriate number of breaks, it is not particularly good at finding their exact locations. This is because if we correctly reject the null hypothesis at location g, we are only guaranteed that there is a break in the region $\{g-w_L+1, ..., g+w_R\}$. We therefore only know the proximity of the breakpoints and the number of breakpoints. When the noise is Gaussian and the number of breaks is known, the best way to find the correct locations is to fit a piecewise constant function that minimizes the mean square error. This is the approach employed by algorithms such as *cghseg*[9]. With dynamic programming, this problem can be solved in $\mathcal{O}(G^2)$, where G is the number of probes in the dataset, which is computationally inefficient for realistic datasets. Nevertheless, when considering only one channel (c-channel), there are heuristics for which computation time is on average highly efficient[8]. For our algoritm, we speed up processing for two channels by restricting the search for breaks within the regions $\{g-w_L+1, ..., g+w_R\}$, where the breaks are likely to be located.

5.3. RESULTS

5.3.1. ESTIMATING THE NUMBER OF BREAKS FOR SIMULATED PROFILES.

For any level of the dendrogram, we know that be is an upper bound for the expected number of false positive breaks $E[B^+]$ in the null (Eq. 5.12). However we are not sure how tight this upper bound is. Therefore, in this section we simulate piecewise constant profiles with Gaussian noise and compare the observed error rates with the predicted error rates.

We simulated copy number profiles as described in the methods (Section 5.2.3). We performed *RUBICseg* on 10 generated profiles for the following SNR levels: $\{0.3, 0.5, 1.0, 2.0\}$. For each simulated profile we vary the FDR level and compute the number of detected breaks (#breaks) and the expected number of false positives E[#errors] = #breaks × FDR. The number of detected breaks

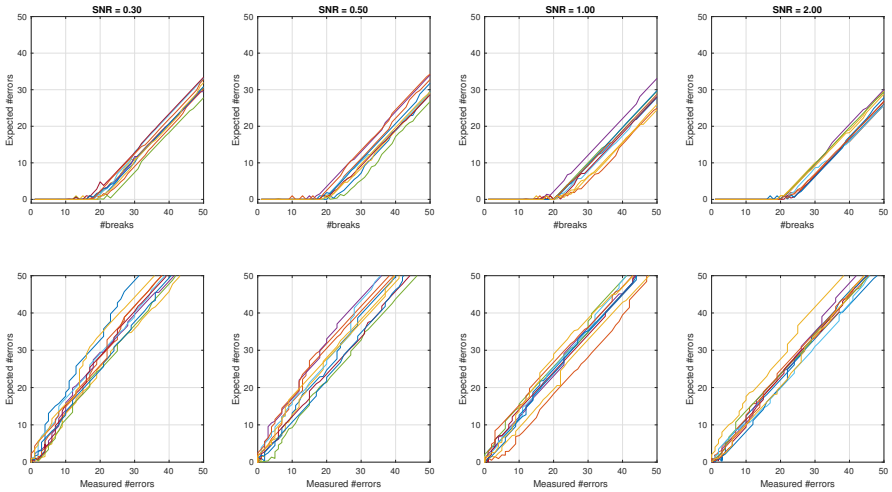

Figure 5.3: Illustrating the breakpoint error rates for simulated copy number profiles with Guassian noise. All results are based on copy number profiles that were generated with 20 breakpoints. Each panel shows 10 lines which correspond to the 10 simulated copy number profiles. Different columns show different signal to noise ratios as defined in the text. The panels in the top row illustrate the number of breakpoints called while changing the FDR on the x-axis. On the y-axis we plot the expected number of errors based on the expected Euler characteristics. In the bottom row of panels, we plot the measured error rates as defined in the text (x-axis) against the expected number of errors (y-axis).

('#breaks' on the x-axis) is plotted against the expected number of false positives ('Expected errors' on the y-axis) in the top panels of Fig. 5.3. Each panel contains 10 curves representing the results of the different profiles and results for different SNR levels are represented in different columns. Note that in all the top panels, the predicted number of errors is close to zero until approximately 20 breaks are chosen (corresponding to the simulated 20 breaks). After this, the predicted number of errors start to increase linearly with a slope of approximately one, as expected. For low SNRs, the predicted error rates start increasing for a lower number of breaks, indicative of the reduced power due to the increased noise variance. This experiment illustrates that *RUBICseg* is effective at selecting the appropriate number of breaks while controlling the FDR parameter.

The bottom row of panels in Fig. 5.3 depicts the predicted number of errors ('Expected errors' on the y-axis), based on the Euler characteristic, as a function of the measured number of errors ('Measured errors' on the x-axis). A break at location g is classified as a measured error if there are in fact no breaks within $\{g-w_L+1,...,g+w_R\}$, where w_L and w_R represent the widths of the left and right flanking regions, respectively. Although this is a fairly relaxed definition, it is directly compatible with the hypothesis testing performed. Due to this relaxed definition, the 'true' number of false breaks will be slightly higher than the measured number. For example, in the top right panel, we see that all predicted error rates are slightly below 30 when 50 breaks are called. Since there can only be 20 correct breaks, we are slightly underestimating the true number of false positives. Nevertheless, with the relaxed definition, the expected error rates are only slightly conservative. This shows that we are not sacrificing unnecessary power due to over-conservative FDR control.

5

5.3.2. *Jointseg* ANALYSIS.

Although the previous section shows that our error control is accurate when considering Gaussian noise, it is important to compare *RUBICseg* against state-of-the-art algorithms on real platforms, where the noise might not be strictly Gaussian or constant for different copy number states. To achieved this, we employ the *Jointseg* comparison pipeline[17] and recreated their experiments with *RUBICseg* included.

With *Jointseg* we simulated profiles with 20000 markers and 20 uniformly selected breakpoints with copy-number states selected as explained by Pierre-Jean *et. al.*[17]. We simulated realistic profiles based on two platforms: the Affymetrix SNP6 microarray and the Illumina HumanCNV370-Duov1 microarrays (called data set 1 and 2 respectively by Pierre-Jean *et. al.*[17]). *Jointseg* also allows one to select the noise level based on the biologically relevant parameter called the tumor fraction (fraction of tumor cells).

In this analysis we considered three algorithms that segment based on both the the c- and d-channel:

- *PSCBS*[12] with R package *PSCBS*. Labeled 'PSCBS' in Fig. 5.4 and 5.5
- *RBS*[13] (recursive binary segmentation) with R package *Jointseg*[17]. Labeled 'RBS+DP:log(c),d (Kmax=200)', with Kmax referring to the maximum number of breaks, set well above the required 20.
- *RUBICseg*. Labeled 'RUBIC:log(c),d', and

four algorithms based on the c- or d-channels alone:

- *CBS*[11] with R package *DNAcopy*. Labeled 'CBS:log(c)'(c-channel) and 'CBS:d'(d-channel)
- *DP*[8] (dynamic programming) with R package *cghseg*. Labeled 'DP:log(c) (Kmax=200)'and 'DP:d (Kmax=200)'.
- *GFLars*[14] with R package *Jointseg*[17]. Labeled 'GFLars+DP:log(c) (Kmax=200)'and 'GFLars+DP:d (Kmax=200)'

- *RBS*[13] with R package *Jointseg*[17]. Labeled 'RBS+DP:log(c) (Kmax=200)'and 'RBS+DP:log(c),d (Kmax=200)'.

Parameter choices for all the algorithms (except for *RUBICseg*) were optimised as in Pierre-Jean *et al.*[17]. A full comparison on the performance of existing algorithms are also discussed by Pierre-Jean *et al.*[17]. For *RUBICseg*, we set the FDR parameter at 5% for all experiments.

For any given algorithm, *Jointseg* defines true positive breakpoints as those calls that are within a tolerance threshold (number of probes) from a true breakpoint. To be exact, each true break is associated with maximally one called break within the tolerance threshold, whereas the rest of the called breaks are considered false positives. *Jointseg* computes a specific type of partial AUC (pAUC) as detailed in the methods.

Fig. 5.4 shows the results of the comparison of *RUBICseg* to the algorithms listed above while varying the tolerance threshold. As expected, algorithms generally perform better for higher tolerance values. The ranking of tested algorithms also does change significantly for different tolerance levels. Note that segmentation algorithms do not generally perform equally well on all platforms [29]. In our experiments, the SNP6 platform is more noisy than the Illumina platform and all the segmentation algorithms we tested performed significantly better on the Illumina platform. One striking (and unintuitive) result is that performance curves saturate at a much lower value for the Illumina platform at 100% tumor fraction (bottom right panel) than for 79% tumor fraction. This occurs because the noise variance in the d-channel estimates are very high within regions with homozygous deletions, resulting in additional false positives for most algorithms. This is especially true for algorithms that segment on the d-channel alone.

Since the order of the algorithms remains the same as a function of tolerance level, we show box plots at a fixed tolerance of five probes for both platforms and all algorithms in Fig. 5.5. Tumor fraction has a clear effect on segmentation performance where high tumor fractions are associated with improved performance. For low tumor percentages, while *RUBICseg* is still amongst the top performing approaches, it performs relatively poorly. This occurs because *RUBICseg* calls too few breaks for noisy profiles (strict FDR control). This directly effects pAUC scores that tend to favor algorithms that overcall breaks[17]. For real data, tumor fractions below 70% are typically regarded as low quality. As the results indicate, *RUBICseg* often outperforms existing algorithms for tumor fractions above 70%.

It is also important to notice that segmentation on the d-channel only performs significantly worse on the SNP6 platform simply because only one in six probes are informative. In contrast, one in three Illumina probes are informative, which indeed improves performance on the d-channel, albeit still worse than the c-channel only. It is also interesting to note that segmentation on the combined c- and d-channel usually only provides marginal improvement when compared to the c-channel only.

RUBICseg and *RBS* (combined c- and d-channel) are the best performing algorithms for high tumor percentages ($\geq 70\%$) on both platforms, with *RUBICseg* outperforming *RBS* in three of the four cases tested.

5.4. DISCUSSION

We developed a new segmentation algorithm called *RUBICseg* that accurately recovers breakpoints in DNA copy number profiles derived from both Affymetrix and Illumina SNP6 platforms. There are a number of reasons why we expect *RUBICseg* to have high statistical power for detecting breaks. First, it is one of the few algorithms that jointly segments on the c- and d-channel and that models the noise covariance directly. Second, *RUBICseg* does not ignore any of the informative markers. For example, some algorithms cannot handle missing values and therefore down-sample the c-channel for compatibility. Third, *RUBICseg* fine tunes the break locations by minimising the

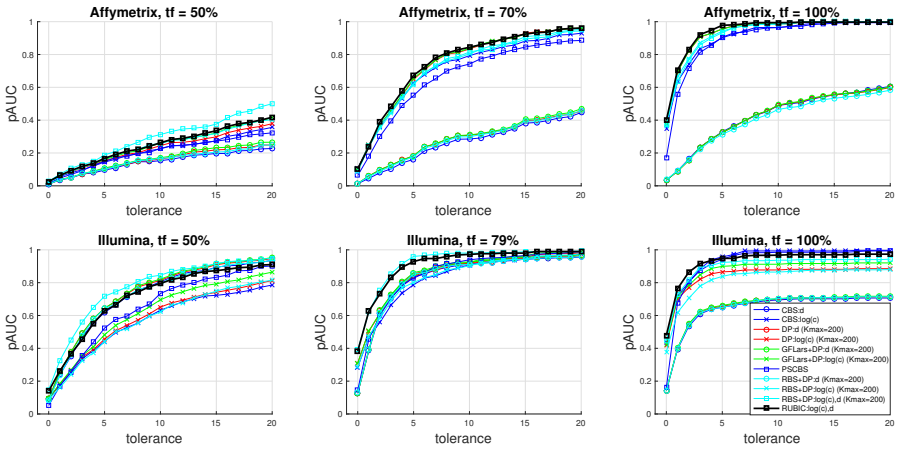

Figure 5.4: Comparison of different segmentation algorithms on *Jointseg* simulations based on data derived from the Affymetrix (top row) and the Illumina (bottom row) platforms. The columns show the results for different levels of the tumor fraction. In each subplot, the x-axis represents a tolerance threshold for calling true positive breaks, i.e. the maximum allowed distance between a true break and a called break. The y-axis represents the partial area under the curve (pAUC) in the ROC. Each experiment was repeated 20 times and the average is reported. Each algorithm is shown in a different color. c-channel, d-channel and combined c- and d-channels are marked with crosses, circles and squares respectively.

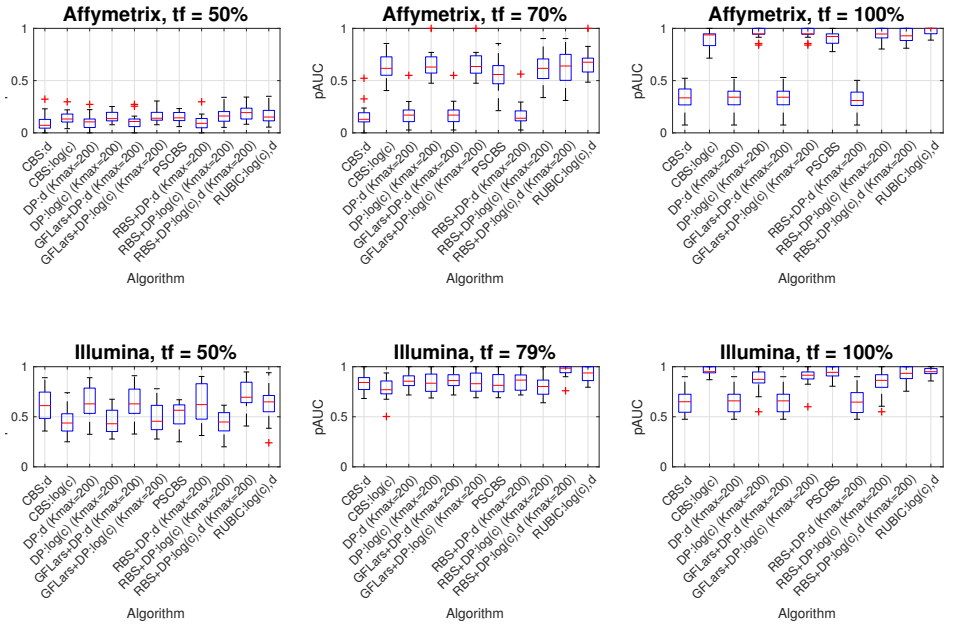

Figure 5.5: Comparison of different segmentation algorithms at a fixed tolerance of five probes. Top (bottom) panels represent *Jointseg* experiments performed on the Affymetrix (Illumina) platform. Columns represent different tumor percentages. In each subplot, the x-axis represents different segmentation algorithms. The y-axis represents the partial area under the curve (pAUC) in the ROC. Each box plot is constructed from 20 repeats.

mean square error (for a fixed number of breaks), which is an approach also employed by one of the top performing methodologies (for example *cghseg*). Finally, *RUBICseg* estimates chi-square values at different and adaptive scales. The larger the widths of segments adjacent to a break, the larger the statistical power (the chi-square test works better for larger sample sizes) and less severe multiple testing since the expected Euler characteristic accounts for the high correlation between noise measurements. For this reason, one should always expect *RUBICseg* to perform better than a methodology that, for instance, uses a fixed wavelet kernel to detect breaks.

This methodology outperforms all existing algorithms, albeit not in all situations, for tumor fractions of 70% or higher for both platforms tested. A tumor fraction above 70% is a reasonable requirement in practice since lower quality profiles drastically decrease the performance of all segmentation algorithms. One of the most desirable properties of *RUBICseg* is that it has only one intuitive parameter: the expected proportion of false positive breaks. While the other top performing algorithm, *RBS*, also only requires a single input paramater (the number of breaks) we are of the opinion that it has more utility to be able to specify the FDR as this immediately provides an objective and useful estimate of the accuracy of the results. Moreover, we showed that in simulations with Gaussian noise, RUBIC does not suffer from unnecessary statistical power loss due to over-conservative error control.

Although *RUBICseg* performs very well for both the Affymetrix and Illumina platforms considered, we highly recommend that users perform extensive tests with the joinseg package for other platforms and/or lower tumor fractions.

5.5. ACKNOWLEDGEMENT

We would like to thank Guillem Rigaill for many discussions.

5.6. AUTHOR CONTRIBUTIONS

L.F.A.W., E.v.D. and M.J.T.R. designed the study; E.v.D. developed the theoretical framework, conducted data analyses and performed the experiments; E.v.D., L.F.A.W. and M.J.T.R. wrote the manuscript; all authors reviewed the manuscript; L.F.A.W. and M.J.T.R. provided supervision.

5.7. CONFLICT OF INTEREST

None of the authors declares any conflicts of interest.

5.8. SUPPLEMENTARY METHODS

5.8.1. ITERATIVE CLUSTERING ON THE GLOBAL NULL.

We show that we can find an upper bound on the expected number of false positive breakpoints at any given stage of iterative clustering in the global null using the statistics:

$$M_{\bar{w}} = \max_{g\in\{1,...,G\}} T^2_{\bar{w}}(g) \tag{5.18}$$

For each scale $\bar{w}$, $M_{\bar{w}}$ is a random variable tracking the maximum value of the random process $T^2_{\bar{w}}$ across all possible locations g.

Suppose we performed iterative clustering until there are b breakpoints (i.e. $b+1$ clusters) in the global null. Label the break locations $G_1, G_2, ..., G_b$ and their respective scales with $\bar{W}_1, \bar{W}_2, ..., \bar{W}_b$. Note that the break locations and scales are random variables (which is why we write them in capital letters) since each realisation of the null will result in different break locations and scales. Furthermore, breaks are associated with break measures $T^2_1, T^2_2, ..., T^2_b$. These are random variables describing realisations of the break measures when applied to the global null. For the next iteration (i.e. jumping from b to $b-1$ breaks), we choose index $I = \arg\max_i p_{\bar{W}_i}(T^2_i)$ and remove the break at location G_I. Then we replace the values T^2_{I-1} and T^2_{I+1} with new values denoted by T'^2_{I-1} and T'^2_{I+1} at scales $\bar{W}'_{I-1} = \langle W_{L,I-1}, W_{R,I-1} + W_{L,I+1}\rangle$ and $\bar{W}'_{I+1} = \langle W_{R,I-1} + W_{L,I+1}, W_{R,I+1}\rangle$ respectively.

In general, the distribution of T^2_{I+1} is not a chi-square distribution, due to the constraint $p_{\bar{W}_{I+1}}(T^2_{I+1}) \le p_{\bar{W}_I}(T^2_I)$. Furthermore, the newly selected value T'^2_{I+1} is positively correlated to T^2_{I+1} and uncorrelated to T^2_I. As a consequence, T'^2_{I+1} will also not be chi-square and will in general result in higher values then that of a chi-square distribution. Similar arguments apply to T'^2_{I-1}. To summarize, the larger the number of iterations we perform (i.e. the smaller the number b), the closer the values T^2_i will get to $M_{\bar{W}_i} = \max_{g\in G} T^2_{\bar{W}_i}(g)$ for all $i \in \{1, ..., b\}$.

Another important point to note is that the scale W'_{I+1} is fully determined by $\bar{W}_{I-1}$ and $\bar{W}_{I+1}$. As a consequence we do not expect the distribution of the maximum statistic $M_{\bar{W}'_{I+1}}$ to be effected by this choice (this is also easy to verify experimentally). Therefore:

$$P[M_{\bar{W}'_{I+1}} \ge \zeta_{\bar{W}'_{I+1}} | \bar{W}'_{I+1} = \bar{w}] = P[M_{\bar{w}} \ge \zeta_{\bar{w}}] \tag{5.19}$$

Since $T'^2_{I+1} \le M_{\bar{W}'_{I+1}}$,

$$\begin{aligned} P[T'^2_{I+1} \ge \zeta_{\bar{W}'_{I+1}} | \bar{W}'_{I+1} = \bar{w}] &\le P[M_{\bar{W}'_{I+1}} \ge \zeta_{\bar{W}'_{I+1}} | \bar{W}'_{I+1} = \bar{w}] \\ &= P[M_{\bar{w}} \ge \zeta_{\bar{w}}] \end{aligned} \tag{5.20}$$

A similar argument applies for T'^2_{I-1}:

$$P[T'^2_{I-1} \ge \zeta_{\bar{W}'_{I-1}} | \bar{W}'_{I-1} = \bar{w}] \le P[M_{\bar{w}} \ge \zeta_{\bar{w}}] \tag{5.21}$$

We now show by induction that for any b (level of clustering):

$$\forall i \in \{1, ..., b\} \qquad P[T^2_i \ge \zeta_{\bar{W}_i} | \bar{W}_i = \bar{w}_i] \le P[M_{\bar{w}_i} \ge \zeta_{\bar{w}_i}], \tag{5.22}$$

This statement is clearly true for $b = G-1$ since all $\bar{W}_i = \langle 1, 1\rangle$, i.e. we are restricted to only one possible scale. Now let us assume Eq.5.22 holds for iterative level b. We show that it also holds for iterative level $b-1$. In this case, we merge a break at index I and consider statistics $T^2_1, ..., T^2_2, ..., T'^2_{I-1}, T'^2_{I+1}, ..., T^2_b$. Clearly Eq. 5.22 remain true for variables $T^2_1, T^2_2, ..., T^2_{I-2}, T^2_{I+2}, ..., T^2_b$,

since they remain unchanged from iterative level b. By Eq. 5.20 and 5.21, the statement also holds true for T'^2_{I-1} and T'^2_{I+1}, completing the proof.

We can now readily see that for any b (level of clustering):

$$E[B^+|\langle \bar{W}_1,...,\bar{W}_b\rangle = \bar{w}_1,...,\bar{w}_b\rangle] \quad \leq \quad \sum_{i\in 1,...,b} P[M_{\bar{w}_i} \geq \zeta_{\bar{w}_i}], \tag{5.23}$$

where B^+ is number of indices i for which $T^2_i \geq \zeta_{\bar{w}_i}$ (i.e. the number of called breaks when clustering). Note that this summation is an upper bound on the expected number of false positive breaks, since we are dealing with the global null (i.e. all called breaks are false).

In summary, our methodology for controlling the expected number of false positive break points in iterative clustering is equivalent modelling the distribution of the maximum statistics of the chi-square random processes $T^2_{\bar{w}}$ for each scale $\bar{w}$ separately. The natural choices for the functions $p_{\bar{w}}$ and thresholds are then represented by:

$$\begin{aligned} p_{\bar{w}}(\tau) &= P[M_{\bar{w}} \geq \tau] \\ \zeta_{\bar{w}} &= f^{-1}(e), \end{aligned} \tag{5.24}$$

for some constant significance threshold e. If we use these functional forms for $p_{\bar{w}}$ and $\zeta_{\bar{w}}$, B^+ is the number of break-points for which $p_{\bar{W}_i}(T^2_i) \leq e$ and Eq. 5.23 simplifies to:

$$E[B^+|\langle \bar{W}_1,...,\bar{W}_b\rangle = \bar{w}_1,...,\bar{w}_b\rangle] \quad \leq \quad be \tag{5.25}$$

Since the right side of the equation is independent on the scales we condition on, we can simplify the equation to:

$$E[B^+] \quad \leq \quad be \tag{5.26}$$

Therefore we can directly control the expected number of false positive breaks with the significance threshold e.

5.8.2. DEALING WITH MISSING DATA.

In the interest of full generality, let us assume that both the c- and d-channel have missing (uninformative) measurements. Suppose that $M = (m_1, m_2,...)$ and $N = (n_1, n_2,...)$ are the index sequences of informative probe measurements at their respective genomic location on the c- and d-channels, respectively. Furthermore, we need indices denoting where both measurements are available $O = M \cap N = (o_1, o_2,...)$. We need to compute new estimates for $\tau_{\bar{w}}(g)$ values in the data. In order to do this, we recompute $\hat{\Sigma}$ (Eq. 5.2), $c_{\bar{w}}$, $d_{\bar{w}}$ (Eq. 5.3) and $\hat{\Sigma}_{\bar{w}}$ (Eq. 5.5). Eq. 5.6 and 5.7 remain the same.

ESTIMATING THE NOISE COVARIANCE

To compute the covariance of the noise $\hat{\Sigma}$, we again use the lag-one difference profile in the c- and d-channel, except that we ignore probes that have missing data-points in the respective channels. To compute the covariance between the two channels, we only consider the probes for which both

measurements are available. The new covariance estimate is as follows:

$$\begin{aligned}
\hat{\Sigma} &= \begin{bmatrix} \hat{\sigma}_c^2 & \hat{\rho} \\ \hat{\rho} & \hat{\sigma}_d^2 \end{bmatrix}, \text{where} \\
\hat{\sigma}_c^2 &= \frac{1}{2(|M|-1)} \sum_{i=1}^{|M|-1} (c_{m_{i+1}} - c_{m_i})^2 \\
\hat{\sigma}_d^2 &= \frac{1}{2(|N|-1)} \sum_{i=1}^{|N|-1} (d_{n_{i+1}} - d_{n_i})^2 \\
\hat{\rho} &= \frac{1}{2(|O|-1)} \times \\
&\quad \sum_{i=1}^{|O|-1} (c_{o_{i+1}} - c_{o_i})(d_{o_{i+1}} - d_{o_i})
\end{aligned} \tag{5.27}$$

DEFINING A BREAK MEASURE BETWEEN ADJACENT SEGMENTS

Similary, when computing $c_{\bar{w}}(g)$ and $c_{\bar{w}}(g)$, we compute averages on the available probes:

$$\begin{aligned}
c_{\bar{w}}(g) &= \frac{1}{|R_{\bar{w}}(g) \cap M|} \sum_{i \in R_{\bar{w}}(g) \cap M} c_i - \\
&\quad \frac{1}{|L_{\bar{w}}(g) \cap M|} \sum_{i \in L_{\bar{w}}(g) \cap M} c_i \\
d_{\bar{w}}(g) &= \frac{1}{|R_{\bar{w}}(g) \cap N|} \sum_{i \in R_{\bar{w}}(g) \cap N} d_i - \\
&\quad \frac{1}{|L_{\bar{w}}(g) \cap N|} \sum_{i \in L_{\bar{w}}(g) \cap N} d_i,
\end{aligned} \tag{5.28}$$

The new estimate for $\hat{\Sigma}_{\bar{w}}$ is more complex than the one proposed in Eq. 5.5, since it now depends on g (the number of informative measures in the left and right segments can vary with g):

$$\begin{aligned}
\hat{\Sigma}_{\bar{w}}(g) &= \begin{bmatrix} \hat{\sigma}^2_{c,\bar{w}}(g) & \hat{\rho}_{\bar{w}} \\ \hat{\rho}_{\bar{w}}(g) & \hat{\sigma}^2_{d,\bar{w}}(g) \end{bmatrix}, \text{where} \\
\hat{\sigma}^2_{c,\bar{w}}(g) &= \left(\frac{1}{|L_{\bar{w}}(g) \cap M|} + \frac{1}{|R_{\bar{w}}(g) \cap M|}\right)\hat{\sigma}_c^2 \\
\hat{\sigma}^2_{d,\bar{w}}(g) &= \left(\frac{1}{|L_{\bar{w}}(g) \cap N|} + \frac{1}{|R_{\bar{w}}(g) \cap N|}\right)\hat{\sigma}_d^2 \\
\hat{\rho}_{\bar{w}}(g) &= \left(\frac{|L_{\bar{w}}(g) \cap O|}{|L_{\bar{w}}(g) \cap M||L_{\bar{w}}(g) \cap N|} + \right. \\
&\quad \left. \frac{|R_{\bar{w}}(g) \cap O|}{|R_{\bar{w}}(g) \cap M||R_{\bar{w}}(g) \cap N|}\right)\hat{\rho}
\end{aligned} \tag{5.29}$$

$\hat{\rho}_{\bar{w}}(g)$ should not be confused with $\rho_{\bar{w}}$ without a hat (Eq. 5.10). $\hat{\rho}_{\bar{w}}(g)$ describes the correlation between $c_{\bar{w}}(g)$ and $d_{\bar{w}}(g)$ at a fixed g, whereas $\rho_{\bar{w}}$ describe the correlation between adjacent measurements.

Eq. 5.6 and 5.7 remain the same, expect that the eigen matrices V and D now depend on g (since $\hat{\Sigma}_{\bar{w}}$ depends on g). Therefore we have everything we need to compute our break measure $\tau_{\bar{w}}(g)$.

DEPENDENCY BETWEEN ADJACENT BREAK MEASURES AND THE EXPECTED EULER CHARACTERISTIC

When dealing with missing data, $\rho_{\bar{w}}$ will depend on g and Eq. 5.16 should be replaced with (see Eq. 4.12):

$$R_{\bar{w}} = \sum_{g=1}^{G-1} \arccos(\rho_{\bar{w}}(g)) \tag{5.30}$$

However, we note that the number of regions crossing a fixed threshold for a realization of the null process $T^2_{\bar{w}}$ (with missing datapoints) cannot, on average, be larger then the number of regions detected at full resolution. In fact, for most kernel widths $\bar{w}$ these expectations are very similar. Therefore it is of limited practical use to compute $\rho_{\bar{w}}$ at every value of g and summing (Eq. 5.30), since this sum will result in values similar to Eq. 5.16.

DEGENERATE CASES

There needs to be at least one informative measure in both the left and right segment when computing $d_{\bar{w}}(g)$, i.e. neither $L_{\bar{w}}(g) \cap N$ nor $R_{\bar{w}}(g) \cap N$ is allowed to be empty. Otherwise we are forced to work with the c-channel alone. In this case, the matrix $\hat{\Sigma}_{\bar{w}}(g)$ will be degenerate, one of the eigen values in D will be zero, and $t_{\bar{w},1}(g)$ (Eq. 5.6) will simply be a scaled version of $c_{\bar{w}}(g)$ (with unit variance in the null). We then compute the expected Euler characteristic for the Gaussian process $T_{\bar{w}}$ instead of the chi-square field $T^2_{\bar{w}}$ (see Eq. 3.11 and Eq. 4.12):

$$\tilde{\chi}_{\bar{w}}(t) = \operatorname{erfc}(\frac{t}{\sqrt{2}}) + \frac{e^{-t^2/2}}{\pi} R_{\bar{w}} \tag{5.31}$$

Note that this equation differs by a factor of two compared to Eq. 3.11 since we are performing a two-tailed test, i.e. we are counting regions above and below the threshold t and $-t$, respectively.

REFERENCES

[1] D. Hanahan and R. A. Weinberg, *Hallmarks of cancer: the next generation,* cell **144**, 646 (2011).

[2] C. Price, *Fluorescence in situ hybridization,* Blood reviews **7**, 127 (1993).

[3] D. Pinkel and D. G. Albertson, *Comparative genomic hybridization,* Annu. Rev. Genomics Hum. Genet. **6**, 331 (2005).

[4] *Affymetrix, inc,* http://www.affymetrix.com/index.affx.

[5] T. I. Zack, S. E. Schumacher, S. L. Carter, A. D. Cherniack, G. Saksena, B. Tabak, M. S. Lawrence, C.-Z. Zhang, J. Wala, C. H. Mermel, *et al.*, *Pan-cancer patterns of somatic copy number alteration,* Nature genetics **45**, 1134 (2013).

[6] P. Neuvial, H. Bengtsson, and T. P. Speed, *Statistical analysis of single nucleotide polymorphism microarrays in cancer studies,* in *Handbook of Statistical Bioinformatics* (Springer, 2011) pp. 225–255.

[7] H. Bengtsson, P. Neuvial, and T. P. Speed, *Tumorboost: Normalization of allele-specific tumor copy numbers from a single pair of tumor-normal genotyping microarrays,* BMC bioinformatics **11**, 245 (2010).

[8] G. Rigaill, *Pruned dynamic programming for optimal multiple change-point detection,* arXiv preprint arXiv:1004.0887 (2010).

[9] F. Picard, S. Robin, M. Lavielle, C. Vaisse, and J.-J. Daudin, *A statistical approach for array cgh data analysis,* BMC bioinformatics **6**, 27 (2005).

[10] D. Mosén-Ansorena and A. M. Aransay, *Bivariate segmentation of snp-array data for allele-specific copy number analysis in tumour samples,* BMC bioinformatics **14**, 84 (2013).

[11] E. Venkatraman and A. B. Olshen, *A faster circular binary segmentation algorithm for the analysis of array cgh data,* Bioinformatics **23**, 657 (2007).

[12] A. B. Olshen, H. Bengtsson, P. Neuvial, P. T. Spellman, R. A. Olshen, and V. E. Seshan, *Parent-specific copy number in paired tumor–normal studies using circular binary segmentation,* Bioinformatics **27**, 2038 (2011).

[13] S. Gey and E. Lebarbier, *Using cart to detect multiple change points in the mean for large sample,* (2008).

[14] C. Levy-leduc and Z. Harchaoui, *Catching change-points with lasso,* in *Advances in Neural Information Processing Systems* (2008) pp. 617–624.

[15] K. Bleakley and J.-P. Vert, *The group fused lasso for multiple change-point detection,* arXiv preprint arXiv:1106.4199 (2011).

[16] H. Chen, H. Xing, and N. R. Zhang, *Estimation of parent specific dna copy number in tumors using high-density genotyping arrays,* PloS computational biology **7**, e1001060 (2011).

[17] M. Pierre-Jean, G. Rigaill, and P. Neuvial, *Performance evaluation of dna copy number segmentation methods,* Briefings in bioinformatics , bbu026 (2014).

[18] J. Fridlyand, A. M. Snijders, D. Pinkel, D. G. Albertson, and A. N. Jain, *Hidden markov models approach to the analysis of array cgh data,* Journal of multivariate analysis **90**, 132 (2004).

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

[20] C. D. Greenman, G. Bignell, A. Butler, S. Edkins, J. Hinton, D. Beare, S. Swamy, T. Santarius, L. Chen, S. Widaa, *et al.*, *Picnic: an algorithm to predict absolute allelic copy number variation with microarray cancer data,* Biostatistics **11**, 164 (2009).

[21] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, *Sparsity and smoothness via the fused lasso,* Journal of the Royal Statistical Society: Series B (Statistical Methodology) **67**, 91 (2005).

[22] A. B. Olshen, E. Venkatraman, R. Lucito, and M. Wigler, *Circular binary segmentation for the analysis of array-based dna copy number data,* Biostatistics **5**, 557 (2004).

[23] R. Edgar, M. Domrachev, and A. E. Lash, *Gene expression omnibus: Ncbi gene expression and hybridization array data repository,* Nucleic acids research **30**, 207 (2002).

[24] J. Von Neumann, R. Kent, H. Bellinson, and B. t. Hart, *The mean square successive difference,* The Annals of Mathematical Statistics **12**, 153 (1941).

[25] Y. Benjamini and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing,* J R Stat Soc B **57**, 289 (1995).

[26] E. van Dyk, M. J. Reinders, and L. F. Wessels, *A scale-space method for detecting recurrent dna copy number changes with analytical false discovery rate control,* Nucleic acids research **41**, e100 (2013).

[27] E. Van Dyk, M. Hoogstraat, J. Ten Hoeve, M. J. Reinders, and L. F. Wessels, *Rubic identifies driver genes by detecting recurrent dna copy number breaks,* Nature communications **7** (2016).

[28] K. J. Worsley, S. Marrett, P. Neelin, A. C. Vandal, K. J. Friston, A. C. Evans, *et al.*, *A unified statistical approach for determining significant signals in images of cerebral activation,* Human brain mapping **4**, 58 (1996).

[29] C. Curtis, A. G. Lynch, M. J. Dunning, I. Spiteri, J. C. Marioni, J. Hadfield, S.-F. Chin, J. D. Brenton, S. Tavaré, and C. Caldas, *The pitfalls of platform comparison: Dna copy number array technologies assessed,* BMC genomics **10**, 588 (2009).

6

DISCUSSION

6.1. LINKING ONCOGENES WITH RECURRENT DNA COPY NUMBER GAINS

The overall theme of this thesis is to discover driver oncogenes and tumor suppressors using DNA copy number data. It is intuitively clear that recurrent copy number gains across tumor samples are associated with oncogenes, but it is important to carefully consider why this happens. When an oncogene is over expressed in a cell due to an amplification, it experiences a selective advantage above other cells in a tumor. In contrast, when only non-oncogenic genes are overexpressed due to an amplification, no such advantage is expected. Therefore, when we consider *multiple* tumor samples, we expect to see an enrichment for amplifications that overlap with oncogenes. However, if we consider only a *single* (highly aberrated) tumor sample, we do not directly know which amplifications actually activate oncogenes. We define amplifications that activate oncogenes as *driver amplifications* and the rest as *passengers*. We emphasise that it is not sufficient for an amplification to overlap with an oncogene to be considered a driver amplification. To be a driver, it needs to activate the oncogene and therefore apply selective pressure. When considering multiple tumor samples, amplifications might show overlap in a fixed common region and if this occurs at a significant frequency, there is a good chance that there will be an oncogene in the common region. Nevertheless, this does not imply that every amplification in every tumor that overlaps with the common region is a driver amplification (Fig. 4.1a). A statistical test that states that a region is significantly frequently amplified only gives us confidence that at least one sample contains a driver amplification and that this amplification overlaps with an oncogene. This realisation is very important since recurrent aberrations vary substantially in width across tumors. Hypothetically it might be that there is only one driver amplification in a single sample that happens to span a whole chromosome arm. In this case it might happen that some overlapping focal region appear to recur significantly (due to the single broad driver event that decreases the p-value) even though the oncogene could be anywhere on the arm. To narrow down the location of the oncogene one might remove such a broad amplification and see if the desired region remains significantly recurrent. GISTIC2, for example, first removes all arm-wide events before performing a recurrence analysis since these aberrations are uninformative with regard to the location of the putative oncogene regardless of whether they are drivers. One strategy for pinpointing the oncogene is to iteratively remove the broadest remaining amplification until recurrence cannot be proved anymore and then calling the union of all remaining amplifications to be sure one does

not miss the oncogene. There are two complications with this scheme. First, amplifications across tumors might not overlap perfectly, i.e. the intersection between two broad aberrations might be very small. If removing both leads to a non significant recurrence of the remainder, one has to accept that either of these two broad events might be the only driver and therefore call a very broad region. The second problem is that in order for this argument to make sense, one needs to reconstruct the occurrence history of somatic copy number events from the segmented profiles. For example, a chromosome with three adjacent amplified segments could be explained in different ways. It might be explained by one broad aberration and a focal amplification inside, or it could be explained by two partially overlapping broad aberrations. An example of such an algorithm used by GISTIC2 is Ziggurat Deconstruction (ZD). As one can not unambiguously deconstruct the history of such events, the strategy of recursively 'peeling' away events might not be reliable.

Most algorithms perform many different steps in order to resolve the above mentioned issues. GISTIC2, for example, integrates roughly four steps in a pipeline.

- Use Ziggurat deconstruction which estimates somatic copy number aberrations.
- Removes copy number aberrations stretching across complete chromosome arms.
- Perform peel-off to iteratively identify independent significant recurrent regions (Fig 4.1b-f).
- Use RegBounder to broaden significant regions to increase the chance of oncogene overlap(Fig. 4.1g).

For ADMIRE and RUBIC, we had the mindset that it is probably better to try and resolve the issues mentioned in one statistical framework instead of performing multiple serialized algorithmic steps (each with associated power losses). For ADMIRE we identified more and more focal recurrent aberrations relative to a hierarchy of recursively updated background models. This recursive procedure eliminates the need for the Ziggurat deconstruction, peel-off and the RegBounder algorithm, but nevertheless requires a large number of parameters (for each background) to be estimated which eventually also leads to a loss of statistical power. By looking for recurrent unidirectional breakpoints (RUBIC), we believe we managed to resolved all the above mentioned difficulties in a single well-defined statistical framework with a single well-defined null-model (and therefore fewer parameters to estimate). We believe that this is one of the main reasons why we enjoy improved statistical power over existing algorithms in not just calling more recurrent events, but also by calling them more focally and therefore pinpointing the desired oncogenes.

6.2. LINKING TUMOR SUPPRESSORS WITH RECURRENT COPY NUMBER LOSSES

We described how oncogenes can be located using copy number amplifications and mentioned that tumor suppressors are found in a similar way. Indeed both ADMIRE and RUBIC treat tumor suppressors in a symmetric way, with few exceptions. Instead of looking for recurrent amplifications (or breaks associated with amplifications) we look for recurrent deletions (or breaks associated with deletions).

The mechanism by which tumor suppressors are inactivated is, however, quite different. With oncogenes, driver amplifications likely overlap with them and as we saw with RUBIC, called regions are likely to (but not always) fully contain such genes. With tumor suppressors, however, only a small subsequence need to be deleted to inactivate a gene. Therefore, in contrast to oncogenes, we often find recurrent regions with RUBIC that only partially overlap tumor suppressors. Sometimes, RUBIC also calls multiple recurrent regions inside a single tumor suppressor. This probably happens because the overall frequency of breaks inside the gene is much higher than

the background break frequency, but nevertheless somewhat random within the gene (to truncate and therefore inactivate it). From a statistical point of view, this is not a severe problem. RUBIC treats multiple called regions (per gene) as a single event and it therefore does not affect FDR control if the goal is to discover tumor suppressors. Focally called regions within a gene are useful for identifying it as a tumor suppressor, but one cannot conclude that there is any special meaning with respect to the region boundaries called by RUBIC. For example, one cannot conclude that a specific domain or exon of the gene is key to tumorigenesis. To do so, one needs to model the copy number break frequency (across samples) within the gene itself and not the overall background frequency.

Unlike amplifications, correctly called deleted regions need not even overlap a tumor suppressor gene. For example, a deleted promoter or enhancer region might also deactivate a tumor suppressor at a different location. At this moment ADMIRE and RUBIC only call overlapping genes as possible tumor suppressors and this list should be extended in future versions.

6.3. THE VALUE OF LARGE DATASETS

Probably only a small proportion of driver genes can be unambiguously identified when using copy number data alone. One obvious reason is that DNA copy number aberrations are only one of many mechanisms by which a driver can be effected. Other possible mechanisms include point mutations and epigenetic factors. So let us rephrase the question and ask: "How many of the drivers that are activated/deactivated by copy number alterations can we pinpoint?" For any given cancer dataset, it is not possible to give an absolute answer to this question since we do not know how many drivers there are in the first place. We can only provide a relative comparison with existing algorithms and perhaps observe convergence with increased dataset sizes.

For very large datasets one might argue that it doesn't matter if we are able to detect all of the (copy number activated) driver genes. The argument is that if an oncogene gets amplified in less than one percent of the tumor samples, how important can it really be? The big problem with this argument is that a very large proportion of genes are actually amplified and/or deleted in a large proportion of samples. A chromosome arm might be amplified in 50% of all samples and therefore activate an important oncogene in half of the tumors, but it is the infrequent focal amplifications that we rely on to pinpoint the gene. As a consequence increasing dataset sizes doesn't only allow one to detect (arguably unimportant) rarely activated oncogenes, but also greatly helps to pinpoint the appropriate and frequently amplified ones. This illustrates how important it is to robustly call regions at their appropriate widths (not too focal or broad) for ever increasing sample sizes. We showed for example how a state-of-the-art algorithm such as GISTIC2 can completely miss EGFR (by calling too focal) in a large Glioblastoma dataset, even though it is by far the most frequently amplified oncogene. We conclude that with ever increasing dataset sizes, robust algorithms such as RUBIC will become more and more important to fine-map the list of important oncogenes and tumor suppressors.

Currently, whole exome sequencing (WES) datasets are growing rapidly from which copy number profiles can be derived. In our experience, the resolution of copy number profiles obtained from WES data currently is frequently much lower that that obtained from SNP6.0 platforms. This is because many (not all) algorithms such as CopywriteR[1] and CNVkit[2] bin reads into 20 kbp (or more) regions in order to acquire high enough coverage for accurate copy number estimates. This means that the density of possible locations at which breaks are called can be low. Nevertheless, for very large WES datasets, we expect RUBIC to perform well compared to SNP6.0 datasets of moderate size, since the major reason for calling wide regions (which leads to driver gene ambiguity) is limited sample size (and therefore rare focal events required for pinpointing the genes) and not the resolution of the copy number profiles. Also, for large datasets, the quality of single

sample profiles become less important, since the aggregate profile 'averages' the noise.

Increased sample sizes are not the only way to improve driver detection from copy number data. Algorithms such as Helios[3] further narrow down the list of potential driver genes by integrating other data sources such as point mutations, high-throughput genetic screens and expression data. The ISAR algorithm[3] which also calls recurrent regions, can easily be replaced with ADMIRE or RUBIC in the Helios pipeline.

6.4. RECURRENCE ANALYSIS AND ITS SENSITIVITY TO SEGMENTATION ALGORITHMS

In a recurrence analysis we roughly consider two types of noise:

- Measurement noise. Noisy copy number measurements are made per probe (or in bins for genomic sequence data). Typically, measurement noise can be considered independent between adjacent probes within a single sample.
- Biological noise. Since we are interested in driver aberrations, all passenger aberrations (albeit real biological events) are considered to be noise as well. For ADMIRE and RUBIC we developed null models to describe the biological noise.

Single sample segmentation algorithms are employed to discover the location of copy number break points on the genome. For our recurrence analysis, these algorithms are nothing more then a step in the pipeline to reduce measurement noise. If a segmentation algorithm calls too many false positive break points but correctly identifies the correct breaks as well, we end up with profiles that contain all the relevant copy number aberrations, but with suboptimal noise reduction. On the other hand, overly conservative segmentation algorithms will smooth away not only measurement noise but also true biological signal (driver aberrations).

It stands to reason that a good recurrence algorithm should be robust against measurement noise and therefore the choice of segmentation algorithm, albeit lose power when smoothing away true breaks. We believe that this is one of the key reasons why RUBIC not only outperforms peel-off based algorithms such as GISTIC2.0, but also performs more robustly across different technological platforms.

GISTIC2, on the other hand, is highly dependent on the segmentation algorithm being employed. Peel-off algorithms iteratively remove aberrations from segment-derived copy number profiles to detect seemingly independent recurrent loci across samples. Such methodologies are inherently sensitive to the segmentation algorithm employed since broad aberrations might appear focal due to false positive break points and hence derail the peel-off strategy. When no segmentation is employed, and GISTIC2 is performed on unsegmented profiles, it is equivalent to telling the algorithm that there is a break located between all adjacent probes on the genome. The algorithm will break down due to the very large number of one probe events being called.

In contrast, nowhere in the methodology of ADMIRE or RUBIC do we assume that the copy number profiles are segmented. Both these algorithms smooth aggregate profiles (with kernels) and therefore reduce both measurement and biological noise. The cyclic null hypotheses, on the other hand, models both uncorrelated measurement and biological noise. On simulated datasets where we add uncorrelated measurement noise to segmented profiles, RUBIC calls very similar recurrent regions, irrespective of whether we segment the samples or not. The fact that RUBIC performs similarly between the two extremes of "perfect" segmentation and trivially liberal segmentation suggests that the algorithm will behave robustly in any realistic "in-between" scenario.

This raises the question, why do we segment our samples before applying ADMIRE or RUBIC? There are multiple reasons for this, but the most important reason is that we believe that single sample segmentation algorithms are more proficient at removing measurement noise than the

smoothing methodologies employed by ADMIRE and RUBIC, since it explicitly accounts for the segmented nature of the data. Second, it is important to remove germline copy number variations before employing ADMIRE or RUBIC. Detecting and removing these copy number variations are usually performed after segmentation. From a practical point of view it is also more convenient to input small segmented files that contain only information about breakpoint locations and the average segment amplitudes than millions of measurement probes.

6.5. THE EXPECTED EULER CHARACTERISTIC

In this thesis we employed the expected Euler characteristic (from differential topology) as a statistic for significance in our algorithms. Our algorithms seek to call peaks in one dimensional datasets relative to a well defined null-model. For any given threshold and any realisation of our null-model (i.e. a realisation of a stochastic process), the Euler characteristic equals the number of up-crossings with respect to the threshold. In other words, it basically counts the number of peaks above the given threshold. The expected Euler characteristic is then defined to be the expected number of "peaks" to cross the threshold (across realisations) in the null. We believe that for a peak calling algorithm, this is a very natural statistic to use since if one performs FDR control on called peaks, it is the expected ratio between the number of false peaks (modelled by the null) and true peaks that is of interest. From a theoretical standpoint, the expected Euler characteristic is also a powerful multiple test corrected statistic since it explicitly accounts for highly correlated test measurements (probes) in close proximity on the genome and therefore reduces the effective number of tests to correct for. The final advantage of the expected Euler characteristic is that we can analytically relate it to our significance thresholds if the null model is multivariate Gaussian (which is often a good assumption in aggregate experiments).The expected Euler characteristic has been used in the past for FWER control in applications such as neuro-imaging, where voxel measurements are highly correlated in space. Our contribution was to extend this theory for scale spaces (dynamic smoothing) and to adapt its usage to (the less restrictive) FDR control.

6.6. SCALE SPACES

Another central theme was to employ scale spaces. In essence we dynamically changed the level of smoothing required when detecting peaks at different widths. The idea here is that we can gain tremendous statistical power for broad peaks since we effectively smooth away (measurement and biological) noise. Throughout our work we constantly used the expected Euler characteristic as a multiple test corrected statistic (across the genome) for calling peaks at a fixed scale (fixed kernel used for smoothing). This statistic on its own was not sufficient in the sense that we needed to correct for multiple testing across scales as well. For ADMIRE, we introduced a resolution parameter α to ensure that FDR control on a single scale also applied to all scales. The introduction of this parameter was somewhat ad-hoc and, although practically useful in this application, needed to be fine-tuned to maximise statistical power while maintaining FDR control. For RUBIC and our single sample segmentation algorithms, this parameter was neatly disposed of. We showed that when using wavelet kernels to segment aggregate profiles, there exists a direct correspondence between the expected Euler characteristic (as employed on a single scale) and the expected number of local maximum segments when considering multiple scales in a clustering scheme.

6.7. CLOSING REMARKS AND OUTLOOK

We developed a statistical framework combining scale spaces and the expected Euler characteristic in the context of peak (or break) calling for DNA copy number data. The scale space methodology allows us to gain statistical power for calling peaks at different widths and the expected Euler

characteristic is, in our opinion, the most natural statistic to use for peak calling. This is because the Euler characteristic, combined with the scale space methodology, allows us to directly control the false discovery rate on called peaks (i.e. the expected proportion of false positive peaks), instead of called probes (i.e. the expected proportion of false positive probes in peaks). Consequently this measure is insensitive to platform resolution, but sensitive to the inherent auto-correlation in the noise, the latter of which greatly alleviates multiple test correction in coloured noise. In addition we derived accurate analytical approximations for the expected Euler characteristic which allows us to completely (partially) eliminate time consuming permutation tests in Gaussian (non-Gaussian) noise profiles respectively.

Although the peak callers in this thesis are applied to copy number data, the theory can be used in any application, provided we can estimate the variance (or covariance for multidimensional measurements) and auto-correlation (correlation between measurements at different genomic positions) parameters of the noise. We hope that this framework will pave the way for future peak callers in other domains including, but not limited to, CIS discovery in insertional mutagenesis, ChIP-seq peak callers and border detection in raw sequencing data derived from Oxford nanopore technologies.

REFERENCES

[1] T. Kuilman, A. Velds, K. Kemper, M. Ranzani, L. Bombardelli, M. Hoogstraat, E. Nevedomskaya, G. Xu, J. de Ruiter, M. P. Lolkema, *et al.*, *Copywriter: Dna copy number detection from off-target sequence data,* Genome biology **16**, 49 (2015).

[2] E. Talevich, A. H. Shain, T. Botton, and B. C. Bastian, *Cnvkit: Copy number detection and visualization for targeted sequencing using off-target reads,* BioRxiv , 010876 (2014).

[3] F. Sanchez-Garcia, P. Villagrasa, J. Matsui, D. Kotliar, V. Castro, U.-D. Akavia, B.-J. Chen, L. Saucedo-Cuevas, R. R. Barrueco, D. Llobet-Navas, *et al.*, *Integration of genomic data enables selective discovery of breast cancer drivers,* Cell **159**, 1461 (2014).

SUMMARY

Cancer is a genetic disease. The activation, alteration or deactivation of cancer genes can stimulate undesirable cell-proliferation. Cancer genes can be subdivided into oncogenes and tumor suppressors. Oncogenes, such as growth factor receptors, are altered and/or overexpressed genes that are causally linked to tumorigenesis. Tumor suppressors, by contrast, are typically underexpressed or deleted in tumors since they would otherwise serve a protective role.

There are two main genetic mechanism that can activate or deactivate cancer genes: mutations and DNA copy number alterations. In this work, we focus on detecting novel cancer genes using somatic DNA copy number data. The philosophy is simple: if independently acquired somatic amplifications or deletions occur frequently across multiple tumor samples, they are likely to harbor oncogenes or tumor-suppressors respectively. With a single tumor DNA copy number profile, it is not possible to know which copy number alterations activate or deactivate cancer genes, since many of the alterations (referred to as passenger aberrations) occur due to genomic instability and do not necessarily provide a selective advantage for cancerous cells. However, when aggregating across many samples, we expect cancer genes to be amplified or deleted more frequently than by chance, which allows us to detect them.

This application can be regarded as a peak calling problem. We aggregate (sum) copy number profiles across many tumors and call peaks that are significantly high. To do this we define a null-model that describes the behavior of an aggregate copy number profile that would arise if only passenger aberrations occurred. The null aggregate profile (also called the noise profile) exhibits high autocorrelation across the genome due to the segmented nature of copy number profiles.

We therefore developed a statistical framework for calling peaks (at varying widths) where the noise profile can exhibit strong autocorrelation. The framework allows us to detect peaks (at varying widths) with high statistical power while controlling the false discovery rate of detected peaks. We employ two concepts. First, we take advantage of the fact that broad peaks can be detected with much higher statistical power when smoothing the profile and we developed techniques for adaptive smoothing. Second, we use a powerful statistic called the expected Euler characteristic that is insensitive to platform resolution, directly compatible with our smoothing methodology and that can be directly used to estimate the expected number of false positive peaks called.

This framework does not rely directly on the inherent properties of DNA copy number profiles and can therefore be applied in many more applications with suitably defined null-models. Although the mathematics we develop in this framework might be taxing at times, we observe that the equations that result and that are ultimately used in our peak calling algorithms are simple and the validity can easily be verified by simulating data and comparing our theoretical expectations with measured observations.

SAMENVATTING

Kanker is een genetische ziekte. De activering, mutatie of deactivering van kankergenen kan ongewenste celdeling stimuleren. Kankergenen kunnen worden onderverdeeld in oncogenen en tumorsuppressoren. Oncogenen, zoals groeifactor-receptoren, zijn veranderde en/of tot overexpressie gebrachte genen die causaal zijn gekoppeld aan tumorigenese. Tumor suppressors, daarentegen, zijn meestal ondervertegenwoordigd in tumoren, of ontbreken volledig. Deze genen hebben immers een beschermende rol.

Er zijn twee belangrijke genetische mechanismen die kankergenen kunnen activeren of deactiveren: veranderingen binnen het gen en wijzigingen in het aantal DNA-kopieën. In dit werk concentreren we ons op het detecteren van nieuwe kankergenen met behulp van somatische DNA-copynumberdata. De filosofie is eenvoudig: als onafhankelijk verworven somatische amplificaties of deleties vaak voorkomen in meerdere tumormonsters, zullen ze waarschijnlijk respectievelijk oncogenen of tumoronderdrukkers bevatten. Met een copynumberprofiel van een enkele tumor is het niet mogelijk om uit te vinden welke copynumberwijzigingen kankergenen activeren of deactiveren. Dit komt doordat veel van de veranderingen (aangeduid als passagiersafwijkingen) optreden als gevolg van genomische instabiliteit en niet noodzakelijkerwijs vanwege een selectief voordeel voor kankercellen. Wanneer we echter aggregeren over veel monsters, verwachten we dat kankergenen vaker dan door kans alleen worden gedupliceerd of verwijderd, waardoor we ze wel kunnen detecteren.

Dit vraagstuk kan als een piekdetectieprobleem worden beschouwd. We sommeren de copynumberprofielen van veel tumoren en ontdekken daarmee significant hoge pieken. Hiertoe definiëren we eerst een nul-model dat het gedrag beschrijft van een geaggregeerd copynumberprofiel dat zou ontstaan als er alleen passagiersafwijkingen zouden optreden. Het nul-aggregaatprofiel (ook wel het ruisprofiel genoemd) vertoont een hoge autocorrelatie binnen het genoom als gevolg van de gesegmenteerde aard van copynumberprofielen.

We hebben daarom een statistisch raamwerk ontwikkeld voor het ontdekken van pieken (met verschillende breedten) waarbij het ruisprofiel een sterke autocorrelatie kan vertonen. Het raamwerk stelt ons in staat pieken te detecteren (met variërende breedtes) met een hoog statistisch vermogen terwijl de *false discovery rate* van gedetecteerde pieken wordt beperkt. We benutten twee concepten. Ten eerste maken we gebruik van het feit dat brede pieken met een veel hoger onderscheidend vermogen kunnen worden gedetecteerd na het gladstrijken of afvlakken van het profiel. Derhalve hebben we technieken voor adaptieve afvlakking ontwikkeld. Ten tweede gebruiken we de verwachte Euler-karakteristiek, een steekproeffunctie met een hoog onderscheidend vermogen. Deze is ongevoelig is voor de platformresolutie, en is direct compatibel met onze afvlakkingsmethodologie. Bovendien kan de verwachte Euler-karakteristiek direct worden gebruikt om het verwachte aantal fout-positieve pieken te schatten.

Onze methodologie wordt niet beperkt door de inherente eigenschappen van DNA-copynumberprofielen en kan daarom in veel meer toepassingen worden gebruikt, mits het nul-model juist is gedefiniëerd. Hoewel de wiskunde die we in dit kader ontwikkelen soms lastig is, zijn de vergelijkingen die daar uit voortvloeien en uiteindelijk in onze piekdetectie-algoritmen worden gebruikt juist eenvoudig. Bovendien kan de juistheid van de vergelijkingen gemakkelijk kan worden geverifiëerd door data te simuleren en onze theoretische verwachtingen te vergelijken met gemeten observaties.

BIBLIOGRAPHY

Shugay, Mikhail, et al. "VDJdb: a curated database of T-cell receptor sequences with known antigen specificity." Nucleic acids research 46.D1 (2017): D419-D427.

Van Dyk, Ewald, et al. "RUBIC identifies driver genes by detecting recurrent DNA copy number breaks." Nature communications 7 (2016): 12159.

Iorio, Francesco, et al. "A landscape of pharmacogenomic interactions in cancer." Cell 166.3 (2016): 740-754.

van Dyk, Ewald, et. al. , “DNA copy number segmentation with clustering on the expected Euler characteristic”, Manuscript in preparation.

Schouten, Philip C., et al. "Robust BRCA1-like classification of copy number profiles of samples repeated across different datasets and platforms." Molecular oncology 9.7 (2015): 1274-1286.

Massink, Maarten PG, et al. "Proper genomic profiling of (BRCA1-mutated) basal-like breast carcinomas requires prior removal of tumor infiltrating lymphocytes." Molecular oncology 9.4 (2015): 877-888.

van Dyk, Ewald, et. al. "A scale-space method for detecting recurrent DNA copy number changes with analytical false discovery rate control." Nucleic acids research 41.9 (2013): e100-e100.

Schouten, Philip C., et al. "Platform comparisons for identification of breast cancers with a BRCA-like copy number profile." Breast cancer research and treatment 139.2 (2013): 317-327.

van Heerden, Charl, et al. "Combining regression and classification methods for improving automatic speaker age recognition." Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on. IEEE, 2010.

Van Dyk, Ewald, and Etienne Barnard. "Binary Naive Bayesian classifiers for correlated Gaussian features: A theoretical analysis." (2008).

Van Dyk, H. O., and Etienne Barnard. "Naive Bayesian classifiers for multinomial features: a theoretical analysis: pattern recognition special edition." South African Computer Journal 2008.40 (2008): 37-43.

CURRICULUM VITÆ

Hendrik Oostewald VAN DYK

09-02-1983	Born in Pretoria, South Africa.

EDUCATION

1996–2001	High School Hoërskool Silverton, Pretoria, South Africa
2002–2006	Undergraduate in Electronic Engineering (BEng) University of Pretoria, South Africa
2007–2009	Master in Computer Engineering (MEng) North-West University, South Africa
Thesis:	Classification in high dimensional feature spaces
Promotor:	Prof. dr. E. Barnard
2009–	Ph.D Computational Cancer Biology Netherlands Cancer Institute & Delft University of Technology
2015–2018	Postdoctoral research Utrecht University
2018–	Postdoctoral research Netherlands Cancer Institute

PROFESSIONAL EXPERIENCE

2006–2007	Contract worker at Rapid Mobile (Pty) Ltd
2007–2009	Human Language Technologies (HLT) at Meraka institute

ACKNOWLEDGEMENTS

Assume there exists a linear order in which I want to thank people. It is hopeless to infer this from data. The number of dimensions in which people touched me far exceeds the number of people that are. To do so would require strict and inherently biasing regularization. I therefore solve this bias-variance tradeoff by avoiding it altogether. Nevertheless, not even a supervised approach is needed to know that my supervisors are at the top of this list:

Lodewyk Wessels: Thank you for allowing me to do my PhD under your supervision and guiding me to grow both as a scientist and as a person through our numerous and fruitful discussions. Breakthroughs were sometimes sparse and far apart, but your patience stretched further. Your strength and appreciation for both the theoretical and practical is invaluable and I greatly admire that. Thank you for everything.

Marcel Reinders: Your critical and unbiased evaluation and contribution to my work (especially in the first 3 years) is also beyond measure. I think it is safe to say that without you and Lodewyk this work would not have been possible.

Obviously without my parents nothing would have been possible. Thank you for teaching me a deep appreciation for science from a young age and for always being supportive. To my brother Jaco: above all, you taught me the value of intellectual integrity, i.e. trying to see things for what they are, and not for what I might want them to be. And for making me realize the awesome power in abstract thinking. My sister, Marlene: thank you for keeping me and my brother down to earth since we spent way too much time watching Star trek.

At this point you might expect me to say that it's impossible to list all the people that influenced me. I shall attempt to name everyone before, during my PhD and up to this point in time:

Guillem Rigaill, Erika Cantelli, Bram Gerritsen, Chelsea McLean, Philip Schouten, Anna Miquel, Gwen Dackus, Jasper Claassen, Miranda van Dongen, Christian Heinzel, Claudia Heinzel, Peter Remeijer, Marlies Pasler, Jasper Nijkamp, Arturo Perez Rivera, Christ Leemans, Stephanie van Hoppe, Ioana Niculescu, Jennifer Chen, Aurora Cerutti, Cesare Lancini, Chiara Cattaneo, Glenn Norton, Lander de Vroede, Akke Pinkster, Jorn Bom, Foteini Tsakou, Fernando Salgado Polo, Jan Ree, Nora Franzen, Torben Wriedt , Anina Swanepoel, Santiago Gisler, Alessandra Buoninfante, Petrit Podrimja, Alvaro Docio, Renato Teixeira, Quirine Kakebeeke, Jelle Kakebeeke, Joanna Kaplon, Maarten Hoekstra, Sedef Iskit, Jeroen Besseling, Judith Muller, Behzad Mombeini, Jeroen Nijwening, Klaas de Lint, Catrin Lutz, Ana Moises da Silva, Stefano Annunziato, Laura Bornes, Melanie Lindenberg, Arnold Bos, Anke Wind, Bruno Vieira, Maria Escala Garcia, Kristian Naydenov, Maarten Slagter, George Damaskos, Andrea Murachelli, Roelof Pruntel, Nikolina Babala, Michal Szczotkowski, Gian-Luca McLelland, Antonio Mulero Sanchez, Bas Pilzecker, Marco Simonetta, Ben Morris, Joao Neto, Darren Sugrue, Tsubasa Matsui, Daniele Giardiello, Rui Lopes, Marcelo Sobral-Leite, Marjolein Droog, Jeroen de Ridder, Christiaan Klijn, Eva Brinkman, Johan Kuiken, Devin Barry, Jolanda Zwagemaker, Rajith Bhaskaran, Eric Pinto Barbera, Nur Biuret, Etienne Barnard, Erik Voets, Laura Bruckner, Guus Heynen, Sandra van den Broek, Ahmed Elbatesh, Anirudh Prahallad, Jacobien Kieffer, Chiara Brambillasca, Giusi Ferone, Lorenzo Bombardelli, Yvonne Geurts, Tom de Wit, Olga Blomberg, Serena Vegna, Leila Akkari, Lorenzo Spagnuolo, Max Wellenstein, Riccardo Mezzadra, Camilla Salvagno, Marieke Bruggemann, Tisee Hau, Lisanne Raeven, Kim Vrijland, Hannah Garner, Anni Laine, Antoinette van Weverwijk, Noor Bakker, Danique Duits, Kevin Kos, Karin de Visser, Daniel de Groot, Ronak Shah, Aldo Spanjaard, Andriy Volkov, Irene van der Haar, Elselien Frijlink,

Patty Lagerweij, Tom Battaglia, Julian de Ruiter, Andreas Schlicker, Sander Canisius, Bram Thijssen, Nanne Aben, Jelle ten Hoeve, Soufiane Mourragui, Daniel Vis, Johann de Jong, Hayssam Soueidan, Jorma de Ronde, Magali Michaut, Marlous Hoogstraat, Gergana Bounova, Tycho Bismeijer, Monique Carreno, Evert Bosdriesz, Joana Goncalves, Kathy Jastrzebski, Tesa Severson, Christian Rausch, Willem Koemans, Petra Nederlof, Esther Lips, Can Kesmir, Rob de Boer, Berend Snel, Bas Dutilh, Sandro Colizzi, Thomas Cuypers, Gijs Schroder, Pieter Visser, Daniel Weise, Ianthe van Belzen, Bram van Dijk, Juliane Schroter, Divyae Prasad and Christiaan van Dorp.

I apologize if I forgot anyone. It is impossible to list all the people that influenced me.