KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Engineering

Development of computational methods to detect copy number alterations in human genomes

Jiqiu Cheng

Supervisory Committee:
Prof. dr. ir. A. Bultheel, chair
Prof. dr. ir. Y. Moreau, supervisor
Prof. dr. ir. J. Vermeesch, co-supervisor
Prof. dr. ir. B. De Moor
Prof. dr. ir. N. Buys
Dr. ir. P. Van Loo
Prof. dr. O.C. Lingjærde
(Oslo university)

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor in Engineering

May 2013

Development of computational methods to detect copy number alterations in human genomes

Jiqiu CHENG

Supervisory Committee:
Prof. dr. ir. A. Bultheel, chair
Prof. dr. ir. Y. Moreau, supervisor
Prof. dr. ir. J. Vermeesch, co-supervisor
Prof. dr. ir. B. De Moor
Prof. dr. ir. N. Buys
Dr. ir. P. Van Loo
Prof. dr. O.C. Lingjærde
(Oslo university)

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor in Engineering

May 2013

© KU Leuven – Faculty of Engineering
Kasteelpark Arenberg 10 - bus 2446, B-3001 Heverlee (Belgium)

Alle rechten voorbehouden. Niets uit deze uitgave mag worden vermenigvuldigd en/of openbaar gemaakt worden door middel van druk, fotocopie, microfilm, elektronisch of op welke andere wijze ook zonder voorafgaande schriftelijke toestemming van de uitgever.

All rights reserved. No part of the publication may be reproduced in any form by print, photoprint, microfilm or any other means without written permission from the publisher.

D/2013/7515/52
ISBN 978-94-6018-666-0

Preface

This thesis summarizes my research work as a PhD student in the bioinformatics group, SCD-SISTA research division, Electrical Engineering Department of the Katholieke Universiteit Leuven. It has been an amazing experience to work in this group and I would like to give acknowledgment to all the people who have helped and supported me.

My motivation to start the doctoral program originated from the course taught by prof. Yves Moreau in the master of bioinformatics program. I was fascinated by the amazing point that computational methods can be used to solve real medical problems. Since then, I have made my decision to do PhD research in bioinformatics.

My deepest gratitude goes to my promoter Prof. Yves Moreau for giving me the opportunity to start my PhD research under his supervision. His joint-expertise and encouragements were a great help for me to find my way out. Enthusiastic discussions with him always inspired me to get new ideas. In addition, Yves himself is active in cooperating with different biomedical groups. I appreciate a lot for Yves' fully understanding and continuous funding to support my research work and travel costs to attend international conferences. Thanks to him, I got the chance to work together with colleagues from Gasthuisberg, Oslo University and Sanger institute.

Next, I would like to express my grateful thank to my co-promoter Prof. Joris Vermeesch, for his illuminating guidance in the field of genome research. The active discussions with him helped me get a comprehensive understanding of human genome and provided me the background knowledge of the biological data I work with. I also appreciate him for being able to propose novel biological questions that can be solved using computational methods.

Furthermore, I like to acknowledge Peter Van Loo as an important supervisor of my cancer projects. I thank him for taking time to answer all my questions and giving insight discussions and dedicated contributions to the manuscript

we prepared together. The scientific texts in the manuscript were considerably improved due to his suggestions. As introduced by him, I got the opportunity to meet and cooperate with excellent researchers from Sanger institute and Oslo University.

May I extend my thanks to the rest of the professors prof. Ole Christian Lingjærde, prof. Nadine Buys, prof. Bart De Moor and prof Adhemar Bultheel, who have accepted to join my doctoral examination committee. I appreciate them for spending time reading my doctoral thesis and giving constructive comments to improve my thesis. In particular, I would like to thank prof. Ole Christian from Oslo University for bringing many innovative suggestions of my previous research work.

Besides professors, I would also like to give my gratitude and respect to my fantastic colleagues. Peter Konings from Yves' group is my first tutor to lead me to start my doctoral program. With Peter Konings' help, I got easily settled in my working environment. By working closely with him, I was impressed by his solid knowledge in statistics and concrete attitude in research, which set a good example for my following research. I would also like to thank Amin Ardeshirdavani, Alejandro Sifrim and Arnaud Installe for helping me being familiar with the use of Avalok cluster. In addition, I appreciate all the help I received from the administrative and professional staff, especially Maarten Truyens for maintaining data on Sista server and Herman Van Biesen for managing Avalok cluster in Gasthuisberg.

My research work was a product of joint collaborations from different research groups. I want to thank Evelyne Vanneste, a colleague from Joris' group in Gasthuisberg, for helping me understand the amazing story of human embryos. I would also like to thank the other colleagues from CME group in Gasthuisberg, prof. Thierry Voet, Niels Van Der Aa, Eftychia Dimitriadou, Masoud ZamaniEsteki and the joint work with Afroditi Mertzanidou and Claudia Spits from Vrije Universiteit Brussel. Furthermore, I would like to express my thank to David Wedge, prof. Peter Campbell, prof. Mike Stratton from Sanger institute, Hege Russnes, Hans Kristian Moen Vollan, Gro Nilsen, Silje Nord, prof. Vessela N. Kristensen and prof. Anne-Lise Børresen-Dale from Oslo University. It was a big pleasure to work with so many people from different backgrounds. Their enthusiastic interactions and efficient feedbacks contributed a lot to the progress of my research work.

Thanks for sharing so many nice memories with me, Prof. Jan Aerts, Griet Laenen, Ryo Sakai, Raf Winand, Sarah Elshal, Pooya Zakeri, Marc Claesen, Tunde Adefioye, Leon-Charles Tranchevent, Nico Verbeeck, Minta Thomas, Dusan Popovic, Georgios Pavlopoulos, Ernesto Iacucci, Charalampos Moschopoulos, Daniela Nitsch, Yu shi, Sonia Leach, Raf Van de Plas, Lieven

Thorrez, Olivier Gevaert, Yousef El Aalamat, Joana Goncalves and Anneleen Daemen. It was a nice joy to study and work with all of you in this friendly and peaceful environment in Esat.

Ultimately, I would like to give thank to my lovely parents in China, my friends and families in Belgium for their comfort and encouragement. Especially thank Xiaolin for helping me learn Latex to write this thesis. I appreciate my father-in-law and sister-in-law for translating the abstract of my thesis into dutch. I would like to give the glory to my husband Bart, who is always patient, interested and supportive for my research work.

Jiqiu Cheng
Leuven, May 2013



Abstract

In the past decades, advanced array technologies have been developed to look into the molecular structure of the whole human genome in order to unravel human diseases. However, the large amount of data resulting from these innovations have brought new challenges to biologists since the traditional ways to look for disease associated regions or genes are time consuming and inefficient. Due to the high complexity of biological data, proper efficient computational approaches and analysis methods are essential to find correct answers. This thesis is directed towards development of computational approaches for copy number alteration analysis of human genomes.

Theoretically, the majority of human genomes contain two copies of chromosomes inherited from each parent. Copy number alteration analysis is the detection of genome regions which contain a copy number different from two. Genome-wide exploration of copy number alteration regions is important since some of these regions are suspected to be relevant to the underlying mechanisms of the diseases.

This thesis consists of two parts. The first part describes the development of computational approaches for the detection of copy number alterations in single-cell genomes. Single-cell genome research is a novel area in which the DNA material is extracted from a single cell instead of a pooling of cells as usual. However, the array signals from single-cell genomes often contain a lot of artifacts caused by the amplification of limited amount of DNA material. We develop a novel channel-clone approach to process this challenging single-cell data and correct genome artifacts. By testing this approach in experimental validated genomes, we show that the accuracy of copy number alteration detection is considerably improved after the channel-clone processing. In addition, we apply this approach to single blastomeres extracted from pre-implanted human embryos and discover chromosome instability in the cleavage-stage of human embryos.

The second part of this thesis focuses on the analysis of allelic-specific copy number alterations in cancer genomes and the identification of cancer-causing genes. Cancer genomes are the genomes located in cancer cells. Since cancer cells are the result of accumulated mutations, it is straight forward to explore mutated regions within cancer genomes and identify mutated genes residing in these aberrant DNA sections. We extend the previous developed allelic-specific copy number analysis approach (ASCAT) to detect copy number alterations in subclonal populations of cancer cells and apply this method to public available SNP array data. This results in a large consortium of patient-specific copy number profiles from in total 12 cancer types. By comparing copy number profiles, we observe the tumor type specific patterns and identify 89 cancer-related regions including known and candidate cancer-causing genes. Our work reveals a spectacular landscape of cancer genomes across different cancer types and provides valuable directions for future cancer research.

In conclusion, this thesis presents different computational approaches to detect aberrations in single-cell genomes as well as cancer genomes and assists biologists and clinicians to create a more comprehensive understanding of human genomes.

Samenvatting

Het afgelopen decennium werden geavanceerde array-technieken ontwikkeld om de moleculaire structuur van het volledige menselijke genoom te onderzoeken. Dit met het oog op het in kaart brengen van gengerelateerde ziekten bij de mens. De hoge complexiteit van de grote hoeveelheid biologische gegevens afkomstig van deze innovaties stelt de medische wereld voor nieuwe uitdagingen. Gerichte en efficiente rekenkundige benaderingen en analysemethoden zijn immers essentieel voor het vinden van de juiste antwoorden. Dit proefschrift richt zich op de ontwikkeling van wiskundige methoden voor het analyseren van wijzigingen in het aantal kopieen in menselijke genomen.

Normaal gesproken bevat het menselijk genoom twee kopieen van eenzelfde chromosoom, geerfd van elke ouder. In het onderzoek naar veranderingen in het aantal kopieen wordt er gezocht naar regio's in het genoom met een kopieaantal dat niet gelijk is aan twee. Genoomverkenning van regio's met wijzigingen in het kopieaantal is belangrijk, omdat een deel van deze zones vermoedelijk relevant zijn met betrekking tot de onderliggende mechanismen van aandoeningen.

Dit proefschrift bestaat uit twee delen. Het eerste deel beschrijft de ontwikkeling van wiskundige methoden voor de detectie van veranderingen in kopieaantal in genomen afkomstig van een enkele cel. Genoom onderzoek waarbij DNA materiaal geextraheerd wordt uit een enkele cel in plaats van uit een grote hoeveelheid aan cellen is een nieuw onderzoeksgebied. Array data van het genoom van een cel bevat vaak veel artefacten veroorzaakt door de amplificatie van een beperkte hoeveelheid DNA materiaal. We ontwikkelen een nieuwe kanaal-kloon normalisatie voor de verwerking van deze complexe een-cel data en corrigeren genoomartefacten. Door het testen van deze technieken aan de hand van experimenteel gevalideerde genomen, tonen we aan dat de nauwkeurigheid van de detectie van kopieaantal wijzigingen aanzienlijk verbeterd wordt na kanaal-kloon normalisatie. We passen deze methode toe op blastomeren die geextraheerd werden uit nog niet geimplanteerde menselijke embryo's om chromosoom-instabiliteit in het delingsstadium van deze embryo's op te sporen.

Het tweede deel van dit proefschrift richt zich op de analyse van allel-specifieke kopieaantal wijzigingen in kankergenomen en de identificatie van kanker gerelateerde genen. Kankergenomen zijn genomen die men kan vinden in kankercellen. Aangezien kankercellen het resultaat zijn van geaccumuleerde mutaties is het vanzelfsprekend om te zoeken naar genen die zich bevinden in gemuteerde regio's binnen kankergenomen. De eerder ontwikkelde ASCAT methode voor de detectie van allel-specifieke kopieaantal wijzigingen breiden we uit naar subklonale populaties van kankercellen en we passen deze methode toe op publiek beschikbare SNP array data. Dit resulteert in een grote verzameling van patientspecifieke profielen met kopieaantallen van in totaal 12 types van kanker. Door het vergelijken van deze profielen observeren we tumortype specifieke patronen en identificeren we 89 kanker gerelateerde regio's, waaronder gekende en kandidaat kankerverwekkende genen. Ons onderzoek toont een spectaculair landschap van kankergenomen verspreid over verschillende vormen van kankers en biedt waardevolle aanwijzingen voor toekomstig kankeronderzoek.

Tenslotte presenteert dit proefschrift verschillende wiskundige methoden om afwijkingen op te sporen via DNA extractie van een-cel genomen en van kankergenomen. Het helpt biologen en artsen om een meer omvattende kennis te verwerven over het menselijk genoom.

Abbreviations

ART	Assisted reproduction treatment
ASCAT	Allelic-specific copy number analysis of tumors
ASPCF	Allele-Specific Piecewise Constant Fitting
BAC	Bacterial artificial chromosomal
BAF	B allele frequency
CA	Channel plus recurrent genome artifact correction without CBS segmentation
CAAI	Complex arm-wise aberration index
CBS	Circular binary segmentation
CG	Channel plus genome composition correction using enlarged window GC contents
CGACBS	Channel plus genome composition correction using enlarged window GC contents plus CBS segmented residuals followed by recurrent genome artifact correction
CGH	Comparative genomic hybridization
CGprobeA	Channel plus genome composition correction using probe GC contents plus recurrent genome artifact correction without CBS segmentation
Channel clone	Channel plus genome composition correction using enlarged window GC contents plus recurrent genome artifact correction without CBS segmentation
CNA	Copy number alteration
CNV	Copy number variation
COSMIC	Catalogue of somatic mutations in cancer
DNA	Deoxyribose nucleic acid

EBV	Epstein-barr virus
EM	Expectation-maximization
FDR	False discovery rate
FISH	Fluorescence in situ hybridization
FPR	False positive rates
HD	Homozygous deletion
HMM	Hidden markov model
IVF	In vitro fertilization
KS	Kolmogorov-Smirnov
LOH	Loss of heterozygousity
LRR	Log R Ratio
MD	molecular dynamics
MDA	Multiple displacement amplification
NGS	Next generation sequencing
PGD	Preimplantation genetic implantation
PGS	Preimplantation genetic screening
PGT	Preimplantation genetic testing
QQ	Quantile-Quantile
SD	Standard deviation
SNP	Single nucleotide polymorphism
TPR	True positive rates

Contents

Abstract v

Contents xi

1 Introduction 1

1.1 Human genome 1

1.2 Mutation 2

1.3 Preimplantation genetic testing (PGT) 4

1.4 Cancer genome 6

1.5 Thesis overview 7

2 Materials and Methods 9

2.1 Protocols of single-cell copy number alteration analysis 9

2.1.1 Array platforms 10

2.1.2 Whole genome amplification technologies 11

2.1.3 Single-cell copy number analysis algorithms 12

2.1.4 Validation methods 13

2.2 Allelic-specific copy number analysis algorithms 15

3 Development of channel-clone normalization for single-cell copy number alteration detection 19

3.1 Summary 19
3.2 Personal contributions 20
3.3 Postscript 48

4 Channel-clone normalization applied to array data from human embryos 51
4.1 Summary 51
4.2 Personal contributions 52
4.3 Postscript 66

5 Allelic-specific copy number analysis across tumor types 73
5.1 Summary 73
5.2 Personal contributions 74

6 Conclusions and Perspectives 103
6.1 Single-cell genome analysis 103
6.1.1 Conclusion 103
6.1.2 Perspectives 104
6.2 Cancer genome 106
6.2.1 Conclusion 106
6.2.2 Perspectives 107

Bibliography 111

List of publications 127

Curriculum Vitae 129

Chapter 1

Introduction

Genome research investigates information from the human genome to gain valuable knowledge and insights into the central role of the genome in human diseases. With the fast development of genome technologies, a large amount of biological data is being produced. To convert this large data ocean into helpful information for diagnostics, prognostics and even treatments, advanced computational approaches are needed to model data combined with appropriate biological or medical interpretations. In our research, we develop different analysis approaches for human blastomere and cancer genome research. Before going into the technical aspects, we will start with some background on the human genome.

1.1 Human genome

Cells are the basic structural and functional units of human organisms. With the exception of red blood cells containing no nucleus, every cell contains a human genome in its nucleus, storing the complete set of genetic information which can be transmitted to offspring. This information is carried by chromosomes, which are stretches of deoxyribose nucleic acid (DNA) sequences, structured in a double helix. Besides regions encoding genetic information, some parts of the chromosome are important in relation to the overall functioning of the chromosome structure during cell division. The two arms of a chromosome are separated by the centromere region. The telomeres are located at the top end of each arm. Defects in these regions can cause mutations during cell division (see section 1.2).

Based on the division types, cells fall into germ cells going through meiosis and giving rise to gamete cells, as well as somatic cells undergoing mitosis and differentiating into tissues of an organism. Different from haploid germ cells which carry a single set of chromosomes, a human genome within somatic cells is diploid, organized into 22 homologous pairs of autosomes and one pair of sex chromosomes. A male has one X and one Y chromosome whereas a female has two X chromosomes.

Nucleotide is a basic unit of a DNA sequence composing of sugar, phosphate backbones and one of 4 types of nucleotide bases: adenine (A), thymine (T), cytosine (C) and guanine (G). The human genome consists of 3 billion nucleotides. Genes are fragments of chromosomes carrying encoded information in subsequences of nucleotides for the assembly of proteins. These proteins are finally the molecules that, when they are produced, determine together with environmental factors the functional traits of the organism, called phenotype. According to sequencing projects, around 20000-25000 protein-coding genes have been reported making up 1-2% of the human genome [1].

An alterative form of a gene or DNA sequence at the same locus is called allele. The combination of two alleles from a pair of chromosomes is called the genotype for that specific locus. Since a human genome contains two sets of chromosomes, alleles are called homozygous if both alleles for a locus are the same and heterozygous if the two alleles are different.

1.2 Mutation

Mutation is a heritable change in a DNA sequence. Mutations can be categorized into small scale and large scale regarding the mutation length of DNA sequence. Small-scale mutations include point mutations, short insertions and deletions in DNA sequence spanning single or several nucleotides. When more than one type of point mutation persists in the population at the same single nucleotide, this point mutation is called single nucleotide polymorphism (SNP). Around 10 million SNPs are present in each human genome and most of them have no influence on the phenotype. However, they are useful as genetic markers for genome research since they may indicate response to therapy and susceptibility to many diseases.

In contrast to small-scale mutations, large-scale mutations refer to chromosomal abnormalities larger than 1kb. In theory, a normal human genome in somatic cells is diploid (2n) containing two copies of each chromosome inherited from each parent. However, chromosomes may show abnormal either by numbers or

by structures. The chromosome constitution of a cell or population of cells is called its karyotype.

Abnormal number of chromosomes other than 46 is called heteroploid. An exact multiple of haploid chromosome number (n) is called euploid whereas the rests are called aneuploid. Thus haploid (n), diploid (2n), triploid (3n) and tetraploid (4n) are all euploid. With the exception of diploid (2n), all the other euploids in somatic cells are abnormal chromosome numbers and frequently observed in cancer genomes, as shown in Chapter 5 [2].

Alternatively, aneuploidy is most commonly observed in human embryos leading to miscarriages and congenital birth defects [3]. For instance, trisomies characterized by one additional chromosome account for 35% of spontaneous abortions and observed in 0.3% of newborns [4]. Down syndrome with an additional copy of chromosome 21 is a well known trisomy disorder affecting 1 in 750 newborns and the affected patients can survive until their adulthoods. Newborns with other trisomies, such as involving chromosomes 13 and 18, rarely survive after the first month of born [4]. Contrast to trisomy, monosomy is characterized by missing chromosomes with total chromosomal number smaller than 46. Females with monosomy on X chromosome are known as Turner's syndrome [5].

Chromosomal structural rearrangements are caused by chromosome breakages. Depending on whether or not there are extra or missing DNA materials, chromosomal structural rearrangements are classified into balanced and unbalanced rearrangements. Unbalanced rearrangements include duplications, deletions, rings, isochromosomes and dicentrics. Duplications and deletions are characterized by a chromosomal region duplicated or deleted leading to partial or complete trisomy, tetrasomy, monosomy or nullisomy of a chromosome. Ring chromosomes form when the telomeres of each chromosome arm break off and the broken arms reunite into a ring. An isochromosome is a chromosome with one missing arm and a duplication of the other arm. Dicentrics arise when centromere-containing fragments fuse together to form a new chromosome with two centromeres [6].

Any unbalanced rearrangements leading to an imbalance like monosomy or trisomy can result in abnormal phenotypes. The unbalanced rearrangements are collectively termed as copy number variations [7]. Copy number variation (CNV) is a widespread and common phenomenon in human genomes accounting for a major proportion of human phenotypic differences [8]. Unlike most of copy number variations which are not associated with the genetic basis of common diseases, copy number alterations are frequently observed in human cleavage-stage embryos contributing to the chromosomal instability [8, 9], as explained in Chapter 3 and 4. In the context of cancer genome, copy number

variations are termed as copy number alterations (CNA) if they originate in somatic cells and are regarded as a hallmark of tumor progression, as explained in Chapter 5 [10]. Unbalanced rearrangements can be detected by fluorescence in situ hybridization (FISH) or array technologies. As introduced in Chapter 2, the resolution of the rearrangements detection depends on the design of the array platforms.

Balanced rearrangements include inversions and translocations. Inversion is a single chromosome undergoing two breaks and one reconstitution. Inversion can be classified into paracentric when both breaks occur in the same arm and pericentric when each break occurs in each arm. Translocation is characterized by exchanging of chromosomal segments between usually two non-homologous chromosomes. Reciprocal translocations, Robertsonian translocations and insertions are three types of translocations. Reciprocal translocations arise when two segments from two non-homologous chromosomes break off and conversely attach to the other chromosomes. Robertsonian translocations occur when the long arms of two chromosomes join at the centromere and lose the short arms. Insertion happens when a segment from one chromosome inserts into a different chromosome [6].

Balanced rearrangements usually lead to few phenotypic abnormalities due to fact that the total amount of DNA material is constant. It accounts for the abnormal phenotypes in about 6% of de novo cases [11]. However, the carriers of balanced rearrangements are associated with a risk of unbalanced gametes and therefore have the risk of giving unbalanced offspring. Focal balanced rearrangements can be detected by conventional karyotype technologies. Genome-wide balanced rearrangements can not be detected by array technologies and require advanced bio-technologies, such as next generation sequencing (NGS). In this thesis, we only discuss the unbalanced rearrangements detected by array platforms.

When the chromosomal abnormalities occur in an organism, they are theoretically present in all of the cells of the organism. However, as shown in Chapter 3, 4, 5, mosaicism is commonly observed in human embryos as well as cancer genomes by showing different chromosomal complements in two or more than two populations of cells.

1.3 Preimplantation genetic testing (PGT)

Preimplantation genetic testing refers to the procedure of a cellular biopsy extraction from a human oocyte or embryo prior to implantation. The purpose of PGT is to evaluate chromosomal abnormalities in biopsy cells and optimize

the selection of good-quality embryos for the implantation. Thus, PGT is a type of prenatal diagnosis and requires in vitro fertilization (IVF) to acquire a biopsy sample [12].

With respect to the differences in goals and patient types, preimplantation genetic tests are divided into two broad categories: preimplantation genetic diagnostic (PGD) and preimplantation genetic screening (PGS) [12, 13]. PGD aims at testing the specific monogenetic disorders in embryos present in one or both of the parents to avoid the delivery of affected children or an abnormality carrier. Patients going through PGD are fertile, carry a known genetic abnormality and have the risk of passing the abnormality to offspring. On the other hand, PGS aims at screening overall chromosome aneuploidies before uterine transfer to improve success rate of live born. Different from PGD, PGS are usually applied to one of the following type of patients: advanced maternal age older than 37 years, history of recurrent miscarriages, repeated implantation failure or severe male infertility [14].

Preimplantation genetic testing removes at least one cell from each embryo and acquires the single cell biopsy in one of the three embryo developmental stages: polar body, cleavage and blastocyst. Polar bodies are produced by meiotic division of oocytes. Both first and second polar bodies are haploid maternal DNA extruded after ovulation and fertilization respectively. Since paternal DNA are absent in polar bodies, the limitation of polar body biopsy is that it only allows the test for maternal abnormality [15, 16].

Cleavage stage biopsy is performed on the developed embryos at the third day after fertilization. Most normally developed embryos consist of six to eight blastomeres at this cleavage stage, among which one or two cells are extracted from the embryo. Currently, cleavage stage biopsy is the most widely used biopsy method for PGT [12, 15]. The underlying assumption is that the genome constitutions of the removed blastomeres are representative for all the blastomeres in embryo. However, the assumption is often violated by a phenomenon called cellular mosaicism indicating that an embryo is the mixture of cell populations with different karyotypes. It has been reported that chromosome instability is common in human cleavage-stage embryos and 40-60% embryos are mosaic [9, 15]. In Chapter 4, we find additional evidence consistent with the latter.

Blastocyst stage biopsy is applied on the fifth day after fertilization. The embryos in this stage develop into inner cell mass and trophectoderm that eventually grow into fetus and placenta. Blastocyst biopsy is performed by extracting cells from the trophectoderm instead of inner cells. Although some studies claim that blastocyst stage biopsy results in a higher successful implantation rate than the other stages, a high proportion of embryos are thrown away before

PGT because only a limited number of embryos achieving blastocyst stage can be arrested in culture [17].

Every biopsy stages have their own advantages and limitations. In this thesis, we only focus on the PGS research performed on cleavage stage biopsy. FISH is traditionally used to perform PGT with a number of designed probes attached to several chromosomes. Recently, array technologies have been developed to enable to screen chromosomal aberrations at a genome wide level. The details of the array platform and its clinical applications are discussed in Chapter 2.

1.4 Cancer genome

Cancer is one of the most common and severe diseases in clinical medicine. Over 100 cancer types have been found in the world originating from around 60 different organs of the human body. According to statistics in 2008, cancer account for 7.6 million deaths among 12.7 million total cancer cases [18].

Although all cancer types show distinct biological features, they are believed to share a common pathogenesis. Like normal cells, cancer cells are the descendents of fertilized germ cells through numeric times of mitosis divisions. Normal cells evolve into cancer cells when they carry cancer-causing mutations which are either inherited from parental genomes known as germline mutations or acquired only in somatic cells termed as somatic mutations. Cancer cells eventually grow into malignant tumors by acquiring an accumulation of mutations which interrupt the normal cell function [10]. Cancer carrying inherited or somatic cancer-causing mutations are called hereditary and sporadic respectively. Most cancers are considered as sporadic cancers and we only focus on these types of cancer in this thesis.

Cancer is a consequence of a cascade of mutations. However, most of the mutations in a cancer genome are random passenger events and only a few of them are cancer-driving mutations, which confer to selective advantage and stimulate cancer progression [19]. Genes carrying these driven mutations are called cancer-driving genes and are divided into oncogenes and tumor suppressor genes. Oncogenes are genes carrying gain-of-function mutations stimulating cancer cell division and proliferation and contributing to the growth of malignant cells. On the other hand, tumor suppressor genes have a repressive or apoptosis effect on cell growth. Thus inactivated suppressor genes caused by loss-of-function mutations can no longer prevent the accumulation of cancer cells.

The identification of cancer-driving mutations or genes has been a big interest

of cancer genome research in the past decades and a number of cancer-driving genes were discovered by the analysis of mutation patterns. The oncogenes are frequently activated by point mutations or amplifications whereas suppressor genes are frequently inactivated by truncating point mutations or deletions. In Chapter 5, we describe a list of candidate tumor suppressor genes by analyzing deletion regions across all the tumor samples. It is known that cancer is a complex disease associated not only with alterations in genomes, but also alterations in transcriptomes, proteomes or even metabolomes. However, in this thesis, we only discuss the identification of genome mutations contributing to cancer development.

1.5 Thesis overview

This thesis starts with an introduction of the biological concepts of human genome, mutation, preimplantation genetic testing and cancer genome in Chapter 1. Subsequently, Chapter 2 describes the experimental protocols used to detect chromosomal aberrations introduced in Chapter 1 including array platforms, amplification technologies, computational approaches and experimental validation approaches.

Chapter 3 to Chapter 5 describes the research studies on the roles of copy number alteration in human embryo and cancer genomes. Chapter 3 proposes a novel channel-clone processing approach to adjust single-cell genomic artifacts and improve the accuracy of single-cell copy number alteration detection. In addition, it also presents a pipeline for single-cell copy number alteration analysis including preprocessing, segmentation and copy number calling. Chapter 4 describes an application of the channel-clone approach proposed in Chapter 3 to human embryos going through preimplantation genetic screening procedures. The results show that high chromosomal instability and mosaicism rate are found in cleavage-stage embryos derived from healthy couples. Chapter 5 provides a computational framework to analyze allelic-specific copy number profiles across tumor types and identify gene roles for tumorigenesis based on deletion patterns.

Finally, Chapter 6 presents a conclusion of our work. By summarizing the current human embryo and cancer genome studies, Chapter 6 also elaborates on the opportunities and directions for future research.

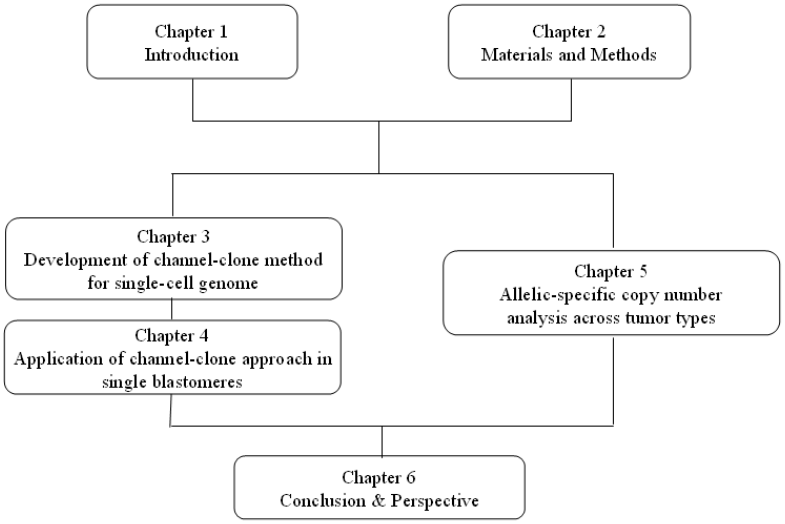

Figure 1.1: Overview of the relationships between different chapters.

Chapter 2

Materials and Methods

In Chapter 1, we introduced copy number alteration as a hallmark of embryogenesis and tumorigenesis. With the advent of array technologies, copy number alteration analysis can be performed at genome wide level.

In general, a DNA array is a solid surface containing spots systematically arranged in rows and columns. Each spot contains DNA sequences, known as probes, targeted to a specific region of a genome. DNAs are extracted from cells, cut in smaller pieces and labeled with fluorescent dyes. When they hybridize to the probes of the array, the dyes emit light that can be captured by an array scanner. The light intensity is measured with specialized software. Some arrays only use one color (or one-channel) to measure relative amounts of DNA from a test sample. Others use two colors (two-channel) for measuring DNA from both a test sample and a reference sample.

In this chapter, we will review the experimental protocols as well as computational approaches used for single-cell copy number alteration analysis. In addition, we will summarize the available computational approaches for the allelic-specific copy number analysis in cancer genomes.

2.1 Protocols of single-cell copy number alteration analysis

In recent years, array technologies have been widely used to detect copy number alterations and a variety of studies have evaluated the performance of array

platforms as well as copy number calling approaches [20]. However, these studies were performed based on genomic DNA extracted from a pooling of cells. It is difficult to evaluate array performance or computational algorithms separately for single-cell genome analysis since the accuracy of single-cell copy number detection is associated with specific array platform designs, amplification technologies and computational algorithms. Thus, the unique combination of specific protocols distinguishes one technology from another. Table 2.1 provides an overview of all the array platforms, amplification and analysis methods used for the current single-cell genome studies and the resolution they have achieved. However, the rapid developments of array technologies could change this information in the near future.

2.1.1 Array platforms

BAC array CGH
Array Comparative Genomic Hybridization (CGH) contains fragmented bacterial artificial chromosomal (BAC) clones, i.e. probes, spaced at different positions of the human genome. DNA from a reference sample and a patient are labeled with Cy3 and Cy5 respectively and hybridized to the probes of the arrays. The fluorescence intensity of the test and reference DNA is measured and the relative copy number changes for a particular genome region is estimated based on the ratios between the intensities of test and reference DNA [21].

The smallest length of the detected aberration, termed as resolution, depends on probe size and distances between probes. The averaged probe length of conventional BAC 4K arrays is 164 kb [9, 22]. Constitutional Chip 4.0 (Perkin Elmer) and 24Sure BAC (Blue Gnome Ltd., Mill Court, reat Shelford, Cambridge, UK) are two recently developed commercial BAC arrays. They comprise of around 5000 and 3000 clones respectively [21]. Both of these two commercial arrays are developed specifically for single-cell copy number analysis and have the potential to give better performance than the existing array platforms.

Oligonucleotide array CGH
The principle of oligonucleotide array CGH is similar to BAC array CGH, i.e. hybridization of test and reference DNA to probes targeted to predesigned genome positions. Instead of using long BAC probes, oligonucleotide probes are between 45 and 85 bp [21]. Oligonucleotide array CGH has been applied for single-cell copy number analysis in a range of platforms, from the early Agilent 44K to current NimbleGen 2.1M [23, 24, 25]. Since oligonucleotide arrays have a higher density of probes than BAC arrays, they are expected to show higher resolution than BAC arrays. However, more genome artifacts are observed due

to the amplification technology combined with high probe density [24]. Different from SNP arrays, one limitation of array CGH is that it can not detect balanced copy number changes such as uniparental disomy, in which both chromosomes are only inherited from one parent.

SNP arrays
SNP array has a different design from array CGH. Instead of targeting a region of a genome, SNP arrays target Single Nucleotide Polymorphism (SNP). It provides measurements of relative copy number changes as well as genotypes of selected SNPs throughout the whole genome.

Affymetrix (Affymetrix Inc. Santa Clara, CA, USA) and Illumina (San Diego, CA, USA) both produce SNP arrays with different SNP typing chemistry. Affymetrix array contains 25bp allele-specific probe sets targeting the SNPs on a glass slide. Instead of two sample hybridization, only the test sample DNAs go through digestion, adaptor ligation, PCR amplification, fragmentation, labeling and are hybridized to the glass slide containing probes. Following wash and scan the signal intensities are compared to intensities of a reference genome to calculate the relative copy number change and give genotype calls [21]. Up to now, the total SNP amount used in Affymetrix inscreased from 50K to 900K (SNP6.0). Among them, Affymetrix 50K and 500K have been applied to single-cell genome analysis [26, 27].

Illumina arrays (San Diego, CA, USA) contain 50 bp locus-specific probes attaching to beads. Test sample DNAs are amplified, labeled and hybridized to the bead array followed by wash and scan. The llumina arrays started from 31700 to current 2.5 million SNPs [21, 28, 29].

After the scanning and preprocessing, SNP arrays provide two types of measurement for each SNP array probe: the Log R Ratio (LRR) measuring total copy number, which is analogous to the log ratio measure in array CGH platforms and the B allele frequency (BAF) measuring the relative intensity of B allele to the total intensity A+B [30]. The chromosomal aberrations and the origin of uniparental disomy can be estimated using LRR and BAF measurements.

2.1.2 Whole genome amplification technologies

Array technologies require hundreds of nanograms of DNA input and single cell only contains 6pg. Therefore, single-cell DNA has to be amplified before they are hybridized to arrays. Amplification can be divided into PCR based and non-PCR based techniques. Existing PCR based techniques include previously developed degenerate oligonucleotide primed PCR (DOC-PCR), primer extension PCR,

ligation-mediated PCR [21], recently developed Genomeplex (Sigma-Aldrich, St.Louis, MO 63103, USA) and Sureplex/Picoplex (Rubicon Genomics Inc., MI 48108, USA/BlueGnome Ltd., Mill Court, Great Shelford, Cambridge, UK) [22, 31]. One commonly used non PCR-based amplification technique is multiple displacement amplification based on a strand displacement amplification at constant temperature (MDA) [32]. The researches described in Chapter 3 and 4 all use the MDA approach to perform single-cell DNA amplification.

2.1.3 Single-cell copy number analysis algorithms

In early studies, single-cell copy number analysis methods are simple intuitive threshold-based methods. A genomic region is considered to be gain or loss if the averaged log-ratios or averaged smoothing values are above or below a certain threshold. The threshold is defined based on the mean and standard deviation of log2Ratios [22, 33].

Vanneste et al. (2009), applied a normal mixture model with 3 components representing gain, neutral and loss of copy numbers respectively. First of all, clone-specific mean and standard deviation are estimated based on non-aberrant chromosomes across several samples. Then the normal mixture model is fitted to each log2Ratio and the probability of each log2Ratio belonging to a component is calculated. The mean and standard deviation of each component are defined as the clone-specific adjusted gain, neutral or loss mean values and scaled clone-specific standard deviations. The main limitation of this approach is that it requires building a reference chromosome table based on DNA from multiple cells with known karyotypes [9, 34, 35].

Cheng et al. (2011) and Mertzanidou et al. (2013) used a combination of Circular Binary Segmentation (CBS) and CGHcall to call copy number changes [25]. Circular Binary Segmentation (CBS) is based on the t-test. By connecting two ends of a chromosome as a circle, the algorithm tests whether two segments divided by probes have different means by performing a two-sample t-test. A breakpoint is considered significant if the p-values are lower than a predefined threshold (By default 0.01). Since the CBS test is based on a statistical test between two groups of signals, less breakpoints will be detected to be significant if the variance of the signals are large and if the mean difference between two groups is relatively small [36]. After CBS, which only provides the segmentations, the CGHcall algorithm uses the breakpoints given by the segmentations as inputs to calculate the probabilities of each segment belonging to a gain, normal or loss of copy number [37].

In Bluegnome arrays, the in-house Bluegnome software is used to call the single-cell copy number changes from Bluegnome data. Briefly, the log2Ratios are

smoothed and followed by a threshold test. The regions above 0.3 or below -0.3 are called gain or loss respectively (Blue Gnome Ltd., Mill Court, reat Shelford, Cambridge, UK). However, the details of the method are not public available.

In the Affymetrix SNP arrays, the commercially available package CNAT (Affymetrix, Santa Clara, CA, USA) is commonly used to call copy number alterations. Log2-Ratios are first smoothed followed by a Hidden Markov Model (HMM) algorithm to call the single-cell copy number changes. In HMM models, copy numbers are considered as hidden states with transition probabilities between them [38]. The limitation of this algorithm is that model results in overestimation of the number of segments.

Since SNP arrays provide SNP genotypes besides copy numbers, SNP arrays are also frequently used to deduce parental-of-origin for aberrations existing in single cells. They can indicate whether the aberration is inherited from the paternal or the maternal chromosome [28, 29, 39]. Johnson et al. (2010) developed a "parental support" approach to adjust single-cell array signals with the help of parental DNA signals [28]. Handyside et al. (2010) enabled the linkage-based inheritance analysis by using informative SNPs in which one parent is heterozygous and the other is homozygous [29]. Knowledge about the exact location of abbarent regions in the parental chromosomes is key for doing linkage based studies. Voet et al. (2011) identified the orgin of the aberrations using informative SNPs which show Mendelian errors [39].

2.1.4 Validation methods

Systematic validations are required to evaluate the protocol performance including amplification, array platform and copy number calling. Currently, the most popular and general validation method is based on genomic DNA derived from cell lines harboring known copy number alterations, for instance, using lymphoblastoid or fibroblast cell lines transformed by the Epstein-Barr virus (EBV) [22]. Genomic DNA isolated from multiple EBV cells carry known chromosomal imbalances. Single-cell DNA from the same EBV cell line is expected to have the same karyotype. Thus the karyotypes of the genomic EBV samples can be used as a reference to evaluate the accuracy of copy number alteration estimation applied to single EBV cells.

An alterative validation method is to perform FISH directly on single-cell DNA. The FISH results are used as a reference to evaluate the protocol performance on the other single cells extracted from the same cell line [40]. However, FISH can not enable testing genome-wide karyotypes since only pre-designed FISH probes can be attached to limited number of chromosomes. In addition, various studies confirm that mosaicisms are frequently observed in cleavage-stage embryos

[9, 40]. Thus the validated single cells probably show different karyotypes from the test ones. These validated karyotypes can not be used as a golden standard for the single-cell copy number alteration test [28].

Table 2.1: Overview of single-cell genome research using array platforms.

Array	**Platform**	**Avg probe length**	**WGA**[1]	**Analysis method**	**Res**[2] **(Mb)**	**Ref**
BAC array CGH	Custom 4K BAC	164kb	MDA-based GenomiPhi	Theshold-based: The mean of chromosome specific log2Ratio plus or minus 3 times the standard deviation	34	[33]
	Custom 32K	164kb	PCR-based Genome-plex	Smoothing + Threshold-based: at least 3 consecutive clones above or below 1.5 times experimental variability	8.3	[22]
	Custom 4K	164kb	MDA-based GenomiPhi	Channel-clone combined with CBS and CGHcall	10	[41]
	Custom 4K	164kb	MDA-based GenomiPhi	Mixture model combined with SNP array	10	[9]
	BlueGnome 3K 24 sure	Not reported	PCR-based Sure Plex	BlueFuse software: Smoothing combined with threshold-based	~2.5	[31]
Oligo array CGH	Agilent 44K	65bp	MDA-based GenomiPhi	Agilent software combined with filter for regions larger than 10Mb	Not reported	[23]
	Nimblegen 385K Nimblegen 2.1M Agilent 240K	60bp	PCR-based GenomePlex	Running means of intensity ratios in different window sizes combined with threshold-based	~3	[24]
	Agilent 244K	60bp	MDA-based GenomiPhi	Channel-clone combined with CBS and CGHcall	10	[25]
SNP array	Affymetrix 50K	25bp	MDA-based GenomiPhi V2	Affymetrix CNAT	Not reported	[26]
	Affymetrix 250K NspI	25bp	PCR-based GenomePlex	Affymetrix CNAT	Not reported	[27]

Continued

Array	Platform	Avg probe length	WGA[1]	Analysis method	Res[2] (Mb)	Ref
	Affymetrix 250K NspI	25bp	MDA-based GenomiPhi	Affymetrix CNAT combined with BAC arrays	10	[9, 35]
	Illumina Human CNV370 Infinium-II CytoSNP-12 chip	50bp	MDA-based PicoPlex	Parental support	Not reported	[28]
SNP array	Illumina Human CNV370 Infinium-II Quad and Duo	50bp	MDA-based Repli-G	Karyomapping: haplotyping using parental and sibling genotype information	Not reported	[29]
	Affymetrix 250K NspI	25bp	MDA-based GenomiPhi V2	Scoring parent-specific manner SNPs which show Mendelian errors	Not reported	[39]

2.2 Allelic-specific copy number analysis algorithms

The majority of copy number alteration algorithms report aberrant genomic regions with total copy number gains or losses [20, 42]. However, total copy number changes can not reveal some underlying genetic mechanisms, such as uniparental disomy or loss of heterozygousity (LOH) [43]. Fortunately, the availability of SNP arrays introduced in 2.1 enables the identification of copy number changes at allele level by taking allele imbalances into account.

Accurately identifying aberrant regions in cancer genomes is complicated because cancer genomes are often non-diploid [2]. The consequence is that the copy number gain or loss regions are wrongly called by simply comparing total copy numbers with ploidy two. In addition, tumor biopsies are always contaminated with non-tumor cells showing no aberrations. Thus, a good approach to call allelic-specific copy numbers should be able to account for ploidy as well as the proportion of non-tumor cells in the model [44].

In this section, we discuss 4 available allelic-specific copy number analysis approaches which simultaneously determine tumor ploidy, tumor purity and allelic-specific copy numbers: OncoSNP [45], GAP [46], ABSOLUTE [47] and

[1]Whole genome amplification
[2]The smallest length of real aberration detected by the array platform

ASCAT [48]. We will introduce ASCAT in more detail since it was used in Chapter 5.

Allelic-specific copy number analysis of tumors (ASCAT) is an algorithm to dissect copy numbers at allele level based on segmented SNP values [48]. First of all, ASCAT estimates the germline heterozygous SNPs from each tumor sample. Subsequently, Allele-Specific Piecewise Constant Fitting (ASPCF) is performed to simultaneously segment LRR and BAF values using germline heterozygous SNPs [49].

$$r_i = \gamma log_2 \left(\frac{2(1-\rho) + \rho(n_{A,i} + n_{B,i})}{2(1-\rho) + \rho\Psi_t} \right) \tag{2.1}$$

$$b_i = \frac{1 - \rho + \rho n_{B,i}}{2 - 2\rho + \rho(n_{A,i} + n_{B,i})} \tag{2.2}$$

$$\begin{aligned} d(\rho, \Psi_t) = \sum_{i=1} w_i \big(&(\hat{n}_{A,i}(\rho, \Psi_t) - round(\hat{n}_{A,i}(\rho, \Psi_t)))^2 \\ &+ (\hat{n}_{B,i}(\rho, \Psi_t) - round(\hat{n}_{B,i}(\rho, \Psi_t)))^2 \big) \end{aligned} \tag{2.3}$$

In the formulas (2.1)-(2.3), $\mathbf{i}$ represents the SNP index; $\mathbf{r_i}$ and $\mathbf{b_i}$ represent the segmented LRR and BAF values of SNP $\mathbf{i}$; $\boldsymbol{\gamma}$ is a constant depends on the platform; $\boldsymbol{\rho}$ is the tumor purity ranging from 0 to 1; $\mathbf{\Psi_t}$ represents the tumor ploidy, which is the average copy number across whole genome, ranging from 1.6 to 5.4; $\mathbf{A}$ and $\mathbf{B}$ represent A allele and B allele; $\mathbf{n_{A,i}}$ and $\mathbf{n_{B,i}}$ are the allelic-specific copy numbers for SNP $\mathbf{i}$; $\mathbf{\hat{n}_{A,i}}$ and $\mathbf{\hat{n}_{B,i}}$ represent the estimated values of allelic-specific copy numbers for SNP $\mathbf{i}$; **round**() function rounds to the nearest non-negative values; $\mathbf{w_i}$ equals to 1 if the BAF segmented value of the ith SNP not equals to 0.5 whereas equals to 0.05 if the BAF segmented value of the ith SNP equals to 0.5.

In formulas (2.1) and (2.2), using the segmented values $\mathbf{r_i}$ and $\mathbf{b_i}$ as inputs, given a tumor ploidy value $\mathbf{\Psi_t}$ and a tumor purity value $\boldsymbol{\rho}$, the estimated allelic-specific copy numbers $\mathbf{\hat{n}_{A,i}}$ and $\mathbf{\hat{n}_{B,i}}$ can be calculated for each SNP $\mathbf{i}$. Assuming that true copy numbers are non-negative rounded values, we search the values of $\mathbf{\Psi_t}$ and $\boldsymbol{\rho}$ such that the estimated allelic-specific copy numbers $\mathbf{\hat{n}_{A,i}}$ and $\mathbf{\hat{n}_{B,i}}$ across all the SNPs are as close as possible to non-negative rounded values. In other words, by trying all the possible combinations of ploidy and tumor purity values, we select the best combination such that $\mathbf{d}(\boldsymbol{\rho}, \mathbf{\Psi_t})$ in formula (2.3) has the minimal value. In this way, tumor ploidy $\mathbf{\Psi_t}$, purity $\boldsymbol{\rho}$, allelic-specific copy numbers $\mathbf{\hat{n}_{A,i}}$ and $\mathbf{\hat{n}_{B,i}}$ are the outputs of the ASCAT algorithm.

OncoSNP is based on hidden Markov models (HMM) whereas the other three approaches are based on numerical models using segmented LRR and BAF values as inputs. OncoSNP assumes that the observed LRR and BAF values follow a probabilistic function of 21 hidden states representing the 21 combinations of allelic-specific copy numbers in tumor cells and normal cells. It infers the tumor state as well as the proportion of contaminated normal cells for each SNP using the expectation-maximization (EM) algorithm. GAP, ABSOLUTE and ASCAT share the feature that they all attempt to search the best allelic-specific copy numbers for overall SNPs by checking all possible combinations of parameters. These three methods differ on which parameters to be inferred and which framework to be used for the optimization. GAP assigns allelic-specific copy numbers through pattern recognition on a bi-dimension segmented and smoothed LRR-BAF scale. ABSOLUTE optimizes the solution using a log-likelihood function given the segmented LRR, BAF values as well as a prior recurrent karyotype model. ASCAT models allelic-specific copy number as a function of segmented LRR, BAF, ploidy and tumor purity and selects the solution closest to non-negative integer copy numbers at all assayed loci in the genome.

Other allelic-specific copy number algorithms are also present. However, we did not discuss them here, either because they do not simultaneously account for ploidy and tumor purity or because they require extra matched normal samples. For instance, PSCN builds the model assuming a diploid cancer genome and requires SNP data from matched normal samples [50]. A comparison study has been made among OncoSNP, GAP, ASCAT and concluded that ploidy estimation based on overall segmented SNP data gives better results than HMM based methods. In addition, OncoSNP results in overestimation of the number of copy number alterations [44].



Chapter 3

Development of channel-clone normalization for single-cell copy number alteration detection

3.1 Summary

Array analysis of single cell genomes is a novel research area with important applications. However, not many computational approaches have been developed for single-cell array data due to experimental limitations. One challenge is the large amount of genome artifacts caused by the single-cell DNA amplification process. Although normalization methods exist to remove undesired biases in regular array data, none of them are suitable to reduce genome artifacts in single-cell array data. These methods lead to inaccurate detection of copy number alterations.

In this chapter, we propose a method based on channel, clone-specific and artifact corrections to preprocess signals from single-cell array data. We apply this technique to simulated as well as array CGH data designed for the discovery of submicroscopic chromosomal imbalances in developing embryos. Our processing strategy significantly improves the detection of single-cell copy number alterations in both simulated and real biological data.

3.2 Personal contributions

The PhD candidate explored different types of artifacts in single-cell array data and developed an appropriate signal processing technique for analyzing those data. The PhD candidate evaluated the signal processing technique using both simulated array CGH data and EBV-transformed single cells. The results proved that this processing technique efficiently corrected artifacts in the data and significantly improved the detection of copy number alterations. Furthermore, the PhD candidate proposed a complete pipeline for single-cell array data which begins with probe filtering, followed by signal processing, segmentation and copy number calling.

This paper was published in Genome Biology 2011, 12: R80.

Single-cell copy number variation detection

Jiqiu Cheng[1,2], Evelyne Vanneste[3], Peter Konings[1,2], Thierry Voet[3], Joris R Vermeesch[3], Yves Moreau[1,2,$]

[1]Department of Electrical Engineering, Esat-SCD, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium.
[2]IBBT-K.U.Leuven Future Health Department, Kasteelpark Arenberg 10, 3001, Leuven, Belgium.
[3]Center for Human Genetics, Katholieke Universiteit Leuven, Herestraat 49, 3000 Leuven, Belgium.
[$]Corresponding author; Tel: +32 16 328644; Fax: +32 16 321970

Abstract

Detection of chromosomal aberrations from a single cell by array comparative genomic hybridization (single-cell array CGH), instead of from a population of cells, is an emerging technique. However, such detection is challenging because of the genome artifacts and the DNA amplification process inherent to the single cell approach. Current normalization algorithms result in inaccurate aberration detection for single-cell data. We propose a normalization method based on channel, genome composition and recurrent genome artifact corrections. We demonstrate that the proposed channel clone normalization significantly improves the copy number variation detection both in simulated and real single-cell array CGH data.

Background

Array analysis of single-cell copy number variations (CNV) is a recently developed experimental technique for the detection of chromosomal rearrangements in single cells [9, 22, 29, 33]. A two-color single-cell array CGH assays the copy number difference between an euploid reference sample from genomic DNA and an unknown test sample from amplified single-cell DNA by comparing signal intensities using log2-ratios [51]. However, the accurate detection of single-cell CNV has been hampered by the noise levels in the log2-ratios caused by the amplification of the minute quantities of DNA present in a single cell. Moreover, since the reference DNA in single-cell array experiments is non-amplified genomic DNA extracted from a large number of cells [9], the biological nature of test and reference sample is different resulting in new genome artifacts [32]. Unfortunately, existing normalization strategies do not provide clear guidelines for checking for these artifacts, nor for handling them appropriately.

Among existing array CGH normalization methods, global loess normalization is commonly used [52]. Global loess normalization regresses the log2-ratios between test and reference sample on intensities using all probes [53]. The snapCGH package commonly used for analyzing array CGH data has included the global loess normalization method [54]. Furthermore, poplowess and CGHnormaliter have been developed for array CGH data [55, 56]. Poplowess attempts to separate normal from aberrant probes using k-means clustering and applies the loess normalization based on the largest group of probes, whereas CGHnormaliter combines a segmentation algorithm with loess normalization iteratively and normalizes data based on segmented normal probes. Although these two methods are supposed to help correctly recognize real chromosomal aberrations, they are not able to correct genome artifacts and could result in false calling of aberrations. Alternatively, the smoothing wave algorithm has been devised

to remove genome artifacts that are either related to the GC-content or other unknown factors [57]. However, this method requires calibrated genome profiles that are typically not available in the single-cell setup. Recently, more advanced algorithms have been proposed based on the combination of normalization, segmentation, and copy number calling [58, 59, 60, 61]. These algorithms allow the simultaneous normalization and segmentation. They are expected to jointly improve the CNV detection performance. However, these advanced algorithms are developed for the genomic array CGH data, but not for single-cell array CGH data, which has an additional property causing artifacts compared to genomic data. All of these normalization methods have in common that they normalize data on the ratio of both channels without taking the single-cell amplification bias and genome artifacts into account.

In this paper, we present a new normalization approach based on channel and clone-specific artifact corrections, named channel clone normalization, to remove the amplification bias caused by the different nature between test and reference sample. Moreover, this approach removes genome artifacts that obscure the detection of real aberrations. The explorations of the amplification bias and genome artifacts are shown in the Results section. Furthermore, we compare our newly developed method to several existing normalization methods: global loess, poplowess, and CGHnormaliter; as well as to the methods combining normalization and segmentation: Haarseg, GADA, and CBS combined normalization [58, 60, 61]. The significant performance improvement of our channel specific normalization method is shown both on simulated and real single-cell array CGH data.

Results

Simulation of single-cell data

To quantify the effect of the channel clone normalization, we simulated 15 samples including 23 artificial aberrations based on 7 real EBV samples as described in the Application section. The simulation details are presented in the Materials and Methods section. This simulation data set is comparable to real genome profile features of the single-cell array CGH data with known artificial aberrations. The overall performance of all normalization methods on the simulation data set is demonstrated in Figure 3.1. The True Positive Rates (TPR) using global loess, CGHnormaliter, poplowess, and channel clone normalization are 0.97, 0.94, 0.92, and 0.96, whereas the False Positive Rates (FPR) are 0.06, 0.08, 0.08 and 0 respectively. Although channel clone normalization missed one out of 23 known aberrations, it offers the best performance in comparison to the other normalization methods with the fewest falsely discovered CNV regions and comparable TPR. Global loess,

CGHnormaliter, and poplowess show similar CNV detection performance in terms of TPR and FPR.

An example shown in Figure 3.2 illustrates the correction of genome artifacts by the channel clone normalization. Chromosome 10 of Sample 4 contains a confirmed duplication on the q-arm. This duplication was correctly detected by all four normalization methods. However, the Chromosome 10 q-terminal region was incorrectly detected as a deletion using global loess, CGHnormaliter, and poplowess. In contrast, this genome artifact was corrected by the channel clone normalization method and detected as a normal region.

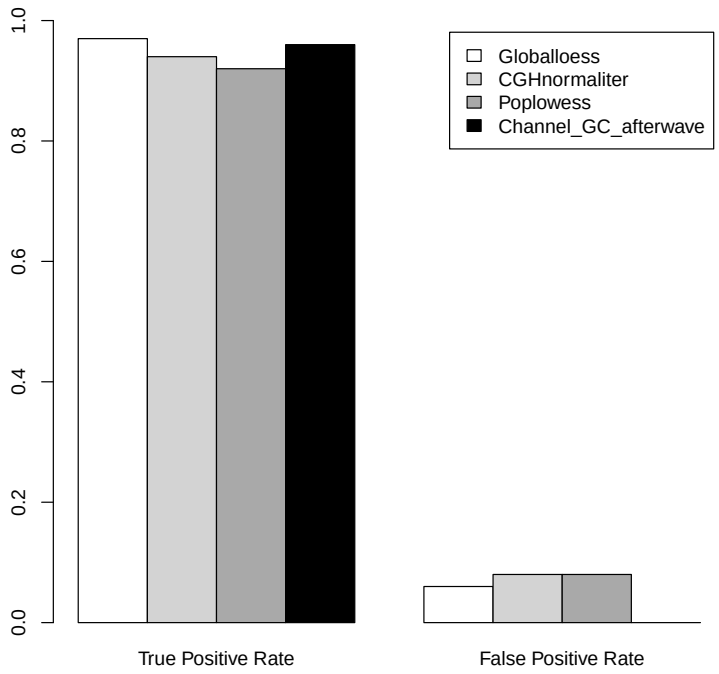

Figure 3.1: Barplot of true positive rate and false positive rate of 15 simulated samples. All the true positive rates (TPRs) and false positive rates (FPRs) were calculated after the global loess, CGHnormaliter, poplowess or channel clone normalization methods.

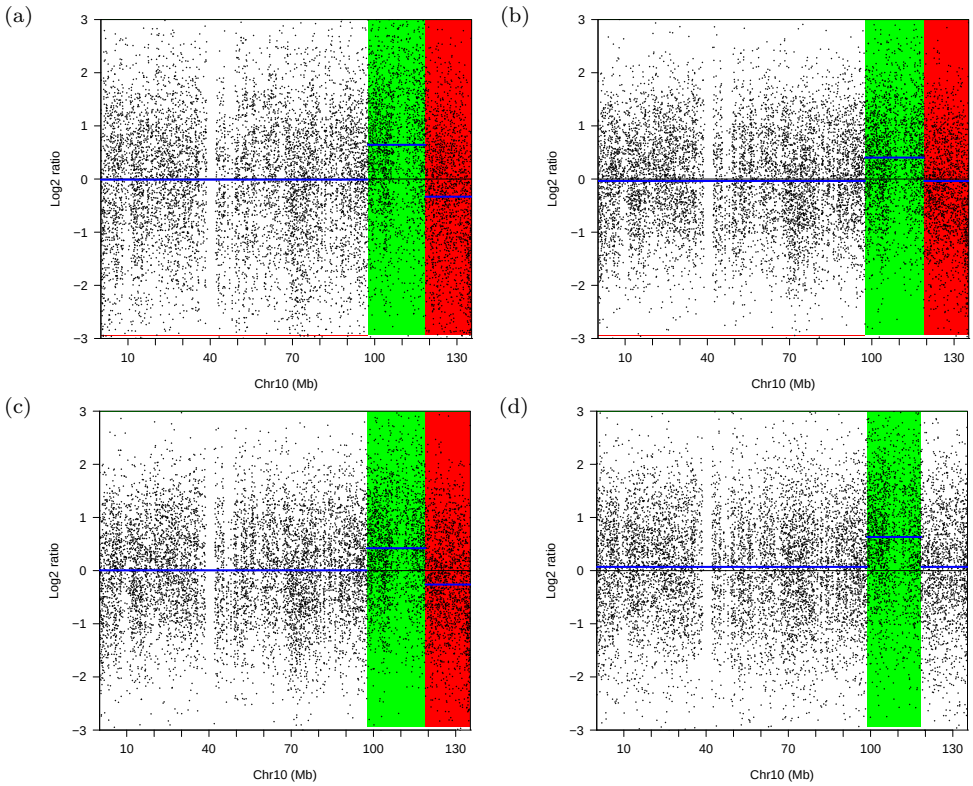

Figure 3.2: Copy number variation detection in Chromosome 10 of a simulated sample. (a-d) CBS segmentation of chromosome 10 from the simulated sample 4 using global loess normalization (a), CGHnormaliter (b), poplowess (c), and channel clone normalization (d). The blue line represents the CBS segmentation line. The red region and green regions represent the deletion and duplication regions called by CGHcall.

Application 1: single EBV transformed lymphoblastoid cell array CGH

We analyzed seven single EBV-transformed lymphoblastoid cells amplified according to the previously described protocol [9]. Each of these amplified single-cell DNA samples was hybridized as a test sample on Agilent 244K arrays against genomic non-amplified DNA derived from a patient with Klinefelter Syndrome (47, XXY). The aberration and diploid regions have been validated by the corresponding genomic DNA using a 250K Affymetrix SNP array with the help of SNP copy number, loss-of-heterozygosity (LOH), and heterozygous

SNPs. The karyotype of each EBV-transformed sample is shown in Table 3.1. We used this data set to quantify our approach and benchmark our data with other methods.

Our normalization approach mainly consists of three steps: channel standardization, genome composition artifacts correction and recurrent genome artifacts correction. All of these three steps are necessary to improve single-cell CNV detection. The investigation of the single-cell amplification bias is covered in the 'Exploration of the amplification bias' section and the exploration of genome artifacts is covered in the "Detection of copy number variation" section.

Exploration of the amplification bias

We first explored the amplification bias caused by the different natures of the test and reference samples with the help of graphical plots. MA, density, and Quantile-Quantile (QQ) plots are used to check for potential artifacts before and after normalization. The y-axis and x-axis of a MA-plot represent the log2-ratios and average log2-intensities between two hybridized samples, respectively. The points of a MA plot should be randomly located around zero in the y-axis if no large aberrations or artifacts exist in the data. The density plot and QQ plot are graphical techniques to show the similarity between intensity distributions from test and reference samples. If the test sample intensities are distributed similarly to reference intensities, the density plot of two hybridized samples should overlap and the QQ plot should be located along the 45-degree line.

An obvious intensity-dependent pattern is observed in the MA plot of all single-cell array CGH experiments (Figure 3.3a; Additional file 1). The pattern visualized using the red lowess smoothing line shows that the log2-ratio increases linearly with the increase of the average intensities in the single-cell array CGH data. In contrast, the MA plot of an array CGH experiment using non-amplified genomic DNA shows no aberrant pattern (Figure 3.3b). Since both array CGH experiments are performed using the same series of Agilent 244K arrays and the only difference between them was the processing of the test samples, we suspect that the intensity-dependent pattern artifact is caused by the amplification of the single-cell DNA. This suspicion is confirmed by the larger standard deviation (SD) of the intensities in the amplified test sample compared to the non-amplified reference sample (Figure 3.4a). Consequently, the median SD of single-cell array CGH log2-ratios is 1.38 with a range from 0.85 to 1.44 across 7 arrays, whereas that of the genomic array CGH experiments is 0.28, ranging from 0.2 to 0.35 across 6 arrays. This larger SD of log2-ratios in the single-cell array CGH experiments hampers the accurate detection of CNVs at the single-cell level. It is thus necessary to remove this amplification bias.

After the data are normalized by the channel standardization step, the pattern between averaged intensity and log2-ratio disappears and the lowess curve fitted

to the data is close to horizontal (Figure 3.3c). The intensity distributions of the reference and test samples are adjusted to have approximate mean zero and SD equal to 1 (Figure 3.4b). The QQ plot in Figure 3.4d shows that most points after the channel-clone normalization are located around the 45-degree reference line, meaning that the intensities of normalized test and reference samples follow similar distributions. We conclude that the amplification bias has been successfully removed by the channel standardization step.

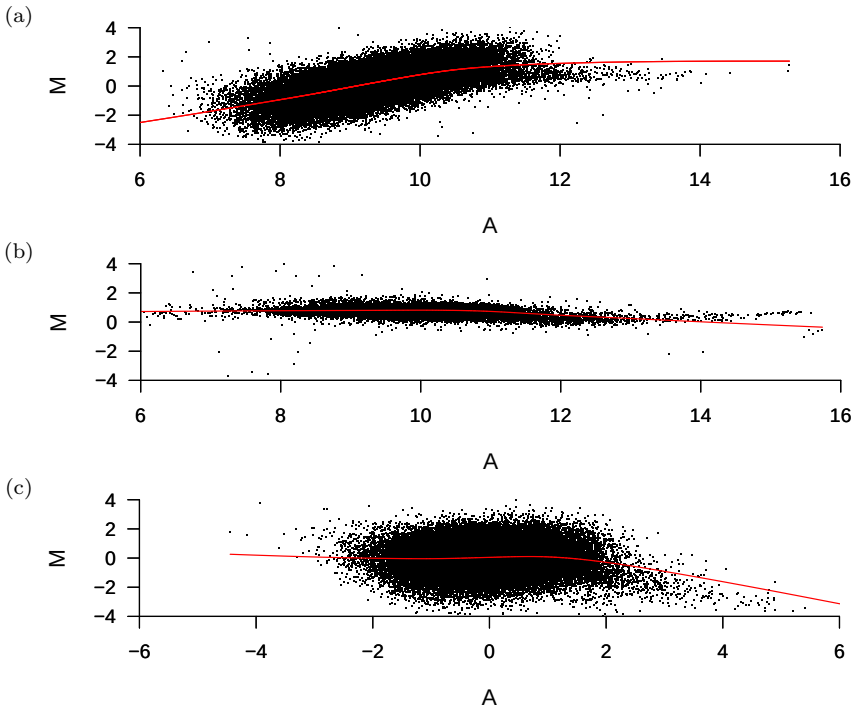

Figure 3.3: MA plot of a single EBV-transformed cell. (a)-(c) MA plot for EBV-transformed single lymphoblastoid cell 1162 before normalization (a), genomic DNA before normalization (b), EBV-transformed single lymphoblastoid 1162 after channel standardization (c). The red line represents a lowess curve fitted to the data. Note that after normalizations, the majority of the log2-ratio values are distributed randomly around zero.

Detection of copy number variation

After the exploration of the amplification bias, we checked the impact of genome composition artifacts and recurrent genome artifacts on the performance of single-cell CNV detection using the CBS algorithm [36]. Genome composition artifacts, appearing as incorrect chromosomal aberrations, are frequently observed in

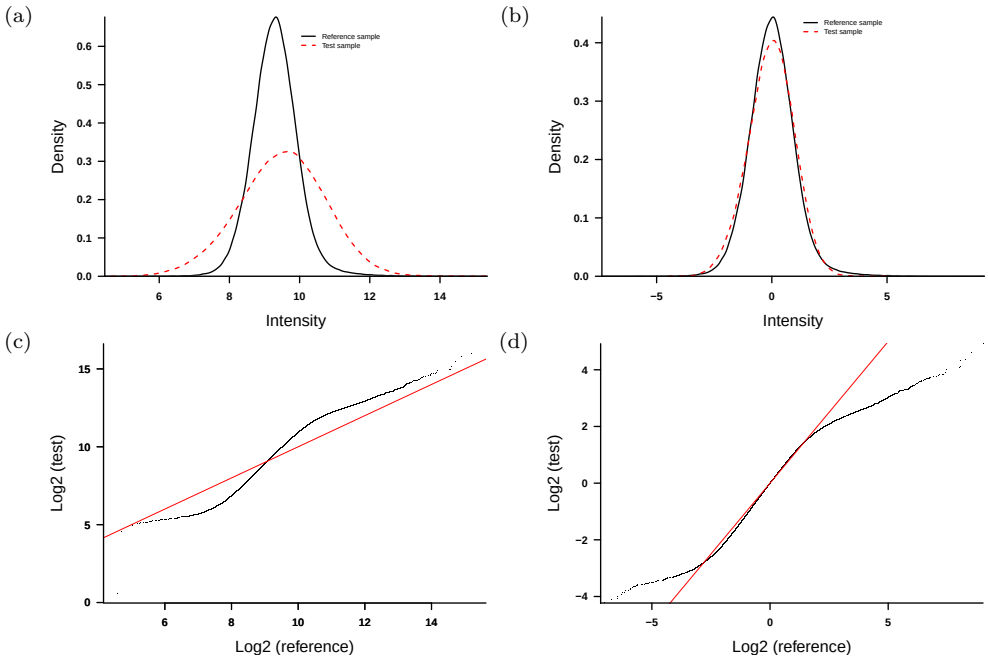

Figure 3.4: Density plot of a single EBV-transformed cell. (a,b) Density plot for EBV-transformed single lymphoblastoid cell 1162 before normalization (a), and after channel standardization (b). The solid line represents the reference sample and the dashed line represents the test sample. Note that the SD of the intensities of the test sample (SD=1.02) is larger than that of the reference sample (SD=0.61). (c,d) QQ plot of the intensities between the test and the reference samples before normalization (c), and after channel standardization (d).

the array CGH data. These artifacts are illustrated in Figure S2 (A,B) of Additional file 2 with the low log2 ratios of the chromosome 1 p terminus and the chromosome 10 q terminus. Studies have shown that these genome composition artifacts could be caused by GC-content as well as other unknown factors [62].

We therefore use a genome composition correction step to correct the artifacts caused by GC-content and a recurrent artifact correction step to correct unknown recurrent artifacts. For the genome composition correction step, we considered two possible methods: correction based on the GC-content of (1) the probe sequence itself or (2) an enlarged window around the probe. Similarly, for the recurrent genome artifact correction we also considered two methods: (1) CBS

segmented residuals followed by the recurrent genome artifact correction and (2) an artifact correction without the CBS segmentation in advance. The details of the channel clone normalization are introduced in the Material and Methods section. We compare our channel clone approach with four sub-methods to show that the combination of channel standardization, genome composition artifact correction and recurrent genome artifact correction together give the best single-cell CNV detection performance: CG (channel plus genome composition correction using enlarged window GC contents); CA (channel plus recurrent genome artifact correction without CBS segmentation); CGprobeA (channel plus genome composition correction using probe GC contents plus recurrent genome artifact correction without CBS segmentation); CGACBS (channel plus genome composition correction using enlarged window GC contents plus CBS segmented residuals followed by recurrent genome artifact correction); channel clone (channel plus genome composition correction using enlarged window GC contents plus recurrent genome artifact correction without CBS segmentation).

The genome profiles before and after genome composition correction are shown in Additional file 2. It is obvious that the GC-content-related artifacts, appearing as a wave pattern in Figure S2 (A,B) in Additional file 2 are adjusted after the genome composition correction shown in Figure S2 (C,D). Similarly, Figure 3.5 shows that the CNV detection performance of CG with a TPR 0.86 and FPR 0.06 respectively, is better than for the methods that do not account for genome composition correction (for example, global loess, CGHnormaliter, and poplowess).

Different studies have used genome composition corrections to correct the genome wave pattern [62]. Array CGH hybridization is influenced not only by the GC content of the probe sequence but also the DNA sequences that lie in an enlarged window around the probe sequence corresponding to a DNA sequence fragment the probe hybridizes to. Diskin et al. [63] used an ordinary linear regression model to regress the Log2Ratio on the GC content of a fixed 1Mb window size around the probe to correct the genome composition artifacts. Since this method was developed for single-channel arrays and cannot be directly implemented for the two-color arrays, we develop a comparable, but more elaborate genome composition correction approach. To account for the GC content of the unknown genome fragments, our method extracts the GC percentage from different window sizes around each probe and elects the window size with the highest correlated GC content to the log2Ratio for the genome composition correction. Secondly, in contrast with Diskin et al.'s method, we use a weighted linear regression model with larger weights for the GC rich probes to avoid the overcorrection of real chromosomal aberrations. Other genome correction methods could also be valid. However, comparison of all GC correction methods is out of the scope of our study. To show that accounting for

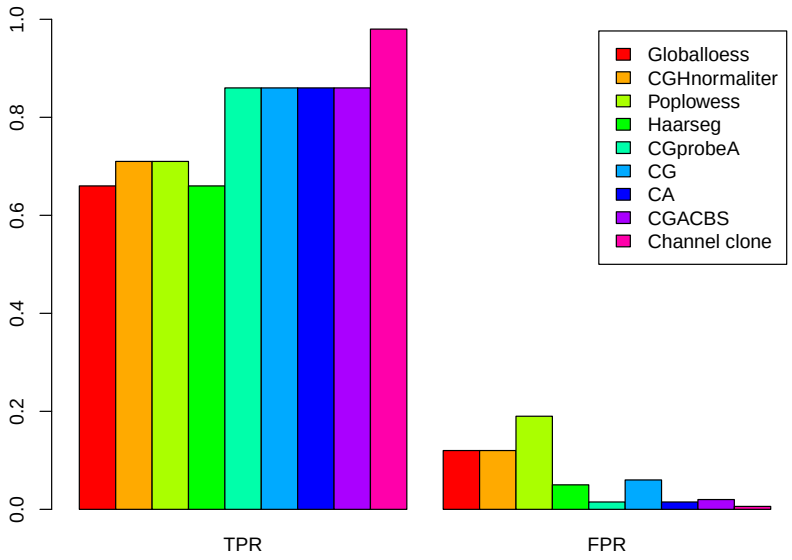

Figure 3.5: Barplot of true positvite rate and false positive rate of 7 EBV-transformed cells. All the TPRs and FPRs were calculated after global loess, CGHnormaliter, poplowess, Haarseg, CG, CA, CGprobeA, CGACBS and channel clone normalization approaches.

the GC content from enlarged window sizes improves the genome composition correction, we also performed the correction based only on the GC content of each probe, as proposed by the CGprobeA normalization. Figure 3.5 shows that the TPR and FPR values are 0.86 and 0.015, respectively, for the CGprobeA normalization method, whereas the values for our channel normalization are 0.98 and 0.006, respectively. This comparison confirms the importance of finding the optimal GC-content window for the genome composition correction.

The impact of the recurrent genome artifact correction of each chromosome is especially explained in Additional file 3 and shown in Additional files 4 to 10. For instance, chromosome 3 of EBV-transformed cell 1168 was experimentally confirmed to have no aberrations. However, two deletions at the location around 50Mb and the q-arm terminal region were observed when no correction was applied (Figure S3 (A) in Additional file 3). The estimated common profile of chromosome 3 (Figure S3 (B) in Additional file 3) shows the artifacts at the same locations as the individual profile of EBV-transformed cell 1168. Since the common profile is estimated across all the EBV-transformed samples, the

artifacts observed in the common profile represent the recurrent genome artifacts existing in multiple EBV-transformed samples. Figure S3 (C) in Additional file 3 shows that after the extraction of the estimated common profile, these two artifacts have been removed and the segmentation line of this chromosomal profile is horizontal around the zero line.

Comparison of the CG and CA methods to channel clone normalization is shown in Figure 3.5 and Table 3.1. Both the CG and CA normalization methods show lower TPRs and larger FPRs for single-cell CNV detection performance. These results confirm our hypothesis that not all genome artifacts can be explained by GC content. Our channel clone normalization method removes genome composition artifacts, as well as unknown recurrent genome artifacts. Therefore, the combination of channel standardization, genome composition and recurrent genome artifact corrections, which we propose, gives the best single-cell CNV detection performance with TPR 0.98 and FPR 0.006.

A recent study suggests that the combination of segmentation with recurrent genome artifact correction can improve aberration detection in genomic array CGH applications [61]. We have tested this CGACBS approach on our single-cell array CGH data. Table 3.1 shows that the TPR and FPR of CGACBS are 0.86 and 0.02, respectively, which is outperformed by channel clone normalization, with values of 0.98 and 0.006, respectively. CGACBS uses CBS segmented residuals for the genome artifact correction to avoid overcorrection of real chromosomal aberrations. However, this method also protects genome artifacts with log2-ratios comparable to real aberrations from being corrected. Consequently, it results in higher false positive calling of aberrations. Therefore, it is a trade-off between keeping real aberration signals and removing undesired genome artifacts.

Moreover, we have compared our normalization approach to global loess, CGHnormaliter, poplowess, Haarseg, and GADA methods. Using the TPR and FPR as given in Figure 3.5 and Table 3.1, we compared the overall CNV detection performance for global loess, CGHnormaliter, poplowess, Haarseg, and channel clone normalization. The TPR values were 0.66, 0.71, 0.71, 0.66, and 0.98 respectively, while the FPRs were 0.13, 0.09, 0.15, 0.05, and 0.006, respectively. Although the recently developed poplowess and CGHnormaliter normalization methods perform better than the original global loess normalization, they have a high FPR as well. The common feature of both methods is the separation of probes with normal log2 ratios from probes with aberrant log2 ratios, as well as the normalization of the data based on the normal probe log2 ratios; however, this is not suitable in single-cell array CGH. The reason is that many genome artifacts appear next to real aberrations caused by amplification bias in the single-cell approach. As a consequence, these genome artifacts are incorrectly segmented or clustered by the CGHnormaliter or poplowess algorithms into

aberrant groups, yielding poor results.

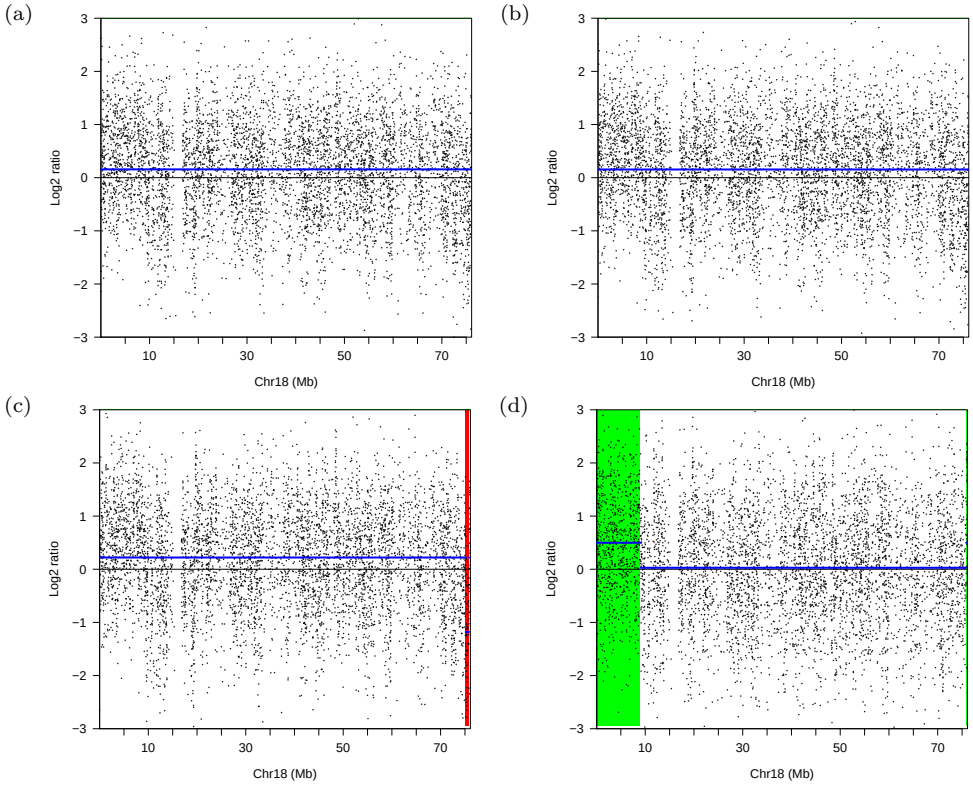

Figure 3.6: Copy number variation detection in chromosome 18 of an EBV-transformed sample. (a-d) CBS segmentation of chromosome 18 from the EBV-transformed single lymphoblastoid cell 1151 using global loess normalization (a), CGHnormaliter (b), poplowess (c), and channel clone normalization (d). The y-axis represents the log2 ratios and the x-axis represents the coordinates along the chromosome. The blue line represents the CBS segmentation line. The green region represents the duplication region called by the CGHcall program.

The channel clone normalization method has shown its advantage in correcting recurrent genome artifacts across samples. Notice that CBS fails to detect a 2.22Mb deletion at the chromosome 20 p terminus of cell 1151 after channel clone normalization (Figure S12 in Additional file 12). The possible reason is that this deletion is located at the terminal region of a chromosome with a short length of 2.22Mb. This aberration thus shows a pattern similar to the artifacts located at the same position and results in an overcorrection by the

channel clone normalization. However, considering the large FPR caused by chromosomal artifacts from the single-cell array CGH, it is worthwhile to reduce FPR from around 10% to 0.6%, even while missing one short aberration.

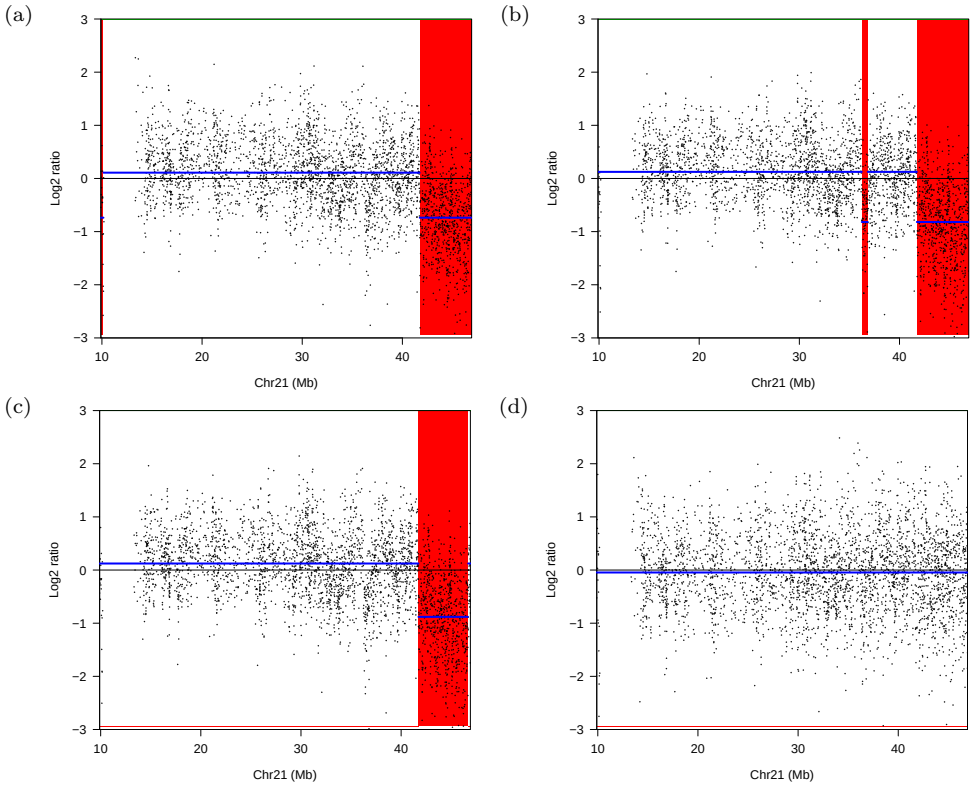

Figure 3.7: Copy number variation detection of chromosome 21 of an EBV-transformed sample. (a-d) CBS segmentation of Chromosome 21 from the EBV-transformed single lymphoblastoid cell 1160 using global loess normalization (a), CGHnormaliter (b), poplowess (c), and channel clone normalization (d). The blue line represents the CBS segmentation line. The red region represents the deletion region called by CGHcall.

The performance of global loess, CGHnormaliter, poplowess, Haarseg and channel clone normalization on each genome profile is shown in Figures 3.6 and 3.7 and Additional files 4 to17. For instance, cell 1151 carries a known terminal 9.3 Mb duplication at the chromosome 18 p terminus (Figure 3.6). This duplication is called after channel clone normalization, but not after the other loess-based methods. Figure 7 illustrates that chromosome 21 of cell 1160

is expected to have no aberration. This is confirmed by SNP-array analysis that revealed no loss-of-heterozygosity for this 21q-ter segment. However, the q-terminal region of this chromosome is detected as a deletion after global loess, CGHnormaliter or poplowess normalizations, thus resulting in a false-positive CNV region.

Haarseg is an algorithm integrating signal smoothing, normalization, segmentation, and copy number calling [58]. However, this algorithm performs somewhat conservatively in calling chromosomal aberrations in the single-cell array CGH data, even though it gives a lower FPR than loess-based normalization methods. We also checked the performance of GADA in the single-cell application. GADA is an iterative procedure with the combining normalization and segmentation by sparse Bayesian learning. Around 800 breakpoints were detected in each EBV-transformed sample by GADA (Additional file 18). This is biologically unrealistic, and we conclude that many false positive aberrations have been detected. Although Haarseg and GADA are suitable in genomic array CGH data [58, 60], the implementation of these methods becomes inappropriate for single-cell array CGH data. The channel clone method outperforms these methods, having the largest TPR (0.98) and smallest FPR (0.006). Clearly, channel clone normalization improves the TPR considerably compared to these other normalization algorithms or normalization integrated algorithms for the single cell array CGH.

Recently, a unified model has been developed by the simultaneous integration of normalization, segmentation and copy number calling [61]. This model has been shown to be efficient for the genomic array CGH data. The advantage of this model is that it can incorporate existing preprocessing methods into one model. It would be attractive to enrich this model by accounting for single-cell data properties for the single-cell CNV detection in the near future.

Table 3.1: True positive rate of each EBV cell followed by different normalizations.

Real aberrations[1]	**Global loess**	**CGH normaliter**	**Pop lowess**	**Haar seg**	**CG probeA**	**CG**	**CA**	**CGA CBS**	**Channel clone**
Cell 617, 14M, 18p ter del	0	13.62	13.72	13.56	14.18	14.86	12.05	14.18	11.99
Cell 1151, 9.3M, 18p, dup	0	0	0	9.04	8.87	6.21	9.06	8.92	8.87
Cell 1151, 1.7M, 20p ter del	1.70	1.70	1.70	0	1.70	1.70	0	1.70	0
Cell 1151, Monosomy X	151.87	151.87	151.87	151.59	151.87	151.87	151.87	151.87	151.87
Cell 1160, 3M, 11 qter del	2.22	1.73	2.56	0	2.20	2.22	0	2.22	2.56
Cell 1162, 47.5M, 14q dup	0	0	0	40.14	31.97	45.78	47.39	39.47	47.39
Cell 1162, 58M, chr X, del	59.94	59.94	59.94	0	59.93	59.93	57.30	59.92	57.30
Cell 1168, trisomy 21	0	0	0	0	0	0	0	0	36.99
TPR	0.66	0.71	0.71	0.66	0.83	0.86	0.86	0.86	0.98

Application 2: human embryo array CGH

In reality, the assumption that only few probes display an aneuploidy copy number and most probes display diploid copy numbers does not hold generally (for example, consider heavily rearranged blastomeres, tumor cells, and so on). It is important, therefore, to test whether channel clone normalization would overcorrect the signals of heavily aberrant samples. We applied the channel clone normalization approach to array CGH of 14 blastomeres from previous published work [9]. All the blastomeres extracted from human embryo 20 carry multiple aberrations. The confirmed karyotype of each blastomere has been described in the previous published paper.

The results show that many artifacts are observed in the genome profile before channel clone normalization (Figure 3.8a,c,e). These artifacts were removed after channel clone normalization and none of the real chromosomal aberrations were over-corrected (Figure 3.8b,d,f). For instance, blastomere A carries aberrations in chromosomes 1, 10, 11, 13, 18, 22 and X, blastomere E carries aberrations in chromosomes 1, 2, 4, 7, 10, 11, and 22 and blastomere G carries aberrations in chromosomes 1, 4, 10, 22 and 23. Figures 3.8b,d,f show that all of these aberrations were detected after the channel clone normalization. Thus, channel clone normalization appears valid for heavily aberrant samples as well.

[1]Real aberrations are validated by Affymetrix 250K array using genomic DNA.The first column represents the true validated aberrations of each EBV cell, followed by the detected aberration length after global loess, CGHnormaliter, poplowess, Haarseg, CGprobeA, CG, CA, CGACBS and channel clone normalization methods. The column with bold numbers shows the detected aberration length and true positive rate after the channel clone normalization. CA, channel standardization followed by recurrent genome artifact correction without CBS segmentation; CG, channel standardization followed by genome composition correction using enlarged window GC contents; CGACBS, channel standardization followed by genome composition correction using enlarged window GC contents and recurrent genome artifact correction with CBS segmentation; CGprobeA, channel standardization followed by genome composition correction using probe GC contents and recurrent genome artifact correction without CBS segmentation; Channel clone, channel standardization followed by genome composition correction using enlarged window GC contents and recurrent genome artifact correction without CBS segmentation.

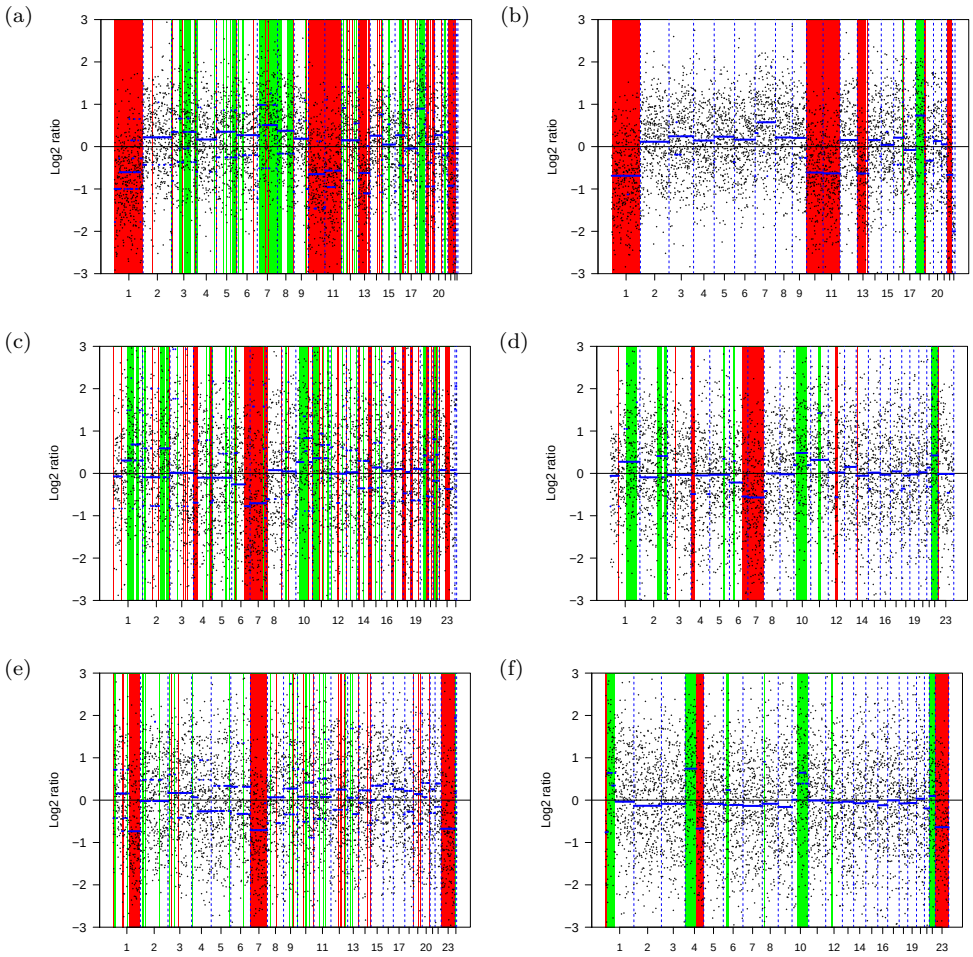

Figure 3.8: Copy number variation detection of 3 blastomere samples. (a,c,e) Genome-wide CNV detection of blastomere A (a), blastomere E (c) and blastomere G (e) from embryo 20 before channel clone normalization. (b,d,f) Genome-wide CNV detection of blastomere A (b), blastomere E (d) and blastomere G (f) from embryo 20 after channel clone normalization. The x-axis represents the coordinate range from chromosome 1 to X and y-axis represents the log2 ratios. The blue line represents the CBS segmentation line. The green regions represent the duplication region and red regions represent the deletion region called by the CGHcall program.

Discussion

The analysis of CNV in single cells using high-density arrays is a novel attractive research technique [24, 26, 27, 64]. It enables genome-wide analyses of blastomeres during early embryogenesis, cell development, and cancer progression [9]. Because the amount of DNA that can be derived from single cells is limited, amplification is necessary. However, amplifying only the test sample results in an amplification bias as well as serious genome artifacts with respect to the log-intensity ratios and leads to poor CNV detection in single-cell array CGH data. So far, no standard procedures have been established to correct this amplification bias and genome artifacts for single cell array CGH. We present a channel clone normalization method that addresses this issue.

The main need for a specific normalization method for single-cell array CGH as opposed to standard genomic array CGH, arises from the fact that the amplification step in the protocol for single-cell array CGH introduces a key difference compared to array CGH using DNA extracted from a large number of cells. Indeed, only the test sample undergoes DNA amplification while the reference sample remains a DNA sample extracted from a large number of cells with normal wild-type karyotype. This introduces a major bias in the distribution of signals between the test (amplified single-cell DNA) and the reference (non-amplified DNA) samples and genome artefacts, which our method aims to correct. Amplification of the reference sample from a single wild-type cell would be difficult because using amplified single-cell reference samples is unlikely to cancel out the biases caused by amplification since the amplification bias appears to be variable between samples in practice.

Our normalization approach is based on standardization of the distributions of the intensities of test and reference samples, genome composition artifact correction and recurrent genome artifact correction across all the samples. We have shown that our channel clone normalization method clearly improves the performance of single-cell CNV detection compared to other normalization methods, as well as the combined normalization segmentation methods, without losing the ability to detect real aberrations.

Conclusions

We have proposed a normalization strategy to handle interchannel variation and genome artifacts in two-color arrays and evaluated its applicability using simulated data and data from real single-cell array CGH experiments. Our method was designed originally for single-cell array CGH experiments, but it can be extended to other two-color array experiments that suffer from interchannel

variation or genome artifacts. Our approach has the following advantages: first, it achieves good performance for the detection of genomic signals; second, it does not require complex experimental designs, which make the experiments less expensive; and finally, it can be easily implemented without requiring long computing times.

Materials and Methods

Channel clone normalization
The pre-processing consists of four steps. Step 1, filter of clones: the internal control, incorrectly annotated and low foreground-intensity clones are filtered out. Step 2, channel standardization: the log2-transformed intensity of test sample and reference sample are standardized based on the trimmed mean and standardized deviation. Step 3, genome composition artifacts correction: log2 ratios are subjected to weighted linear regression on the highest correlated GC content, with larger weights for the GC-rich clones. Step 4, recurrent genome artifacts correction: a profile is generated using the trimmed mean of log2 ratios for each probe across all the samples. Subsequently, the common profile trend is estimated by applying a spline model to the generated profile. Finally, the estimated common profile trend is subtracted from each individual genome profile.

The channel clone normalization approach was implemented in R 2.12.1 and the code is available in the Additional file 19 [65]. The last three steps (channel standardization, genome composition correction and recurrent genome artifact correction) are the core steps of our approach. The impact of each normalization step is discussed in the Results section and the details of each step are explained below.

Filtering of clones
First, internal control and clones with incomplete physical annotations are removed. Second, the median background intensities of each array across all the spots are calculated. Subsequently, clones with intensities more than five-fold smaller than the median background intensities as a threshold are filtered out [9]. The threshold is chosen with the help of the MA plot of raw intensities excluding internal control and incomplete physical annotated clones. For instance, Additional file 1 shows the MA plot of the raw intensity of EBV-transformed cell 1160 with the red spots corresponding to the clones with intensities more than five-fold smaller than the median background intensity of this array. These low intensity clones show higher variability than the other clones [66] and are thus excluded.

Channel standardization
The log2-transformed intensity of test sample and reference sample are standardized based on the trimmed mean and standard deviation:

$$\mathrm{Test}_{\mathrm{ij_standardize}} = \frac{\mathrm{Test}_{\mathrm{ij}} - \mathrm{trimmedmean}(\mathrm{Test}_{\mathrm{j}})}{\mathrm{sd}(\mathrm{Test}_{\mathrm{j}})}$$

$$\mathrm{Ref}_{\mathrm{ij_standardize}} = \frac{\mathrm{Ref}_{\mathrm{ij}} - \mathrm{trimmedmean}(\mathrm{Ref}_{\mathrm{j}})}{\mathrm{sd}(\mathrm{Ref}_{\mathrm{j}})}$$

where $\mathrm{Test}_{\mathrm{ij}}$ represents the log2-transformed intensities of the ith probe of the jth array derived from a test sample; $\mathrm{trimmedmean}(\mathrm{Test}_{\mathrm{j}})$ represents the trimmed mean of the log2-transformed intensities of the jth array derived from the test sample; $\mathrm{sd}(\mathrm{Test}_{\mathrm{j}})$ represents the standard deviation of the log2-transformed intensities of the jth array of the test sample; $\mathrm{Test}_{\mathrm{ij_standardize}}$ represents the standardized intensities of the test sample. The parameters to calculate the standardized intensities for the reference sample $\mathrm{Ref}_{\mathrm{ij_standardize}}$ are defined in a similar way as the test sample $\mathrm{Test}_{\mathrm{ij_standardize}}$.

In this step, the amplification bias is expected to be removed by adjusting most of the intensities of the reference and test samples to follow similar distributions without reducing the correlation between them. To make the normalization robust to outliers, the trimmed mean instead of the global mean is calculated. The difference between the mean and trimmed mean is that the mean is calculated using all of the observations whereas the trimmed mean is based on observations excluding a percentage threshold of extreme observations. The trimmed mean is thus less influenced by extreme values than the mean and more robust to outliers. The percentage threshold is determined from a Quantile-Quantile (QQ) plot between the intensities of the test and reference samples. For instance, the QQ plot in Figure 3.4 shows that, in our case, approximately 20% of the points have extreme values, located at the two ends of the plot.

Genome composition artifact correction
This step aims to correct for genome composition related artifacts by the weighted linear regression of log2 ratios on the GC content of an enlarged window around probes. The model is stated as follows:

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

To estimate the parameter β, the following expression needs to be minimized:

$$L_w(\beta) = \sum_{i=1}^{n} w_i(Y_i - \beta_0 - \beta_1 x_i - \beta_2 x_i^2)$$

where $\mathbf{w_i}$ is 1000 if $\mathbf{x_i}$>0.5 and 0.01 in other cases; $\mathbf{Y_i}$ represents the log2 ratio of probe i obtained from the 'channel standardization' step; $\mathbf{x_i}$ represents the GC content of a certain window size around each probe.

GC contents of different window sizes around probes ranging from 0 to 1 Mb are extracted from the human genome sequence. Next, the correlation between GC content with different window size and log2 ratio is calculated. The window size with the highest correlation is selected to fit the model. Thirdly, a large weight (1000) was assigned to the clones with large GC content whereas a small weight (0.01) was assigned to the clones with low GC content. The residual of the model is the log2-ratio after the genome composition correction.

Recurrent genome artifact correction
This step is corrects recurrent genome artifacts. The recurrent genome artifacts are expected to be represented by the estimated common profile across all the samples. Therefore, a common profile is generated by calculating a trimmed mean of log2 ratios for each probe across all the samples. The common profile trend is estimated using a spline smoothing function [67]:

$$S(g) = \sum_{i=1}^{n} \{Y_i - g(x_i)\}^2 + \alpha \int_{x_1}^{x_n} \{g''(x)\}^2 d(x)$$

Where $\mathbf{i}$ represents the i-th probe of an array; $\mathbf{x_i}$ represents the genome physical position of probe i; $\mathbf{Y_i}$ represents the log2ratio of probe i after genome composition correction ; $\mathbf{n}$ represents the number of knots; $\mathbf{g}$ represents a twice-differentiable function; $\mathbf{g(t_i)}$ represents the estimated smoothing value of log2 ratio from probe i; α represents a smoothing parameter balancing the model fitting and the model complexity. The maximum of knots which are equal to the total number of probes were used to fit the model.

The smoothing Spline estimation of $\mathbf{g(t_i)}$ is the minimizer of $\mathbf{s(g)}$. It represents the common profile trend across all the samples including recurrent genome artifacts. Thus the subtraction of the estimated common profile trend from each individual genome profile can remove the recurrent genome artifacts.

Simulation
A simulation data set was generated based on seven real EBV-transformed samples (described below). Firstly, the EBV data set was processed by setting

the true aberrant intensities as empty values. Subsequently, 15 sample intensities were simulated by replacing individual probe intensities from corresponding processed EBV probe intensities. These two steps ensure that all the simulated intensities are non-aberrant. In addition, the simulated genome profiles represent the real single-cell genome profile features, including recurrent genome artifacts. Thirdly, 23 aberrations were artificially added to the simulated data, with the mean intensities of the simulated aberrations setting as the ones of the true aberrant regions from the real EBV-transformed samples. The length of the aberrations was set to around 20 Mb.

Single EBV-transformed lymphoblastoid cell array CGH
Seven EBV-transformed cells derived from patients carrying known unbalanced chromosomal rearrangements were isolated, lysed and amplified following a multiple displacement amplification approach using Genomi Phi V2 [32]. Amplified single cell and non-amplified genomic DNA (500 ng) derived from a patient with Klinefelter syndrome was labeled for 2 hours by random primer labeling using Cy5 and Cy3 dCTPS and hybridized according to the manufacturer's instructions to the genome-wide Agilent 244K array. Slides were scanned by the Feature Extraction software using Agilent protocol CGH-v4_10_Apr08. As a validation, genomic DNA isolated from multiple cells of the corresponding EBV-transformed lines was karyotyped as well as analyzed on 250K Affymetrix SNP array to confirm real aberrant regions. A deleted region on a SNP array presents only a single allele and is indicated by loss-of-heterozygosity. Diploid regions were confirmed by heterozygous SNPs [9]. The karyotype of each EBV sample is listed in Table 3.1.

Human embryo array CGH
Fourteen blastomeres derived from human embryo 6, 8, 15, 16, 19 and 20 carrying known chromosomal rearrangements were hybridized to the Agilent 244K array. The experiment protocol and validation were similar to the single EBV-transformed cell array CGH and were explained in detail in previous published work [9]. Most of these blastomeres carry multiple aberrations within one cell. The complete karyotype of each blastomere is reported in [9].

Gene Expression Omnibus accession number
All the single-cell data from this study are public accessible in the Gene Expression Omnibus under SuperSeries [GSE31219]. GSE31219 contains Single EBV-transformed lymphoblastoid cell array and human blastomeres array data. The previously published human blastomeres data are accessible through GEO series accession number [GSE 11663].

Evaluation of CNV calling in single-cell array CGH experiments
The parameters of the CBS algorithm were optimized to detect validated known CNVs of EBV-transformed single cells. The CGHcall program was used to call

the CNVs in single-cells. It fits each CBS segment to a mixture model with four states and calls each segment as a duplication, deletion, amplification or normal state [37]. We calculated the TPR and FPR to evaluate the CNV detection. TPR was defined as the length of CGHcall CNVs within the true aberrant regions divided by the total length of true aberrant regions. FPR was defined as the length of CGHcall CNVs outside the aberrant regions divided by the total non-aberrant region lengths [42]. The CBS algorithm was implemented using the R package snapCGH [54].

Acknowledgements

We are grateful to the editor and two reviewers for the reviewing the manuscript and providing precious comments. This work was supported by Research Council K.U.Leuven: ProMeta, GOA Ambiorics, GOA MaNet, CoE EF/05/007 SymBioSys, START 1, several PhD/postdoc & fellow grants; Flemish government: FWO-PhD/postdoc grants, FWO-projects, G.0318.05 (subfunctionalization), G.0553.06 (VitamineD), G.0302.07 (SVM/Kernel), FWO-research communities (ICCoS, ANMMM, MLDM), G.0733.09 (3UTR), G.082409 (EGFR), IWT- PhD Grants, Silicos, SBO-BioFrame, SBO-MoKa, SBO-60848, TBM-IOTA3, FOD-Cancer plans; Belgian Federal Science Policy Office: IUAP P6/25 (BioMaGNet, Bioinformatics and Modeling: from Genomes to Networks, 2007-2011); EU-RTD: ERNSI: European Research Network on System Identification; FP7-HEALTH CHeartED. We would like to thank Agilent for providing us with the Agilent 244K arrays. We would also like to thank Sigrun Jackmaert for preparing and performing the array experiments. We appreciated theoretical discussions with Leon Charles Tranchevent and Kristof Engelen. E.V. was supported by the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen).

Additional files

Additional file 1: Figure S1 - MA plot of single-cell array CGH. MA plot of EBV-transformed cell 1160. The spots in the plot are the clones excluding internal control and incomplete physical annotated clones. The red spots represent clones with intensities more than five-fold lower than the median background intensity.

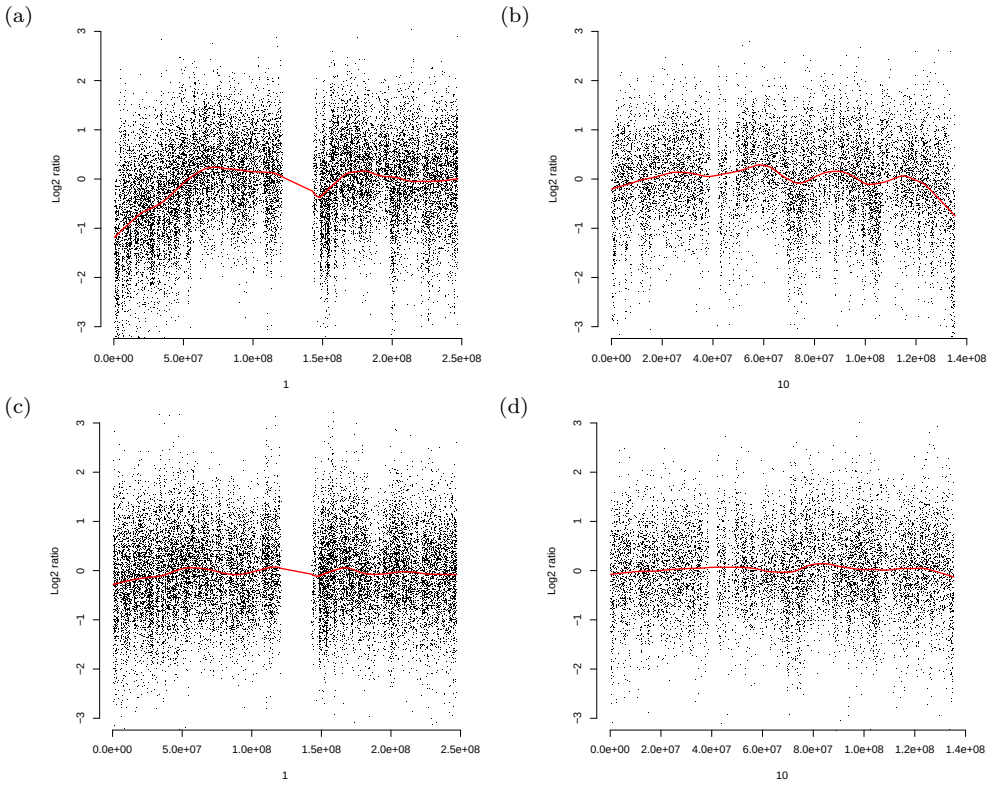

Figure S2 Genome profile of single-cell array CGH before and after genome composition correction. Genome plots of EBV-transformed cell 1168. (a,b) Genome plots of chromosomes 1 and 10 before genome composition correction. (c,d) Genome plots of chromosomes 1 and 10 after genome composition correction. The red line represents a lowess curve fitted to the data.

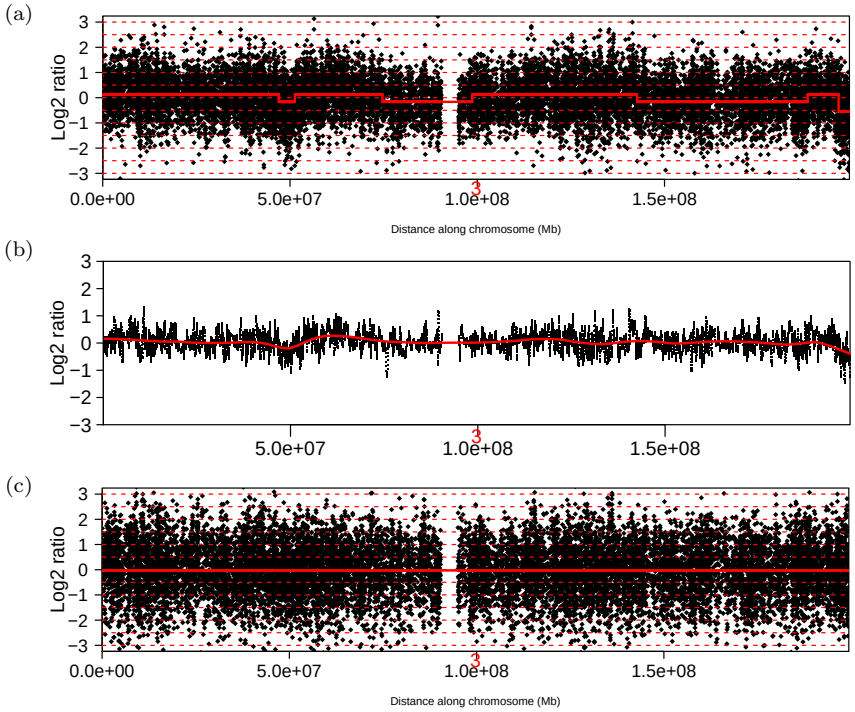

Figure S3 Genome profile of single-cell array CGH before and after recurrent genome artifacts correction. (a) Genome plot of chromosome 3 from EBV-transformed cell 1168 before recurrent genome artifact correction. The red line represents the CBS segmentation. (b) Estimated common profile trend of chromosome 3 across all the EBV-transformed cells. The red line represents a lowess curve. (c) Genome plot of chromosome 3 from EBV-transformed cell 1168 after recurrent genome artifact correction. The red line represents the CBS segmentation.

Additional file 4: Figure S4 - genome-wide copy number variation detection of single EBV-transformed cell 1168 using existing normalization methods. (a-d) Single-cell CNV detection of EBV- transformed cell 1168 after global loess (a), CGHnormaliter (b), poplowess (c) and Haarseg normalization (d). The y-axis represents the log2 ratios and the x-axis the probe position along the chromosome. The blue line represents the CBS segmentation line. The red region represents the deletion and the green region represents the duplication called by CGHcall.

Additional file 5: Figure S5 - genome-wide copy number variation

detection of single EBV-transformed cell 1151 using existing normalization methods. (a-d) Single-cell CNV detection of EBV-transformed cell 1151 after global loess (a), CGHnormaliter (b), poplowess (c) and Haarseg normalization (d). The y-axis represents the log2 ratios and the x-axis the probe position along the chromosome. The blue line represents the CBS segmentation line. The red region represents the deletion and the green region represents the duplication called by CGHcall.

Additional file 6: Figure S6 - genome-wide copy number variation detection of single EBV-transformed cell 1160 using existing normalization methods. (a-d) Single-cell CNV detection of EBV- transformed cell 1160 after global loess (a), CGHnormaliter (b), poplowess (c) and Haarseg normalization (d). The y-axis represents the log2 ratios and the x-axis the probe position along the chromosome. The blue line represents the CBS segmentation line. The red region represents the deletion and the green region represents the duplication called by CGHcall.

Additional file 7: Figure S7 - genome-wide copy number variation detection of single EBV-transformed cell 1162 using existing normalization methods. (a-d) Single-cell CNV detection of EBV- transformed cell 1162 after global loess (a), CGHnormaliter (b), poplowess (c) and Haarseg normalization (d). The y-axis represents the log2 ratios and the x-axis the probe position along the chromosome. The blue line represents the CBS segmentation line. The red region represents the deletion and the green region represents the duplication called by CGHcall.

Additional file 8: Figure S8 - genome-wide copy number variation detection of single EBV-transformed cell 614 using existing normalization methods. (a-d) Single-cell CNV detection of EBV- transformed cell 614 after global loess (a), CGHnormaliter (b), poplowess (c) and Haarseg normalization (d). The y-axis represents the log2 ratios and the x-axis the probe position along the chromosome. The blue line represents the CBS segmentation line. The red region represents the deletion and the green region represents the duplication called by CGHcall.

Additional file 9: Figure S9 - genome-wide copy number variation detection of single EBV-transformed cell 617 using existing normalization methods. (a-d) Single-cell CNV detection of EBV- transformed cell 617 after global loess (a), CGHnormaliter (b), poplowess (c) and Haarseg normalization (d). The y-axis represents the log2 ratios and the x-axis the probe position along the chromosome. The blue line represents the CBS segmentation line. The red region represents the deletion and the green region represents the duplication called by CGHcall.

Additional file 10: Figure S10 - genome-wide copy number variation detection of single EBV-transformed cell 1013 using existing normalization methods. (a-d) Single-cell CNV detection of EBV-transformed cell 1013 after global loess (a), CGHnormaliter (b), poplowess (c) and Haarseg normalization (d). The y-axis represents the log2 ratios and the x-axis the probe position along the chromosome. The blue line represents the CBS segmentation line. The red region represents the deletion and the green region represents the duplication called by CGHcall.

Additional file 11: Figures S11 - genome-wide copy number variation detection of single EBV-transformed cell 1168 using the channel clone normalization method. Single-cell CNV detection of EBV-transformed cell 1168 after channel clone normalization. The y-axis represents the log2 ratios and the x-axis the probe position along the chromosome. The blue line represents the CBS segmentation line. The red region represents the deletion and the green region represents the duplication called by CGHcall.

Additional file 12: Figures S12 - genome-wide copy number variation detection of single EBV-transformed cell 1151 using the channel clone normalization method. Single-cell CNV detection of EBV-transformed cell 1151 after channel clone normalization. The y-axis represents the log2 ratios and the x-axis the probe position along the chromosome. The blue line represents the CBS segmentation line. The red region represents the deletion and the green region represents the duplication called by CGHcall.

Additional file 13: Figures S13 - genome-wide copy number variation detection of single EBV-transformed cell 1160 using the channel clone normalization method. Single-cell CNV detection of EBV-transformed cell 1160 after channel clone normalization. The y-axis represents the log2 ratios and the x-axis the probe position along the chromosome. The blue line represents the CBS segmentation line. The red region represents the deletion and the green region represents the duplication called by CGHcall.

Additional file 14: Figures S14 - genome-wide copy number variation detection of single EBV-transformed cell 1162 using the channel clone normalization method. Single-cell CNV detection of EBV-transformed cell 1162 after channel clone normalization. The y-axis represents the log2 ratios and the x-axis the probe position along the chromosome. The blue line represents the CBS segmentation line. The red region represents the deletion and the green region represents the duplication called by CGHcall.

Additional file 15: Figures S15 - genome-wide copy number variation detection of single EBV-transformed cell 614 using the channel clone normalization method. Single-cell CNV detection of EBV-transformed cell

614 after channel clone normalization. The y-axis represents the log2 ratios and the x-axis the probe position along the chromosome. The blue line represents the CBS segmentation line. The red region represents the deletion and the green region represents the duplication called by CGHcall.

Additional file 16: Figures S16 - genome-wide copy number variation detection of single EBV-transformed cell 617 using the channel clone normalization method. Single-cell CNV detection of EBV-transformed cell 617 after channel clone normalization. The y-axis represents the log2 ratios and the x-axis the probe position along the chromosome. The blue line represents the CBS segmentation line. The red region represents the deletion and the green region represents the duplication called by CGHcall.

Additional file 17: Figures S17 - genome-wide copy number variation detection of single EBV-transformed cell 1013 using the channel clone normalization method. Single-cell CNV detection of EBV-transformed cell 1013 after channel clone normalization. The y-axis represents the log2 ratios and the x-axis the probe position along the chromosome. The blue line represents the CBS segmentation line. The red region represents the deletion and the green region represents the duplication called by CGHcall.

Additional file 18: Figure S18 - genome-wide copy number variation detection of single EBV-transformed cells using the GADA algorithm. Single-cell CNV detection of all seven EBV-transformed cells. Each row represents the profile of one EBV-transformed cell and each column represents one probe across all the EBV-transformed samples. Different colors in the profile represent the breakpoints of single-cell CNVs detected by GADA.

Additional file 19: R code to implement channel clone normalization approach. Due to a large amount of supplementary figures, only selected Figure S2 and S3 are shown here, the rest can be found via the following link: http://genomebiology.com/2011/12/8/R80/additional

3.3 Postscript

This paper was published in Genome Biology 2011, 12: R80. The term "copy number variation" was mentioned in the paper. However, copy number variation is mostly used to indicate the common copy number changes present in germline DNA. This is not the case for single cells used in this paper. It would have been more precise if we used the term "single-cell copy number alteration" in the concerned paper. The term "genomic DNA" mentioned in the paper represents DNA extracted from a pooling number of cells. In contrast, "single-cell DNA"

is extracted from one single cell. "Median background intensities" represent the median of intensities of each background pixel of the array.

This chapter introduced a novel processing technique to correct the biases caused by amplification, GC rich regions and recurrent genome artifacts. The channel standardization step aims at correcting the intra-sample and inter-sample biases. Quantile normalization of the channel intensities across all the samples can also achieve this purpose. This step assumes that the majority of the chromosomes within one sample have no aberrations.

The channel standardization and the GC correction steps were performed based on each sample individually while the recurrent artifact correction was performed on a batch of samples together. The recurrent artifact correction step assumes that only genome artifacts are recurrently present at the same positions across a batch of samples. However, this step needs to be adjusted for samples carrying recurrent aberrations because otherwise it risks overcorrecting those aberrations. By using the segmented Log2Ratios prior to the recurrent artifact correction step, the real aberrations can be protected. This adjustment of the algorithm still requires further evaluation in the future.

The channel-clone normalization is suitable for two-channel (two-color) as well as single-channel (one-color) array data. For the single-channel array data, we can replace the channel intensities with log2Ratios across all the samples to perform the channel standardization step. The implementation of the rest of the channel-clone steps are equal to the process followed for two-channel array data.



Chapter 4

Channel-clone normalization applied to array data from human embryos

4.1 Summary

We propose a channel-clone normalization method based on channel standardization, genome composition and recurrent genome artifact correction to reduce single-cell genome artifacts, as described in the previous chapter. The results show that by applying a channel-clone processing technique, the accuracy to detect single-cell copy number changes improves considerably.

In this chapter, we apply this channel-clone approach to BAC array data of single blastomere biopsies aspirated from 14 normally developing human cleavage-stage embryos. The goal of this study is to investigate the chromosomal instabilities during early developmental of human embryos. Among these 14 human embryos, we found 10 embryos with different karyotypes for individual blastomeres. Particularly, around half of the analyzed blastomeres carry chromosomal aberrations and around 13% of them carry structural rearrangements. Our results confirm that chromosomal instabilities are already frequently present in early development of human embryos with good-quality.

4.2 Personal contributions

The PhD candidate first evaluated the performance of the channel-clone normalization using BAC array data from EBV-transformed cells. The results are discussed in the postscript section. Subsequently, the PhD candidate performed the complete data analysis of the 70 cleavage-stage blastomeres, including probes filtering, channel intensity standardization, GC correction, recurrent genome artifact correction, CBS segmentation and copy number calling. Finally, the other joint first authors summarized the karyotype of each blastomere and interpreted the overall results from all blastomeres.

This paper was published in Human Reproduction 2013, 28(1): 256-64.

Microarray analysis reveals abnormal chromosomal complements in over seventy percent of fourteen normally developing human embryos

Running title: aCGH reveals abnormalities on good-quality human embryos

A. Mertzanidou[1,*], L. Wilton[2,*], J. Cheng[3,4,*], C. Spits[1], E. Vanneste[5], Y. Moreau[3,4], J. R. Vermeesch[5],K. Sermon[1,$]

[1]Research Group Reproduction & Genetics, Faculty of Medicine and Pharmacy, Vrije Universiteit Brussel (VUB), 1090, Brussels, Belgium.
[2]Preimplantation Genetics, Melbourne IVF, Suite 10/320 Victoria Parade, Victoria, Australia.
[3]Department of Electrical Engineering, Esat-SCD, Katholieke Universiteit Leuven, 3001 Leuven, Belgium.
[4]IBBT-K.U.Leuven Future Health Department, 3001, Leuven, Belgium.
[5]Centre for Human Genetics, University Hospital Gasthuisberg, B-3000, Leuven, Belgium.
[*]The authors consider that the first three authors should be regarded as joint First Authors.

[$]Corresponding author; Tel: +32 (0)2 477 46 35; E-mail: karen.sermon@uzbrussel.be.

Abstract

Study question: What are the aneuploidy rates and incidence of mosaicism in good-quality human preimplantation embryos?

Summary answer: High-level mosaicism and structural aberrations are not restricted to arrested or poorly developing embryos but are also common in good-quality IVF embryos.

What is known already: Humans, compared with other mammals, have a poor fertility rate, and even IVF treatments have a relatively low success rate. It is known that human gametes and early preimplantation embryos carry chromosomal abnormalities that are thought to lower their developmental potential.

Study design, size and duration: The embryos studied came from nine young (age, 35 years old) IVF patients and were part of a cohort of embryos that all resulted in healthy births. These 14 embryos inseminated by ICSI and cryopreserved on Day 2 of development were thawed, cultured overnight and allowed to succumb by being left at room temperature for 24 h. Following removal of the zona pellucida, blastomeres were disaggregated and collected.

Participants/materials, setting and methods: There were 91 single blastomeres collected and amplified by multiple displacement amplification. Array-comparative genomic hybridization was performed on the amplified DNA. Array-data were normalized and aneuploidy was detected by the circular binary segmentation method.

Main results and the role of chance: The good-quality embryos exhibited high rates of aneuploidy, 10 of 14 (71.4%) of the embryos being mosaic. While none of the embryos had the same aneuploidy pattern in all cells, 4 of 14 (28.6%) were uniformly diploid. Of the 70 analysed blastomeres, 55.7% were diploid and 44.3% had chromosomal abnormalities, while 29% of the abnormal cells carried structural aberrations.

Wider implications of the findings: Finding such a high rate of aneuploidy and mosaicism in excellent quality embryos from cycles with a high implantation rate warrants further research on the origin and significance of chromosomal abnormalities in human preimplantation embryos.

Study funding/competing interest(s): This research was supported by the Instituut voor de aanmoediging van innovatie door Wetenschap en Technologie in Vlaanderen (IWT-Vlaanderen). A.M. is a PhD student at the IWT-Vlaanderen. C.S. is a postdoctoral fellow at the FWO Vlaanderen. There are no competing interests.

Key words: aneuploidy / array-CGH / mosaicism / preimplantation embryo / chromosomal abnormalities

Introduction

Preimplantation genetic screening (PGS) has been used over the last 15 years to test the chromosome content of human IVF embryos and select for uterine transfer of those that may have a better potential to lead to a healthy birth [68]. Typically, PGS is done by performing fluorescence in situ hybridization (FISH) for 5-10 chromosomes on one or two cells of a Day-3 IVF embryo. The large number of patients who have been treated worldwide with this technology has generated solid data sets on the ploidy status of human preimplantation embryos. Even if FISH only gives limited information on the chromosomal constitution of a cell, it is obvious from these data that human embryos often carry numerical chromosome errors and that cells within one embryo can have different karyotypes.

Studies using high-resolution methods for the complete karyotyping of a cell highlight the inherent limitations of FISH in detecting aneuploidy and the actual incidence of mosaicism [31, 69, 70, 71]. In the studies that use FISH as a method of analysis, the cytogenetic assessment involves a limited number of chromosomes, while array-comparative genomic hybridization (array-CGH) enables analysis of the whole genome of the cell and more abnormalities can be detected than with FISH. Moreover, the size and type of aberrations that can be detected by FISH are restricted. Recently, Vanneste et al. (2009) used an array-based approach on the remaining blastomeres from embryos after preimplantation genetic diagnosis and showed that not only aneuploidy but also chromosome breakage is common in cleavage-stage embryos and often leads to segmental aberrations. The presence of segmental imbalances in good-quality cleavage-stage embryos was confirmed via single-nucleotide polymorphism (SNP) based parental origin analyses during the clinical implementation of array-CGH [39]. Johnson et al. (2010) found similar results using SNP-arrays on multiple cells of cleavage-stage embryos from patients of advanced maternal age [28]. Aetiological models underlying this structural instability as well as the link to genomic profiles of healthy and diseased newborns mechanisms were reviewed in [72].

Until now, most of the data on full chromosomal content of human preimplantation embryos using either CGH or array-CGH were extrapolated from the information obtained from only one or two blastomeres [31, 73, 74]. Although these studies give a good impression of the overall aneuploidy rates in embryos, their main limitation is that they do not provide information on

the embryo as a whole. Two studies provide data from all blastomeres from good-quality embryos, using metaphase CGH, which has a lower resolution than array-based CGH [75, 76]. One study used high-resolution SNP arrays to analyze multiple cells from good-quality embryos from older women and two other studies performed array-CGH on the remaining blastomeres from preimplantation genetic diagnosis and screening embryos, giving a nearly complete image of the embryos [9, 28, 33].

In this study, using array-CGH, we analyzed all the blastomeres of normally developing cleavage-stage embryos that were part of a cohort of embryos that resulted in healthy births, to provide a high resolution full chromosomal analysis of entire human embryos.

Materials and Methods

Embryos
Embryos used in this study were surplus to Melbourne IVF patients' needs and had been declared excess by the couple according to Australian Federal and Victorian State legislation and National Health and Medical Research Council requirements. Embryos were from six couples who had achieved a live birth from the same IVF cycle. The mean age of the women was 31.3 (range 29-35) years and a maximum of four embryos were used from any one couple. All the embryos had been inseminated by intracytoplasmic sperm injection and cryopreserved on Day 2 of development [77]. Fourteen embryos were thawed, cultured overnight and then allowed to succumb by being left at room temperature for 24 h. We have Institutional Review Board approval (Human Research Ethics Committee, Epworth and Freemason's Hospital, Melbourne, study number S/05/11/2) to use excess, succumbed embryos for research aimed at improving and developing novel technologies for PGD and IVF. The zona pellucida was removed from embryos by brief exposure to acidified (pH 2.4) media. Blastomeres were disaggregated by gentle pipetting.

Whole genome amplification and array-CGH of single blastomeres
Multiple displacement amplification was performed as described in [32].

Array-CGH was carried out using bacterial artificial chromosome/P1-derived artificial chromosome (BAC/PAC) arrays of 1 Mb resolution. All the BAC and PAC sequences are printed in duplicate and their number varies between the used batches. In particular, the array design contains in total 7744 features printed in a layout of 4 meta rows, 4 meta columns, 22 rows per block and 22 columns per block. This grid is printed twice on the glass slides allowing two microarray assays per slide. Array-CGH was carried out as described previously [33], with minor modifications.

Equal amounts (1750 ng) of the Cy3 and Cy5-labeled probes were combined with 150 μg of Cot-1DNA (humanCot-1 DNA; Invitrogen) followed by ethanol precipitation. Resuspension of the pelletwas done in 20 μl of hybridization buffer containing 200 μg of yeastRNA (Invitrogen) to hybridize a spotting area of 24×24 mm. The slide was blocked with 50 μg of Cot-1 DNA and 300 μg of salmon tested DNA (Sigma-Aldrich, St. Louis, USA) dissolved in 25 μl of hybridization buffer. The slides were then washed and scanned using an Agilent dual laser DNA microarray scanner G2566AA (Agilent technologies, Palo Alto, CA, USA). The scan images were processed with Agilent Feature Extraction Software v9.5 and data were further analyzed. Arrays with autosomal standard deviation larger than 0.75 were excluded from the analysis. Array-data were sent blinded to be preprocessed using the channel clone normalization method [25]. In brief, array data were normalized by the channel standardization, genome composition artifact correction and recurrent genome artifact correction steps. Subsequently, aneuploidy was detected by the circular binary segmentation (CBS) method and copy number state was called using CGH call [36, 37]. As described before, at least 18 consecutive BAC clones (half the clones of the smallest chromosome) with signal intensities 2-fold above the autosomal median background intensity was set as the threshold to consider the minimum length of an aberration [9].

Results

Figure 4.1 shows the chromosomal complements of each blastomere and embryo. A flow-chart of the experimental design can be found in Figure 4.2. Table 4.1 shows the detailed chromosomal complements per cell. The CBS plots for all chromosomes of all blastomeres analysed can be found in Supplementary data. The GEO accession number for the BAC-array data reported in this paper is GSE34290. According to our results, good-quality cleavage-stage embryos exhibit high rates of aneuploidy. Only 4 of 14 embryos (28.6%) were completely normal, while 10 embryos (71.4%) were mosaic and structural aberrations were found in 12 cells. No embryo contained the same aneuploidy in all of its cells.

Of the analysed blastomeres, 39/70 (55.7%) were diploid, 5 of which contained structural aberrations and 31/70 (44.3%) had numerical abnormalities. More specifically, 18.6% (13/70) of the blastomeres had a single monosomy. Additionally, cell 4.6 had a monosomy and a structural aberration that might be a ring chromosome. There were six blastomeres (8.6%) with a single trisomy. Additionally, cell 8.4 had a trisomy and also contained a structural aberration that might be a ring chromosome. Furthermore, 2.9% (2/70) of the blastomeres had two aneuploidies and 11.4% (8/70) of the blastomeres had a complex aneuploidy. Whole chromosome abnormalities were found in almost

all chromosomes except for 4, 6, 11, 16 and 17. The chromosomes 2, 19, 22 and X were most frequently involved in aneuploidy.

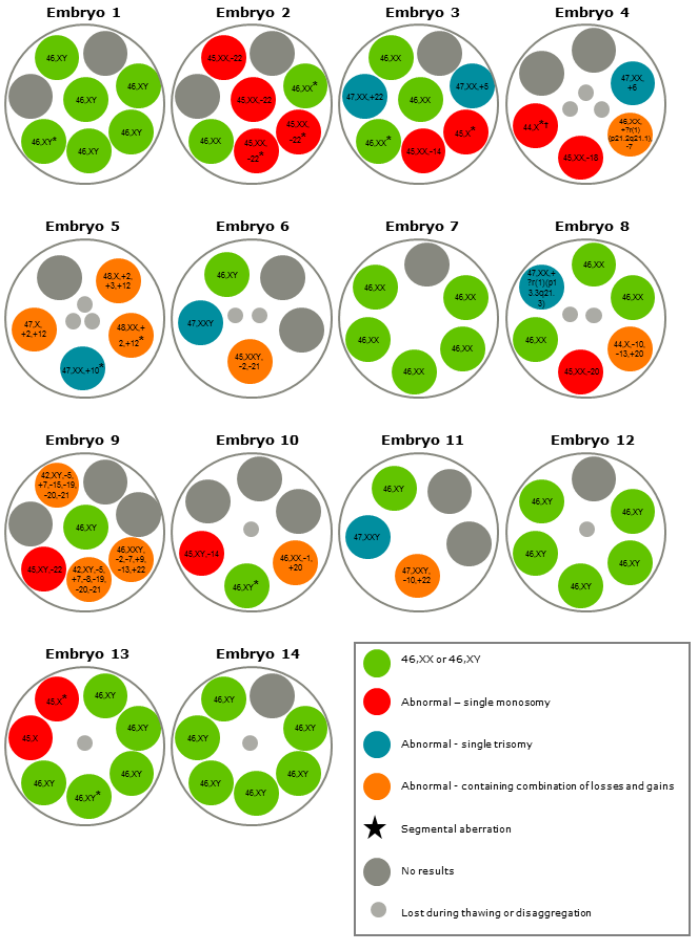

Figure 4.1: Chromosome complements of the blastomeres analyzed by array-CGH. †Chromosome 16 has a segmental aberration spanning the centromere. The acentric fragments may be translocated to other chromosomes resulting in a loss of chromosome 16.

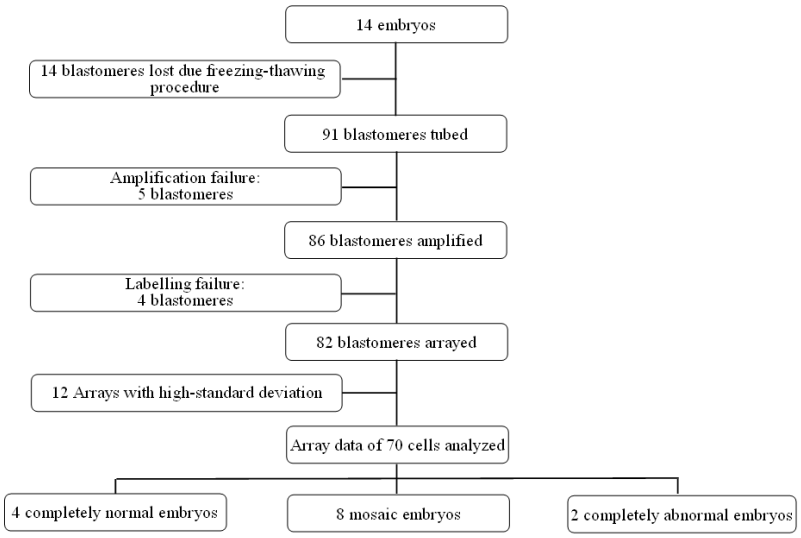

Figure 4.2: Flow-chart of the experimental design.

There were five blastomeres that had only structural aberrations (7.1%), while structural aberrations co-existed with numerical abnormalities in seven other blastomeres. The breakage events involved chromosomes 1, 2, 3, 5, 8, 9, 14, 16, 18, 22 and X. Out of a total of 16 unbalanced segments detected, 9 were interstitial, 6 were likely to be terminal and one was likely to involve the whole arm of the chromosome. Of the structural aberrations, eight were deletions and seven were duplications. Two blastomeres (4.6 and 8.4) had interstitial duplications of chromosome 1 that included the centromeric region, possibly being ring chromosomes. The smallest aberration detected was a 14 Mb deletion of the short arm of chromosome 18 at bands p11.32 through p11.21 (cell 2.4).

Table 4.1: Array results of the analyzed blastomeres.

Cell	**Chr status**	**Single copy loss**	**Single copy gain**	**Euploid (Mb)**	**Start (Mb)**	**End (Mb)**	**Euploid (Mb)**	**Size (Mb)**
1.1	Euploid, male	-	-	-	-	-	-	-
1.2	Euploid, male	-	-	-	-	-	-	-
1.3	Euploid, male	-	-	-	-	-	-	-
1.4	Euploid, male	-	-	-	-	-	-	-

Continued

Cell	Chr status	Single copy loss	Single copy gain	Euploid (Mb)	Start (Mb)	End (Mb)	Euploid (Mb)	Size (Mb)
1.5	Unbalanced, male	-	Xp22.33 p21.1	0.48	31.44	31.99	31.0	
1.6	Euploid, male	-	-	-	-	-	-	-
2.1	Unbalanced, female	22q11.1 q13.33	-	-	15.62	49.46	-	33.8
2.2	Unbalanced, female	-	3p24.1 p21.31	27.53	28.65	50.01	51.56	21.4
		22q11.1 q13.33	-	-	15.62	49.46	-	33.8
2.3	Unbalanced, female	22q11.1 q13.33	-	-	15.62	49.46	-	33.8
2.4	Unbalanced, female	18p11.32 p11.21	-	-	0.13	14.14	14.84	14.0
2.5	Unbalanced, female	Xp21.3 p22.31	-	5.42	6.81	28.28	28.78	21.5
		22q11.1 q13.33	-	-	15.62	49.46	-	33.8
2.6	Euploid, female	-	-	-	-	-	-	-
3.1	Unbalanced, female	5q12.1 q35.3	-	60.90	61.66	180.63	-	119.0
3.2	Unbalanced, female	Xp22.33 q28	-	-	0.48	154.39	-	153.9
		2q32.2 q33.3	-	189.97	190.40	206.36	206.94	16.0
		-	5q31.3 q35.3	140.12	141.04	180.63	-	39.6
		14q12 q24.3	-	31.60	32.38	74.48	74.59	42.1
3.3	Unbalanced, female	-	22q11.1 q13.33	-	15.62	49.46	-	33.8
3.4	Unbalanced, female	-	5p15.33 q35.3	-	2.57	180.63	-	178.1
3.5	Euploid, female	-	-	-	-	-	-	-
3.6	Euploid, female	-	-	-	-	-	-	-
3.7	Unbalanced, female	14q11.2 q32.33	-	-	19.57	105.44	-	85.9
4.1	Unbalanced, female	18p11.32 q23	-	-	0.13	75.94	-	75.8
4.4	Unbalanced, female	Xp22.33 q28	-	-	0.48	154.39	-	153.9
		16p12.2 q13	-	21.47	21.94	57.14	56.95	35.2
4.5	Unbalanced, female	-	6p25.3 q27	-	0.09	170.80	-	170.7
4.6	Unbalanced, female	-	1p21.2 q21.1	99.78	101.01	142.93	144.19	41.9
		7p22.3 q36.3	-	-	0.19	15.46	-	158.3
5.2	Unbalanced, female	Xp22.33 q28	-	0.48	154.39	-	153.9	
		-	2p25.3 q37.3	-	0.22	242.62	-	242.4
		-	12p13.33 q24.33	-	0.15	132.14	-	132.0

Continued

Cell	Chr status	Single copy loss	Single copy gain	Euploid (Mb)	Start (Mb)	End (Mb)	Euploid (Mb)	Size (Mb)
5.3	Unbalanced, female	8q21.3 q24.3	-	89.40	90.08	146.17	-	56.1
		-	10p15.3 q26.3	-	0.26	135.29	-	135.0
5.4	Unbalanced, female	-	2p25.3 q37.3	-	0.22	242.62	-	242.4
		-	3q13.33 q26.33	118.33	119.24	182.47	183.17	63.2
		-	12p13.33 q24.33	-	0.15	132.14	-	132.0
5.5	Unbalanced, female	Xp22.33 q28	-	-	0.48	154.39	-	153.9
		-	2p26.3 q37.3	-	0.22	242.62	-	242.4
		-	3p25.3 q29	-	0.19	199.16	-	199.0
		-	12p13.33 q24.33	-	0.15	132.14	-	132.0
6.1	Unbalanced, male	-	Xp22.33 q28	-	0.48	154.39	-	153.9
		2p25.3 q37.3	-	-	0.22	242.62	-	242.4
		21q11.2 q22.3	-	-	13.46	46.93	-	33.5
6.2	Unbalanced, male	-	Xp22.33 q28	-	0.48	154.39	-	153.9
6.3	Euploid, male	-	-	-	-	-	-	-
7.1	Euploid, female	-	-	-	-	-	-	-
7.2	Euploid, female	-	-	-	-	-	-	-
7.3	Euploid, female	-	-	-	-	-	-	-
7.4	Euploid, female	-	-	-	-	-	-	-
7.6	Euploid, female	-	-	-	-	-	-	-
8.1	Unbalanced, female	Xp22.33 q28	-	-	0.48	154.39	-	153.9
		10p15.3 q26.3	-	-	0.26	135.29	-	135.0
		13q12.11 q34	-	-	19.14	113.93	-	94.8
		-	20p13 q13.33	-	0.18	62.39	-	62.2
8.2	Euploid, female	-	-	-	-	-	-	-
8.3	Unbalanced, female	20p13 q13.33	-	-	0.18	62.40	-	62.2
8.4	Unbalanced, female	-	1p13.3 q21.3	106.79	107.30	150.65	150.88	43.4
8.5	Euploid, female	-	-	-	-	-	-	-
8.6	Euploid, female	-	-	-	-	-	-	-
9.1	Euploid, male	-	-	-	-	-	-	-

Continued

Cell	Chr status	Single copy loss	Single copy gain	Euploid (Mb)	Start (Mb)	End (Mb)	Euploid (Mb)	Size (Mb)
9.2	Unbalanced, male	22q11.1 q13.33	-	-	15.62	49.46	-	33.8
9.3	Unbalanced, male	5p15.33 q35.3	-	-	2.57	180.63	-	178.1
		-	7p22.3 q36.3	-	0.19	158.46	-	158.3
		15q11.2 q26.3	-	-	20.36	100.02	-	79.7
		19p13.3 q13.43	-	-	0.21	63.77	-	63.6
		20p13 q13.33	-	-	0.18	62.39	-	62.2
		21q11.2 q22.3	-	-	13.46	46.93	-	33.5
9.5	Unbalanced, male	5p15.33 q35.3	-	-	2.57	180.63	-	178.1
		-	7p22.3 q36.3	-	0.19	158.46	-	158.3
		8p23.3 q24.3	-	-	0.48	146.17	-	145.7
		19p13.3 q13.43	-	-	0.21	63.77	-	63.6
		20p13 q13.33	-	-	0.18	62.39	-	62.2
		21q11.2 q22.3	-	-	13.46	46.93	-	33.5
9.6	Unbalanced, male	-	Xp22.33 q28	-	0.48	154.39	-	153.9
		2p25.3 q37.3	-	-	0.22	242.62	-	242.4
		7p22.3 q36.3	-	-	0.19	158.46	-	158.3
		-	9p24.3 q34.2	-	0	136.36	-	136.4
		13q12.11 q34	-	-	19.14	113.93	-	94.8
		-	22q11.21 q13.33	-	15.62	49.46	-	33.8
10.3	Unbalanced, male	14q11.2 q32.33	-	-	19.57	105.44	-	85.9
10.5	Unbalanced	Yp11.32 q12	-	-	0.59	57.50	-	56.9
		-	Xp22.33 q28	-	0.48	154.39	-	153.9
		1p36.33 q44	-	-	0.97	245.41	-	244.4
		-	20p13 q13.33	-	0.18	62.39	-	62.2
10.6	Unbalanced, male	9q21.11 q34.3	-	69.68	70.53	138.27	-	67.7
11.2	Unbalanced, male	Xp22.33 q28	-	0.48	154.39	-	153.9	
		10p15.3 q26.3	-	0.26	135.29	-	135.0	
		22q11.1 q13.33	-	15.62	49.46	-	33.8	
11.3	Euploid, male	-	-	-	-	-	-	-

Continued

Cell	Chr status	Single copy loss	Single copy gain	Euploid (Mb)	Start (Mb)	End (Mb)	Euploid (Mb)	Size (Mb)
11.5	Unbalanced, male	-	Xp22.33 q28	-	0.48	154.39	-	153.9
12.1	Euploid, male	-	-	-	-	-	-	-
12.2	Euploid, male	-	-	-	-	-	-	-
12.4	Euploid, male	-	-	-	-	-	-	-
12.5	Euploid, male	-	-	-	-	-	-	-
12.6	Euploid, male	-	-	-	-	-	-	-
13.1	Unbalanced, male	8q21.3 q24.13	-	88.30	89.22	122.88	123.53	33.7
13.2	Euploid, male	-	-	-	-	-	-	-
13.3	Unbalanced	Yp11.32 q12	-	-	0.59	57.50	-	56.9
		-	22q11.21 q12.3	-	15.62	34.95	35.47	19.3
13.4	Euploid, male	-	-	-	-	-	-	-
13.5	Unbalanced	Yp11.32 q12	-	-	0.59	57.50	-	56.9
13.6	Euploid, male	-	-	-	-	-	-	-
13.7	Euploid, male	-	-	-	-	-	-	-
14.1	Euploid, male	-	-	-	-	-	-	-
14.2	Euploid, male	-	-	-	-	-	-	-
14.3	Euploid, male	-	-	-	-	-	-	-
14.4	Euploid, male	-	-	-	-	-	-	-
14.6	Euploid, male	-	-	-	-	-	-	-
14.7	Euploid, male	-	-	-	-	-	-	-

Discussion

In our results, half of the cells were normal, which is in line with older studies that used CGH on cells of good-quality embryos. Voullaire et al. (2000) carried out CGH on 63 succumbed blastomeres and found that 41% of the cells were normal [75]. From the remaining ones, 27% had a single monosomy and 3% a single trisomy, while ~5% had two aneuploidies and 16% of the blastomeres carried complex abnormalities. Additionally, 5% of the cells had partial chromosome losses and gains. The findings of Wells and Delhanty (2000) on 64 single blastomeres from fresh embryos were comparable [76]. They reported 56% of

normal cells, 8% with single monosomy and 17% with single trisomy, while 14% of the cells had two or more aneuploid chromosomes and 6.3% of the cells had structural aberrations. Comparable data were obtained after SNP array analysis of single blastomeres. Johnson et al. (2010) found that 48% of cells were euploid, 32% had aneuploidy of one or two chromosomes, 20% had complex aneuploidy of three or more chromosomes and 12% had structural abnormalities [28].

A large number of chromosomal abnormalities in single blastomeres were also reported when array-CGH was performed on single blastomeres from non-PGS embryos not suitable for transfer and from PGS embryos [40]. In this study, 65% of the tested cells were found to be abnormal and 19.4% of the abnormal cells had a single monosomy and 14.7% a single trisomy. The same study reported that chromosomes 15, 16, 21 and 22 were most frequently involved in aneuploidy. Similarly, in a study performing array-CGH and FISH on blastocysts, chromosomes 15, 16, 21, 22 and X were found to be most frequently abnormal [70]. In our hands, chromosomes X and 22 were also frequently abnormal along with chromosomes 2 and 19.

With regard to mosaicism, a recent systematic review on the chromosomal constitution of human preimplantation embryos found that 73% of the embryos were mosaic, 22% diploid and the remaining 5% contained other abnormalities [78]. Almost all of the studies included in this meta-analysis used FISH to detect aneuploidy and so a limited number of chromosomes were analyzed. Moreover, it has recently been reported that FISH may well overestimate the true frequency of mosaicism in early embryos [79]. Despite these limitations, our results are in agreement showing 71.4% mosaic embryos and 28.6% diploid embryos.

The introduction of higher resolution techniques such as the single-cell metaphase and array-CGH has enabled the detection of de novo partial chromosome losses and gains. Chromosome breakage leading to chromosomal imbalance is reported in several studies using CGH [73, 75, 76, 80]. Daphnis et al. (2008) reported that in a group of 17 embryos where CGH revealed at least one cell with abnormal chromosomal complement, 28% of the events leading to mosaicism were due to partial chromosome breakage [81]. Using an array-based approach, Vanneste et al. (2009, 2011) reported that 31-70% of the embryos carried structural deletions, duplications or amplifications [9, 82]. In our hands, structural aberrations were found in 29% of the abnormal cells. Although it is now clear that structural aberrations are a common occurrence in preimplantation embryos, their true frequency, underlying mechanisms and biological significance are yet to be elucidated [39, 72].

The embryos chosen for this study came from young women whose assisted

reproduction treatment (ART) cycles were successful. Sibling embryos had a clinical pregnancy rate of 100% (9/9), as evidenced by the presence of a fetal heartbeat at 6 weeks gestation, and all the nine pregnancies progressed to a live birth. The implantation rate (number of fetal hearts/embryos transferred) was 11/15 (73%). There was no difference in embryo cell number or quality between the embryos that were transferred or analyzed by array-CGH. The frequency of abnormalities detected in these embryos is thus difficult to reconcile with the outcomes of the sibling embryos that were transferred in the same ART cycles, if it is assumed that only embryos with a completely normal chromosome complement could have implanted. The thorough and previous validation of the technology used in this study rules out that the results were technical artifacts [9, 25, 33]. This brings up the question of the minimum required number of karyotypically normal cells in an embryo to be viable. Given the high implantation rate in the transferred sibling embryos, it could be assumed that some of the mosaic embryos also have reasonable implantation potential. For instance, 70% of the cells of embryo 13 were normal, while embryo 3 had two normal cells out of the seven analyzed. There is evidence that embryos that are diploid-aneuploid mosaic at the cleavage stage but are still developing with a normal cleavage rate and pattern can reach the blastocyst stage, although this does not imply that they all would implant [82, 83, 84]. It has been suggested that embryos might 'self-correct' their chromosome complement as they develop towards the blastocyst stage [9, 85, 86, 87]. Several studies show that the proportion of aneuploid cells in embryos diminish as the embryos go through the cleavage, morula and blastocyst stage [83, 88]. Several mechanisms have been suggested to explain the 'self-correction', such as preferential allocation of diploid cells to the inner cell mass, loss of aneuploid cells due to apoptosis or trisomic rescue by anaphase lagging or non-disjunction [87, 89, 90]. Although our data do not shed light on this question, they represent a reference set for further studies.

In conclusion, in this study we present evidence that around 70% of good-quality embryos carry chromosomal abnormalities, including structural aberrations. The main strength of this work compared with other published data is that: (i) we analyzed the majority of the blastomeres of (ii) top-quality embryos from a cohort of embryos with high implantation and developmental potential and for (iii) all chromosomes. This makes this data set unique and a touchstone for future experiments. These experiments should elucidate the true frequency and biological significance of chromosomal instability and the natural course of aneuploid cells in a normally developing embryo.

Funding

This research was supported by the Instituut voor de aanmoediging van innovatie door Wetenschap en Technologie in Vlaanderen (IWT-Vlaanderen). A.M. is a PhD student at the IWT-Vlaanderen. C.S. is a postdoctoral fellow at the FWO Vlaanderen. Y.M. is funded by GOA MaNet, KUL PFV/10/016 SymBioSys, Hercules III PacBio RS, FP7-HEALTH CHeartED, Action BM1006: NGS Data analysis network.

Conflict of interest

Y.M. holds stock in Cartagenia, which develops and distributes software for analysis of copy number variation.

Supplementary data

CBS plots for all chromosomes of blastomeres C1.1 to C14.7 analyzed by array-CGH. The threshold to consider the minimum length of an aberration was set at at least 18 consecutive BAC clones with signal intensities two-fold above the autosomal median background intensity. In green are called the duplications and in red the deletions. From the sex chromosomes only chromosome X is represented in the plots. The clones from the Y chromosome were not analyzed with the CBS algorithm because they are too few in number. The presence or absence of the Y chromosome was established using the mean of log2 ratio of all the clones.

Supplementary data are available at http://humrep.oxfordjournals.org/.

4.3 Postscript

This paper was published in Human Reproduction 2013, 28(1): 256-64.

In previous chapter, we evaluated the performance of the channel-clone normalization method using Agilent 244K arrays of EBV-transformed cells. In this study however, we evaluated this technique using the same EBV-transformed cell lines, but hybridized to BAC arrays. The results indicate that, considering the limited resolution of BAC arrays, the channel-clone approach enables the correct identification of aberrations without false callings (see Additional Figures). "Numerical abnormalities" and "Structural aberrations" represent that

chromosomes show abnormal either by numbers or structures, as introduced in Chapter 1, section 1.2.

One limitation of the BAC arrays used in this study is that they have low resolutions. This is caused by the long BAC clones which are 164kb on average. The platform is not able to detect reliable small aberrations because only a few numbers of clones are targeted to small aberrant regions. Therefore the segmentation algorithms do not have enough power to identify small aberrations.

Currently, advanced versions of BAC arrays, called Bluegnome arrays, have been developed(Blue Gnome Ltd., Mill Court, reat Shelford, Cambridge, UK). One study claimed to successfully detect 2.5 Mb aberrations in single cells using the Bluegnome arrays combined with PCR-based SurePlex amplifications and threshold-based copy number calling methods [31]. By adapting the channel-clone approach for Bluegnome arrays, we expect to improve accuracies that have achieved by the Bluegnome softwares.

We introduced the concept of preimplantation genetic screening (PGS) technology in Chapter 1. The original purpose of PGS is to preselect good-quality human embryos ahead of uterine transfer by screening the chromosomal aneuploidies of one or two random extracted blastomeres. However, our finding indicates that the majority of normally developed cleavage-stage embryos are composed of normal as well as aberrant blastomeres, although they are expected to produce healthy births. This implies that cleavage-stage biopsy is probably not the best stage to perform PGS, because blastomeres generally show chromosomal instabilities during this period. However, as mentioned in Chapter 1, using polar body or blastocyst stages for PGS also has its limitations. Thus, this is an ongoing clinical topic that requires the development of new strategies.

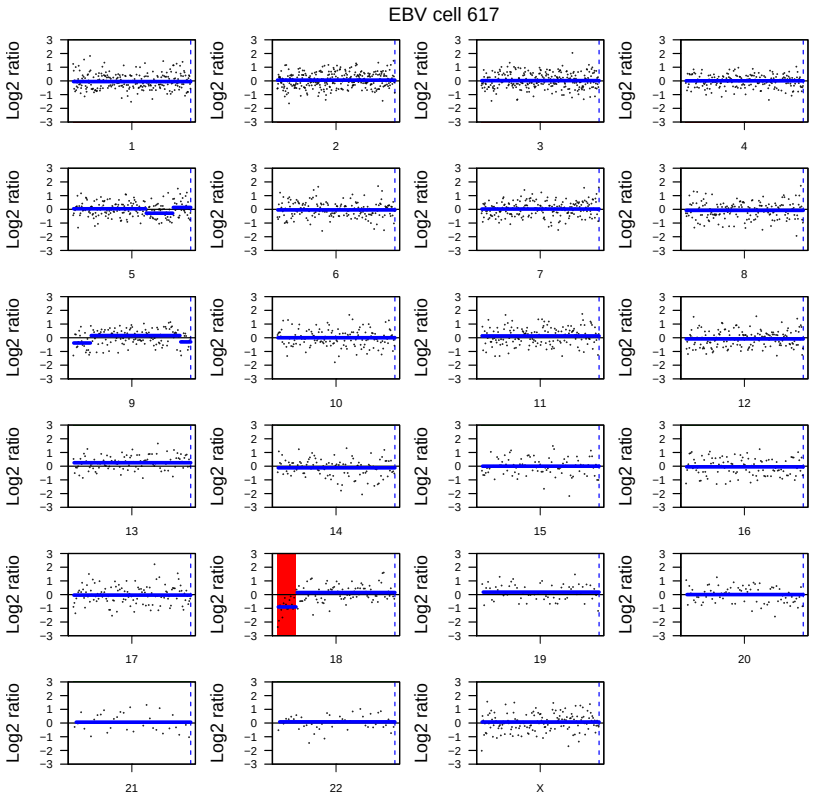

Figure 4.3: Genome-wide copy number alteration detection in EBV-transformed cell 617 after channel clone normalization. The blue line represents the CBS segmentation line. The green regions represent the duplication and red regions represent deletion called by the CGHcall program. The karyotype of cell 617 is shown in Table 3.1 of Chapter 3.

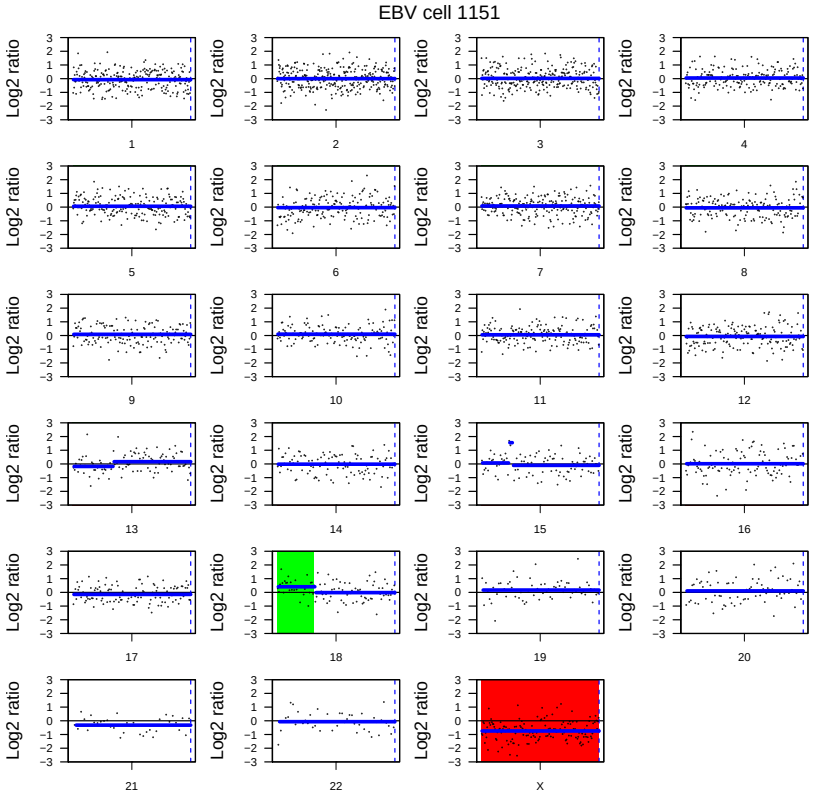

Figure 4.4: Genome-wide copy number alteration detection in EBV-transformed cell 1151 after channel clone normalization. The blue line represents the CBS segmentation line. The green regions represent the duplication and red regions represent deletion called by the CGHcall program. The karyotype of cell 1151 is shown in Table 3.1 of Chapter 3.

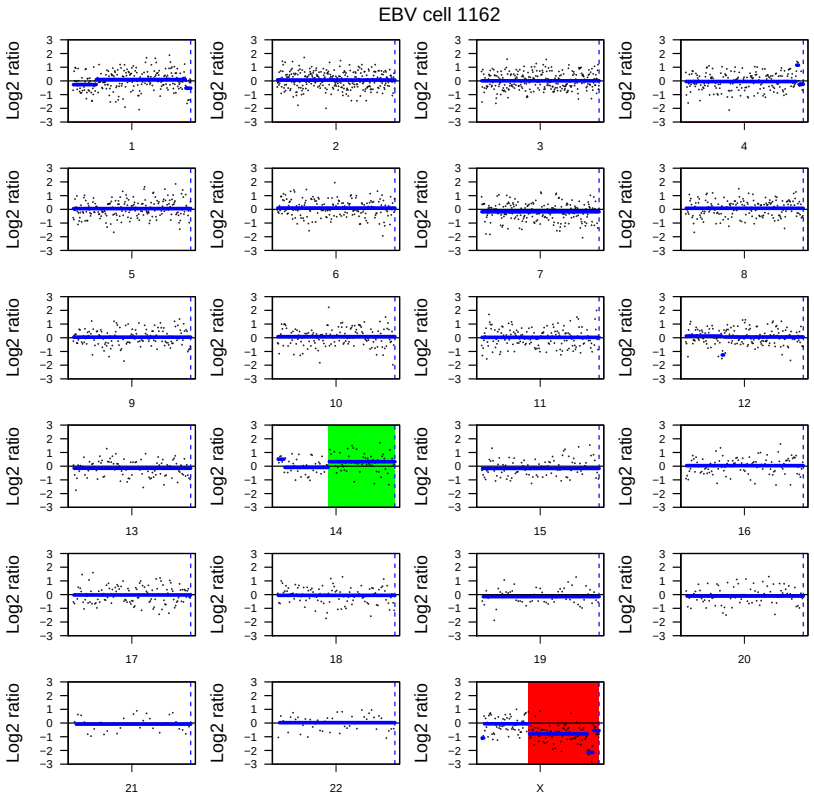

Figure 4.5: Genome-wide copy number alteration detection in EBV-transformed cell 1162 after channel clone normalization. The blue line represents the CBS segmentation line. The green regions represent the duplication and red regions represent deletion called by the CGHcall program. The karyotype of cell 1162 is shown in Table 3.1 of Chapter 3.

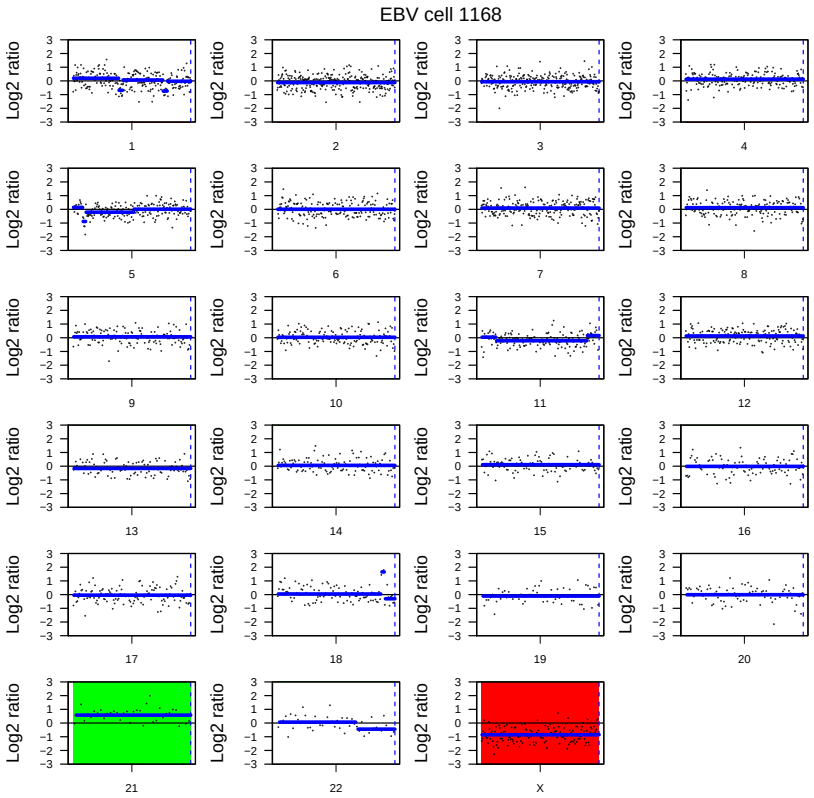

Figure 4.6: Genome-wide copy number alteration detection in EBV-transformed cell 1168 after channel clone normalization. The blue line represents the CBS segmentation line. The green regions represent the duplication and red regions represent deletion called by the CGHcall program. The karyotype of cell 1168 is shown in Table 3.1 of Chapter 3.



Chapter 5

Allelic-specific copy number analysis across tumor types

5.1 Summary

Cancer is a broad group of diseases characterized by unregulated growth of cancer cells. Cancer cells are decedents of regular somatic cells which accumulated mutations over time. Among these somatic mutations, copy number alterations (CNA) are commonly observed and play an important role in cancer development. Although various studies have been conducted using arrays to detect somatic copy number alterations in cancer genomes, the explorations of these somatic alterations are mostly limited to total copy number gains or losses, regardless of allele frequencies. In addition, the methods used in these studies can not distinguish diploid from other types of euploid genomes, which are often found in cancer cells. Cancer tissues also consist of a mixture of contaminating normal cells and tumor cells. Therefore, it is necessary to take the tumor ploidy and purity into account to infer allelic-specific copy numbers.

In this chapter, we construct genome-wide allelic-specific copy number (ASCAT) profiles of 2240 tumor samples derived from 12 tumor types by simultaneously inferring tumor ploidy, purity and allelic-specific copy numbers. Based on these inferred ASCAT profiles, we reveal the specific regions of gains, losses, loss of heterogeneity (LOH) and homozygous deletions and explore different genome architectures between tumor types. In addition, analyzing recurrent homozygous deleted regions enables us to identify fragile sites and regions harboring cancer-driving genes. In this way, we build a spectacular caner genome landscape

across 12 tumor types. Our findings provide important biological and clinical insights into cancer development.

5.2 Personal contributions

This manuscript is submitted to the journal "Cancer Discovery". The PhD candidate collected the 2240 public available SNP array samples and performed the complete data analysis including extraction of LRR and BAF values using PennCNV, genomic wave correction of LRR values and construction of ASCAT profiles [30]. In addition, the PhD candidate worked together with the corresponding author to improve the original ASCAT approach and extended the method to detect subclonal copy number alterations. Lastly, the PhD candidate inferred genome complexities, gains, losses, LOH profiles for each tumor type and optimized the rules to characterize candidate tumor suppressor genes from fragile sites.

This manuscript is going to be submitted to journal "Cancer Discovery"

Copy number meta-analysis across primary tumors highlights distinct genomic architecture and identifies tumor suppressors

Jiqiu Cheng[1,2], David C. Wedge[3], Jason J. Pitt[4], Hege G. Russnes[6,7,8], Hans Kristian Moen Vollan[6,8,9], Bina P. Pandey[1], Gro Nilsen[6,10], Silje Nord[6,8], Graham R. Bignell[3], Kevin P. White[4,11,12], Anne-Lise Børresen-Dale[6,8], Peter J. Campbell[3], Vessela N. Kristensen[6,8], Michael R. Stratton[3], Ole Christian Lingjærde[8,10], Yves Moreau[1,2], Peter Van Loo[3,13,14,*]

[1]SCD-SISTA, Department of Electrical Engineering-ESAT,University of Leuven, Leuven, Belgium;
[2]iMinds Future Health Department, University of Leuven, Leuven, Belgium;
[3]Wellcome Trust Sanger Institute, Hinxton Cambridge, UK;
[4]Institute for Genomics and Syytems Biology, University of Chicago, Chicago, IL, USA;
[5]Committee on Genetics, Genomics, and Systems Biology, University of Chicago, Chicago, IL, USA;
[6]Department of Genetics, Institute for Cancer Research, Oslo University Hospital, Oslo, Norway;
[7]Department of Pathology, Institute for Cancer Research, Oslo University Hospital, Oslo, Norway;
[8]The K.G. Jebsen Center for Breast Cancer Research, Institute for Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway;
[9]Department of Oncology, Division of Cancer Medicine, Surgery and Transplantation, Oslo University Hospital, Oslo, Norway;
[10]Biomedical Informatics, Department of Informatics, University of Oslo, Norway;
[11]Department of Ecology and Evolution, University of Chicago, Chicago, IL, USA;
[12]Department of Human Genetics, University of Chicago, Chicago, IL, USA;
[13]Center for the Biology of Disease, VIB, Leuven, Belgium;
[14]Department of Human Genetics, University of Leuven, Leuven, Belgium;
[*]To whom correspondence may be addressed: Wellcome Trust Sanger Institute, Hinxton, CB10 1SA, United Kingdom. Tel: +44 (0) 1223 834 244, e-mail: pvl@sanger.ac.uk.

Abstract

We performed a copy number meta-analysis of 12 different human cancers based on a compendium of 2240 primary tumors, to characterize their copy number landscape and identify novel tumor suppressors. In broad terms, our results demonstrate genomic complexity, genomic architecture and intra-tumor heterogeneity, highlighting key differences between cancer types in their adoption of copy number changes over evolutionary time. Our analysis defines 89 genomic regions recurrently targeted by homozygous deletions. In addition to established tumor suppressors, we identify driver homozygous deletions of *FAT1*, *MGMT*, *MXRA5*, *BIRC2/BIRC3*, *TET1*, *RAD17*, *CASP3*, *CASP9*, *SETD1B*, *BAZ1A*, *KAT6B*, *CPEB3*, *LINC00662* and *KIAA1551*. Furthermore, our findings imply that homozygous loss of several ubiquitin specific proteases (*USP8*, *USP29*, *USP34*, *USP44* and *USP50*) may contribute to cancer development. Our report highlights the utility of meta-analysis of existing large-scale genomic data sets in furthering our understanding of cancer genome architecture and of driver events of human cancer.

Significance

This large-scale meta-analysis highlights the unique nature of each tumor type and suggests that each tumor type favors a particular evolutionary path. This study identifies tumor suppressors through recurrent homozygous deletions, significantly advancing the landscape of tumor suppressors.

Introduction

The genomes of cancer cells are shaped by different types of somatic mutations, including base substitutions, small insertions and deletions, genomic rearrangements and copy number changes [10]. Many of these somatic changes are neutral passenger events, yet some confer clonal selective advantage on cancer cells and drive oncogenesis. Elucidation of these driver events provides insight into cancer biology and in some cases can directly lead to focused treatment strategies [91, 92]. For identification of cancer genes that when mutated drive oncogenesis, recurrence has proven to be a highly successful criterion [19, 93, 94]. These cancer genes can crudely be stratified into oncogenes (dominant cancer genes) and tumor suppressor genes (recessive cancer genes). Oncogenes are frequently hit by activating point mutations or amplifications and tumor suppressors are recurrently hit by inactivating mutations such as truncating point mutations or homozygous deletions.

Analyses of copy number through CGH arrays or SNP arrays have been frequently used in the past 15 years in endeavors to identify cancer genes, though it has proven difficult to separate drivers from passengers. The most successful approaches have exploited recurrence through very large sample sizes, concentrating on recurrent focal gains and losses [93, 95]. As primary tumor samples contain both tumor cells and admixed normal cells in unknown proportions, it can be difficult to identify focal losses in primary tumors as either homozygous or hemizygous deletions, hampering the identification of tumor suppressors. Cancer cell lines represent a simplified model system that does not show this normal cell admixture and a comprehensive catalog of homozygous deletions in cancer cell lines has been constructed [96]. However, this model system has important limitations: cancer cell lines are biased towards late stage cancers and metastases, they have accumulated mutations to adapt to growth on plastic, their numbers are limited and some cancer cell types do not lend themselves to growth in culture.

With the advent of SNP array technology came the realization that it is theoretically possible to infer the fraction of admixed normal cells in primary tumors from array data [97]. This prompted the development of several computational methods and mathematical models, leading to mature approaches that can infer tumor purity and ploidy and separate copy number profiles from tumor cells and admixed normal cells [45, 46, 47, 48], opening novel avenues for studying tumorigenesis and tumor evolution [98]. These methods are able to reliably separate homozygous from hemizygous deletions, and in a previous study of driver mutations in breast cancer, we have been able to pinpoint driver homozygous deletions in known and newly identified tumor suppressors [94].

Large-scale features derived from copy number profiles of cancer genomes, capturing genomic architecture or complexity, can provide additional insight into cancer biology. Some of these features have been shown to be clinically relevant [99, 100, 101, 102]. Copy number profiles can also provide insight into genetic intratumor heterogeneity and we and others have demonstrated the significance of a tumor's subclonal architecture in its evolution [103, 104, 105, 106, 107]. However, the variation of genetic heterogeneity across tumor types remains unexplored. Many studies on genomic complexity focus on a single cancer type [99, 100, 101], whereas pan-cancer studies concentrate mostly on ploidy, whole-arm aberrations and chromothripsis [47, 93, 95]. Hence, a comprehensive analysis of genomic complexity and genetic intratumor heterogeneity across cancer types is currently lacking.

Here, we perform a copy number meta-analysis of human cancer, based on a compendium of more than 2000 SNP arrays across 12 cancer types. We study genomic complexity and genetic intratumor heterogeneity across cancer types, showing that different cancer types adopt copy number changes in different

ways to shape their genomes. We find 89 genomic regions recurrently targeted by homozygous deletions, many of which hit known tumor suppressors or fragile sites. Our results add to the emerging evidence of several genes recently proposed in the literature as tumor suppressors (including *FAT1*, *BIRC2/BIRC3*, *TET1*, *MGMT* and *USP44*), and bring forward several candidate tumor suppressors (including *CASP3*, *CASP9*, *RAD17*, *BAZ1A*, *CPEB3* and *SETD1B*). In addition, we hypothesize a tumor suppressive role for several ubiquitin specific proteases (*USP8*, *USP29*, *USP34*, *USP44* and *USP50*).

Results

Allelic-specific copy number analysis across 12 tumor types
We constructed a compendium of 2240 publically available primary tumor samples hybridized to Affymetrix 250K StyI SNP arrays, encompassing cancers arising in 12 broadly defined tissue types (Supplementary Tables S1 and S2). We employed ASCAT [48] to infer tumor purity and ploidy and derive whole-genome allele-specific copy number profiles. 81 samples (3.6%) failed ASCAT analysis, mostly due to excessive noise, and were not included in further analyses. In addition, of the 2159 cases that passed ASCAT analysis, 292 (13.5%) showed little to no copy number aberrations, and therefore tumor purity estimates can be considered less accurate. However, as we observed clear driver homozygous deletions in known tumor suppressor *CDKN2A* in 6 of these cases, we opted to include these samples in all further analyses, except analyses of tumor purity. As expected, the proportion of non-aberrant samples varied considerably between tumor types, with leukemia showing the highest proportion of non-aberrant cases (55.6%), followed by lymphoma (22.5%) and sarcoma (20.1%) (Supplementary Fig. S1).

Most cancer types showed extensive variation in purity and ploidy (Fig. 5.1A, B). The majority of leukemias and lymphomas presented with near-perfect diploidy. Multiple myeloma and hepatocellular carcinoma showed prominent diploid and hyperdiploid peaks as well as some tetraploid cases. All other tumor types had both a significant proportion of approximately diploid cases and a significant proportion of aneuploid cases showing a wide ploidy range, with esophageal carcinomas displaying the highest proportion of aneuploid cases (Fig. 5.1A, Supplementary Table S2). Individual cases from all tumor types varied extensively in purity values (Fig. 5.1B), indicating large variation in normal cell admixture between individual cancers of the same cancer type. Multiple myelomas were enriched in $CD138^+$ tumor cells during sample collection and indeed showed high purity values. Of the non-purified cancers, the median purity was the lowest in lung cancer (46%) and the highest in brain cancers (70%). These findings are broadly consistent with previous ploidy measurements

from chromosome counts [2], as well as with a previous SNP array study on an overlapping series of samples [47].

Genomic architecture and genetic intratumor heterogeneity across cancer types

We evaluated genomic architecture and genetic heterogeneity across cancer types using several complementary measures (Fig. 5.1C-F). We captured the amount of somatic copy number change by calculating an 'aberration index', the fraction of the genome showing copy number changes (Fig. 5.1C). Similarly, we calculated an 'LOH index' that quantifies the fraction of the genome showing loss-of-heterozygosity (LOH) (Fig. 5.1D). To measure local complexity, we used the previously developed Complex Arm-wise Aberration Index (CAAI) [100] (Fig. 5.1E) that captures short windows showing highly complex rearrangements such as firestorms [99]. Finally, we developed a novel method to assay whether a copy number aberration is clonal or subclonal and we calculate a 'heterogeneity index', using the fraction of genomic aberrations that are subclonal as a measure of genetic intratumor heterogeneity (Fig. 5.1F).

Our results indicate considerable differences in the extent and manner in which a cancer's genome is shaped by copy number changes, both within and across cancer types. Each cancer type includes cases with few aberrations and thus low complexity by any measure, as well as considerably more complex cases. As expected, we observed significant differences in the distributions of the aberration index values between different tumor types (Fig. 5.1C). Esophageal carcinomas are the most aberrant (on average, 44.0% of the genome is affected by copy number changes), while leukemias show very few copy number changes (2.9%). The LOH index broadly followed the aberration index (Fig. 5.1D), but highlights subtle differences in the extent to which different tumor types employ LOH to shape their genomes. Although multiple myelomas and sarcomas show similar aberration index values, multiple myelomas display a markedly lower fraction of LOH (6.6% vs. 17.7%). Similarly, hepatocellular carcinomas exhibit higher aberration index values than colorectal and renal cell carcinomas, but lower frequencies of LOH. The maximum CAAI score, measuring local complexity, also highlights considerable differences between cancer types (Fig. 5.1E). Focal regions of high complexity are very common in breast cancers, esophageal cancers and sarcomas, but very rare in liver and kidney cancers and in hematopoietic cancers. When we assayed copy number changes for clonality and subclonality, a complementary picture emerged (Fig. 5.1F). Of the solid tumors, renal cancers have the lowest fraction of subclonal copy number changes (on average, 14.7% of copy number changes are subclonal), while all other cancer types show averages above 20%, with colorectal cancers exhibiting the highest genetic heterogeneity (37.7%).

Complementary to these measures of genomic complexity and genetic intratumor

heterogeneity, we deduced frequency patterns of gains, losses and LOH along the genome, as well as their statistical significance (Supplementary Fig. S2). The frequency patterns of gains and losses essentially reproduce those described previously [95, 108] and highlight several known properties such as the frequent gain of odd-numbered chromosomes in multiple myelomas and the gain of 1q and 8q and loss of 8p, 11q and 17p in breast cancer. These patterns also roughly correspond to the aberration index patterns shown in Fig. 5.1C. The genome-wide patterns of LOH in contrast are less well studied. As expected, the frequency patterns of LOH correspond well to the frequency pattern of losses. However, a limited number of regions show considerably higher frequencies of LOH than losses, indicating these regions are prone to copy-number neutral LOH events. These regions include chromosome arm 5q in colorectal carcinoma, harboring the APC tumor suppressor, and chromosome 17 in esophageal carcinoma, harboring several known tumor suppressors, including *TP53*.

A systematic screen for homozygous deletions in primary tumors
Homozygous deletions (HDs) result in complete inactivation of gene function. Two independent hits are required to form a HD, and in addition, negative selection efficiently eliminates any HD targeting genes necessary for cell survival. Therefore, HDs are very rare in most cancers. A subset of HDs is under positive selection, as they contain a tumor suppressor gene. We therefore expect that systematic screens for HDs in large series of cancer samples represent very powerful ways to identify known and novel tumor suppressors. Such a systematic screen has so far been performed only in cell lines [96], as non-aberrant cell infiltration has historically hindered the reliable identification of HDs in primary tumors. As recent methods such as ASCAT can effectively deconvolute copy number profiles of tumor cells from those of admixed normal cells [46, 48, 45, 47] and reliably identify HDs [94], we here embarked on such a screen for HDs in primary tumors.

We detected 1884 HDs in our compendium of 2159 primary cancers (Supplementary Table S3). The median length of these HDs was 315 kb, with the 10% largest HDs encompassing 51% of the total homozygously deleted sequence. 1324 cases did not show any HDs at the resolution of our screen, and in the remaining 835 cases, on average only 2.0 Mb was found homozygously deleted per case. There were marked differences between cancer types in the number of HDs and the amount of sequence homozygously deleted. HDs are very rare in renal cancer and more common in esophageal cancer, lung cancer and sarcoma (Table 5.1).

These HDs were highly non-randomly distributed across the genome. Permutation tests showed that 967 genes were more frequently hit by HDs than expected by chance (FDR < 0.05) (Supplementary Table S4). Due to the high selectivity

of our screen, 4 HDs targeting one gene were often sufficient to be able to assign statistical significance. In total, we observed 89 distinct peaks of HDs across the genome (Fig. 5.2, Table 5.2, Supplementary Fig. S3), over these 967 significant genes or over intergenic regions hit by at least 5 HDs. As has been shown in cell lines [96], recurrent HDs in cancer genomes occur either over tumor suppressor or over regions of increased genomic instability called fragile sites.

We observed HDs over 11 established tumor suppressors (Table 5.2, Fig. 5.3). *CDKN2A* was evidently the dominant homozygously deleted tumor suppressor, with 108 HDs across 9 cancer types. We detected 16 HDs over *PTEN* in multiple cancer types and 5 HDs over *NF1*. HDs of *RB1* were most frequently found in sarcomas. Similarly, *CYLD* HDs are exclusive to multiple myeloma, whereas *SMARCB1* HDs are exclusive to brain cancers. We further observed 4 HDs over *TP53* (3 sarcomas), 6 HDs over *MAP2K4* (3 breast carcinomas) and 6 HDs over *CDH1* (5 lung cancers). Interestingly, the 4 HDs over *BRCA2* included 2 sarcomas, a cancer type not previously linked to *BRCA2* mutations. In addition, we observed 6 HDs over the *VHL* tumor suppressor, known to be very frequently hit by truncating mutations in renal cancer. Strikingly however, 4 of these HDs were found in lung cancers. Although we cannot exclude involvement of other genes, most notably the nearby tumor suppressor *FANCD2*, the observation that *VHL* is the only gene targeted by all 6 HDs suggests HDs of *VHL* may be driver events in lung cancer.

The 89 regions showing frequent HDs also include the T-cell receptor locus, as well as both the immunoglobulin heavy and light chain loci. HDs of these regions are predominantly found in hematological cancers. These represent somatic HDs in precursors of normal T and B lymphocytes that later developed into tumor cells. These HDs per se most likely do not play a role in oncogenesis.

HDs also occur frequently over fragile sites, chromosomal regions showing high rates of breakage. We observed 13 peaks of HDs over known (named) fragile sites (Fig. 5.4A, Table 5.2). As previously described [96], the pattern of HDs compared to small hemizygous deletions is distinctive in fragile sites compared to regions harboring tumor suppressors. Regions harboring tumor suppressors are enriched in HDs due to positive selection. Fragile sites, on the other hand, are enriched in both homozygous and hemizygous deletions due to a local increase in genomic instability. Therefore, the ratio of HDs to small hemizygous deletions over a region can be used to discriminate fragile sites from regions harboring tumor suppressors [96]. Based on this ratio, we constructed a statistical model and calculated a posterior probability that each region with recurring HDs is a fragile site (Table 5.2). For 6 of the 11 regions containing known tumor suppressors, we obtained a significant p-value, confirming that the increased occurrence of HDs was due to positive selection. No significant positive selection for HDs could be inferred for any of the known fragile sites. Although this

statistical model is suitable to detect a sizable subset of tumor suppressors, there are limitations. Two of the established tumor suppressors, *CDH1* and *MAP2K4*, are within known fragile sites (*FRA16B* and *FRA17B* respectively). In addition, some tumor suppressors show very frequent biallelic inactivation by a hemizygous deletion combined with a point mutation, while HDs are rare. A possible reason is that HDs, which often also affect neighboring genes or regulatory regions, are selected against, whereas point mutations may be better tolerated. For these genes, our statistical model of positive selection for HDs may not be applicable. The most notable example of this is *TP53*, which is only rarely inactivated by HDs, while point mutations of *TP53* occur frequently in most cancer types.

Identification of fragile sites and regions of (sub)telomeric instability
Among the remaining 64 regions showing recurring HDs, we identified 37 fragile sites. These include 22 fragile sites not involving telomeres (Table 5.2, Fig. 5.4B), a sizable subset of which has been described previously [96], as well as 15 (sub)telomeres showing enrichment of both homozygous and hemizygous deletions (Table 5.2, Fig. 5.4C). Hemizygous and homozygous deletions in these latter regions represent the signature of a prior telomere crisis triggering genomic instability. Telomere attrition in dividing cells without telomerase expression gradually leads to removal of the telomeres and later of subtelomeric sequences. These naked chromosome ends are often resolved by end-to-end chromosome fusions, with loss of subtelomeric sequences [109]. The resulting dicentric chromosome is mitotically unstable and initiates a number of breakage-fusion-bridge cycles in the subsequent cell divisions, creating exponentially increasing numbers of subclones [110, 111]. These serve as a substrate for natural selection, enabling significant steps towards oncogenesis within a limited time. Hence, although the observed homozygous and hemizygous deletions at these subtelomeric sites are by themselves most likely not oncogenic, they represent archeological traces of likely oncogenic events initiated by telomere loss.

These breakage-fusion-bridge cycles initiated by telomere attrition will often cause local complex copy number aberrations. We indeed observed that tumors with hemizygous or homozygous (sub)telomeric losses show higher local complexity compared to cases that do not show such (sub)telomeric losses (Fig. 5.4D), and this difference was statistically significant ($p < 10^{-15}$).

Identification of candidate tumor suppressors
In addition to the 11 established tumor suppressors, we identified 27 regions showing signatures of positive selection (Table 5.2). From the patterns of the HDs, the patterns of mutations in COSMIC [112] and the literature, we were able to propose candidate tumor suppressors in 20 of these regions (Fig. 5.5). One region was intergenic, and for the 6 remaining regions, the

targets remain unknown (Table 5.2). Our list of tumor suppressors includes genes already proposed to play a role in oncogenesis (e.g., *FAT1*, *MXRA5*, *MGMT*, *BIRC2/BIRC3*, *TET1* and *RAD17*), as well as novel candidate tumor suppressors (e.g., *CASP3*, *CASP9*, *SETD1B*, *CPEB*, *BAZ1A*, *KAT6B*, *USP29*, *USP34*, *LINC00662* and *KIAA1551*).

Several candidate tumor suppressors show HDs primarily in specific tumor types, a feature shared with 7 of our established tumor suppressors, adding to the evidence that these candidates are genuine cancer genes. In contrast, this tumor-type specificity is uncommon in known fragile sites (Table 5.2). The *FRA16B/CDH1* and *FRA17A/MAP2K4* loci, two known fragile sites that contain established tumor suppressors, advocate that tumor-type specificity, even in the absence of a signature for selection, may be used as a suggestion that a region contains a tumor suppressor. The fragile site containing the *USP44* gene is exclusive to lung cancers, and *USP44* knockout mice have recently been shown to spontaneously form lung tumors [113]. The known fragile site FRAXC contains the gene *DMD*, encoding dystrophin involved in Duchenne muscular dystrophy, and is homozygously deleted specifically in sarcomas. We observed only 25 HDs in 73 renal cancers, 3 of which overlap a fragile site over the *PTPRD* gene, a pattern that is highly significant ($p = 3.0 \times 10^{-6}$ by permutation analysis). In total, we were able to propose a candidate tumor suppressor for 6 fragile sites (Table 5.2). There are however other possible explanations for tumor-type specificity in the absence of clear positive selection. Most notably, copy number changes may be caused by mechanisms that are correlated with gene expression, such as transcription-associated recombination [114], which may explain e.g. the high rate of HDs affecting *DMD* in sarcomas. Therefore, copy number analyses alone are not sufficient to conclusively establish candidate genes without a clear signature of selection as tumor suppressors.

Our systematic screen for tumor suppressors in primary tumors discovered 11 established tumor suppressors, 20 candidate tumor suppressors with clear signatures of selection and 6 candidate genes with tumor-type specificity but no clear signature of selection, as well as several regions where no clear candidate gene could be assigned. To our knowledge this is the first such screen performed on primary tumors, preceded only by a smaller study on cancer cell lines [96]. Cell lines however often do not completely represent the spectrum of primary tumors. We assayed our regions of HDs in the cancer cell line panel [96]. This indicated that cancer cell lines compared to primary tumors show a considerably higher rate of HDs across *CDKN2A*, as well as of a small number of other genes and a limited number of fragile sites. Many other tumor suppressors are rarely or never found homozygously deleted in cancer cell lines (Supplementary Table S5).

Discussion

Distinct genomic architecture and evolutionary paths in different cancer types

Our results highlight important biological properties of different cancer types. As expected, renal cancers show relatively simple copy number profiles. Chromosome arm 3p is nearly always lost, and focal complex events are almost never found. Unexpectedly, however, we found that at the copy number level, renal cancer is predominantly a clonal disease, in sharp contrast with all other solid cancer types we profiled. However, a recent study that sampled multiple sites of renal cancer indicated that point mutations reveal quite complex subclonal architectures, with interesting patterns of convergent evolution in different subclones [104]. We suggest two possible models to explain both findings. Firstly, renal cancers may be highly geographically stratified, such that any single biopsy is clonal, but the complete tumor shows a complex pattern of subclones. Alternatively, the large majority of copy number changes may occur early in the development of a renal cancer, before the emergence of the most recent common ancestor from which all cancer cells are derived, while in contrast, point mutations occur throughout the lifetime of the tumor, and in later stages lead to subclonal variegation and ongoing selection and evolution.

The genomes of colorectal cancers are only slightly more complex than those of renal cancers, although the regions affected are quite different. However, many more colorectal cancers show focal complex events, and most colorectal cancers show intricate patterns of subclonal copy number changes.

Esophageal cancers are prone to copy number neutral LOH on chromosomes 9, 13, 17 and X. They have highly aberrant copy number profiles with frequent complex focal events and high genetic heterogeneity. This indicates that genomic instability initiates early in esophageal cancer development, before emergence of the most recent common ancestor, and continues driving evolution and selection throughout the lifetime of the tumor.

Although we observed extensive variation between individual tumors of the same cancer type, these patterns highlight the unique nature of each tumor type, suggesting that each tumor type favors a particular evolutionary path.

Advancing the landscape of tumor suppressors

This is the first study that systematically assays HDs in primary tumors. Our screen detected 11 established tumor suppressors, 3 immune regions, 13 known (named) fragile sites, 22 additional intrachromosomal fragile sites, 15 regions of telomeric instability and 27 regions showing clear signatures of positive selection for HDs. For 20 of the latter regions, we could assign a clear candidate tumor suppressor. Finally, through patterns of tumor-type specificity, we propose 6

further potential tumor suppressors in fragile sites.

Our results further establish several genes already proposed in the literature to be tumor suppressors. We found HDs encompassing the protocadherin *FAT1* in ovarian carcinoma, lung cancer and sarcoma. *FAT1* was recently proposed as a tumor suppressor [115] through the observation of HDs and point mutations in multiple cancer types. Functional validation showed that loss of *FAT1* function drives oncogenesis through activation of the Wnt pathway [115]. We identified 5 HDs in multiple myelomas encompassing *BIRC2* and *BIRC3*, encoding the ubiquitin ligases cIAP1 and cIAP2 respectively. HDs affecting these neighboring genes have been observed previously, and linked to aberrant NF-κB activation, similar to HDs of *CYLD* in multiple myeloma [116]. We observed 5 HDs of *MXRA5*, a poorly-studied gene that was recently suggested to be frequently mutated in non-small cell lung carcinoma [117].

Several peaks of HDs highlight candidate tumor suppressors that are related to known cancer genes. We found HDs affecting *CASP3* and *CASP9*, or regions close by, in various cancer types. Both genes are members of the caspase family that plays a key role in executing apoptosis. We previously established that *CASP8* is a tumor suppressor gene in breast cancer [94]. Both *CASP8* and *CASP9* induce apoptosis through cleavage of *CASP3*. Therefore, complete inactivation of any of these three caspases may abrogate apoptosis.

We detected HDs encompassing *TET1* in 7 lung cancers and one sarcoma, all larger than 1 Mb. *TET1*, *TET2* (a known tumor suppressor in myelodysplastic syndromes) and *TET3* are epigenetic regulators that have been suggested to catalyze the initial step of active DNA demethylation [118]. Mutations of *TET1* have been reported recently in leukemia [119], but an involvement in lung cancer has so far not been described. Another peak region of HDs in lung cancer is located close to *SETD1B*, a member of a family of histone methyltransferases that also include *SETD2*, known to be involved in renal carcinoma. *SETD1B* specifically methylates *H3K4*, thereby playing a role in epigenetic control of transcription. We identified several large and smaller HDs targeting *ARHGEF10*, a Rho guanine nucleotide exchange factor that has been proposed as a tumor suppressor [120]. Interestingly, we also observed HDs affecting its close relative ARHGEF10L, suggesting that both genes may be tumor suppressors.

We discovered several HDs affecting *MGMT*, encoding the DNA-repair enzyme O^6-methyl-guanine-DNA methyltransferase that is found frequently inactivated in gliomas (as well as multiple other cancer types) by promoter methylation, and that is an important marker for therapy response [121]. Although *MGMT* is extensively studied, to our knowledge we are the first to report that HDs represent an alternative mechanism to inactivate this DNA repair gene in cancer cells, confirming it is a tumor suppressor.

We found HDs of *RAD17* in 5 lung cancer and one sarcoma. *RAD17* is involved in DNA damage repair and acts as a cell cycle checkpoint gene. An RNAi screen indicated that *Rad17* acts as a haploinsufficient tumor suppressor in a mouse lymphoma model [122]. A third gene involved in DNA damage repair, *BAZ1A*, was also found homozygously deleted in lung cancer and sarcoma. *BAZ1A*, also known as *ACF1*, has been recently shown to be involved in double-strand break repair [123], but has so far not been linked to cancer.

Some of the peaks of HDs may drive oncogenesis through other mechanisms than inactivation of a tumor suppressor gene. One sarcoma and 3 lung cancers showed HDs of *ABL1*, encoding a tyrosine kinase constitutionally activated in myeloproliferative neoplasms. As it is highly unlikely that *ABL1* is inactivated by HDs and given that we observe several breakpoints within *ABL1*, we hypothesize that a potential underlying mechanism may be the formation of a constitutionally activated ABL1 protein through gene fusions or other means. We observed several small intragenic HDs within *GPC5*, encoding glypican 5. These potentially represent partial loss-of-function, or gain-of-function mutations removing one or a few exons of this 1.5 Mb gene.

Several peaks of HDs point to genes not previously implicated as tumor suppressors. In several cancer types, we found HDs of the intergenic non-protein coding RNA 662 (*LINC00662*), a non-protein coding gene in a gene-poor region for which nearly nothing is known. We observed 5 large HDs in lung cancer where the uncharacterized gene *KIAA1151* (also known as *c12orf15*) appears to be the only gene targeted by all HDs. We also found 5 HDs targeting *KAT6B* (also known as *MYST4*) in various cancer types. *KAT6B* encodes a histone acetyltransferase and has been shown to be recurrently translocated to the *CREBBP* locus in acute myeloid leukemia [124]. Several truncating mutations in cancers have been found in this gene that, combined with our HDs, implicate this gene as a tumor suppressor. Another peak of HDs targets *CPEB3*, mostly in lung cancer. *CPEB3* encodes a sequence-specific RNA-binding protein that has been shown to transcriptionally repress *EGFR* [125]. Therefore, our results suggest that, in addition to amplification of *EGFR*, lung cancers can also adopt HDs of *CPEB3* as an alternative mechanism to overexpress *EGFR*.

We observed four peaks of HDs around the ubiquitin specific proteases *USP34*, *USP29*, *USP44* and the neighboring genes *USP8* and *USP50*, predominantly in lung cancer. These are all family members of a large group of deubiquitinating enzymes that prevent protein degradation by the proteasome. The range of targets of each of these ubiquitin specific proteases are poorly described although, interestingly, *p53* has been described as a target of *USP29* [126] and *USP34* has been linked to Wnt signaling [127]. In addition, *USP44* was recently found to regulate the mitotic cell cycle checkpoint and *Usp44* knockout mice spontaneously formed tumors, particularly in the lung [113]. Therefore,

our results further establish USP44 as a tumor suppressor in lung cancer, and implicate other ubiquitin specific proteases in oncogenesis as well, particularly in lung cancer.

We extended this series of candidate tumor suppressors showing signatures of positive selection for HDs with 6 potential tumor suppressors within fragile sites that exhibit tumor-type specificity. These include the candidate tumor suppressors *USP44* and *USP50/USP8* mentioned above. In addition, the *LSAMP* gene, observed to be homozygously deleted in sarcomas, has previously been linked to osteosarcoma [128], and the *PTPRD* gene, homozygously deleted in lung and kidney cancers has been suggested to be a tumor suppressor as well [129]. One region contained 4 large HDs in lung cancer, 3 of which overlap the known tumor suppressor *CDKN2C*. While the HDs of *CDKN2C* likely play a role in oncogenesis of these cases, due to the large size of these HDs, a more complicated picture with possible involvement of other genes cannot be excluded. Finally, we observe 7 HDs in sarcomas affecting *DMD*. Although *DMD* is the largest gene in the human genome and is part of known fragile site FRAXC, our results suggest HDs of *DMD* may play a role in sarcoma. In support of that hypothesis, a recent study observed a high incidence of rhabdomyosarcoma in dystrophin/dysferlin double mutant mice [130].

In summary, our results advance the landscape of tumor suppressors, identify several novel players and propose a role for several ubiquitin specific proteases in oncogenesis.

Methods

Copy number analysis of a compendium of 2240 primary tumor samples

SNP array studies were selected from the CaSNP database [131], the MetaCGH database [95] and from literature searches, with the aim of collecting all studies of primary tumors performed on Affymetrix 250K StyI arrays with publically available raw data. All pre-cancerous lesions and metastases were excluded. Duplicate samples and cases with multiple tumors from the same patient were identified using Pearson correlation of genotypes and subsequently removed. Samples were stratified into 12 broadly defined tumor types, each containing >50 samples. Other tumor types with less than 50 samples available were excluded. Our final dataset included 2240 unique primary tumor samples from 28 different studies (Supplementary Tables S1 and S2). As the samples from each cancer type in our compendium are derived from multiple centers, we expect that cross-tumor comparisons are generalizable and do not reflect local ascertainment biases and referral procedures.

LogR and B allele frequency (BAF) values were obtained from CEL files using PennCNV-Affy [30]. GRCh37/hg19 probe annotation (version 32) was obtained from the Affymetrix website (http://www.affymetrix.com). GC wave correction was performed as previously described [25]. Copy number analysis was performed using ASCAT [48] version 2.2, using default parameters for tumor samples without matched normal data. In order to remove any adverse influence of germline CNVs, all SNPs within known CNVs were removed prior to ASCAT analysis. Positions of CNVs were obtained from the Database of Genomic Variants [132], version hg19.v10, from which all CNVs identified on Affymetrix 500k and SNP6 platforms were selected. This reduced the number of analyzed probes from 228,586 to 208,786.

Indexes of genomic architecture and genetic heterogeneity
The aberration index indicates the fraction of the genome showing copy number changes in a particular sample. In order to account for endoreduplication events and large-scale tumor aneuploidy, total copy number values for each locus were compared to the tumor's ploidy rather than to the normal state of 2 copies, as previously described [48]. Loci with copy number values differing by at least 0.6 from the tumor's ploidy were considered aberrant, and the aberration index was calculated as the fraction of probes that were aberrant according to that definition. The LOH index of a sample indicates the fraction of the genome showing LOH. All probes with predicted germline heterozygosity were selected and the fraction of those probes that show LOH in the tumor cells (i.e. zero copies for at least one allele) was calculated. The Complex Arm-wise Aberration Index, which captures local complexity, was calculated as previously described (16), using the LogR values after GC wave correction.

The heterogeneity index expresses the fraction of alterations that are subclonal. This procedure builds upon the tumor's purity $\boldsymbol{\rho}$ and ploidy $\boldsymbol{\Psi_t}$ as derived from ASCAT, the LogR and BAF values $\mathbf{r_i}$ and $\mathbf{b_i}$ as well their segmentation by the ASPCF algorithm [48, 49]. The allele-specific copy number values for a probe i can be expressed as [48]:

$$n_{M,i} = \frac{\rho - 1 + 2^{\frac{r_i}{\gamma}}(1 - b_i)(2(1-\rho) + \rho\Psi_t)}{\rho} \tag{5.1}$$

$$n_{m,i} = \frac{\rho - 1 + 2^{\frac{r_i}{\gamma}} b_i(2(1-\rho) + \rho\Psi_t)}{\rho} \tag{5.2}$$

$\mathbf{n_{M,i}}$ and $\mathbf{n_{m,i}}$ represent the copy number values for the allele with the higher and lower copy number respectively. BAFs were mirrored around 0.5 [48], such that $\mathbf{b_i < 0.5}$. $\boldsymbol{\gamma}(= 0.55)$ is a platform-specific constant [48] whose value we determined using an X-chromosome titration series obtained from the Affymetrix

website. To determine if a segment S shows a clonal or subclonal copy number state, the following rules were used. If the segmented BAF value equals 0.5, both alleles are present in equal proportions, and hence the segment is considered clonal. If the segmented BAF value differs from 0.5, the values $\mathbf{n_{M,i}}$ and $\mathbf{n_{m,i}}$ are evaluated. For each allele, a t-test is performed (comparing to the closest non-negative integer as this represents the closest clonal state), using a significance cutoff of 0.01. To account for potential inaccuracies in the estimates of $\boldsymbol{\rho}$ and $\mathbf{\Psi_t}$, segments with an average deviation from non-negative integers less than 0.2 were considered clonal. For the calculation of the heterogeneity index, all probes in segments with copy number aberrations were aggregated, and the fraction of these that are within subclonal copy number segments constitutes the heterogeneity index value.

Genomic distributions of gains, losses and LOH

Genomic distributions of gains, losses and LOH were obtained as previously described [48]. The significance of enrichment of gains, losses or LOH along the genome was evaluated by comparing the frequency in the tumor type under study to that of all other cancers. For each array probe, aberration type (gain, loss or LOH) and tumor type, a 2 by 2 frequency table was constructed, comparing the number of cases with the aberration to those without the aberration in the tumor type under study compared to all other cancers. A logistic regression model was subsequently fitted for each frequency table and the p-value of the log-odds ratio was calculated.

Identification of genomic regions with frequent HDs over fragile sites or tumor suppressors

After ASCAT analysis, HDs were straightforwardly identified as regions having zero copies of both alleles in the tumor cells. Genes showing enrichment of HDs were identified by permutation analysis, where 107 permutations were performed randomizing the positions (but not the size) of HDs. A Benjamini-Hochberg FDR < 0.05 was used as the significance threshold. Neighboring genes with significant enrichment were combined into peaks. To account for regions of frequent HDs not overlapping any gene, genomic regions exhibiting 5 or more HDs were added. The position of each peak region was defined as the region showing the largest number of HDs.

A statistical model was constructed quantifying the probability that a peak region is observed due to positive selection for HDs, based on the ratio between the number of HDs and the number of small hemizygous deletions. Small hemizygous deletions were defined as regions of LOH smaller than 1 Mb. A normal mixture model was fitted, assuming the log-transformed ratios of HDs and small hemizygous deletions follow a normal distribution:

$$Y_i \sim p_{i,FS} N_{FS}(x \mid \mu_{FS}, \sigma_{FS}) + p_{i,TS} N_{TS}(x \mid \mu_{TS}, \sigma_{TS}) \tag{5.3}$$

$\mathbf{Y_i}$ denotes the log-transformed ratio between homozygous and hemizygous deletion counts in the i-th peak region; $\mathbf{p_{i,FS}}$ and $\mathbf{p_{i,TS}}$ represent the probability that the i-th peak region harbors a fragile site or a tumor suppressor respectively. $\boldsymbol{\mu_{FS}}$, $\boldsymbol{\sigma_{FS}}$ and $\boldsymbol{\mu_{TS}}$, $\boldsymbol{\sigma_{TS}}$ represent the mean and standard deviation of the log-transformed ratios over fragile sites and tumor suppressors, respectively. The starting values for these variables were calculated based on known fragile sites and regions known to contain tumor suppressors, excluding regions that harbor both a fragile site and a tumor suppressor. $\mathbf{p_{i,FS}}$ and $\mathbf{p_{i,TS}}$ are estimated for each peak region i, and regions where $\mathbf{p_{i,FS} < 0.05}$ were considered to show significant signatures of positive selection.

Table 5.1: Distribution of homozygous deletions across cancer types.

Cancer type	**Nr of cases**	**Nr of HDs**	**HD sequence (kb)**	**Nr of HDs per case (range)**	**HD sequence per case (kb)**	**Median length of HDs (kb)**
Breast	193	138	95,284	0.72 (0-8)	494	165
Ovary	136	113	82,794	0.83 (0-6)	609	207
Colon	125	98	48,580	0.78 (0-9)	389	205
Liver	106	80	46,658	0.75 (0-11)	440	116
Kidney	73	25	6,891	0.34 (0-4)	94	97
Lung	402	535	536,890	1.33 (0-49)	1,336	404
Brain	309	138	145,431	0.45 (0-6)	471	360
Esophage	58	66	74,515	1.14 (0-6)	1,285	285
Sarcoma	244	353	425,480	1.45 (0-15)	1,744	455
Myeloma	242	113	83,946	0.47 (0-14)	347	343
Leukemia	151	99	57,756	0.66 (0-10)	382	231
Lymphoma	120	126	95,242	1.05 (0-12)	794	251
All	2,159	1,884	1,699,469	0.87 (0-49)	787	315

Table 5.2: Peaks of homozygous deletions over tumor suppressors and fragile sites, and their predicted target genes. HemiDs: small hemizygous deletions; P_{FS}: probability of fragile site ($P_{FS} < 0.05$: significant signature of positive selection for homozygous deletions).

Known tumor suppressors						
Peak region (kb)	**Nr of HDs**	**Nr of HemiDs**	**Ratio**	**P_{FS}**	**Cancer type specificity**	**Tumor suppressor**
chr3: 10,181-10,199	6	10	0.60	0.66	lung	VHL
chr9: 22,018-22,020	108	15	7.20	3.0×10^{-12}	-	CDKN2A
Chr10: 89,738-89,835	16	0	∞	0	-	PTEN
Chr13: 32,921-33,051	4	0	∞	0	-	BRCA2
Chr13: 49,036-49,089	23	1	23.00	8.0×10^{-24}	sarcoma	RB1
Chr16: 50,913-50,942	6	2	3.00	4.5×10^{-6}	myeloma	CYLD
[1]Chr16: 68,635-69,945	6	12	0.50	0.74	lung	CDH1
chr17: 7,579-7,583	4	7	0.57	0.69	sarcoma	TP53
[1]Chr17: 11,957-12,088	6	6	1.00	0.17	breast	MAP2K4
Chr17: 29,546-29,830	5	3	1.67	3.7×10^{-3}	-	NF1
Chr22: 24,193-24,483	6	5	1.20	0.056	brain	SMARCB1
T-cell receptor and immunoglobulin regions						
Peak region (kb)	**Nr of HDs**	**Nr of HemiDs**	**Ratio**	**P_{FS}**	**Cancer type specificity**	**Region**
Chr14: 22,350-22,447	5	8	0.63	0.63	hematological	TCR
chr14: 106,973-107,350	8	2	4.00	7.4×10^{-8}	hematological	IGH
Chr22: 22,914-23,139	18	9	2.00	6.0×10^{-4}	hematological	IGL
Known (named) fragile sites						
Peak region (kb)	**Nr of HDs**	**Nr of HemiDs**	**Ratio**	**P_{FS}**	**Cancer type specificity**	**Fragile site**
Chr1: 214,965-214,971	17	19	0.89	0.28	-	FRA1H
Chr2: 141,942-142,076	6	10	0.60	0.66	-	FRA2F
chr3: 60,491-60,536	6	18	0.33	0.76	-	FRA3B
chr6: 14,588-14,607	5	9	0.56	0.70	-	FRA6A

Continued

chr6: 33,055-33,071	6	21	0.29	0.72	-	FRA6H
chr6: 74,112-74,317	6	9	0.67	0.58	-	FRA6D
Chr6: 109,587-109,587	6	10	0.60	0.66	-	FRA6F
chr12: 114,597-114,628	7	21	0.33	0.76	-	FRA12C
[1]Chr16: 68,635-69,945	6	12	0.50	0.74	lung	FRA16B
Chr16: 78,718-78,818	5	6	0.83	0.36	-	FRA16D
[1]Chr17: 11,957-12,088	6	6	1.00	0.17	breast	FRA17A
chrX: 6,895-7,067	5	0	NA	NA	-	FRAXB
[1]chrX: 32,809-33,160	7	0	NA	NA	sarcoma	FRAXC

Predicted intrachromosomal fragile sites

Peak region (kb)	**Nr of HDs**	**Nr of HemiDs**	**Ratio**	**P_{FS}**	**Cancer type specificity**
[1]chr1: 51,581-53,533	4	7	0.57	0.69	lung
chr3: 73,004-73,041	4	6	0.67	0.58	-
[1]Chr3: 116,686-116,819	4	9	0.44	0.77	sarcoma
chr4: 78,168-78,214	5	8	0.63	0.63	-
chr6: 18,174-18,207	4	8	0.50	0.74	lung
[2]Chr8: 3,321-3,394	10	11	0.45	0.77	-
[2]Chr8: 6,377-6,427	7	5	0.40	0.78	-
chr8: 29,969-29,980	4	9	0.44	0.77	lung
chr8: 98,841-98,849	10	19	0.53	0.72	-
[1]Chr9: 9,423-9,640	5	8	0.63	0.63	kidney
Chr9: 121,644-121,645	8	14	0.57	0.69	colon
Chr10: 12,208-12,464	6	6	1.00	0.17	lung
[1]Chr12: 95,881-96,267	4	8	0.50	0.74	lung
Chr13: 28,251-28,260	5	8	0.63	0.63	-
Chr13: 39,536-39,537	6	9	0.67	0.58	-
Chr14: 74,073-74,548	4	10	0.40	0.78	lung
[1]Chr15: 50,655-51,110	6	8	0.75	0.47	lung
chr16: 6,740-6,764	7	22	0.32	0.75	colon

Continued

Chr16: 10,071-10,083	5	23	0.22	0.53	-
chr19: 9,058-9,066	4	12	0.33	0.76	-
Chr20: 15,079-15,107	8	11	0.73	0.50	-
chrX: 9,105-10,074	4	3	NA	NA	-

Telemeric regions showing increased instability

Peak region (kb)	**Nr of HDs**	**Nr of HemiDs**	**Ratio**	**P_{FS}**	**Cancer type specificity**
chr2: 923-941	5	14	0.36	0.77	-
chr3: 0-431	3	3	1.00	0.17	-
Chr4: 189,885-191,154	5	6	0.83	0.36	sarcoma
Chr6: 170,763-171,115	4	7	0.57	0.69	-
Chr7: 158,911-159,139	5	11	0.45	0.77	-
[2]chr8: 0-784	13	10	0.80	0.40	-
chr9: 0-876	6	15	0.40	0.78	-
chr10: 133,978-135,535	3	2	1.50	9.7×10^{-3}	-
chr13: 113,087-115,170	5	5	1.00	0.17	-
chr16: 0-318	2	8	0.25	0.64	sarcoma
chr17: 0-53	4	4	1.00	0.17	no
Chr17: 80,943-81,195	6	16	0.38	0.78	no
Chr18: 76,712-77,795	4	11	0.36	0.77	no
chr19: 0-557	3	15	0.20	0.43	no
chrX: 0-419	4	1	NA	NA	lung

Candidate tumor suppressors

Peak region (kb)	**Nr of HDs**	**Nr of HemiDs**	**Ratio**	**P_{FS}**	**Cancer type specificity**	**Candidate tumor suppressor**
chr1: 15,901-15,919	7	3	2.33	1.1×10^{-4}	-	CASP9
chr1: 17,576-17,634	4	2	2.00	6.0×10^{-4}	lymphoma	ARHGEF10L
chr2: 61,832-61,958	6	4	1.50	9.7×10^{-3}	lung	USP34
Chr4: 185,595-185,655	5	2	2.50	4.6×10^{-5}	-	CASP3
Chr4: 187,753-187,899	4	1	4.00	7.4×10^{-8}	-	FAT1
chr5: 68,396-68,692	6	3	2.00	6.0×10^{-4}	lung	RAD17

Continued

[2]Chr8: 1,771-1,938	13	1	8.00	3.9×10^{-13}	-	ARHGEF10
Chr9: 133,599-133,623	4	3	1.33	0.026	lung	ABL1
Chr10: 70,051-70,952	8	4	2.00	6.0×10^{-4}	lung	TET1
Chr10: 76,718-76,809	5	1	5.00	2.1×10^{-9}	-	KAT6B
Chr10: 93,995-94,029	5	3	1.67	3.7×10^{-3}	lung	CPEB3
chr10: 131,425-131,490	5	2	2.50	4.6×10^{-5}	-	MGMT
chr11: 101,946-102,050	6	4	1.50	9.7×10^{-3}	myeloma	BIRC2/BIRC3
Chr12: 32,145-32,242	5	4	1.25	0.042	lung	KIAA1551
chr12: 122,297-122,375	6	2	3.00	4.5×10^{-6}	lung	SETD1B
Chr13: 92,445-92,445	10	3	3.33	1.1×10^{-6}	-	GPC5
Chr14: 35,087-35,315	6	4	1.50	9.7×10^{-3}	-	BAZ1A
Chr19: 28,142-28,151	6	4	1.50	9.7×10^{-3}	-	LINC00662
Chr19: 57,986-58,026	3	1	3.00	4.5×10^{-6}	-	USP29
chrX: 3,219-4,122	5	0	NA	NA	-	MXRA5
chr8: 34,285-34,296	9	6	1.50	9.7×10^{-3}	-	intergenic
Chr3: 185,442-185,529	4	0	∞	0	-	??
chr4: 39,084-39,152	12	6	2.00	6.0×10^{-4}	lung	??
chr4: 83,681-83,682	4	0	∞	0	lung	??
Chr13: 20,337-20,451	4	2	2.00	6.0×10^{-4}	-	??
Chr13: 95,391-95,460	4	3	1.33	0.026	myeloma	??
Chr14: 27,180-28,038	4	2	2.00	6.0×10^{-4}	-	??
[1]chr1: 51,581-53,533	4	7	0.57	0.69	lung	CDKN2C
[1]Chr3: 116,686-116,819	4	9	0.44	0.77	sarcoma	LSAMP
[1]Chr9: 9,423-9,640	5	8	0.63	0.63	kidney	PTPRD
[1]Chr12: 95,881-96,267	4	8	0.50	0.74	lung	USP44
[1]Chr15: 50,655-51,110	6	8	0.75	0.47	lung	USP8/USP50
[1]chrX: 32,809-33,160	7	0	NA	NA	sarcoma	DMD

[1]Fragile sites that show a (candidate) tumor suppressor are shown twice in the table.

[2]The region 1-8Mb on chromosome 8 contains 4 peaks. In order to separate the different regions, 5 large homozygous deletions (> 5 Mb) overlapping all 4 peaks are not included for calculation of the ratio of homozygous deletions compared to small hemizygous deletions.

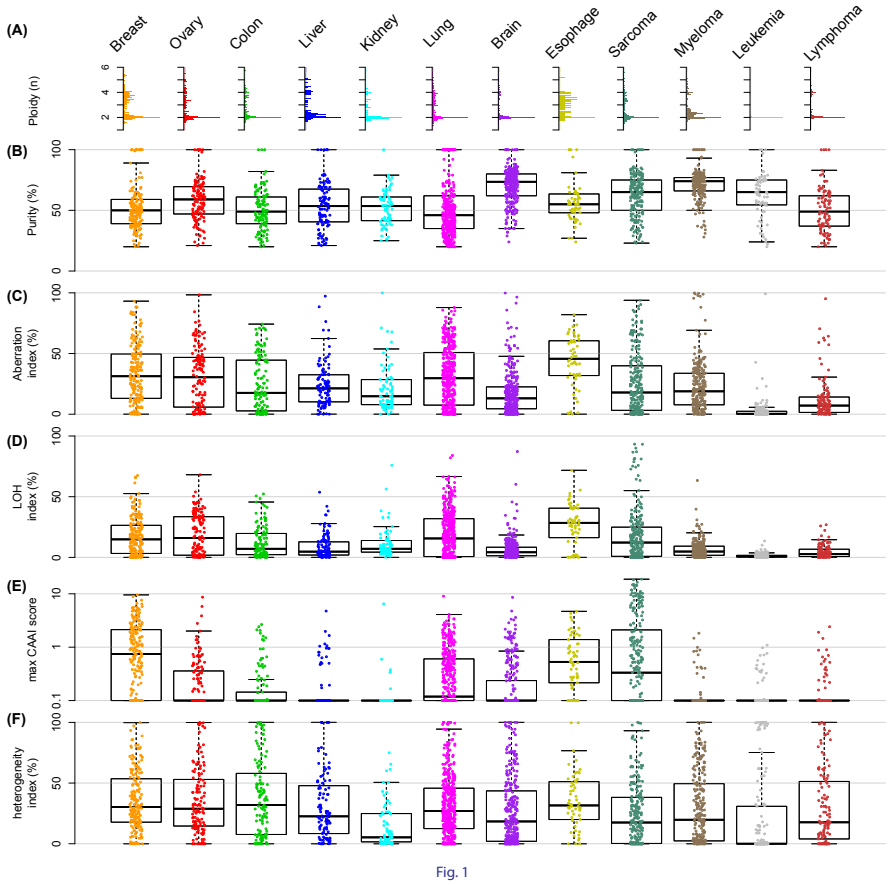

Figure 5.1: Genomic architecture and genetic intra-tumor heterogeneity across 12 tumor types. (A) Tumor ploidy, (B) tumor purity (excluding nonaberrant samples), (C) aberration index values (fraction of the genome showing copy number aberrations, expressed as a percentage of the genome), (D) LOH index values (fraction of the genome showing LOH), (E) maximum Complex Arm Aberration Index (CAAI) values, measuring local complexity (values below 0.1 are rounded up to 0.1), (F) heterogeneity index values (quantifying the fraction of genomic aberrations that are inferred to be subclonal, expressed as a percentage of the sequence showing copy number aberrations shown in (C)). Cancers were stratified into 12 broadly defined tumor types: breast carcinoma, ovarian carcinoma, colorectal carcinoma, hepatocellular carcinoma, renal cell carcinoma, lung adenoma and carcinoma, brain cancer, esophageal carcinoma, sarcoma, multiple myeloma, leukemia and lymphoma.

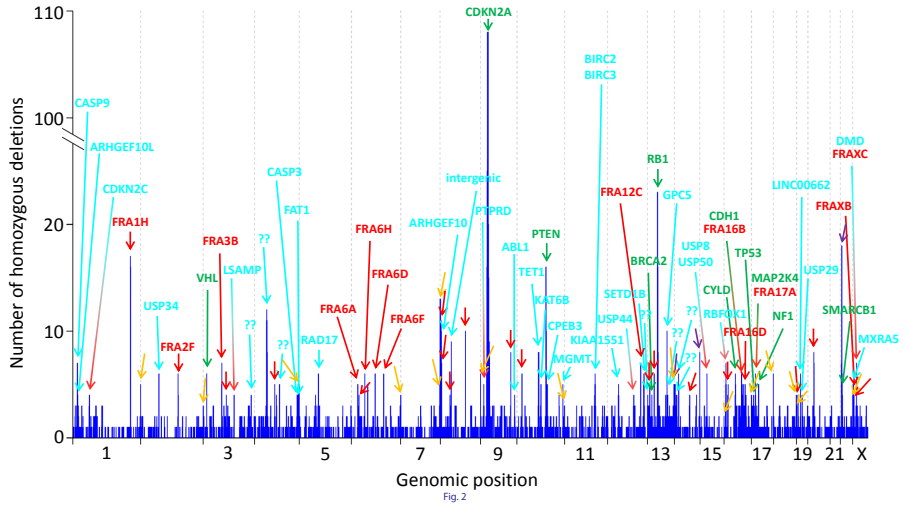

Figure 5.2: Genome-wide frequency pattern of homozygous deletions across a compendium of 2159 cancers. 89 significant peak regions are annotated as established tumor suppressor genes (green), candidate tumor suppressor genes (light blue), T-cell receptor and immunoglobulin loci (purple), known (named) and predicted intra-chromosomal fragile sites (red) and telomeric regions with increased instability (orange).

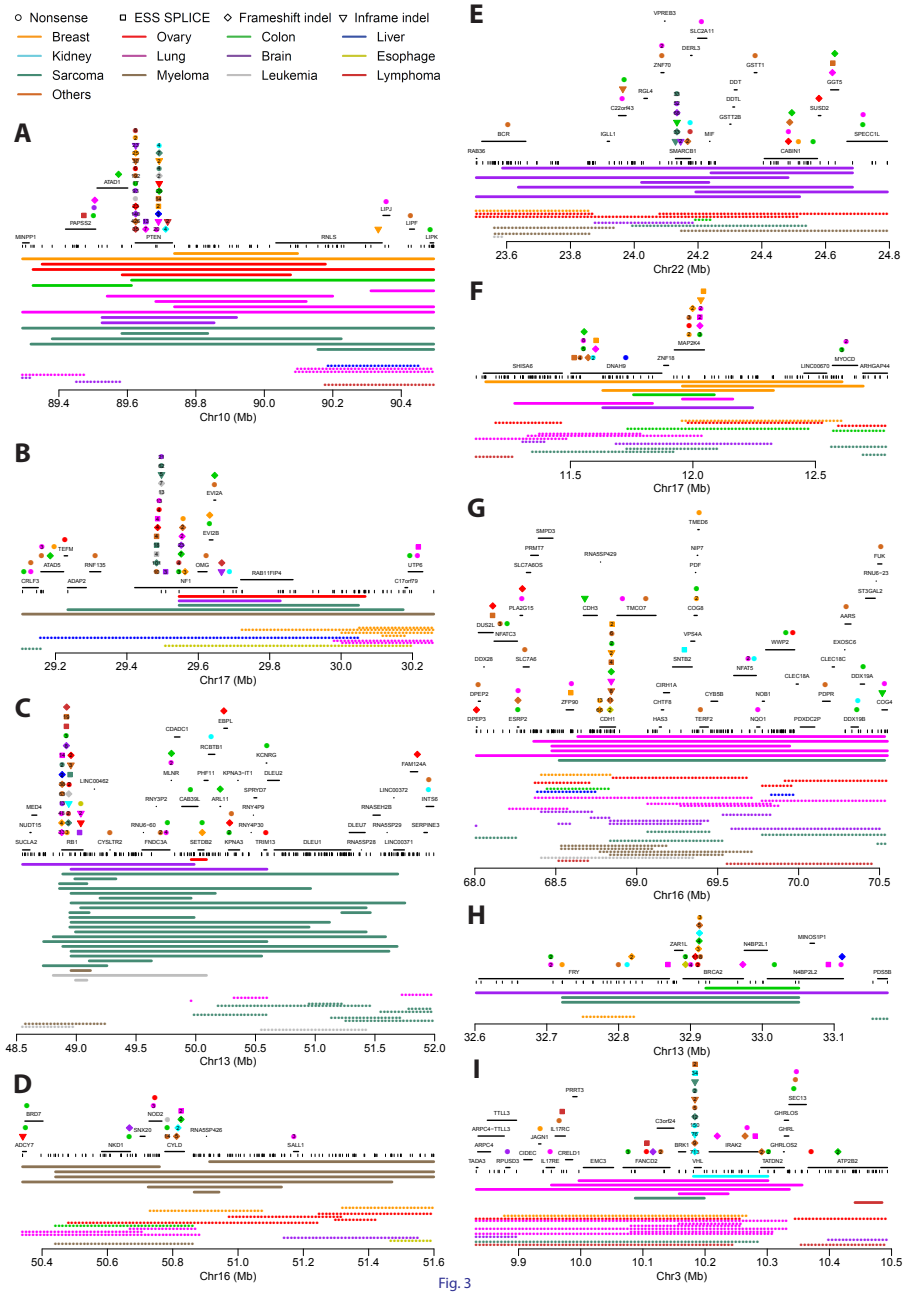

Fig. 3

Figure 5.3: Homozygous deletions targeting 9 known tumor suppressors. Positions of genes are indicated (black lines), as well as truncating mutations annotated in COSMIC, colored according to tumor type and symbols according to mutation type (nonsense, essential splice site, frame-shift insertion or deletion, in-frame insertion or deletion). When multiple somatic mutations in the same tumor type are annotated close together, their numbers are shown. Homozygous deletions are shown as bold lines and small hemizygous deletions as dotted lines, both colored according to tumor type. (A) *PTEN*, (B) *NF1*, (C) *RB1*, (D) *CYLD*, (E) *SMARCB1*, (F) *MAP2K4*, (G) *CDH1*, (H) *BRCA2*, (I) *VHL*.

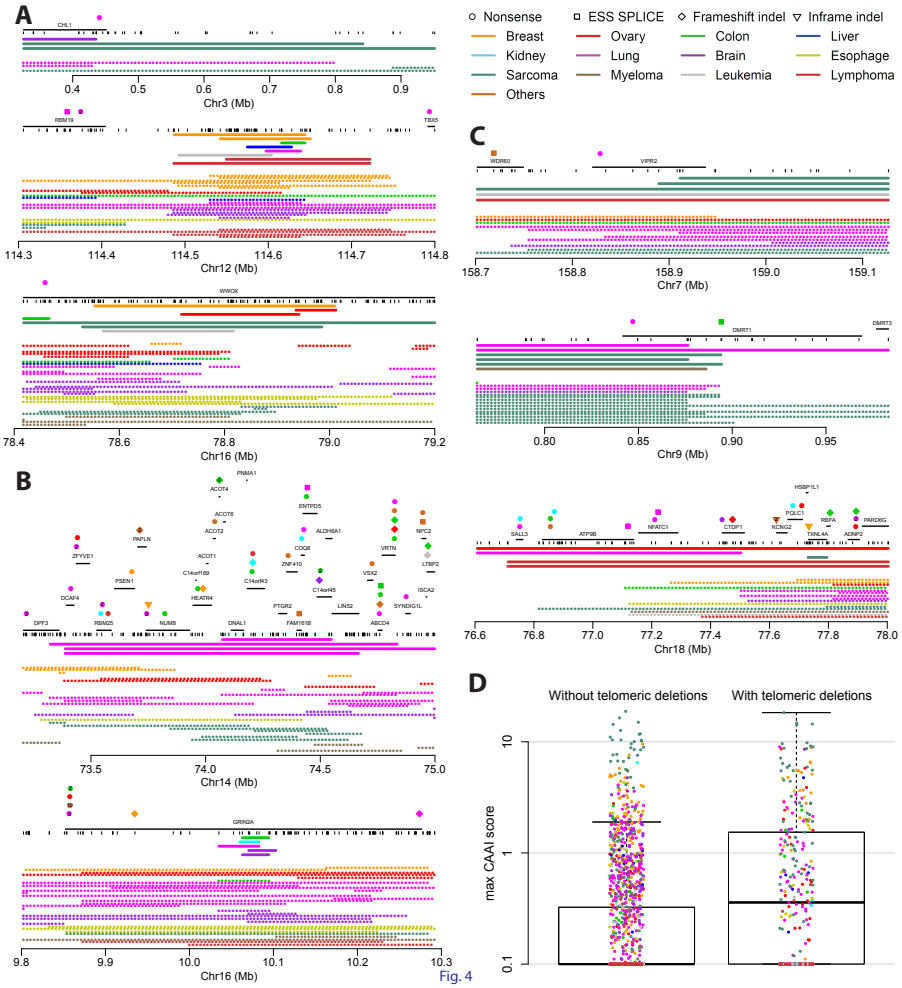

Fig. 4

Figure 5.4: Homozygous and hemizygous deletions over regions of increased genomic instability. Positions of genes are indicated (black lines), as well as truncating mutations annotated in COSMIC, colored according to tumor type and with symbols according to mutation type. When multiple somatic mutations in the same tumor type are annotated close together in COSMIC, their numbers are shown. Array probe positions are depicted below the genes. Homozygous deletions are shown as bold lines and small hemizygous deletions as dotted lines, both colored according to tumor type. (A) Three known (named) fragile sites FRA3B, FRA12C and FRA16D, (B) two additional regions predicted to be intrachromosomal fragile sites, (C) three (sub)telomeric regions showing increased instability. (D) Maximum CAAI scores, stratifying tumors into those with (homozygous or hemizygous) telomeric deletions, confirming that cases with telomeric deletions have more focal complex copy-number alterations ($p < 10^{-15}$).

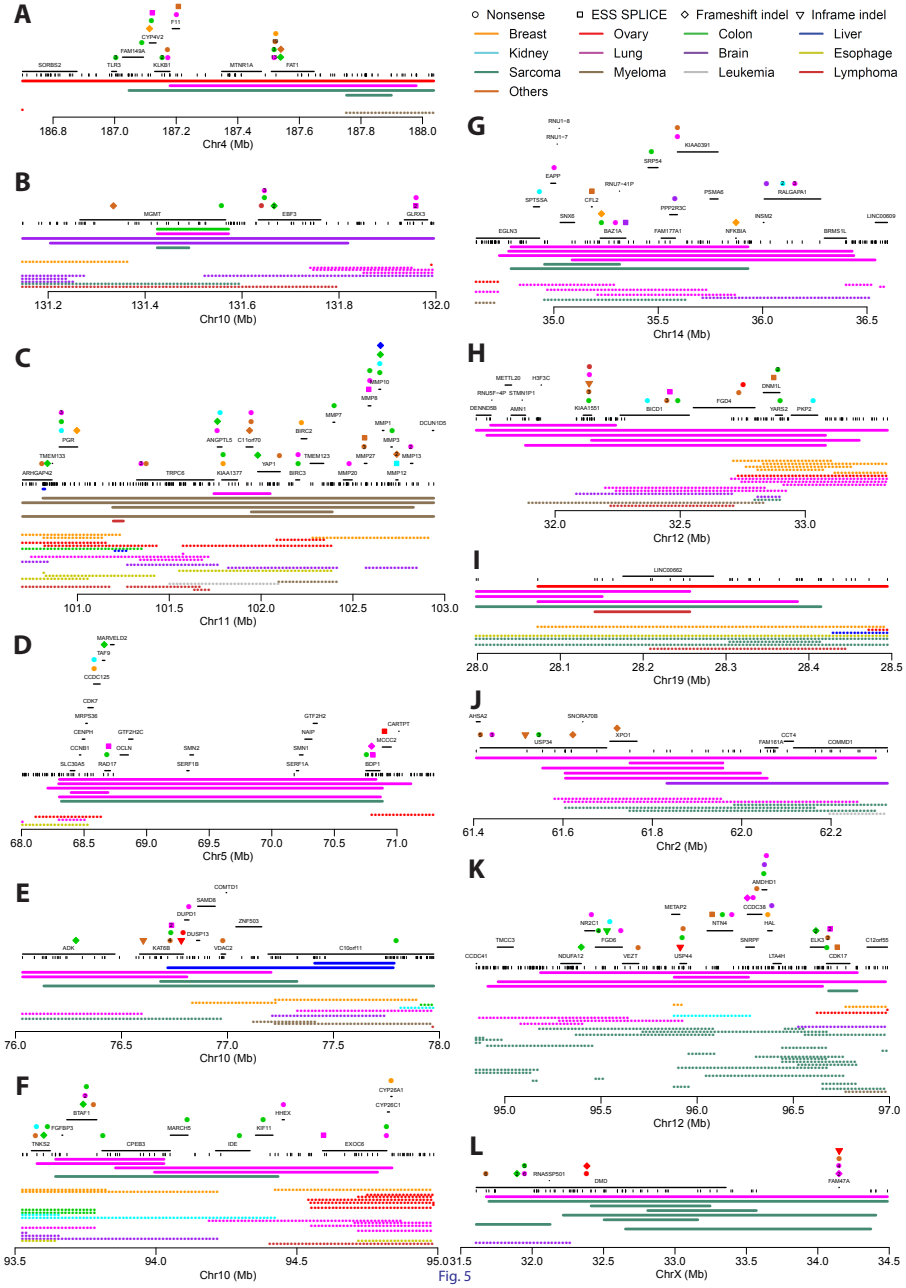

Fig. 5

Figure 5.5: Examples of candidate tumor suppressors identified in this study. Positions of genes are indicated (black lines), as well as truncating mutations annotated in COSMIC, colored according to tumor type and with symbols according to mutation type. When multiple somatic mutations in the same tumor type are annotated close together in COSMIC, their numbers are shown. Homozygous deletions are shown as bold lines and small hemizygous deletions as dotted lines, both colored according to tumor type. (A) *FAT1*, (B) *MGMT*, (C) *BIRC2/BIRC3*, (D) *RAD17*, (E) *KAT6B*, (F) *CPEB3*, (G) *BAZ1A*, (H) *KIAA1551*, (I) *LINC00662*, (J) *USP34*, (K) *USP44*, (L) *DMD*.

Acknowledgements

This research is supported by project grant G.0687.12N (genetic intra-tumor heterogeneity) from the Research Foundation-Flanders, as well as GOA MaNet, KUL PFV/10/016 SymBioSys, iMinds 2013, Hercules III PacBio RS, IWT TBM Haplotyping, Action BM1006: NGS Data analysis network. PVL is supported by a postdoctoral research fellowship of the FWO. PJC is personally funded through a Wellcome Trust Senior Clinical Research Fellowship. JJP is supported in part by grants from the National Science Foundation (Grant OISE - 1129076) and the Gordon and Betty Moore Foundation. SN is a post-doctoral fellow of the Norwegian Cancer Association (PK01-2007-0356). HGR is partly funded from the Norwegian Cancer Association. We thank Sam Behjati for valuable comments on the manuscript.

Supplementary data

Supplemental Figure S1 - Overview of aberrant and nonaberrant samples after ASCAT analysis across 12 cancer types, and cases that failed ASCAT analysis.

Supplementary Figure S2 - Genome-wide profiles of gains, losses and LOH across 12 tumor types. The frequency of gains (red, upwards), losses (green, downwards, foreground) and LOH events (blue, downwards, background) are plotted along the genome (chromosome 1 - chromosome X) across 12 tumor types. For each tumor type, the local significance of enrichment of gains (red), losses (green) and LOH (blue) is calculated compared to all other tumor types using a logistic regression model, and plotted along the genome in 3 bars below the frequency profiles.

Supplementary Figure S3 - Eighty-nine peak regions of frequent homozygous deletions. Positions of genes are indicated (black lines), as well as truncating mutations annotated in COSMIC, colored according to tumor type and with symbols according to mutation type. When multiple somatic mutations in the same tumor type are annotated close together in COSMIC, their numbers are shown. Array probe positions are depicted below the genes. Homozygous deletions are shown as bold lines and small hemizygous deletions as dotted lines, both colored according to tumor type. (A) 11 established tumor suppressors, (B) 3 immune regions, (C) 13 known (named) fragile sites (D) 22 predicted intrachromosomal fragile sites, (E) 15 telomeric regions showing increased instability, (F) 20 regions showing signatures of positive selection for homozygous deletions and their predicted target genes, (G) one intergenic region and 6 other regions showing signatures of positive selection but without

a clear target gene, (H) 6 fragile sites containing candidate tumor suppressors, predicted due to tumor-type specificity of homozygous deletions. Note: fragile sites that contain (established or candidate) tumor suppressor are shown twice, as they fall in two categories.

Supplementary Table S1 - Overview of studies included in our compendium of 2240 unique primary cancer samples.

Supplementary Table S2 - Samples included in this study, results of ASCAT analysis (aberrant, non-aberrant, failed; tumor purity and ploidy), genomic architecture and intratumor heterogeneity analysis (aberration index, LOH index, max CAAI score and heterogeneity index values), and summary statistics of homozygous deletions identified per sample.

Supplementary Table S3 - List of homozygous deletions identified in this study.

Supplementary Table S4 - Results of permutation analysis to identify genes showing enrichment of homozygous deletions.

Supplementary Table S5 - Comparison of frequencies of homozygous deletions over the 89 peak regions identified in this study, between a series of 940 cancer cell lines and our series of 2159 primary tumors.



Chapter 6

Conclusions and Perspectives

This thesis describes copy number analysis of DNA with the help of array technologies in two types of human genomes: single-cell genomes and cancer genomes. These different data sources have distinct signal patterns and thus require specific computational approaches to analyze them.

6.1 Single-cell genome analysis

6.1.1 Conclusion

Copy number alteration analysis in single blastomere cells is important to investigate chromosomal aberrations in early stages of human embryos. Previously, the detection of chromosomal aberrations relied on probe-specific FISH technologies targeting a limited number of chromosomal loci. Thus, a comprehensive view of the chromosomal constitution was lacking. The recent development of array-based technologies enables screening of aberrations in all blastomere chromosomes. Different from the conventional copy number analysis of genomic DNA extracted from multiple cells, single-cell genome research has difficulties with genome artifacts caused by the amplification of single-cell DNA. These artifacts show similar patterns to some of the real chromosomal aberrations. Thus, specific computational approaches are needed to separate real aberrations from genome artifacts. After exploring genome artifact patterns, we developed a channel-clone normalization approach to reduce artifacts present in single-cell genomes. The performance of the channel-clone technique was evaluated in advance using simulated array CGH data and validated EBV-

transformed single cells with known karyotypes. Our results show that this method outperforms other existing methods which are originally developed for processing unamplified DNA data. Applying our channel-clone normalization to array data from cleavage-stage embryos revealed a high mosaicism rate. Structural aberrations are not restricted to poorly developing embryos, but are also common in early stage embryos of good-quality.

Among all single-cell genome array studies, the highest resolution is achieved by the Bluegnome platform using amplified DNA resulting from PCR-based Sureplex technology. Although this Bluegnome protocol achieves a higher resolution than other single-cell protocols, the detection accuracy is still low compared to the accuracy when using unamplified DNA with the same array platforms. This reduced accuracy is probably due to amplification bias and array platform limitations. Thus, further improvements in design of array platforms, amplification technologies as well as copy number calling algorithms seem necessary. In clinical applications, single-cell genome research has been applied to preimplantation genetic diagnostic (PGD) and preimplantation genetic screening (PGS). SNP arrays are often used in PGD studies, as they provide measurements of allele imbalances, as well as relative copy number changes. When the parental genotypes and parental abnormalities are known, SNP arrays allow the identification of the origin of the mutations in offspring blastomeres, indicating whether the mutation is inherited from the maternal or paternal chromosome. This feature of SNP arrays is particularly useful for PGD applications because it assists transplantation of healthy embryos that do not carry parental affected alleles [28, 39].

Besides PGD, single-cell genome research also provides new views on PGS. PGS is more challenging than PGD because it aims at selecting good embryos by screening the whole genome in a single blastomere without any prior knowledge of parental abnormalities. However, current findings reveal that chromosome instabilities are commonly present in good-quality cleavage-stage embryos, even from healthy couples [9, 41]. Thus, the usefulness of PGS at cleave-stage of human embryos is in doubt now and other strategies will be required.

6.1.2 Perspectives

Recently, the automated Sanger sequencing ('first-generation' technology) was followed by next-generation sequencing (NGS). Compared to first-generation technology, NGS is more efficient in reading genome-wide nucleotide sequences [133]. It outperforms all the array technologies in that it enables the identification of a full catalogue of mutations including point mutations, copy number changes and complex structural rearrangements with high resolution [10].

These newer technologies constitute various strategies relying on a combination of template preparation, sequencing, imaging, genome alignment and assembly methods [133].

The comprehensive information provided by NGS platforms makes this new technology popular for many applications. The major application of NGS is the sequencing of human genomes to enhance our understanding of genetic mechanisms underlying diseases, such as Mendelian disorders. The easiest way to discover causal mutations for genetic diseases is to sequence the genome of affected individuals followed by filtering out non-suspicious regions, such as common variants, silent mutations and non-coding regions, and search for rare mutations existing in most affected individuals. An alternative approach to identify de causal mutations of Mendelian disorders is to find de novo mutations based on sequencing of parent-child trios [134]. The key point is to construct a workflow to efficiently filter out all the un-relevant mutations. In addition, NGS technology is also used a lot to distinguish somatic mutations from large amounts of germline variants by comparing mutations from cancer genomes with matched normal genomes [134]. More applications of NGS technology on cancer genomes will be introduced in section 6.2.

Among all the NGS applications, sequencing in single-cell genomes would be the most challenging area. Besides the challenges in terms of preprocessing billions of NGS reads which are the same for all the NGS data, single-cell sequencing data also contains a large amount of artifacts caused by single-cell DNA amplification resulting in difficulties of distinguishing real biological mutations from genome artifacts and sequencing errors [135].

The most explored type of mutation in single-cell sequencing is unbalanced structural rearrangement resulting in copy number changes. To detect copy number alterations, single-cell sequencing data can be transferred to array-like data by calculating the log-transformed ratios between read depths of test sample and reference sample. In Navin and Hicks (2011), the single-cell sequencing reads were firstly grouped into variable bins across the genome and the log-transformed ratios were calculated based on the depths within each bin. They claim that bins with variable lengths outperform fixed length bins to correct for amplification and mapping artifacts. Subsequently, the variable bins were segmented using the Kolmogorov-Smirnov (KS) statistical test and converted into integer copy numbers [136]. With the development of appropriate binning methods, the previously developed processing methods for single-cell array data could be adapted for single-cell sequencing data. In summary, advanced computational approaches are necessary to keep pace with the continuing developments in NGS technologies.

6.2 Cancer genome

6.2.1 Conclusion

The second part of the thesis describes the copy number alteration analysis of cancer genomes. Somatic copy number alterations in human genome are hallmarks of tumorigenesis [93]. Accurate detection of somatic copy number alterations is a vital step to gain insights into cancer genome development. However, the explorations of these somatic copy number alterations are mostly limited to total copy number gains or losses regardless of allele imbalances. Moreover, the bottlenecks towards cancer genome analysis include tumor polyploidies and unknown proportions of contaminated normal tissues in tumor samples.

The previously developed Allele-Specific Copy number Analysis of Tumors (ASCAT) approach estimates copy number alterations on allele level as well as ploidy and tumor purity [48]. We extended the original ASCAT approach to dissect copy number changes within each subclonal population. By applying the extended approach to 2240 publicly available tumor samples hybridized to SNP arrays, we built allelic-specific copy number profiles and grouped them into 12 tumor types.

The inferred allelic-specific copy number profiles for each tumor type allowed us to reveal chromosomal aberration patterns reflected in recurrent regions of gain, loss and loss of heterogeneity (LOH) between tumor types. The dissected tumor ploidy, purity and subclonal fractions also provided us a view of the heterogeneity pattern for each tumor type. Our results show that different cancer types are characterized by different aberration profiles and intra-tumor heterogeneities.

The most important goal of cancer genome research is to identify tumor suppressor genes from copy number alteration regions. It is known that mutations in tumor suppressor genes are loss-of-function mutations characterized by recurrent deletions. By analyzing recurrent deletion patterns, we applied a normal mixture model to distinguish the regions harboring candidate tumor suppressor genes from fragile sites. In conclusion, the computational methods we developed as well as the comprehensive information we discovered in this study will be valuable for future cancer genome research.

6.2.2 Perspectives

6.2.2.1 Allelic skewness

Allelic skewness is a specific mutation pattern observed in germline heterozygous SNPs across multiple tumor samples. If a SNP at a locus shows an allelic skewness pattern, one allele of the SNP is preferentially amplified or deleted across multiple tumor samples compared to the other allele. The hypothesis is that if a germline heterozygous allele is preferentially skewed in cancer genomes, this skewed allele is subject to a tumor selection pressure [137]. Using the allelic-specific copy number profiles, we are able to look into SNPs that are preferentially skewed across tumor samples and pick up the candidate disease genes mapped by those preferentially allelic skewed SNPs.

Allelic skewness can be measured using a binomial test by comparing the number of times one allele is gained or lost against the other. Under the null hypothesis, the probability of either allele to be preferentially gained or lost is 50%. The SNPs with gain frequencies of two alleles which significantly deviate from a 1:1 ratio are considered allelic skewed SNPs [137]. We tried the allelic skewness test on the 2240 tumor samples in our study and identified a list of allelic-skewed SNPs. However, the results are not trustful enough due to the fact that there are no matched normals of these tumor samples and the germline heterozygous SNPs are estimated based on the B allele frequencies (BAF) of tumor samples. Because the number of germline heterozygous SNPs used for the binomial test is not completely accurate, the result of this test becomes inaccurate as well.

A powerful way to identify allelic skewed SNPs would be as follows: Firstly, identify the germline heterozygous SNPs using matched normal samples. Subsequently, remove a number of SNPs affected by possible platform biases using a number of filters, such as genotyping quality. Thirdly, perform the binomial test on each of the rest of SNPs across tumor samples and select the SNPs showing significant allelic skewness. It would be more convincing to validate the results on an independent dataset using the same workflow and selecting the SNPs that show significance in both studies. In addition, it would also be interesting to check the influence of DNA allelic-skewness on expression data.

6.2.2.2 Identification of cancer-driven genes and follow up studies

Sporadic cancer genomes develop by acquiring somatic mutations, such as point mutations, structural rearrangements and copy number alterations [10]. Somatic mutations can be stratified into driver mutations and passenger mutations. Driver mutations are the result of selection pressure during tumor expansion whereas passenger mutations are randomly present and do not contribute to tumor growth [19]. The identification of driver mutations has been a central goal

of cancer research for a decade. Remarkable successes have been achieved by the detection of recurrent somatic copy number alterations using array technologies and mapping genes within these recurrent regions [93, 138, 139, 140]. To date, over 100000 mutations in human genes are listed in the Catalogue of Somatic Mutations in Cancer (COSMIC) and a census of 418 cancer-driven genes is implicated in cancer development [112, 141].

In our study, we found 89 regions harboring candidate tumor suppressor genes or fragile sites. However, discovery of candidate cancer-driven mutations is only the first step to explore cancer genomes. Follow-up studies are required to validate those mutations. Since suppressor genes require mutations in both alleles, one straightforward way is to re-sequence the targeted regions covering putative cancer-driven genes from a number of cancer genomes and check the mutation pattern [142]. An alternative way is to investigate the biological consequences of putative driver mutations, such as functional studies in animals.

Provision of full genome sequencing technologies offers new strategies and brings new eras for surveying cancer genomes. With the availability of complete nucleotide sequences of human genomes, advanced computational approaches are required to accurately identify the full catalogue of somatic mutations, such as point mutations and complex structural rearrangements [133].

The availability of a full library of somatic mutations helps the analysis of mutation patterns. Nik-Zainal et al. (2012) extracted five distinct mutation processes in 21 breast cancer genomes using nonnegative matrix factorization. They revealed a complex relationship between mutagenesis and transcription [143]. With increased production of sequencing data and expected advances in bioinformatics in the coming years, a full range of somatic mutations from more than fifty classes of cancers can be profiled and characterized [10].

6.2.2.3 Intra-tumor heterogeneity and metastasis

Analysis of intra-tumor heterogeneity is another attractive area in cancer research. Cancer arises from a primary monoclone, progresses dynamically into subclonal populations bearing different somatic mutations [103]. Identification of subclonal populations in cancer cells provides us a picture of evolutionary processes and will be particular useful for the discovery of drug-resistance or chemotherapy-resistance mutations hidden in subclonal populations [10]. The arrival of full-genome sequencing provides the ability to identify subclonal genetic diversity within a population of cancer cells.

Intra-tumor heterogeneity can be detected by deep sequencing. Accessing the exact mutations within each subclone is challenging for the current computational methods because the mutations within subclonal populations can be obscured by sequencing errors. Nik-Zainal et al. (2012) developed a

novel computational framework to identify copy number alterations and point mutations within each subclonal population and built a cancer evolution tree for 21 breast cancer genomes [103]. This was a pilot study showing the ability to accurately decipher intra-tumor heterogeneities. More efforts are needed to adapt the framework for more cancer genomes derived from different cancer types in the near future.

An alterative way to explore intra-tumor heterogeneity is to quantify somatic mutations in single cells extracted from different parts of the tumor tissue. Navin et al. (2011) performed copy number alteration analysis in 100 single-cell genomes from a primary tumor and an additional analysis in 100 single-cell genomes from another primary and its liver metastasis sample. This study identified the subclonal populations from single cells isolated from macrodissected tumor tissues and provided a direct comparison of mutation patterns between primary and metastasis. They confirmed that a primary tumor originates from a single clonal expansion. The comparable copy number alteration patterns between primary tumor and paired metastasis indicated that metastatic cells emerge from a main advanced expansion and arise late in tumor development [144]. The limitation of this study is that they only explored the intra-tumor heterogeneity for copy number alterations and ignored the point mutations present in each subclonal population of cells. A comprehensive analysis of all mutation types will be more appreciated in the future.



Bibliography

[1] F. Collins, E. Lander, J. Rogers, R. Waterston, and I. Conso, "Finishing the euchromatic sequence of the human genome," *Nature*, vol. 431, no. 7011, pp. 931–945, 2004. pages 2

[2] Z. Storchova and C. Kuffer, "The consequences of tetraploidy and aneuploidy," *Journal of cell science*, vol. 121, no. 23, pp. 3859–3866, 2008. pages 3, 15, 79

[3] T. Hassold, H. Hall, and P. Hunt, "The origin of human aneuploidy: where we have been, where we are going," *Human molecular genetics*, vol. 16, no. R2, pp. R203–R208, 2007. pages 3

[4] T. Hassold and P. Hunt, "To err (meiotically) is human: the genesis of human aneuploidy," *Nature Reviews Genetics*, vol. 2, no. 4, pp. 280–291, 2001. pages 3

[5] C. O'Connor, "Chromosomal abnormalities: Aneuploidies," *Nature Education*, vol. 1, no. 1, 2008. pages 3

[6] R. Nussbaum, R. R. McInnes, and H. F. Willard, *Thompson & Thompson genetics in medicine*. Saunders, 2007. pages 3, 4

[7] C. Alkan, B. P. Coe, and E. E. Eichler, "Genome structural variation discovery and genotyping," *Nature Reviews Genetics*, vol. 12, no. 5, pp. 363–376, 2011. pages 3

[8] N. Craddock, M. E. Hurles, N. Cardin, R. D. Pearson, V. Plagnol, S. Robson, D. Vukcevic, C. Barnes, D. F. Conrad, E. Giannoulatou, *et al.*, "Genome-wide association study of cnvs in 16,000 cases of eight common diseases and 3,000 shared controls," *Nature*, vol. 464, no. 7289, pp. 713–720, 2010. pages 3

[9] E. Vanneste, T. Voet, C. Le Caignec, M. Ampe, P. Konings, C. Melotte, S. Debrock, M. Amyere, M. Vikkula, F. Schuit, *et al.*, "Chromosome instability is common in human cleavage-stage embryos," *Nature medicine*, vol. 15, no. 5, pp. 577–583, 2009. pages 3, 5, 10, 12, 14, 15, 22, 25, 36, 38, 39, 42, 56, 57, 64, 65, 104

[10] M. R. Stratton, P. J. Campbell, and P. A. Futreal, "The cancer genome," *Nature*, vol. 458, no. 7239, pp. 719–724, 2009. pages 4, 6, 76, 104, 107, 108

[11] C. Sismani, S. Kitsiou-Tzeli, M. Ioannides, C. Christodoulou, V. Anastasiadou, G. Stylianidou, E. Papadopoulou, E. Kanavakis, Z. Kosmaidou-Aravidou, P. C. Patsalis, *et al.*, "Cryptic genomic imbalances in patients with de novo or familial apparently balanced translocations and abnormal phenotype," *Mol Cytogenet*, vol. 1, no. 1, p. 15, 2008. pages 4

[12] P. R. Brezina, D. S. Brezina, and W. G. Kearns, "Preimplantation genetic testing," *BMJ: British Medical Journal*, vol. 345, 2012. pages 5

[13] A. R. Thornhill, J. Geraedts, J. Harper, G. Harton, S. Lavery, C. Moutou, M. Robinson, A. Schmutzler, P. Scriven, K. Sermon, *et al.*, "Eshre pgd consortium 'best practice guidelines for clinical preimplantation genetic diagnosis (pgd) and preimplantation genetic screening (pgs)'," *Human Reproduction*, vol. 20, no. 1, pp. 35–48, 2005. pages 5

[14] L. Wilton, "Preimplantation genetic diagnosis and chromosome analysis of blastomeres using comparative genomic hybridization," *Human reproduction update*, vol. 11, no. 1, pp. 33–41, 2005. pages 5

[15] G. Harton, M. Magli, K. Lundin, M. Montag, J. Lemmen, and J. Harper, "ESHRE PGD Consortium/Embryology Special Interest Group-best practice guidelines for polar body and embryo biopsy for preimplantation genetic diagnosis/screening (PGD/PGS)," *Human Reproduction*, vol. 26, no. 1, pp. 41–46, 2011. pages 5

[16] J. C. Harper and S. B. SenGupta, "Preimplantation genetic diagnosis: State of the art 2011," *Human genetics*, vol. 131, no. 2, pp. 175–186, 2012. pages 5

[17] S. J. McArthur, D. Leigh, J. T. Marshall, K. A. de Boer, and R. P. Jansen, "Pregnancies and live births after trophectoderm biopsy and preimplantation genetic testing of human blastocysts," *Fertility and sterility*, vol. 84, no. 6, pp. 1628–1636, 2005. pages 6

[18] A. Jemal, F. Bray, M. M. Center, J. Ferlay, E. Ward, and D. Forman, "Global cancer statistics," *CA: a cancer journal for clinicians*, vol. 61, no. 2, pp. 69–90, 2011. pages 6

[19] C. Greenman, P. Stephens, R. Smith, G. L. Dalgliesh, C. Hunter, G. Bignell, H. Davies, J. Teague, A. Butler, C. Stevens, *et al.*, "Patterns of somatic mutation in human cancer genomes," *Nature*, vol. 446, no. 7132, pp. 153–158, 2007. pages 6, 76, 107

[20] D. Pinto, K. Darvishi, X. Shi, D. Rajan, D. Rigler, T. Fitzgerald, A. C. Lionel, B. Thiruvahindrapuram, J. R. MacDonald, R. Mills, *et al.*, "Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants," *Nature biotechnology*, vol. 29, no. 6, pp. 512–520, 2011. pages 10, 15

[21] E. Vanneste, L. Bittman, N. Van der Aa, T. Voet, and J. R. Vermeesch, "New array approaches to explore single cells genomes," *Frontiers in genetics*, vol. 3, 2012. pages 10, 11, 12

[22] H. Fiegler, J. B. Geigl, S. Langer, D. Rigler, K. Porter, K. Unger, N. P. Carter, and M. R. Speicher, "High resolution array-cgh analysis of single cells," *Nucleic acids research*, vol. 35, no. 3, pp. e15–e15, 2007. pages 10, 12, 13, 14, 22

[23] A. Hellani, K. Abu-Amero, J. Azouri, and S. El-Akoum, "Successful pregnancies after application of array-comparative genomic hybridization in pgs-aneuploidy screening," *Reproductive biomedicine online*, vol. 17, no. 6, pp. 841–847, 2008. pages 10, 14

[24] J. B. Geigl, A. C. Obenauf, J. Waldispuehl-Geigl, E. M. Hoffmann, M. Auer, M. Hormann, M. Fischer, Z. Trajanoski, M. A. Schenk, L. O. Baumbusch, *et al.*, "Identification of small gains and losses in single cells after whole genome amplification on tiling oligo arrays," *Nucleic acids research*, vol. 37, no. 15, pp. e105–e105, 2009. pages 10, 11, 14, 38

[25] J. Cheng, E. Vanneste, P. Konings, T. Voet, J. R. Vermeesch, and Y. Moreau, "Single-cell copy number variation detection," *Genome Biology*, vol. 12, no. 8, p. R80, 2011. pages 10, 12, 14, 57, 65, 88

[26] K. Iwamoto, M. Bundo, J. Ueda, Y. Nakano, W. Ukai, E. Hashimoto, T. Saito, and T. Kato, "Detection of chromosomal structural alterations in single cells by snp arrays: a systematic survey of amplification bias and optimized workflow," *PloS one*, vol. 2, no. 12, p. e1306, 2007. pages 11, 14, 38

[27] N. R. Treff, J. Su, X. Tao, B. Levy, and R. T. Scott Jr, "Accurate single cell 24 chromosome aneuploidy screening using whole genome amplification and single nucleotide polymorphism microarrays," *Fertility and sterility*, vol. 94, no. 6, pp. 2017–2021, 2010. pages 11, 14, 38

[28] D. Johnson, G. Gemelos, J. Baner, A. Ryan, C. Cinnioglu, M. Banjevic, R. Ross, M. Alper, B. Barrett, J. Frederick, *et al.*, "Preclinical validation of a microarray method for full molecular karyotyping of blastomeres in a 24-h protocol," *Human Reproduction*, vol. 25, no. 4, pp. 1066–1075, 2010. pages 11, 13, 14, 15, 55, 56, 64, 104

[29] A. H. Handyside, G. L. Harton, B. Mariani, A. R. Thornhill, N. Affara, M.-A. Shaw, and D. K. Griffin, "Karyomapping: a universal method for genome wide analysis of genetic disease based on mapping crossovers between parental haplotypes," *Journal of medical genetics*, vol. 47, no. 10, pp. 651–658, 2010. pages 11, 13, 15, 22

[30] K. Wang, M. Li, D. Hadley, R. Liu, J. Glessner, S. F. Grant, H. Hakonarson, and M. Bucan, "Penncnv: an integrated hidden markov model designed for high-resolution copy number variation detection in whole-genome snp genotyping data," *Genome research*, vol. 17, no. 11, pp. 1665–1674, 2007. pages 11, 74, 88

[31] F. Fiorentino, L. Spizzichino, S. Bono, A. Biricik, G. Kokkali, L. Rienzi, F. Ubaldi, E. Iammarrone, A. Gordon, and K. Pantos, "Pgd for reciprocal and robertsonian translocations using array comparative genomic hybridization," *Human Reproduction*, vol. 26, no. 7, pp. 1925–1935, 2011. pages 12, 14, 55, 67

[32] C. Spits, C. Le Caignec, M. De Rycke, L. Van Haute, A. Van Steirteghem, I. Liebaers, and K. Sermon, "Whole-genome multiple displacement amplification from single cells," *Nature protocols*, vol. 1, no. 4, pp. 1965–1970, 2006. pages 12, 22, 42, 56

[33] C. Le Caignec, C. Spits, K. Sermon, M. De Rycke, B. Thienpont, S. Debrock, C. Staessen, Y. Moreau, J.-P. Fryns, A. Van Steirteghem, *et al.*, "Single-cell chromosomal imbalances detection by array cgh," *Nucleic acids research*, vol. 34, no. 9, pp. e68–e68, 2006. pages 12, 14, 22, 56, 65

[34] M. Ampe, G. Verbeke, E. Vanneste, and J. R. Vermeesch, "Ojb tm," *Online Journal of Bioinformatics*, vol. 11, no. 2, pp. 224–244, 2010. pages 12

[35] P. Konings, E. Vanneste, S. Jackmaert, M. Ampe, G. Verbeke, Y. Moreau, J. R. Vermeesch, and T. Voet, "Microarray analysis of copy number variation in single cells," *Nature Protocols*, vol. 7, no. 2, pp. 281–310, 2012. pages 12, 15

[36] E. Venkatraman and A. B. Olshen, "A faster circular binary segmentation algorithm for the analysis of array cgh data," *Bioinformatics*, vol. 23, no. 6, pp. 657–663, 2007. pages 12, 27, 57

[37] M. A. Van De Wiel, K. I. Kim, S. J. Vosse, W. N. Van Wieringen, S. M. Wilting, and B. Ylstra, "Cghcall: calling aberrations for array cgh tumor profiles," *Bioinformatics*, vol. 23, no. 7, pp. 892–894, 2007. pages 12, 43, 57

[38] 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*, vol. 90, no. 1, pp. 132–153, 2004. pages 13

[39] T. Voet, E. Vanneste, N. Van der Aa, C. Melotte, S. Jackmaert, T. Vandendael, M. Declercq, S. Debrock, J.-P. Fryns, Y. Moreau, *et al.*, "Breakage–fusion–bridge cycles leading to inv dup del occur in human cleavage stage embryos," *Human mutation*, vol. 32, no. 7, pp. 783–793, 2011. pages 13, 15, 55, 64, 104

[40] C. Gutiérrez-Mateo, P. Colls, J. Sánchez-García, T. Escudero, R. Prates, K. Ketterson, D. Wells, and S. Munné, "Validation of microarray comparative genomic hybridization for comprehensive chromosome analysis of embryos," *Fertility and sterility*, vol. 95, no. 3, pp. 953–958, 2011. pages 13, 14, 64

[41] A. Mertzanidou, L. Wilton, J. Cheng, C. Spits, E. Vanneste, Y. Moreau, J. Vermeesch, and K. Sermon, "Microarray analysis reveals abnormal chromosomal complements in over 70% of 14 normally developing human embryos," *Human Reproduction*, vol. 28, no. 1, pp. 256–264, 2013. pages 14, 104

[42] W. R. Lai, M. D. Johnson, R. Kucherlapati, and P. J. Park, "Comparative analysis of algorithms for identifying amplifications and deletions in array cgh data," *Bioinformatics*, vol. 21, no. 19, pp. 3763–3770, 2005. pages 15, 43

[43] 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*, vol. 7, no. 1, p. e1001060, 2011. pages 15

[44] D. Mosén-Ansorena, A. Aransay, and N. Rodríguez-Ezpeleta, "Comparison of methods to detect copy number alterations in cancer using simulated and real genotyping data," *BMC bioinformatics*, vol. 13, no. 1, p. 192, 2012. pages 15, 17

[45] C. Yau, D. Mouradov, R. N. Jorissen, S. Colella, G. Mirza, G. Steers, A. Harris, J. Ragoussis, O. Sieber, C. C. Holmes, *et al.*, "A statistical approach for detecting genomic aberrations in heterogeneous tumor

samples from single nucleotide polymorphism genotyping data," *Genome Biol*, vol. 11, no. 9, pp. R92–R92, 2010. pages 15, 77, 80

[46] T. Popova, E. Manié, D. Stoppa-Lyonnet, G. Rigaill, E. Barillot, M. H. Stern, *et al.*, "Genome alteration print (gap): a tool to visualize and mine complex cancer genomic profiles obtained by snp arrays," *Genome Biol*, vol. 10, no. 11, pp. R128–R128, 2009. pages 15, 77, 80

[47] 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*, vol. 30, no. 5, pp. 413–421, 2012. pages 15, 77, 79, 80

[48] 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*, vol. 107, no. 39, pp. 16910–16915, 2010. pages 16, 77, 78, 80, 88, 89, 106

[49] G. Nilsen, K. Liestøl, P. Van Loo, H. K. M. Vollan, M. B. Eide, O. M. Rueda, S.-F. Chin, R. Russell, L. O. Baumbusch, C. Caldas, *et al.*, "Copynumber: Efficient algorithms for single-and multi-track copy number segmentation," *BMC genomics*, vol. 13, no. 1, p. 591, 2012. pages 16, 88

[50] M. Rasmussen, M. Sundstrom, H. G. Kultima, J. Botling, P. Micke, H. Birgisson, B. Glimelius, A. Isaksson, *et al.*, "Allele-specific copy number analysis of tumor samples with aneuploidy and tumor heterogeneity," *Genome Biol*, vol. 12, no. 10, pp. R108–R108, 2011. pages 17

[51] J. G. Hacia, K. Edgemon, B. Sun, F. S. Collins, D. Stern, and S. P. Fodor, "Two color hybridization analysis using high density oligonucleotide arrays and energy transfer dyes," *Nucleic acids research*, vol. 26, no. 16, pp. 3865–3866, 1998. pages 22

[52] N. E. Wineinger, R. E. Kennedy, S. W. Erickson, M. K. Wojczynski, C. Bruder, and H. Tiwari, "Statistical issues in the analysis of dna copy number variations," *International journal of computational biology and drug design*, vol. 1, no. 4, pp. 368–395, 2008. pages 22

[53] W. S. Cleveland, "Robust locally weighted regression and smoothing scatterplots," *Journal of the American statistical association*, vol. 74, no. 368, pp. 829–836, 1979. pages 22

[54] M. Smith, J. Marioni, T. Hardcastle, and N. Thorne, "snapcgh: Segmentation, normalization and processing of acgh data users' guide," *Bioinformatics*, vol. 22, pp. 1144–1146, 2006. pages 22, 43

[55] J. Staaf, G. Jonsson, M. Ringner, and J. Vallon-Christersson, "Normalization of array-cgh data: influence of copy number imbalances," *BMC genomics*, vol. 8, no. 1, p. 382, 2007. pages 22

[56] B. P. van Houte, T. W. Binsl, H. Hettling, W. Pirovano, and J. Heringa, "Cghnormaliter: an iterative strategy to enhance normalization of array cgh data with imbalanced aberrations," *BMC genomics*, vol. 10, no. 1, p. 401, 2009. pages 22

[57] M. A. van de Wiel, R. Brosens, P. H. Eilers, C. Kumps, G. A. Meijer, B. Menten, E. Sistermans, F. Speleman, M. E. Timmerman, and B. Ylstra, "Smoothing waves in array cgh tumor profiles," *Bioinformatics*, vol. 25, no. 9, pp. 1099–1104, 2009. pages 23

[58] E. Ben-Yaacov and Y. C. Eldar, "A fast and flexible method for the segmentation of acgh data," *Bioinformatics*, vol. 24, no. 16, pp. i139–i145, 2008. pages 23, 34

[59] P. Hupe, N. Stransky, J.-P. Thiery, F. Radvanyi, and E. Barillot, "Analysis of array cgh data: from signal ratio to gain and loss of dna regions," *Bioinformatics*, vol. 20, no. 18, pp. 3413–3422, 2004. pages 23

[60] R. Pique-Regi, A. Ortega, and S. Asgharzadeh, "Joint estimation of copy number variation and reference intensities on multiple dna arrays using gada," *Bioinformatics*, vol. 25, no. 10, pp. 1223–1230, 2009. pages 23, 34

[61] F. Picard, E. Lebarbier, M. Hoebeke, G. Rigaill, B. Thiam, and S. Robin, "Joint segmentation, calling, and normalization of multiple cgh profiles," *Biostatistics*, vol. 12, no. 3, pp. 413–428, 2011. pages 23, 31, 34

[62] M. A. van de Wiel, F. Picard, W. N. van Wieringen, and B. Ylstra, "Preprocessing and downstream analysis of microarray dna copy number profiles," *Briefings in bioinformatics*, vol. 12, no. 1, pp. 10–21, 2011. pages 28, 29

[63] S. J. Diskin, M. Li, C. Hou, S. Yang, J. Glessner, H. Hakonarson, M. Bucan, J. M. Maris, and K. Wang, "Adjustment of genomic waves in signal intensities from whole-genome snp genotyping platforms," *Nucleic acids research*, vol. 36, no. 19, pp. e126–e126, 2008. pages 29

[64] J. B. Geigl and M. R. Speicher, "Single-cell isolation from cell suspensions and whole genome amplification from single cells to provide templates for cgh analysis," *Nature protocols*, vol. 2, no. 12, pp. 3173–3184, 2007. pages 38

[65] R. C. Team *et al.*, "R: A language and environment for statistical computing," *R Foundation Statistical Computing*, 2008. pages 39

[66] M. E. Ritchie, J. Silver, A. Oshlack, M. Holmes, D. Diyagama, A. Holloway, and G. K. Smyth, "A comparison of background correction methods for two-colour microarrays," *Bioinformatics*, vol. 23, no. 20, pp. 2700–2707, 2007. pages 39

[67] C. H. Reinsch, "Smoothing by spline functions," *Numerische mathematik*, vol. 10, no. 3, pp. 177–183, 1967. pages 41

[68] S. Munné, M. Alikani, G. Tomkin, J. Grifo, and J. Cohen, "Embryo morphology, developmental rates, and maternal age are correlated with chromosome abnormalities," *Fertility and sterility*, vol. 64, no. 2, pp. 382–391, 1995. pages 55

[69] E. Fragouli, D. Wells, A. Doshi, S. Gotts, J. C. Harper, and J. D. Delhanty, "Complete cytogenetic investigation of oocytes from a young cancer patient with the use of comparative genomic hybridisation reveals meiotic errors," *Prenatal diagnosis*, vol. 26, no. 1, pp. 71–76, 2006. pages 55

[70] E. Fragouli, S. Alfarawati, D. D. Daphnis, N.-n. Goodall, A. Mania, T. Griffiths, A. Gordon, and D. Wells, "Cytogenetic analysis of human blastocysts with the use of fish, cgh and acgh: scientific data and technical evaluation," *Human Reproduction*, vol. 26, no. 2, pp. 480–490, 2011. pages 55, 64

[71] L. Voullaire, V. Collins, T. Callaghan, J. McBain, R. Williamson, and L. Wilton, "High incidence of complex chromosome abnormality in cleavage embryos from patients with repeated implantation failure," *Fertility and sterility*, vol. 87, no. 5, pp. 1053–1058, 2007. pages 55

[72] T. Voet, E. Vanneste, and J. Vermeesch, "The human cleavage stage embryo is a cradle of chromosomal rearrangements," *Cytogenetic and genome research*, vol. 133, no. 2-4, pp. 160–168, 2011. pages 55, 64

[73] L. Voullaire, L. Wilton, J. McBain, T. Callaghan, and R. Williamson, "Chromosome abnormalities identified by comparative genomic hybridization in embryos from women with repeated implantation failure," *Molecular human reproduction*, vol. 8, no. 11, pp. 1035–1041, 2002. pages 55, 64

[74] S. Alfarawati, E. Fragouli, P. Colls, and D. Wells, "First births after preimplantation genetic diagnosis of structural chromosome abnormalities using comparative genomic hybridization and microarray analysis," *Human Reproduction*, vol. 26, no. 6, pp. 1560–1574, 2011. pages 55

[75] L. Voullaire, H. Slater, R. Williamson, and L. Wilton, "Chromosome analysis of blastomeres from human embryos by using comparative genomic

hybridization," *Human genetics*, vol. 106, no. 2, pp. 210–217, 2000. pages 56, 63, 64

[76] D. Wells and J. D. Delhanty, "Comprehensive chromosomal analysis of human preimplantation embryos using whole genome amplification and single cell comparative genomic hybridization," *Molecular Human Reproduction*, vol. 6, no. 11, pp. 1055–1062, 2000. pages 56, 63, 64

[77] H. Jericho, L. Wilton, D. A. Gook, and D. H. Edgar, "A modified cryopreservation method increases the survival of human biopsied cleavage stage embryos," *Human Reproduction*, vol. 18, no. 3, pp. 568–571, 2003. pages 56

[78] J. van Echten-Arends, S. Mastenbroek, B. Sikkema-Raddatz, J. C. Korevaar, M. J. Heineman, F. van der Veen, and S. Repping, "Chromosomal mosaicism in human preimplantation embryos: a systematic review," *Human reproduction update*, vol. 17, no. 5, pp. 620–627, 2011. pages 64

[79] N. R. Treff, B. Levy, J. Su, L. E. Northrop, X. Tao, and R. T. Scott, "Snp microarray-based 24 chromosome aneuploidy screening is significantly more consistent than fish," *Molecular human reproduction*, vol. 16, no. 8, pp. 583–589, 2010. pages 64

[80] D. Wells, J. K. Sherlock, J. D. Delhanty, and A. H. Handyside, "Detailed chromosomal and molecular genetic analysis of single cells by whole genome amplification and comparative genomic hybridisation," *Nucleic acids research*, vol. 27, no. 4, pp. 1214–1218, 1999. pages 64

[81] D. Daphnis, E. Fragouli, K. Economou, S. Jerkovic, I. Craft, J. Delhanty, and J. Harper, "Analysis of the evolution of chromosome abnormalities in human embryos from day 3 to 5 using cgh and fish," *Molecular human reproduction*, vol. 14, no. 2, pp. 117–125, 2008. pages 64

[82] E. Vanneste, C. Melotte, T. Voet, C. Robberecht, S. Debrock, A. Pexsters, C. Staessen, C. Tomassetti, E. Legius, T. D'Hooghe, *et al.*, "Pgd for a complex chromosomal rearrangement by array comparative genomic hybridization," *Human Reproduction*, vol. 26, no. 4, pp. 941–949, 2011. pages 64, 65

[83] E. Gonzalez-Merino, S. Emiliani, G. Vassart, M. Van den Bergh, A.-S. Vannin, M. Abramowicz, D. Delneste, and Y. Englert, "Incidence of chromosomal mosaicism in human embryos at different developmental stages analyzed by fluorescence in situ hybridization," *Genetic testing*, vol. 7, no. 2, pp. 85–95, 2003. pages 65

[84] E. B. Baart, E. Martini, I. van den Berg, N. Macklon, R. H. Galjaard, B. Fauser, and D. Van Opstal, "Preimplantation genetic screening reveals a high incidence of aneuploidy and mosaicism in embryos from young women undergoing ivf," *Human Reproduction*, vol. 21, no. 1, pp. 223–233, 2006. pages 65

[85] C. Rubio, L. Rodrigo, A. Mercader, E. Mateu, P. Buendia, T. Pehlivan, T. Viloria, M. J. De los Santos, C. Simon, J. Remohi, *et al.*, "Impact of chromosomal abnormalities on preimplantation embryo development," *Prenatal diagnosis*, vol. 27, no. 8, pp. 748–756, 2007. pages 65

[86] S. Barbash-Hazan, T. Frumkin, M. Malcov, Y. Yaron, T. Cohen, F. Azem, A. Amit, and D. Ben-Yosef, "Preimplantation aneuploid embryos undergo self-correction in correlation with their developmental potential," *Fertility and sterility*, vol. 92, no. 3, pp. 890–896, 2009. pages 65

[87] C. Robberecht, E. Vanneste, A. Pexsters, T. D'Hooghe, T. Voet, and J. R. Vermeesch, "Somatic genomic variations in early human prenatal development," *Current genomics*, vol. 11, no. 6, p. 397, 2010. pages 65

[88] M. Bielanska, S. L. Tan, and A. Ao, "Chromosomal mosaicism throughout human preimplantation development in vitro: incidence, type, and relevance to embryo outcome," *Human Reproduction*, vol. 17, no. 2, pp. 413–419, 2002. pages 65

[89] D. K. Kalousek, "Pathogenesis of chromosomal mosaicism and its effect on early human development," *American journal of medical genetics*, vol. 91, no. 1, pp. 39–45, 2000. pages 65

[90] F. J. Los, D. Van Opstal, and C. van den Berg, "The development of cytogenetically normal, abnormal and mosaic embryos: a theoretical model," *Human reproduction update*, vol. 10, no. 1, pp. 79–94, 2004. pages 65

[91] B. J. Druker, M. Talpaz, D. J. Resta, B. Peng, E. Buchdunger, J. M. Ford, N. B. Lydon, H. Kantarjian, R. Capdeville, S. Ohno-Jones, *et al.*, "Efficacy and safety of a specific inhibitor of the bcr-abl tyrosine kinase in chronic myeloid leukemia," *New England Journal of Medicine*, vol. 344, no. 14, pp. 1031–1037, 2001. pages 76

[92] D. J. Slamon, B. Leyland-Jones, S. Shak, H. Fuchs, V. Paton, A. Bajamonde, T. Fleming, W. Eiermann, J. Wolter, M. Pegram, *et al.*, "Use of chemotherapy plus a monoclonal antibody against her2 for metastatic breast cancer that overexpresses her2," *New England Journal of Medicine*, vol. 344, no. 11, pp. 783–792, 2001. pages 76

[93] R. Beroukhim, C. H. Mermel, D. Porter, G. Wei, S. Raychaudhuri, J. Donovan, J. Barretina, J. S. Boehm, J. Dobson, M. Urashima, *et al.*, "The landscape of somatic copy-number alteration across human cancers," *Nature*, vol. 463, no. 7283, pp. 899–905, 2010. pages 76, 77, 106, 108

[94] P. J. Stephens, P. S. Tarpey, H. Davies, P. Van Loo, C. Greenman, D. C. Wedge, S. Nik-Zainal, S. Martin, I. Varela, G. R. Bignell, *et al.*, "The landscape of cancer genes and mutational processes in breast cancer," *Nature*, vol. 486, no. 7403, pp. 400–404, 2012. pages 76, 77, 80, 85

[95] T.-M. Kim, R. Xi, L. J. Luquette, R. W. Park, M. D. Johnson, and P. J. Park, "Functional genomic analysis of chromosomal aberrations in a compendium of 8000 cancer genomes," *Genome research*, vol. 23, no. 2, pp. 217–227, 2013. pages 77, 80, 87

[96] G. R. Bignell, C. D. Greenman, H. Davies, A. P. Butler, S. Edkins, J. M. Andrews, G. Buck, L. Chen, D. Beare, C. Latimer, *et al.*, "Signatures of mutation and selection in the cancer genome," *Nature*, vol. 463, no. 7283, pp. 893–898, 2010. pages 77, 80, 81, 82, 83

[97] D. A. Peiffer, J. M. Le, F. J. Steemers, W. Chang, T. Jenniges, F. Garcia, K. Haden, J. Li, C. A. Shaw, J. Belmont, *et al.*, "High-resolution genomic profiling of chromosomal aberrations using infinium whole-genome genotyping," *Genome Research*, vol. 16, no. 9, pp. 1136–1148, 2006. pages 77

[98] P. Van Loo and P. J. Campbell, "Absolute cancer genomics," *Nature biotechnology*, vol. 30, no. 7, pp. 620–621, 2012. pages 77

[99] J. Hicks, A. Krasnitz, B. Lakshmi, N. E. Navin, M. Riggs, E. Leibu, D. Esposito, J. Alexander, J. Troge, V. Grubor, *et al.*, "Novel patterns of genome rearrangement and their association with survival in breast cancer," *Genome research*, vol. 16, no. 12, pp. 1465–1479, 2006. pages 77, 79

[100] H. G. Russnes, H. K. Vollan, O. C. Lingjaerde, A. Krasnitz, P. Lundin, B. Naume, *et al.*, "Genomic architecture characterizes tumor progression paths and fate in breast cancer patients," *Science Translational Medicine*, vol. 2, no. 38, p. 38ra47, 2010. pages 77, 79

[101] F. Magrangeas, H. Avet-Loiseau, N. C. Munshi, and S. Minvielle, "Chromothripsis identifies a rare and aggressive entity among newly diagnosed multiple myeloma patients," *Blood*, vol. 118, no. 3, pp. 675–678, 2011. pages 77

[102] N. J. Birkbak, Z. C. Wang, J.-Y. Kim, A. C. Eklund, Q. Li, R. Tian, C. Bowman-Colin, Y. Li, A. Greene-Colozzi, J. D. Iglehart, *et al.*, "Telomeric allelic imbalance indicates defective dna repair and sensitivity to dna-damaging agents," *Cancer Discovery*, vol. 2, no. 4, pp. 366–375, 2012. pages 77

[103] S. Nik-Zainal, P. Van Loo, D. C. Wedge, L. B. Alexandrov, C. D. Greenman, K. W. Lau, K. Raine, D. Jones, J. Marshall, M. Ramakrishna, *et al.*, "The life history of 21 breast cancers," *Cell*, 2012. pages 77, 108, 109

[104] M. Gerlinger, A. J. Rowan, S. Horswell, J. Larkin, D. Endesfelder, E. Gronroos, P. Martinez, N. Matthews, A. Stewart, P. Tarpey, *et al.*, "Intratumor heterogeneity and branched evolution revealed by multiregion sequencing," *New England Journal of Medicine*, vol. 366, no. 10, pp. 883–892, 2012. pages 77, 84

[105] N. Navin, A. Krasnitz, L. Rodgers, K. Cook, J. Meth, J. Kendall, M. Riggs, Y. Eberling, J. Troge, V. Grubor, *et al.*, "Inferring tumor progression from genomic heterogeneity," *Genome research*, vol. 20, no. 1, pp. 68–80, 2010. pages 77

[106] K. Anderson, C. Lutz, F. W. van Delft, C. M. Bateman, Y. Guo, S. M. Colman, H. Kempski, A. V. Moorman, I. Titley, J. Swansbury, *et al.*, "Genetic variegation of clonal architecture and propagating cells in leukaemia," *Nature*, vol. 469, no. 7330, pp. 356–361, 2010. pages 77

[107] F. Notta, C. G. Mullighan, J. C. Wang, A. Poeppl, S. Doulatov, L. A. Phillips, J. Ma, M. D. Minden, J. R. Downing, and J. E. Dick, "Evolution of human bcr-abl1 lymphoblastic leukaemia-initiating cells," *Nature*, vol. 469, no. 7330, pp. 362–367, 2011. pages 77

[108] M. Baudis, "Genomic imbalances in 5918 malignant epithelial tumors: an explorative meta-analysis of chromosomal cgh data," *BMC cancer*, vol. 7, no. 1, p. 226, 2007. pages 80

[109] D. Gisselsson, T. Jonson, Å. Petersén, B. Strömbeck, P. Dal Cin, M. Höglund, F. Mitelman, F. Mertens, and N. Mandahl, "Telomere dysfunction triggers extensive dna fragmentation and evolution of complex chromosome abnormalities in human malignant tumors," *Proceedings of the National Academy of Sciences*, vol. 98, no. 22, pp. 12683–12688, 2001. pages 82

[110] G. R. Bignell, T. Santarius, J. C. Pole, A. P. Butler, J. Perry, E. Pleasance, C. Greenman, A. Menzies, S. Taylor, S. Edkins, *et al.*, "Architectures of

somatic genomic rearrangement in human cancer amplicons at sequence-level resolution," *Genome research*, vol. 17, no. 9, pp. 1296–1303, 2007. pages 82

[111] B. McClintock, "The stability of broken ends of chromosomes in zea mays," *Genetics*, vol. 26, no. 2, p. 234, 1941. pages 82

[112] 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*, vol. 39, no. suppl 1, pp. D945–D950, 2011. pages 82, 108

[113] Y. Zhang, O. Foreman, D. A. Wigle, F. Kosari, G. Vasmatzis, J. L. Salisbury, J. van Deursen, P. J. Galardy, *et al.*, "Usp44 regulates centrosome positioning to prevent aneuploidy and suppress tumorigenesis," *The Journal of clinical investigation*, 2012. pages 83, 86

[114] N. Kim and S. Jinks-Robertson, "Transcription as a source of genome instability," *Nature Reviews Genetics*, vol. 13, no. 3, pp. 204–214, 2012. pages 83

[115] L. G. Morris, A. M. Kaufman, Y. Gong, D. Ramaswami, L. A. Walsh, Ş. Turcan, S. Eng, K. Kannan, Y. Zou, L. Peng, *et al.*, "Recurrent somatic mutation of fat1 in multiple human cancers leads to aberrant wnt activation," *Nature genetics*, vol. 45, no. 3, pp. 253–261, 2013. pages 85

[116] C. M. Annunziata, R. E. Davis, Y. Demchenko, W. Bellamy, A. Gabrea, F. Zhan, G. Lenz, I. Hanamura, G. Wright, W. Xiao, *et al.*, "Frequent engagement of the classical and alternative nf-κb pathways by diverse genetic abnormalities in multiple myeloma," *Cancer cell*, vol. 12, no. 2, pp. 115–130, 2007. pages 85

[117] D. Xiong, G. Li, K. Li, Q. Xu, Z. Pan, F. Ding, P. Vedell, P. Liu, P. Cui, X. Hua, *et al.*, "Exome sequencing identifies mxra5 as a novel cancer gene frequently mutated in non–small cell lung carcinoma from chinese patients," *Carcinogenesis*, vol. 33, no. 9, pp. 1797–1805, 2012. pages 85

[118] S. Ito, L. Shen, Q. Dai, S. C. Wu, L. B. Collins, J. A. Swenberg, C. He, and Y. Zhang, "Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine," *Science*, vol. 333, no. 6047, pp. 1300–1303, 2011. pages 85

[119] Z. K. Atak, K. De Keersmaecker, V. Gianfelici, E. Geerdens, R. Vandepoel, D. Pauwels, M. Porcu, I. Lahortiga, V. Brys, W. G. Dirks, *et al.*, "High accuracy mutation detection in leukemia on a selected panel of cancer genes," *PLoS One*, vol. 7, no. 6, p. e38463, 2012. pages 85

[120] S. L. Cooke, J. C. Pole, S.-F. Chin, I. O. Ellis, C. Caldas, and P. A. Edwards, "High-resolution array cgh clarifies events occurring on 8p in carcinogenesis," *BMC cancer*, vol. 8, no. 1, p. 288, 2008. pages 85

[121] M. Esteller, J. Garcia-Foncillas, E. Andion, S. N. Goodman, O. F. Hidalgo, V. Vanaclocha, S. B. Baylin, and J. G. Herman, "Inactivation of the dna-repair gene mgmt and the clinical response of gliomas to alkylating agents," *New England Journal of Medicine*, vol. 343, no. 19, pp. 1350–1354, 2000. pages 85

[122] A. Bric, C. Miething, C. U. Bialucha, C. Scuoppo, L. Zender, A. Krasnitz, Z. Xuan, J. Zuber, M. Wigler, J. Hicks, *et al.*, "Functional identification of tumor-suppressor genes through an in vivo rna interference screen in a mouse lymphoma model," *Cancer cell*, vol. 16, no. 4, pp. 324–335, 2009. pages 86

[123] L. Lan, A. Ui, S. Nakajima, K. Hatakeyama, M. Hoshi, R. Watanabe, S. M. Janicki, H. Ogiwara, T. Kohno, S.-i. Kanno, *et al.*, "The acf1 complex is required for dna double-strand break repair in human cells," *Molecular cell*, vol. 40, no. 6, pp. 976–987, 2010. pages 86

[124] J. L. Vizmanos, M. J. Larráyoz, I. Lahortiga, F. Floristán, C. Alvarez, M. D. Odero, F. J. Novo, and M. J. Calasanz, "t (10; 16)(q22; p13) and morf-crebbp fusion is a recurrent event in acute myeloid leukemia," *Genes, Chromosomes and Cancer*, vol. 36, no. 4, pp. 402–405, 2003. pages 86

[125] S.-C. Peng, Y.-T. Lai, H.-Y. Huang, H.-D. Huang, and Y.-S. Huang, "A novel role of cpeb3 in regulating egfr gene transcription via association with stat5b in neurons," *Nucleic acids research*, vol. 38, no. 21, pp. 7446–7457, 2010. pages 86

[126] J. Liu, H.-J. Chung, M. Vogt, Y. Jin, D. Malide, L. He, M. Dundr, and D. Levens, "Jtv1 co-activates fbp to induce usp29 transcription and stabilize p53 in response to oxidative stress," *The EMBO journal*, vol. 30, no. 5, pp. 846–858, 2011. pages 86

[127] T. T. Lui, C. Lacroix, S. M. Ahmed, S. J. Goldenberg, C. A. Leach, A. M. Daulat, and S. Angers, "The ubiquitin-specific protease usp34 regulates axin stability and wnt/β-catenin signaling," *Molecular and cellular biology*, vol. 31, no. 10, pp. 2053–2065, 2011. pages 86

[128] S. H. Kresse, H. O. Ohnstad, E. B. Paulsen, B. Bjerkehagen, K. Szuhai, M. Serra, K.-L. Schaefer, O. Myklebost, and L. A. Meza-Zepeda, "Lsamp, a novel candidate tumor suppressor gene in human osteosarcomas, identified by array comparative genomic hybridization," *Genes, Chromosomes and Cancer*, vol. 48, no. 8, pp. 679–693, 2009. pages 87

[129] S. Veeriah, C. Brennan, S. Meng, B. Singh, J. A. Fagin, D. B. Solit, P. B. Paty, D. Rohle, I. Vivanco, J. Chmielecki, *et al.*, "The tyrosine phosphatase ptprd is a tumor suppressor that is frequently inactivated and mutated in glioblastoma and other human cancers," *Proceedings of the National Academy of Sciences*, vol. 106, no. 23, pp. 9435–9440, 2009. pages 87

[130] V. Hosur, A. Kavirayani, J. Riefler, L. Carney, B. Lyons, B. Gott, G. A. Cox, and L. D. Shultz, "Dystrophin and dysferlin double mutant mice: a novel model for rhabdomyosarcoma," *Cancer Genetics*, vol. 205, no. 5, pp. 232–241, 2012. pages 87

[131] Q. Cao, M. Zhou, X. Wang, C. A. Meyer, Y. Zhang, Z. Chen, C. Li, and X. S. Liu, "Casnp: a database for interrogating copy number alterations of cancer genome from snp array data," *Nucleic acids research*, vol. 39, no. suppl 1, pp. D968–D974, 2011. pages 87

[132] A. J. Iafrate, L. Feuk, M. N. Rivera, M. L. Listewnik, P. K. Donahoe, Y. Qi, S. W. Scherer, and C. Lee, "Detection of large-scale variation in the human genome," *Nature genetics*, vol. 36, no. 9, pp. 949–951, 2004. pages 88

[133] M. L. Metzker, "Sequencing technologies-the next generation," *Nature Reviews Genetics*, vol. 11, no. 1, pp. 31–46, 2009. pages 104, 105, 108

[134] J. Shendure *et al.*, "Next-generation human genetics," *Genome Biol*, vol. 12, no. 9, p. 408, 2011. pages 105

[135] T. Kalisky and S. R. Quake, "Single-cell genomics," *Nature methods*, vol. 8, no. 4, pp. 311–314, 2011. pages 105

[136] N. Navin and J. Hicks, "Future medical applications of single-cell sequencing in cancer," *Genome medicine*, vol. 3, no. 5, pp. 1–12, 2011. pages 105

[137] T. LaFramboise, N. Dewal, K. Wilkins, I. Pe'er, and M. L. Freedman, "Allelic selection of amplicons in glioblastoma revealed by combining somatic and germline analysis," *PLoS genetics*, vol. 6, no. 9, p. e1001086, 2010. pages 107

[138] B. A. Weir, M. S. Woo, G. Getz, S. Perner, L. Ding, R. Beroukhim, W. M. Lin, M. A. Province, A. Kraja, L. A. Johnson, *et al.*, "Characterizing the cancer genome in lung adenocarcinoma," *Nature*, vol. 450, no. 7168, 2007. pages 108

[139] P. M. Haverty, J. Fridlyand, L. Li, G. Getz, R. Beroukhim, S. Lohr, T. D. Wu, G. Cavet, Z. Zhang, and J. Chant, "High-resolution genomic and expression analyses of copy number alterations in breast tumors," *Genes, Chromosomes and Cancer*, vol. 47, no. 6, pp. 530–542, 2008. pages 108

[140] M. Kadota, M. Sato, B. Duncan, A. Ooshima, H. H. Yang, N. Diaz-Meyer, S. Gere, S.-I. Kageyama, J. Fukuoka, T. Nagata, *et al.*, "Identification of novel gene amplifications in breast cancer and coexistence of gene amplification with an activating mutation of pik3ca," *Cancer research*, vol. 69, no. 18, pp. 7357–7365, 2009. pages 108

[141] P. A. Futreal, L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and M. R. Stratton, "A census of human cancer genes," *Nature Reviews Cancer*, vol. 4, no. 3, pp. 177–183, 2004. pages 108

[142] M. R. Stratton, "Exploring the genomes of cancer cells: progress and promise," *Science Signaling*, vol. 331, no. 6024, p. 1553, 2011. pages 108

[143] S. Nik-Zainal, L. B. Alexandrov, D. C. Wedge, P. Van Loo, C. D. Greenman, K. Raine, D. Jones, J. Hinton, J. Marshall, L. A. Stebbings, *et al.*, "Mutational processes molding the genomes of 21 breast cancers," *Cell*, 2012. pages 108

[144] N. Navin, J. Kendall, J. Troge, P. Andrews, L. Rodgers, J. McIndoo, K. Cook, A. Stepansky, D. Levy, D. Esposito, *et al.*, "Tumour evolution inferred by single-cell sequencing," *Nature*, vol. 472, no. 7341, pp. 90–94, 2011. pages 109

List of publications

Journal papers

- **Cheng J.**, Janssens S., and Buys N. (2009). Full sib pens of pigs are not suitable to identify variance component of associative effect: a simulation study using Gibbs Sampling. BMC genetics, 10:9.
- Van Calster B., Vanden Bempt I., Drijkoningen M., Pochet N., **Cheng J.**, Van Huffel S., Hendrickx W., Decock J., Huang H.-J., Leunen K., Amant F., Berteloot P., Paridaens R., Wildiers H., Van Limbergen E., Weltens C., Timmerman D., Van Gorp T., Smeets A., Van den Bogaert W., Vergote I., Christiaens M.-R., Neven P. (2009). Axillary lymph node status of operable breast cancers by combined steroid receptor and HER-2 status: triple positive tumors are more likely lymph node positive. Breast Cancer Research and Treatment, 113 (1), 181-187.
- **Cheng J.**, Vanneste E., Konings P., Voet T., Vermeesch J., Moreau Y. (2011). Single-cell copy number variation detection. Genome Biology, 12(8), R80.
- Mertzanidou A.*, Wilton L.*, **Cheng J.***, Spits C., Vanneste E., Moreau Y., Vermessch J.R., Sermon K. (2012). Microarray analysis reveals abnormal chromosomal complements in over 70% of 14 normally developing human embryos. Human Reproduction, ahead of print (*Joint first author).
- Van der Aa N., **Cheng J.**, Mateiu L., Zamani Esteki M., Dimitriadou E., Kumar P., Vanneste E., Moreau Y., Vermeesch J., Voet T. (2013). Genome-wide copy number profiling of single cells in S-phase reveals DNA-replication domains. Nucleic Acids Research, ahead of print.
- **Cheng J.**, Wedge D.C., Pitt J.J., Russnes H.G., Moen Vollan H.K., Pandey, B.P., Nilsen G., Nord S., Bignell G.R., White K. P., Børresen-

Dale A.L., Campbell P.J., Kristensen V.N., Stratton M.R., Lingjærde O.C., Moreau Y., Van Loo P. (2013). Copy number meta-analysis across primary tumors highlights distinct genomic architecture and identifies tumor suppressors. (Submited to "Cancer Discovery").

- Zamani Esteki M., Dimitriadou E., Mateiu L., Melotte C., Jackmaert S., **Cheng J.**, Van der Aa N., Kumar P., Mertzanidou A., Legius E., Moreau Y., Debrock S., D'Hooghe T., De Rycke M., Sermon K., Vermeesch J.R. and Voet T. (2013) Whole-genome single-cell Haplotyping, a generic method for preimplantation genetic diagnosis (In preparation).
- Yen J., White R.M., Wedge D.C., Van Loo P., De Ridder J., Cappe A., Richardson J., Jones D., Raine K., Watson I., Wu C.J., **Cheng J.**, Martincorena I., et al. (2013). Defining the genetic heterogeneity of engineered melanoma using zebrafish models. (Submitted to "Proceedings of the National Academy of Sciences of the United States of America")

Oral presentations

- Cheng J., Thienpont B., Konings P., Vanneste E., Voet T., Vermeesch J., Moreau Y. (2010). Using statistical model to correct genomic wave in high density oligo arrays. Oral presentation in a contributed Oral 05 Session at XXVth International Biometric Conference, Floripa, Brazil.
- Cheng J., Wedge D.C., Pitt J.J., Pandey, B.P., Lingjærde O.C., Moreau Y., Van Loo P. (2013). Allelle-specific copy number analysis across primary tumors to identify gene roles. Oral presentation at 13th Annual meeting of the Belgian Society of Human Genetics, Brussels, Belgium.

Prize

- 2012 Chinese government award for Outstanding Self-financed Students Abroad. (10 winners in 2012, Belgium.)

Curriculum Vitae

Jiqiu Cheng was born in Shenyang, China on August 31th, 1982. In the year 2004, she obtained a Bachelor degree in Computer Science from the University Aeronautical Engineering, Shenyang, China. In the year 2005, she obtained a Master degree in Bio-informatics from the university K. U. Leuven, Belgium. In the year 2007, January, she obtained an advanced Master degree in Statistics from the university K. U. Leuven, Belgium with Cum laude. From January, 2007 to October, 2008, she worked for the project "Incorporating competitive effects in animal model" as a BAP in Division of Gene Technology, Department of Biosystems, K.U. Leuven. In November, 2008, she started a PhD program in the bioinformatics group of the Department of Electrical Engineering (Esat-SCD) of the Katholieke Universiteit Leuven. Her research topic is "Development of computational approaches to detect copy number alterations in human genomes" under the supervision of Prof. Yves Moreau.







FACULTY OF ENGINEERING
DEPARTMENT OF ELECTRICAL ENGINEERING (ESAT)
SISTA, BIO-INFORMATICS GROUP
Kasteelpark Arenberg 10 - bus 2446
B-3001 Heverlee
jiqiu.cheng@esat.kuleuven.be
http://www.esat.kuleuven.be/scd/index.php