Study of complex chromosomal rearrangements in cancer

The role of extrachromosomal circular DNA as a genome remodeler in neuroblastoma

Elias Rodriguez Fos

Study of complex chromosomal rearrangements in cancer

The role of extrachromosomal circular DNA as a genome remodeler in neuroblastoma

Elias Rodriguez Fos

Programa de doctorat Biomedicina HDK05

Facultat de Biologia, Universitat de Barcelona

Study of complex chromosomal rearrangements in cancer

The role of extrachromosomal circular DNA as a genome remodeler in neuroblastoma

Tesi realitzada al

Barcelona Supercomputing Center (BSC)

Memòria presentada per Elias Rodriguez Fos

per optar al grau de doctor per la Universitat de Barcelona

Doctorand

Elias Rodriguez Fos

Director

David Torrents Arenales

Tutor

Josep Lluís Gelpí Buchaca

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

LIBERTAS PERFVNDET · OMNIA LVCE
UNIVERSITAT DE BARCELONA

Acknowledgements

En estos años de tesis, han sido muchas las personas que, de un modo u otro, han contribuido al desarrollo de esta y a todo lo que la rodea. Me han apoyado en las decepciones y celebrado los éxitos. Éste es mi pequeño agradecimiento.

En primer lugar agradezco a Núria López-Bigas y a Xavier de la Cruz su labor de seguimiento, así como sus preguntas y consejos; a Alex Kentsis y Anton Henssen el haberme brindado la oportunidad de colaborar con ellos en los dos estudios que representan la parte central de esta tesis. Dar las gracias también a mis supervisores Josep Lluís Gelpi y David Torrents, con una mención especial a este último ya que, como director de tesis, me ha enseñado a no tirar la toalla cuando no salen las cosas y a entender que la investigación no se basa únicamente en tener buenos resultados. Me ha aconsejado cómo tratar con colegas, afrontar colaboraciones, revisiones de artículos y mi carrera científica. De no ser por él, no habría podido poner un pie en este campo y no creo que pueda agradecérselo lo suficiente.

Entre la gente que ya no está en el grupo, pero con la que he tenido la suerte de compartir momentos de trabajo y no-trabajo, se encuentra Josep Maria Mercader, mi supervisor de prácticas durante mi primera etapa en el BSC. Txema me orientó hacia la bioinformática y, al igual que David, fue el responsable de mi entrada en el mundo de la investigación y por ello siempre le estaré agradecido. A lo largo del tiempo, se ha convertido en un amigo al que admiro y con el que he aprendido que ser perseverante y dejarse la piel trabajando en algo que te gusta vale la pena.

Otra mención especial es para Santi, con el que empecé a trabajar en la genética del cáncer. Mientras escribía su tesis, encontró tiempo para enseñarme y aconsejarme. Siempre he admirado lo pragmático que es a la hora de encarar un problema. Por todo ello, lo considero un mentor de quien sigo aprendiendo.

Agradezco a Sílvia la paciencia que tuvo cuando entré en el grupo sin haber abierto una terminal y cómo se sentó durante semanas para repasar scripts conmigo. Forma parte de la que recuerdo como la mejor época del doctorado, cuando las reuniones en su casa y las discusiones sobre política estaban a la orden del día. La determinación que pone en todo que hace es increíble.

No imagino un doctorado sin Marta y Mercè al lado. Son dos personas excepcionales que han hecho que durante estos años ir a trabajar sea un placer. Cuando los resultados no salían, siempre han estado allí con una sonrisa y un momento para levantarme el ánimo. Han demostrado ser las mejores compañeras y su amistad representa una de las mejores cosas que me llevo de esta etapa. Espero que, a pesar de que el doctorado acabe, nuestros caminos se vuelvan a encontrar.

Quiero agradecer al resto del grupo su apoyo y *feedback* durante las reuniones, en especial a Montse porque sin ella el grupo estaría perdido y yo no podría haber corrido y analizado todos los datos que presento en esta tesis.

Durante estos años, muchos amigos han aportado su granito de arena a esta tesis al recordarme que no todo es trabajar o pensar en ciencia. Agradezco a Zepo, con el que he crecido y compartido cine, literatura y música, y que ha seguido paso a paso cada submission, review, rechazo y aceptación de mis artículos. A Marcel que, como un hermano, me ha acompañado siempre en todo sin poner un pero. Ojalá podamos seguir arreglando el mundo como lo hemos hecho muchas noches. "*Somos...*". A Álex y Anna por esos viajes y fines de semana, porque hay vida más allá del doctorado. Si ya quería a Galicia, los *petits* hacen que la tenga siempre presente. A mi familia de Santa Coloma, en especial a Boleas, Miki y Ferri, que han conseguido que sienta esa ciudad como mi casa y con los que he pasado algunos de los mejores momentos de estos años. También quiero agradecer a Marc, Alberto, Sergi y Edu, que me han visto crecer y me han apoyado y aconsejado en cada paso. A los *Sissy Bars* por acompañarme en la aventura de cruzar España y Europa en unos viajes que me han ayudado a desconectar y pensar.

Quiero agradecer a mis padres el apoyo constante y la libertad para escoger mi camino. A mi hermana Andrea, que me ha levantado el ánimo y sacado una sonrisa siempre, y que a pesar de que hace ya un año que cruzó el charco, pienso cada día en ella.

Por último, quiero darle las gracias a Irene, que ha sido mi compañera estos años. Ella me aconsejó y animó para enviar el mail a Txema y David que me permitió empezar las prácticas en el BSC y posteriormente el doctorado. Ha sido un apoyo constante, no sólo con la tesis, sino con todo. Ha aguantado envíos de papers, rechazos, nervios antes de congresos, la redacción de esta tesis y todo ello animándome y sacándome una sonrisa. No puedo imaginarme haber compartido todo esto y lo que queda con nadie que no sea ella.

No quiero acabar sin antes darle las gracias a *Mme*. Torres, mi profesora de bachillerato que con su amabilidad y apoyo hizo, sin querer, que eligiera estudiar biología.

Abbreviations

ASCII	American Standard Code for Information Interchange
ARMD	*Alu* Recombination-Mediated Deletion
bp	base pair
BSC	Barcelona Supercomputing Center
CGP	Cancer Genome Project
CNV	Copy Number Variation/Copy Number Variant
COSMIC	Catalogue Of Somatic Mutations In Cancer
ddNTPs	modified di-deoxynucleotidetriphosphates
DDR	DNA Damage Response
DDX1	DEAD-Box Helicase 1 gene
DNA	Deoxyribonucleic Acid
dNTPs	normal deoxynucleotidetriphosphates
DSBs	Double Strand Breaks
eccDNA	small copy neutral extrachromosomal circular DNA
ecDNA	large amplified extrachromosomal circular DNA
FoSTeS	Fork Stalling and Template Switching
HERVs	Human Endogenous Retroviruses
HGP	Human Genome Project
HR	Homologous Recombination
ICGC	International Cancer Genome Consortium
indel	insertion and deletion
kb	kilobase pair
LINEs	Long Interspersed Nuclear Elements
LOH	Loss Of Heterozygosity
LTRs	Long Terminal Repeats
Mb	Mega base pair
MDM2	*MDM2* proto-oncogene

MMBIR	Microhomology-Mediated Break-Induced Replication
MMEJ	Microhomology-Mediated End Joining
MYCN	*MYCN* proto-oncogene
NAHR	Non-Allelic Homologous Recombination
NGS	Next-Generation Sequencing
NHEJ	Non-Homologous End Joining
p53	tumor protein p53
PCAWG	Pan-Cancer Analysis of Whole Genomes
PCR	Polymerase Chain Reaction
PGBD5	*PiggyBac* transposable element derived 5 gene
RNA	Ribonucleic Acid
SINEs	Short Interspersed Nuclear Elements
SMRT	Single-molecule real-time sequencing
SNV	Single Nucleotide Variant
ssDNA	single strand Deoxyribonucleic Acid
SVs	Structural Variants
TCGA	The Cancer Genome Atlas
TERT	Telomerase Reverse Transcriptase gene
TEs	Transposable Elements
WGS	Whole-Genome Sequencing
WHO	World Health Organization

Contents

Prologue 14

Introduction 18

1 19
Global cancer burden

2 22
Studying the human genome

2.1 22
From classic genetics to genomics: a historical overview of the emergence of genomics

2.2 26
The role of sequencing technologies

2.2.1 Sanger sequencing

2.2.2 Next-generation sequencing

2.2.3 Mapping / assembly of the genome

3 32
The cancer genome

3.1 32
Genomic instability in the cancer cell

3.1.1 The role of driver and passenger mutations

3.2 36
Somatic variation in cancer

3.2.1 Types of somatic variants

3.2.2 The role of structural variation

3.2.3 Mechanisms generating structural variants

3.2.4 Transposable elements: an example of SV generators

The particular case of Alu elements.

3.2.5 Identification of structural variants

Variant calling in cancer.

3.3 49
Tumor evolution and drug resistance

3.4 51
Cancer projects

4 54
Circular DNA elements in cancer

4.1 54
Types of extrachromosomal circular DNA

4.2 55
Generation of circular DNA elements

4.3 56
Function and impact of circular DNA in cancer

4.3.1 The role of circular DNA in oncogene amplification and tumor heterogeneity

4.3.2 Circular DNA as a vehicle for oncogenic overexpression

4.3.3 Other implications of circular DNA in genomic instability

4.4 61
Identification of circular DNA

5 63
Neuroblastoma: a cancer example

5.1 65
Proto-oncogene amplification in high-risk neuroblastoma

5.2 65
Other important genomic alterations in high-risk neuroblastoma

Objectives 68

Contribution to publications 70

PGBD5 publication 78

1 79 Abstract

2 80 Main

3 82 Results

3.1 82 Human rhabdoid tumors exhibit genomic rearrangements associated with PGBD5-specific signal-sequence breakpoints

3.2 86 PGBD5 is physically associated with human genomic pss sequences that are sufficient to mediate DNA rearrangements in rhabdoid tumor cells

3.3 89 PGBD5 expression in genomically stable primary human cells is sufficient to induce malignant transformation in vitro and in vivo

3.4 90 PGBD5-induced cell transformation requires DNA transposase activity

3.5 92 Transient expression of PGBD5 is sufficient for PGBD5-induced cell transformation

3.6 93 PGBD5-induced transformation requires DNA end-joining repair

3.7 94 PGBD5-induced cell transformation involves site-specific genomic rearrangements associated with PGBD5-specific signal-sequence breakpoints

4 98 Discussion

5 102 Methods

6 112 References

7 116 Authors information

8 117 Supplementary information

PGBD5 PCAWG analysis 118

1 Abstract 119

2 Introduction 120

3 Results 121

3.1 PGBD5-motif-related deletions are present in different cancer types conserving its characteristic features 121

3.2 PGBD5-motif-related deletions present high recurrence across different patients and cancer types 124

3.3 A fraction of deletions not presenting the PGBD5 motif show recurrence and different homologous sequences around the breakpoints 126

3.4 Manual curation of a subset of recurrent motif-related deletions revealed fluctuation in the number of reads supporting the rearrangements 128

3.5 Classification of motif-related deletions and artifact-related deletions 131

4 Discussion and conclusions 132

5 Methods 135

6 References 138

7 Supplementary information 139

Neuroblastoma publication 142

1 143 Abstract

2 144 Main

3 155 Methods

4 165 References

5 167 Authors information

6 167 Supplementary information

Results and Discussion 168

Conclusions 172

Bibliography 174

Appendix 180

1 181 Supplementary figures of the PGBD5 publication

2 193 Supplementary figures of the neuroblastoma publication

3 207 Appendix publication 1

Enhancer hijacking determines intraand extrachromosomal circular *MYCN* amplicon architecture in neuroblastoma

4 236 Appendix publication 2

Re-analysis of public genetic data reveals a rare X-chromosomal variant associated with type 2 diabetes

Prologue

This thesis illustrates the work I have developed as a Ph.D. student in the computational genomics group lead by Dr. David Torrents at the Barcelona Supercomputing Center. The group's expertise in the analysis of biological data and the detection of variants to gain more knowledge about the genetic and molecular implications of human diseases, such as cancer, has allowed me to learn and conduct my research.

Focusing on the analysis of structural variation in cancer, I have been able to apply different methodologies for sequencing data, retrieving, filtering, and determining the mutational profile for each of the studied samples. Moreover, I have characterized new patterns of genomic rearrangements related to transposase-derived genes and extrachromosomal circular DNA elements in cancer. Therefore, this thesis is centered in the study of the genomic variation and mechanisms associated with oncogenic processes together with the analysis of elements of the human genome that are not generally included in comprehensive cancer studies, such as circular DNA elements.

In summary, starting with the introduction, I give an overview of the methodological aspects of the study of cancer through the impact of sequencing technologies, the biological and molecular causes and consequences of this disease, focusing on structural variation, and the description of the circular DNA genomic component and its known implications in cancer. Finally, I introduce neuroblastoma, an example of how structural variants and circular DNA drive tumorigenesis.

Next, I present the results of this thesis in three blocks, all of which have in common the study of structural variation in cancer. Two of the blocks correspond to the *PGBD5* and neuroblastoma publications, and one corresponds to the continuation of the *PGBD5* analysis in ICGC-Pan-Cancer data.

As an overview of the trajectory of this thesis, I started with my involvement in a project focused on analyzing the role of *PGBD5* —a transposase-derived gene—

as an oncogenic mutator with an associated mechanism for site-specific DNA rearrangements. In this study, we describe how the expression of this gene promotes cell transformation and the generation of recurrent rearrangements, presenting a conserved motif in cell lines and childhood tumors. As a logical continuation of this publication and thanks to the access of our group to ICGC-PCAWG data, we expanded the study of these characteristic *PGBD5*-motif-related rearrangements to different patients and tumor types.

The following part of this thesis is focused on the analysis, description, and classification of the genomic somatic rearrangements in neuroblastoma. With the aim of better grouping the patients with different clinical outcomes, we searched for differential patterns of structural variants across the samples. From this analysis, we were able to describe a new phenomenon that connects circular DNA with different integration sites around the genome through complex rearrangement clusters providing evidence on how circular DNA can act as a driver of genomic remodeling in neuroblastoma.

To finalize, I present the general discussion of the results and questions addressed in this work to, then, end up disclosing the final conclusions of this thesis.

Introduction

1 Global cancer burden

Social improvements, economic growth, and advances in health care over the last decades have a direct impact on the rise of life expectancy and the changes in the causes of death across the world. In this period, cancer has become the second major cause of mortality in the globe[1]. Its numbers are increasing every year, with more than 8.9 million deaths in 2016 and 9.6 million deaths in 2018, according to the World Health Organization (WHO), and affect populations from all regions and different incomes[2,3]. Although cancer's major impact is in low and middle-income countries, where infections, poverty, and difficult access to health care limit the diagnosis and treatment, its extent is worldwide. Contrary to what it might be expected, the incidence of different types of cancer is also growing in high-income countries, due to the adoption of new lifestyles where tobacco, alcohol, processed foods, pollution, and sedentarism are increasingly present. In the last decade, the incidence of cancer has grown from 14.1 million people affected in 2012 to 18.1 million people affected in 2018 (according to the last World Cancer Report from WHO, Fig. 1). The current trend predicts that cancer mortality and morbidity will increase by 70% in the next years, worldwide[2,3].

Focusing on childhood cancers, which appear before the age of 15 years and are distinct from cancers arising in adults, the overall incidence rates globally follow the same general trend. Childhood cancers are rare diseases representing 0.5-4.6% of the total number of cancer cases in the world, of which 20% are associated with embryonal tumors such as neuroblastoma, which is one of the subjects of this thesis. Unfortunately, the differences in mortality rates between distinct income regions are more pronounced when it comes to these specific types of cancer. In the last 50 years, the rates of survival in high-income countries have risen from 30% to 80%, contrary to what happens in lower-income countries where, for example, 93% of the childhood cancers related deaths occurred in 2012[3]. The marked differences in children's survival between regions point to a necessary worldwide collaboration in cancer research, which nowadays is limited to some countries.

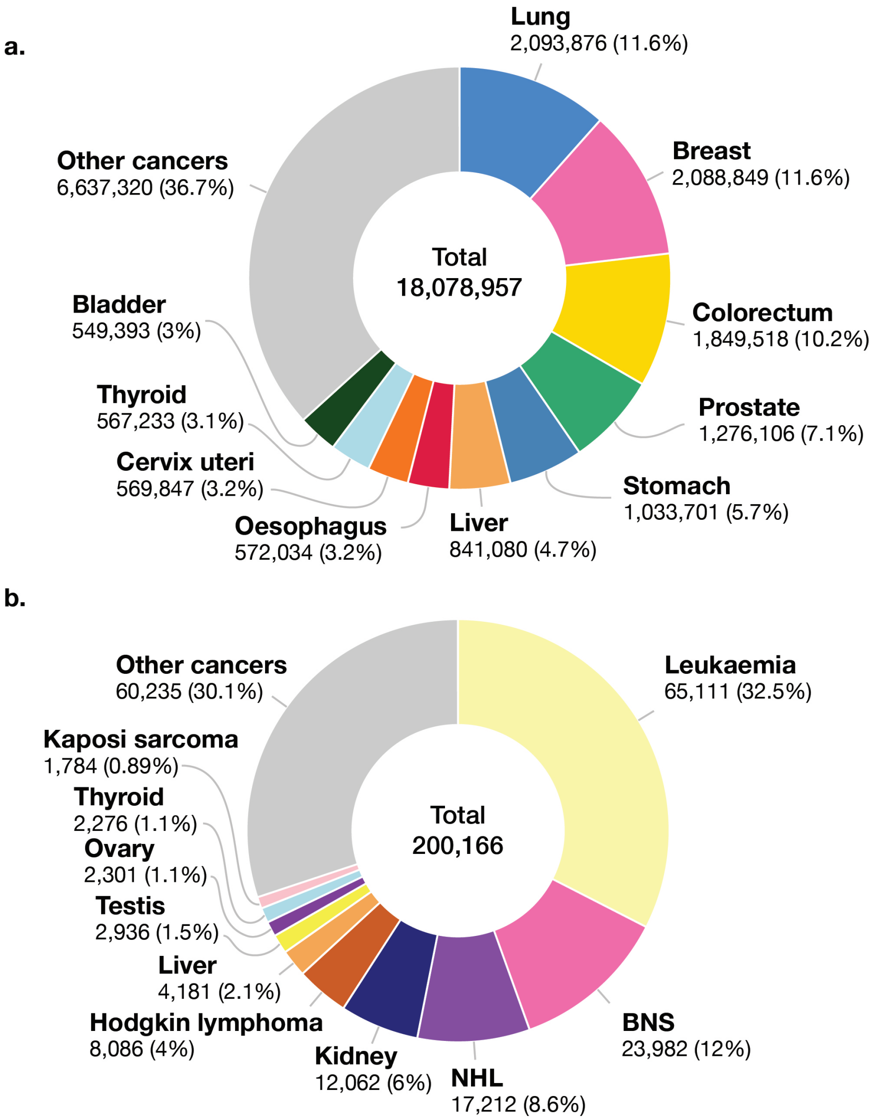


Figure 1. Incidence of adult and childhood cancers worldwide.

(**a**), Estimated number of cancer cases worldwide in 2018, taking into account a population from both sexes and ages from 0 to 85+. The three types of cancer with more incidence are lung, breast, and colorectum cancer. (**b**), Estimated number of childhood cancer cases worldwide in 2018, taking into account a population from both sexes and ages from 0 to 15 (not included). The three types of cancer with more incidence are leukemia, BNS (brain, central nervous system), which includes neuroblastoma, and NHL (Non-Hodgkin Lymphoma). *Data source: GLOBOCAN 2018. Adapted from: Global Cancer Observatory (http://gco.iarc.fr), International Agency for Research on Cancer 2020.*

The importance of cancer is unquestionable, and all the efforts are made to meet the objective of reversing the rise of these diseases globally. In order to change the worst predictions, global combined work focused on the prevention, diagnosis, and treatment of cancer is mandatory. To better achieve and execute each of these innovations, more extensive knowledge of the diseases is needed. Following this need, scientific and medical communities have acquired a leading role in understanding the causes of cancer and how it affects individuals and populations. Through research and collaboration in the different medical and omics fields, they have been able to achieve therapeutic advantages through improving cancer prevention —finding new opportunities for early detection—, diagnosis —identifying mutations related to the different diseases—, and treatment —determining new molecular targets—. Unfortunately, there still are limitations in dealing with cancer, such as access to effective health care, affordable drugs, or data sharing, that cannot be fully addressed from research. In summary, more comprehensive knowledge, control, and advantage over cancer are obtained through research and clinical innovation, which are essential steps in reducing the impact of the disease worldwide. However, in order to help biomedical research reach the patients, a serious social, legal, and political compromise is also needed.

2 Studying the human genome

The study of complex diseases, and particularly cancer, has changed over time with the rise of genetics, and the emergence of genomics[4], made possible by the development of next-generation sequencing (NGS) techniques. Traditionally, human genetic studies were conceived from function to genetics. They first started with the analysis of the physiology of the disease, characterizing the molecular and cellular function of a given protein, searching for candidate genes or regions, to associate them to the particular disease, and finally demonstrate this association through the analysis of patient's genomes. Nowadays, with the improvement of sequencing techniques, which confer the ability to look at the mutational spectrum of an organism at a genome-wide scale, studies are planned and developed the other way around, from genomics to function. They start with the sequencing of the whole genome from the patient to identify mutations in candidate genes or regions. Finally, the functional implications of these candidates are determined in order to associate the molecular findings to the malignancy. Nevertheless, NGS not only has been a significant contribution to the genomic characterization of cancer and other diseases but has also played a major role in the definition of a comprehensive picture of human genome variation[5,6], which helps us better understand our biology and how it is affected by diseases.

Following this idea, a historical overview of the evolution of genetics and genomics is presented here below, along with the contribution of the different sequencing technologies to the human genome analyses.

2.1 From classic genetics to genomics: a historical overview of the emergence of genomics

In the 19th century, Gregor Mendel[7,8] (1822-1884) studied the patterns of inheritance in pea plants and how different characteristics such as color, shape, and size are passed down across generations following mathematical models. From his work, he established specific patterns to predict the traits in the progeny according to the parental characteristics. Based on these findings, he formulated the laws that established the principles of genetic inheritance and are still considered the base of modern genetics. Mendel's findings started the path for the study of genetics and the subsequent study and characterization of the DNA, the molecule that encompasses all the genetic information of the organism, and that is responsible for the transfer of the hereditary traits.

The first approximation of what it would be later identified as DNA was made in 1869 by Swiss physiological chemist Friedrich Miescher[9] (1844-1895) inside the nucleus of human white blood cells. In 1866 and 1889, respectively, Mendel and Hugo De Vries[9] (1848-1935), mentioned in their corresponding publications the existence of transmissible elements or "pangenes" describing the presence of inheritable particles in the organism. Following this idea, in 1909, Wilhelm Johannsen[9] (1857-1927) finally coined the term gene[10] to designate the hereditary elements of the cell. Genetics became, then, the study of genes and heritability. Johannsen also introduced the concepts of genotype and phenotype, which nowadays refer to the whole set of genetic elements and observable traits of a living organism. Although the concepts of inheritance, gene, and the presence of nucleic acids in the cell nucleus were already established, we had to wait until the mid-20th century to connect heritability and genes to the DNA molecule.

Based on the contributions made in the 20th century by scientists such as Phoebus Levene[9] (1869-1940), who characterized the structure of the nucleotides and its components, and Erwin Chargaff[9,11] (1905-2002), who described the relationship between the different nitrogenous bases, James Watson (1928-present) and Francis Crick (1916-2004) with the decisive contribution of Rosalind Franklin (1920-1958) were able to propose in 1953, the well-known three-dimensional, double-helical structure of DNA[9,12] (Fig. 2). In the same period, Alfred Hershey (1908-1997) and Martha Chase (1927-2003) confirmed the DNA's role in heredity, describing it as the carrier of the genetic information in the cell[13]. Relying on all these discoveries, a few years later, Crick stated the central dogma of molecular biology[14] in which he

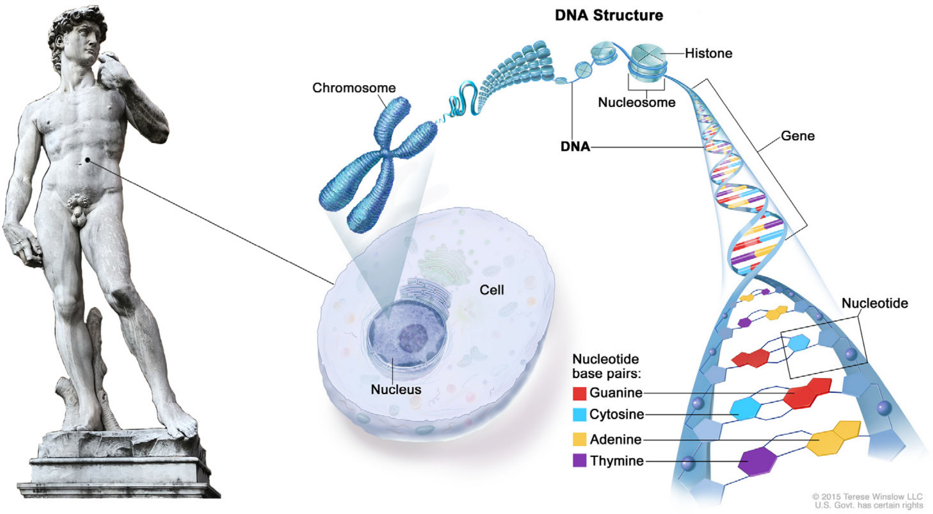


Figure 2. From the cell to the DNA sequence.

Representation of the DNA molecule located in the cell nucleus through its different compaction states. It illustrates the DNA states from the highly compacted chromosomal structure to the relaxed double-helix showing the nucleotide composition of the DNA and its characteristic quaternary code (A, C, G, T). *Image source: For the National Cancer Institute. Copyright 2015. Terese Winslow LLC, U.S Govt. has certain rights.*

described the flow of the genetic information, from genes to proteins inside a living organism. His description of the transcription from the stable DNA molecule to the RNA and its posterior translation to functional proteins is still the basis on which all molecular biology study relies. The discovery and characterization of the DNA structure, sequence, and function redefined the concept of the gene at a molecular level and positioned the DNA molecule as the central element of life.

An excellent example of a collection of genetic diseases is cancer, which is primarily associated with the accumulation of changes that affect DNA and genes. Traditionally, genetic studies are typically focused on the role of individual genes and their relationship with the disease. However, in recent years, research, in general, and cancer studies, in particular, have evolved towards more integrative and broad strategies in what is known as omics studies. In order to gain more knowledge about the genetics of an organism, or a disease such as cancer and how it affects the whole organism from a genome-wide perspective, genomics studies emerged.

The genomic field made its big appearance at the end of the 20th century with the Human Genome Project (HGP)[15-17] and the development of sequencing technologies, and it is still expanding. The HGP was a collaborative project between 2,800 researchers from different universities and research centers all over the world that started in 1990 and ended in 2003 with the complete release of the entire human genome. It was the first project to determine the sequence of all the bases in human DNA, map and locate the genes of our genome in the different chromosomes, and produce linkage maps to study inherited traits. The HGP also represents one of the most significant examples of how collaboration in research is essential to generate new fundamental knowledge that helps the whole community and humanity understanding life and diseases. Before its final publication, in 2001, Francis Collins (1950-present) illustrated the importance of studying our genome with the following words: "*It's a history book - a narrative of the journey of our species through time. It's a shop manual, with an incredibly detailed blueprint for building every human cell. And it's a transformative textbook of medicine, with insights that will give health care providers immense new powers to treat, prevent and cure disease*"[18].

Genomics, therefore, represents the study of the entire organism's genetic material, taking into account not only a specific gene and its heritability but all the genes and the different functional elements of the genome. One of the crucial aspects of genomic studies is their interdisciplinarity. They focus on different characteristics of genomes such as sequence, structure, function, regulation, and the interaction of those elements with each other and the environment. This approach is especially useful in the study of complex diseases such as cancer since they are caused by multiple genetic and environmental factors, which can only be addressed from a genomic approach.

Overall, genomic research expands genetic studies by analyzing the structure and function of the whole genome, using large datasets mainly formed by sequencing data to find changes, characteristics, and interrelations in the DNA that can be associated with cancer or other diseases. Due to its greater complexity and comprehensive aspect, genomics can, in some ways, be considered the future of genetics. Paraphrasing Collins, genomic studies take a look into the entire blueprint of the cell.

2.2 The role of sequencing technologies

Watson, Crick, and Franklin described the structure of DNA, depicting a long sequence of base pairs that encompass all the genetic information of our organism (Fig. 2). All the information carried by the DNA molecule, and by extension by the whole genome, is codified using a quaternary code with four bases: A, C, G, T. We can think about this code the same way we think about the binary ASCII code with 0 and 1 or even the alphabet which is a code with 26 different symbols. For example, by ordering the letters of the alphabet, we can write words, sentences that make more or less sense, or even beautiful masterpieces such as *East of Eden* (John Steinbeck, 1952). It is fascinating to think of the genomic code the same way. For this reason, in an effort to try to read and decipher the whole genome, sequencing technologies were developed.

With the publication of the human genome sequence, we entered the post-genomic era[19], where increasing amounts of genomic data are generated, analyzed, stored, and shared. One of the most important processes of this workflow is the generation of the data, involving different steps such as sample preparation, sequencing, and alignment or *de novo* assembly. The generation of genomic data, cannot be understood without the emergence of sequencing technologies and has changed and evolved alongside the transition from first-generation to next-generation sequencing, the release of new builds of reference genomes (i.e., The last human genome version —hg38— corresponds to the 20th release), and the advance in alignment algorithms such as the Burrows-Wheeler Aligner (BWA)[20], among others. DNA sequencing processes have granted us access to the whole DNA sequence from an organism, which can be read and analyzed in order to classify species, do genomic annotation, understand the transcriptome, discover mutations and rearrangements that have applications in cancer research, and in summary, understand the function and complexity of an organism's genome.

2.2.1 Sanger sequencing

In the 1970s, Frederick Sanger (1918-2013) developed what we still call the Sanger sequencing method, the first robust and accurate technique to sequence DNA[21,22]. Later, in the 1980s, automated DNA sequencing machines based on the Sanger method were manufactured, providing sequencing data cheaper and faster than the traditional method by generating up to 96 long reads of DNA at a time. Sanger sequencing is based on the chain-termination technique, which makes use of chemical analogs of the DNA nucleotides. The first steps of the method include the amplification of the DNA we want to sequence, currently using polymerase chain reaction (PCR), followed by denaturalization in order to obtain single-strand DNA (ssDNA) molecules. Using the ssDNA molecules as the template along with a polymerase, primers, and normal and modified nucleotides (dNTPs and ddNTPs), the new complementary chain is elongated one nucleotide at a time until a marked ddNTP is incorporated terminating the extension. These nucleotides are tagged with fluorescent dyes that can be read directly after ordering the elongated fragments by size using electrophoresis. Once the reaction is over, the whole sequence of DNA can be read by detecting the marked nucleotides in each terminal position of each of the different resulting chains.

Although the Sanger method remained the prevalent sequencing method for 30 years and those sequencing machines were the ones allowing us to obtain the human genome for the HGP[23], new technologies more time and cost-effective have arisen in the past years relegating Sanger sequencing to smaller projects or validation purposes. In structural variation analysis, for example, validation with

Sanger sequencing is common due to its production of long reads from 500bp to 1kb in length. Long reads are useful in these analyses, especially when it comes to verifying intrachromosomal rearrangements such as deletions and insertions, and rearrangements within repetitive regions of the genome. The greater length of Sanger sequencing reads allows better alignment and mapping in those regions and more accurate detection of variants larger than the average short-read size.

2.2.2 Next-generation sequencing

New sequencing technologies described as high-throughput are named next-generation sequencing (NGS) as opposed to Sanger sequencing, which belongs to the first generation. The different companies developing NGS machines present different methodologies of sequencing such as pyrosequencing, sequencing by synthesis, and sequencing by ligation, among others. It is also interesting to notice that NGS can be performed not only on the whole genome but also on the whole exome, specific genes, and RNA[6], opening up a wide range of data that can be analyzed in different omic studies.

Sequencing by synthesis technique, provided by Illumina, offers the highest throughput per run and lowest cost per-base[24] and therefore is the most popular NGS method[25,26]. The first stage in NGS corresponds to template preparation, which starts with the ligation of adaptors at the end of each ssDNA fragment that has to be sequenced. The adaptors are used, among other things, to attach the DNA fragments that act as the template to the flow cell that serves as solid support for an efficient sequencing process. Commonly, the next step would be the clonal amplification of the DNA molecules by PCR, although, in the past years, due to the discovery of PCR-related artifacts, this step has become less and less popular, especially in genetic variation studies which some opt for a PCR-free sequencing methodology[26].

The following process, analogous to Sanger method, is the sequencing itself. This process begins with the incorporation of fluorescently labeled nucleotides, which block the newly generated DNA fragment to ensure that only a single base at a time is incorporated per cycle. Each cycle starts with the deblocking of the last base of the extended DNA fragment in order to accept another marked base. Cycles are then followed by a reading step that identifies each added nucleotide in each of the cycles in a massive parallel process occurring through all the flow cell obtaining the nucleotide sequence for all the millions of short reads.

The most significant advantage of NGS over the old techniques is the vast generation of data. NGS has the possibility of carrying out millions of sequencing reactions at the same time. It has been established that an NGS machine can sequence 15 individuals in a little more than three days. The reduced time improvement is directly associated with the reduced cost per sample[24,27], allowing the faster and cheaper sequencing of more samples, paving the way to the rise of international large-scale genomic projects such as ICGC-Pan-Cancer[28]. However, NGS is not exempted from limitations[25,29]. The most important limitation is the generation of short reads from 75bp to 150bp in length due to the increase of its sequencing error rate when generating longer reads. Because of the smaller read length, the detection of intrachromosomal rearrangements larger that the read size, and the mapping and assembly of the genome, which we will discuss further on, become more challenging. In consequence, new long-read NGS machines have been developed in the last years in order to overcome the short-read length limitation. However, unfortunately, techniques such as Single-Molecule Real-Time sequencing (SMRT) from PacBio with reads from 30kb to 100kb, still represent expensive options with higher sequencing error rates[24] and are nowadays integrated into the different studies as a complement to short-read NGS data.

2.2.3 Mapping/assembly of the genome

Once the sequencing process terminates, millions of reads have been produced and have to be ordered and assembled. Following the book analogy, we end up with a bag full of millions of shredded sentences from the book, which we have to reorder in a way that we can read it as it was. In order to achieve the reordering task, two different strategies are available depending on the application purposes: mapping or *de novo* assembly[30]. *De novo* assembly, as its name implies, is based on the assembly of the sequenced reads without any template or support with the aim of reconstructing the whole genome. It is a common strategy for the assembly of genomes for which we do not have a reference such as bacterial genomes or artificial chromosomes. The application of this type of assembly to the human genome is still limited, mainly due to the size and complexity of this genome and the small length of the reads produced by NGS.

On the other hand, mapping strategies are based on the alignment of the sequenced reads to a known reference genome that is used as a template. Ideally, it would be like rebuilding the book using a nearly identical book as a template. Unfortunately, it is not that simple, and it has more limitations[25] that we might think, although it works well enough to be the preferred strategy for identifying genomic variants in cancer research[29].

One of its major limitations is the difficulty of mapping short-reads within repetitive or poorly characterized regions around the genome such as SINEs, LINEs, transposable elements, satellites, centromeric and pericentromeric regions, for example. Pair-end read sequencing can resolve a percentage of these cases, as long as one of the reads of the pair maps in a unique region, but is still a restraint. Another limitation is the mapping within regions that may not exist in the reference genome, such as gaps or structural variants. This limitation is crucial in the study of cancer genomes, which are known to be highly rearranged. Nevertheless, NGS has been, and still is, the most prevalent method of sequencing for studies of genomic variation in cancer.

In summary, thanks to the advances in DNA-sequencing technologies, which facilitate the sequencing, assembling, and analysis of whole-genomes of different organisms, including humans, more and more genomic studies are carried out generating massive amounts of data[17]. In order to analyze and understand this data, the development of new fields such as bioinformatics has been necessary.

3 The cancer genome

Cancer has become a central topic in biomedical research as a consequence of its high impact on society. Although it may sometimes be treated like it, cancer is not a unique entity. It corresponds to a set of more than 100 distinct diseases with diverse risk factors, symptoms, incidence, and prevalence. Those diseases can originate from most tissues and cell types of the organism. Actually, cancer can start from a normal cell located almost anywhere in our body. Although all these diseases are different, they also present main characteristics that are common in all affected individuals[31].

All cancer diseases, as established in 1976 by Peter Nowell, share two essential complementary mechanisms that define cancer as a Darwinian evolutionary process[32]: the acquisition of genetic changes in the DNA of the cell and the natural selection of altered cells[31,33]. Those two mechanisms working together may confer survival advantage to cells that acquire beneficial mutations, allowing them to grow, proliferate, and invade other tissues. In essence, cancer can be defined as the clonal expansion or proliferation of abnormal cells in the organism. For this reason, understanding the complexity of cancer through the study of complex changes in the genetic material of the oncogenic cell, the mechanisms involved in the generation of the genomic instability, and the selective pressure undergone by those cells is critical to fighting the disease[34].

3.1 Genomic instability in the cancer cell

Contrary to what we might think, genomic instability is not exclusive to cancer cells. During the cell cycle, normal cells undergo cell division where they have to replicate the genome in order to provide a full copy of it to each of the two resulting daughter cells or clones. In the case of the human genome, each copy has a size of 3.2 billion base pairs, which corresponds to the complete set of DNA from the organism[35]. Generally, we delimit its composition to 23 pairs of chromosomes —22 pairs of autosomes and one pair of sexual chromosomes XX or XY—. However, as it will be illustrated in this thesis, this delimitation does not represent the full complexity of the genetic material of the cell. In addition to autosomes and sexual chromosomes, other elements in the cell carry genetic information that is also part of our genome and is also passed on during cell division, such as mitochondrial DNA, extrachromosomal linear DNA, or extrachromosomal circular DNA structures[36]. The replication and cell division processes ideally ensure that all cells from the organism have exactly the same genetic material.

Unfortunately, the steps taking place in the cell cycle are not entirely error-free, resulting in the presence and accumulation of alterations in the genome of the daughter cells. One type of change that can occur in the resulting daughter cells due to genome instability is the modification of the number of chromosomal structures —including linear and circular DNA— due to missegregation during cell division. However, the most prevalent change in the DNA corresponds to rearrangements in its sequence, driven by mutagenic processes of both internal and external origin, such as replication errors, DNA repair errors, transposition, viral integration, and exposure to carcinogens like tobacco smoke, among others[37]. Nevertheless, most of the changes that affect DNA are successfully repaired by in-cell mechanisms that operate to maintain genomic integrity. Still, a fraction of the acquired mutations remains unrepaired and ends up fixed in the genome.

The changes in the genome accumulated during our lifetime, which correspond to the ones not inherited from our parents, are known as somatic mutations. They contribute to the differences in the genetic profile of the cells of our organism. These mutations contain different types of DNA changes and rearrangements such as point mutations, deletions, insertions, and translocations, among others. However, not all somatic mutations have consequences in cancer transformation and development.

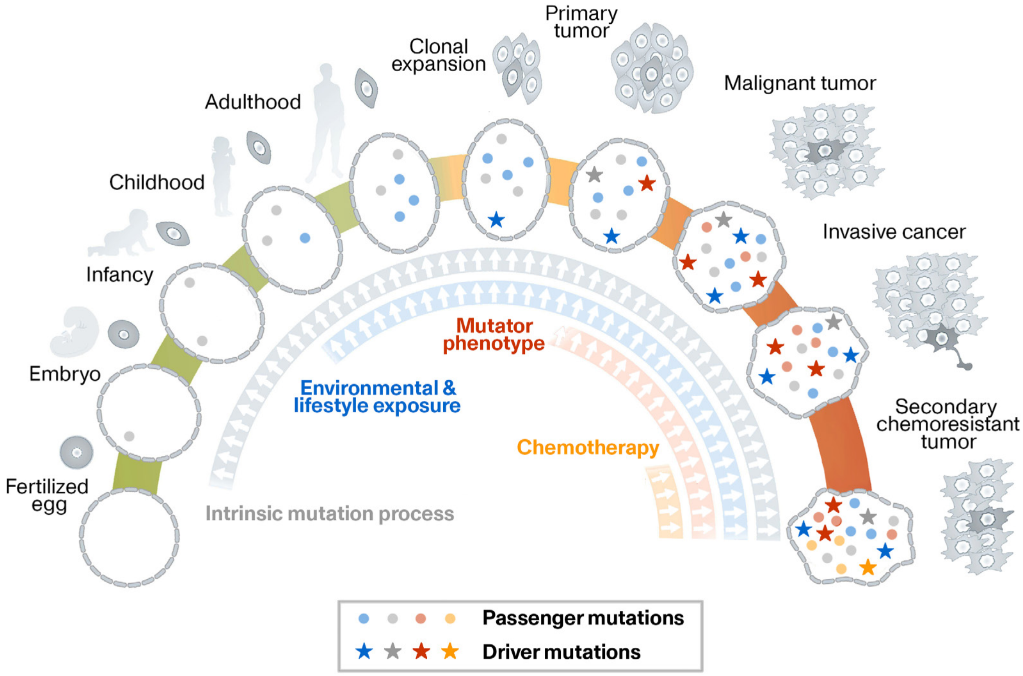


Figure 3. Genome instability in the cell: from the embryo to the malignant tumor.

Representation of the accumulation of passenger and driver (dots, stars) somatic mutations over a lifetime. Each color represents different mutational processes, such as intrinsic mutations (grey), environmental and lifestyle exposure (blue), mutator phenotype related to the acquisition of driver mutations starting the development of a tumor (red) and chemotherapy exposure, promoting resistance to the malignant cells (orange). *Adapted from reference*[108].

3.1.1 The role of driver and passenger mutations

There are two classes of somatic mutations that accumulate in the normal cell genome: passenger and driver mutations[31,38,39]. Passenger mutations are variants that have no phenotypic effect or do not confer selective advantage to the cell. We use to think of these mutations as harmless, although it is known that some of them can be mildly deleterious and can even have anticancer effects[40]. On the contrary, driver mutations are variants that confer a selective advantage to the cell, leading to its growth and survival. In the case of driver mutations, they can affect genes, regulatory elements, and lead to phenotypic consequences such as the malignant transformation of the cell. The differences between passenger and driver mutations show the fact that not all somatic mutations present in the genome are involved in the oncogenic transformation of the cell. In fact, from the thousands of mutations existing in the DNA, the vast majority are passenger or neutral for the cell, and only a few correspond to driver mutations implicated in oncogenesis[39]. However, passenger mutations, even if they do not actively take part in the oncogenic transformation, they do contribute to increasing the heterogeneity[41] of the genetic background of cancer cells (Fig. 3).

The genomic differences, accumulated between cancer cells within a single tumor, have been described as intratumor heterogeneity[42]. Moreover, variation in the landscape of mutations between patients harboring tumors of the same type has long been established. This mutational diversity has been associated with patient-specific factors and is defined as intertumor heterogeneity[42]. The existence of inter and intratumor heterogeneity entails an increase in the complexity in the study of genomic instability in cancer and has been associated with poor clinical outcome[43]. To the evident differences between tumor types, tumor heterogeneity adds the uneven distribution of passenger and driver alterations across different regions of the same tumor and across different tumors from the same cancer type.

The malignant transformation of the normal cell is defined by the existence of different driver mutations in its DNA. This can happen by a gradual accumulation of genetic alterations through each clone obtained from cell division or by a single catastrophic event that generates all the alterations[33]. Either way, the presence of these mutations in the cell acts as a driving event for tumor development. For this reason, the study of somatic mutations existing in cancer cells represents a record and source of information of all mutational processes that these cells, and by extension, the patient, have experienced during their lifetime[31]. From the study of these mutations, we can elucidate the primary mechanisms that play a role in the oncogenic process and find new genomic markers to improve cancer diagnosis.

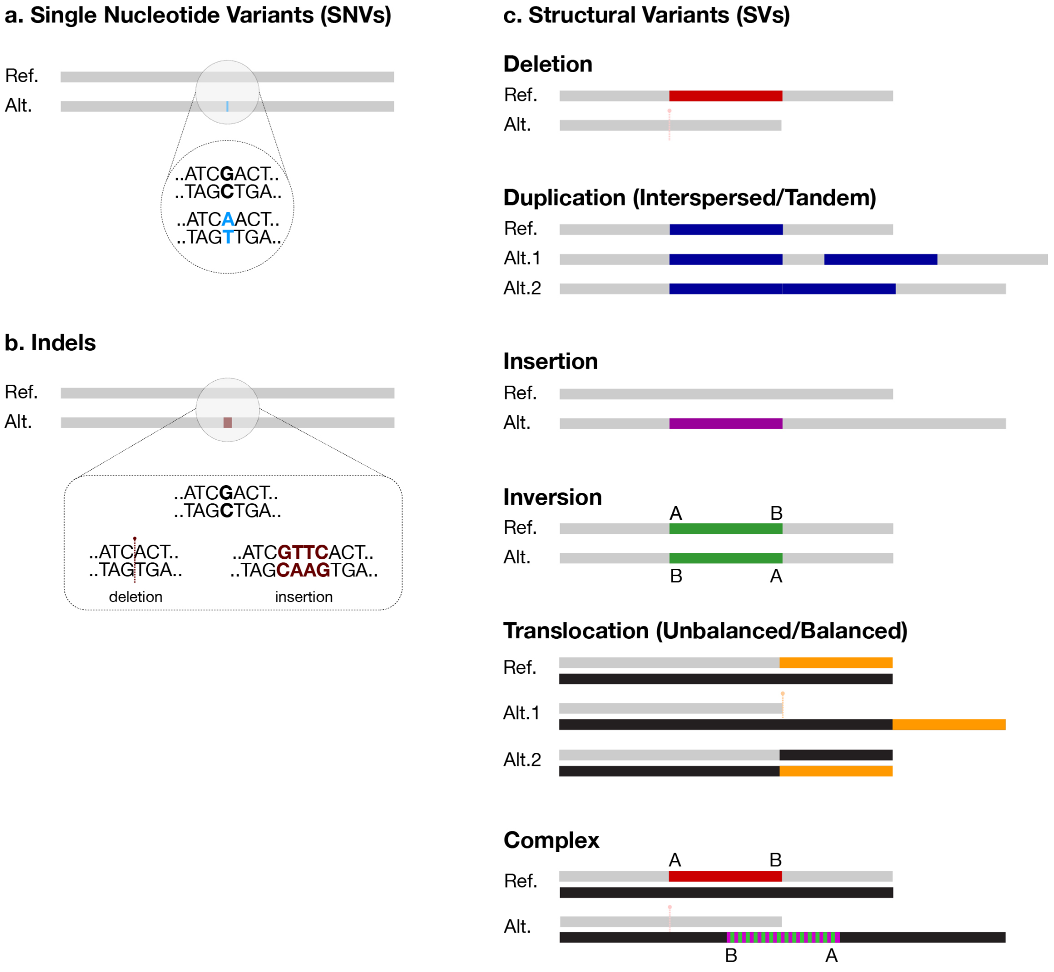


Figure 4. Types of somatic variants.

(**a**), Single Nucleotide Variants (SNVs). (**b**), Indels, corresponding to small deletions or insertions. (**c**), Structural Variants (SVs), corresponding to large genomic rearrangements including deletions, duplications (Alt.1: Interspersed; Alt.2: Tandem), insertions, inversions, translocations (Alt.1: Unbalanced; Alt.2: Balanced), and complex rearrangements.

3.2 Somatic variation in cancer

The emergence of NGS technologies has allowed a better and more comprehensive detection and characterization of the genomic variation occurring in human cells. In cancer genomics, somatic variation is defined as the genomic changes accumulated in the cancer cell genome, which are not present in the normal cell genome of the same individual[44]. The study of somatic variation in cancer has provided the biomedical community with different catalogs, such as COSMIC[45] (Catalogue Of Somatic Mutations In Cancer), collecting and describing the landscape of genomic mutations in cancer. These studies lay the groundwork for the discovery of novel therapeutic targets and the understanding of the different DNA mutational profiles and mechanisms that are active in human cells and are relevant to cancer development.

3.2.1 Types of somatic variants

Based on their size, two major categories in the classification of genomic rearrangements have been defined[46,47].

The first category (Fig. 4a,b) corresponds to small variants that include (a) single-nucleotide variants (SNVs), and (b) indels, which are short insertions and deletions smaller than 50bp-100bp, a variant size that can be detected within a single next-generation sequencing short-read.

The second category (Fig. 4c) corresponds to large rearrangements of the genome, ranging from 50-100bp to megabases in length, known as structural variants (SVs), which include chromosomal rearrangements. Structural variants can be classified[48] as intrachromosomal, involving only one chromosome, or interchromosomal, involving two different chromosomes. Intrachromosomal rearrangements can be unbalanced, associated with a copy number loss or gain, such as (c) deletions —loss of a segment of DNA—, (d) insertions —insertion of a segment of DNA into the genome—, and (e) duplications —where a segment of DNA is inserted in variable number of copies—, or balanced, which are copy number neutral, such as (f) inversions —segments of DNA reversed in orientation—. In the case of interchromosomal rearrangements, which correspond to (g) translocations —where a segment of DNA changes its position in the genome between chromosomes—, they are generally balanced but can also be unbalanced, associated with copy number variation (CNV) and resulting in trisomies or monosomies of chromosome ends.

Besides the well-established types of structural variants, other genomic rearrangements are also classified as SVs[44,49]. It is the case of amplifications —gain of a high number of copies of a segment of DNA (CNV>9[45])—, or the formation of isochromosomes —circular or linear chromosomes including a centromeric structure—, among others. However, SVs do not always occur individually. We can find, for example, translocations associated with copy number variants such as deletions or insertions. The combinations of different rearrangements in the same mutational event having three or more breakpoints, generate what are known as complex structural variants that can also be balanced or unbalanced. Particular cases of complex structural variants have been described[49], being chromothripsis —massive number of SVs clustered in a chromosome—, chromoplexy —several balanced translocations reshuffling multiple chromosomes— and SVs associated with transposition events, the primary examples.

3.2.2 The role of structural variation

As introduced earlier, structural variants are genetic alterations of the genome that, alongside SNVs, contribute to genetic variation among healthy individuals and are known to play a significant role in cancer development and progression[50]. For this reason, it is crucial to characterize the spectrum of SVs in the human cancer genome in order to provide new insights into the understanding of structural variation impact, mechanisms of generation, and patterns that can explain the recurrence of these abnormalities in the chromosome structure.

Structural variants are known to affect coding and non-coding regions of the genome with diverse functional consequences[47,49,51] that affect molecular and cellular processes. It has been described that these types of rearrangements can alter the expression of genes, such as oncogenes, tumor suppressor genes, or others, by truncating or amplifying their loci[51]. According to these findings, amplified regions have been proved to be enriched in oncogenes, while on the other side, deleted regions have been observed to be enriched in tumor suppressor genes. One of the most representative examples of oncogene amplification in cancer is the case of *MYCN* gene in neuroblastoma. Moreover, SVs have shown to have a position effect[48] in genes, altering the expression of the ones that are intact but located close to the variant's breakpoints. Structural variation also has the capacity to generate fusion genes, which have the potential to acquire novel functions differing from the genes in origin and to rearrange regulatory elements of the genome, as in enhancer hijacking events.

Through the emergence of the different national and international collaborative projects to generate catalogs of mutations in cancer, variability in the number and type of SVs across different types of cancer has been described[52]. Still, all tumors accumulate different levels of structural variants and SNVs. For this reason, genome-wide studies looking for recurrent patterns of SVs are performed in order to better characterize these types of rearrangements and their associated mechanisms. Depending on the cancer type, some patterns of SVs are more prevalent than others, indicating signatures of structural variation that are characteristic of different tumors[44], following a similar idea to the one established for SNVs. These findings agree with the reported variation in the mutational spectra across cancer types[51]. Although structural variation has been extensively studied in the past years and the functional and clinical impact of a significant portion of those variants has been well established, there is still work to do, notably in the characterization of non-coding, complex, and recurrent structural variants[51,53].

3.2.3 Mechanisms generating structural variants

As discussed earlier, the appearance of DNA damage such as double-strand breaks (DSBs) in the genome of a normal or tumor dividing cell is highly deleterious and sufficient to trigger cell death. For this reason, the cell present active response machinery to DNA damage, known as DDR, to avoid and solve this emerging

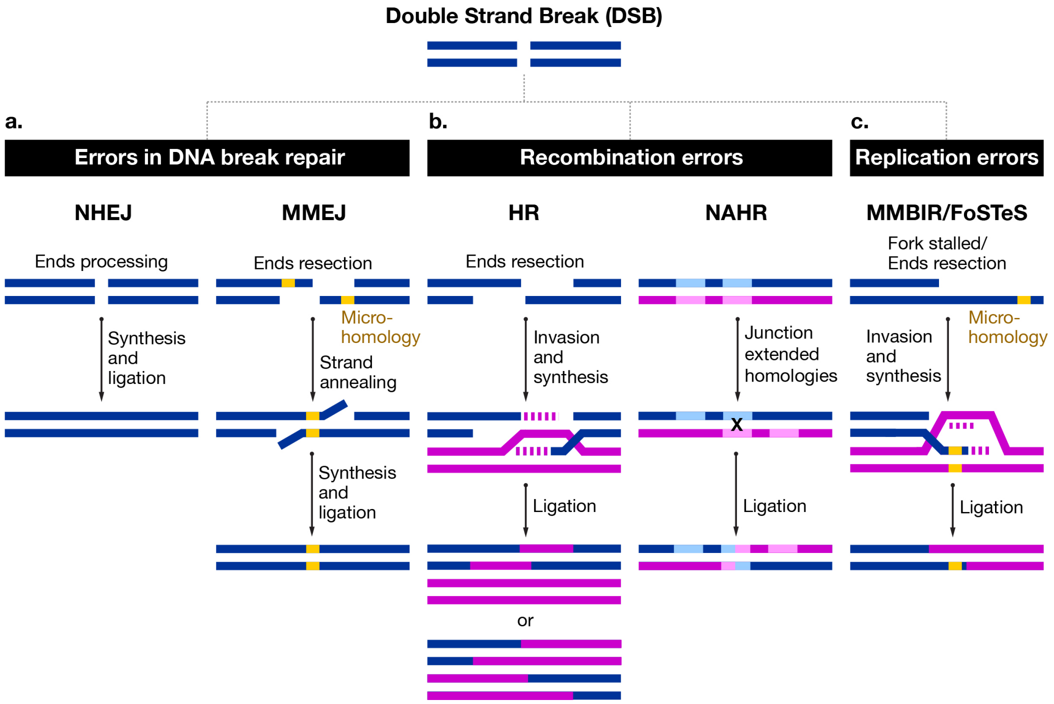


Figure 5. Mechanisms generating structural variants.

Representation of the different mechanisms for the generation of structural variants, classified as (**a**) Errors in DNA break repair, including Non-Homologous End Joining (NHEJ) and Microhomology-Mediated End Joining (MMEJ), (**b**) Recombination Errors, including Homologous Recombination (HR) and Non-Allelic Homologous Recombination (NAHR), and (**c**) Replication errors, including Microhomology-Mediated Break-Induced Replication (MMBIR)/Fork Stalling and Template Switching (FoSTeS). *Adapted from reference*[109].

alterations[54]. Basically, the cell shows checkpoints[55] in its cycle to monitor genome integrity and ensure its viability while avoiding the transmission of the DNA damage to the next generation of daughter cells. Unfortunately, the repair machinery is not entirely error-free, and mutations can end up generated and fixed in the genome due to the action of those mechanisms of DNA break repair.

In the case of structural variants in cancer, there are several mechanisms[48,53-57] that can lead to the generation and shaping of the different types of chromosomal rearrangements.

The first group of mechanisms corresponds to the ones generating recombination errors (Fig. 5b). Non-allelic homologous recombination (NAHR) is one of those mechanisms, which is defined as an alternative form of homologous recombination (HR) but far more error-prone and consequently related to an increase in genome instability in cancer. NAHR is a DNA repair process based on the recombination between regions with high sequence similarity, which are used as the repair template. Unlike HR, the sequence used in NAHR is highly identical (i.e., 95% of homology) but incorrectly homologous since it comes from a different region of the genome that is indeed misaligned. This difference is what makes NAHR a mutational mechanism that can form all kinds of structural variants, such as duplications or deletions, between long homologous segments.

The next group of mechanisms corresponds to the ones generating errors in DNA break repair (Fig. 5a), starting with non-homologous end joining (NHEJ), which is another alternative to HR and the most common method of double-strand break repair in mammals. NHEJ fuses both ends of a double-strand break without requiring any homology, potentially generating small insertions or deletions at the breakpoint junction. Due to the absence of sequence homology acting as a template, this process is also error-prone and has the potential to generate different kinds of structural variants. It is the same case for microhomology-mediated end joining (MMEJ), which follows a similar mechanism as NHEJ, yet involving microhomology sites. MMEJ is highly mutagenic, associated with the generation of deletions containing microhomology at their breakpoints, and often described as a cause for translocations.

The third group of mechanisms corresponds to the ones generating replication errors (Fig. 5c), which mainly are microhomology-mediated break-induced replication (MMBIR) and fork stalling and template switching (FoSTeS). Those two mechanisms have been associated with different kinds of “simple” and complex SVs. Their processes involve the stalling of the replication fork and the invasion of a nearby replication fork by the DNA polymerase via the presence of microhomology.

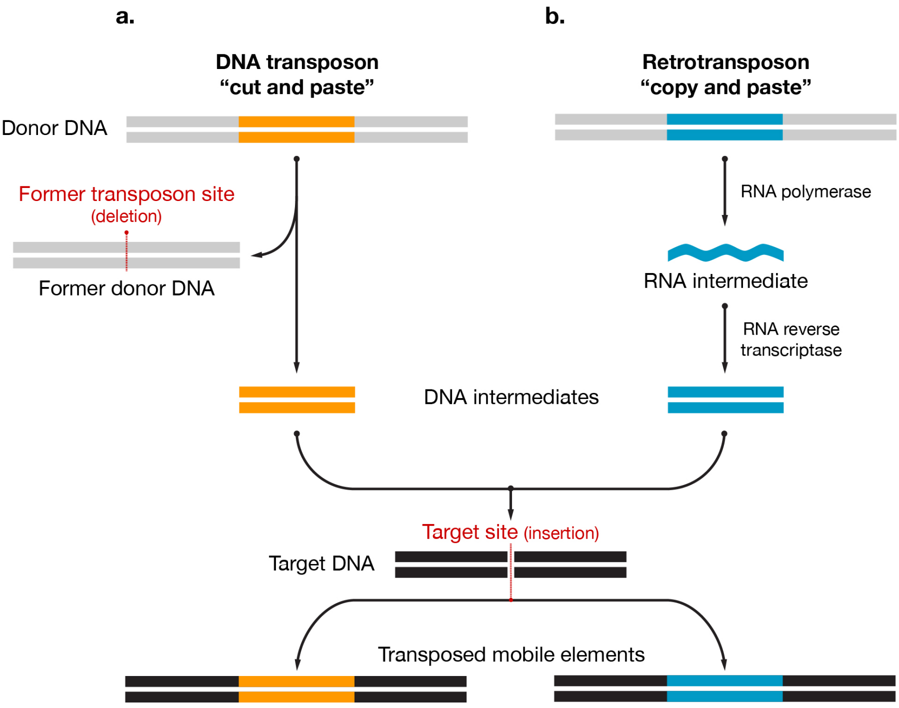


Figure 6. Mechanisms of transposition.

(**a**), DNA transposons (Class II), present a "*cut and paste*" mechanism of transposition. The transposable element is cut from its former site, generating a deletion, and inserted in the target site. (**b**), Retrotransposons (Class I), present a "*copy and paste*" mechanism of transposition. The transposable elements are copied through an RNA intermediate and inserted into the target site.

Interestingly, the generation of genomic amplifications has been associated[49] with these replication-based mechanisms.

3.2.4 Transposable elements: an example of SV generators

Studies suggest[58] that over 60% of the human genome is composed of repetitive sequences derived from transposable elements (TEs). Transposons or mobile elements are DNA sequences that are able to move from one location of the genome to another. It exists two major classes of transposons based on their mobilization mechanism[59], Class I, and Class II.

Class I corresponds to retrotransposons that present a "*copy and paste*" mechanism of transposition (Fig. 6b). This mechanism is based on the mobilization of the transposable element via an intermediate RNA molecule[60], which then is transformed again to DNA using reverse transcriptase to end up inserted in a different genomic location. Retrotransposons are divided into long terminal repeats (LTRs), which include human endogenous retroviruses (HERVs), and non-LTRs. Non-LTR elements include long interspersed nuclear elements (LINEs) —known for retaining the functions needed for retrotransposition through autonomous mechanisms such as L1s—, short interspersed nuclear elements (SINEs) —known to be non-autonomous such as *Alus*— and SVA elements. In the case of non-autonomous elements, they "parasitize" L1s machinery of retrotransposition to actively mobilize themselves in the genome.

Class II mobile elements, on the other hand, correspond to DNA transposons characterized by a "*cut and paste*" mechanism of transposition (Fig. 6a) in which the transposable element is cut from its original region, generating a deletion, and inserted elsewhere in the genome. Although there is evidence of higher activity[61] of retrotransposons than DNA transposons in human cells, a recent study[62] has shown the conservation in humans of the activity of *PGBD5*, a transposase-derived gene from the *PiggyBac* subfamily of transposable elements. This gene is expressed in the early stages of embryo development as well as in some cancers, raising questions about the role of *PGBD5* in the human genome and whether its activity is relevant in oncogenesis.

Interestingly, transposable elements are known to drive evolution but also have the potential to generate double-strand breaks in the DNA, pointing to a direct connection between the activity of these elements and genome instability. It is not for nothing that Irene Scarfo *et al.*, in 2016, described the transposable elements as the "*enemies within*"[59]. As presented earlier in this thesis, genome instability is one of the driver processes in cancer development. Following this definition, differences in the transposition activity of mobile elements between normal and cancer cells

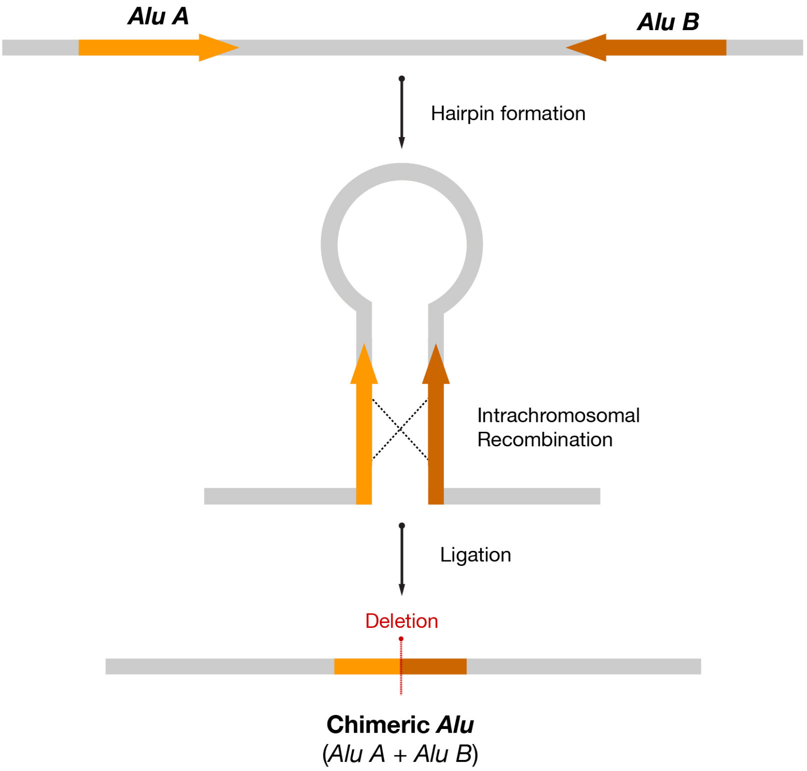


Figure 7. *Alu* recombination-mediated deletion mechanism.

Example of the generation of a deletion resulting in a chimeric *Alu* by the recombination of highly homologous *Alu* elements. Adapted from reference[66].

have been reported. In the case of normal cells, mobile elements present a nearly silent activity, while in cancer cells, TEs show high activity, notably L1 and *Alu* elements. Additionally, many somatic transposable element insertions have been described in different tumors with functional implications such as altering gene function, indicating[61] the involvement of transposon-driven structural variation in promoting human oncogenesis. However, there is still a lack of information on the patterns of rearrangements and mechanisms associated with the activity of transposable elements in cancer.

Withal, the mobilization of transposable elements in the genome and its consequent insertion is not the only process in which mobile elements promote genomic instability and generate structural variation. In addition to TEs mobilization, the high number of these elements in the genome acts as a substrate[63] for different DSBs repair mechanisms resulting in deletions, duplications, and other genomic rearrangements. The most common rearrangements associated with mobile elements besides transposition-related SVs are *Alu*-mediated deletions.

The particular case of Alu elements.

Alu repetitive elements are non-autonomous retrotransposons classified as SINEs that have an average size of 300bp. They correspond to one of the most abundant TEs in humans, with 1.1 million of them distributed across 11% of the genome. Interestingly, numerous studies have described[64,65] the enrichment of genomic rearrangements, notably deletions, in the proximity of *Alu* elements, presenting a specific *Alu-Alu* recombination mechanism (Fig. 7).

Unique features of *Alu* elements conferring the ability to interact and form DNA structures between each other during repair processes, contribute actively to the generation of *Alu* recombination-mediated deletions (ARMDs). One of the essential features[66] in *Alu-Alu* recombination is the presence of homologous sequences between different Alu elements that make them prone to recombine. As we could expect, knowing that the sequence divergence[65] between those elements range from 0% to 30%, *Alus* that have higher homology manifest higher recombination rates between them. Other critical features[61] in the mechanism of recombination between *Alus* are the density[66], proximity[67], and orientation of the two elements involved in the deletion. *Alu* elements normally are 3.000bp away from each other, although we can find regions of the genome presenting higher density of these repetitive elements. In those regions, *Alus* are closer to each other, around 450bp, for example. Those elements in close proximity and inverted orientation are more prone to undergo recombination through the formation of a hairpin structure[66].

ARMDs are usually small to moderate size[60] rearrangements, going from around 300bp to kilobases in length and have the potential to generate chimeric *Alus*. The generation of those genomic rearrangements has been associated with different DNA repair mechanisms[60,61,66], including NAHR, NHEJ, and MMEJ, whereby the two partly homologous *Alu* elements can undergo recombination and produce deletions. *Alu-Alu* recombination has been described as a recurrent mechanism associated with genome instability in different diseases such as colorectal cancer[63]. In the same way as other structural variants, the ones generated by the recombination between these repetitive elements can also affect gene expression and gene functionality. Interestingly the generation of ARMDs has been linked with the absence of functionality of *p53*[61], the most important tumor suppressor gene in humans, reinforcing the idea of transposable elements playing a role in tumorigenesis and suggesting an association between *Alu-Alu* recombination and gene function.

The prevalence of *Alu* elements in the human genome, as well as their ability to undergo recombination, generate structural variants, contribute to recombination-associated mutagenic events, genetic instability, and tumorigenesis, consolidate[61] *Alu* repetitive elements as a mutational hotspot of the genome.

3.2.5 Identification of structural variants

Since the first use of whole-genome sequencing (WGS) data in cancer research, in 2008[6], for the comparison of normal and tumor genomes, next-generation sequencing represents the baseline for cancer studies. The transition from focused to genome-wide approaches and the access to the complete genomic sequence of a vast collection of cancer samples through the use of NGS has benefited the discovery of novel somatic changes related with tumor progression and to the better understanding of the mechanisms of cancer pathogenesis leading to improvement in the diagnosis and treatment of the disease. These advances have been possible due to the potential of NGS data in the discovery of a diversity of DNA rearrangements such as deletions, complex structural variants, translocations, and transposition-related events, among others. Furthermore, the inclusion of non-coding regions, traditionally poorly understood[68], alongside the coding ones in the analysis of the cancer genome, has led to an expansion in the knowledge of the disease and more precisely of the important role of the non-coding genome in cancer processes in particular, and in the human genome in general.

Nonetheless, the sequencing of tumor samples and the comparison between normal and tumor genomes in order to extract the landscape of genetic variation associated with the oncogenic processes is not straight forward. The high heterogeneity present in tumors, which is reflected in the purity and ploidy of the cancer sample and the difficult sampling of tumor cells, make the study of genetic variation more challenging and must be taken into account in the analysis. In addition to this, the size of structural variants typically spanning multiple NGS reads makes its detection more difficult than single-nucleotide variants, requiring computational inference.

Variant calling in cancer.

The purpose of variant calling in cancer is to find the somatic alterations that can be related to the oncogenic processes. It is essential to remove from the equation the germline variants that are present in all cells of the organism and hence are not associated with cancer development. By removing these variants, we will prevent the introduction of noise and false positives in our analysis. For this reason, the identification of somatic mutations in cancer is based on the comparison of matched normal and tumor DNA samples from the same individual.

Once the normal and tumor samples have been extracted from the patient —ideally from the same tissue to reduce heterogeneity, although the use of peripheral blood as the normal sample is highly extended—, both samples are processed in parallel through the different NGS steps explained earlier, obtaining whole-genome sequencing data, for instance. As previously established in this thesis, NGS is the most prevalent method of sequencing data for cancer analysis, and it has proved to allow the detection of known and novel structural variants, balanced and unbalanced at base-pair resolution[29].

In order to process the sequencing data, a good number of variant calling programs have been developed to compare simultaneously normal and tumor samples and identify somatic mutations present only in the tumor. With the purpose of a better and more comprehensive identification of the whole range of somatic rearrangements, variant callers design adopt diverse strategies and computational approaches with notable differences in sensitivity and specificity.

These approaches[50,69] can be reference-free —reads not mapped to the reference genome (i.e., SMuFin[70])—, or mapping-based, being the last the more popular ones. Mapping based algorithms detect variants in the genome using read pairs —looking at mapping discordance between sample read and reference—, read depth —counting reads across the genome, useful for CNV (i.e., ASCAT[71])—, and split reads information —looking at reads not mapped in its entirety, soft-clipped—. They can also incorporate strategies of *de novo* assembly of rearranged regions (i.e., SvABA[72], BRASS[73]), or consist of combined methods such as Delly[74] —based on read pairs and split reads—. Curiously it has been proved that different algorithms used on the same dataset generate different results. For this reason, we not only have to use the appropriate caller depending on the variants we are looking for, but it is also recommended to combine different algorithms to improving detection[69].

The combination of the results from different variant calling algorithms is typically done in two different ways[75] depending on the balance of sensitivity and specificity we want. If it is specificity what we are looking for, we should go for the intersection of results, keeping all the variants that are common in at least *n* callers, being n>1. On the other hand, if we are looking for sensitivity, we should go for the union of results from the different callers. In this case, to avoid a drop of specificity, results of this union might be filtered, taking into account the number of supporting reads in each of the variants, for example. Overall, the selection and combination of variant calling methods correspond to critical stages in cancer analysis and have to be carried out accordingly.

In general, the improvement in variant calling algorithms, notably SV detection, has allowed the comprehensive characterization of the prevalence, complexity, and importance of structural variants in human genomes. Current analyses link more than 20,000 SVs per human genome[75]. However, there are still numerous SVs that are poorly detected and described, particularly the small variants, bigger than the short-read size and smaller than the insert sizes, and the variants associated with repetitive elements of the genome. For this reason, new sequencing methodologies such as long reads are becoming more and more frequent in structural variation analysis. Nevertheless, the detection of SVs is a crucial first step to assign functional impact to these variants finally. The integration of structural variation results with gene expression, epigenetic, or three-dimensional structure data is what allows us to interpret the functional consequences and reveal new mechanisms related to this significant fraction of the human genetic variation.

3.3 Tumor evolution and drug resistance

After accumulating mutations, the cell with an aberrant genotype continues its clonal expansion that may lead to the development of a tumor. As an essential part of the oncogenic process, natural selection acts in the clonal division of this cell. The same way ecosystems affect the survival and reproduction of species, cell and tumor environments play a major role in the survival and proliferation of the abnormal cell[32,33].

Oncogenic cells exploit normal cell properties such as telomeric stabilization, cell proliferation, migration, and invasion for its own evolutionary benefit. It is known that the majority of cancer cells get their mitotic machinery stalled or suffer from apoptosis before they can start dividing[33]. However, depending on the advantages conferred by the genomic instability, and the selective environment where they grow, cancer cells can be positively selected and undergo high proliferation rates forming a mass of cells known as tumor or neoplasm. One of the general features for the selection of cancer cells is its capacity for self-renewal[76,77]. It is also interesting to notice that cancer cells have higher proliferation rates than normal cells[78], acquiring more mutations in a shorter period of time[41], overcoming normal tissue development, generating genetic diversity between clones, and essentially, providing heterogeneity and growth advantages to tumors. In consequence, cancer is defined as a dynamic disease.

Genetic diversity within tumors is essential in the understanding of neoplastic processes but even more important in its treatment. It is accepted that intratumor heterogeneity not only drives the evolution of cancer but also favors drug resistance. Conventional cancer treatments such as chemotherapy or radiotherapy, known for its unspecific and broad profile, modify the tumor environment of cancer clones, and in some cases facilitate the selection of the clones that are more resistant to treatment, increase the speed of clonal evolution, and therefore make the remission of the disease more difficult[79] (Fig. 3). For this reason, cancer therapies go towards the development of specific target treatments specifically focused on the tumor cells using proteomic and genomic information. However, selective pressure to resistant clones can also arise from these newer therapies[41]. Thus, a good characterization of the genetic diversity of tumors, including their structural variation profile, and the use of combination therapies are necessary.

Unfortunately, the generation of the neoplasm may not be the last step of oncogenesis. As stated before, by taking advantage of the mechanism of the normal cell, a fraction of the tumor cells acquires the capacity to migrate to other tissues and organs of the body. Migrating cells invade new habitats, in what is known as a metastatic process, generating the seed for the formation of new tumors and the acquisition of more diversity among cancer cells[33,77,79]. Metastatic events make even more challenging the diagnosis and treatment of cancer and, in consequence, decrease the survival of the patient.

3.4 Cancer projects

In the early 2000s, encouraged by the release of the assembly of the human genome by the HGP and the appearance of NGS game-changing techniques, different international projects emerged to provide new insights into the biology of cancer.

In the case of the United Kingdom, we find the Cancer Genome Project[80], which started in the early 2000s managed by the Wellcome Trust Sanger Institute. This project aims to improve cancer diagnosis, treatment, and prevention through a better understanding of the disease. To do so, they started to build a publicly available database of genomic changes present in different cancer types. All the somatic mutations discovered through their analysis and other related projects are currently included in COSMIC[45], one of the most extensive resources available for exploring the impact of somatic mutations in human cancer.

Going to the other side of the Atlantic, in the United States, we find The Cancer Genome Atlas[81], which started in 2006 and is managed by sections of the National Institutes of Health (NIH). The program of TCGA is centered in the study of cancer genomics and the generation of comprehensive genomic datasets. Within this study, they analyzed more than 20,000 matched primary cancer and normal samples from 33 different cancer types using WGS generating over 2.5 petabytes of genomic, epigenomic, transcriptomic, and proteomic data. Through the integrative analysis of all this data, researchers have gained insight into the biology of somatic mutations, discovering new biomarkers for prognosis and therapies, and germline mutations, identifying genes related to hereditary malignancies. Importantly, the translation of the molecular results to clinics is the key point of this initiative.

Both TCGA and CGP operate within the scope of the International Cancer Genome Consortium[82], launched in 2007, to coordinate large-scale cancer genome studies across the globe, including the ones presented here. ICGC's main goal is common to TCGA or CGP, to generate and provide comprehensive catalogs of the full range

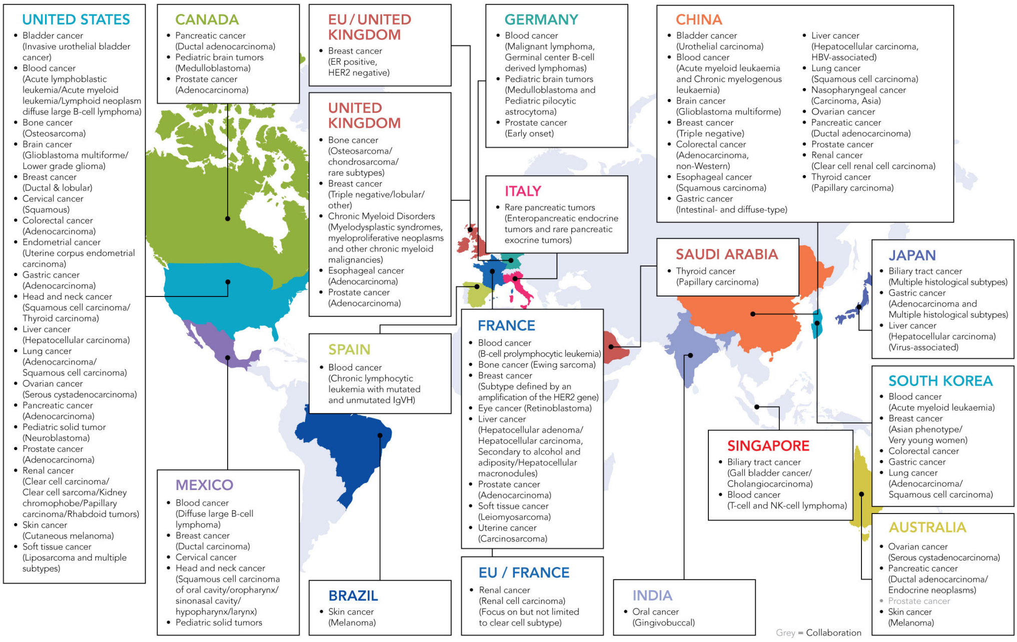

UNITED STATES
• Bladder cancer (Invasive urothelial bladder cancer)
• Blood cancer (Acute lymphoblastic leukemia/Acute myeloid leukemia/Lymphoid neoplasm diffuse large B-cell lymphoma)
• Bone cancer (Osteosarcoma)
• Brain cancer (Glioblastoma multiforme/ Lower grade glioma)
• Breast cancer (Ductal & lobular)
• Cervical cancer (Squamous)
• Colorectal cancer (Adenocarcinoma)
• Endometrial cancer (Uterine corpus endometrial carcinoma)
• Gastric cancer (Adenocarcinoma)
• Head and neck cancer (Squamous cell carcinoma/ Thyroid carcinoma)
• Liver cancer (Hepatocellular carcinoma)
• Lung cancer (Adenocarcinoma/ Squamous cell carcinoma)
• Ovarian cancer (Serous cystadenocarcinoma)
• Pancreatic cancer (Adenocarcinoma)
• Pediatric solid tumor (Neuroblastoma)
• Prostate cancer (Adenocarcinoma)
• Renal cancer (Clear cell carcinoma/ Clear cell sarcoma/Kidney chromophobe/Papillary carcinoma/Rhabdoid tumors)
• Skin cancer (Cutaneous melanoma)
• Soft tissue cancer (Liposarcoma and multiple subtypes)
CANADA
• Pancreatic cancer (Ductal adenocarcinoma)
• Pediatric brain tumors (Medulloblastoma)
• Prostate cancer (Adenocarcinoma)
EU / UNITED KINGDOM
• Breast cancer (ER positive, HER2 negative)
UNITED KINGDOM
• Bone cancer (Osteosarcoma/ chondrosarcoma/ rare subtypes)
• Breast cancer (Triple negative/lobular/ other)
• Chronic Myeloid Disorders (Myelodysplastic syndromes, myeloproliferative neoplasms and other chronic myeloid malignancies)
• Esophageal cancer (Adenocarcinoma)
• Prostate cancer (Adenocarcinoma)
GERMANY
• Blood cancer (Malignant lymphoma, Germinal center B-cell derived lymphomas)
• Pediatric brain tumors (Medulloblastoma and Pediatric pilocytic astrocytoma)
• Prostate cancer (Early onset)
ITALY
• Rare pancreatic tumors (Enteropancreatic endocrine tumors and rare pancreatic exocrine tumors)
CHINA
• Bladder cancer (Urothelial carcinoma)
• Blood cancer (Acute myeloid leukaemia and Chronic myelogenous leukaemia)
• Brain cancer (Glioblastoma multiforme)
• Breast cancer (Triple negative)
• Colorectal cancer (Adenocarcinoma, non-Western)
• Esophageal cancer (Squamous carcinoma)
• Gastric cancer (Intestinal- and diffuse-type)
• Liver cancer (Hepatocellular carcinoma, HBV-associated)
• Lung cancer (Squamous cell carcinoma)
• Nasopharyngeal cancer (Carcinoma, Asia)
• Ovarian cancer
• Pancreatic cancer (Ductal adenocarcinoma)
• Prostate cancer
• Renal cancer (Clear cell renal cell carcinoma)
• Thyroid cancer (Papillary carcinoma)
SAUDI ARABIA
• Thyroid cancer (Papillary carcinoma)
JAPAN
• Biliary tract cancer (Multiple histological subtypes)
• Gastric cancer (Adenocarcinoma and Multiple histological subtypes)
• Liver cancer (Hepatocellular carcinoma) (Virus-associated)
SPAIN
• Blood cancer (Chronic lymphocytic leukemia with mutated and unmutated IgVH)
FRANCE
• Blood cancer (B-cell prolymphocytic leukemia)
• Bone cancer (Ewing sarcoma)
• Breast cancer (Subtype defined by an amplification of the HER2 gene)
• Eye cancer (Retinoblastoma)
• Liver cancer (Hepatocellular adenoma/ Hepatocellular carcinoma, Secondary to alcohol and adiposity/Hepatocellular macronodules)
• Prostate cancer (Adenocarcinoma)
• Soft tissue cancer (Leiomyosarcoma)
• Uterine cancer (Carcinosarcoma)
SOUTH KOREA
• Blood cancer (Acute myeloid leukaemia)
• Breast cancer (Asian phenotype/ Very young women)
• Colorectal cancer
• Gastric cancer
• Lung cancer (Adenocarcinoma/ Squamous cell carcinoma)
SINGAPORE
• Biliary tract cancer (Gall bladder cancer/ Cholangiocarcinoma)
• Blood cancer (T-cell and NK-cell lymphoma)
MEXICO
• Blood cancer (Diffuse large B-cell lymphoma)
• Breast cancer (Ductal carcinoma)
• Cervical cancer
• Head and neck cancer (Squamous cell carcinoma of oral cavity/oropharynx/ sinonasal cavity/ hypopharynx/larynx)
• Pediatric solid tumors
BRAZIL
• Skin cancer (Melanoma)
EU / FRANCE
• Renal cancer (Renal cell carcinoma) (Focus on but not limited to clear cell subtype)
INDIA
• Oral cancer (Gingivobuccal)
AUSTRALIA
• Ovarian cancer (Serous cystadenocarcinoma)
• Pancreatic cancer (Ductal adenocarcinoma/ Endocrine neoplasms)
• Prostate cancer
• Skin cancer (Melanoma)
Grey = Collaboration

Figure 8. International Cancer Genome Consortium map of projects.

Map depicting the coordination of international cancer projects from 16 different countries and two European consortia carried by the ICGC. *Image source: https://icgc.org.*

of genomic variation —SNVs and SVs—, mainly somatic mutations but also germline, across different cancer types. ICGC is coordinating 90 projects from 16 different countries and two European consortia (Fig. 8) in which they have analyzed more than 25,000 matched normal and tumor genomes from around the world from 50 different cancer types at a genomic, epigenomic, and transcriptomic level.

Definitely, the formation of an international force that coordinates the most important cancer projects in the world is not trivial. The main reasons and goals for the creation of ICGC are: the will of avoiding duplicated efforts in different independent and uncoordinated studies to prevent redundancy in the analysis; the necessity of standardization between studies to facilitate the merging and comparison of resulting datasets from different projects; and the requirement of a faster and better dissemination of the datasets and methodologies in the global scientific community to accelerate cancer research around the globe.

Following these goals, in the last decade, a novel TCGA-ICGC initiative described as the Pan-Cancer Analysis of Whole Genomes (PCAWG)[28], was launched. This project analyzed 2,658 matched normal and cancer whole-genomes from 38 distinct tumor types. The interesting thing about it is the implementation of standardized alignment and variant calling pipelines for all the different research groups that were participating in order to facilitate the comparison and sharing of the results, according to the ICGC philosophy. Pursuing this idea, a catalog of coding and non-coding somatic and germline variation is publicly available in their data portal[83]. This project has provided the molecular profiling of tumors at a DNA, RNA, epigenetic, and protein levels with an emphasis on non-coding driver events, patterns of structural variation, and other cancer-specific molecular alterations.

In summary, all the big international cancer projects, such as the ones presented above, aim to provide new and better insights into the molecular biology of the oncogenic processes by analyzing and providing new data highly valuable for the field. However, they centered their analysis in the traditionally-known part of the genome —the 23 pairs of linear chromosomes— and the mitochondrial chromosome, omitting the analysis of other extrachromosomal genomic structures associated with cancer, such as the circular DNA elements.

4 Circular DNA elements in cancer

In the early 1960s, Alix Bassel and Yasuo Hotta[84] found evidence, for the first time, of the presence of circular DNA mixed with linear chromosomes in the nucleus of mammalian cells. The following findings of circular DNA with variable sizes in humans and other organisms' healthy tissues, together with the discovery of double-minute chromosomes in cancer, started changing the perception of the cellular genome, which was historically restricted to the 23 pairs of linear chromosomes. The presence and the ability to detect circular elements in human cells opened the path to the study of the role of this long-forgotten portion of the human genome existing as extrachromosomal circles of DNA, which we now refer to as the circulome[85]. Curiously, following the gain in the importance of circular DNA in the biology of the cell due to the discovery of its functional impact, notably in tumors, researchers have coined a new term to designate this novel area of study in genomics, which is termed circulomics[86].

4.1 Types of extrachromosomal circular DNA

The nomenclature to design extrachromosomal circular DNA elements in the genome has changed over the years with the evolution of the field and is still in the process of being fully established. Actually, all types of extrachromosomal circular DNA are commonly defined as circular DNA elements that lack a centromere and are found outside the chromosomes. Following the last reviews[87], which classify circular DNA based on its copy number, which seems to be related with size, we can define at least two major classes of extrachromosomal circular DNA with unique features, in human cells.

The first class, designated as eccDNA, correspond to copy number neutral, small, extrachromosomal circular DNA, usually with sizes up to 1kb, although they can reach 50kb or more. eccDNA does not contain entire genes, even though it can contain parts of them, and is common in normal and tumor human cells.

microDNAs, defined as circular elements formed by non-repetitive sequences up to 400bp long, and telomeric circles are included in this class.

On the other hand, ecDNA, the second class of circular DNA considered in this thesis, corresponds to high copy number large extrachromosomal circles of DNA, usually with sizes around 1 to 3Mb, and containing one or multiple entire genes and/or regulatory regions. ecDNA, as opposed to eccDNA, is rarely found in healthy cells but is highly prevalent in cancer. Double-minute chromosomes, which were the first extrachromosomal circular DNA elements identified in tumor cells, are included in this class.

4.2 Generation of circular DNA elements

The sequences forming the extrachromosomal circular DNA elements found on human somatic tissues are generally homologous to the linear genome, accounting for 12.6% of it. The evidence that the circulome derives[36] from all types of genomic structures from any part of the linear human genome, including coding —genes—, non-coding —intergenic regions, regulatory elements—, and repetitive regions —SINEs, LINEs, and others—, explains this homology. In essence, circular DNA can include genes/regulatory elements or not and can be composed of repetitive and/or unique genomic sequences.

The exact mechanisms leading to the formation of circular DNA have yet to be determined. However, different characteristics[88] have been described to be significantly associated with the circularization of DNA, suggesting the implication of distinct processes in the origin of the circular elements. Those characteristics include the complexity of the DNA sequence within the circular structure, notably the one flanking the circle junction, the three-dimensional structure of DNA, and the damage repair mechanisms involved. Pursuing this idea, analyses have shown a higher tendency to circularization of coding and repetitive regions, pointing to the presence of genomic loci that can be considered hotspots[36] for the origin of the circular structures. The generation of circular DNA containing repetitive elements, such as inverted repeats flanking the circle junction, is generally attributed[89] to the activity of HR or MMEJ DNA repair mechanisms. However, a significant fraction of the circles of DNA does not contain repeats, suggesting the action of other mechanisms as parts of random mutational processes.

Overall, two differentiated mechanisms[88] are proposed to produce circular DNA elements. The most popular mechanism in the literature corresponds to the excision of chromosomal DNA, which is then circularized and is clearly associated with a deletion in the chromosomal region of origin. Still, it has been described more than 100,000 sites of circularization with different sizes in the human genome, suggesting the improbable number of 100,000 genomic deletions. Moreover, no deletions were found in many chromosomal positions originating DNA circles. Following these ideas, a replication-dependent process has also been proposed[86]. In that case, the copying mechanism would generate a copy of the region that will then be circularized without any loss of chromosomal DNA.

In summary, circularization from repetitive, non-repetitive, or mixed sequences —unique sequence with repeats flanking the circle junction— can be carried out by replication-dependent —copy— or -independent —deletion— mechanisms (Fig. 9).

4.3 Function and impact of circular DNA in cancer

While little is known about the impact and function of eccDNA in cancer, ecDNA, in contrast, has been the center subject of numerous studies on the role of extrachromosomal circular DNA in tumor cells. This is due to the fact that ecDNA has been detected[90] in around 40% of tumor cell lines, and although the levels of this type of circular DNA vary across cancer types, it has been found to be a common phenomenon in this disease. ecDNA represents the principal vehicle for oncogene amplification in tumor cells with important examples in neuroblastoma and glioblastoma, among others.

4.3.1 The role of circular DNA in oncogene amplification and tumor heterogeneity

The increase of the copy number of oncogenes is one of the most frequent alterations in cancer. It has largely been associated[87] with overexpression of these genes, and consequently, with tumor heterogeneity and evolution through conferring growth advantages to the cell. The rise in oncogene copy number is attributed to two distinct DNA structures: ecDNA and homogeneous staining regions (HSR) (Fig. 9a,b). HSR correspond to large regions of the chromosomes formed by several copies of smaller genomic regions containing cancer-related genes. Interestingly, HSR can be formed by the reintegration of various copies of ecDNAs in the same region. However, the amplification of genes via homogeneous staining regions is significantly less frequent than the one via extrachromosomal circular DNA, which has been described to be widespread in cancer. In this line, numerous studies reveal ecDNA as a prevalent mechanism by which oncogenes are extensively amplified in tumor cells achieving copy numbers significantly higher than HSRs[88,90].

The propagation and accumulation of ecDNA elements containing oncogene copies in the dividing cells are not only responsible for the overexpression of these genes but are also responsible for the increase of tumor heterogeneity. Due to the fact that circular DNA employs the same replication mechanisms as the linear genome, and principally, due to the missegregation of circular elements during cell division, ecDNA containing oncogenes may be accumulated asymmetrically in the daughter cells (Fig. 9c). This process is believed to start with the replication of the circulome, which takes place during cell division the same way as the linear genome. However, the lack of centromeres characteristic of extrachromosomal circular DNA promotes the random unequal segregation[88,91] of these elements during cell division into the daughter cells. As a consequence, a large number of oncogene copies may end up accumulating in one cell, conferring proliferative advantages associated with an increase of its growth and development within the tumor cell population.

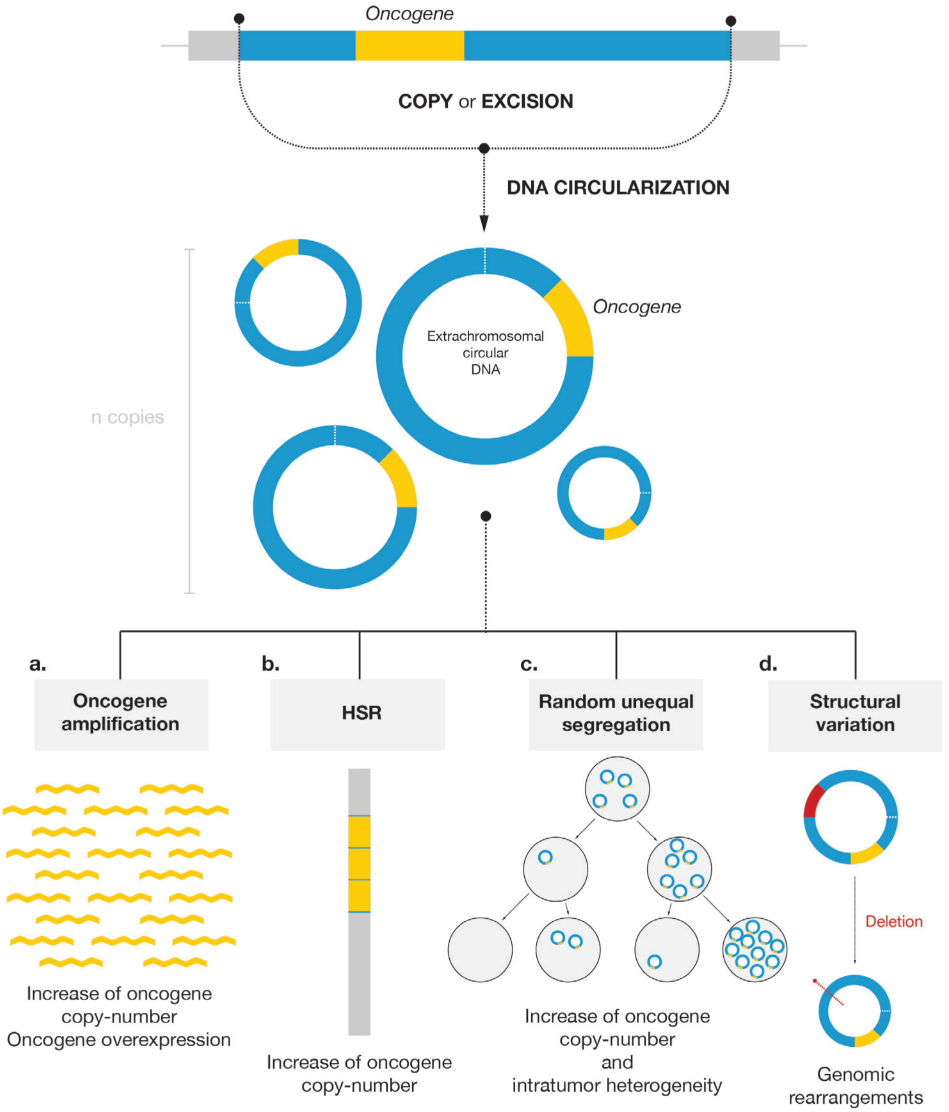

Oncogene
COPY or EXCISION
DNA CIRCULARIZATION
Oncogene
Extrachromosomal circular DNA
n copies
a.
Oncogene amplification
b.
HSR
c.
Random unequal segregation
d.
Structural variation
Increase of oncogene copy-number
Oncogene overexpression
Increase of oncogene copy-number
Deletion
Increase of oncogene copy-number and intratumor heterogeneity
Genomic rearrangements

Figure 9. Functions of ecDNA in cancer.

Representation of the major functions associated with ecDNA in cancer. It starts with the generation of the circular DNA elements containing an oncogene, by copy or excision. (**a**), These elements can be amplified several times, increasing the number of copies of the oncogene promoting its overexpression. (**b**), Circular DNA can also be re-integrated in the genome as Homogeneous Staining Regions (HSRs), stabilizing the copies of the oncogene and increasing its expression. (**c**), Circular DNA suffers from random unequal segregation during cell division, increasing the intratumor heterogeneity and the potential accumulation of oncogene copies in daughter cells. (**d**), Circular DNA elements can also undergo structural variation pre- or post-circularization, modifying the genomic elements inside them.

The enhancement of the fitness of cells with high levels of ecDNA containing oncogenes, while expanding tumor heterogeneity[87], would result in a rise of oncogene copy numbers within the tumor. As presented earlier in this thesis, intratumor heterogeneity drives the evolution of cancer and favors response to changes in the cell environment, making heterogeneous tumors more difficult to treat. Following this idea, tumors with presence of extrachromosomal oncogene amplification associated with higher tumor heterogeneity may adapt more efficiently to changes in its environment such as therapies and, in consequence, may become more aggressive and hard to treat. On the contrary, intratumor heterogeneity promoted by oncogene amplification as HSR would stabilize faster due to the no-unequal segregation of linear chromosomes. For these reasons, ecDNA has been defined as a driving force in tumor evolution, rapidly promoting and maintaining intratumor heterogeneity in cancer.

4.3.2 Circular DNA as a vehicle for oncogenic overexpression

In correlation with the role of ecDNA as a vehicle for oncogene amplification, it has been found that cancer-related genes encoded on those high copy number circular elements correspond to the top highly expressed genes in tumors. This is something we could expect since it is not strange that the number of copies correlates with the transcription levels of genes. But, curiously, the newest studies have confirmed[91] that the amount of DNA template available, or in other words, the amount of ecDNA containing oncogenes, is not the only factor determining the levels of transcription of these genes.

Contrarily to what happens in the linear genome where the DNA is packed around histones, making it inaccessible to the transcription machinery, ecDNA present prevalent highly accessible chromatin. This accessibility facilitates the increased levels of transcription of cancer genes related to oncogenesis and again, with tumor heterogeneity and potential therapeutic resistance. Moreover, the finding of transcription factors and other regulatory elements captured[88] inside circular DNA supports the role of extrachromosomal circles of DNA in modulating gene expression.

4.3.3 Other implications of circular DNA in genomic instability

Although the extrachromosomal amplification of oncogenes represents a major event driving tumorigenesis, it is not the only type of mutational process associated with circular DNA. There is evidence of the presence of DNA rearrangements inside extrachromosomal circular DNA elements (Fig. 9d), which can happen during or post- circularization. The structural variants occurring inside the circles, such as deletions, act as circulome remodelers that can be driver mutations or can even operate as a substrate for the assembly[36] of larger circles from a combination of smaller ones, generating what is known as chimeric circles.

Extrachromosomal circular DNA has been proved[87] to be a common mutational element in human cancer cells. It has been associated with genetic variation — SVs, deletions, amplifications—, tumor development and evolution —increasing the copies and expression of oncogenes—, and tumor heterogeneity —unequal segregation of circular DNA in cell division—. In addition to this, the hypothesis about a possible role of cell-free circular DNA elements as vehicles for transferring driver genes in the formation of metastatic tumors has been proposed[86].

In conclusion, extrachromosomal circular DNA has been proved to have an essential role in the increase of genomic instability and tumor progression, emphasizing the need for including the circulome in genomic cancer studies. Through the analysis of the human circular genome in cancer, researchers[88] have also shown the relevance of circular elements in the identification and prognosis of the disease, being neuroblastoma one of the more evident examples. These findings point to the potential use of extrachromosomal circular DNA as a cancer biomarker[86], opening a new way of addressing genomic studies. However, there are still many unresolved questions[87] concerning circular elements, such as their mechanisms of formation, maintenance, and segregation, the existence of different functional impacts, notably as potential genome remodelers by interacting with linear chromosomes, gene expression modulators, and the clinical impact and utility of these elements.

4.4 Identification of circular DNA

As mentioned above, extrachromosomal circular DNA is commonly forgotten from human genome analysis, although it has been proven to be present in healthy cells and to have tumorigenic implications in cancer. This is due not only to the need of a change of paradigm in genomic studies, starting to think about the human genome as a more dynamic entity than the classic 23 linear chromosomes configuration, but is also due to the fact that standard pipelines for whole-genome analysis have to be improved[87] to address the circulome.

Standard WGS in cancer studies, for example, permit the analysis of the whole genome of the individual, as a mix of DNA from the linear genome and the circulome. Even if this technique captures the whole genomic fraction of the cell, it cannot distinguish between circular and linear DNA elements, with its consequent negative effect on the results. For this reason, new bioinformatic algorithms[90,92] have been created to infer circularity from whole-genome sequencing data based on the orientation of the paired-end reads once mapped to the reference genome. Currently, as a complement of NGS short reads, long-read sequencing is also included[87] in circle DNA analysis, facilitating a better resolution of ecDNA and eccDNA structures, being especially useful for assembling chimeric circular elements.

However, these algorithms are still used in mixed linear and circular data. With the aim of purifying the samples for extrachromosomal circular DNA studies, new wet-lab protocols have been incorporated, such as circle-seq[36]. This protocol has been proved to be useful in the detection and analysis of circular DNA elements at a genomic scale. It is mainly based on the digestion of linear DNA molecules to isolate and purify circular DNA from the sample. This way, we obtain the sequencing of the circulome only.

As the circular DNA field grows, more and more techniques and algorithms arise allowing us to detect and analyze the portion of our genome that is circularized in order to gain more insights in the role of those elements in different diseases such as cancer but also in the normal behavior of our cells.

5 Neuroblastoma: a cancer example

As established earlier in this thesis, cancer is a collection of different diseases that can affect adults and children. One of those diseases corresponds to neuroblastoma, a pediatric malignancy of the developing sympathetic nervous system. This type of cancer spreads into lymph nodes, bone, and bone marrow forming tumors from non-differentiated cells, precursors of neurons, called neuroblasts (Fig. 10a). Neuroblastoma is the most common diagnosed cancer during infancy, accounting for 7-10% of detected childhood cancers and 15% of childhood cancer deaths[93]. Only 1% of neuroblastoma cases have a hereditary origin, while the majority arise sporadically accumulating somatic genomic rearrangements. It has not been reported that environmental factors involving the patient act in the development of the disease[94]. Therefore, neuroblastoma represents an excellent example of a malignancy fundamentally driven by somatic rearrangements in the cell.

One of its principal traits is its heterogeneity, which translates into a strong therapeutic stratification. Depending on the complexity of the genetic and genomic events existing in the tumor, neuroblastoma presents two completely different development and progression profiles that might even seem two entirely different diseases[93-95].

The first form of neuroblastoma, which corresponds to the lower stage disease and low-risk, displays whole chromosome gains without chromosomal rearrangements. Patients in this form are usually hyperploid or near-triploid without incorporating structural variation. The singularity of this stage is the excellent prognosis of the patients, presenting high cure rates and even spontaneous regression of the disease. On the other hand, the second form of neuroblastoma defined as high-risk, is more aggressive, correlating with poor prognosis, and a cure rate of less than 30% of cases. The aggressive form of neuroblastoma is characterized by a diploid karyotype with the presence of somatic structural rearrangements, notably with the amplification of *MYCN* oncogene. This second form is also known to present metastatic events in the liver and lung[96].

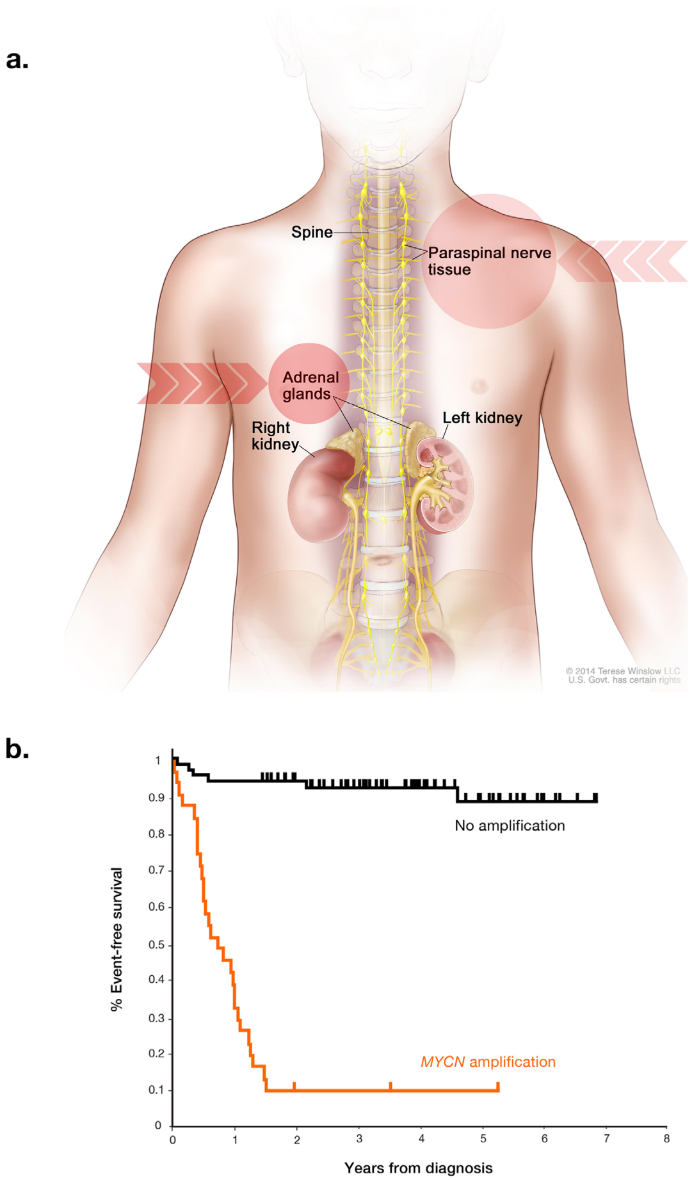


Figure 10. Neuroblastoma primary development and *MYCN*-amplified survival.

(**a**), Neuroblastoma begins in the neuroblasts —immature nerve cells— outside the brain of infants and young children. Normal neuroblasts mature into nerve cells or cells from the adrenal gland. When neuroblasts do not mature, they can continue to grow, forming a tumor. This tumor formation can start in the spinal nerve tissue, chest, abdomen, or pelvis, but most commonly begins in the adrenal glands. *Image source: For the National Cancer Institute. Copyright 2014. Terese Winslow LLC, U.S Govt. has certain rights.* (**b**), Survival of infants with metastatic neuroblastoma based on *MYCN* status. The 3-year event-free survival (EFS) of infants whose tumors lacked *MYCN* amplification was 93%, whereas those with tumors that had *MYCN* amplification had only a 10% EFS. *Image source from reference*[94.]

5.1 Proto-oncogene amplification in high-risk neuroblastoma

The genetic aberration most commonly associated with the aggressive form of neuroblastoma is *MYCN* gene amplification, described for the first time in 1983 by Schwab *et al.* using in situ hybridization[97]. *MYCN* gene is a proto-oncogene from the *Myc* family of transcription factors located in the short arm of chromosome 2. It is primarily expressed in healthy developing embryos and is thought to be critical in the brain and other neural development. Oncogenes, as opposed to tumor suppressor genes, which code for proteins that operate to restrict the cell cycle or even promote programmed cell death or apoptosis, code for proteins involved in cell regulation that operate by stimulating cellular growth and division[98]. In the case of *MYCN*, the amplification resulting in an average of 50 to 400 copies of the gene per cell corresponds to the genetic change that transforms *MYCN* from a proto-oncogene to an actively expressed oncogene[93]. This high copy number gain is present in 25% of primary tumors and correlates with advanced stage and aggressive form of the disease, and treatment failure (Fig. 10b). *MYCN* amplification represents the only example of high-frequency oncogene activation in this disease[94]. For this reason, it is still used as the principal poor prognosis predictor for neuroblastoma.

As introduced earlier in this thesis, it is currently established that the mechanism behind *MYCN* amplification significantly involves large extrachromosomal circular DNA elements with high copy number containing the gene, defined as ecDNA. For this reason, neuroblastoma represents one of the best examples to illustrate the role and importance of extrachromosomal circular DNA elements in cancer development and prognosis.

5.2 Other important genomic alterations in high-risk neuroblastoma

Despite the relevance of *MYCN* amplification in aggressive neuroblastoma, there are other recurrent genomic changes associated with this form of the disease that have not been directly associated with ecDNA. In the case of structural variation, and more specifically on loss or gain of genetic material, we find recurrent chromosomal rearrangements in 1p, 17q, and 11q among others[93-95,99]. The deletion of the short arm of chromosome 1 is present in 25-35% of neuroblastoma cases and correlates with *MYCN* amplification and poor patient outcome. Moreover, this loss of material in chromosome 1 is commonly associated with a gain in the long arm of chromosome 17 through a recurrent translocation, which also correlates with *MYCN* amplification and a more aggressive form of neuroblastoma. Interestingly there is another loss of material related to the chromosome 17 translocation mechanism, which is the loss in 11q. Curiously this deletion is inversely correlated with 1p deletion and *MYCN* amplification, although it is also associated with poor clinical outcome.

As other examples of amplification in neuroblastoma, we have the cases of *NBAS*[100-102] —neuroblastoma amplified sequence gene— and *DDX1*[95,101,102] —RNA helicase gene— which are frequently co-amplified with *MYCN* due to their close upstream location to its loci. As other genetic alterations in this type of cancer, they have also been described amplifications in *MDM2*[103,104] —proto-oncogene—, and frequent rearrangements around *TERT*[105,106] locus —telomerase reverse transcriptase gene— a gene that contributes to the maintenance of the telomeres which is crucial to oncogenic processes.

From the neuroblastoma example and the comparison of its two aberrant genomic profiles, we can extract the importance of the complex structural variants as critical events in the disease and by extension in patient's evolution. Therefore the study of structural variants, such as deletions or translocations and amplification of different genes, can give us essential insight into the development of cancer, specifically neuroblastoma, and the prognosis and clinical outcome of the different patients. However, not all high-risk neuroblastoma patients presenting *MYCN* amplification have the same clinical outcome. For this reason, a comprehensive analysis and characterization of the structural variation patterns and its associated mechanisms in neuroblastoma is essential to gain more knowledge about the development of the disease and the outcome of patients.

Objectives

Under the general goal of our group to find and describe the relationship between genomic structural variation and cancer, in this particular thesis we have followed this idea through more specific objectives.

1. Find patterns of genomic rearrangements associated with the activity of *PGBD5* transposase-derived gene in transformed cell lines.

2. Analyze the prevalence of *PGBD5*-specific structural variants and other transposase-like related patterns of variation in 2,706 normal and tumor pairs across different cancer types from ICGC-PCAWG data.

3. Characterize the potential mechanism associated with recurrent patterns of complex structural variation in 93 neuroblastoma patients.

Contribution to publications

The director of this thesis, Dr. David Torrents Arenales, informs that:

Elias Rodriguez Fos is presenting his Ph.D. thesis entitled "Study of complex chromosomal rearrangements in cancer. The role of extrachromosomal circular DNA as a genome remodeler in neuroblastoma", which has been developed at the Barcelona Supercomputing Center (BSC). During his Ph.D., Elias has contributed to two published studies, including one as a co-first author. These publications correspond to the main work of his thesis. Besides, he has taken part in a third study that has been sent to Nature Communications, and in a fourth publication from a different research topic. Both are included in the appendix of this thesis. In general, due to his biological background, Elias' contribution to the publications consists of performing the majority of bioinformatic analysis, directy focusing on the answering of the underlying biological questions related to structural variation in cancer.

Here below, you can find the scientific contribution made by the Ph.D. student in each of the published articles, as well as the impact factor of the journals.

PGBD5 publication

Title

PGBD5 promotes site-specific oncogenic mutations in human tumors

Authors

Anton G Henssen, Richard Koche, Jiali Zhuang, Eileen Jiang, Casie Reed, Amy Eisenberg, Eric Still, Ian C MacArthur, **Elias Rodríguez-Fos**, Santiago Gonzalez, Montserrat Puiggròs, Andrew N Blackford, Christopher E Mason, Elisa de Stanchina, Mithat Gönen, Anne-Katrin Emde, Minita Shah, Kanika Arora, Catherine Reeves, Nicholas D Socci, Elizabeth Perlman, Cristina R Antonescu, Charles WM Roberts, Hanno Steen, Elizabeth Mullen, Stephen P Jackson, David Torrents, Zhiping Weng, Scott A Armstrong, Alex Kentsis.

Journal	Year	Impact factor	Citations
Nature Genetics	2017	27.125	29

Contribution

This publication in which Elias took part in his first year in the group, started as a collaboration with Alex Kentsis group from the Memorial Sloan Kettering Cancer Center in New York. In 2015, they described the role of *PGBD5*, a transposase-derived gene expressed in the developing embryo and particular areas of the brain, that was found to conserve its transposition activity in the human genome. Knowing that PGBD5 was a fully working transposase with potential mutagenic activity and it was expressed in the majority of childhood solid tumors, they hypothesized that PGBD5 might act as an oncogenic mutator providing a mechanism for site-specific DNA rearrangements in childhood and adult solid tumors.

Elias' contribution to this study was focused on the analysis of the ability of *PGBD5* expression to promote cell transformation and the generation of PGBD5-related rearrangements. The identification of the genomic changes associated with *PGBD5* expression, started with the comparison of the whole-genome sequencing data from *PGBD5*-transformed RPE cells —cells that expressed the gene— and GFP-control RPE cells, using the variant caller previously developed in our group, SMuFin. Based on these results, he was able to identify significant enrichment of small deletions with a small average size associated with the expression of *PGBD5*. Those deletions showed specific characteristics, such as the presence of a highly conserved PSS motif around the breakpoints. Interestingly, these results were in line with the ones found in the analysis of structural variants in rhabdoid tumors within the same study. Elias' work shows evidence that *PGBD5*-induced cell transformation involves site-specific genomic rearrangements that are associated with PGBD5-specific motif breakpoints.

Neuroblastoma publication

Title

Extrachromosomal circular DNA drives oncogenic genome remodeling in neuroblastoma

Authors

Richard P Koche*, **Elias Rodriguez-Fos***, Konstantin Helmsauer*, Martin Burkert, Ian C MacArthur, Jesper Maag, Rocio Chamorro, Natalia Munoz-Perez, Montserrat Puiggròs, Heathcliff Dorado Garcia, Yi Bei, Claudia Röefzaad, Victor Bardinet, Annabell Szymansky, Annika Winkler, Theresa Thole, Natalie Timme, Katharina Kasack, Steffen Fuchs, Filippos Klironomos, Nina Thiessen, Eric Blanc, Karin Schmelz, Annette Künkele, Patrick Hundsdörfer, Carolina Rosswog, Jessica Theissen, Dieter Beule, Hedwig Deubzer, Sascha Sauer, Joern Toedling, Matthias Fischer, Falk Hertwig, Roland F Schwarz, Angelika Eggert, David Torrents*, Johannes H Schulte*, Anton G Henssen*.

*These authors contributed equally. As Elias' thesis director, I certify that this publication has not been used by any other first co-author in their Ph.D. thesis.

Journal	Year	Impact factor	Citations
Nature Genetics	2020	25.455	4

Contribution

Anton Henssen, former postdoc in Alex Kentsis group, who just started a group at the Max Delbrück Center for Molecular Medicine and Charité-Universitätsmedizin in Berlin, contacted us to follow our last collaboration with a new study analyzing whole-genome sequencing and circular DNA sequencing data from 93 neuroblastoma patients.

The contribution of Elias to this study as co-first author was focused on the analysis and classification of the genomic rearrangements in neuroblastoma, looking for patterns of SVs that could further be associated with different clinical outcomes. From his analysis, he discovered recurrent patterns of translocations associated with the integration of circular DNA elements into the linear genome.

For this study, the calling of variants was performed using five different algorithms. Inter- and intrachromosomal rearrangements were merged and post-filtered with the aim to obtain a standardized final set of variants. Using copy number variant information, Elias was able to categorize the circles of DNA inferred from whole-genome sequencing data by their level of amplification. Moreover, with the merged data from structural variation and extrachromosomal circular DNA elements, he managed to classify the landscape of variants depending on their association with circularized regions of the genome. This classification showed a direct association between specific clusters of translocations and the regions of circularization in the genome. Following this idea, he proposed examples of an integration mechanism of circular DNA elements into the linear genome. He also took an active part in the general design of the manuscript, notably with the design of main figures.

Appendix publication 1

Title

Enhancer hijacking determines intra-and extrachromosomal circular *MYCN* amplicon architecture in neuroblastoma

Authors

Konstantin Helmsauer, Maria Valieva, Salaheddine Ali, Rocio Chamorro Gonzalez, Robert Schöpflin, Claudia Roeefzaad, Yi Bei, Heathcliff Dorado Garcia, **Elias Rodriguez-Fos**, Montserrat Puiggros, Katharina Kasack, Kerstin Haase, Luis P

Kuschel, Philipp Euskirchen, Verena Heinrich, Michael Robson, Carolina Rosswog, Joern Toedling, Annabell Szymansky, Falk Hertwig, Matthias Fischer, David Torrents, Angelika Eggert, Johannes H Schulte, Stefan Mundlos, Anton G Henssen, Richard P Koche

Journal	Year	Impact factor	Citations
Under revision in Nature Communications	2019	11.878	-

Contribution

This study was based on the same data as the previously listed neuroblastoma publication. Elias provided the variant calling, filtering and merging of structural variants from the 93 neuroblastoma patients, expanding the analysis to ten additional patients and two relapse samples.

Appendix publication 2

Title

Re-analysis of public genetic data reveals a rare X-chromosomal variant associated with type 2 diabetes

Authors

Sílvia Bonàs-Guarch, Marta Guindo-Martínez, Irene Miguel-Escalada, Niels Grarup, David Sebastian, **Elias Rodriguez-Fos**, Friman Sánchez, Mercè Planas-Fèlix, Paula Cortes-Sánchez, Santi González, Pascal Timshel, Tune H Pers, Claire C Morgan, Ignasi Moran, Goutham Atla, Juan R González, Montserrat Puiggros, Jonathan Martí, Ehm A Andersson, Carlos Díaz, Rosa M Badia, Miriam Udler, Aaron Leong, Varindepal Kaur, Jason Flannick, Torben Jørgensen, Allan Linneberg, Marit E Jørgensen, Daniel

R Witte, Cramer Christensen, Ivan Brandslund, Emil V Appel, Robert A Scott, Jian'an Luan, Claudia Langenberg, Nicholas J Wareham, Oluf Pedersen, Antonio Zorzano, Jose C Florez, Torben Hansen, Jorge Ferrer, Josep Maria Mercader, David Torrents.

Journal	Year	Impact factor	Citations
Nature Communications	2018	11.878	27

Contribution

Elias' contribution in this study was centered in the pathway and gene set enrichment analysis to prioritize likely causal genes, highlighting enriched pathways and identifying the most relevant tissues/cell types involved in Type 2 Diabetes. Additionally, he analyzed the effect of indels using the Ensembl Variant Effect Predictor (VEP). Elias also took part in the generation of supplementary figures for this publication.

Signature

David Torrents Arenales, Ph.D.

PGBD5 publication

Nature Genetics

PGBD5 promotes site-specific oncogenic mutations in human tumors

Authors

Anton G Henssen[1,21,22], Richard Koche[2,22], Jiali Zhuang[3,22], Eileen Jiang[1], Casie Reed[1], Amy Eisenberg[1], Eric Still[1], Ian C MacArthur[1], **Elias Rodríguez-Fos**[4], Santiago Gonzalez[4], Montserrat Puiggròs[4], Andrew N Blackford[5], Christopher E Mason[6], Elisa de Stanchina[7], Mithat Gönen[8], Anne-Katrin Emde[9], Minita Shah[9], Kanika Arora[9], Catherine Reeves[9], Nicholas D Socci[10], Elizabeth Perlman[11], Cristina R Antonescu[12], Charles W M Roberts[13], Hanno Steen[14], Elizabeth Mullen[15], Stephen P Jackson[5,16,17] , David Torrents[4,18], Zhiping Weng[3], Scott A Armstrong[2,19,20] & Alex Kentsis[1,19,20].

Citation

Henssen, A.G., Koche, R., Zhuang, J., Jiang, E., Reed, C., Eisenberg, A., Still, E., Macarthur, I.C., **Rodríguez-Fos, E**., Gonzalez, S., et al. (2017). PGBD5 promotes site-specific oncogenic mutations in human tumors. Nature Genetics 49, 1005–1014.

1 Abstract

Genomic rearrangements are a hallmark of human cancers. Here, we identify the piggyBac transposable element derived 5 (*PGBD5*) gene as encoding an active DNA transposase expressed in the majority of childhood solid tumors, including lethal rhabdoid tumors. Using assembly-based whole-genome DNA sequencing, we found previously undefined genomic rearrangements in human rhabdoid tumors. These rearrangements involved PGBD5-specific signal (PSS) sequences at their breakpoints and recurrently inactivated tumor-suppressor genes. PGBD5

was physically associated with genomic PSS sequences that were also sufficient to mediate PGBD5-induced DNA rearrangements in rhabdoid tumor cells. Ectopic expression of PGBD5 in primary immortalized human cells was sufficient to promote cell transformation in vivo. This activity required specific catalytic residues in the PGBD5 transposase domain as well as end-joining DNA repair and induced structural rearrangements with PSS breakpoints. These results define PGBD5 as an oncogenic mutator and provide a plausible mechanism for site-specific DNA rearrangements in childhood and adult solid tumors.

2 Main

Whole-genome analyses have now produced near-comprehensive topographies of coding mutations for certain human cancers, thus enabling detailed molecular studies of cancer pathogenesis and providing potential for precisely targeted therapies[1,2,3,4,5]. For certain childhood cancers, recent studies have begun to identify the essential functions of complex noncoding structural variants that induce aberrant expression of cellular proto-oncogenes[6,7]. However, for many aggressive childhood cancers, including solid tumors, such studies have identified distinct cancer subtypes that have no discernible coding mutations[8,9,10,11]. In addition, whereas defects in DNA-damage repair have been suggested to explain the increased incidence of some cancers in relatively young people, the causes of complex genomic rearrangements in other cancers in young children without apparent widespread genomic instability remain largely unknown.

Rhabdoid tumors are a prototypical example of this phenomenon. These tumors occur in the developing tissues of infants and children, and exhibit neuroectodermal, epithelial, and mesenchymal components in the brain, liver, kidney, and other organs[10,12,13]. Rhabdoid tumors that cannot be cured through surgery are generally chemotherapy resistant and are almost always lethal[14]. Rhabdoid tumors exhibit inactivating mutations of *SMARCB1*, generally as a result of genomic rearrangements of the 22q11.2 chromosomal locus[15]. These mutations may be inherited as part of the rhabdoid tumor predisposition syndrome but are not thought to involve chromosomal instability[13]. Whereas *SMARCB1* mutations are sufficient to cause rhabdoid tumors in mice[16], human rhabdoid tumors have been observed to have multiple molecular subtypes and rearrangements of additional chromosomal loci that are poorly understood[9,10,17,18]. These findings suggest that additional genetic elements and molecular mechanisms may contribute to the pathogenesis of rhabdoid tumors.

In humans, nearly half of the genome comprises sequences derived from transposons, including both autonomous and nonautonomous mobile genetic elements[19]. Most human genes encoding enzymes that might mobilize transposons appear to be catalytically inactive, with the exception of L1 long interspersed repeated sequences, which appear to induce structural genomic variation in human neurons and adenocarcinomas[20,21,22]; Mariner transposase-derived SETMAR, which functions in DNA repair[23]; and Transib-like DNA transposase RAG1/2, which catalyzes somatic recombination of V(D)J receptor genes in lymphocytes[24]. In particular, aberrant activity of RAG1/2 in lymphoblastic leukemias and lymphomas can induce the formation of chromosomal translocations that generate transforming fusion genes[25,26,27]. The identities of similar genomic rearrangements and the mechanisms by which they may be formed in childhood and adult solid tumors are unknown, but the existence of additional human recombinases that can induce somatic-DNA rearrangements has long been hypothesized[28].

Recently, human PGBD5 and THAP9 have been found to catalyze transposition of synthetic DNA transposons in human cells[29,30]. The physiologic functions of these activities are currently not known. PGBD5 is distinguished by its deep evolutionary conservation among vertebrates (~500 million years) and developmentally restricted expression in tissues from which several childhood solid tumors, including rhabdoid tumors, are thought to originate[30,31]. *PGBD5* is transcribed as a multi-intronic and nonchimeric transcript from a gene encoding a full-length transposase that has become immobilized on human chromosome 1 (refs. 30,31). Genomic transposition activity of PGBD5 requires distinct aspartate residues in its transposase domain as well as specific DNA sequences containing inverted terminal repeats similar to those of *piggyBac* transposons from the lepidopteran *Trichoplusia ni*[30]. These findings, combined with the recent evidence that PGBD5 can induce genomic rearrangements that inactivate the *HPRT1* gene[32], prompted us to investigate whether PGBD5 might induce site-specific DNA rearrangements in human rhabdoid tumors that share developmental origin with cells that normally express *PGBD5*.

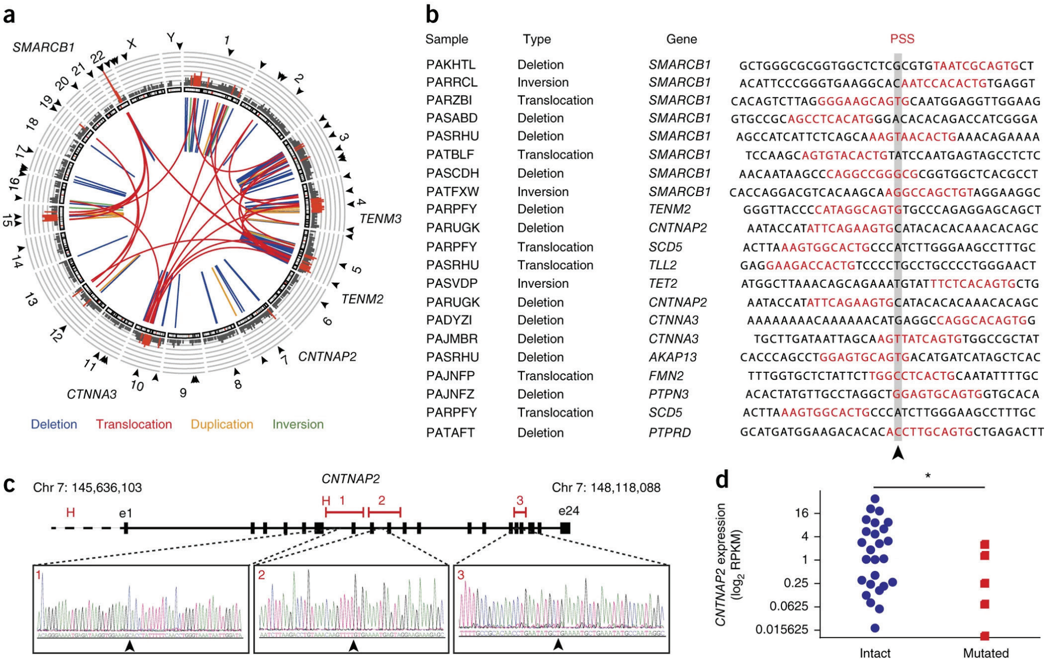
3 Results

3.1 Human rhabdoid tumors exhibit genomic rearrangements associated with PGBD5-specific signal-sequence breakpoints

First, we analyzed the expression of *PGBD5* in large well-characterized cohorts of primary childhood and adult tumors (Supplementary Fig. 1a). We observed that *PGBD5* was highly expressed in a variety of childhood and adult solid tumors, including rhabdoid tumors, but not in acute lymphoblastic or myeloid leukemias (Supplementary Fig. 1a). The expression of *PGBD5* in rhabdoid tumors was similar to that in the embryonal tissues from which these tumors are thought to originate, but it was not significantly associated with currently defined molecular subgroups or patient age at diagnosis (Supplementary Fig. 1a–f). To investigate potential PGBD5-induced genomic rearrangements in primary human rhabdoid tumors, we performed *de novo* structural-variant analysis of whole-genome paired-end Illumina sequencing data for 31 individually matched tumors and normal paired blood specimens from children with extracranial rhabdoid tumors, which are generally characterized on the basis of inactivating *SMARCB1* mutations[10]. Owing to their repetitive nature, sequences derived from transposons present challenges to genome analysis. Thus, we reasoned that genome analysis approaches that do not rely on short-read alignment algorithms, such as the local assembly-based algorithm laSV and the tree-based sequence-comparison algorithm SMuFin, might identify genomic rearrangements that otherwise might escape conventional algorithms[33,34].

Using this assembly-based approach, we observed recurrent rearrangements of the *SMARCB1* gene on chromosome 22q11 in nearly all cases examined, in agreement with the established pathogenic function of inactivating mutations of *SMARCB1* in rhabdoid tumorigenesis (Fig. 1a). In addition, we observed previously unrecognized somatic deletions, inversions, and translocations involving focal regions of chromosomes 1, 4, 5, 10, and 15 (median of three per tumor), which were recurrently altered in more than 20% of cases (Fig. 1a and Supplementary Data Set 1). These results indicated that, in addition to the pathognomonic mutations of *SMARCB1*, human rhabdoid tumors are characterized by additional distinct and recurrent genomic rearrangements.

To determine whether any of the observed genomic rearrangements might be related to PGBD5 DNA transposase or recombinase activity, we first used a forward genetic screen to identify PSS sequences that were specifically found at the breakpoints of PGBD5-induced deletions, inversions, and translocations that caused inactivation of the *HPRT1* gene in a thioguanine resistance assay[32]. Using these PSS sequences as templates for supervised analysis of the somatic genomic rearrangements in primary human rhabdoid tumors, we identified specific PSS sequences associated with the breakpoints of genomic rearrangements in rhabdoid tumors ($P = 1.1 \times 10^{-10}$, hypergeometric test; Fig. 1b and Supplementary Fig. 2). By contrast, we observed no enrichment of the RAG1/2-recombination signal (RSS) sequences at the breakpoints of somatic rhabdoid tumor genomic rearrangements, although the RSS and PSS sequences were equally sized, a result consistent with the lack of RAG1/2 expression in rhabdoid tumors. Likewise, we did not find significant enrichment of PSS motifs at the breakpoints of structural variants and genomic rearrangements in breast carcinomas lacking *PGBD5* expression, even though these breast carcinoma genomes were characterized by high rates of genomic instability (Supplementary Data Set 1). The PSS sequences



Sample	Type	Gene	PSS
PAKHTL	Deletion	*SMARCB1*	GCTGGGCGCGGTGGCTCTCGCGTGTAATCGCAGTGCT
PARRCL	Inversion	*SMARCB1*	ACATTCCCGGGTGAAGGCACAATCCACACTGTGAGGT
PARZBI	Translocation	*SMARCB1*	CACAGTCTTAGGGGAAGCAGTGCAATGGAGGTTGGAAG
PASABD	Deletion	*SMARCB1*	GTGCCGCAGCCTCACATGGGACACACAGACCATCGGGA
PASRHU	Deletion	*SMARCB1*	AGCCATCATTCTCAGCAAAGTAACACTGAAACAGAAAA
PATBLF	Translocation	*SMARCB1*	TCCAAGCAGTGTACACTGTATCCAATGAGTAGCCTCTC
PASCDH	Deletion	*SMARCB1*	AACAATAAGCCCAGGCCGGGCGCGGTGGCTCACGCCT
PATFXW	Inversion	*SMARCB1*	CACCAGGACGTCACAAGCAAGGCCAGCTGTAGGAAGGC
PARPFY	Deletion	*TENM2*	GGGTTACCCCATAGGCAGTGTGCCCAGAGGAGCAGCT
PARUGK	Deletion	*CNTNAP2*	AATACCATATTCAGAAGTGCATACACACAAACACAGC
PARPFY	Translocation	*SCD5*	ACTTAAAGTGGCACTGCCCATCTTGGGAAGCCTTTGC
PASRHU	Translocation	*TLL2*	GAGGAAGACCACTGTCCCCTGCCTGCCCCTGGGAACT
PASVDP	Inversion	*TET2*	ATGGCTTAAACAGCAGAAATGTATTTCTCACAGTGCTG
PARUGK	Deletion	*CNTNAP2*	AATACCATATTCAGAAGTGCATACACACAAACACAGC
PADYZI	Deletion	*CTNNA3*	AAAAAAAACAAAAAAACATGAGGCCAGGCACAGTGG
PAJMBR	Deletion	*CTNNA3*	TGCTTGATAATTAGCAAGTTATCAGTGTGGCCGCTAT
PASRHU	Deletion	*AKAP13*	CACCCAGCCTGGAGTGCAGTGACATGATCATAGCTCAC
PAJNFP	Translocation	*FMN2*	TTTGGTGCTCTATTCTTGGCCTCACTGCAATATTTTGC
PAJNFZ	Deletion	*PTPN3*	ACACTATGTTGCCTAGGCTGGAGTGCAGTGGTGCACA
PARPFY	Translocation	*SCD5*	ACTTAAAGTGGCACTGCCCATCTTGGGAAGCCTTTGC
PATAFT	Deletion	*PTPRD*	GCATGATGGAAGACACACACCTTGCAGTGCTGAGACTT

Figure 1. Human rhabdoid tumors exhibit genomic rearrangements associated with PGBD5-specific signal-sequence breakpoints.

(**a**), Aggregate Circos plot of somatic structural variants identified in 31 human rhabdoid tumors by using laSV, as marked for PSS-containing breakpoints (outer ring, arrowheads), recurrence (middle-ring histogram, rearrangements occurring in ≥3 of 31 samples and highlighted in red for rearrangements with recurrence frequency >13%), and structural-variant type (inner lines, as color-coded). Recurrently rearranged genes are labeled. (**b**), Representation of 21 structural-variant breakpoints in rhabdoid tumors identified to contain PSS sequences (red) within 10 bp of the breakpoint junction (arrowhead). (**c**), Recurrent structural variants of *CNTNAP2* (red) with gene structure (black) and Sanger sequencing of the rearrangement breakpoints. Chr, chromosome (**d**), *CNTNAP2* mRNA expression in primary rhabdoid tumors, as measured through RNA sequencing in *CNTNAP2*-mutant (red) compared with *CNTNAP2*-intact (blue) specimens (*P = 0.017 by one-sided t-test for 26 intact versus 5 mutant *CNTNAP2* individual specimens). RPKM, reads per kilobase of transcript per million mapped reads

.

observed in human rhabdoid tumors exhibited both similarities and differences as compared with those found in the forward genetic screen (Supplementary Fig. 2), thus suggesting that context-dependent factors may control PGBD5 activity. In total, 580 (52%) of 1,121 somatic genomic rearrangements detected in rhabdoid tumors contained PSS sequences near their rearrangement breakpoints (Supplementary Data Set 1).

Overall, the majority of the observed rearrangements were deletions and translocations (Fig. 1a and Supplementary Fig. 3a). Notably, we found recurrent PSS-containing genomic rearrangements affecting the *CNTNAP2*, *TENM2*, *TENM3*, and *TET2* genes (Fig. 1a–c, Supplementary Fig. 3c and Supplementary Data Set 1). Using allele-specific PCR followed by Sanger DNA sequencing, we confirmed three of the observed intragenic *CNTNAP2* deletions and rearrangement breakpoints (Fig. 1c). Likewise, we confirmed the somatic nature of mutations of *CNTNAP2* and *TENM3* by allele-specific PCR in matched tumor and normal primary patient specimens (Supplementary Fig. 3d–h).

CNTNAP2, a member of the neurexin family of signaling and adhesion molecules, has previously been found to function as a tumor-suppressor gene in gliomas[35]. In agreement with the potential pathogenic functions of the apparent *CNTNAP2* rearrangements in rhabdoid tumors identified in our analysis, *CNTNAP2* has recently been reported to be recurrently deleted in an independent cohort of rhabdoid tumor patients[18]. By using comparative RNA-sequencing gene expression analysis in our cohort, we found that primary tumors containing recurrent genomic rearrangements of *CNTNAP2*, as compared with those lacking *CNTNAP2* rearrangements, were indeed associated with a significant decrease in *CNTNAP2* mRNA expression (P = 0.017, t-test; Fig. 1d). Additional mechanisms, including as-yet-undetected mutations or silencing[35], may contribute to the loss of *CNTNAP2* expression in apparently nonrearranged cases (Fig. 1d).

Interestingly, some of the observed genomic rearrangements with PSS-containing breakpoints in rhabdoid tumors involved *SMARCB1* deletions (Fig. 1a,b and Supplementary Data Set 1), thus suggesting that in a subset of rhabdoid tumors, PGBD5 activity itself may contribute to the somatic inactivation of *SMARCB1* in rhabdoid tumorigenesis. Similarly, we observed recurrent interchromosomal translocations and complex structural rearrangements containing breakpoints with the PSS motifs that involved *SMARCB1* (Fig. 1b and Supplementary Data Set 1), including chromosomal translocations, as previously observed through cytogenetic methods[17]. For example, we verified the t(5;22) translocation by using allele-specific PCR followed by Sanger sequencing of the translocation breakpoint (Supplementary Fig. 3i,j). Together, these results indicated that human rhabdoid tumors exhibit recurrent genomic rearrangements that are defined by PSS breakpoint sequences specifically associated with PGBD5, at least some of which appear to be pathogenic and may be coupled with inactivating mutations of *SMARCB1* itself.

3.2 PGBD5 is physically associated with human genomic pss sequences that are sufficient to mediate DNA rearrangements in rhabdoid tumor cells

In prior studies, human PGBD5 has been found to localize to cell nuclei[31]. To test whether PGBD5 in rhabdoid tumor cells is physically associated with genomic PSS-containing sequences, as would be predicted for a DNA transposase that induces genomic rearrangements, we used chromatin immunoprecipitation followed by DNA sequencing (ChIP–seq) to determine the genomic localization of endogenous PGBD5 in human G401 rhabdoid tumor cells. We observed that human DNA regions bound by PGBD5 were significantly enriched in PSS motifs ($P = 2.9 \times 10^{-29}$, hypergeometric test), in contrast to scrambled PSS sequences of identical composition or functionally unrelated RSS sequences of equal size, neither of which showed significant enrichment ($P = 0.28$ and 1.0, respectively, hypergeometric test; Fig. 2a).

To test the hypothesis that PGBD5 can act directly on human PSS-containing DNA sequences and mediate their genomic rearrangement, we used the previously established DNA transposition reporter assay[30]. Human embryonic kidney (HEK) 293 cells were transiently transfected with plasmids for expression of human GFP-PGBD5, hyperactive lepidopteran *T. ni* GFP-piggyBac DNA transposase or control GFP in the presence of reporter plasmids for expression of the neomycin-resistance gene (*NeoR*) flanked by a human PSS sequence, as identified from rhabdoid tumor rearrangement breakpoints (Supplementary Figs. 2 and 3 and

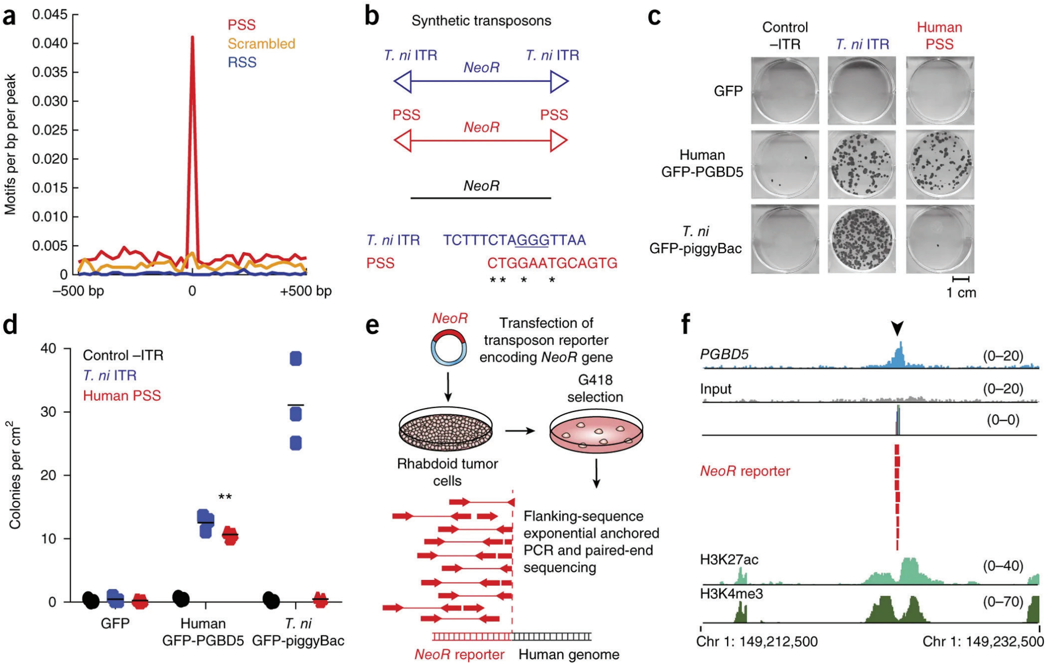


Figure 2. PGBD5 is physically associated with human genomic PSS sequences that are sufficient to mediate DNA rearrangements in rhabdoid tumor cells.

(**a**), Genomic distribution of PGBD5 protein in G401 rhabdoid tumor cells as a function of enrichment of PSS (red), compared with scrambled PSS (orange) and RSS (blue) controls, as measured through PGBD5 ChIP–seq ($P = 2.9 \times 10^{-29}$ for PSS, $P = 0.28$ for scrambled PSS, $P = 1.0$ for RSS, by hypergeometric test for 622 sites). (**b**), Top, schematic of synthetic transposon substrates used for DNA transposition assays, including transposons with *T. ni* ITR (blue triangles), transposons with PSSs (red triangles), and transposons lacking ITRs (black). Bottom, sequence alignment of *T. ni* ITR compared with human PSS (bottom). (**c**), Representative photographs of crystal violet–stained colonies obtained after G418 selection in the transposon reporter assay for three independent experiments. (**d**), Genomic DNA transposition assay, as measured on the basis of neomycin-resistance clonogenic assays in HEK293 cells cotransfected with human *GFP-PGBD5* or control *GFP* and *T.ni GFP-piggyBac*, and transposon reporters encoding the *NeoR* gene flanked by human PSS (red), as compared with control reporters lacking inverted terminal repeats (–ITR, black) and *T. ni piggyBac* ITR (blue). $**P = 5.0 \times 10^{-5}$, two-sided t-test of three independent experiments. Lepidopteran *T. ni PiggyBac* DNA transposase and its *piggyBac* ITR served as specificity controls. Errors bars, s.d. of three independent experiments. (**e**), Schematic model of transposition reporter assay in G401 rhabdoid tumor cells and subsequent FLEA PCR and Illumina paired-end sequencing. (**f**), Genomic integration of synthetic NeoR transposons (red) by endogenous PGBD5 in G401 rhabdoid tumor cells at the PSS site (arrowhead). ChIP–seq genome tracks of *PGBD5* (blue) compared with its sequencing input (gray), and Lys27-acetylated (H3K27ac) and Lys4-trimethylated (H3K4me3) histone H3 (bottom), suggesting bound PGBD5 transposase protein complex, are shown.

Supplementary Data Set 1), lepidopteran *piggyBac* inverted terminal repeat (ITR) transposon sequence[30], or control plasmids lacking flanking transposon elements (Fig. 2b). Clonogenic assays of transfected cells in the presence of G418 to select for neomycin-resistant cells with genomic reporter integration demonstrated that GFP-PGBD5, but not control GFP, exhibited efficient activity toward reporters containing terminal repeats with the human PSS sequences but not control reporters lacking flanking transposon elements ($P = 5.0 \times 10^{-5}$, t-test; Fig. 2c,d). This activity was specific, because the lepidopteran GFP-piggyBac DNA transposase, which efficiently mobilizes its own *piggyBac* transposons, did not mobilize reporter plasmids containing human PSS sequences (Fig. 2c,d).

To determine whether endogenous PGBD5 can mediate genomic rearrangements in rhabdoid cells, we transiently transfected human G401 rhabdoid cells with the *NeoR* transposon reporter plasmids and determined their chromosomal integration by using flanking-sequence exponential anchored (FLEA) PCR to amplify and sequence-specific segments of the human genome flanking transposon-integration sites[30] (Fig. 2e and Supplementary Fig. 4). Similar assays in HEK293 cells lacking *PGBD5* expression did not induce measureable genomic integration of reporter transposons (Fig. 2c,d). In contrast, we observed that endogenous PGBD5 in G401 rhabdoid tumor cells was sufficient to mediate integration of transposon-containing DNA into human genomic PSS-containing sites (Fig. 2f and Supplementary Tables 1 and 2). This activity was specifically observed for transposon reporters with intact transposons but not those in which the essential 5'-GGGTTAACCC-3' hairpin structure was mutated to 5'-ATATTAACCC-3' (location of mutation underlined; Supplementary Table 1). Thus, PGBD5 physically associates with human genomic PSS sequences that are sufficient to mediate DNA rearrangements of synthetic reporters in rhabdoid tumor cells.

3.3 PGBD5 expression in genomically stable primary human cells is sufficient to induce malignant transformation in vitro and in vivo

Recurrent somatic genomic rearrangements in primary rhabdoid tumors associated with PGBD5-specific signal-sequence breakpoints, their targeting of tumor-suppressor genes, and their specific activity as genomic-rearrangement substrates suggest that PGBD5 DNA transposase activity might be sufficient to induce tumorigenic mutations that contribute to malignant cell transformation. To determine whether PGBD5 can act as a human-cell-transforming factor, we used established transformation assays of primary human foreskin BJ and retinal pigment epithelial (RPE) cells immortalized with telomerase[36]. Primary RPE and BJ cells at passage 3–5 are immortalized by the expression of human *TERT* telomerase *in vitro*, undergo growth arrest after contact inhibition, and do not form tumors after transplantation into immunodeficient mice *in vivo*[36]. Prior studies have established the essential requirements for their malignant transformation through the concomitant dysregulation of p53, Rb, and Ras pathways[36]. Thus, transformation of primary human RPE and BJ cells enables detailed studies of human PGBD5 genetic mechanisms that cannot be performed in mouse or other heterologous model systems.

To test whether *PGBD5* has transforming activity in human cells, we used lentiviral transduction to express *GFP-PGBD5* and control *GFP* transgenes in telomerase-immortalized RPE and BJ cells, at levels 1.1- to 5-fold and 1.5- to 8-fold higher than those in primary rhabdoid tumor specimens and cell lines, respectively (Fig. 3a,b). We observed that *GFP-PGBD5*-expressing, but not nontransduced or *GFP*-expressing, RPE and BJ cells formed retractile colonies in monolayer cultures and exhibited anchorage-independent growth in semisolid cultures, a hallmark of cell transformation (Fig. 3c,d). When transplanted into immunodeficient mice, *GFP-PGBD5*-expressing RPE and BJ cells formed subcutaneous tumors with latency and penetrance similar to those observed in cells expressing both mutant *HRAS* and the SV40 large T antigen, which dysregulates both p53 and Rb pathways (Fig.

3f,g and Supplementary Fig. 5). Importantly, both RPE and BJ cells transformed with *GFP-PGBD5* had stable diploid karyotypes when they were passaged *in vitro* (Supplementary Fig. 6). By contrast, expression of the distantly related lepidopteran GFP-piggyBac DNA transposase, which exerts specific and efficient transposition activity on lepidopteran *piggyBac* transposon sequences (Fig. 2d), did not transform human RPE cells (Fig. 3e), in spite of being equally expressed (Supplementary Fig. 7a). These results indicated that the PGBD5 transposase can specifically transform human cells in the absence of chromosomal instability both *in vitro* and *in vivo*.

3.4 PGBD5-induced cell transformation requires DNA transposase activity

To test whether the cell-transforming activity of PGBD5 requires the enzymatic activity of its transposase, we used PGBD5 point mutants that are either proficient or deficient in DNA transposition reporter assays[30]. Thus, we compared p.Glu373Ala and p.Glu365Ala PGBD5 mutants, which retain wild-type transposition activity[30], with p.Asp168Ala, p.Asp194Ala, and p.Asp386Ala, or their double mutant (DM) p.[Asp194Ala]+[Asp386Ala] and triple mutant (TM) p.[Asp168Ala]+[Asp194Ala]+[Asp386Ala], which occur at residues required for efficient DNA transposition *in vitro*, in agreement with their evolutionary conservation and putative function as the DDD/E catalytic triad for phosphodiester-bond hydrolysis[30]. After confirming stable and equal expression of these PGBD5 mutants in RPE cells by protein blotting (Fig. 4a), we assessed their transforming activity with contact inhibition assays in monolayer cultures and transplantation into immunodeficient mice. Whereas ectopic expression of wild-type GFP-PGBD5 induced efficient and fully penetrant cell transformation, the p.Asp168Ala, p.Asp194Ala, DM, and TM deficient in transposition function in reporter assays did not induce contact inhibition in vitro or tumor formation *in vivo* (Fig. 4b,d). By contrast, transposition-proficient p.Glu373Ala and p.Glu365Ala mutants exhibited transforming activity equivalent to that of wild-type GFP-PGBD5 (Fig. 4b,d). Importantly, we confirmed that the catalytic mutants of GFP-PGBD5 on average retained their chromatin localization compared with that of wild-type PGBD5, as

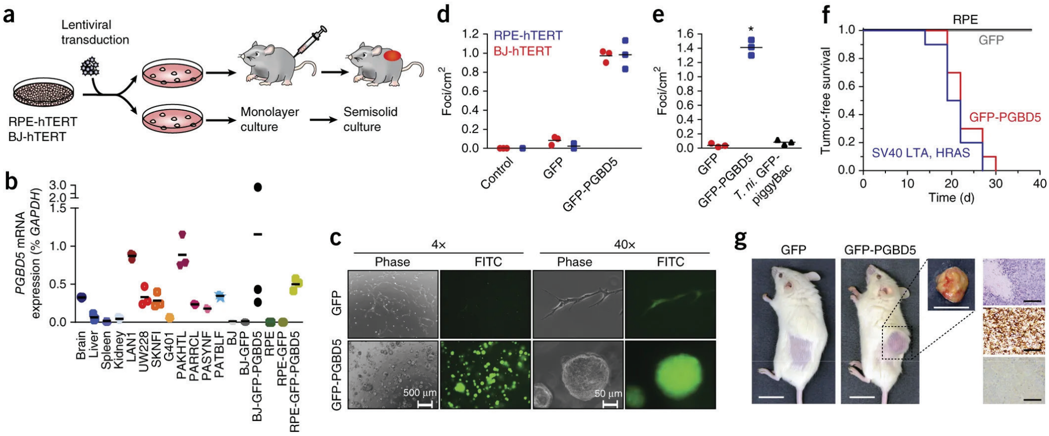


Figure 3. Ectopic expression of *PGBD5* in human cells leads to oncogenic transformation both *in vitro* and i*n vivo.*

(**a**), Schematic for testing transforming activity of PGBD5. (**b**), Relative *PGBD5* mRNA expression measured by quantitative RT–PCR in normal mouse tissues (brain, liver, spleen, and kidney), as compared with human tumor cell lines (rhabdoid G401, neuroblastoma LAN1 and SK-N-FI, and medulloblastoma UW-228 cells), primary human rhabdoid tumors (PAKHTL, PARRCL, PASYNF, and PATBLF), and BJ and RPE cells stably transduced with *GFP-PGBD5* and *GFP*. Error bars, s.d. of three independent measurements. (**c**), Representative images of *GFP-PGBD5*-transduced RPE cells grown in semisolid medium after 10 d of culture in three independent experiments, as compared with control GFP-transduced cells. FITC, fluorescein isothiocyanate. (**d**), Number of refractile foci formed in monolayer cultures of RPE and BJ cells expressing *GFP-PGBD5* or *GFP*, as compared with nontransduced cells (P = 3.6 × 10^{-5} and 3.9 × 10^{-4} by two-sided t-test for *GFP-PGBD5* versus *GFP* for BJ and RPE cells, respectively, in three independent experiments). (**e**), Expression of *T. ni GFP-piggyBac* does not lead to the formation of anchorage-independent foci in monolayer culture (*P = 3.49 × 10^{-5} for *GFP-PGBD5* versus *T. ni GFP-piggyBac*). Error bars, s.d. of three independent experiments. (**f**), Kaplan–Meier analysis of tumor-free survival of mice with subcutaneous xenografts of RPE cells expressing *GFP-PGBD5* or *GFP* control, as compared with nontransduced cells or cells expressing SV40 large T antigen (LTA) and *HRAS* (n = 10 mice per group; P < 0.0001 by log-rank test). (**g**), Representative photographs (from left) of mice with shaved flanks bearing RPE xenografts (scale bar, 1 cm), with 10 mice per treatment group. Tumor excised from mouse bearing a *GFP-PGBD5*-expressing tumor (scale bar, 1 cm). Photomicrographs of *GFP-PGBD5*-expressing tumors (from top to bottom, hematoxylin and eosin stain, vimentin, and cytokeratin; scale bars, 1 mm).

assessed with ChIP–seq (Fig. 4c). Although the p.Asp386Ala mutant exhibited decreased transposition activity in reporter assays *in vitro*[30], its expression induced wild-type transforming activity *in vivo* (Fig. 4d). This result suggested that the transforming activity of PGBD5 may involve noncanonical DNA transposition or recombination reactions, in agreement with the dispensability of some catalytic residues for certain types of DNA transposase–induced DNA rearrangements[37,38]. Thus, cell transformation induced by PGBD5 requires its nuclease activity.

3.5 Transient expression of PGBD5 is sufficient for PGBD5-induced cell transformation

If PGBD5 can induce transforming genomic rearrangements, then transient exposure to PGBD5 should be sufficient to heritably transform human cells. To test this prediction, we generated doxycycline-inducible *PGBD5*-expressing RPE cells and performed protein blotting, which confirmed a lack of detectable expression of the enzyme in the absence of doxycycline and its induction after exposure to doxycycline *in vitro* (Supplementary Fig. 7b). The transduced cells, which were transplanted into immunodeficient mice whose doxycycline chow treatment (−Dox) was stopped after the appearance of macroscopic signs of tumor formation (Fig. 5a and Supplementary Fig. 7c), produced essentially the same tumorigenicity as that observed in continuously treated (+Dox) animals or in animals transplanted with cells constitutively expressing *GFP-PGBD5* (Supplementary Fig. 7c). Importantly, we confirmed the absence of measureable PGBD5 expression in tumors harvested from −Dox animals by protein blotting (Fig. 5a). In agreement with cell transformation by transient expression of *PGBD5*, −Dox and +Dox tumors were histopathologically indistinguishable (Fig. 5b). To investigate the potential irreversibility and heritability of cell transformation induced by transient PGBD5 expression, we transplanted tumors harvested from −Dox and +Dox animals into secondary recipients and observed that tumors were induced with the same latency and penetrance in both −Dox and +Dox animals (Fig. 5a).In agreement with this model of PGBD5-induced cell transformation, endogenous PGBD5 in established G401 and A204 rhabdoid tumor cells was observed to be dispensable for cell survival, as assessed though use of small hairpin RNA (shRNA) interference with two different shRNA vectors and a control shRNA targeting GFP (Fig. 5c,d). Thus, transient expression of *PGBD5* is sufficient to transform cells, as would be predicted from the ability of a catalytically active transposase to induce heritable cellular alterations.

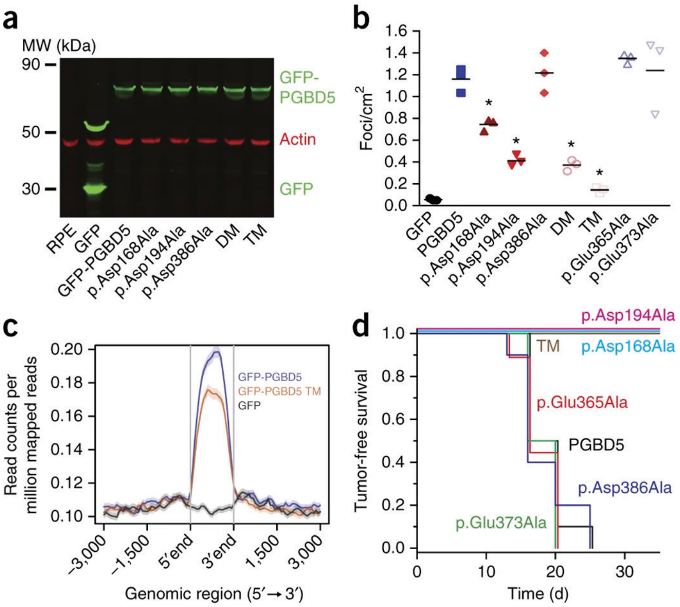


Figure 4. PGBD5 transposase activity is necessary to transform human cells.

(**a**), Protein blot of GFP in RPE cells expressing *GFP-PGBD5*, *GFP-PGBD5* mutants, or *GFP*, compared with RPE cells (DM, double mutant p.[Asp194Ala]+[Asp386Ala]; TM, triple mutant p.[Asp168Ala]+[Asp194Ala]+[Asp386Ala]); MW, molecular weight. Actin, loading control. (**b**), Number of refractile foci formed in monolayer culture in RPE and BJ cells stably expressing *GFP-PGBD5* or control *GFP*, as compared with nontransduced cells and cells expressing *GFP-PGBD5* mutants (red, transposase-deficient mutants; blue, transposase-proficient mutants; *P = 2.1 × 10^{-4} for p.Asp168Ala versus GFP-PGBD5; P = 2.7 × 10^{-6} for p.Asp194Ala versus GFP-PGBD5; P = 1.8 × 10^{-6} for p.[Asp194Ala]+[Asp386Ala] versus GFP-PGBD5; P = 2.4 × 10^{-7} for p.[Asp168Ala]+[Asp194Ala]+[Asp386Ala] versus GFP-PGBD5 by two-sided t-test). Error bars, s.d. of three independent experiments. (**c**), Composite plot of ChIP–seq of GFP-PGBD5 (blue), as compared with the GFP-PGBD5 p.[Asp168Ala]+[Asp194Ala]+[Asp386Ala] catalytic TM (orange) and GFP control (Black). (**d**), Kaplan–Meier analysis of tumor-free survival of mice with subcutaneous xenografts of RPE cells expressing *GFP-PGBD5*, as compared with cells expressing *GFP-PGBD5* mutants (n = 10 mice per group; P < 0.0001 by two-tailed log-rank test).

3.6 PGBD5-induced transformation requires DNA end-joining repair

If PGBD5-induced cell transformation involves transposase-mediated genomic rearrangements, then this process should depend on the repair of DNA double-strand breaks (DSBs) generated by the DNA-recombination reactions[39]. Genomic rearrangements induced by transposases of the DDD/E superfamily involve transesterification reactions, which generate DSBs that are predominantly repaired by DNA nonhomologous end-joining (NHEJ) in somatic cells[40], as is the case for human V(D)J rearrangements induced by the RAG1/2 recombinase[38]. To test whether PGBD5-induced cell transformation requires NHEJ, we used isogenic RPE cells that were wild type or deficient in the NHEJ cofactor PAXX (encoded by *C9orf142*), which stabilizes the NHEJ repair complex and is required for efficient DNA repair[41]. In contrast to defects in other NHEJ components, such as LIG4, PAXX deficiency does not appreciably alter cell growth or viability but significantly decreases NHEJ efficiency without requiring TP53 inactivation to survive[41]. Thus, we generated RPE cells that expressed doxycycline-inducible *PGBD5* and were *C9orf142*$^{+/+}$ or *C9orf142*$^{-/-}$, and confirmed the induction of PGBD5 and lack of PAXX expression by protein blotting (Fig. 6a). Doxycycline-induced expression of PGBD5 in C9orf142$^{-/-}$ but not isogenic *C9orf142*$^{+/+}$ RPE cells caused the accumulation of DNA-damage-associated phosphorylated histone H2AX (γH2AX) (Fig. 6b and Supplementary Fig. 8b), apoptosis-associated cleavage of caspase 3 (Fig. 6c and Supplementary Fig. 8a), and cell death (Supplementary Fig. 8c). We confirmed the requirement of NHEJ for the repair of PGBD5-induced rearrangements by using *Xrcc5*-deficient mouse embryonic fibroblasts (data not shown). Importantly, PGBD5-mediated induction of DNA damage and cell death in NHEJ-deficient *C9orf142*$^{-/-}$ cells, as compared with isogenic NHEJ-proficient *C9orf142*$^{+/+}$ cells, was nearly completely rescued by the p.[Asp168Ala]+[Asp194Ala]+[Asp386Ala] alteration of residues required for the transposase activity of PGBD5 (Fig. 6d). Thus, NHEJ DNA repair is required for the survival of cells expressing active PGBD5.

3.7 PGBD5-induced cell transformation involves site-specific genomic rearrangements associated with PGBD5-specific signal-sequence breakpoints

The requirements for PGBD5 enzymatic transposase activity, cellular NHEJ DNA repair, and the ability of transient *PGBD5* expression to promote cell transformation

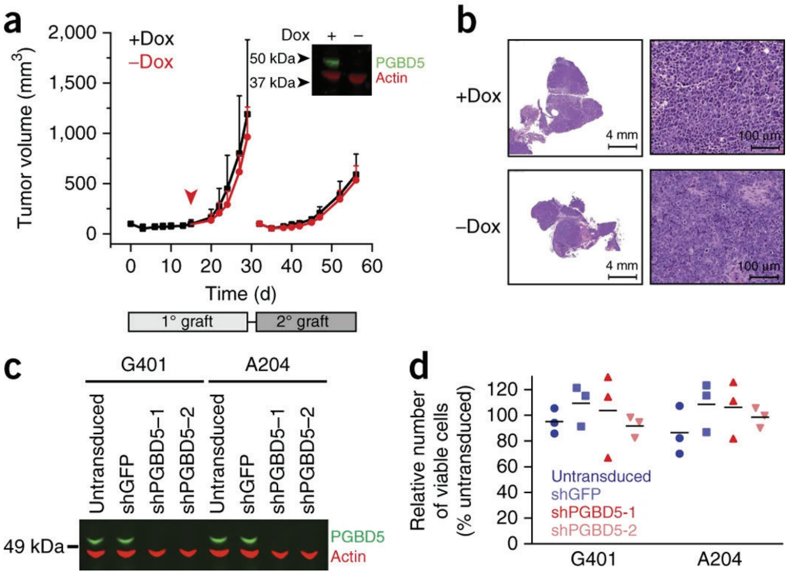


Figure 5. Transient PGBD5 transposase expression is sufficient to transform human cells.

(**a**), Tumor volume of RPE cells as a function of time in primary (1°, light-gray box) and secondary (2°, dark-gray box) transplants, with *PGBD5* expression induced by doxycycline (black), as indicated. RPE cells were treated with doxycycline *in vitro* for 10 d before transplantation. Red arrowhead denotes withdrawal of doxycycline from the diet. Inset, protein blot of PGBD5 protein and actin control, in cells derived from tumors after primary transplant. (**b**), Representative photomicrographs of hematoxylin- and eosin-stained tumor sections from doxycycline-inducible *PGBD5*-expressing RPE tumors after continuous (+Dox) and discontinuous (–Dox) doxycycline treatment, with 10 mice per experimental group. (**c**), Protein blot of PGBD5 in G401 and A204 rhabdoid tumor cells after depletion of *PGBD5* with two independent shRNAs, as compared with nontransduced cells and control cells expressing shGFP. The prefix 'sh' denotes shRNAs. Actin, loading control. (**d**), Relative number of viable G401 and A204 cells 72 h after *PGBD5* shRNA depletion. Errors bars, s.d. of three independent experiments.

are all consistent with the generation of heritable genomic rearrangements that mediate PGBD5-induced tumorigenesis. To determine the genetic basis of PGBD5-induced cell transformation, we sequenced whole genomes of PGBD5-induced tumors as well as control GFP-expressing and nontransduced RPE cells, by using massively parallel paired-end Illumina sequencing at a coverage in excess of 80-fold for over 90% of the genome (Supplementary Data Set 1). As for the rhabdoid tumor genome analysis, we used the assembly-based algorithm laSV as well as conventional techniques[33,34] (Supplementary Table 3, Supplementary Figs. 9, 10, 11 and Supplementary Data Set 1). This analysis led to the identification of distinct genomic rearrangements, specifically in PGBD5-induced tumor cell genomes, as compared with those of control GFP-transduced and nontransduced RPE cells (Fig. 7a). The identified rearrangements were characterized by intrachromosomal deletions with a median length of 183 bp, in agreement with their apparent limited detectability through conventional genome analysis methods, as well as inversions, duplications, and translocations (Supplementary Fig. 12a–c and Supplementary Data Set 1). As with genomic rearrangements found in primary human tumors (Fig. 1), the genomic rearrangements found in PGBD5-transformed RPE cells revealed significant enrichment of PSS motifs at the breakpoints of PGBD5-induced tumor structural variants ($P = 7.2 \times 10^{-3}$, hypergeometric test; Fig. 7b and Supplementary Data Set 1). By contrast, the breakpoints of structural variants in the GFP-control RPE-cell genomes, presumably at least in part because of normal genetic variation, exhibited no enrichment in PSS motifs ($P = 0.37$). We independently verified these findings by using the direct tree-graph-based comparative SMuFin analysis method (Supplementary Fig. 12a and Supplementary Data Set 1). In addition, we validated five of these rearrangements by using variant and wild-type allele-specific PCR followed by Sanger DNA sequencing of rearrangement breakpoints, to confirm that they were specifically present in PGBD5-transformed but not control GFP-transduced RPE cells (Supplementary Fig. 12d–h). Additionally, we did not find genomic-rearrangement breakpoints containing RSS sequences that were targeted by the RAG1/2 recombinase, which is not expressed in RPE cells. We also did not find evidence of structural alterations of the annotated human *MER75* and *MER85 piggyBac*-like transposable elements, in agreement with the distinct evolutionary history of human *PGBD5* (ref. 30). We found that the genomic rearrangements and structural variants observed in PGBD5-induced RPE tumors were significantly enriched in regulatory DNA elements important for normal human embryonal as opposed to adult tissue development (Fig. 7c and Supplementary Table 4).

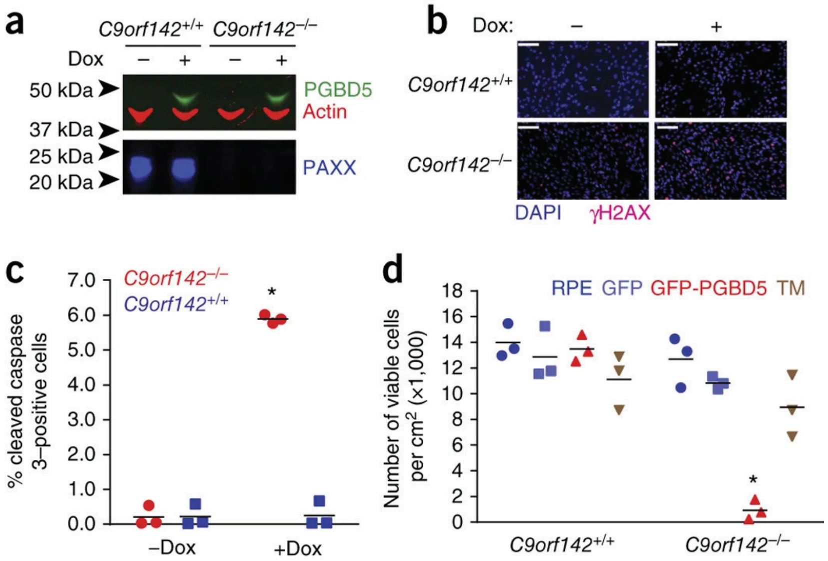


Figure 6. DNA end-joining repair is required for survival of cells expressing active PGBD5.

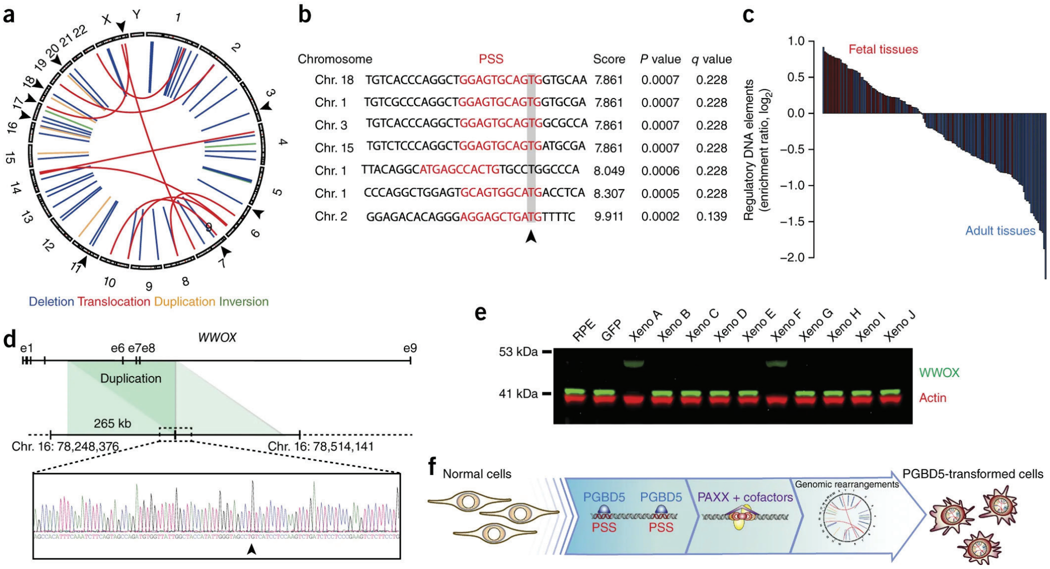
(**a**), Protein blot of PGBD5 protein after 24 h of doxycycline (500 ng/ml) treatment of isogenic *C9orf142*$^{+/+}$ and *C9orf142*$^{-/-}$ RPE cells stably expressing doxycycline-inducible *PGBD5*. Actin, loading control. (**b**), Representative photomicrograph of *C9orf142*$^{+/+}$ and *C9orf142*$^{-/-}$ RPE cells after 48 h treatment with doxycycline (500 ng/ml) or vehicle control, stained for DAPI (blue) and γH2AX (red) in three independent experiments. Scale bars, 100 μm. (**c**), Fraction of apoptotic cells, as determined by cleaved caspase-3 staining and flow cytometric analysis of *C9orf142*$^{+/+}$ and *C9orf142*$^{-/-}$ RPE cells after treatment with doxycycline or vehicle control. *$P = 8.7 \times 10^{-4}$ by two-sided t-test for *C9orf142*$^{+/+}$ versus *C9orf142*$^{-/-}$ with doxycycline in three independent experiments. (**d**), Number of viable *C9orf142*$^{+/+}$ and *C9orf142*$^{-/-}$ RPE cells per cm^2 in monolayer culture, as measured by trypan blue staining after 48 h of expression of GFP-PGBD5, as compared with GFP-PGBD5 p.[Asp168Ala]+[Asp194Ala]+[Asp386Ala]–mutant and GFP-expressing control cells. *$P = 7.4 \times 10^{-5}$ by two-sided t-test for *C9orf142*$^{-/-}$ GFP-PGBD5 versus GFP control. Error bars, s.d. of three independent experiments.

To identify genomic rearrangements that might be functionally responsible for PGBD5-induced cell transformation, we analyzed the recurrence of PGBD5-induced genomic rearrangements in ten different RPE tumors from independent transduction experiments in individual mouse xenografts. We detected 59 PGBD5-induced structural variants per tumor, 42 (71%) of which were deletions, 36 (61%) of which affected regulatory intergenic elements, and 13 (22%) of which contained PSS motifs at their breakpoints (Supplementary Data Set 1). In particular, we identified recurrent and clonal PSS-associated rearrangements of *WWOX*, including duplication of exons 6–8 (Fig. 7d). *WWOX* is a tumor-suppressor gene that controls p53 signaling[42]. We confirmed the duplication of exons 6–8 of *WWOX* by PCR and Sanger DNA sequencing (Fig. 7d), and tested its functional consequence on WWOX protein expression by protein blotting (Fig. 7e). Remarkably, this mutation resulted in low-level expression of the extended mutant form of the WWOX protein, which is associated with loss of wild-type *WWOX* expression, in agreement with the dominant-negative or gain-of-function activity of mutant *WWOX* in RPE-cell transformation. We observed this mutation in two out of ten independent RPE tumors, a result consistent with its probable pathogenic function in PGBD5-induced cell transformation.

To determine its function in PGBD5-induced RPE-cell transformation, we depleted endogenous WWOX and ectopically expressed wild-type WWOX in nontransformed wild-type and *WWOX*-mutant PGBD5-induced RPE-cell tumors (Supplementary Fig. 13a,d). In agreement with the tumorigenic function of PGBD5-induced mutations of *WWOX*, we found that *WWOX* inactivation was necessary but not sufficient to maintain clonogenicity of PGBD5-transformed RPE tumor cells *in vitro* (Supplementary Fig. 13b,c,e,f). Thus, PGBD5-induced cell transformation involves site-specific genomic rearrangements that are associated with PGBD5-specific signal-sequence breakpoints that recurrently target regulatory elements and tumor-suppressor genes (Fig. 7f).

4 Discussion

Here, we found that primary human rhabdoid tumor genomes exhibit signs of PGBD5-mediated DNA recombination involving recurrent mutations of previously elusive rhabdoid tumor-suppressor genes (Fig. 1). These genomic rearrangements involve breakpoints associated with the PSS sequences, which are sufficient to mediate DNA rearrangements in rhabdoid tumor cell lines and physical recruitment



Chromosome	PSS	Score	*P* value	*q* value
Chr. 18	TGTCACCCAGGCTGGAGTGCAGTGGTGCAA	7.861	0.0007	0.228
Chr. 1	TGTCGCCCAGGCTGGAGTGCAGTGGTGCGA	7.861	0.0007	0.228
Chr. 3	TGTCACCCAGGCTGGAGTGCAGTGGCGCCA	7.861	0.0007	0.228
Chr. 15	TGTCTCCCAGGCTGGAGTGCAGTGATGCGA	7.861	0.0007	0.228
Chr. 1	TTACAGGCATGAGCCACTGTGCCTGGCCCA	8.049	0.0006	0.228
Chr. 1	CCCAGGCTGGAGTGCAGTGGCATGACCTCA	8.307	0.0005	0.228
Chr. 2	GGAGACACAGGGAGGAGCTGATGTTTTC	9.911	0.0002	0.139

Figure 7. PGBD5-induced cell transformation involves site-specific genomic rearrangements associated with PGBD5-specific signal-sequence breakpoints.

(**a**), Circos plot of structural variants discovered in RPE-GFP-PGBD5 tumor cells through assembly-based genome analysis. Black arrows on outer circle indicate the presence of PSS at variant breakpoints. (**b**), Representation of seven breakpoints identified to contain PSS sequences (red) within 10 bp of the breakpoint junction (arrowhead) of structural variants in PGBD5-expressing RPE cells. Genomic sequence is annotated 5' to 3', as presented in the reference-genome (+) strand. (**c**), Waterfall plot of enrichment of ENCODE regulatory DNA elements with structural variants in fetal (red) and adult tissues (blue) in PGBD5-transformed RPE cells ($P = 1.9 \times 10^{-5}$ by hypergeometric test for 59 variants in 154 cell types). (**d**), Schematic of the *WWOX* gene and its intragenic duplication in *GFP-PGBD5*-transformed RPE cells (top), with Sanger sequencing chromatogram of the rearrangement breakpoint (bottom). The black arrowhead marks the breakpoint. (**e**), Protein blot analysis of WWOX in ten independent *GFP-PGBD5*-transformed RPE-cell tumor xenografts (xeno), as compared with control *GFP*-transduced and nontransduced RPE cells. Actin, loading control. (**f**), Schematic model of the proposed mechanism of PGBD5-induced cell transformation, involving association of PGBD5 with genomic PSS sequences, their remodeling dependent on PAXX-meditated end-joining DNA repair, and generation of tumorigenic genomic rearrangements.

of endogenous PGBD5 transposase (Fig. 2). The enzymatic activity of PGBD5 is both necessary and sufficient to promote similar genomic rearrangements in primary human cells, thus causing their malignant transformation (Figs. 3, 4, 5, 6, 7).

PGBD5-induced genomic rearrangements exhibit a defined architecture, including characteristic deletions, inversions, and complex rearrangements distinct from those generated by other known mutational processes. We observed an imprecise relationship between PSS sequences and genomic-rearrangement breakpoints, with evidence of incomplete 'cut and paste' DNA transposition, in agreement with potentially aberrant targeting of PGBD5 nuclease activity. Although our structure–function studies suggested that PGBD5 induces genomic rearrangements in conjunction with the canonical NHEJ apparatus, it is possible that PGBD5 activity may also promote other DSB-repair pathways, such as alternative microhomology-mediated end-joining (Supplementary Fig. 14). We confirmed that the putative catalytic aspartate mutants of PGBD5 on average maintained the chromatin localization of wild-type PGBD5. It is also possible that these residues contribute to cell transformation, owing to their interaction with cellular cofactors or assembly of DNA-regulatory complexes, or yet-unknown nuclease-independent functions that contribute to cell transformation.

PSS-associated genomic rearrangements induced by PGBD5 in rhabdoid tumors are reminiscent of McClintock's 'mutable loci' induced by DNA transposase–mediated mutation of the *Ds* locus, which controls position-effect variegation in maize[24,43]. Insofar as nuclease substrate accessibility is controlled by chromatin structure and conformation, PGBD5-induced genomic rearrangements indeed may be coupled to developmental regulatory programs that control gene expression and specification of cell fate, as suggested by their strong association with developmental regulatory DNA elements in our analysis. The association of PGBD5-induced rearrangements may involve sequence-specific recognition of human genomic PSS sequences or alternatively may be determined by their accessibility or the presence of cellular cofactors, as determined by the cellular developmental states.

Importantly, the spectrum of PGBD5-induced genomic rearrangements and their PSS sequences identified in this study should provide a useful approach for the functional characterization of tumor genomes and identification of cancer-causing genomic alterations. In the case of rhabdoid tumors, the association of *SMARCB1* mutations with additional recurrent genomic lesions, such as structural alterations in *CNTNAP2*, *TENM2*, and *TET2* genes, which regulate developmental and epigenetic cell-fate specification, may lead to the identification of additional mechanisms of childhood cancer pathogenesis, including those that cooperate with the dysregulation of SWI–SNF–BAF-mediated chromatin and nucleosome remodeling induced by *SMARCB1* loss. While the current study was under review, an additional genome analysis of rhabdoid tumors was described, and the results independently identified recurrent mutations at *CNTNAP2* and other loci in human rhabdoid tumors[44]. Notably, the recurrence patterns of PGBD5-induced genomic rearrangements in rhabdoid tumors indicate that, even for rare cancers, comprehensive tumor genome analyses will be necessary to define the spectrum of causal genomic lesions and potential therapeutic targets. Our results also indicated that improved genome-analysis methods, such as SMuFin and laSV used in our work[33,34], and confirmation of their sensitivity and specificity will be needed to elucidate tumorigenic genome rearrangements. Similarly, given the existence of distinct molecular subtypes of rhabdoid tumors[9,10], it will be important to determine the extent to which PGBD5-induced genome remodeling contributes to this phenotypic diversity.

In summary, PGBD5 defines a distinct class of oncogenic mutators that contribute to cell transformation due not to mutational activation but instead to their aberrant induction and chromatin targeting, thereby inducing site-specific transforming genomic rearrangements. Our data identified *PGBD5* as an endogenous human DNA transposase that is sufficient to fully transform primary immortalized human cells in the absence of chromosomal instability[36]. Given the expression of *PGBD5* in various childhood and adult solid tumors, owing to its aberrant or co-opted tissue expression, we anticipate that PGBD5 may also contribute to the pathogenesis of these cancers. Similarly, it will be important to investigate the functions of PGBD5 in normal vertebrate and mammalian development, given its ability to induce site-specific somatic genomic rearrangements in human cells. Finally, the functional requirement for cellular NHEJ DNA repair in PGBD5-induced cell transformation might facilitate the development of rational therapeutic strategies for rhabdoid and other tumors involving endogenous DNA transposases.

5 Methods

Reagents.

All reagents were obtained from Sigma-Aldrich if not otherwise specified. Synthetic oligonucleotides were obtained from Eurofins (Eurofins MWG Operon) and were purified by HPLC, as listed in Supplementary Table 5. Antibodies are listed in Supplementary Table 6.

Plasmid constructs.

Human PGBD5 cDNA (NM_024554.3) was cloned into the lentiviral vector in frame with N-terminal GFP to generate pRecLV103-GFP-PGBD5 (GeneCopoeia). pReceiver-Lv103 encoding GFP was used as a negative control in all experiments. Plasmid encoding the hyperactive *T. ni* piggyBac transposase, as originally cloned by N. Craig and colleagues[45], was obtained from System Biosciences and was cloned into pReceiver-Lv103. The plasmids pBABE-neo-largeT, pBABE-puro-H-Ras, psPAX2, and pMD2.G were obtained from Addgene. Missense GFP-PGBD5 mutants were generated through site-directed mutagenesis, according to the manufacturer's instructions (QuikChange Lightning), as previously described[30]. Doxycycline-inducible pINDUCER21 vector was a kind gift from T. Westbrook[46] and was used to generate pINDUCER21-PGBD5 through Gateway cloning, according to the manufacturer's instructions (Fisher Scientific). Lentiviral shRNA and doxycycline-inducible WWOX expression vectors were a kind gift from M. Aldaz[47]. pLKO.1 shRNA vectors targeting PGBD5 (TRCN0000138412, TRCN0000135121) and control shGFP were obtained from the RNAi Consortium (Broad Institute). The PB-EF1-IRES-NEO transposon reporter plasmid was used as previously described[30]. pBS-EF1-IRES-NEO was created by cloning the EF1-IRES-NEO cassette from PB-EF1-IRES-NEO into the pBluescript plasmid and was modified by PCR mutagenesis to replace the *T. ni piggyBac* inverted terminal repeat with the PGBD5 signal sequence (CTGGAATGCAG). All newly generated plasmids are available from Addgene (URLs).

Production and purification of anti-PGBD5 antibody.

Synthetic peptide from human PGBD5 (NM_024554.3) ELQLLSIVPGRDLQPSDSFTGPTRC was used to immunize mice (Lampire Biological Products). Hybridoma clones were screened through enzyme-linked immunosorbent assays, and hybridoma supernatants were purified with Protein A affinity chromatography to generate the 10A8-11-7-P-5 antibody used in protein blotting (Supplementary Table 6).

Lentivirus production and cell transduction.

Lentivirus production was carried out as previously described[48]. Briefly, HEK293T cells were transfected with TransIT-LT1 with a 2:1:1 ratio of the lentiviral vector and psPAX2 and pMD2.G packaging plasmids, according to the manufacturer's instructions (Mirus). Virus supernatant was collected 48 and 72 h after transfection, pooled, filtered, and stored at −80 °C. RPE and BJ cells were transduced with virus particles at a multiplicity of infection (MOI) of 5 in the presence of 8 μg/ml hexadimethrine bromide. Transduced cells were selected for 2 d with puromycin hydrochloride (RPE cells at 10 μg/ml and BJ cells at 2 μg/ml) or G418 sulfate (2 mg/ml), depending on the vector-mediated resistance. For pINDUCER21 viruses, cells were transduced at an MOI of 1 and were isolated through fluorescence-activated cell sorting (FACSAria III, BD Bioscience). For inducible expression of WWOX, RPE cells were transduced with lentiviruses encoding tetOn-advanced-WWOX and selected with G418 sulfate (2 mg/ml) for 10 d. For shRNA depletion of WWOX, cells were transduced with lentiviruses encoding pGIPZ-shWWOX or pGIPZ-shScramble control and were selected with puromycin hydrochloride (10 μg/ml) for 2 d.

Cell culture.

Low-passage RPE and BJ cells, and human tumor cell lines were obtained from the American Type Culture Collection (ATCC). *C9orf142*$^{-/-}$ RPE cells have been described previously[41]. The identity of all cell lines was verified by STR analysis (Genetica DNA Laboratories), and absence of *Mycoplasma sp.* contamination was determined with a Lonza MycoAlert system. Cell lines were cultured in 5% CO2 in a humidified atmosphere at 37 °C in Dulbecco's Modified Eagle's medium with high glucose (DMEM-HG) supplemented with 10% FBS and antibiotics (100 U/ml penicillin and 100 μg/ml streptomycin). Clonogenic assays of RPE cells were carried out in DMEM/F-12 medium. To assess the number of viable cells, cells were trypsinized, resuspended in medium and sedimented at 500g for 5 min. Cells were then resuspended in PBS, and 10 μL was mixed in a 1:1 ratio with 0.4% trypan blue (Thermo Fisher) and counted with a hemocytometer (Hausser Scientific).

Transposon reporter assay.

The transposon reporter assay was performed with the pBS-EF1-IRES-NEO vector in HEK293 cells, as described previously[30].

Quantitative RT–PCR.

RNA was isolated with an RNeasy Mini kit according to the manufacturer's instructions (Qiagen). cDNA was synthesized with a SuperScript III First-Strand Synthesis System according to the manufacturer's instructions (Invitrogen). qRT–PCR was performed with KAPA SYBR FAST PCR polymerase with 20 ng template and 200 nM primers, according to the manufacturer's instructions (Kapa Biosystems). PCR primers are listed in Supplementary Table 5. Ct values were calculated with ROX normalization in ViiA 7 software (Applied Biosystems).

Protein blotting.

To analyze protein expression by protein blotting, 1 million cells were suspended in 80 µl of lysis buffer (4% SDS, 7% glycerol, 1.25% β-mercaptoethanol, 0.2 mg/ml bromophenol blue, and 30 mM Tris-HCl, pH 6.8) and incubated at 95 °C for 10 min. Cell suspensions were lysed with a Covaris S220 adaptive focused sonicator, according to the manufacturer's instructions. Lysates were cleared by centrifugation at 16,000g for 10 min at 4 °C. Clarified lysates (30 µl) were resolved with SDS−PAGE and electroeluted on Immobilon FL PVDF membranes (Millipore). Membranes were blocked with Odyssey Blocking buffer (Li-Cor) and blotted with the antibodies listed in Supplementary Table 6. Blotted membranes were visualized on an Odyssey CLx fluorescence scanner, according to the manufacturer's instructions (Li-Cor), with goat secondary antibodies conjugated to IRDye 800CW or IRDye 680RD (Supplementary Table 6).

Flow cytometry of cleaved caspase-3.

Cells were fixed with neutral-buffered formalin for 10 min on ice, washed with PBS, resuspended in 0.1% Triton X-100 in PBS, and incubated for 15 min at room temperature. Permeabilized cells were washed twice with PBS and resuspended in 100 µl of Hank's balanced salt solution (HBSS) with 0.1% bovine serum albumin and 2 µl of Alexa Fluor 647–conjugated antibody against cleaved caspase-3 (Supplementary Table 6). Cells were incubated for 30 min at room temperature in the dark, washed twice with PBS and stained with 1 µg/ml DAPI. Cells were analyzed on a Fortessa LSR, as previously described (BD Bioscience)[49,50].

Histological staining.

Histologic processing and staining was done as previously described[51,52]. Briefly, cell lines were plated on eight-well glass Millicell EZ chamber slides at 5,000 cells/well, grown for 24 h, and fixed with 4% paraformaldehyde for 10 min at room temperature (Millipore). Tumor xenograft tissue was fixed with 4% paraformaldehyde for 24 h at room temperature. Tissues were embedded in paraffin with an ASP6025 tissue processor (Leica), sectioned at 5 µm with a RM2265 microtome (Leica), and collected on SuperfrostPlus slides (Fisher Scientific). Tissue sections were deparaffinized with EZPrep buffer (Ventana Medical Systems). Antigen retrieval was performed with Cell Conditioning 1 buffer (Ventana Medical Systems), and sections were blocked for 30 min with Background Buster solution (Innovex). Primary antibodies were applied for 5 h at 1 µg/ml (Supplementary Table 6). Secondary antibodies were applied for 60 min.

For immunohistochemistry staining, diaminobenzidine (DAB) detection was performed with a DAB detection kit according to the manufacturer's instructions (Ventana Medical Systems). Slides were counterstained with hematoxylin, and a cover slip was mounted with Permount (Fisher Scientific).

For immunofluorescence staining, the detection was performed with streptavidin-HRP D (Ventana Medical Systems) and subsequent incubation with tyramide Alexa Fluor 647, as prepared according to the manufacturer's instructions (Invitrogen). Slides were then counterstained with 5 µg/ml DAPI for 10 min, and a cover slip was mounted with Mowiol (Sigma-Aldrich).

Image acquisition.

Bright-field images were acquired on an Axio Observer microscope (Carl Zeiss Microimaging). Epifluorescence images were acquired with an EVOS FL microscope (Thermo Fisher). Slides were scanned with a Pannoramic 250 slide scanner, and images were analyzed with the Pannoramic Viewer (3DHistech).

Karyotype analysis.

Five million cells were grown for 24 h before harvesting. Cultures were treated with 0.005 µg/ml colcemid for 1 h at 37 °C, resuspended in 75 mM KCl for 10 min at 37 °C, and fixed in methanol/acetic acid (3:1). Cells were transferred onto slides, stained in 0.08 µg/ml DAPI in citric acid buffer for 3 min, and mounted in Vectashield solution

(Vector Labs). For each cell line, a minimum of 15 metaphases were counted.

Anchorage independence assay.

One million RPE and BJ cells stably transduced with lentiviral vectors were expanded in 10-cm tissue culture plates until fully confluent. At confluence, cells were microscopically inspected for the occurrence of refractile colonies within the cell monolayer. For growth in semisolid medium, one million cells were resuspended in 2 ml of medium mixed with 2 ml of Matrigel (BD Bioscience). Cell suspensions were plated in 12-well tissue culture plates (200 µl per well). Semisolid suspensions were cultured for 10 d before scoring.

Xenografts.

All mouse experiments were carried out in accordance with institutional animal protocols, as approved by the Memorial Sloan Kettering Cancer Center Institutional Animal Care and Use Committee. Ten million RPE and BJ cells were suspended in 200 µl Matrigel (BD Bioscience) and injected subcutaneously into the left flanks of 6-week-old female NOD.Cg-Prkdc(scid)Il2rg(tm1Wjl)/SzJ mice (Jackson Laboratory). Tumor growth was monitored with caliper measurements, and tumor volume was calculated with the formula 3.14159 × length × width^2/6,000. Mice were sacrificed by CO_2 asphyxiation 35 d after transplantation or when tumor size exceeded 2,000 mm^3. For secondary xenografts, primary xenografts were manually dissected and dissociated with 2 mg/ml collagenase in PBS for 30 min at 37 °C. Dissociated cell suspensions were filtered with 40-µm nylon-mesh filters and cryopreserved with 10% DMSO, 40% FBS, and 50% DMEM-HG. For doxycycline treatment of mice, animals were fed 625 doxycycline chow, which was replaced weekly (Harlan). Photographs of mice and tumors were taken with a Nikon D3100 camera (Minato). Mouse experimental sample sizes were determined to achieve 80% power to detect a five-fold difference, by using the K-sample rank test. In mouse experiments with doxycycline treatment, we used randomization to assign animals to treatment groups. Mouse tumor size measurements were performed with blinding.

Analysis of published gene expression arrays.

The R2 visualization and analysis platform (URLs) was used to reanalyze published HG-U133 Plus 2.0 microarray gene expression data from normal and tumor human tissues. The analyzed gene expression data sets are listed in Supplementary Table 7.

Flanking sequence exponential anchored (FLEA) PCR.

Transposon mapping with FLEA PCR was done as previously described[53].

Chromatin immunoprecipitation and sequencing (ChIP–seq).

ChIP was performed as previously described[54]. Briefly, cells were fixed in 1% formalin in PBS for 10 min at room temperature. Glycine (125 mM final concentration) was added to the cells, and cells were washed twice in ice-cold PBS and resuspended in SDS lysis buffer (1% SDS, 10 mM EDTA, and 50 mM Tris-HCl, pH 8.1). Lysates were sonicated with a Covaris S220 adaptive focused sonicator to obtain 100- to 500-bp chromatin fragments (Covaris). Lysates containing sheared chromatin fragments were resuspended in 0.01% SDS, 1.1% Triton X-100, 1.2 mM EDTA, 16.7 mM Tris-HCl, pH 8.1, and 167 mM NaCl. Rabbit anti-PGBD5 antibody was coupled to Protein A and Protein G Dynabeads according to the manufacturer's protocol (Thermo Fisher Scientific). Lysates and antibody-coupled beads were incubated overnight at 4 °C. Precipitates were washed sequentially with ice-cold low-salt washing solution (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris-HCl, pH 8.1, and 150 mM NaCl), high-salt washing solution (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris-HCl, pH 8.1, and 500 mM NaCl), LiCl washing solution (0.25 M LiCl, 1% IGEPAL CA-630, 1% deoxycholic acid, 1 mM EDTA, and 10 mM Tris-HCl, pH 8.1), and Tris-buffered EDTA washing solution (1 mM EDTA and 10 mM Tris-HCl, pH 8.1), then eluted in elution buffer (1% SDS and 0.1 M $NaHCO_3$). ChIP–seq libraries were generated with the NEBNext ChIP–seq Library Prep kit according to the manufacturer's protocol (New England BioLabs). Libraries were sequenced on Illumina HiSeq 2500 instruments, with 30 million 2 × 50-bp paired-end reads.

ChIP–seq analysis.

Reads were trimmed for both quality and adaptor sequences, and paired reads were removed if either read length became <20 nt. Bowtie2 (v2.2.2) with default parameters was used to align the reads to the human reference assembly hg19, and PCR and optical duplicates were removed with Picard (URLs). Genomic segments enriched for ChIP over input signal were classified with MACS (v1.4) with the default parameters, and genomic 'blacklisted' regions were subsequently filtered (URLs). Signals in enriched regions were then normalized to segment length and sequencing depth.

Whole-genome DNA sequencing.

Genomic DNA was extracted with a PureLink Genomic DNA Mini Kit according to the manufacturer's instructions (Thermo Fisher Scientific). Genome sequencing libraries were constructed with a TruSeq Nano library kit, according to the manufacturer's protocol (Illumina). Genomes were sequenced on Illumina HiSeq X instruments, with 2 × 150-bp paired-end reads. For analysis of primary patient rhabdoid tumor genomes, sequencing files were downloaded from the TARGET Data Matrix, as previously described[10]. Reads were aligned to the GRCh37 human reference with the Burrows–Wheeler Aligner (BWA aln and BWA MEM for GATK and laSV analyses, respectively) and processed with the best-practices pipeline, which included marking of duplicate reads with Picard tools, realignment around indels, and base recalibration via Genome Analysis Toolkit (GATK) (ver. 3.2.2)[55,56].

Alignment-based mutational and structural variant analysis.

MuTect (v1.1.4)[57], LoFreq (v2.0.0)[58] (single-nucleotide variants (SNVs) only), Strelka (v1.0.13)[59] (both SNVs and indels), Pindel (v0.2.5), and Scalpel (v0.4) (indels only) were used with the default filtering criteria, as implemented in each of the programs. Triallelic SNVs and common germline variants (>1% MAF in 1000 Genomes Project release 3 or the Exome Aggregation Consortium server (URLs)), and a blacklist of recurrent artifactual calls seen in HapMap samples sequenced and analyzed with the same methodology, were filtered out. The union of all SNV and indel calls was annotated with snpEff, snpSift[60] and GATK VariantAnnotator according to the annotations from ENSEMBL, COSMIC, 1000 Genomes Project, and ExAC[61,62]. Copy-number variants (CNVs) were detected with BIC-seq2 (ref. 63). DELLY (v0.6.1)[64], CREST (v1.0)[65], and BreakDancer (v1.4.0)[66] were used to detect structural variants (SVs). Bedtools pairtopair[67] was used to merge structural variants. Germline variants from the 1000 Genomes Project call set, Database of Genomic Variants and a blacklist of SVs seen in HapMap genomes were filtered out. SplazerS was used for the analysis of split reads[68], and SV breakpoints were annotated with coinciding BIC-seq2 CNV changepoints. SVs with split-read support (tumor only), with at least one coinciding (within 500 bp) CNV changepoint called by two or more tools or called by CREST, are marked as higher confidence. The annotation with gene overlap (RefSeq, Cancer Gene Census), including prediction of potential effects on genes (for example, disruptive/exonic, intronic, and intergenic) and with annotated transposons, was done with bedtools[67].

laSV.

De novo assembly-based laSV[33] was used with the following parameters: -s 15 -k 63 -p 3. Structural variants supported by fewer than four reads or with allele frequencies below 10% were filtered. Variant recurrence was measured in 100-kb bins with bedtools[67]. Circos plots were generated with Circos (version 0.67-4)[69].

SMuFin.

SMuFin was used with default parameters, as previously described[34]. SMuFin results included, SNVs as well as small (indels) and SVs. Large SVs were defined as SVs identified with a single breakpoint, for which the SV length exceeded the length of the underlying variant block called by SMuFin. Breakpoints supported by fewer than four reads were filtered. SV size was estimated on the basis of the assumption that SVs were caused by single genomic events.

Regulatory element analysis.

Annotated regulatory elements were compiled from both ENCODE and NIH Roadmap Epigenomics Consortium (URLs). The analysis focused on distal DNase I–hypersensitivity sites, because distal sites have been shown to vary in a cell-type-specific manner, and DNase I sensitivity covers both active and poised regulatory elements. Cancer cell line data sets were removed, and the overlap of at least 1 bp was calculated between breakpoints and DNase I–hypersensitivity peaks in each cell type. To account for cell types with variable DNase I–hypersensitive sites, the overlap count for each cell type was normalized to the total number of regulatory sites in that cell type.

PGBD5 signal sequence (PSS) analysis.

The position weight matrix (PWM) for the PSS and RSS were generated as previously described[32]. These PWMs were used to scan sequences around variant breakpoints (± 50 bp) for both PSS and RSS with the sequence-motif-matching algorithm FIMO[70]. Additionally, PGBD5 signal-sequence motifs associated with structural variants were detected by analysis of 20-bp windows around variant breakpoints with MEME with default parameters[71]. Matches with a false discovery rate <0.1 and within 15 bp from the variant breakpoints were retained and counted. All variants associated with PSS motifs were manually verified. To construct the position-scrambled PSS, the perl rand function was used to generate ten independent position-scrambled PWMs.

Statistical analysis.

All experiments were performed a minimum of three times with a minimum of three independent measurements. For comparisons between two sample sets, statistical analysis of means was performed with two-tailed unpaired Student’s t-tests. Survival analysis was done with the Kaplan–Meier method, as assessed with a log-rank test. For gene expression analysis, statistical significance was assessed with paired t-tests. False discovery was assessed at the 0.05 level with the step-down Dunnett method, as extended to general parametric models[72,73]. The significance of sequence-motif enrichment was assessed with hypergeometric tests. For significance analysis of association of structural variants with regulatory elements, Welch's t-test was used. Calculations were performed with R statistical computing software[74].

Code availability.

Scripts used in this analysis are openly available at github (URLs).

Data availability.

Genome and chromatin immunoprecipitation sequencing data have been deposited in the NCBI Sequence Read Archive and Gene Expression Omnibus databases (Bioproject 320056 and Data Set GSE81160, respectively). Analyzed data are openly available at the Zenodo digital repository (http://dx.doi.org.sire.ub.edu/10.5281/zenodo.50633), as summarized in Supplementary Table 8.

URLs.

Plasmids available from Addgene, https://www.addgene.org/Alex_Kentsis/; annotated regulatory elements, http://www.encodeproject.org/data/annotations/v2/; R2 visualization and analysis platform, http://r2.amc.nl/; Picard, https://broadinstitute.github.io/picard/; hg19 blacklist, http://www.broadinstitute.org/~anshul/projects/encode/rawdata/blacklists/hg19-blacklist-README.pdf; Exome Aggregation Consortium server, http://exac.broadinstitute.org/; scripts used in this work, https://github.com/kentsisresearchgroup/Rhabdoid_PGBD5_MSK_paper/.

6 References

1. Vogelstein, B. et al. Cancer genome landscapes. Science 339, 1546–1558 (2013).
2. Alexandrov, L.B. et al. Signatures of mutational processes in human cancer. Nature 500, 415–421 (2013).
3. Weinstein, J.N. et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet. 45, 1113–1120 (2013).

4. Futreal, P.A. et al. A census of human cancer genes. Nat. Rev. Cancer 4, 177–183 (2004).
5. Huether, R. et al. The landscape of somatic mutations in epigenetic regulators across 1,000 paediatric cancer genomes. Nat. Commun. 5, 3630 (2014).
6. Northcott, P.A. et al. Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. Nature 511, 428–434 (2014).
7. Mansour, M.R. et al. Oncogene regulation: an oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. Science 346, 1373–1377 (2014).
8. Molenaar, J.J. et al. Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes. Nature 483, 589–593 (2012).
9. Johann, P.D. et al. Atypical teratoid/rhabdoid tumors are comprised of three epigenetic subgroups with distinct enhancer landscapes. Cancer Cell 29, 379–393 (2016).
10. Chun, H.J. et al. Genome-wide profiles of extra-cranial malignant rhabdoid tumors reveal heterogeneity and dysregulated developmental pathways. Cancer Cell 29, 394–406 (2016).
11. Jones, D.T. et al. Dissecting the genomic complexity underlying medulloblastoma. Nature 488, 100–105 (2012).
12. Fischer, H.P., Thomsen, H., Altmannsberger, M. & Bertram, U. Malignant rhabdoid tumour of the kidney expressing neurofilament proteins: immunohistochemical findings and histogenetic aspects. Pathol. Res. Pract. 184, 541–547 (1989).
13. Lee, R.S. et al. A remarkably simple genome underlies highly malignant pediatric rhabdoid cancers. J. Clin. Invest. 122, 2983–2988 (2012).
14. van den Heuvel-Eibrink, M.M. et al. Malignant rhabdoid tumours of the kidney (MRTKs), registered on recent SIOP protocols from 1993 to 2005: a report of the SIOP renal tumour study group. Pediatr. Blood Cancer 56, 733–737 (2011).
15. Versteege, I. et al. Truncating mutations of hSNF5/INI1 in aggressive paediatric cancer. Nature 394, 203–206 (1998).
16. Roberts, C.W., Leroux, M.M., Fleming, M.D. & Orkin, S.H. Highly penetrant, rapid tumorigenesis through conditional inversion of the tumor suppressor gene Snf5. Cancer Cell 2, 415–425 (2002).
17. Rousseau-Merck, M.F., Fiette, L., Klochendler-Yeivin, A., Delattre, O. & Aurias, A. Chromosome mechanisms and INI1 inactivation in human and mouse rhabdoid tumors. Cancer Genet. Cytogenet. 157, 127–133 (2005).
18. Takita, J. et al. Genome-wide approach to identify second gene targets for malignant rhabdoid tumors using high-density oligonucleotide microarrays. Cancer Sci. 105, 258–264 (2014).
19. Smit, A.F. Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr. Opin. Genet. Dev. 9, 657–663 (1999).
20. Kazazian, H.H. Jr. Mobile elements: drivers of genome evolution. Science 303, 1626–1632 (2004).
21. Rodic´, N. et al. Retrotransposon insertions in the clonal evolution of pancreatic ductal adenocarcinoma. Nat. Med. 21, 1060–1064 (2015).
22. Muotri, A.R. et al. Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature 435, 903–910 (2005).
23. Shaheen, M., Williamson, E., Nickoloff, J., Lee, S.H. & Hromas, R. Metnase/ SETMAR: a domesticated primate transposase that enhances DNA repair, replication, and decatenation. Genetica 138, 559–566 (2010).
24. Hiom, K., Melek, M. & Gellert, M. DNA transposition by the RAG1 and RAG2 proteins:a possible source of oncogenic translocations. Cell 94, 463–470 (1998).
25. Navarro, J.M. et al. Site- and allele-specific polycomb dysregulation in T-cell leukaemia. Nat. Commun. 6, 6094 (2015).
26. Papaemmanuil, E. et al. RAG-mediated recombination is the predominant driver of oncogenic rearrangement in ETV6-RUNX1 acute lymphoblastic leukemia. Nat. Genet. 46, 116–125 (2014).

27. Halper-Stromberg, E. et al. Fine mapping of V(D)J recombinase mediated rearrangements in human lymphoid malignancies. BMC Genomics 14, 565 (2013).
28. Dreyer, W.J., Gray, W.R. & Hood, L. The genetics, molecular, and cellular basis of antibody formation: some facts and a unifying hypothesis. Cold Spring Harb. Symp Quant. Biol. 32, 353–367 (1967).
29. Majumdar, S., Singh, A. & Rio, D.C. The human THAP9 gene encodes an active P-element DNA transposase. Science 339, 446–448 (2013).
30. Henssen, A.G. et al. Genomic DNA transposition induced by human PGBD5. eLife 4, e10565 (2015).
31. Pavelitz, T., Gray, L.T., Padilla, S.L., Bailey, A.D. & Weiner, A.M. PGBD5: a neuralspecific intron-containing piggyBac transposase domesticated over 500 million years ago and conserved from cephalochordates to humans. Mob. DNA 4, 23 (2013).
32. Henssen, A.G. et al. Forward genetic screen of human transposase genomic rearrangements. BMC Genomics 17, 548 (2016).
33. Zhuang, J. & Weng, Z. Local sequence assembly reveals a high-resolution profile of somatic structural variations in 97 cancer genomes. Nucleic Acids Res. 43, 8146–8156 (2015).
34. Moncunill, V. et al. Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads. Nat. Biotechnol. 32, 1106–1112 (2014).
35. Bralten, L.B. et al. The CASPR2 cell adhesion molecule functions as a tumor suppressor gene in glioma. Oncogene 29, 6138–6148 (2010).
36. Hahn, W.C. et al. Creation of human tumour cells with defined genetic elements. Nature 400, 464–468 (1999).
37. Landree, M.A., Wibbenmeyer, J.A. & Roth, D.B. Mutational analysis of RAG1 and RAG2 identifies three catalytic amino acids in RAG1 critical for both cleavage steps of V(D)J recombination. Genes Dev. 13, 3059–3069 (1999).
38. Lu, C.P., Posey, J.E. & Roth, D.B. Understanding how the V(D)J recombinase catalyzes transesterification: distinctions between DNA cleavage and transposition Nucleic Acids Res. 36, 2864–2873 (2008).
39. Gellert, M. V(D)J recombination: RAG proteins, repair factors, and regulation. Annu. Rev. Biochem. 71, 101–132 (2002).
40. Mitra, R., Fain-Thornton, J. & Craig, N.L. piggyBac can bypass DNA synthesis during cut and paste transposition. EMBO J. 27, 1097–1109 (2008).
41. Ochi, T. et al. DNA repair. PAXX, a paralog of XRCC4 and XLF, interacts with Ku to promote DNA double-strand break repair. Science 347, 185–188 (2015).
42. Aldaz, C.M., Ferguson, B.W. & Abba, M.C. WWOX at the crossroads of cancer, metabolic syndrome related traits and CNS pathologies. Biochim. Biophys. Acta 1846, 188–200 (2014).
43. McClintock, B. The origin and behavior of mutable loci in maize. Proc. Natl. Acad Sci. USA 36, 344–355 (1950).
44. Torchia, J. et al. Integrated (epi)-genomic analyses identify subgroup-specific therapeutic targets in CNS rhabdoid tumors. Cancer Cell 30, 891–908 (2016).
45. Li, X. et al. piggyBac transposase tools for genome engineering. Proc. Natl. Acad.Sci. USA 110, E2279–E2287 (2013).
46. Westbrook, T.F., Stegmeier, F. & Elledge, S.J. Dissecting cancer pathways and vulnerabilities with RNAi. Cold Spring Harb. Symp. Quant. Biol. 70, 435–444 (2005).
47. Ferguson, B.W. et al. The cancer gene WWOX behaves as an inhibitor of SMAD3 transcriptional activity via direct binding. BMC Cancer 13, 593 (2013).
48. Kentsis, A. et al. Autocrine activation of the MET receptor tyrosine kinase in acute myeloid leukemia. Nat. Med. 18, 1118–1122 (2012).
49. Fox, R. & Aubert, M. Flow cytometric detection of activated caspase-3. Methods Mol. Biol. 414, 47–56 (2008).

50. Sordet, O. et al. Specific involvement of caspases in the differentiation of monocytes into macrophages. Blood 100, 4446–4453 (2002).
51. Yarilin, D. et al. Machine-based method for multiplex in situ molecular characterization of tissues by immunofluorescence detection. Sci. Rep. 5, 9534 (2015).
52. Fujisawa, S., Turkekul, M., Barlas, A., Fan, N. & Manova, K. Double in situ detection of sonic hedgehog mRNA and pMAPK protein in examining the cell proliferation signaling pathway in mouse embryo. Methods Mol. Biol. 717, 257–276 (2011).
53. Henssen, A., Carson, J.R. & Kentsis, A. Transposon mapping using flanking sequence exponential anchored (FLEA) PCR. Protocol Exchange http://dx.doi.org/10.1038/protex.2015.071 (2015).
54. Krivtsov, A.V. et al. H3K79 methylation profiles define murine and human MLL-AF4 leukemias. Cancer Cell 14, 355–368 (2008).
55. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760 (2009).
56. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303 (2010).
57. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol. 31, 213–219 (2013).
58. Wilm, A. et al. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets.Nucleic Acids Res. 40, 11189–11201 (2012).
59. Saunders, C.T. et al. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics 28, 1811–1817 (2012).
60. Ye, K., Schulz, M.H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25, 2865–2871 (2009).
61. Narzisi, G. et al. Accurate de novo and transmitted indel detection in exome-capture data using microassembly. Nat. Methods 11, 1033–1036 (2014).
62. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6, 80–92 (2012).
63. Xi, R. et al. Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. Proc. Natl. Acad. Sci. USA 108, E1128–E1136 (2011).
64. Rausch, T. et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 28, i333–i339 (2012).
65. Wang, J. et al. CREST maps somatic structural variation in cancer genomes with base-pair resolution. Nat. Methods 8, 652–654 (2011).
66. Chen, K. et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat. Methods 6, 677–681 (2009).
67. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842 (2010).
68. Emde, A.K. et al. Detecting genomic indel variants with exact breakpoints in singleand paired-end sequencing data using SplazerS. Bioinformatics 28, 619–627 (2012).
69. Krzywinski, M. et al. Circos: an information aesthetic for comparative genomics Genome Res. 19, 1639–1645 (2009).
70. Grant, C.E., Bailey, T.L. & Noble, W.S. FIMO: scanning for occurrences of a given motif. Bioinformatics 27, 1017–1018 (2011).
71. Bailey, T.L. et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 37, W202–W208 (2009).

72. Hothorn, T., Bretz, F. & Westfall, P. Simultaneous inference in general parametric models. Biom. J. 50, 346–363 (2008).
73. Dunnett, C.W. & Tamhane, A.C. Step-down multiple tests for comparing treatments with a control in unbalanced one-way layouts. Stat. Med. 10, 939–947 (1991).
74. R Development Core Team R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, 2013).

7 Authors information

Affiliations

[1]Molecular Pharmacology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, New York, USA. [2]Cancer Biology & Genetics Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, New York, USA. [3]Program in Bioinformatics and Integrative Biology, Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, Massachusetts, USA. [4]Joint BSC-CRG-IRB Research Program in Computational Biology, Barcelona Supercomputing Center (BSC-CNS), Barcelona, Spain. [5]The Wellcome Trust/Cancer Research UK Gurdon Institute, University of Cambridge, Cambridge, UK. [6]Institute for Computational Biomedicine, Weill Cornell Medical College, New York, New York, USA. [7]Antitumor Assessment Core Facility, Memorial Sloan Kettering Cancer Center, New York, New York, USA. [8]Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, New York, USA. [9]New York Genome Center, New York, New York, USA. [10]Bioinformatics Core, Memorial Sloan Kettering Cancer Center, New York, New York, USA. [11]Northwestern University Feinberg School of Medicine, Ann & Robert H. Lurie Children's Hospital of Chicago, Chicago, Illinois, USA. [12]Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, New York, USA. [13]Department of Oncology, St. Jude Children's Research Hospital, Memphis, Tennessee, USA. [14]Department of Pathology, Boston Children's Hospital, Boston, Massachusetts, USA. [15]Department of Pediatric Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [16]Department of Biochemistry, University of Cambridge, Cambridge, UK. 17The Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK. [18]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain. [19]Weill Cornell Medical College, Cornell University, New York, New York, USA. [20]Department of Pediatrics, Memorial Sloan Kettering Cancer Center, New York, New York, USA. [21]Present address: Department of Pediatric Oncology/Hematology, Charité-Universitätsmedizin Berlin, Berlin, Germany. [22]These authors contributed equally to this work. Correspondence should be addressed to A.K. (kentsisresearchgroup@gmail.com).

Contributions

A.G.H. contributed to the study design and collection and interpretation of the data. R.K. performed ChIP–seq, whole-genome sequencing, and FLEA PCR data analysis. J.Z. analyzed tumor genome-sequencing data with laSV. E.J., C. Reed, A.E., and E.S. performed in vitro transformation assays and vector design and cloning. I.C.M. performed experiments and analyzed data. E.R.-F., S.G., M.P., C.E.M., A.-K.E., M.S., K.A., C. Reeves, N.D.S., D.T., and Z.W. analyzed genome-sequencing data. A.N.B. and S.P.J. contributed to creation of PAXX-deficient cells and study design. E.d.S. contributed to mouse-xenograft study design. M.G. performed statistical analysis of data sets. C.R.A. performed histological analysis of tumor samples. E.P., C.W.M.R., H.S., E.M., and S.A.A. contributed to study design. A.K. contributed to study design, data analysis, and interpretation. A.K. and A.G.H. wrote the manuscript, to which all authors contributed.

8 Supplementary information

Supplementary figures in the appendix section.

PGBD5
PCAWG analysis

Not published

Characterization of recurrent deletions promoted by transposase-derived genes across different human tumor types

Authors

Elias Rodriguez-Fos, Santi González, Montserrat Puiggròs, Anton G. Henssen, Alex Kentsis, David Torrents.

Background/Motivation

The analysis I present below is the continuation of the work I have previously described, where we identified a pattern of structural rearrangements associated with the expression of *PGBD5*. Based on these findings, we continued our study in ICGC-PCAWG[1] data, intending to explore the presence of similar rearrangements in different tumor types. This study is currently on hiatus.

1 Abstract

Through the analysis of whole-genome sequencing data from cells expressing *PGBD5* —a known transposase-derived gene—, we identified significant enrichment of small deletions with a specific motif around the breakpoints (see PGBD5 publication section, p. 78). Interestingly, we also found a recurrent pattern of confronted *Alu* sequences flanking the deletion, which could be acting as a substrate in mechanisms triggered by this gene.

Following this line, we have expanded the same type of analysis to different tumor and cancer types. We have analyzed the landscape of somatic deletions within 2,706 WGS tumor samples across 37 tumor types from the ICGC-Pan-

Cancer cohort. Based on this analysis, we defined the presence of 5,331 motif-related deletions associated with PGBD5 specific motif, and 3,568 other deletions associated with different motifs sharing the same characteristics: microhomology around the breakpoints, small average size, and enrichment in repetitive elements flanking the deletion. The enrichment in repetitive elements and the presence of microhomology around the breakpoints suggest the potential role of an homology-mediated mechanism in the generation of these deletions. Interestingly, from all 8,899 motif-related deletions, we observed 90.3% of recurrence across patients and tumor types.

Though, when manually curating a subset of motif-related deletions to confirm the high recurrence associated with these somatic rearrangements, we found high fluctuation in the number of reads supporting the deletions in the tumor and normal samples, making its classification difficult. This new information added uncertainty in the characterization of these recurrent structural variants.

2 Introduction

DNA transposases are enzymes that recognize and catalyze the movement of mobile elements in the human genome known as transposons. There are abundant transposase-derived genes in the genome that have been conserved through evolution, and some of them maintain their enzymatic activity in human cells. As presented in the general introduction of this thesis, this is the case of *PGBD5*[2], a transposase-derived gene that belongs to the subfamily of *PiggyBac* transposable elements. The main characteristic of the transposases from the *PiggyBac* family is that they efficiently mobilize DNA in the genome via a "*cut and paste*" mechanism[3].

A recent study published[4] in Nature Genetics in 2017 in which we have participated, link the expression of *PGBD5* with the generation of genomic rearrangements in human cancer, notably in rhabdoid tumors. This study defines PGBD5 as an oncogenic mutator associated with rearrangements involving PGBD5-specific sequences at their breakpoints. Focusing on the analysis of *PGBD5*-transformed cells, included in the study, we identified significant enrichment of small deletions, with sizes around 183bp, presenting the same PGBD5-specific motif around the breakpoints, found in rhabdoid tumors. As additional characteristics of PGBD5-motif-related deletions, we also identified a recurrent pattern of confronted *Alu* sequences flanking this type of deletions.

There is still a lack of information on the characteristics of the rearrangements and mechanisms involved in the transformation of normal cells driven by the expression of transposase-derived genes. Nevertheless, transposase-derived genes are known to be expressed in different cancer types[4,5]. Following this idea, we hypothesize that the expression of *PGBD5* and other transposase-derived genes can generate specific genomic rearrangements associated with the presence of motif sequences around the breakpoints and have potential mutagenic activity in different types of cancer. Thanks to the accessibility of our group to ICGC-Pan-Cancer[1] data, we were able to analyze the landscape of recurrent genome modifications with specific characteristics such as the ones described above, expanding our analysis from the previous PGBD5 study to 2,706 cancer patients.

3 Results

3.1 PGBD5-motif-related deletions are present in different cancer types conserving its characteristic features

In our previous study, we characterized PGBD5-motif-related deletions as a recurrent type of small somatic deletions presenting a conserved motif sequence CACTGCA/TGCAGTG around the breakpoints. Interestingly, all the deletions found with this motif flanking the breakpoints registered the reconstruction of the exact same 7bp sequence around the breakpoint junction (Fig. 1, Supplementary Fig. 1a). Based on this feature, we analyzed 2,706 whole-genome sequencing matched normal and tumor samples from 47 different projects encompassing 37 different cancer types included in the ICGC-PCAWG study using SMuFin[6] (Somatic Mutation Finder), the variant caller developed in our group employed in the previously exposed PGBD5 publication. The execution of the variant calling steps has been performed with great support from Montserrat Puiggròs.

Focusing on the results from SMuFin, we found 5,331 somatic deletions —5,310 detected as structural variants; 21 detected as indels— that present the PGBD5 associated CACTGCA/TGCAGTG motif flanking the breakpoints in 806 patients from 41 different cancer projects. We detected this specific type of deletions in 87.2% of the cancer projects, and 29.8% of the patients included in this study, with an irregular distribution of these variants across the different studies and cancer types. These results indicate a higher prevalence of PGBD5-motif-related deletions than previously expected, showing the presence of these variants in the majority of cancer types from the PCAWG study, with marked examples in liver and ovarian cancers (Fig. 2), indicating that these rearrangements are not restricted to rhabdoid tumors.

Following the study of somatic PGBD5-motif-related deletions in Pan-Cancer, we determined their median size as 294bp (min.=12bp, first quartile=190bp, third quartile=401bp, max.=199,272,244bp), which is consistent with our previous findings (Supplementary Fig. 2). Moreover, we identified enrichment in the presence of repetitive elements flanking the deleted region, notably in *Alus*, which are present in 99.3% of these rearrangements. These elements are reconstructed after the loss of DNA, resulting in chimeric *Alus* (Fig. 1). This analysis supports not only the features already described in our previous publication but also the hypothesis of a potential role of repetitive elements, *Alus* particularly, in the generation of these deletions[7,8].

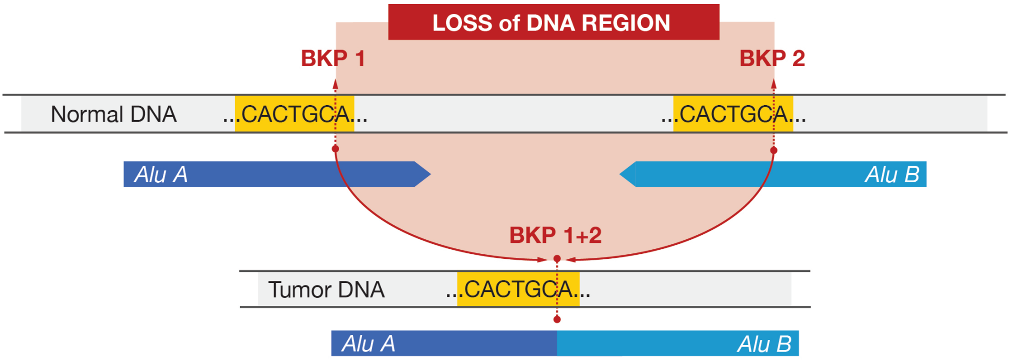


Figure 1. PGBD5-motif-related deletion mechanism.

Schema of the potential mechanism for the generation of PGBD5-motif-related deletions. Motifs and *Alu* elements flanking the breakpoints are reconstructed after the deletion obtaining a chimeric *Alu*.

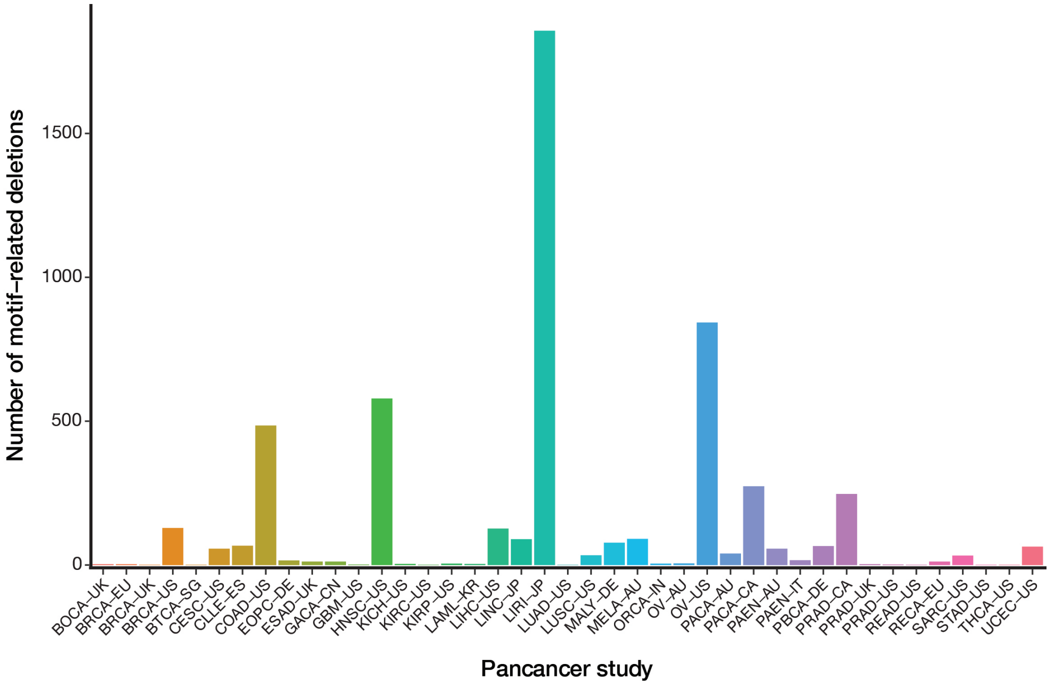


Figure 2. Distribution of PGBD5-motif related deletions across the different Pan-Cancer studies.

Number of PGBD5-motif-related deletions detected in each of the 41 ICGC-PCAWG studies presenting this type of deletions. Six studies are not represented due to their lack of motif-related deletions. The two studies showing more deletions related to the PGBD5 motif are LIRI-JP (liver cancer from Japan) and OV-US (ovarian cancer from the USA).

3.2 PGBD5-motif-related deletions present high recurrence across different patients and cancer types

Through the exploration of the genomic regions affected by PGBD5-motif-related deletions, we discovered high levels of recurrence among them. Surprisingly, 4,824 deletions, related to PGBD5 motif, showed the same exact breakpoint positions in more than one tumor sample, evidencing that 90.5% of this type of deletions are recurrent (Fig. 3a). In other words, we found exactly the same PGBD5-motif-related deletions with the same genomic positions across different patients and different types of cancer. From a total of 5,331 detected deletions presenting the CACTGCA/TGCAGTG motif, 507 are unique, and 810 are present in more than one patient. The degrees of recurrence exhibited by these specific deletions go from two —detected in two different patients— to 57 with an average recurrence level of 5.96 different patients sharing the same exact rearrangement.

Looking deeper into the regions of the genome that present PGBD5-motif-related deletions, we observed regions recurrently mutated in all the human chromosomes (Fig. 3b). However, some genomic loci accumulate more recurrent mutations than others. This is the case of chr2:160559499-160559649, chr4: 85453295-85453392, chr8:48662211-48662312, chr12:56589021-56589115, and chr19:3003601-3003774 loci, which appear to be hotspots for this type of deletions, being found mutated in more than 30 different patients from different cancer types. Exploring these high recurrent mutations, we observed the presence of *Alus* flanking all the affected regions. Nevertheless, looking for recurrently affected genes, we have not seen a significant enrichment of recurrent deletions in coding regions of the genome.

Our results show high degrees of recurrence of small deletions across different tumor genomes, which is surprising given their somatic nature. This raises the following questions: Have the rest of the deletions —no-PGBD5-motif-related— the same proportion of recurrence? Do the high degrees of recurrence exhibited by these deletions correlate with the presence of homologous sequences in the breakpoints?

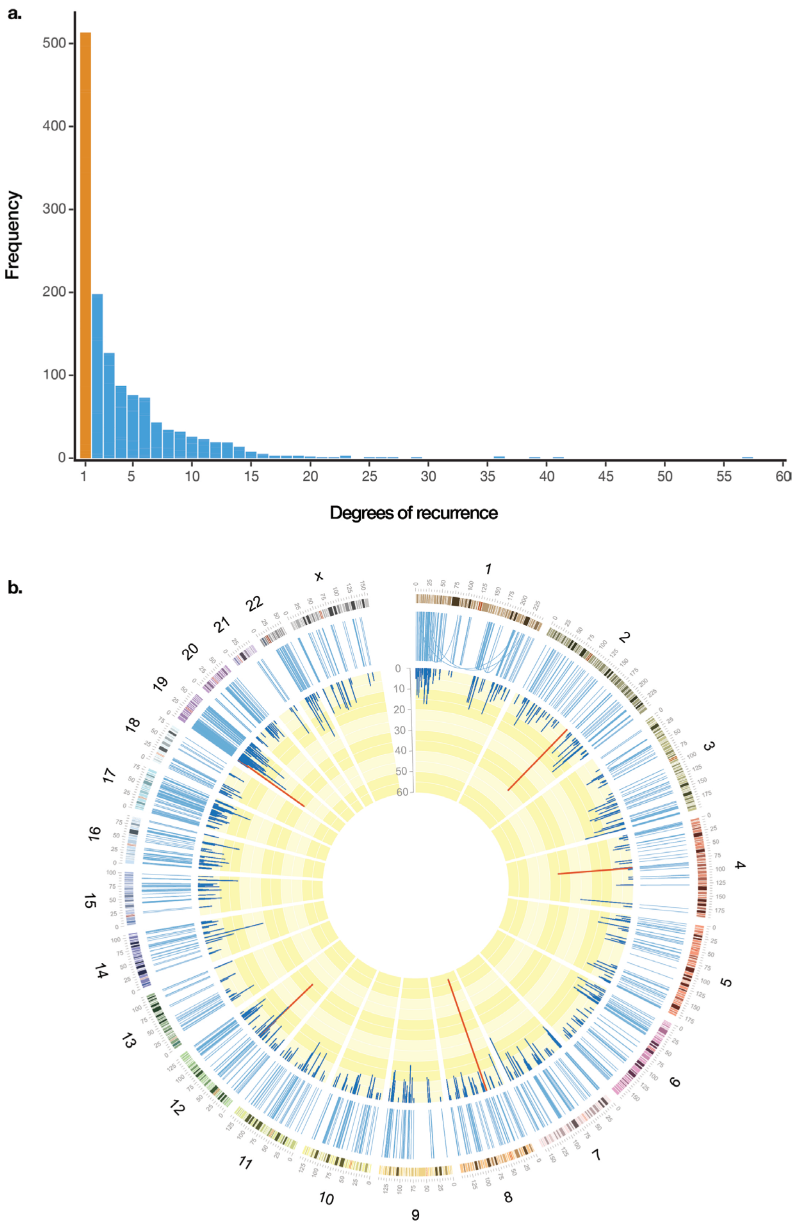


Figure 3. Recurrence levels of PGBD5-motif related deletions in the human genome.

(**a**), Frequency of PGBD5-motif related deletions by its degree of recurrence. In orange, unique deletions occurring in one patient. In blue, recurrent deletions sharing the same genomic positions, present in at least two different patients. (**b**), Circos plot showing the genomic position of each PGBD5-motif related deletion detected in ICGC-PCAWG (light blue) and their associated degrees of recurrence (dark blue). Deletions present in more than 30 patients are highlighted in red (chr2:160559499-160559649, chr4: 85453295-85453392, chr8:48662211-48662312, chr12:56589021-56589115, and chr19:3003601-3003774).

3.3 A fraction of deletions not presenting the PGBD5 motif show recurrence and different homologous sequences around the breakpoints

As introduced earlier, 5,331 PGBD5-motif-related deletions were detected in the structural variant analysis from PCAWG data, corresponding to 2.84% of the total number of deletions called. In order to evaluate the recurrence of PGBD5 and no-PGBD5-motif-related deletions, we explored the 182,043 remaining deletions that do not present the CACTGCA/TGCAGTG motif flanking the DNA loss. In this case, we found that 16.3% of no-PGBD5-motif-related deletions are recurrent across different patients and types of cancer, while 90,5% of deletions associated with the PGBD5 motif present this type of recurrence (Fig. 4). Our results indicate that PGBD5-motif-related deletions are significantly (p-value<2.2e-16) more recurrent than the rest of the deletions we detect in PCAWG data.

Still, we did not expect a 16.3% of the somatic deletions not associated with the PGBD5 motif to be present with the same exact breakpoints in different patients. For this reason, we expanded our analysis to this new set of recurrent mutations in order to find if they share some common characteristics with PGBD5-motif-related rearrangements, starting with the presence of a conserved sequence or motif around the breakpoints.

Performing the motif discovery analysis in the recurrent no-PGBD5-motif-related deletions, we significantly determined six highly conserved motifs with similar length as the one associated with *PGBD5* expression: TC(C,T)CAGC, CACTTTGGGA, CAGGTGG, GCCTG(G,T)(C,A), GC(C,T)TCCCA, TAGCTGGG (Supplementary Fig. 1b-g). Our results hence indicate that recurrent no-PGBD5-motif-related deletions also present conserved motifs around their breakpoints.

In summary, we detected 8,899 recurrent and non-recurrent deletions displaying seven specific sequences in their breakpoints, including the motif associated with PGBD5. Moreover, we observed that 90.3% of all the identified somatic motif-

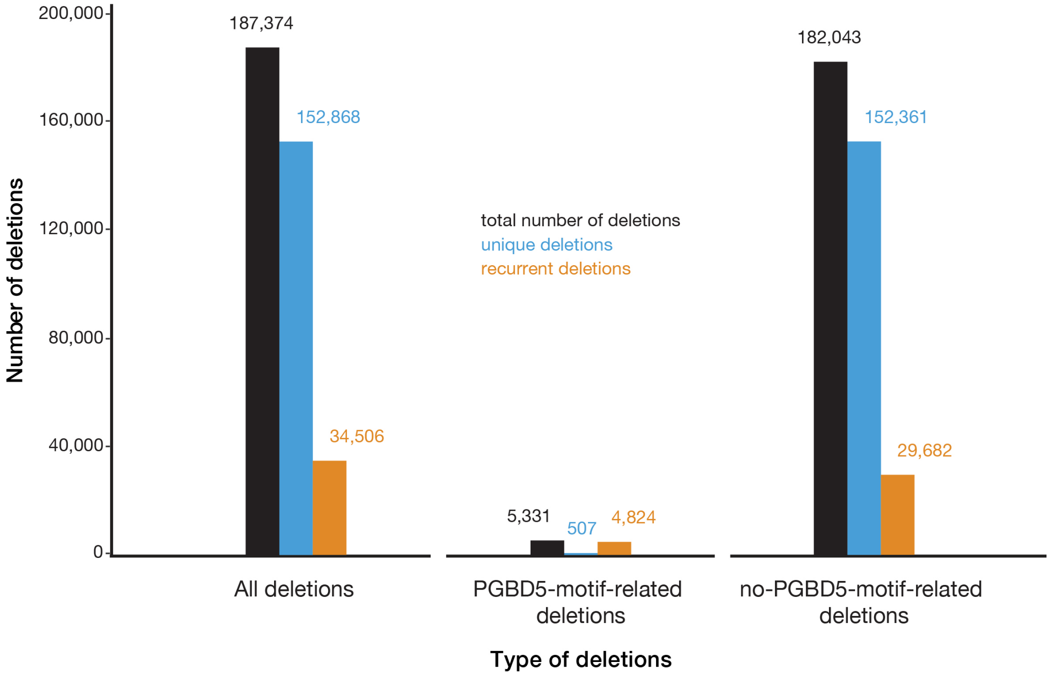


Figure 4. Frequency of recurrent and no-recurrent deletions.

Comparison between the number of recurrent and no-recurrent deletions in all deletions detected in Pan-Cancer, PGBD5-motif-related deletions, and no-PGBD5-motif-related deletions. In black, the total number of deletions. In blue, the number of unique deletions. In orange, the number of recurrent deletions. The proportion of recurrent deletions in the PGBD5-motif-related group is higher.

related deletions —encompassing all seven motifs— are recurrent across patients and tumor types and present an enrichment in repetitive elements flanking the breakpoints, principally *Alus* (94.6%), supporting our previous observations (Fig. 5). Interestingly, our results suggest a correlation between the presence of *Alus* and the recurrence of motif-related deletions.

At this point, we had characterized a novel set of small deletions, corresponding to 4.75% of all the somatic deletions detected in PCAWG data using SMuFin. As already defined, these rearrangements present seven conserved motifs around its breakpoints along with *Alus* in inverted orientation[9] flanking the deleted region. The described *Alu* configuration has been reported to act as a substrate for frequent recombination in the human genome, generating deletions and other genomic rearrangements. Therefore, the presence of homologous elements —*Alus*— and microhomologous sequences around the breakpoints point to a major role of a homology-related mechanism driving the generation of these deletions. These findings suggest the potential activity of mechanisms such as MMEJ[10], which has already been described to generate chimeric *Alus*.

However, some facts raised certain uncertainty about our conclusions. First of all, the high recurrence levels across different patients and cancer types were totally unexpected. Second of all, recurrent motif-related deletions have an average size bigger than the read length and smaller than the library size and correlate with repetitive sequences of the genome, increasing the difficulty for their accurate detection[11,12]. Nevertheless, it is known that the presence of *Alus* in close proximity promotes[13] their recombination, resulting in structural variants. Additionally, this mechanism has been associated with the expression of genes such as *p53*[13]. Following this idea, could a similar mechanism promoted by the expression of transposase-derived genes generate the exact same somatic rearrangements in more than 30 different patients, for example?

3.4 Manual curation of a subset of recurrent motif-related deletions revealed fluctuation in the number of reads supporting the rearrangements

In order to evaluate the high degree of recurrence observed in somatic motif-related deletions, we first explored the results from the official Pan-Cancer somatic variant calling pipelines[1,11]. Surprisingly, while the first published set of somatic variants from independent callers supported our results showing the presence of 2,067 PGBD5-motif-related deletions of which 86.36% were recurrent across different patients, the official consensus set of variants, which was the one to be used, filtered out the type of variants we were studying according to their size (Fig. 6).

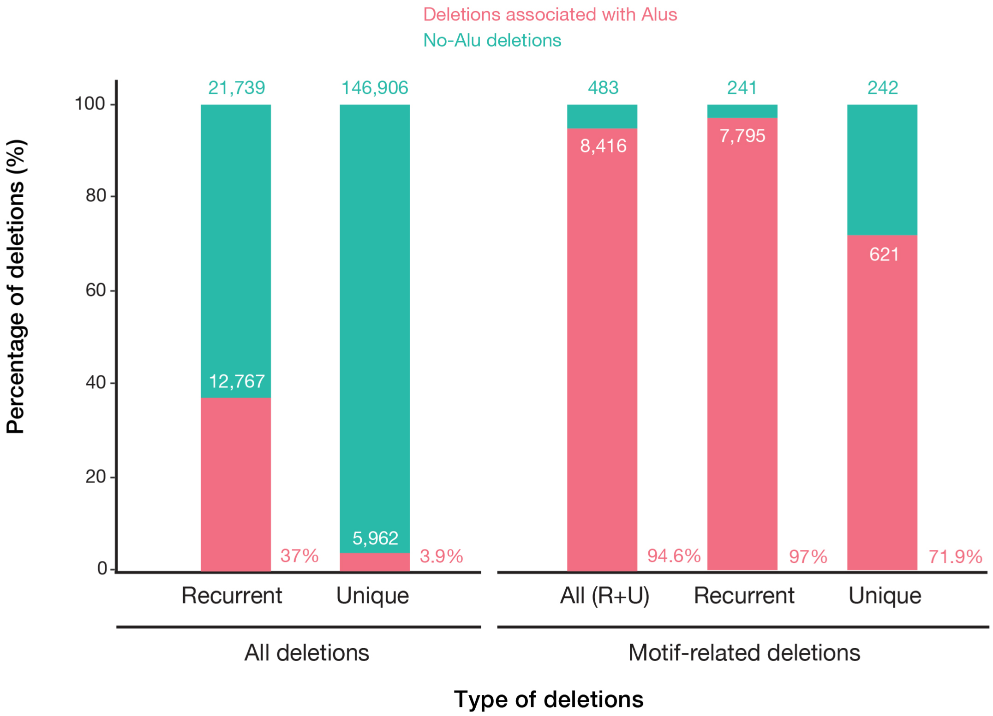


Figure 5. Proportion of *Alu* elements in motif-related deletions.

Comparison of the presence of *Alu* elements in recurrent and no-recurrent deletions from two different groups of variants: all deletions and motif-related deletions. Recurrent deletions and motif-related deletions have a higher prevalence of *Alus*.

Due to these inconsistencies in the data, and the fact that we did not expect such levels of recurrence in somatic variants, we chose to generate a random subset of deletions that we curated by inspecting manually the FastQ and BAM files to retrieve the reads supporting the rearrangements and discard possible false positives. Surprisingly, when looking at the subset of 65 PCAWG recurrent motif-related deletions in different patients, we found mutated reads not only in the tumor sample, as we expected, but also in the normal sample varying from patient to patient. We observed three different scenarios depending on the presence of mutated reads in each sample: a) no mutated reads in normal and mutated reads in the tumor; b) mutated reads in normal and mutated reads in the tumor; c) mutated reads in normal and few mutated reads in the tumor. Overall, we clearly saw evidence of supporting reads in the tumor sample for all those recurrent deletions. However, the presence of reads in the normal sample threw more uncertainty on the somatic classification of these variants.

Focusing on the deletion chr12:58310053-58310360 as an example, we identified this same motif-related deletion with the same exact breakpoints in 20 different patients using SMuFin, in 11 different patients using Pindel[14] from the official pipeline and in no patients using the official consensus calls, as expected due to its size of 307bp. This deletion was called as somatic in 31 different patients sharing the same breakpoints by two independent variant callers. Clearly, this sort of high recurrence is not expected from somatic variants, which raised the question about a possible incorrect classification of motif-related deletions, which could be germline variants, in fact.

However, in this particular case for example, nor all the patients present supporting reads in the normal sample, nor the variant was detected by the calling of germline variants done in the PCAWG study, nor it was included in reference panels of genetic variation such as UK10K[15] or 1000G[16], facts that we would expect from a germline variant exhibiting these levels of recurrence. In summary, motif-related deletions show high recurrence levels across patients unusual for somatic variants, but besides a fluctuating number of mutated reads in the normal samples, they do not display any characteristics from germline variants either.

These results lead us to new questions. We know that there are mutated reads supporting the deletions in normal and tumor samples, indicating that the deletions are real. We also know that previous studies[17,18] describe that PCR library preparation prior to sequencing has been associated with the generation of DNA rearrangements, notably deletions. With this in mind, could motif-related deletions, for instance, be generated as methodological artifacts and be consequently detected as somatic?

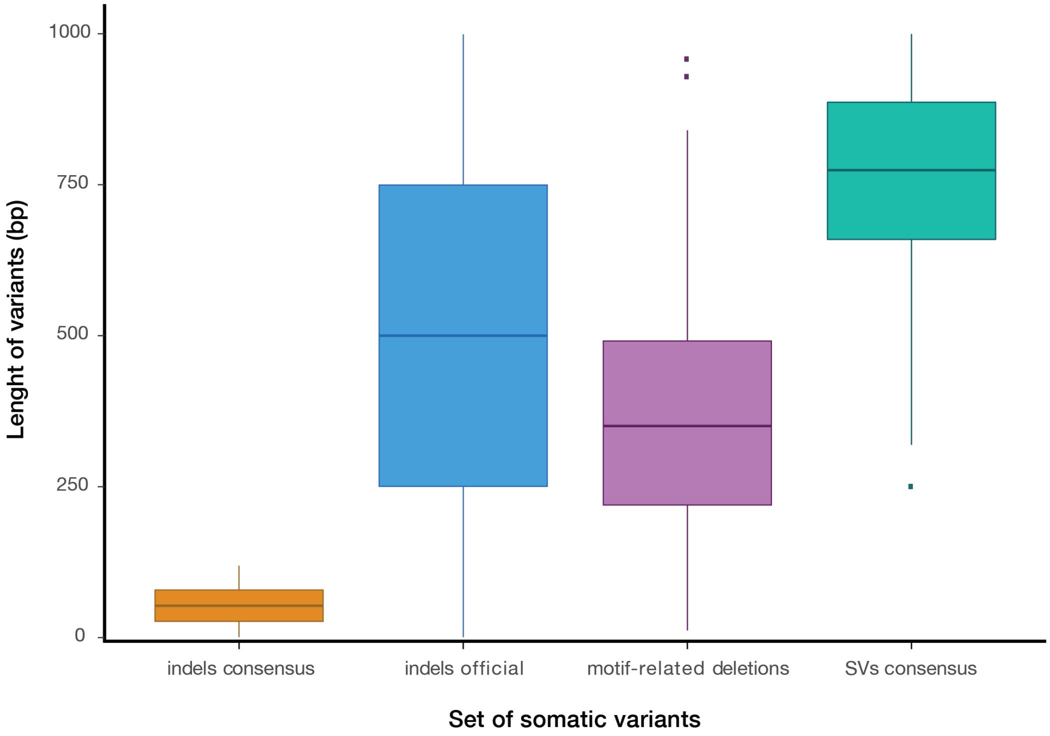


Figure 6. Distribution of the length of deletions in the different somatic calls from Pan-Cancer.

Comparison between the distribution of the length of all the somatic deletions included in the official indels consensus call, the official indels call from independent callers (Pindel), the motif-related deletions detected with SMuFin and the official SVs consensus call. The window size, which contains motif-related deletions, has been excluded from the two consensus calls from Pan-Cancer. Plot limited to 1000bp variants.

Sequence homology, and more specifically, *Alu* elements, have been proved to act as a substrate[19,20] in the recombination mechanism by template switching driven by PCR. In our case, the presence of motifs and repetitive elements flanking the loss of DNA along the usage of PCR for library preparation in PCAWG[21] certainly suggest that a fraction of recurrent motif-related deletions could be PCR-artifacts.

3.5 Classification of motif-related deletions and artifact-related deletions

At this point, the continuation of our analysis became more challenging. The obvious next step was to evaluate whether a fraction of motif-related deletions were, in fact, due to methodological artifacts or not. Luckily for us, we learned that one of the studies included in PCAWG, specifically the CLL-ES (Chronic lymphocytic leukemia) study, had a set of patients sequenced PCR-free[22], in an attempt to avoid the problems related with the usage of PCR library preparation in cancer analyses.

To better understand the role of PCR in the formation of recurrent deletions we classified 99 CLL-ES patients in two groups depending on the use of PCR for library preparation —66 PCR-free; 33 PCR+ (positive)—. Interestingly, the majority of patients affected by motif-related deletions were PCR+ (Fig. 7a) Looking at the distribution of the type of deletions between the two groups of patients, we determined a significant association between the library preparation method and the presence of motif-related deletions (p-value=1.847e-12; Fig. 7b). Our results support the idea of a potential artefactual origin of a subset of these deletions, since we also detected motif-related deletions, notably PGBD5-motif-related, in PCR-free patients.

4 Discussion and conclusions

There are two major classes of structural variants that are commonly excluded from cancer analysis due to the difficulty associated with their correct detection with short-read sequencing data: small SVs bigger than the read size and smaller than the library size, and rearrangements between highly-homologous sequences. In this analysis, we were dealing with deletions that combine these two classes as their main characteristics, with its consequent difficulty.

We have been able to detect small deletions displaying homologous specific sequences around their breakpoints in the majority of cancer studies from ICGC-PCAWG. Moreover, we have associated unexpected high recurrence levels to motif-related deletions exhibiting the same somatic deletions in different patients and cancer types. In an effort to explain this recurrence, we have evaluated the

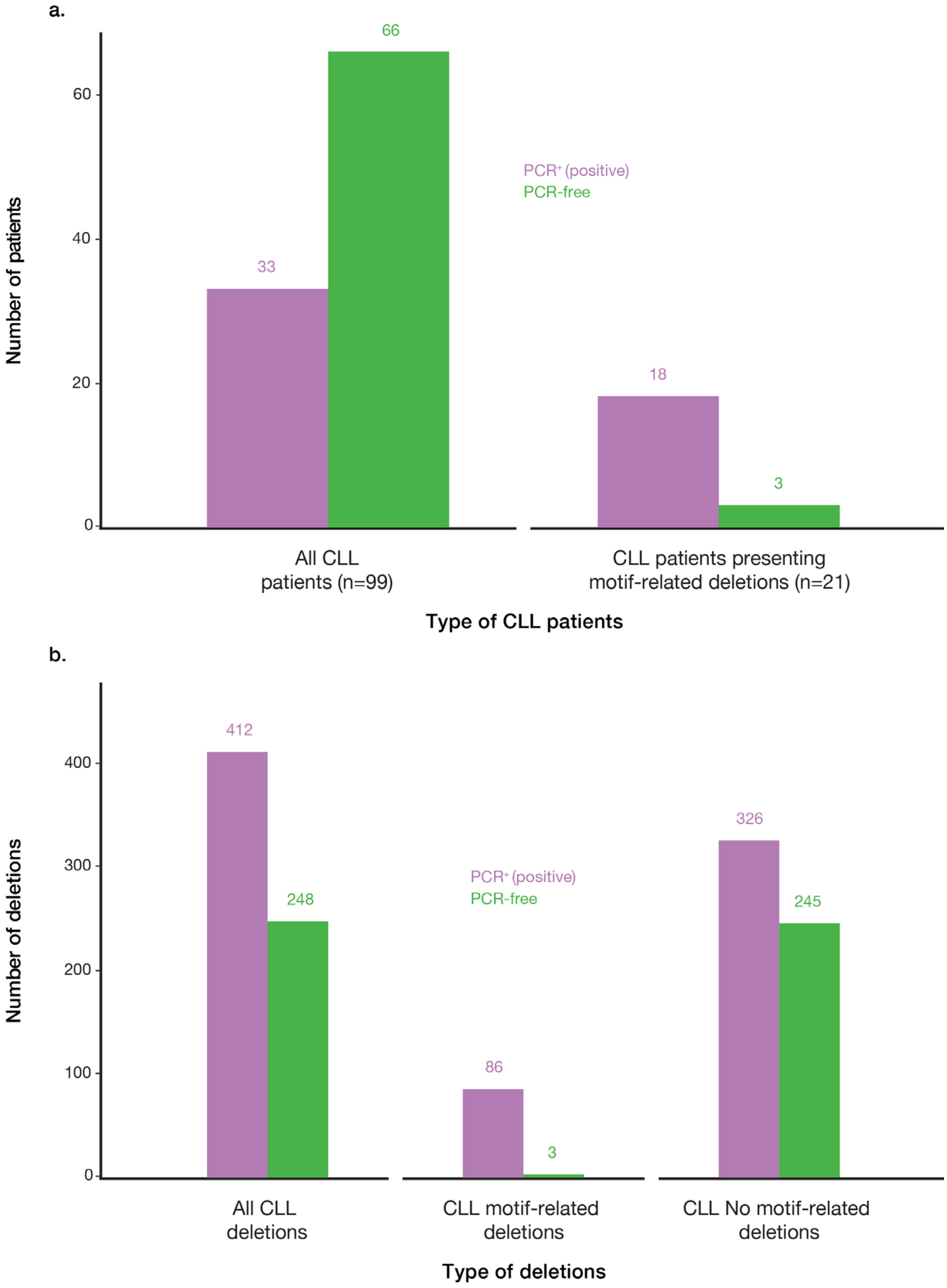


Figure 7. Study of the association between the use of PCR in library preparation and the presence of motif-related deletions.

(**a**), Number of CLL patients (n=99) PCR+ (purple) and PCR-free (green) presenting motif related-deletions (n=21). (**b**), Number of deletions detected in PCR+ and PCR-free patients associated or not with motif. Motif-related deletions are more present in PCR+ patients than PCR-free.

association of motif-related deletions with the methodology —PCR— used in the library preparation prior to the sequencing step. Although, our results suggest that PCR could be related with the generation of this type of deletions, the fact that we find motif-related deletions, notably the ones associated to PGBD5 motif, in PCR-free samples from CLL-ES suggest that not all those variants have an artifactual nature.

To better address this question, an ideal approach would be to sequence a subset of patients with both methods, PCR+ and PCR-free. This way, we could compare the results in order to quantify the effect of the library preparation methodology in the generation of structural variants. The perfect addition to this approach would be to include long-read sequencing data to validate motif-related deletions, especially the *Alu-Alu* recombination mechanism. In this line, we contacted CNAG, which was the center responsible for the sequencing of CLL patients from Pan-Cancer, trying to find a set with samples both PCR+/PCR-free, but the outcome was not favorable. For this reason, and because of the challenges associated with generating this sample set, this study was postponed. Further investigation is needed on this front to be able to draw final conclusions.

During the rest of my thesis, I had the opportunity to work in other projects that were sequenced PCR+ or PCR-free, obtaining, curiously, divergent results. In the case of the PCR+ project corresponding to the neuroblastoma study, which is exposed later in this thesis, we looked for PGBD5-motif-related deletions in the first stages of the analysis and did not find any significant results. On the other hand, in the last months of the thesis, we have analyzed PCR-free sequencing data from healthy adult and embryonal brain mice samples finding enrichment in PGBD5-motif-related deletions, although we did not find recurrence between samples. These last analyses contradict the association between PCR usage and the generation of PGBD5-motif-related deletions opening new questions to our PCAWG findings.

One of these questions relates precisely to the mice study. Having found PGBD5-motif-related deletions in healthy embryonal brains, could this type of deletions be related to cancer but also be generated in the first steps of the development, which is why we found mutated reads in the tumor but also a fluctuation in mutated reads in the normal samples?

To conclude, in this work, we aimed to study the landscape of PGBD5-related structural variation across multiple tumor types. By the analysis of PCAWG data, we have characterized the homology patterns related to small-size deletions (median size of 294bp), in order to determine the potential mechanism behind these rearrangements. The high level of repetitive elements and motifs identified flanking the deletion breakpoints —*Alu* and microhomologies—, strongly points to a homology-mediated mechanism of formation. Furthermore, we have also detected a fraction of these annotated deletions that could actually derive from methodological artifacts —generated, for example, from PCR amplification during library preparation—. These results seed light on the fraction of small size deletions and question the value of variant calling results without additional examination.

5 Methods

PCAWG whole-genome sequencing dataset.

We analyzed whole-genome sequencing data from the ICGC-PCAWG study for 2,834 tumor and matched normal pairs across 38 cancer types, of which 2,706 pairs from 37 cancer types that passed our quality-control criteria were selected for further analysis. The complete information for all tumor samples and patients is provided in the recently published overview of Pan-Cancer[1]. Sequencing reads were aligned to the hg19 reference genome using BWA-MEM v.2.6.0[23], and BioBamBam v.0.0.138[24] was used to mark duplicates.

Structural variant and indel detection.

Variant calling was performed on matching normal and tumor genomes using

SMuFin v.0.9.4[6] with default parameters, obtaining single nucleotide variants, indels, and structural variants. Breakpoints supported by fewer than four reads were filtered. For the evaluation of SMuFin calls, we used the results from the callers included in the three official established PCAWG pipelines. Each of the pipelines consisted in multiple software packages, in our case we used the ones for calling somatic indels and SVs: cgpPindel[14], BRASS[25] —from Sanger—, DELLY[26] —from EMBL/DKFZ—, and dRanger[27], SnowMan (a.k.a. SvABA[28]) —from Broad Institute—. We worked with the results from the independent callers as well as the officially released consensus set of SVs.

Motif discovery analysis.

The discovery of new motifs associated with recurrent deletions in our data has been performed using MEME v.4.10.0[29] (-nmotifs 10 -minw 6 -maxw 10 -revcomp) in 20bp sequences around each of the breakpoints. In order to not bias our results due to the high recurrence of the breakpoint positions, we got rid of the duplicated recurrent deletions selecting one hit per recurrent rearrangement. We selected the longest motifs that were more conserved with an e-value<0.001.

Motif-related deletions analysis.

Once we detected all the motifs, we developed a script to look for each of them in 20bp sequences around the original breakpoints and the breakpoints junction. We consider a motif-related deletion a deletion that conserves the same motif in each of the three points, in the original breakpoints and the reconstructed junction after the loss of the DNA region. We validated different sets of motif-related deletions for each of the motifs using BLAT[30] and manual inspection of split reads in the BAM[31] file.

Repetitive elements and genes analysis.

For this analysis, we retrieve the table of repetitive elements of the hg19 human genome from the UCSC[32], annotated using Repeatmasker[33]. In the case of genes, we employed the genes annotated by NCBI RefSeq human GRCh37/hg19. In both analyses, we used BEDTOOLS v.2.25.0[34] to intersect both tables with the detected motif-related deletions. From the Repeatmasker's table and visual inspection in the UCSC-Genome Browser[35], we evaluated the orientation of *Alu* elements.

Recurrence levels detection.

We developed a script to count all the hits of the same variant across our cohort —with the exact same breakpoint positions—, establishing the recurrence levels for each of the deletions.

Evaluation of the fluctuating number of reads.

We have detected the different supporting reads scenarios: a) no mutated reads in normal and mutated reads in tumor; b) mutated reads in normal and mutated reads in tumor; c) mutated reads in normal and few mutated reads in tumor by manually inspecting and counting the reads for a subset of variants in the BAM file. In addition, we have expanded our analysis to the rest of the variants using Pindel, which provided us with the number of supporting reads for each one of them.

Germline validation

To evaluate whether the motif-related deletions with mutated reads in the normal sample were germline variants detected as somatic, we searched the variants in the results from the official PCAWG germline variant calling[1] pipeline and two reference panels of human genetic variation that include indels and structural variants: UK10K[15] and 1000G[16].

Statistical analysis.

The significance of the association between PGBD5-motif-related deletions and recurrence was performed with a proportion test. The association between the library preparation method and the presence of motif-related deletions was performed using a chi-squared test. For both tests, the significant threshold was established at 0.05.

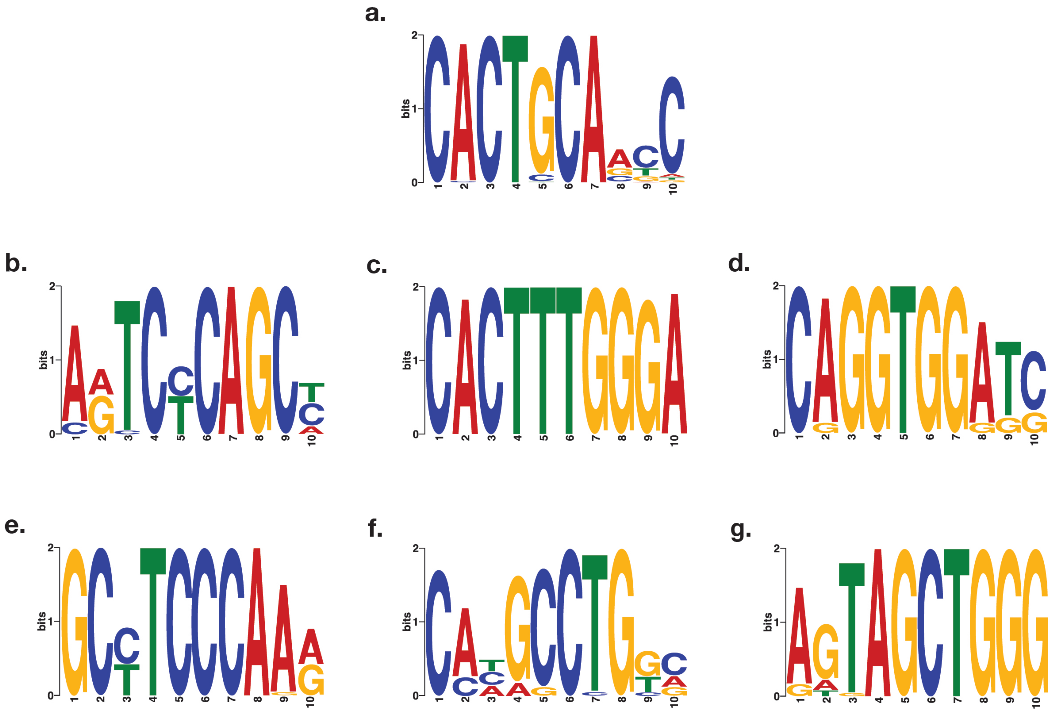
6 References

1. Consortium, I. T. P.-C. A. o. W. G. Pan-cancer analysis of whole genomes. Nature 578, 82-93, doi:10.1038/s41586-020-1969-6 (2020).
2. Henssen, A. G. et al. Genomic DNA transposition induced by human PGBD5. Elife 4, doi:10.7554/eLife.10565 (2015).
3. Zhao, S. et al. PiggyBac transposon vectors: the tools of the human gene encoding. Transl Lung Cancer Res 5, 120-125, doi:10.3978/j.issn.2218-6751.2016.01.05 (2016).
4. Henssen, A. G. et al. PGBD5 promotes site-specific oncogenic mutations in human tumors. Nat Genet 49, 1005-1014, doi:10.1038/ng.3866 (2017).
5. Kong, Y. et al. Transposable element expression in tumors is associated with immune infiltration and increased antigenicity. Nat Commun 10, 5228, doi:10.1038/s41467-019-13035-2 (2019).
6. Moncunill, V. et al. Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads. Nat Biotechnol 32, 1106-1112, doi:10.1038/nbt.3027 (2014).
7. Payer, L. M. & Burns, K. H. Transposable elements in human genetic disease. Nat Rev Genet 20, 760-772, doi:10.1038/s41576-019-0165-8 (2019).
8. Sen, S. K. et al. Human genomic deletions mediated by recombination between Alu elements. Am J Hum Genet 79, 41-53, doi:10.1086/504600 (2006).
9. Morales, M. E. et al. The contribution of alu elements to mutagenic DNA double-strand break repair. PLoS Genet 11, e1005016, doi:10.1371/journal.pgen.1005016 (2015).
10. White, T. B., Morales, M. E. & Deininger, P. L. Alu elements and DNA double-strand break repair. Mob Genet Elements 5, 81-85, doi:10.1080/2159256X.2015.1093067 (2015).
11. Li, Y. et al. Patterns of somatic structural variation in human cancer genomes. Nature 578, 112-121, doi:10.1038/s41586-019-1913-9 (2020).
12. Guan, P. & Sung, W. K. Structural variation detection using next-generation sequencing data: A comparative technical review. Methods 102, 36-49, doi:10.1016/j.ymeth.2016.01.020 (2016).
13. Belancio, V. P., Roy-Engel, A. M. & Deininger, P. L. All y'all need to know 'bout retroelements in cancer. Semin Cancer Biol 20, 200-210, doi:10.1016/j.semcancer.2010.06.001 (2010).
14. Raine, K. M. et al. cgpPindel: Identifying Somatically Acquired Insertion and Deletion Events from Paired End Sequencing. Curr Protoc Bioinformatics 52, 15 17 11-15 17 12, doi:10.1002/0471250953.bi1507s52 (2015).
15. Consortium, U. K. et al. The UK10K project identifies rare variants in health and disease. Nature 526, 82-90, doi:10.1038/nature14962 (2015).
16. Genomes Project, C. et al. A map of human genome variation from population-scale sequencing. Nature 467, 1061-1073, doi:10.1038/nature09534 (2010).
17. Hommelsheim, C. M., Frantzeskakis, L., Huang, M. & Ulker, B. PCR amplification of repetitive DNA: a limitation to genome editing technologies and many other applications. Sci Rep 4, 5052, doi:10.1038/srep05052 (2014).
18. Marton, A., Delbecchi, L. & Bourgaux, P. DNA nicking favors PCR recombination. Nucleic Acids Res 19, 2423-2426, doi:10.1093/nar/19.9.2423 (1991).
19. Kurahashi, H., Shaikh, T. H. & Emanuel, B. S. Alu-mediated PCR artifacts and the constitutional t(11;22) breakpoint. Hum Mol Genet 9, 2727-2732, doi:10.1093/hmg/9.18.2727 (2000).
20. Ji, W., Zhang, X. Y., Warshamana, G. S., Qu, G. Z. & Ehrlich, M. Effect of internal direct and inverted Alu repeat sequences on PCR. PCR Methods Appl 4, 109-116, doi:10.1101/gr.4.2.109 (1994).
21. Whalley, J. P. et al. Framework for quality assessment of whole genome, cancer sequences. bioRxiv, 140921, doi:10.1101/140921 (2017).
22. Puente, X. S. et al. Non-coding recurrent mutations in chronic lymphocytic leukaemia. Nature 526, 519-524, doi:10.1038/nature14666 (2015).

23. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).
24. Tischler G, L. S. biobambam: tools for read pair collation based algorithms on BAM files. Source Code Biol Med 2014;9:13, doi:10.1186/1751-0473-9-13 (2014).
25. BRASS, <https://github.com/cancerit/BRASS> (
26. Rausch, T. et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 28, i333-i339, doi:10.1093/bioinformatics/bts378 (2012).
27. Drier, Y. et al. Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. Genome Res 23, 228-235, doi:10.1101/gr.141382.112 (2013).
28. Wala, J. A. et al. SvABA: genome-wide detection of structural variants and indels by local assembly. Genome Res 28, 581-591, doi:10.1101/gr.221028.117 (2018).
29. Bailey, T. L., Williams, N., Misleh, C. & Li, W. W. MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res 34, W369-373, doi:10.1093/nar/gkl198 (2006).
30. Kent, W. J. BLAT--the BLAST-like alignment tool. Genome Res 12, 656-664, doi:10.1101/gr.229202 (2002).
31. Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).
32. Karolchik, D. et al. The UCSC Table Browser data retrieval tool. Nucleic Acids Res 32, D493-496, doi:10.1093/nar/gkh103 (2004).
33. Smit, A., Hubley, R & Green, P. RepeatMasker Open-4.0, <http://www.repeatmasker.org> (2013-2015).
34. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841-842, doi:10.1093/bioinformatics/btq033 (2010).
35. Karolchik, D., Hinrichs, A. S. & Kent, W. J. The UCSC Genome Browser. Curr Protoc Bioinformatics Chapter 1, Unit1 4, doi:10.1002/0471250953.bi0104s28 (2009).

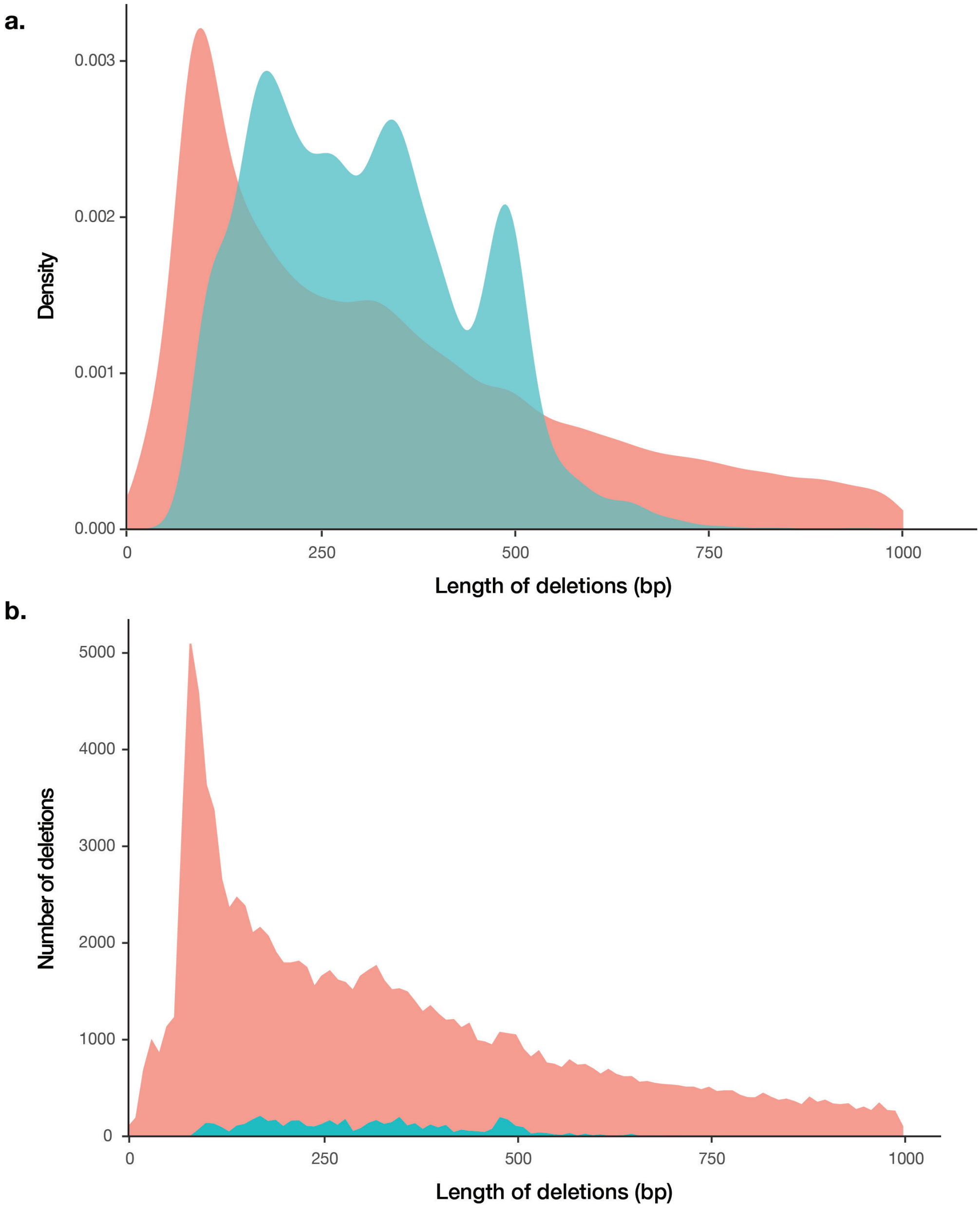
7 Supplementary information

(next pages)



Supplementary Figure 1. Specific sequences associated with motif-related deletions in PCAWG data.

Logos of the seven highly conserved motifs (**a**, **b**, **c**, **d**, **e**, **f**, **g**) significantly associated with recurrent deletions. In this figure, (**a**). corresponds to the already described PGBD5-related motif.



Supplementary Figure 2. Distribution of the length of PGBD5-motif-related deletions compared to all PCAWG detected deletions.

(**a**), Density plot showing the different length distributions between PGBD5-motif-related and all the detected deletions. (**b**), Plot showing the number of deletions per length for PGBD5-motif-related and all PCAWG detected deletions. Both plots have been limited to represent variants smaller than 1000bp. These plots illustrate the different distribution of sizes of PGBD5-motif-related deletions, manifesting a median size of 294bp.

Neuroblastoma publication

Nature Genetics

Extrachromosomal circular DNA drives oncogenic genome remodeling in neuroblastoma

Authors

Richard P. Koche[1,12*], **Elias Rodriguez-Fos**[2,12], Konstantin Helmsauer[3,12], Martin Burkert[4,5,6], Ian C. MacArthur[3], Jesper Maag[1], Rocio Chamorro[3], Natalia Munoz-Perez[3], Montserrat Puiggròs[2], Heathcliff Dorado Garcia[3], Yi Bei[3], Claudia Röefzaad[3], Victor Bardinet[3], Annabell Szymansky[3], Annika Winkler[3], Theresa Thole[3], Natalie Timme[3], Katharina Kasack[7], Steffen Fuchs[3,5,7], Filippos Klironomos[3], Nina Thiessen[5], Eric Blanc[5], Karin Schmelz[3], Annette Künkele[3,5,7], Patrick Hundsdörfer[3,5,7], Carolina Rosswog[7], Jessica Theissen[7], Dieter Beule[5], Hedwig Deubzer[3,7,8], Sascha Sauer[5], Joern Toedling[3], Matthias Fischer[9,10], Falk Hertwig[3,7], Roland F. Schwarz[6,7], Angelika Eggert[3,5,7], David Torrents[2,11,12], Johannes H. Schulte[3,5,7,12] and Anton G. Henssen[3,5,7,8,12*]

Citation

Koche, R.P., **Rodriguez-Fos, E.**, Helmsauer, E., Burkert, M., MacArthur, I.C., Maag, J., Chamorro, R., Munoz-Perez, N., Puiggròs, M., Dorado Garcia, H., et al. (2020). Extrachromosomal circular DNA drives oncogenic genome remodeling in neuroblastoma. Nature Genetics 52, 29-34.

1 Abstract

Extrachromosomal circularization of DNA is an important genomic feature in cancer. However, the structure, composition and genome-wide frequency of extrachromosomal circular DNA have not yet been profiled extensively. Here, we combine genomic and transcriptomic approaches to describe the landscape of extrachromosomal circular DNA in neuroblastoma, a tumor arising in childhood from primitive cells of the sympathetic nervous system. Our analysis identifies and characterizes a wide catalog of somatically acquired and undescribed extrachromosomal circular DNAs. Moreover, we find that extrachromosomal circular DNAs are an unanticipated major source of somatic rearrangements, contributing to oncogenic remodeling through chimeric circularization and reintegration of circular DNA into the linear genome. Cancer-causing lesions can emerge out of circle-derived rearrangements and are associated with adverse clinical outcome. It is highly probable that circle-derived rearrangements represent an ongoing mutagenic process. Thus, extrachromosomal circular DNAs represent a multihit mutagenic process, with important functional and clinical implications for the origins of genomic remodeling in cancer.

2 Main

Recent studies have shown that circular DNA is more prevalent in human tissues than previously anticipated[1,2,3,4,5]. Based on size and copy number, at least three classes of circular DNA exist in human cells: (1) small extrachromosomal circular DNA (including microDNA; referred to as eccDNA throughout the text)[3,6]; (2) large, copy number–amplified extrachromosomal circular DNA (ecDNA)[1], and (3) ring and/or neochromosomes[7,8]. ecDNA can lead to oncogene amplification and is a powerful driver of intratumoral heterogeneity[1,9,10,11,12]. Whether ecDNA has other cancer-causing functions is unknown, and the impact circularization has on genome remodeling is unclear.

Neuroblastoma is one of the first tumor entities where extrachromosomal oncogene circularization in the form of *MYCN* proto-oncogene double-minute chromosomes was detected[10,13]. Since the first descriptions in 1965 (refs. [14,15]), the extent of DNA circularization has not been accurately quantified in neuroblastoma. We hypothesized that DNA circularization could represent a genome-wide, driving mutagenic process in neuroblastoma with functional consequences beyond oncogene amplification. We set out to systematically describe the spectrum and impact of circular DNA in neuroblastoma by using different genomic and transcriptomic approaches (Supplementary Fig. 1).

Since DNA circularity can be computationally inferred from whole-genome sequencing (WGS) data[3,16,17], we applied an algorithm using paired-end read orientation to detect circularity to WGS from 93 neuroblastomas paired with normal blood specimens (Fig. 1a,b). This approach detected a large tumor-specific circular DNA catalog, including MYCN double-minute chromosomes, mitochondrial DNA and many previously undescribed ecDNAs and eccDNAs (Fig. 1c,d and Supplementary Fig. 2a,b). This suggests a greater prevalence and complexity of circular DNA in neuroblastoma than previously anticipated.

To achieve complementary and more sensitive detection and characterization of circular DNA in neuroblastoma, we adapted and modified the Circle sequencing (Circle-seq) method (Supplementary Figs. 1 and 2c,d)[6]. We achieved specific DNA circle enrichment through $>10^{10}$-fold depletion of linear genomic DNA (gDNA; Fig. 1c and Supplementary Figs. 2c and 3a–c). Applying Circle-seq to endonuclease-treated gDNA significantly reduced read mapping to circularized genomic regions by 474-fold ($P = 7.566 \times 10^{-11}$, Welch's t-test; Fig. 1c and Supplementary Fig. 3d,e), confirming specific enrichment of circular DNA. Sequence composition was analyzed and genomic origin inferred combining massive parallel paired-end sequencing with long-read Nanopore and single-molecule real-time sequencing (SMRT-seq). Circular head-to-tail junctions predicted computationally were confirmed by PCR and Sanger sequencing (Supplementary Fig. 3a–c). De novo sequence assembly of long reads spanning the entirety of circles allowed further physical confirmation of their circular structure in 65% of cases (Supplementary Fig. 4a–c). Circle-seq confirmed 100% of ecDNAs and 30% of eccDNAs predicted from WGS and identified on average 0.82 ecDNAs and 5,673 eccDNAs per neuroblastoma (Fig. 1c–e and Supplementary Fig. 4d–f). Although ecDNA was

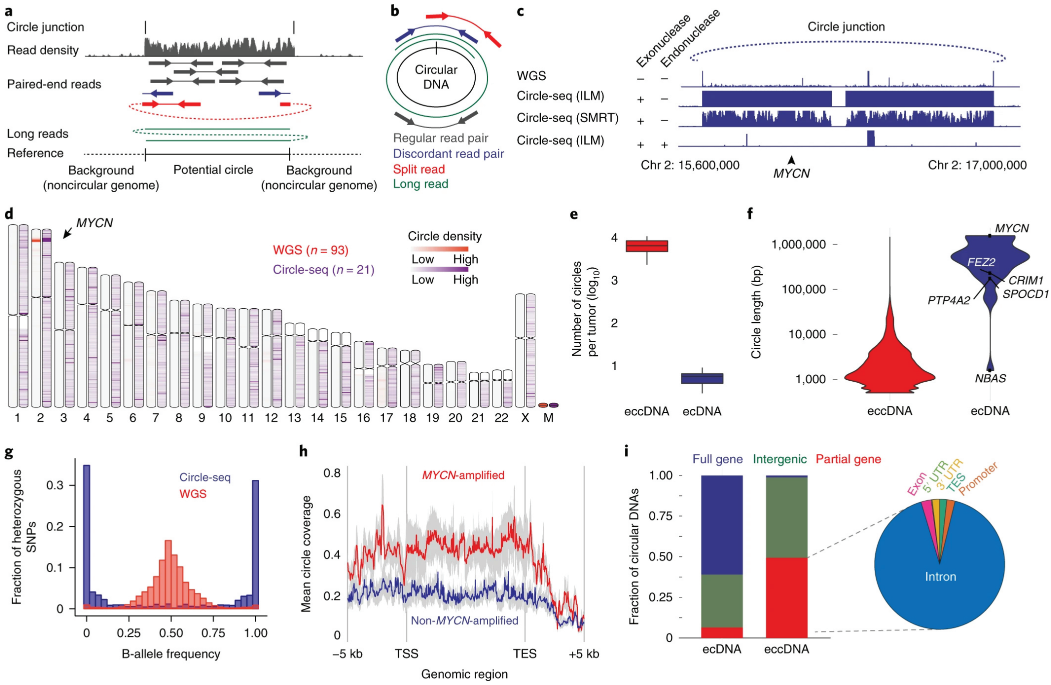


Figure 1. A genome-wide map of circular DNA in neuroblastoma.

(**a**), Schematic representation of sequencing reads as predicted for circular genomic regions. Background indicated noncircular genome. (**b**), Schematic representation of sequencing read positions on circular DNA. (**c**), Genome tracks comparing sequencing read densities on an ecDNA as detected via WGS (only circle-specific head-to-tail reads are depicted), Circle-seq followed by Illumina paired-end sequencing (ILM) and SMRT-seq in neuroblastoma cells. DNA digestion with an exonuclease and/or endonuclease is indicated (+/−). The dashed blue line indicates the predicted circle junction. Interruption of read density profile is due to lack of read alignment (*y* axis: 0–30 reads). (**d**,) Chromosome ideogram with genome-wide somatic circular DNA density as inferred from WGS (red) compared to Circle-seq (blue). M, circular mitochondrial DNA. (**e**), Number of ecDNAs and eccDNAs per neuroblastoma (*n* = 21 tumors, *n* = 96,436 eccDNAs, *n* = 14 ecDNAs). (**f**), Size distribution of ecDNAs and eccDNAs identified using Circle-seq in neuroblastomas (*n* = 21 tumors, *n* = 96,436 eccDNAs, *n* = 14 ecDNAs). (**g**), Alternative B-allele frequencies (BAF) in the sequencing reads from Circle-seq (*n* = 21 tumors) and WGS (*n* = 93 tumors). (**h**), Density of circular DNA detected using Circle-seq over genic compared to gene-surrounding regions in *MYCN*-amplified and nonamplified neuroblastomas (*n* = 7 *MYCN*-amplified tumors, *n* = 14 nonamplified tumors). The lines represent the mean signal and the shaded area represents the s.e.m. TES, transcription end site; TSS, transcription start site. (**i**), Fraction of genomic regions affected by eccDNA compared to ecDNA (*n* = 21 tumors).

accurately predicted from WGS with high sensitivity (100%), our results highlight the advantages of using additional and more sensitive approaches, such as Circle-seq, to obtain a comprehensive characterization of circular DNAs in tumors.

The structure of circularized genomic loci in neuroblastoma varied considerably, with mean sizes of 680,200 base pairs (bp; ecDNA) and 2,403 bp (eccDNA) in tumors, reproducing the oscillating length distribution observed in lymphoma cancer cell lines[3] (Fig. 1f and Supplementary Fig. 4g–j). In agreement with cytogenetic reports[18], no ring chromosomes were detected in neuroblastoma. Notably, both ecDNAs and eccDNAs were of monoallelic origin, as determined by haplotype phasing (Fig. 1g). Inspection of circle junction sequences (ecDNA and eccDNA) indicated the probable mechanism(s) of generation, since 2.8% contained nontemplate insertions indicative of nonhomologous end joining repair or replication-associated mechanisms (Supplementary Fig. 4k). In line with reports in human lymphoma cell lines[19], 6.3% of circle junctions contained sequence microhomologies (minimally 5 bp), suggesting the involvement of microhomology-mediated DNA repair (Supplementary Fig. 4l). Notably, eccDNA and ecDNA were significantly enriched in genic regions, particularly in *MYCN*-amplified neuroblastomas (Fig. 1h and Supplementary Fig. 5a–c). Whereas ecDNAs regularly contained entire genes (62.5%), eccDNAs mostly included fractions of genes (Fig. 1i). Our genome-wide map of circular DNA in neuroblastoma shows that DNA circularization is not restricted to proto-oncogenes but also affects various coding and noncoding regions with yet unknown functional consequences.

Extrachromosomal circularization and amplification are associated with increased oncogene expression. It is unclear whether circularization itself or subsequent circle copy number amplification drives overexpression. The majority of genomic amplifications (85.7%) identified using WGS coincided with ecDNAs, as confirmed by Circle-seq, suggesting that ecDNAs contribute to genomic amplifications. Moreover, haplotype phasing showed that ecDNAs were exclusively derived from the amplified allele, confirming extrachromosomal circularization as a potential driver of high-level focal genomic amplifications (Fig. 2a,b). Notably, circle length was significantly associated with a higher copy number of circularized regions (Supplementary Fig. 5d; $P < 1 \times 10^{-4}$), implicating circle length as a determining factor for subsequent amplification/propagation of circular DNA (Supplementary

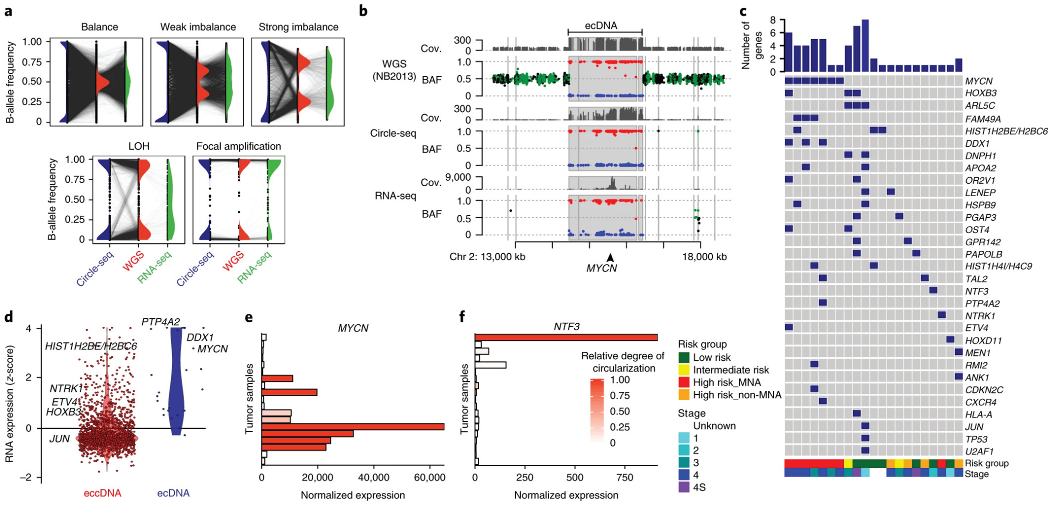


Figure 2. Monoallelic large circular DNAs are an origin of oncogene amplification and overexpression in neuroblastoma.

(**a**), B-allele frequency of all circular DNAs involving genes (both ecDNA and eccDNA) detected using Circle-seq (blue) compared to the corresponding genomic loci in WGS (red) and mRNA expressed from genes affected by DNA circularization measured using RNA-seq (green; the gray lines indicate the corresponding measurements from Circle-seq, WGS and RNA-seq, n = 18 tumors). (**b**), Genome track with phased reads from WGS of NB2013, Circle-seq and RNA-seq at the region of extrachromosomal circularization on chromosome 2 affecting *MYCN*. The blue and red colored dots represent reads from different haplotypes. Cov., read coverage. (**c**), Genes (rows) affected by circularization in neuroblastoma samples (columns) as detected using Circle-seq (n = 21 tumors). (**d**), Relative mRNA expression (z-scores) of genes affected by DNA circularization in the form of eccDNA (n = 1,696) compared to ecDNA (n = 24) as measured using total RNA-seq (n = 21 tumors). (**e**,**f**), Normalized gene expression (mRNA) for *MYCN* proto-oncogene, basic helix-loop-helix transcription factor (**e**) and *NTF3* (**f**) in neuroblastomas (*MNA*, *MYCN*-amplified neuroblastoma). The degree of gene circularization is indicated in red (see color scale).

Fig. 5d–f). In agreement with its prominent role in neuroblastoma genesis, *MYCN* was the most recurrently extrachromosomally amplified and overexpressed gene in our cohort (Fig. 2b–e and Supplementary Fig. 5a–c). Other cancer-related genes listed in the Catalogue Of Somatic Mutations In Cancer (COSMIC) database[20] were also circularized in tumors and neuroblastoma cell lines, including the *JUN* and *MDM2* proto-oncogenes and *SOX11* and *TAL2* transcription factor genes (Fig. 2c and Supplementary Fig. 5a–c). However, the genomic copy number of oncogenes contained in the majority of eccDNAs was not altered (Supplementary Fig. 5g,h), suggesting that extrachromosomal circularization may be required but insufficient for oncogene amplification.

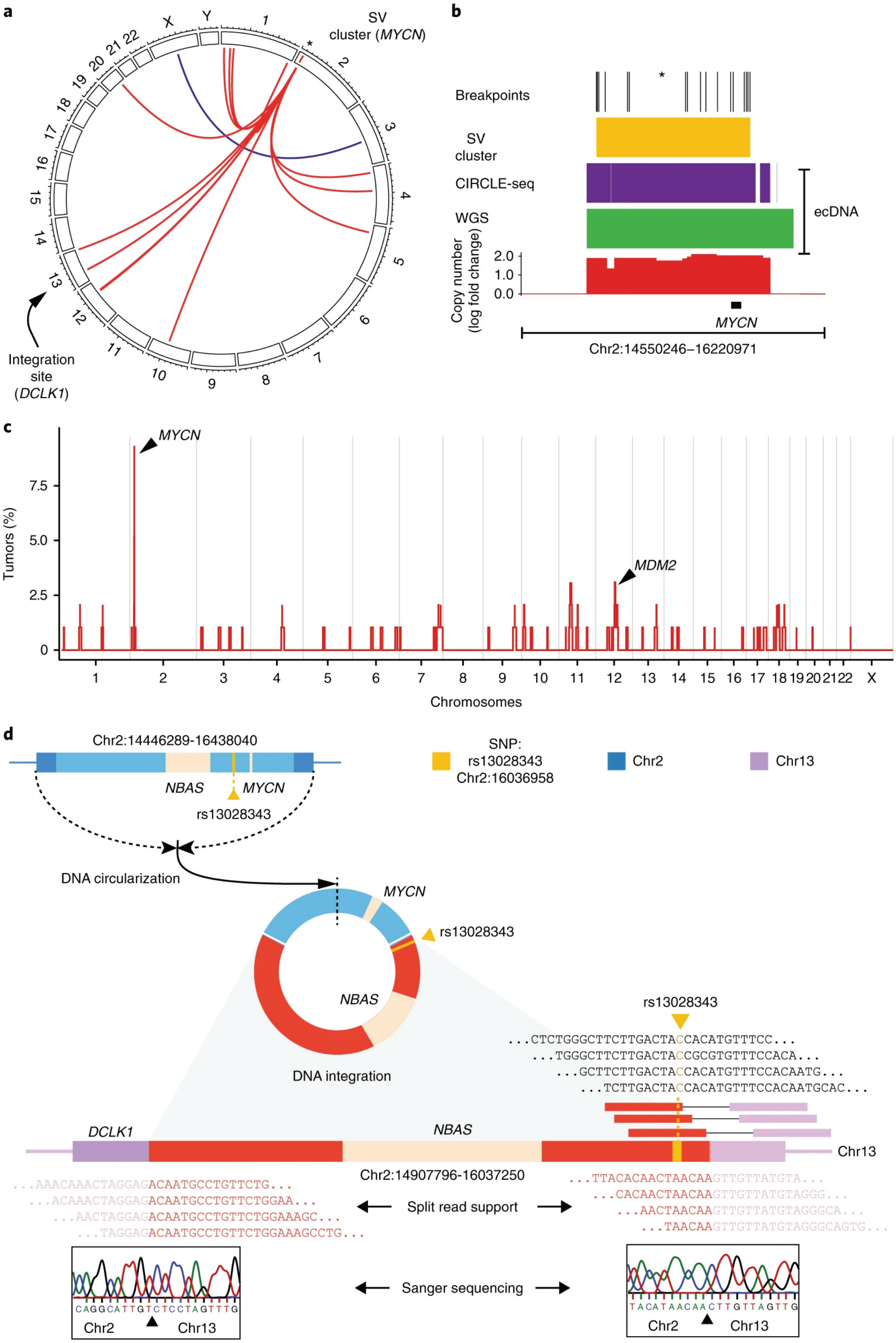
To determine the consequences of DNA circularization on gene expression, we performed total RNA sequencing (RNA-seq) on our neuroblastoma cohort. Whereas differences in gene expression were not observed for most genes affected by circularization in the form of small eccDNA (Fig. 2d and Supplementary Fig. 5i–j), massive increases in expression occurred for a small subset of genes entirely incorporated on circularized DNA and amplified as ecDNA (Fig. 2d–f). For example, *NTF3*, a gene encoding a neurotrophic factor with known importance in neuroblastoma[21], was strongly expressed from amplified ecDNA (Fig. 2f). Allele-specific messenger RNA expression (allele-specific expression (ASE)) analysis confirmed that increased gene expression originated from the circular allele (Fig. 2a,b). In contrast, ASE from copy number–neutral extrachromosomal circles did not differ from noncircular counterparts (Supplementary Fig. 5g,i,j; binomial test for equal probability, $P = 0.24$), suggesting that DNA circularization was insufficient to induce high-level gene expression. Thus, even though DNA circularization is a major route to gene amplification, it appears insufficient alone (without combined amplification) to increase gene expression. Given this observation, we hypothesized that circular DNA may have additional, cancer-relevant functions.

The genome-wide frequency and functional impact of circle-derived structural rearrangements, such as chimeric circle formation (circular DNA including parts from different chromosomes)[17,22], and circular DNA reintegration[23], in neuroblastomas are currently unknown. We hypothesized that beyond their ability to drive gene amplification, circular DNAs may serve as substrates for oncogenic genome remodeling. We sought evidence of genomic rearrangements at circularization loci (ecDNA and eccDNA) in WGS data (Supplementary Fig. 1). Strikingly, most intrachromosomal and interchromosomal rearrangements detected in neuroblastoma genomes coincided with regions of extrachromosomal circularization, supporting the idea of circle-mediated genome remodeling

(Supplementary Fig. 6a,b). Visual inspection of Circos plots from each tumor showed that interchromosomal rearrangements at circularization loci often formed a tree-shaped pattern, defined as clusters of at least three interchromosomal rearrangements with the same origin and branches reaching other distant genomic regions (Fig. 3a,b and Supplementary Fig. 7a–l). Tree-shaped rearrangement cluster origins significantly overlapped with ecDNAs, with hot spots on chromosomes 2 (including *MYCN*) and 12 (Fig. 3c and Supplementary Fig. 7i). Only 10.5% of *MYCN*-amplified neuroblastomas displayed homogenously staining regions (Supplementary Table 1), consistent with their rarity in neuroblastomas[14,24,25]. Thus, the majority of *MYCN*-derived tree-shaped rearrangements did not represent homogenously staining regions. Tree-shaped rearrangement patterns indicative of circle-derived rearrangements were detected in 9% of pediatric tumors in the analysis of an independent dataset of structural rearrangements in 546 pediatric cancer genomes[26], confirming that this pattern is neither entity-specific nor dependent on variant detection methods (Supplementary Fig. 7j). Our data reveal an unanticipated association between circular DNA and somatic genomic rearrangements in neuroblastoma.

Figure 3. The majority of structural rearrangements involve sites of DNA circularization and form clustered rearrangement patterns in neuroblastoma.

(**a**), Circos plot of interchromosomal rearrangements identified using five variant detection algorithms in one neuroblastoma genome (CB2013), shown exemplarily. The tree-shaped clustered rearrangement pattern (red), originating at a region of *MYCN* circularization (asterisk) is highlighted (SV, structural variants). (**b**), Detailed view of genomic breakpoint localizations (black) at the base of the tree-shaped rearrangement cluster (SV cluster) for the neuroblastoma shown in a define a region of clustered breakpoints (yellow) and overlaps with the region of DNA circularization, as detected using Circle-seq (pink) and WGS (green). The copy number changes are highlighted in red. (**c**), Genome-wide frequency of tree-shaped clusters of rearrangements in 93 primary neuroblastoma samples. The pattern is recurrently identified on chromosome 2 (at the *MYCN* locus), and chromosomes 11 and 12 (at the *MDM2* locus). (**d**), Schematic representation of circle integration in an example of neuroblastoma (CB2013). The genomic region, including *MYCN* (blue), is circularized; parts of the extrachromosomal circle are integrated (red) into chromosome 13 (pink) leading to a disruption of *DCLK1*. The sequencing reads supporting a circle-specific SNP as well as split reads supporting circle integration are shown below. Sanger sequencing of integration breakpoints is shown in the boxes. *(next page)*


a
X
Y
1
SV cluster (MYCN)
*
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Integration site (DCLK1)
b
Breakpoints
*
SV cluster
CIRCLE-seq
WGS
ecDNA
Copy number (log fold change)
2.0
1.0
0.0
MYCN
Chr2:14550246–16220971
c
MYCN
MDM2
Tumors (%)
7.5
5.0
2.5
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Chromosomes
d
Chr2:14446289-16438040
NBAS
MYCN
rs13028343
SNP: rs13028343 Chr2:16036958
Chr2
Chr13
DNA circularization
MYCN
rs13028343
NBAS
DNA integration
rs13028343
...CTCTGGGCTTCTTGACTACCACATGTTTCC...
...TGGGCTTCTTGACTACCGCGTGTTTCCACA...
...GCTTCTTGACTACCACATGTTTCCACAATG...
...TCTTGACTACCACATGTTTCCACAATGCAC...
DCLK1
NBAS
Chr13
Chr2:14907796-16037250
...AAACAAACTAGGAGACAATGCCTGTTCTG...
...ACAAACTAGGAGACAATGCCTGTTCTGGAA...
...AACTAGGAGACAATGCCTGTTCTGGAAAGC...
...TAGGAGACAATGCCTGTTCTGGAAAGCCTG...
Split read support
...TTACACAACTAACAAGTTGTTATGTA...
...CACAACTAACAAGTTGTTATGTAGGG...
...AACTAACAAGTTGTTATGTAGGGCA...
...TAACAAGTTGTTATGTAGGGCAGTG...
Sanger sequencing
CAGGCATTGTCTCCTAGTTTG
Chr2
Chr13
TACATAACAACTTGTTAGTTG
Chr2
Chr13

We reasoned that circle-derived tree-shaped rearrangements could either represent chromosomal circle integrations or the formation of chimeric circles, incorporating different chromosomal parts. To test this, we inspected the rearrangement recipient sites for signs of extrachromosomal circularization and integration and performed de novo assembly of circular DNAs (ecDNA and eccDNA). Extrachromosomal circular DNAs (identified using Circle-seq) appeared in 5.5% of rearrangement recipient sites (tree branch intercepts), indicating chimeric circle formation (Supplementary Fig. 6). This was confirmed by long-read Nanopore sequencing and assembly-based circle reconstruction, determining chimeric structures in 2.1% of eccDNAs and 84% of ecDNAs with on average 2.2 and 4.8 chimeric segments, respectively. Chromosomal circle integration was defined as interchromosomal rearrangements connecting extrachromosomal circles with intrachromosomal sites (that is, not detected by Circle-seq). The majority of rearrangement recipient sites (83.3%) were classified as circle integrations (Fig. 3d and Supplementary Fig. 6), which were validated by visual inspection of split reads, allele-specific PCR and Sanger sequencing (Fig. 3d and Supplementary Fig. 8). Phased heterozygous SNPs near integration breakpoints further confirmed extrachromosomal DNA circles as the origin of the integrations (Fig. 3d). Thus, circle-derived, tree-shaped rearrangement clusters represent (1) formation of chimeric circles and (2) chromosomal circle integrations.

To test the functional impact of circle-derived, tree-shaped rearrangements in neuroblastoma, we inspected the rearrangement recipient sites for the presence of cancer-relevant genes and changes in gene expression (Fig. 4a). Circle integration sites and sites included in chimeric circles were significantly enriched for cancer-relevant genes ($P = 0.033$) and particularly for tumor suppressor genes ($P = 0.033$), whose expression varied from tumors where the same gene was not involved in circle-derived rearrangements (Fig. 4b,c and Supplementary Fig. 9). For example, integration of an extrachromosomal circle fragment into the DCLK1 gene (shown in Fig. 3d) led to loss of heterozygosity (LOH) and was associated with significant repression of *DCLK1* expression (Fig. 4b). In agreement with a tumor suppressor

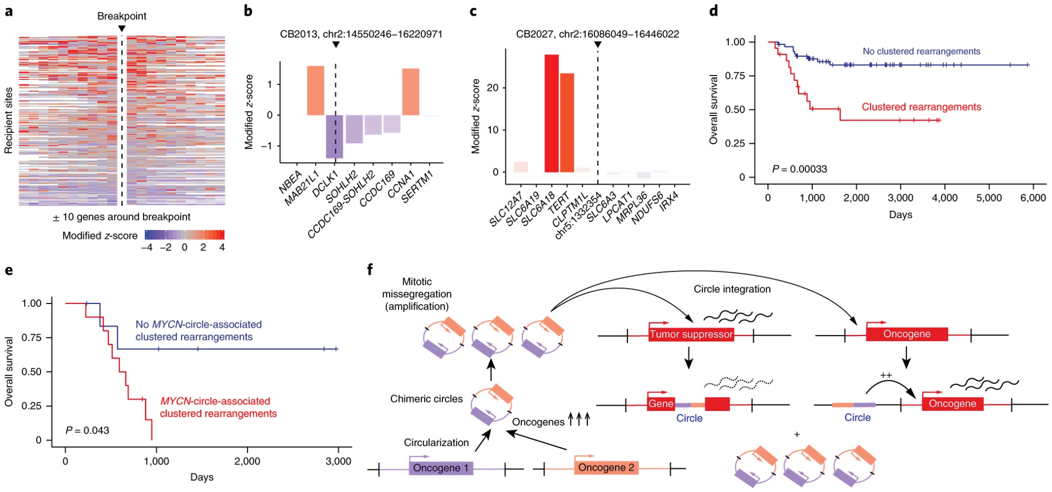


Figure 4. Rearrangement of circular DNAs drives transcriptional deregulation and dismal prognosis in neuroblastoma.

(**a**), Heatmap showing differential expression of up to ten genes located both upstream and downstream or a maximal distance of 2 Mb from each circle-derived rearrangement breakpoint ($n = 259$ breakpoints, $n = 24$ tumors). (**b**,**c**), The modified z-scores for the expression of the cancer-relevant genes *DCLK1* (**b**) and *TERT* (**c**) affected by circle-derived rearrangements are shown for two representative genomic loci (in two neuroblastomas). (**d**), Kaplan–Meier analysis comparing the neuroblastoma patient survival of patients with neuroblastomas affected by circle-derived clustered rearrangements ($n = 22$ patients) to patients with tumors lacking such rearrangements ($n = 59$ patients, $P = 0.00033$, two-sided log-rank test). (**e**), Kaplan–Meier analysis comparing neuroblastoma patient survival with *MYCN*-amplified tumors affected by *MYCN*-circle-derived clustered rearrangements ($n = 10$) to patients with tumors lacking such rearrangements ($n = 7$, $P = 0.043$, two-sided log-rank test). (**f**), Schematic diagram of the proposed mechanism of circle-mediated genome remodeling.

function in neuroblastoma, low *DCLK1* expression was associated with adverse patient prognosis and short hairpin RNA-mediated *DCLK1* knockdown significantly increased clonogenicity in neuroblastoma cell lines (Supplementary Fig. 10a–i). Notably, circle integration also occurred proximal to the *TERT* gene and was associated with enhanced *TERT* expression (Fig. 4c). It is tempting to speculate that enhancer hijacking[27] or disruption of other *cis*-regulatory elements could explain such expression changes. Chimeric circle formation, on the other hand, often resulted in simultaneous amplification of multiple proto-oncogenes and aberrant circle-specific fusion transcript expression in a subset of cases (Supplementary

Fig. 11). Thus, circle-derived rearrangements can contribute to aberrant expression of cellular tumor suppressors and proto-oncogenes.

Seemingly genetically identical *MYCN*-amplified neuroblastomas can produce strong clinical heterogeneity, representing a conundrum in the field. We hypothesized that circle-derived oncogenic lesions could functionally cooperate with extrachromosomal circular *MYCN* amplification, explaining some of the clinical heterogeneity observed. Indeed, the presence of circle-derived rearrangements was associated with adverse patient outcome (Fig. 4d). In line with our hypothesis, patients with *MYCN*-amplified neuroblastomas and circle-derived rearrangement clusters involving *MYCN* had significantly worse overall survival compared to patients with *MYCN*-amplified tumors lacking such rearrangements (Fig. 4e). Contrastingly, the number of rearrangements in *MYCN*-amplified tumors did not correlate with survival (Supplementary Fig. 12a–c). This implicates circle-derived rearrangements as clinically relevant genomic alterations in neuroblastoma.

Our work provides a comprehensive map of extrachromosomal DNA circularization in neuroblastoma, revealing this mutagenic process to be more frequent than previously anticipated. We demonstrate that the majority of genomic rearrangements in neuroblastoma involve circular DNA, challenging our current understanding about cancer genome remodeling. Such rearrangements have previously gone largely undetected or underestimated in WGS analyses because integrative, sequencing-based methods identifying circular DNA in tumor samples were lacking. In contrast to previous cytogenetic reports describing homogenously staining region-based circle integration and chimeric circle formation as a means of stable gene amplification, we conclude that extrachromosomally circularized DNA can actively contribute to genome remodeling with important functional and clinical consequences (Fig. 4f). It is tempting to speculate that factors exist, such as recently described oncogenic transposases[28,29,30], that could induce a mutator phenotype in the presence of circular DNA, driving circle-mediated genome remodeling. We envision that our findings extend to other cancers and that further detailed analyses of circle-derived rearrangements will shed new insights into our understanding of cancer genome remodeling.

3 Methods

Reagents.

The synthetic oligonucleotides listed in Supplementary Table 2 were obtained from Eurofins Genomics and were salt-free purified. pLKO.1 shRNA vectors targeting *DCLK1* (TRCN0000002145, TRCN0000002146) and control short hairpin green fluorescent protein were obtained from the RNAi Consortium (Broad Institute).

Cell culture.

Human tumor cell lines were obtained from the DSMZ-German Collection of Microorganisms and Cell Cultures (Leibniz Institute), from *ATCC* or were a gift from C. J. Thiele. The identity of all cell lines was verified by short tandem repeat STR genotyping (Genetica DNA Laboratories and/or IDEXX BioResearch). Absence of *Mycoplasma* contamination was determined with a MycoAlert system (Lonza). Neuroblastoma cell lines were cultured in RPMI 1640 medium (Thermo Fisher Scientific) supplemented with penicillin, streptomycin and 10% FCS. To assess the number of viable cells, cells were trypsinized, resuspended in medium and sedimented at 500*g* for 5 min. Cells were then resuspended in medium, mixed in a 1:1 ratio with 0.02% Trypan Blue Solution (Thermo Fisher Scientific) and counted with a TC20 Automated Cell Counter (Bio-Rad Laboratories). Lentiviral production and transduction were performed as described previously[28]. Clonogenicity was assessed as described previously[28]. Kelly and IMR-5 cells were plated in 24-well microplates at a concentration of 5,000 cells per well and incubated for 7 d. Clonogenicity was quantified using methods described previously[31].

Protein blotting.

Protein blotting was performed as described previously[28] using antibodies directed against mouse anti-β-actin (clone 8H10D10; Cell Signaling Technology), mouse anti-α-tubulin (clone DM1A; Cell Signaling Technology) and rabbit anti-DCLK1/DCAMKL1 (clone D2U3L; Cell Signaling Technology).

PCR and Sanger sequencing.

PCR reactions were performed on 50–100 ng of gDNA using 0.4U Phusion Hot Start II High-Fidelity DNA Polymerase (Thermo Fisher Scientific), 0.5 µM forward and reverse primers (Supplementary Table 2), 200 µM deoxyribonucleotide triphosphates (Bio-Budget Technologies) and 4 µl 5× Phusion Green buffer (Thermo Fisher Scientific). PCR products were resolved on 1% agarose gels. PCR amplicons were purified using the PureLink PCR Purification Kit (Thermo Fisher Scientific). Sanger sequencing was carried out by capillary sequencing using standard procedures (Eurofins Genomics).

Quantitative PCR (qPCR).

qPCR was performed using 50 ng or 1.5 µl of template DNA and 0.5 µM primers with SYBR Green PCR Master Mix (Thermo Fisher Scientific) in FrameStar 96-well PCR plates (4titude). Reactions were run and monitored on a StepOnePlus Real-Time PCR System (Thermo Fisher Scientific) and Ct values were calculated with the StepOne Plus software v.2.3 (Thermo Fisher Scientific).

Circular DNA isolation, purification and sequencing.

Circular DNA isolation and purification was performed on the samples described in Supplementary Tables 3 and 4 similarly to previous reports of Circle-seq[6]. A detailed step-by-step protocol for circular DNA isolation has been deposited on the Nature Protocol Exchange server[32]. DNA content was measured with a NanoDrop 2000 Spectrophotometer (Thermo Fisher Scientific) and a Qubit 3.0 Fluorometer (Thermo Fisher Scientific). Amplified circular DNA was sheared to an average fragment size of 150–200 bp using an S220 focused ultrasonicator (Covaris). Libraries for next-generation sequencing were prepared using the NEBNext Ultra DNA Library Kit for

Illumina according to the manufacturer's protocol (New England Biolabs). Libraries were sequenced on MiSeq instruments with 2 × 150 bp paired-end reads, HiSeq 4000 instruments with 2 × 125 bp paired-end reads or NextSeq 500 instruments with 2 × 150 bp paired-end reads (all Illumina). SMRT-seq was performed on a PacBio RS II instrument according to the manufacturer's protocol (Pacific Biosciences). Nanopore sequencing was performed on a MinION instrument according to the manufacturer's protocol (Oxford Nanopore).

Circle-seq analysis.

Reads were 3' trimmed for both quality and adapter sequences, with reads removed if the length was less than 20 nucleotides. Burrows–Wheeler Aligner MEM v.0.7.15 with default parameters was used to align the reads to human reference assembly hg19; PCR and optical duplicates were removed with Picard v.2.16.0. The aligned BAM files were then analyzed in two ways. First, all read pairs and split reads containing any outward-facing read orientation, indicating potential circles, were placed in a new BAM file. Second, genomic segments enriched for signal over background were detected in the 'all reads' BAM file using variable-width windows from Homer v.4.11 findPeaks (http://homer.ucsd.edu/), and the edges of these enriched regions were intersected with the 'circle only' BAM file to quantify the number of circle-supporting reads. To determine the thresholds for significance of real circles versus background noise, matched WGS data were used to determine the background distribution of circle-oriented reads in non-circle-enriched regions that were matched for length and nucleotide composition. An empirical P value of 0.01 was used to filter putative circles and regions passing this filter were then used for downstream analysis.

Circle analysis in WGS data.

Alignments to hg19 were created as outlined earlier, with read trimming, Burrows–Wheeler Aligner MEM and duplicate removal. Discovery of putative tumor-specific circular DNA relied on the filtering of false positives from genomic sequence as well as circles from normal tissue. This was classified with the following approach: (1) alignments with an outward-facing read orientation served as markers of putative circle boundaries projected onto the linear genome; (2) all such regions were merged if their edges occurred within 500 bp on both ends; (3) regions not meeting the empirically defined background threshold were filtered out ($P < 0.01$; see Circle-seq analysis); (4) lastly, these putative circles were classified as tumor-specific once filtered against circles discovered in the matched normal genome (using steps 1–3). To allow for the detection of copy number–neutral DNA circles, copy number information was not used for this analysis. We confirmed that tandem duplications identified using variant calling algorithms did not identify the same number of circular DNA from the WGS data (Supplementary Fig. 4).

De novo assembly of extrachromosomal circular DNA.

De novo assembly of long-read data (SMRT and Nanopore) was accomplished using two approaches. First, for long-read data alone, the Flye v.2.5 assembler (http://github.com/fenderglass/Flye) was used in '-meta' mode with circle junctions evaluated after polishing. Second, for hybrid assemblies using both long and short read data, Unicycler v.0.4.7 (http://github.com/rrwick/Unicycler) was used with racon v.1.3.3 and SPAdes v.3.13.0 and polished with Pilon v.1.23. In all cases, circle assembly was inspected visually using Bandage v.0.8.1 (http://rrwick.github.io/Bandage/). Genic overlap with de novo assemblies was evaluated in two ways. First, by building a BLAST database of all assembled contigs and scoring matches to human genes with at least 70% of gene length covered. Second, each contig, independent of genic overlap, was mapped to hg19 using minimap2 v.2.17 (http://github.com/lh3/minimap2).

SMRT-seq analysis.

Reads from the SMRT-seq data were aligned to hg19 using the Burrows–Wheeler Aligner MEM with the, pacbio flag (-k17 -W40 -r10 -A1 -B1 -O1 -E1 -L0). Since these data are single-ended, outward-facing read pairs cannot be used; thus, classification of circle junctions depended on split reads. Segments of the genome enriched for circular DNA were discovered by scanning 10-kilobase (kb) windows and calculating the false discovery rate (FDR)-adjusted P value from the Poisson distribution of the randomized reads.

Circle classification.

Genome-wide distribution was calculated by dividing each chromosome into 1-megabase (Mb) bins and overlapping with quality-filtered circles. The number of circle reads overlapping each bin was divided by the total number of circle reads per patient, calculated separately for Circle-seq and WGS data. Genic circles were classified with bedtools v.2.25.0 intersect (http://bedtools.readthedocs.io/) against all protein-coding genes, with gene bodies covered at least 20% being used for downstream analysis. Recurrence across samples was calculated from a high-confidence set of genic circles created from genes with at least four circle-supporting reads covering at least 80% of the shortest transcript. Patients with matched Circle-seq, WGS and RNA-seq ($n = 16$) were used to investigate the relationship between circles, amplification and expression with a focus on circles with genic overlap. Correlation plots were computed per patient based on circle coverage, RNA expression and copy number variation fold change. Concordance between gene expression and circles was discovered by converting normalized read counts to z-scores and correlating with circle coverage across patient samples. For further methods, see the Supplementary Note.

Circle chimerism.

Circle chimerism was evaluated using split reads from Nanopore sequencing ($n = 21$) that either bridged another chromosome or linked to a region separated by at least 4 Mb on the same chromosome. A minimum of 5 reads at a mapping quality (MAPQ) > 30 were required for a region to be considered chimeric; all such regions within the circle length ±500 bp were merged using pgltools v.1.2.0 (http://github.com/billgreenwald/pgltools). The resulting chimeric circles were further used as a secondary metric to evaluate the FDR of clustered tree-shaped rearrangement contacts in the WGS data.

Structural variant detection.

Copy number variation was detected using Control-FREEC[33] v.10.6 with contamination adjustment based on a contamination of 0.4 (that is, samples are 60% tumor), a minimalSubclonePresence of 0.244 and with ASCAT v.4.0.1 using default parameters[33,34]. Regions in the genome with a total copy number ≥9 were considered amplified regions following COSMIC copy number variant definition[20]. Amplifications were intersected with regions of circularization using the bedtools v.2.25.0; circular DNAs identified over these amplified regions were classified as ecDNAs. All remaining circular DNAs were classified as eccDNAs. Structural variation was done on matched tumor/normal genomes using novoBreak v.1.1.3 (ref. 35), SvABA v.1.1.1 (ref. [36]), Delly2 v.0.7.7 (ref. [37]), BRASS v.6.0.5 (https://github.com/cancerit/BRASS) and SMUFIN v.0.9.4 (ref. [38]) using default parameters. From 97 initial neuroblastoma genomes, 4 of them (NBL47, NBL53, NBL54 and NBL61) were excluded from the analysis due to their abnormal high number of breakpoints and amplified regions. The 93 genomes left were analyzed with at least 4 variant callers each. Focusing on interchromosomal rearrangements, merging and filtering of the results from different variant calling algorithms was performed. Filtering for all variants was performed with a Brass Assembly Score (BAS) ≥99 and at least 6 variant-supporting reads with an MAPQ > 60. All rearrangements that did not have a minimum of 6 aligned supporting reads with an MAPQ > 60 at each breakpoint were discarded. For the merging of interchromosomal rearrangements, all results from different variant callers were joined after filtering. Variants with breakpoints within a window of 500 bp where collapsed. Only intrachromosomal variants supported by at least two different callers were included. Two additional samples (NBL49 and NBL50), which had exceptionally high numbers of rearrangements (z-score > 2) were discarded. A 1-Mb genomic region was blacklisted due to its high number of recurrent, visually confirmed false positive breakpoints (z-score > 2 within the 10 highest-ranking bins). Structural variant calls from an independent cohort of WGS data of 546 pediatric cancer genomes was obtained from the DKFZ Pediatric Pan Cancer dataset (https://hgserver1.amc.nl/cgi-bin/r2/main.cgi?&dscope=DKFZ_PED&option=about_dscope). For further methods, see the Supplementary Note.

Regions of clustered rearrangements.

A region of clustered, tree-shaped rearrangement pattern was defined as having three or more interchromosomal rearrangements within a 4-Mb sliding window. The outermost breakpoints defined the boundaries of a cluster region. When five or more interchromosomal rearrangements connected the same two chromosomes, these were flagged and not considered for cluster detection. When 2 or more interchromosomal rearrangements connected 2 regions <10 Mb in size, only one rearrangement was counted for cluster detection. All chromosomes with >25 interchromosomal rearrangements were not considered. All structural variants detected in our dataset, as well as regions of clustered rearrangements detected using the methods described, can be visually inspected in an openly accessible website39. To estimate the FDRs, we randomly redistributed breakpoints of each sample across the mappable genome before counting the number of rearrangements within 4-Mb sliding windows. Five hundred such randomized datasets were created. The FDR was estimated as the mean fraction of rearrangement cluster-positive samples in this randomized dataset. For the chosen threshold of 3 or more rearrangements, the estimated FDR was 0.13. The analysis of circle integration was carried out by detecting the rearrangements connecting a circularized region with a candidate insertion site. Integration sites were defined by two main characteristics: both recipient breakpoints being located on the same chromosome and at a distance between breakpoints smaller than the circularized region inserted. Visual inspection of BAM files was performed for each candidate integration site. For further methods, see the Supplementary Note.

Circle length analysis.

To identify the length preferences for circles depending on the copy number state of the underlying genomic segment, we derived a zero-sum score, following common enrichment test strategies such as gene set enrichment analysis[40,41]. For a given copy number category (balanced, weak imbalance, strong imbalance, LOH and focal amplification), each circle was assigned a score of $1/k$ if the circle belonged to the category and $-1/(n-k)$ otherwise, where k is the total number of circles in that category and n is the total number of circles. Circles were ranked by length and cumulative scores along the list were calculated. The absolute maximum cumulative score was tested against 10,000 random permutations of the ranked list to determine the approximate enrichment P values. For further methods, see the Supplementary Note.

Circle breakpoint analysis.

Base-pair accurate circle junctions were reassembled using SvABA v.1.1.1 with default parameters and only read pairs and split reads containing any outward-facing read orientation as input. Each precise head-to-tail rearrangement call was considered a circle junction. Homology and insertion sequences were taken from the SvABA output directly.

To screen for motifs enriched at circle junction breakpoints, hg19 reference sequences for 41-bp windows around each circle junction breakpoint were obtained. MEME v.5.0.2 (parameters -objfun de -revcomp -nmotifs 5) was used to assess these sequences for motif enrichment with respect to a set of 1 million length-matched sequences randomly sampled from hg19 (excluding poorly or nonassembled regions and the ENCODE DAC blacklist). We compared reference sequence-derived microhomology lengths for actual breakpoints versus a random permutation of breakpoint partners using a two-sided t-test. For further methods, see the Supplementary Note.

Structural variant breakpoint analysis.

Base-pair accurate structural rearrangement calls from the merged structural variant set were considered for detailed breakpoint analysis. The hg19 reference sequence was obtained for a 61-bp window around each breakpoint. MEME v.5.0.2 (parameters -objfun de -revcomp -nmotifs 10) was used to identify motifs that were enriched with regard to a set of 1 million length-matched sequences randomly sampled from hg19 (excluding poorly or nonassembled regions and the ENCODE DAC blacklist). Differential enrichment was equally assessed to compare subsets of rearrangements (clustered rearrangements versus nonclustered rearrangements, circle–circle versus other, circle–genome versus other, genome–genome versus other). Only SvABA rearrangement could be readily analyzed for homology and inserted sequences at breakpoints. We compared reference sequence-derived microhomology lengths for actual breakpoints versus a random permutation of breakpoint partners using a two-sided t-test. For further methods, see the Supplementary Note.

Statistical analysis.

The enrichment of rearrangements in circularization loci was done using a two-sample test for equality of proportions with continuity correction. The enrichment of interchromosomal rearrangement breakpoint clusters within circularized regions was assessed using the union of interchromosomal rearrangements detected by all variant callers and at regions of circularization determined using Circle-seq and WGS separately. The relative overlap of each region of clustered breakpoints with circularized regions in the respective sample was computed. The distribution of overlap was then compared to the distribution expected by chance. For each region of clustered rearrangements, 2,000 random intervals of matching length were randomly positioned over a masked genome that excluded poorly or nonassembled regions and the ENCODE DAC blacklist. The relative overlap of each random interval with circular DNA in the matching patient was then assessed. A hypothesis test was derived from considering the mean relative overlap for the set of observed cluster regions with regard to the distribution of the mean relative overlap for the 2,000 synthetic sets of cluster regions. The one-sided empirical *P* value was calculated and Benjamini–Hochberg-corrected for multiple comparisons (circle classes and circle calling methods). We investigated the distance of distal breakpoints of tree-shaped clustered rearrangements. We tested whether these breakpoints were closer to certain classes of genes than expected by chance. We looked at three gene classes: all COSMIC v.87 genes versus only COSMIC v.87 oncogenes versus only COSMIC v.87 tumor suppressor genes. For each breakpoint, we calculated the distance to the closest gene of the particular gene class and calculated the class-wise median of distances. Each median was assigned a one-tailed *P* value based on the distribution of medians in 500 synthetic datasets with breakpoint positions randomly drawn from the nonblacklisted genome. *P* values were corrected for multiple testing using the Benjamini–Hochberg method. To assess gene expression changes around rearrangement breakpoints, expression of protein-coding genes within 2 Mb of each breakpoint were analyzed. The differential RNA expression of genes in each sample compared to the rest of the cohort was quantified and the modified z-score of their transcripts per million was calculated. Two-sided log-rank tests were used for survival analysis across subgroups. To assess the effect of rearrangement clusters at the *MYCN* amplicon locus, *MYCN*-associated clusters were defined as all clusters that overlapped the ±1-Mb window around the *MYCN*. All violin plots depict the smoothed distribution using a Gaussian kernel with bandwidth selected according to Silverman's rule. The box plots depict the first and third quartiles, segmented by the median; the whiskers depict the points within the 1.5× interquartile range beyond the box edges. All cell culture experiments were conducted at least three independent times, unless otherwise stated. For further details, see the Nature Research Reporting Summary. For further methods, see the Supplementary Note.

Patient samples and clinical data access.

This study comprised the analyses of tumor and blood samples of patients diagnosed with neuroblastoma between 1991 and 2016. Patients were registered and treated according to the trial protocols of the German Society of Pediatric

Oncology and Hematology (GPOH). This study was conducted in accordance with the World Medical Association Declaration of Helsinki (2013) and good clinical practice; informed consent was obtained from all patients or their guardians. The collection and use of patient specimens was approved by the institutional review boards of Charité-Universitätsmedizin Berlin and the Medical Faculty, University of Cologne. Specimens and clinical data were archived and made available by Charité-Universitätsmedizin Berlin or the National Neuroblastoma Biobank and Neuroblastoma Trial Registry (University Children's Hospital Cologne) of the GPOH. The *MYCN* gene copy number was determined as a routine diagnostic method using FISH. DNA and total RNA were isolated from tumor samples with at least 60% tumor cell content as evaluated by a pathologist. For further methods, see the Supplementary Note

Data availability.

The WGS and RNA-seq data that support the findings of this study have been deposited with the European Genome-phenome Archive (https://www.ebi.ac.uk/ega/) under accession nos. EGAS00001001308 and EGAS00001004022. The Circle-seq data that support the findings of this study are available from the corresponding author upon request. Source data for Fig. 1 are available online.

Code availability.

The scripts used to analyze the sequencing data have been uploaded to www.github.com/henssenlab. Data on tree-shaped rearrangements can be accessed and visualized online (https://kons.shinyapps.io/trees/).

4 References

1. Turner, K. M. et al. Extrachromosomal oncogene amplifcation drives tumour evolution and genetic heterogeneity. Nature 543, 122–125 (2017).
2. Møller, H. D. et al. Circular DNA elements of chromosomal origin are common in healthy human somatic tissue. Nat. Commun. 9, 1069 (2018).
3. Shibata, Y. et al. Extrachromosomal microDNAs and chromosomal microdeletions in normal tissues. Science 336, 82–86 (2012).

4. Pennisi, E. Circular DNA throws biologists for a loop. Science 356, 996 (2017).
5. Verhaak, R. G. W., Bafna, V. & Mischel, P. S. Extrachromosomal oncogene amplifcation in tumour pathogenesis and evolution.Nat. Rev. Cancer 19, 283–288 (2019).
6. Møller, H. D., Parsons, L., Jørgensen, T. S., Botstein, D. & Regenberg, B. Extrachromosomal circular DNA is common in yeast. Proc. Natl Acad. Sci. USA 112, E3114–E3122 (2015).
7. Tjio, J. H. & Levan, A. Te chromosome number of man. Hereditas 42, 1–6 (1956).
8. Garsed, D. W. et al. Te architecture and evolution of cancer neochromosomes. Cancer Cell 26, 653–667 (2014).
9. Rausch, T. et al. Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. Cell 148, 59–71 (2012).
10. Kohl, N. E. et al. Transposition and amplifcation of oncogene-related sequences in human neuroblastomas. Cell 35, 359–367 (1983).
11. deCarvalho, A. C. et al. Discordant inheritance of chromosomal and extrachromosomal DNA elements contributes to dynamic disease evolution in glioblastoma. Nat. Genet. 50, 708–717 (2018).
12. Nikolaev, S. et al. Extrachromosomal driver mutations in glioblastoma and low-grade glioma. Nat. Commun. 5, 5690 (2014).
13. Schwab, M. et al. Amplifed DNA with limited homology to myc cellular oncogene is shared by human neuroblastoma cell lines and a neuroblastoma tumour. Nature 305, 245–248 (1983).
14. Balaban-Malenbaum, G. & Gilbert, F. Double minute chromosomes and the homogeneously staining regions in chromosomes of a human neuroblastoma cell line. Science 198, 739–741 (1977).
15. Cox, D., Yuncken, C. & Spriggs, A. I. Minute chromatin bodies in malignant tumours of childhood. Lancet 1, 55–58 (1965).
16. Sanborn, J. Z. et al. Double minute chromosomes in glioblastoma multiforme are revealed by precise reconstruction of oncogenic amplicons. Cancer Res. 73, 6036–6045 (2013).
17. Deshpande, V. et al. Exploring the landscape of focal amplifcations in cancer using AmpliconArchitect. Nat. Commun. 10, 392 (2019).
18. Avet-Loiseau, H. et al. Morphologic and molecular cytogenetics in neuroblastoma. Cancer 75, 1694–1699 (1995).
19. Dillon, L. W. et al. Production of extrachromosomal microDNAs is linked to mismatch repair pathways and transcriptional activity. Cell Rep. 11, 1749–1759 (2015).
20. Forbes, S. A. et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. 45, D777–D783 (2017).
21. Bouzas-Rodriguez, J. et al. Neurotrophin-3 production promotes human neuroblastoma cell survival by inhibiting TrkC-induced apoptosis. J. Clin. Invest. 120, 850–858 (2010).
22. Xu, K. Structure and evolution of double minutes in diagnosis and relapse brain tumors. Acta Neuropathol. 137, 123–137 (2019).
23. Storlazzi, C. T. et al. Gene amplifcation as double minutes or homogeneously staining regions in solid tumors: origin and structure. Genome Res. 20, 1198–1206 (2010).
24. Villamón, E. et al. Genetic instability and intratumoral heterogeneity in neuroblastoma with MYCN amplifcation plus 11q deletion. PLoS ONE 8, e53740 (2013).
25. Marrano, P., Irwin, M. S. & Torner, P. S. Heterogeneity of MYCN amplifcation in neuroblastoma at diagnosis, treatment, relapse, and metastasis. Genes Chromosomes Cancer 56, 28–41 (2017).
26. Gröbner, S. N. et al. Te landscape of genomic alterations across childhood cancers. Nature 555, 321–327 (2018).
27. Northcott, P. A. et al. Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. Nature 511, 428–434 (2014).
28. Henssen, A. G. et al. PGBD5 promotes site-specifc oncogenic mutations in human tumors. Nat. Genet. 49, 1005–1014 (2017).

29. Henssen, A. G. et al. Genomic DNA transposition induced by human PGBD5. eLife 4, e10565 (2015).
30. Henssen, A. G. et al. Forward genetic screen of human transposase genomic rearrangements. BMC Genomics 17, 548 (2016)

5 Authors information

Affiliations

[1]Center for Epigenetics Research, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [2]Barcelona Supercomputing Center, Joint Barcelona Supercomputing Center-Centre for Genomic Regulation-Institute for Research in Biomedicine Research Program in Computational Biology, Barcelona, Spain. [3]Department of Pediatric Oncology/Hematology, Charité-Universitätsmedizin Berlin, Berlin, Germany. [4]Department of Biology, Humboldt University, Berlin, Germany. [5]Berlin Institute of Health, Berlin, Germany. [6]Max Delbrück Center for Molecular Medicine, Berlin, Germany. [7]German Cancer Consortium (DKTK), partner site Berlin, and German Cancer Research Center (DKFZ), Heidelberg, Germany. [8]Experimental and Clinical Research Center, Max Delbrück Center for Molecular Medicine and Charité-Universitätsmedizin Berlin, Berlin, Germany. [9]Department of Experimental Pediatric Oncology, University Children's Hospital of Cologne, Cologne, Germany. [10]Center for Molecular Medicine Cologne, Medical Faculty, University of Cologne, Cologne, Germany. [11]Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain. [12]These authors contributed equally: Richard P. Koche, Elias Rodriguez-Fos, Konstantin Helmsauer, David Torrents, Johannes H. Schulte, Anton G. Henssen.

Contributions

R.P.K., E.R.F., K.H., J.M., F.H., I.C.M., R.C., A.W., M.B., M.P., C. Röefzaad, T.T., H.D., R.F.S., C. Rosswog, J. Theissen, V.B., N.M.P., H.D.G., Y.B., A.S., N. Timme, K.K., S.F., N. Thiessen, E.B., K.S, A.K., P.H., J. Toedling, M.F., D.B., S.S., A.E., D.T., J.H.S. and A.G.H. contributed to the study design and the collection and interpretation of the data. R.P.K. performed the analysis of Circle-seq and WGS. E.R.F. performed the data analysis of the WGS data. I.C.M., R.C., N.M.P. and H.D.G. performed the circular DNA extraction and PCR-based validation. V.B., C. Röefzaad, A.W., P.H., K.S., M.F., A.S. and F.H. collected and prepared the patient samples. C. Rosswog. and J. Theissen performed and analyzed the FISH. M.B. and R.F.S. performed the haplotype phasing analyses. M.P. and J. Toedling. analyzed the tumor genome sequencing data. M.B. performed the allele-specific analysis of Circle-seq, WGS and RNA-seq. S.F., F.K., R.P.K, K.H. and J.M. performed the RNA-seq data analysis. P.H., H.D.G., N.M.P., A.S., D.B. and K.S. performed the experiments and analyzed the data. R.P.K., A.K., A.E. and J.H.S. contributed to the study design. R.P.K. and A.G.H. led the study design, performed the data analysis and wrote the manuscript. All authors contributed to manuscript drafting.

6 Supplementary information

Supplementary figures in the appendix section.

Results and Discussion

The biomedical field has undergone a paradigm shift with the emergence of genomic studies. The combination in cancer research and therapy of different disciplines such as molecular biology, genetics, and bioinformatics, among others, is pushing this field from a classic function-to-genotype perspective to a genomics-to-function approach. Next-generation sequencing technologies have contributed to this shift allowing us to obtain the genomic profile of the patient. This is especially relevant in cancer treatment where traditionally, two patients with the same tumor type were treated the same way regardless of their genotype. However, this scenario is changing with the inclusion of genomic and mutational profiling of patients, which helps to understand and treat the disease with greater precision. Personalized or precision medicine[107], take into account the profile of the genetic variants of the patient to guide the selection of treatment, in order to minimize harmful effects and maximize the favorable outcomes.

Following the goal of precision medicine, a better understanding and characterization of the mutational patterns of the tumor is needed. By identifying novel rearrangements associated with cancer, we expand our knowledge about tumor development and help group the patients for further personalized treatment. In this line, this thesis represents a contribution to the task of characterizing new patterns of genomic rearrangements in cancer. Centering this work in three studies, we have been able not only to identify recurrent patterns of structural variation in tumors but also to propose potential associated mechanisms. Moreover, another particular aspect of this thesis is the analysis of variants and elements of the human genome that are generally omitted or present a challenge in cancer studies.

In summary, through the analysis of *PGBD5*-transformed cell lines, we have identified small deletions associated with specific sequences or motifs, and *Alu* elements flanking the breakpoints. The same motifs have been found in rhabdoid tumors, which expressed this gene, supporting the association between PGBD5 and the generation of specific rearrangements in the human genome, notably in cancer. In this sense, our work makes a contribution to the long-standing question about the role of transposase-derived genes in cancer.

The exploration of the deletions related to the PGBD5 specific motif in ICGC-PCAWG data revealed different classes of motif-related rearrangements sharing the same characteristics. It also revealed an unexpectedly high recurrence of these somatic deletions across patients and tumor types. The exploration of this recurrence resulted in the possibility that a subset of motif-related deletions were indeed artifacts generated from methodological processes. The more interesting aspect of the Pan-Cancer analysis is not our results *per se*, but the questions they raise. All things considered, small-sized somatic structural variants involving repetitive regions such as *Alus* are present in our genome and represent a portion of our genetic variation that has, as in the case of PGBD5, been associated with cancer[61,63-65]. Although the study of this type of mutations represents a challenge, we have to think about how we address them, rather than just omit these variants from our analyses. From this work, we learned that not taking these rearrangements into account excludes a fraction of the mutational profile of the tumor, while taking them into account without further validation could have serious consequences such as defining artifact-related deletions as real.

On another note, in the study of structural variants associated with extrachromosomal circular DNA elements in neuroblastoma, we have been able to identify a recurrent pattern of translocations related to circularized regions of the genome. These rearrangements describe the interaction between linear chromosomes and extrachromosomal circular elements through a mechanism of re-integration of circular DNA, and the interaction between different circular DNA elements through a mechanism of chimeric circle formation. Interestingly, we have reported the functional and clinical impact of the re-integration of circles, which has been associated with poor patient outcome. These findings indicate that extrachromosomal circular DNA elements have a genome remodeling role in neuroblastoma. In this line, knowing that circular DNA structures have been detected in at least 40% of cancers[90], we can consider that these elements could play a similar role in different tumor types. With that in mind, if we want to gain more insight about oncogenic processes, we cannot further omit the extrachromosomal circular fraction of our genome in cancer studies. Not only because we are missing a part of the picture but because we can also be misclassifying genomic changes associated with circular DNA.

This idea became evident in the first circular DNA conference held in Berlin, which I have been lucky to attend, and in which we saw the emergence of a new field in genomics, based in the study of these elements. Traditionally circular DNA has been examined in other organisms such as yeast, but in the last years, more cancer studies have been focusing their analysis on those DNA structures. It seems kind of obvious, but the more we include these elements in cancer studies, the more knowledge about the genetic variation associated with them and, by extension, with oncogenic processes we will have.

Overall, this thesis also illustrates the necessary collaboration between computational and experimental groups, notably in cancer research. An increasingly common situation in the field as a consequence of the change of paradigm explained above. On our side, the resources we have at the Barcelona Supercomputing Center (BSC), and the expertise of our group in the detection of structural variants, allowed us to analyze the genomic data presented in this thesis. However, without the collection and sequencing of this data and the experimental validation of our results together with the functional and clinical studies, it would not be possible to achieve all the findings presented here.

Conclusions

In this thesis, we have examined and described patterns of structural variation associated with mutagenic processes in cancer through the analysis of sequencing data from transformed cell lines and different cancer patients, such as neuroblastoma. Based on this work, we have come to the following conclusions:

1. The expression of *PGBD5* in transformed cell lines is associated with the generation of recurrent rearrangements.
2. The rearrangements associated with PGBD5 correspond to small deletions presenting 183bp average length, a specific microhomology motif around the breakpoints and a pattern of confronted *Alus* flanking the deletion site.
3. Extending this analysis to 37 tumor types confirmed the presence of motif-related deletions, similar to the ones associated with PGBD5, with a frequency beyond expectation.
4. The recurrence levels identified within the motif-related deletions do not match with purely somatic rearrangements.
5. A detailed analysis of this recurrence shows potential methodological artifacts (i.e., PCR) behind a fraction of these rearrangements.
6. The analysis of structural variation in neuroblastoma reveals a recurrent pattern of clustered translocations, the origin of which coincide with circularized regions of the genome such as *MYCN* loci.
7. The pattern of clustered translocations describes mechanisms of chimeric circle formation and re-integration of circles into the linear genome, indicating that extrachromosomal circular DNA actively contributes to genome remodeling in neuroblastoma.
8. The presence of these patterns is associated with functional and clinical consequences, especially with poor patient outcome, pointing to his potential use as a marker of disease prognosis.

Bibliography

1. Ritchie, H. & Roser, R. Causes of Death, <https://ourworldindata.org/causes-of-death> (2019).
2. Bray, F. et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 68, 394-424, doi:10.3322/caac.21492 (2018).
3. Stewart, B. W., Wild, C., International Agency for Research on Cancer & World Health Organization. World cancer report 2014. (International Agency for Research on Cancer WHO Press, 2014).
4. Branco, A. T. The evolution of genetics to genomics. Journal of Human Growth and Development 26(1), 28-32, doi:https://dx.doi.org/10.7322/jhgd.113710 (2016).
5. Mardis, E. R. The impact of next-generation sequencing technology on genetics. Trends Genet 24, 133-141, doi:10.1016/j.tig.2007.12.007 (2008).
6. Mardis, E. R. The Impact of Next-Generation Sequencing on Cancer Genomics: From Discovery to Clinic. Cold Spring Harb Perspect Med 9, doi:10.1101/cshperspect.a036269 (2019).
7. Gayon, J. From Mendel to epigenetics: History of genetics. C R Biol 339, 225-230, doi:10.1016/j.crvi.2016.05.009 (2016).
8. Miko, I. Gregor Mendel and the Principles of Inheritance. Nature Education 1(1):134 (2008).
9. Pray, L. Discovery of DNA Structure and Function: Watson and Crick. Nature Education 1(1):100 (2008).
10. Pearson, H. Genetics: what is a gene? Nature 441, 398-401, doi:10.1038/441398a (2006).
11. Chargaff, E. Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. Experientia 6, 201-209, doi:10.1007/bf02173653 (1950).
12. Watson, J. D. & Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 171, 737-738, doi:10.1038/171737a0 (1953).
13. O'Connor, C. Isolating hereditary material: Frederick Griffith, Oswald Avery, Alfred Hershey, and Martha Chase. Nature Education 1(1):105 (2008).
14. Morange, M. R. The Central Dogma of molecular biology. A retrospective after fifty years. Resonance 14, 236–247, doi:https://doi.org/10.1007/s12045-009-0024-6 (2009).
15. International Human Genome Sequencing, C. Finishing the euchromatic sequence of the human genome. Nature 431, 931-945, doi:10.1038/nature03001 (2004).
16. Lander, E. S. et al. Initial sequencing and analysis of the human genome. Nature 409, 860-921, doi:10.1038/35057062 (2001).
17. Tyler-Smith, C. et al. Where Next for Genetics and Genomics? PLoS Biol 13, e1002216, doi:10.1371/journal.pbio.1002216 (2015).
18. Richards, J. & Hawley, R. S. in The Human Genome (ed Academic Press) 416 (2010).
19. Perbal, L. The case of the gene: Postgenomics between modernity and postmodernity. EMBO Rep 16, 777-781, doi:10.15252/embr.201540179 (2015).
20. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).
21. Adams, J. DNA sequencing technologies. Nature Education (2008).
22. Heather, J. M. & Chain, B. The sequence of sequencers: The history of sequencing DNA. Genomics 107, 1-8, doi:10.1016/j.ygeno.2015.11.003 (2016).
23. Chial, H. DNA sequencing technologies key to the Human Genome Project. Nature Education (2008).
24. van Dijk, E. L., Auger, H., Jaszczyszyn, Y. & Thermes, C. Ten years of next-generation sequencing technology. Trends Genet 30, 418-426, doi:10.1016/j.tig.2014.07.001 (2014).
25. Metzker, M. L. Sequencing technologies - the next generation. Nat Rev Genet 11, 31-46, doi:10.1038/nrg2626 (2010).
26. Ulahannan, D., Kovac, M. B., Mulholland, P.J., Cazier, J. B. & Tomlinson, I. Technical and implementation issues in using next-generation sequencing of cancers in clinical practice. Br J Cancer 109, 827-835, doi:10.1038/bjc.2013.416 (2013).

27. Sboner, A., Mu, X. J., Greenbaum, D., Auerbach, R. K. & Gerstein, M. B. The real cost of sequencing: higher than you think! Genome Biol 12, 125, doi:10.1186/gb-2011-12-8-125 (2011).
28. Consortium, I. T. P.-C. A. o. W. G. Pan-cancer analysis of whole genomes. Nature 578, 82-93, doi:10.1038/s41586-020-1969-6 (2020).
29. Meyerson, M., Gabriel, S. & Getz, G. Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet 11, 685-696, doi:10.1038/nrg2841 (2010).
30. Bao, S. et al. Evaluation of next-generation sequencing software in mapping and assembly. J Hum Genet 56, 406-414, doi:10.1038/jhg.2011.43 (2011).
31. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. Nature 458, 719-724, doi:10.1038/nature07943 (2009).
32. Nowell, P. C. The clonal evolution of tumor cell populations. Science 194, 23-28, doi:10.1126/science.959840 (1976).
33. Greaves, M. & Maley, C. C. Clonal evolution in cancer. Nature 481, 306-313, doi:10.1038/nature10762 (2012).
34. Cell editorial, t. Cancer: The Road Ahead. Cell 168, 545-546, doi:10.1016/j.cell.2017.01.036 (2017).
35. Rosenberg, E. in It's in Your DNA, From Discovery to Structure, Function and Role in Evolution, Cancer and Aging (ed E. Rosenberg) Ch. 11, 95-104 (Academic Press, 2017).
36. Moller, H. D. et al. Circular DNA elements of chromosomal origin are common in healthy human somatic tissue. Nat Commun 9, 1069, doi:10.1038/s41467-018-03369-8 (2018).
37. Shen, Z. Genomic instability and cancer: an introduction. J Mol Cell Biol 3, 1-3, doi:10.1093/jmcb/mjq057 (2011).
38. Martincorena, I. & Campbell, P. J. Somatic mutation in cancer and normal cells. Science 349, 1483-1489, doi:10.1126/science.aab4082 (2015).
39. Martincorena, I. et al. Universal Patterns of Selection in Cancer and Somatic Tissues. Cell 173, 1823, doi:10.1016/j.cell.2018.06.001 (2018).
40. McFarland, C. D. et al. The Damaging Effect of Passenger Mutations on Cancer Progression. Cancer Res 77, 4763-4772, doi:10.1158/0008-5472.CAN-15-3283-T (2017).
41. Dagogo-Jack, I. & Shaw, A. T. Tumour heterogeneity and resistance to cancer therapies. Nat Rev Clin Oncol 15, 81-94, doi:10.1038/nrclinonc.2017.166 (2018).
42. Liu, J., Dang, H. & Wang, X. W. The significance of intertumor and intratumor heterogeneity in liver cancer. Exp Mol Med 50, e416, doi:10.1038/emm.2017.165 (2018).
43. Jamal-Hanjani, M. et al. Tracking the Evolution of Non-Small-Cell Lung Cancer. N Engl J Med 376, 2109-2121, doi:10.1056/NEJMoa1616288 (2017).
44. Tubio, J. M. Somatic structural variation and cancer. Brief Funct Genomics 14, 339-351, doi:10.1093/bfgp/elv016 (2015).
45. Tate, J. G. et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res 47, D941-D947, doi:10.1093/nar/gky1015 (2019).
46. Carvalho, C. M. & Lupski, J. R. Mechanisms underlying structural variant formation in genomic disorders. Nat Rev Genet 17, 224-238, doi:10.1038/nrg.2015.25 (2016).
47. Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J. O. Phenotypic impact of genomic structural variation: insights from and for human disease. Nat Rev Genet 14, 125-138, doi:10.1038/nrg3373 (2013).
48. Weckselblatt, B. & Rudd, M. K. Human Structural Variation: Mechanisms of Chromosome Rearrangements. Trends Genet 31, 587-599, doi:10.1016/j.tig.2015.05.010 (2015).
49. Yi, K. & Ju, Y. S. Patterns and mechanisms of structural variations in human cancer. Exp Mol Med 50, 98, doi:10.1038/s12276-018-0112-3 (2018).
50. Tattini, L., D'Aurizio, R. & Magi, A. Detection of Genomic Structural Variants from Next-Generation Sequencing Data. Front Bioeng Biotechnol 3, 92, doi:10.3389/fbioe.2015.00092 (2015).

51. Ewing, A. & Semple, C. Breaking point: the genesis and impact of structural variation in tumours. F1000Res 7, doi:10.12688/f1000research.16079.1 (2018).
52. Watson, I. R., Takahashi, K., Futreal, P. A. & Chin, L. Emerging patterns of somatic mutations in cancer. Nat Rev Genet 14, 703-718, doi:10.1038/nrg3539 (2013).
53. Yang, L. et al. Diverse mechanisms of somatic structural variations in human cancer genomes. Cell 153, 919-929, doi:10.1016/j.cell.2013.04.010 (2013).
54. Trenner, A. & Sartori, A. A. Harnessing DNA Double-Strand Break Repair for Cancer Treatment. Front Oncol 9, 1388, doi:10.3389/fonc.2019.01388 (2019).
55. Jeggo, P. A., Pearl, L. H. & Carr, A. M. DNA repair, genome stability and cancer: a historical perspective. Nat Rev Cancer 16, 35-42, doi:10.1038/nrc.2015.4 (2016).
56. Currall, B. B., Chiang, C., Talkowski, M. E. & Morton, C. C. Mechanisms for Structural Variation in the Human Genome. Curr Genet Med Rep 1, 81-90, doi:10.1007/s40142-013-0012-8 (2013).
57. Escaramis, G., Docampo, E. & Rabionet, R. A decade of structural variants: description, history and methods to detect structural variation. Brief Funct Genomics 14, 305-314, doi:10.1093/bfgp/elv014 (2015).
58. de Koning, A. P., Gu, W., Castoe, T. A., Batzer, M. A. & Pollock, D. D. Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet 7, e1002384, doi:10.1371/journal.pgen.1002384 (2011).
59. Scarfo, I., Pellegrino, E., Mereu, E., Inghirami, G. & Piva, R. Transposable elements: The enemies within. Exp Hematol 44, 913-916, doi:10.1016/j.exphem.2016.06.251 (2016).
60. Xing, J. et al. Mobile elements create structural variation: analysis of a complete human genome. Genome Res 19, 1516-1526, doi:10.1101/gr.091827.109 (2009).
61. Belancio, V. P., Roy-Engel, A. M. & Deininger, P. L. All y'all need to know 'bout retroelements in cancer. Semin Cancer Biol 20, 200-210, doi:10.1016/j.semcancer.2010.06.001 (2010).
62. Henssen, A. G. et al. Genomic DNA transposition induced by human PGBD5. Elife 4, doi:10.7554/eLife.10565 (2015).
63. Payer, L. M. & Burns, K. H. Transposable elements in human genetic disease. Nat Rev Genet 20, 760-772, doi:10.1038/s41576-019-0165-8 (2019).
64. Bose, P., Hermetz, K. E., Conneely, K. N. & Rudd, M. K. Tandem repeats and G-rich sequences are enriched at human CNV breakpoints. PLoS One 9, e101607, doi:10.1371/journal.pone.0101607 (2014).
65. White, T. B., Morales, M. E. & Deininger, P. L. Alu elements and DNA double-strand break repair. Mob Genet Elements 5, 81-85, doi:10.1080/2159256X.2015.1093067 (2015).
66. Morales, M. E., Servant, G., Ade, C. M. & Deininger, P. in Human Retrotransposons in Health and Disease (ed Gael Cristofari) 239-257 (Springer International Publishing, 2017).
67. Reddy, R. et al. The genomic architecture of NLRP7 is Alu rich and predisposes to disease-associated large deletions. Eur J Hum Genet 24, 1516, doi:10.1038/ejhg.2016.96 (2016).
68. Puente, X. S. et al. Non-coding recurrent mutations in chronic lymphocytic leukaemia. Nature 526, 519-524, doi:10.1038/nature14666 (2015).
69. Guan, P. & Sung, W. K. Structural variation detection using next-generation sequencing data: A comparative technical review. Methods 102, 36-49, doi:10.1016/j.ymeth.2016.01.020 (2016).
70. Moncunill, V. et al. Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads. Nat Biotechnol 32, 1106-1112, doi:10.1038/nbt.3027 (2014).
71. Van Loo, P. et al. Allele-specific copy number analysis of tumors. Proc Natl Acad Sci U S A 107, 16910-16915, doi:10.1073/pnas.1009843107 (2010).
72. Wala, J. A. et al. SvABA: genome-wide detection of structural variants and indels by local assembly. Genome Res 28, 581-591, doi:10.1101/gr.221028.117 (2018).

73. BRASS, <https://github.com/cancerit/BRASS> (

74. Rausch, T. et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 28, i333-i339, doi:10.1093/bioinformatics/bts378 (2012).

75. Ho, S. S., Urban, A. E. & Mills, R. E. Structural variation in the sequencing era. Nat Rev Genet, doi:10.1038/s41576-019-0180-9 (2019).

76. Al-Hajj, M. & Clarke, M. F. Self-renewal and solid tumor stem cells. Oncogene 23, 7274-7282, doi:10.1038/sj.onc.1207947 (2004).

77. Reya, T., Morrison, S. J., Clarke, M. F. & Weissman, I. L. Stem cells, cancer, and cancer stem cells. Nature 414, 105-111, doi:10.1038/35102167 (2001).

78. Feitelson, M. A. et al. Sustained proliferation in cancer: Mechanisms and novel therapeutic targets. Semin Cancer Biol 35 Suppl, S25-S54, doi:10.1016/j.semcancer.2015.02.006 (2015).

79. Ibragimova, M. K., Tsyganov, M. M. & Litviakov, N. V. Natural and Chemotherapy-Induced Clonal Evolution of Tumors. Biochemistry (Mosc) 82, 413-425, doi:10.1134/S0006297917040022 (2017).

80. Institute, W. S. Cancer Genome Project | Cancer Genetics & Genomics, <https://www.sanger.ac.uk/science/groups/cancer-genome-project> (

81. Cancer Genome Atlas Research, N. et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 45, 1113-1120, doi:10.1038/ng.2764 (2013).

82. International Cancer Genome, C. et al. International network of cancer genome projects. Nature 464, 993-998, doi:10.1038/nature08987 (2010).

83. Consortium, I. C. G. ICGC data portal, <https://dcc.icgc.org/> (2019).

84. Hotta, Y. & Bassel, A. Molecular Size and Circularity of DNA in Cells of Mammals and Higher Plants. Proc Natl Acad Sci U S A 53, 356-362, doi:10.1073/pnas.53.2.356 (1965).

85. Pennisi, E. Circular DNA throws biologists for a loop. Science 356, 996, doi:10.1126/science.356.6342.996 (2017).

86. Tandon, I., Pal, R., Pal, J. K. & Sharma, N. K. Extrachromosomal circular DNAs: an extra piece of evidence to depict tumor heterogeneity. Future Sci OA 5, FSO390, doi:10.2144/fsoa-2019-0024 (2019).

87. Verhaak, R. G. W., Bafna, V. & Mischel, P. S. Extrachromosomal oncogene amplification in tumour pathogenesis and evolution. Nat Rev Cancer 19, 283-288, doi:10.1038/s41568-019-0128-6 (2019).

88. Paulsen, T., Kumar, P., Koseoglu, M. M. & Dutta, A. Discoveries of Extrachromosomal Circles of DNA in Normal and Tumor Cells. Trends Genet 34, 270-278, doi:10.1016/j.tig.2017.12.010 (2018).

89. Vogt, N. et al. Amplicon rearrangements during the extrachromosomal and intrachromosomal amplification process in a glioma. Nucleic Acids Res 42, 13194-13205, doi:10.1093/nar/gku1101 (2014).

90. Turner, K. M. et al. Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. Nature 543, 122-125, doi:10.1038/nature21356 (2017).

91. Wu, S. et al. Circular ecDNA promotes accessible chromatin and high oncogene expression. Nature 575, 699-703, doi:10.1038/s41586-019-1763-5 (2019).

92. Deshpande, V. et al. Reconstructing and characterizing focal amplifications in cancer using AmpliconArchitect. bioRxiv, 457333, doi:10.1101/457333 (2018).

93. Capasso, M. & Diskin, S. J. Genetics and genomics of neuroblastoma. Cancer Treat Res 155, 65-84, doi:10.1007/978-1-4419-6033-7_4 (2010).

94. Brodeur, G. M. Neuroblastoma: biological insights into a clinical enigma. Nat Rev Cancer 3, 203-216, doi:10.1038/nrc1014 (2003).

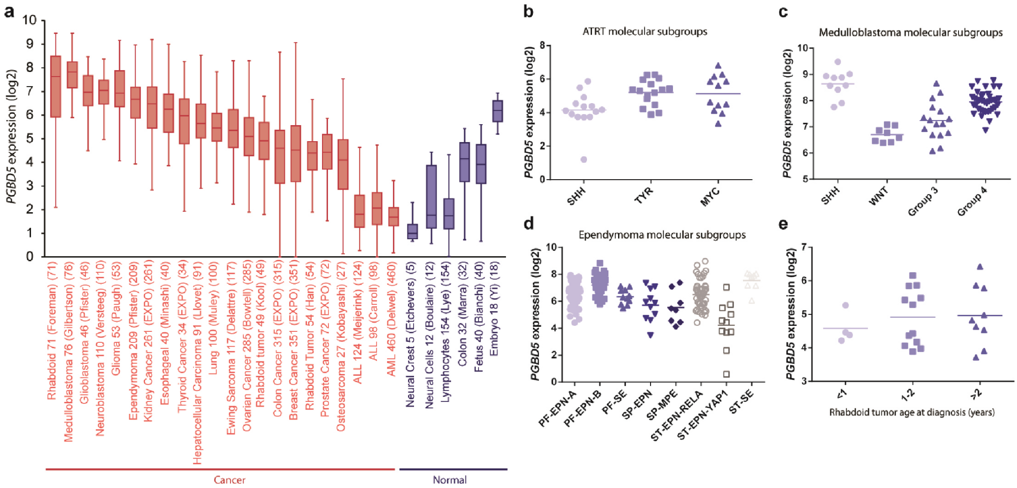
95. Bown, N. Neuroblastoma tumour genetics: clinical and biological aspects. J Clin Pathol 54, 897-910, doi:10.1136/jcp.54.12.897 (2001).

96. DuBois, S. G. et al. Metastatic sites in stage IV and IVS neuroblastoma correlate with age, tumor biology, and survival. J Pediatr Hematol Oncol 21, 181-189, doi:10.1097/00043426-199905000-00005 (1999).

97. Schwab, M. et al. Amplified DNA with limited homology to myc cellular oncogene is shared by human neuroblastoma cell lines and a neuroblastoma tumour. Nature 305, 245-248, doi:10.1038/305245a0 (1983).
98. Chow, A. Y. Cell Cycle Control by Oncogenes and Tumor Suppressors: Driving the Transformation of Normal Cells into Cancerous Cells. Nature Education (2010).
99. Speleman, F., De Preter, K. & Vandesompele, J. Neuroblastoma genetics and phenotype: a tale of heterogeneity. Semin Cancer Biol 21, 238-244, doi:10.1016/j.semcancer.2011.07.003 (2011).
100. Pugh, T. J. et al. The genetic landscape of high-risk neuroblastoma. Nat Genet 45, 279-284, doi:10.1038/ng.2529 (2013).
101. Scott, D. K. et al. The neuroblastoma amplified gene, NAG: genomic structure and characterisation of the 7.3 kb transcript predominantly expressed in neuroblastoma. Gene 307, 1-11, doi:10.1016/s0378-1119(03)00459-1 (2003).
102. Wimmer, K. et al. Co-amplification of a novel gene, NAG, with the N-myc gene in neuroblastoma. Oncogene 18, 233-238, doi:10.1038/sj.onc.1202287 (1999).
103. Momand, J., Jung, D., Wilczynski, S. & Niland, J. The MDM2 gene amplification database. Nucleic Acids Res 26, 3453-3459, doi:10.1093/nar/26.15.3453 (1998).
104. Schwab, M. Oncogene amplification in solid tumors. Semin Cancer Biol 9, 319-325, doi:10.1006/scbi.1999.0126 (1999).
105. Peifer, M. et al. Telomerase activation by genomic rearrangements in high-risk neuroblastoma. Nature 526, 700-704, doi:10.1038/nature14980 (2015).
106. Valentijn, L. J. et al. TERT rearrangements are frequent in neuroblastoma and identify aggressive tumors. Nat Genet 47, 1411-1414, doi:10.1038/ng.3438 (2015).
107. Vogenberg, F. R., Isaacson Barash, C. & Pursel, M. Personalized medicine: part 1: evolution and development into theranostics. P T 35, 560-576 (2010).
108. Stratton, M. R. Journeys into the genome of cancer cells. EMBO Molecular Medicine 5, 169-172, doi:10.1002/emmm.201202388 (2013).
109. Vu, G. T. et al. Repair of Site-Specific DNA Double-Strand Breaks in Barley Occurs via Diverse Pathways Primarily Involving the Sister Chromatid. Plant Cell 26, 2156-2167, doi:10.1105/tpc.114.126607 (2014).

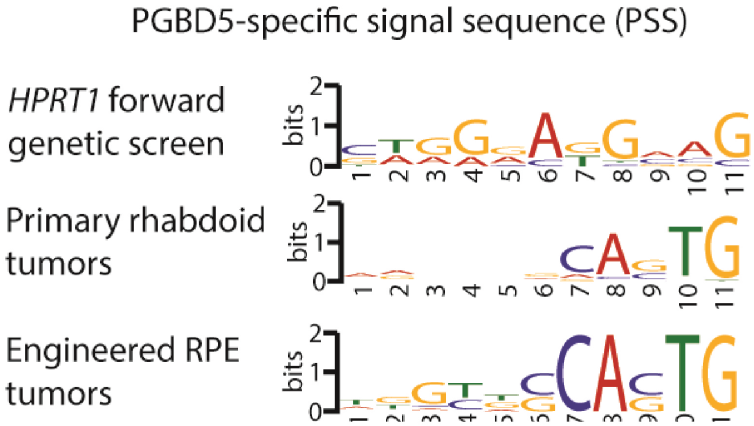
Appendix

Appendix 1
Supplementary figures of the PGBD5 publication



Supplementary Figure 1. *PGBD5* is highly expressed in rhabdoid and other pediatric and childhood solid tumors.

(**a**), Bar graph showing relative expression of *PGBD5* in tumors (red), as compared to normal tissues (blue). Median expression is indicated by horizontal line, boxes indicate 25% and 75% quartiles; whiskers indicate minimum and maximum values. (**b**), Dot plot showing the relative *PGBD5* mRNA expression in atypical teratoid/rhabdoid tumor (ATRT) molecular subgroups (SHH, Sonic hedgehog pathway activation; TYR, tyrosinase overexpression; *MYC*, *MYC* and *HOX* overexpression). Bars denote mean. (**c**), Dot plot showing the relative *PGBD5* mRNA expression in medulloblastoma tumor molecular subgroups. (**d**), Dot plot showing the relative *PGBD5* mRNA expression in ependymoma tumor molecular subgroups. (**e**), Dot plot showing the relative *PGBD5* mRNA expression in ATRT tumors relative to the age of patients at diagnosis. Bars denote mean.


PGBD5-specific signal sequence (PSS)
HPRT1 forward genetic screen
Primary rhabdoid tumors
Engineered RPE tumors
bits
0
1
2
1
2
3
4
5
6
7
8
9
10
11

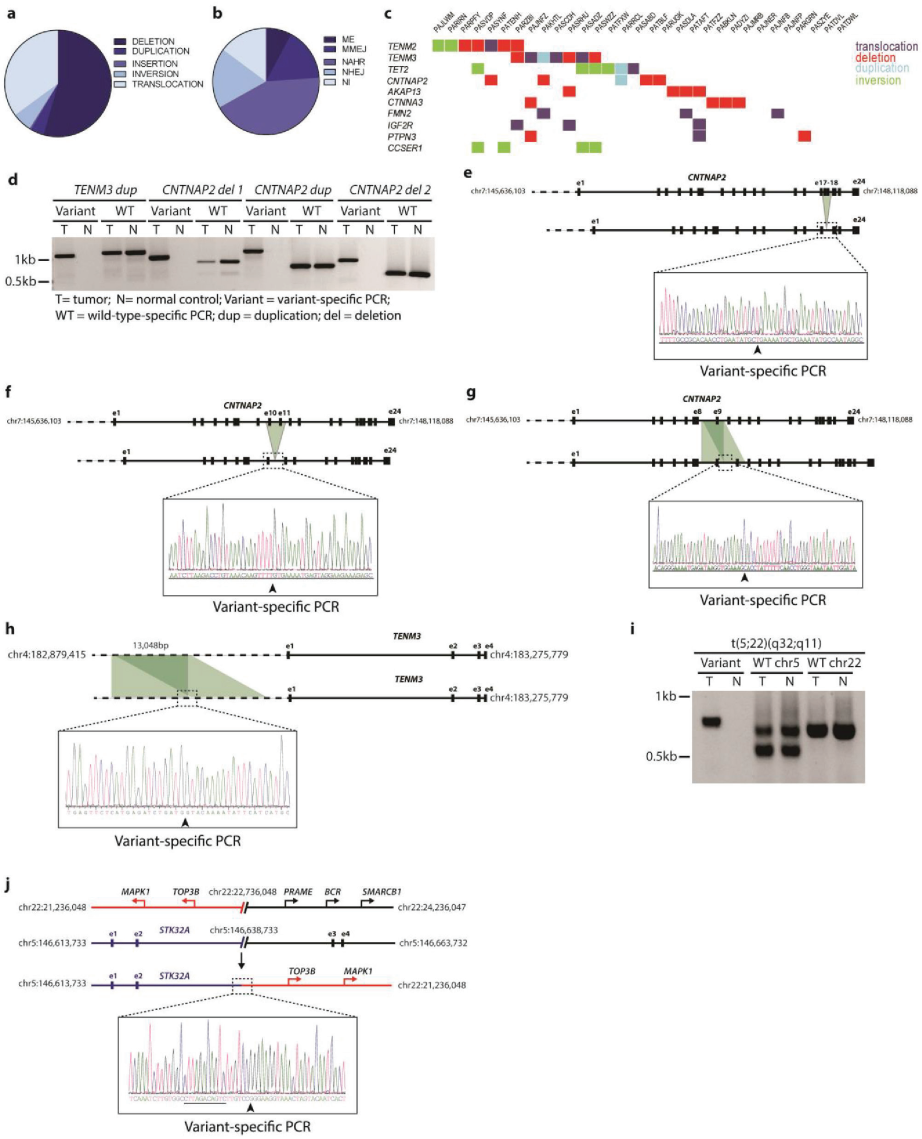
Supplementary Figure 2. PGBD5-specific signal sequences..

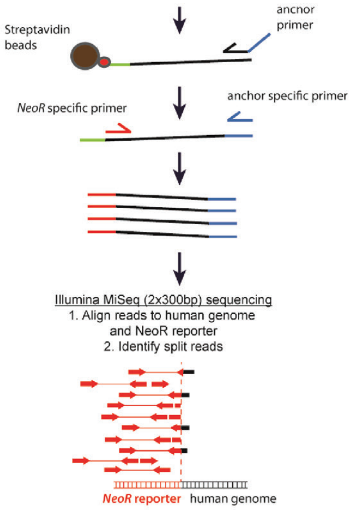
Sequence logos detected near the breakpoints of genomic rearrangements in the *HPRT1* forward genetic screen (top)[32], as compared to those observed in primary rhabdoid (middle) and engineered RPE cell tumors (bottom).

Supplementary Figure 3. Distribution and structure of somatic genomic rearrangements in primary rhabdoid tumors.

(**a**), Distribution of somatic deletions, duplications, insertions, inversions and translocations observed in 31 primary rhabdoid tumors. (**b**), Distribution of predicted mechanisms at the rearrangement breakpoints as homologous recombination (HR), microhomology-mediated end joining (MMEJ), mobile element rearrangements (ME), and non-template insertions (NI). (**c**), Tile plot showing recurrence of somatic translocations (blue), deletions (red), duplications (light blue), and inversions (green) affecting specific genes, excluding *SMARCB1*, in individual rhabdoid tumor specimens. (**d**) Validation of specific somatic rearrangements of *TENM3* and *CNTNAP2* genes, as assessed using variant and wild-type allele-specific PCR in matched tumor and normal primary patient specimens. (**e-h**), Schematics of gene structure of *CNTNAP2* and *TENM3* before and after rearrangements, and Sanger DNA sequencing chromatograms of the individual rearrangement breakpoints detected by variant allele-specific PCR in individual primary rhabdoid tumor specimens (arrowheads mark the breakpoints). (**i**), Validation of t(5;22) translocation using variant and allele-specific PCR. (**j**), Schematic of the chromosomes 5 and 22 before and after rearrangement, leading to the translocation breakpoint detected by variant allele-specific PCR (arrowhead marks the breakpoint).

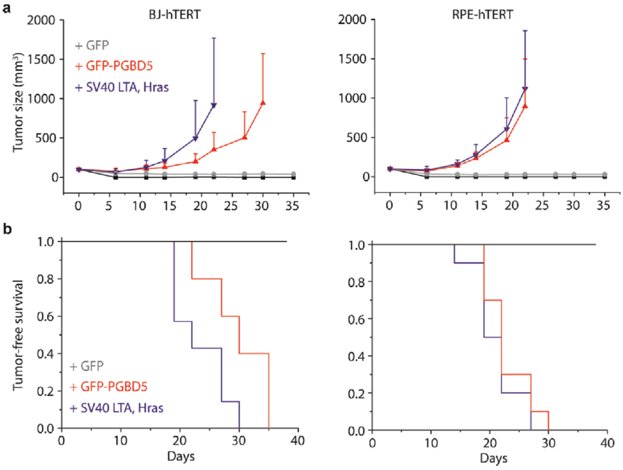
(next page)


a
DELETION
DUPLICATION
INSERTION
INVERSION
TRANSLOCATION
b
ME
MMEJ
NAHR
NHEJ
NI
c
PAJLWM PARIRN PARPFY PASVDP PASYNF PATENH PARZBI PAJNFZ PAKHTL PASCDH PASRHU PASADZ PASWZZ PATFXW PARRCL PASABD PATBLF PARUGK PASDLA PATAFT PATFZZ PABKLN PADYZI PAJMRB PAJNER PAJNFB PAJNFP PARGRN PASZYE PATDVL PATDWL
TENM2
TENM3
TET2
CNTNAP2
AKAP13
CTNNA3
FMN2
IGF2R
PTPN3
CCSER1
translocation
deletion
duplication
inversion
d
TENM3 dup
CNTNAP2 del 1
CNTNAP2 dup
CNTNAP2 del 2
Variant
WT
T N
1kb
0.5kb
T= tumor; N= normal control; Variant = variant-specific PCR;
WT = wild-type-specific PCR; dup = duplication; del = deletion
e
CNTNAP2
chr7:145,636,103
chr7:148,118,088
e1
e17-18
e24
Variant-specific PCR
f
CNTNAP2
chr7:145,636,103
chr7:148,118,088
e1
e10 e11
e24
Variant-specific PCR
g
CNTNAP2
chr7:145,636,103
chr7:148,118,088
e1
e8 e9
e24
Variant-specific PCR
h
chr4:182,879,415
13,048bp
TENM3
e1
e2
e3 e4
chr4:183,275,779
Variant-specific PCR
i
t(5;22)(q32;q11)
Variant
WT chr5
WT chr22
T N
1kb
0.5kb
j
chr22:21,236,048
MAPK1
TOP3B
chr22:22,736,048
PRAME
BCR
SMARCB1
chr22:24,236,047
chr5:146,613,733
e1 e2
STK32A
chr5:146,638,733
e3 e4
chr5:146,663,732
TOP3B
MAPK1
chr22:21,236,048
Variant-specific PCR



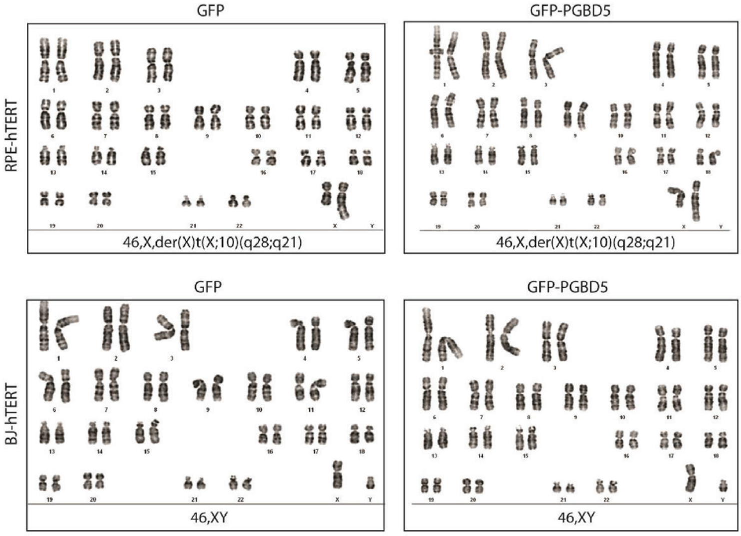
Supplementary Figure 4. Schematic of flanking-sequence exponential anchored polymerase chain reaction (FLEA PCR).

Biotinylated primer specific for the *NeoR* cassette is used for linear extension, followed by streptavidin purification, and nested PCR to amplify integration breakpoints, followed by DNA sequencing.



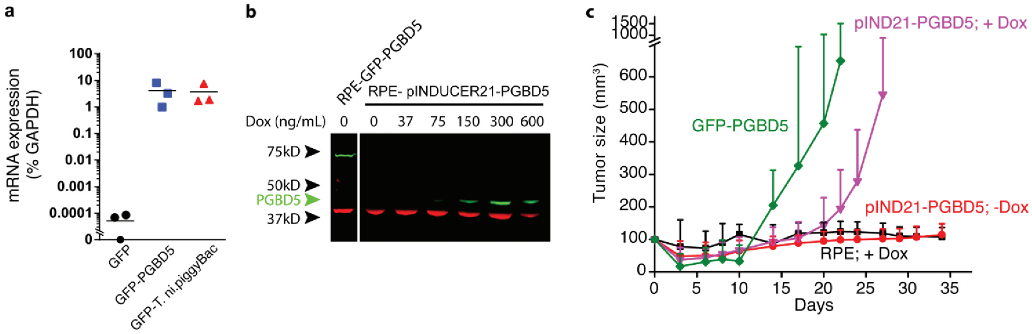
Supplementary Figure 5. Ectopic expression of PGBD5 transforms immortalized BJ and RPE cells in vivo.

(**a**) Tumor volume as a function of time of RPE (right) and BJ cells (left) stably expressing *GFP-PGBD5* and *GFP* only, compared to non-transduced cells and cells expressing SV40 large T antigen (LTA) and *HRAS* (n = 10 per group). (**b**) Kaplan-Meier analysis of tumor-free survival of mice with subcutaneous xenografts of RPE and BJ cells expressing *GFP-PGBD5* or *GFP* only, as compared to non-transduced cells or cells expressing SV40 LTA and HRAS (n = 10 per group, P < 0.0001 by log-rank test).



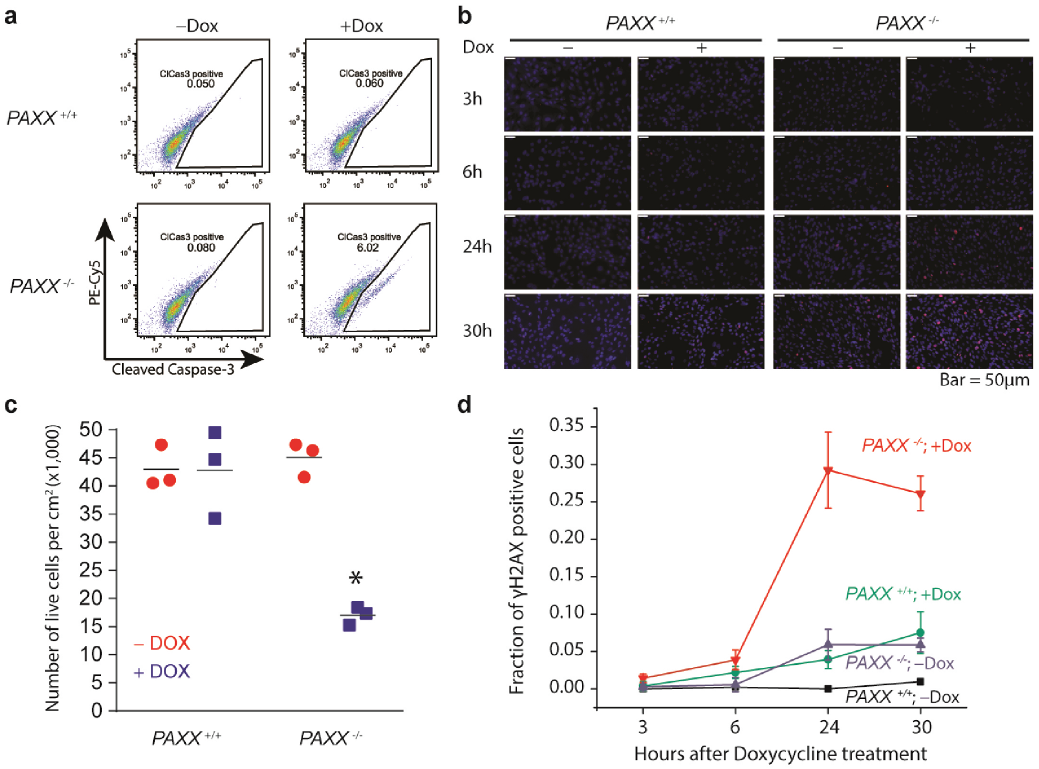
Supplementary Figure 6. *GFP-PGBD5* expression does not induce global chromosomal instability.

Representative karyotype of BJ (lower panel) and RPE cells (upper panel) stably expressing *GFP-PGBD5* (right) and *GFP* (left).



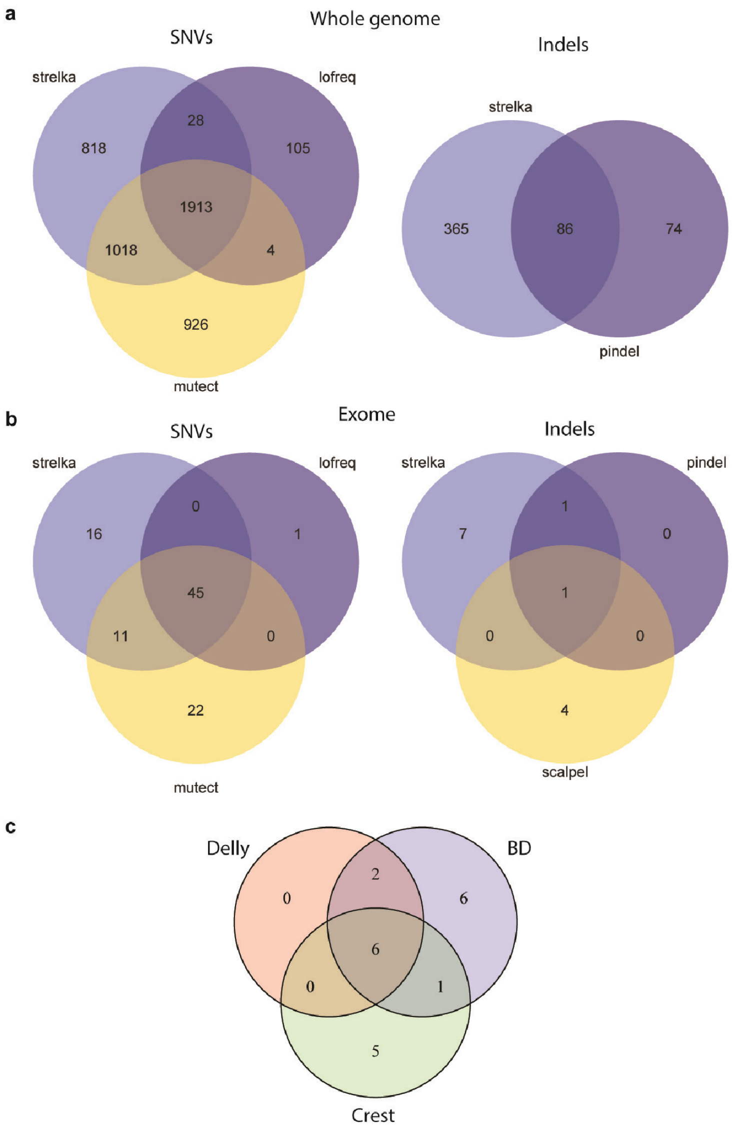
Supplementary Figure 7. Doxycycline-inducible *PGBD5* expression in RPE cells leads to penetrant subcutaneous tumor formation in xenograft models.

(**a**), *GFP-T. ni piggyBac* is expressed at similar relative mRNA levels as *GFP-PGBD5* in RPE cells as measured by quantitative RT-PCR (n = 3, P = 0.79 for *GFP-PGBD5* vs. *GFP-T. ni piggyBac)*. (**b**), Western blot against PGBD5 showing inducible expression of PGBD5 protein in RPE cells stably transduced with *pINDUCER21-PGBD5* after 48 h of treatment with doxycycline (0-600 ng/mL) compared to RPE cells stably expressing *GFP-PGBD5*. (**c**), Tumor size of RPE xenografts as a function of time, with *PGBD5* expression induced using doxycycline (+/- Dox) in RPE cells stably transduced with *pINDUCER21-PGBD5* compared to *GFP-PGBD5* expressing RPE cells and non-transduced cells. Cells were treated with doxycycline for 10 days prior to subcutaneous injection (n = 10 per group).



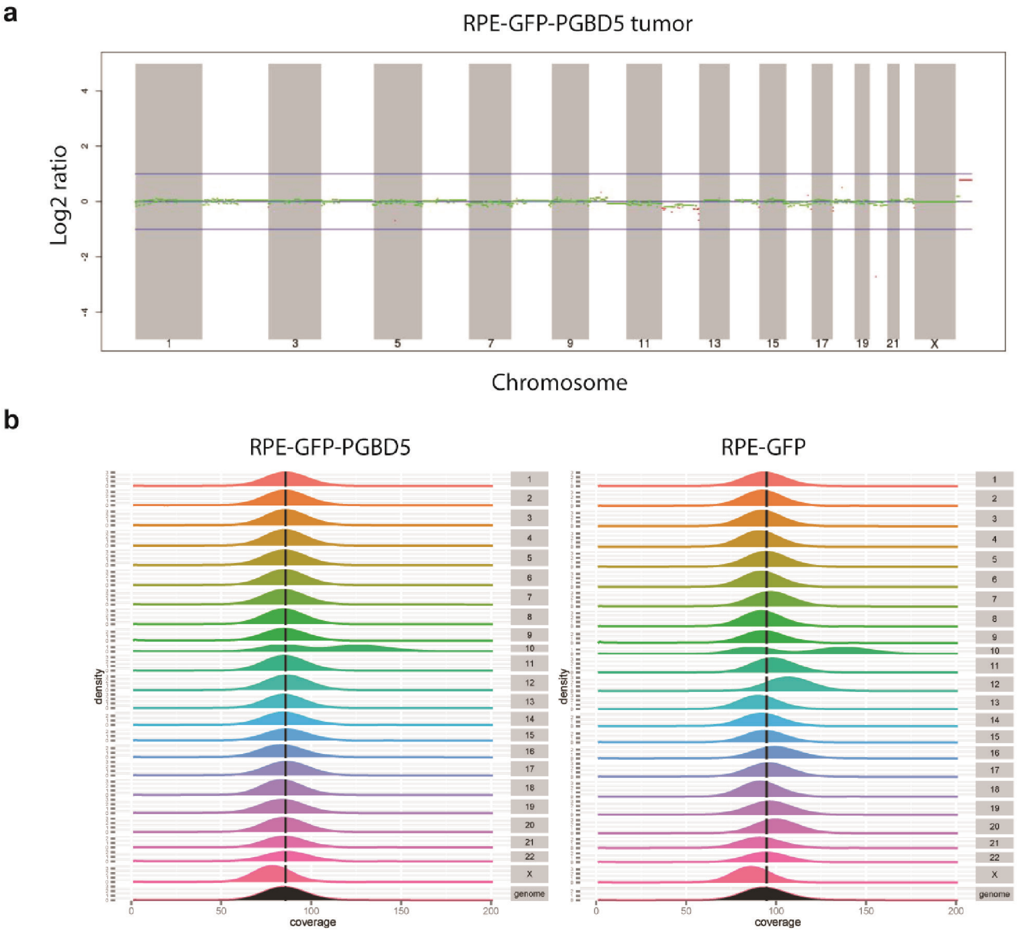
Supplementary Figure 8. PGBD5-mediated genome remodeling requires NHEJ repair.

(**a**), Flow cytometric analysis of cleaved caspase-3 expression in PAXX+/+ and PAXX−/− RPE cells before and after 48 h of doxycycline-induced *PGBD5* expression (500 ng/ml doxycycline). (**b**), Representative images of PAXX+/+ and PAXX−/− RPE cells stained for DAPI (blue) and γ-H2AX (red) 3 h, 6 h, 24 h and 30 h after doxycycline-induced PGBD5 expression (500 ng/ml doxycycline, scale bar = 50 μm). (**c**), Number of viable PAXX+/+ and PAXX−/− RPE cells per cm2 in monolayer culture as measured by trypan blue staining after 72 h of doxycycline-induced expression of *PGBD5*, as compared to untreated control cells (n = 3). **P* = 1.52 × 10^{-4} for PAXX−/−; +Dox vs. PAXX−/−; -Dox. Error bars represent standard deviations of three independent experiments. (**d**), Fraction of γ-H2AX-positive cells over time in PAXX+/+ and PAXX−/− RPE cells before and after doxycycline-induced PGBD5 expression (500 ng/ml doxycycline, n = 3 per group).



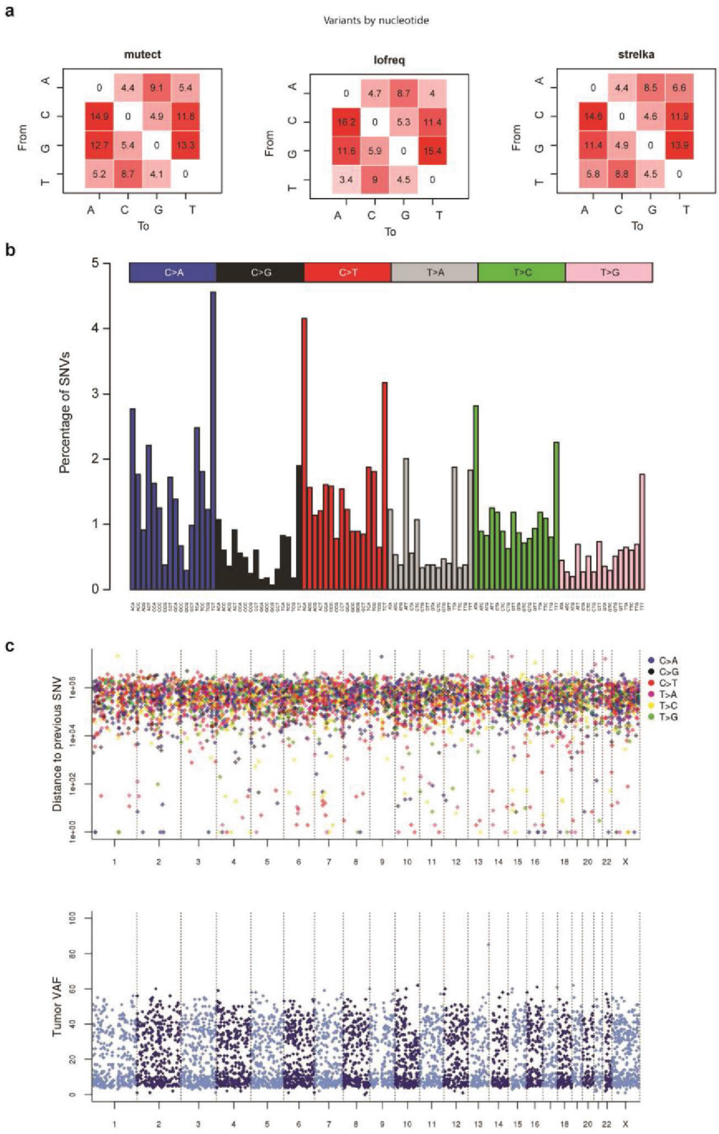
Supplementary Figure 9. Conventional alignment-based variant analysis of structural variants in *PGBD5*-transformed RPE cells.

(**a**), Venn diagrams showing the number of identified SNVs and indels detected by Strelka, LoFreq and Pindel in genomes of RPE cells expressing *GFP-PGBD5* compared to *GFP*. (**b**), Venn diagrams showing the number of identified exonic SNVs and indels detected by Strelka, loFreq and Pindel in *GFP-PGBD5* expressing RPE cells. (**c**), Venn diagrams showing the number of identified large structural variants detected by DELLY, BreakDancer (BD) and CREST (filtered high-confidence set) in *GFP-PGBD5* expressing RPE cells.



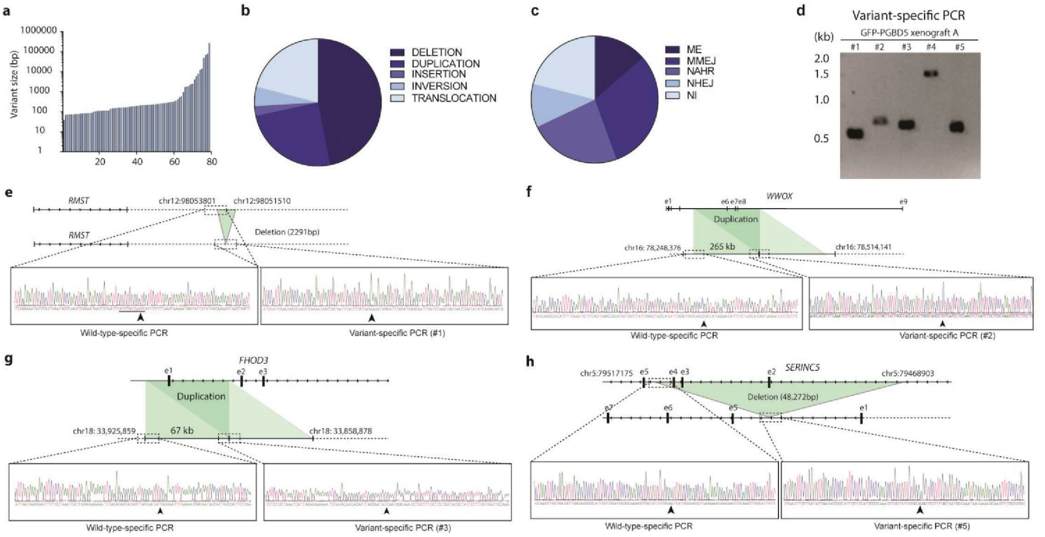
Supplementary Figure 10. *GFP-PGBD5*-expressing cells exhibit a low frequency of copy-number variants across the genome.

(**a**), Copy number profile in RPE cells expressing *GFP-PGBD5* compared to *GFP* expressing cells, computed by BIC-Seq2. (**b**), Relative chromosomal sequence coverage in *GFP-PGBD5* expressing cells (left) compared to *GFP* expressing cells (right) as a function of chromosome number.



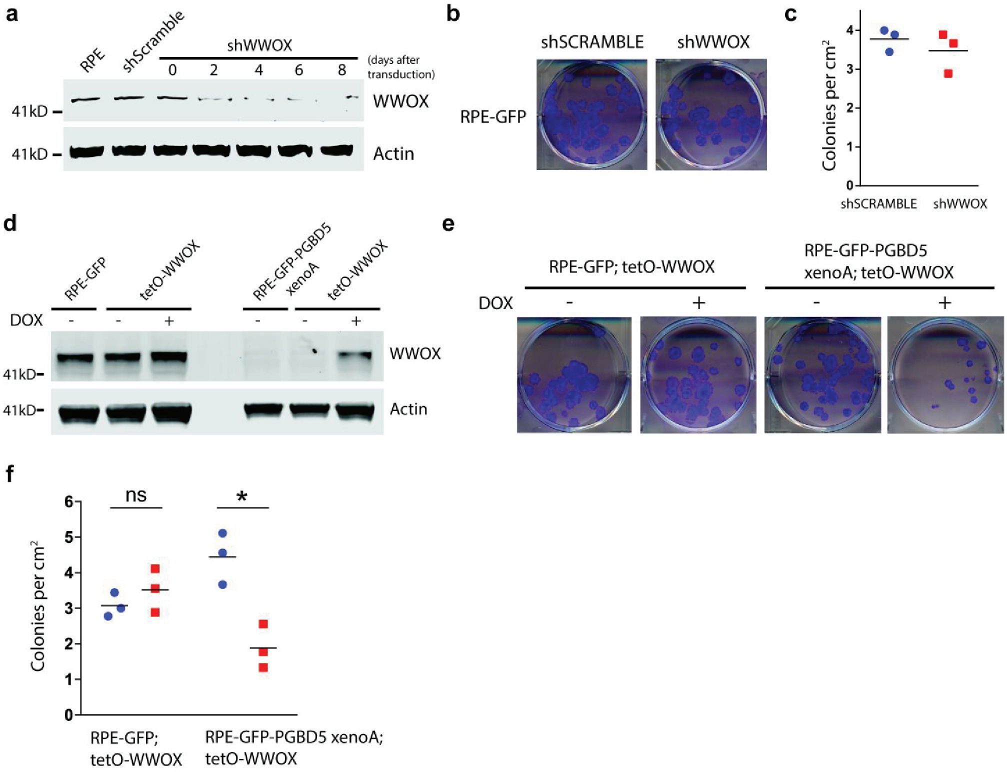
Supplementary Figure 11. Single-nucleotide-variant mutational signatures of *GFP-PGBD5*-expressing cells.

(**a**), Fraction of SNVs involving each nucleotide in *GFP-PGBD5* expressing RPE cells compared to *GFP* expressing cells as detected by Mutect, LoFreq and Strelka (left to right). (**b**), Mutational signature in *GFP-PGBD5* expressing RPE cells measured as the relative fraction of SNVs (union of Mutect, LoFreq and Strelka) in each substitution class and sequence context immediately 3' and 5' to the mutated base. (**c**), Genomic distribution of SNVs in *GFP-PGBD5* expressing RPE cells according to their mutational class (upper panel) and variant allele frequency (lower panel).



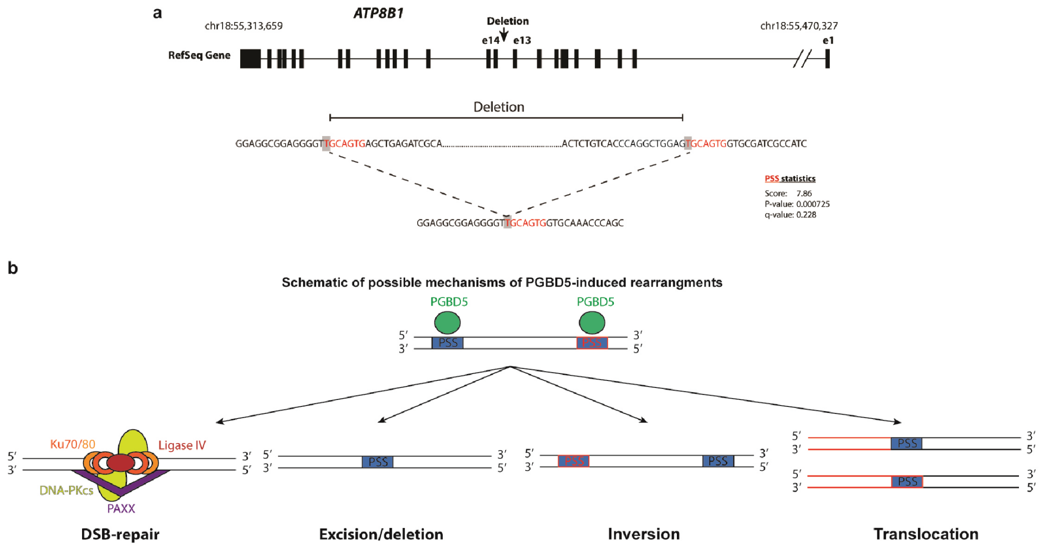
Supplementary Figure 12. PGBD5-induced genomic rearrangements in RPE cells and primary malignant rhabdoid tumors.

(**a**), Histogram showing the genomic size distribution of deletions (excluding small indels) detected by SMuFin in *PGBD5*-transformed RPE cells. (**b**), Distribution of somatic deletions, duplications, insertions, inversions and translocations observed in PGBD5-expressing RPE cell tumors. (**c**), Distribution of predicted mechanisms at the rearrangement breakpoints as homologous recombination (HR), microhomology-mediated end joining (MMEJ), mobile element rearrangements (ME), and non-template insertions (NI). (**d**), Variant allele-specific PCR of genomic rearrangements detected in PGBD5-expressing RPE cell tumors of *RMST* (#1), *WWOX* (#2), *FHOD3* (#3), *XRN2* (#4), and *SERINC5* (#5). (**e-h**), Schematics of gene structure of *RMST*, *WWOX*, *FHOD3*, and *SERINC5* genes before and after rearrangements, and Sanger DNA sequencing chromatograms of the individual rearrangement breakpoints detected by variant allele-specific PCR in individual primary RPE cell tumor specimens (arrowheads mark the breakpoints)



Supplementary Figure 13. Inactivation of WWOX is necessary but not sufficient for clonogenic maintenance of PGBD5-transformed RPE tumor cells.

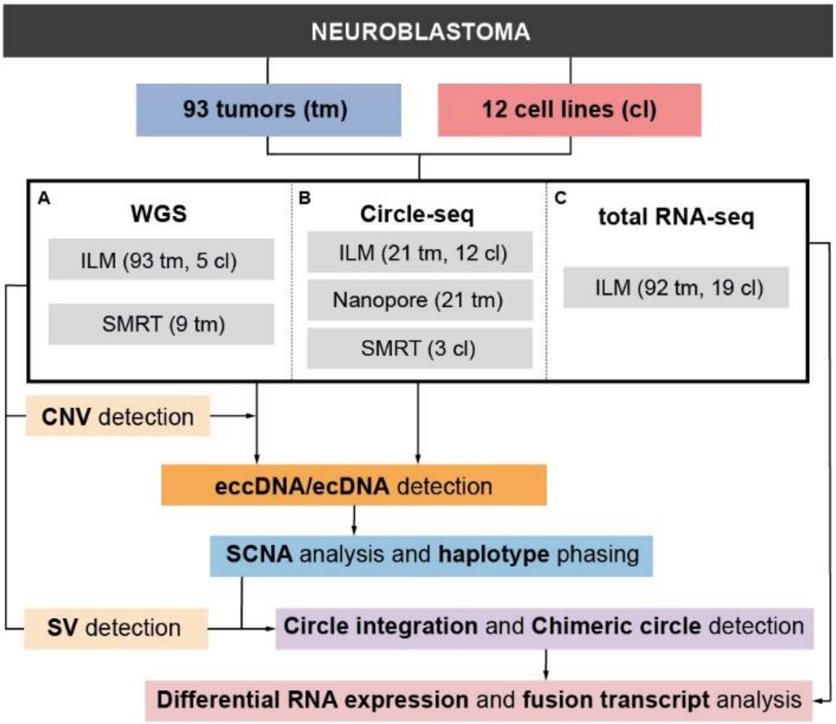
(**a**), Western blot of WWOX showing shRNA-mediated depletion of WWOX in RPE-GFP cells stably transduced with pGIPZ-shWWOX, as compared to pGIPZ-shScramble control. Actin serves as loading control. (**b**,**c**), Representative photographs of Crystal violet-stained colonies (**b**) and clonogenic efficiency (**c**) of RPE-GFP cells expressing pGIPZ-shWWOX, as compared to pGIPZ-shScramble control. (*P* = 0.44). (**d**), Western blot of WWOX showing doxyclycline-induced expression of wild-type WWOX in RPE-GFP cells stably transduced with tetOn-advanced-WWOX vector, as compared to RPE-GFP-PGBD5 xenograft tumor-derived cells with PGBD5-induced WWOX mutation. (**e**,**f**), Representative photographs of Crystal violet-stained colonies (**e**) and clonogenic efficiency (**f**) of RPE-GFP cells and RPE-GFP-PGBD5 xenograft tumor-derived cells stably transduced with tetOn-advanced-WWOX and treated with doxycycline (500 ng/ml) or vehicle control. PGBD5-transformed cells with WWOX mutations, but not control GFP cells, exhibit significantly reduced clonogenic efficiency upon ectopic expression of wild-type WWOX (**P* = 0.0098). Error bars represent standard deviations of three independent experiments.



Supplementary Figure 14. Schematic of PGBD5-induced genomic rearrangement mechanisms.

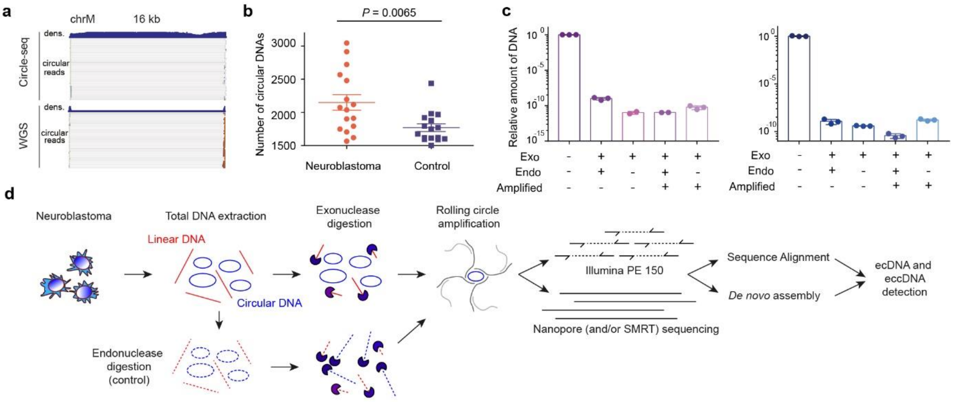
(**a**), Schematic of intragenic deletion with the PSS sequences colored in red. (**b**), Schematic of possible mechanisms of PGBD5-induced rearrangements.

Appendix 2
Supplementary figures of the neuroblastoma publication



Supplementary Figure 1. Flow chart of the data analysis strategy.

(WGS: Whole genome sequencing; ILM: Illumina sequencing; SMRT: Single molecule real-time sequencing; SCNA: Somatic copy number alterations; SV: Structural variant; RNA-seq: RNA sequencing).



Supplementary Figure 2. Circle-seq enables efficient enrichment of extrachromosomal circular DNA.

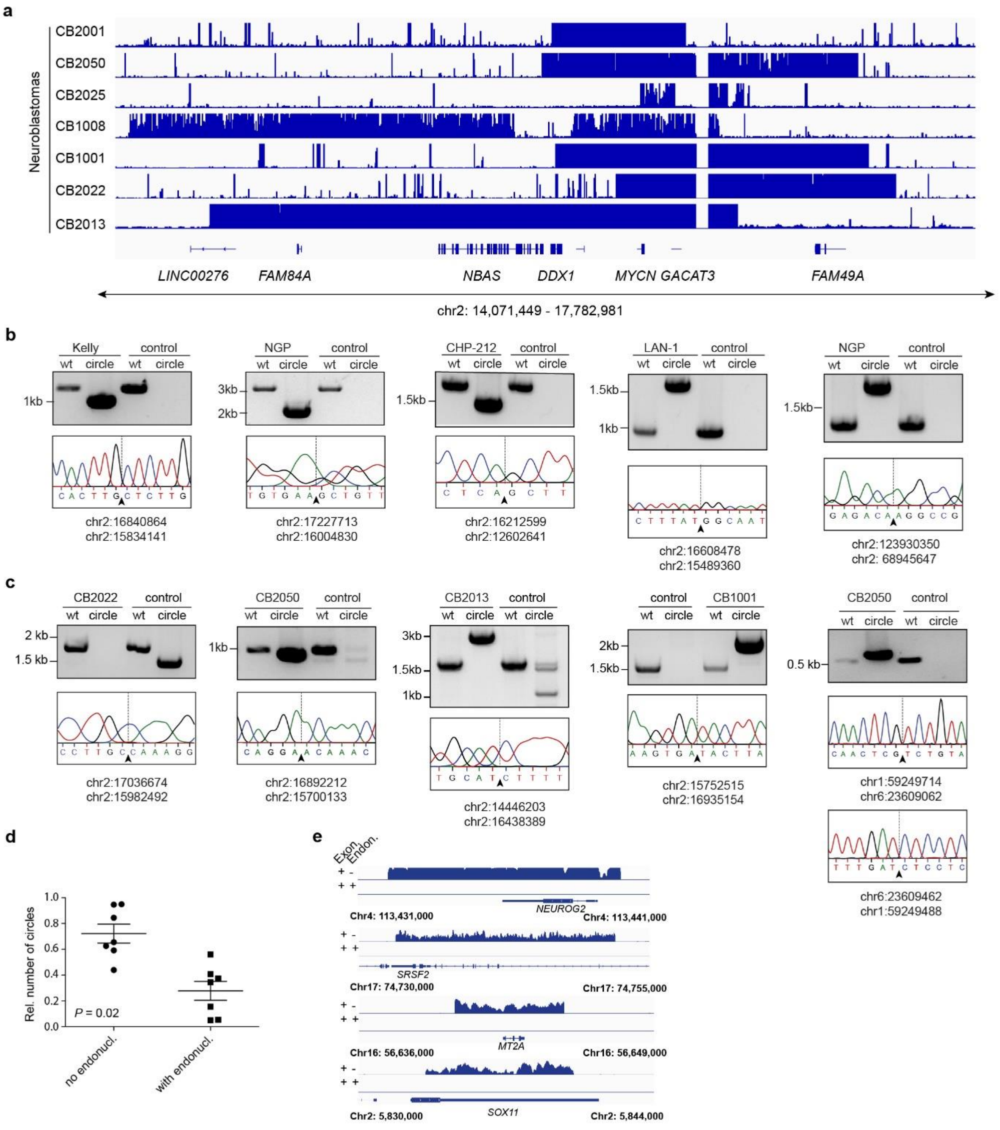
(**a**), Circle read density from Circle-seq (top) and whole genome sequencing (bottom) from mitochondrial DNA (ChrM) in one exemplary patient. (**b**), Number of circular DNAs detected in 16 neuroblastoma tumors-normal pairs (center line indicates mean, error bars indicate standard deviation from the mean, two sided t-test, *P*=0.0065). (**c**), Relative amount of linear genomic DNA (β-globin) before and after exonuclease +/- endonuclease treatment as measured using quantitative PCR (qPCR) in two independent cell lines (right and left; center indicates mean, error bars represent standard deviation of at least two independent qPCR measurements). (**d**), Detailed schematic of the Circle-seq method (A step by step protocol is available online).

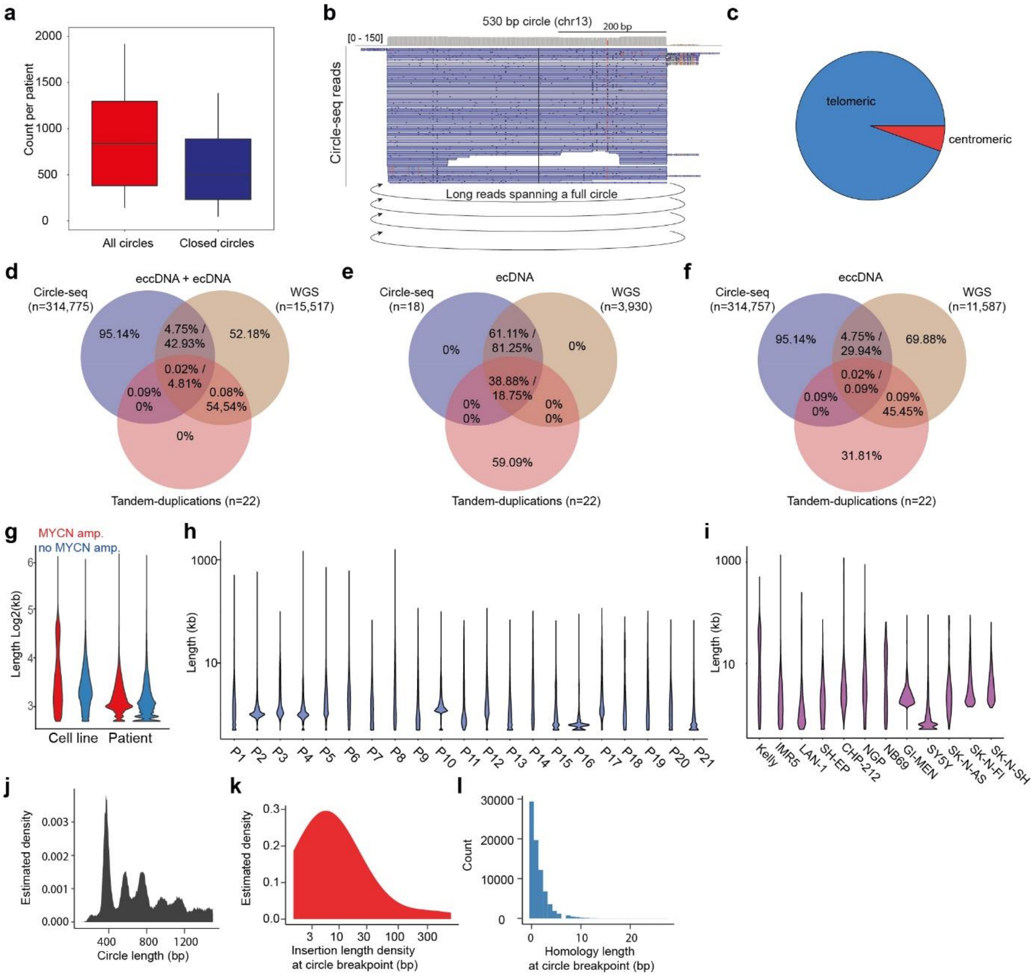
.

Supplementary Figure 3. Circle-seq reliably detects extrachromosomal circular DNA.

(**a**), Extrachromosomal circular DNA read density at chromosome 2 near *MYCN* in primary neuroblastomas as detected using Circle-seq. Extrachromosomal circle junction and wild-type allele-specific PCR in neuroblastoma cell lines (**b**) and matched tumor and normal primary patient specimens (**c**) repeated independently at least three times. (**d**), Relative number of extrachromosomal circular DNAs detected using Circle-seq before and after treatment with a rare-cutting endonuclease in 7 independent neuroblastoma cell lines (Error bars represent standard deviation, mean is indicated by horizontal line, unpaired two-sided t-test, *P*=0.02). (**e**), Exemplary genome tracks at sites of extrachromosomal circularization detected via Circle-seq in each tumor before and after treatment with a rare-cutting endonuclease (number of reads in spanning all circular DNAs was reduced 474 fold, *P* = 7.566 × 10^{-11}, Welch two sample two sided t-test).

(next page)


a
Neuroblastomas
CB2001
CB2050
CB2025
CB1008
CB1001
CB2022
CB2013
LINC00276
FAM84A
NBAS
DDX1
MYCN GACAT3
FAM49A
chr2: 14,071,449 - 17,782,981
b
Kelly
control
wt circle wt circle
1kb
chr2:16840864
chr2:15834141
NGP
control
3kb
2kb
chr2:17227713
chr2:16004830
CHP-212
control
1.5kb
chr2:16212599
chr2:12602641
LAN-1
control
1.5kb
1kb
chr2:16608478
chr2:15489360
NGP
control
1.5kb
chr2:123930350
chr2: 68945647
c
CB2022
control
2 kb
1.5 kb
chr2:17036674
chr2:15982492
CB2050
control
1kb
chr2:16892212
chr2:15700133
CB2013
control
3kb
1.5kb
1kb
chr2:14446203
chr2:16438389
control
CB1001
2kb
1.5kb
chr2:15752515
chr2:16935154
CB2050
control
0.5 kb
chr1:59249714
chr6:23609062
chr6:23609462
chr1:59249488
d
Rel. number of circles
1.0
0.8
0.6
0.4
0.2
0.0
P = 0.02
no endonucl.
with endonucl.
e
Exon
Endon.
+ -
+ +
NEUROG2
Chr4: 113,431,000
Chr4: 113,441,000
SRSF2
Chr17: 74,730,000
Chr17: 74,755,000
MT2A
Chr16: 56,636,000
Chr16: 56,649,000
SOX11
Chr2: 5,830,000
Chr2: 5,844,000


a
2000
1500
1000
500
0
Count per patient
All circles
Closed circles
b
530 bp circle (chr13)
200 bp
[0 - 150]
Circle-seq reads
Long reads spanning a full circle
c
telomeric
centromeric
d
eccDNA + ecDNA
Circle-seq (n=314,775)
WGS (n=15,517)
95.14%
4.75% / 42.93%
52.18%
0.02% / 4.81%
0.09% 0%
0.08% 54,54%
0%
Tandem-duplications (n=22)
e
ecDNA
Circle-seq (n=18)
WGS (n=3,930)
0%
61.11% / 81.25%
0%
38.88% / 18.75%
0% 0%
0% 0%
59.09%
Tandem-duplications (n=22)
f
eccDNA
Circle-seq (n=314,757)
WGS (n=11,587)
95.14%
4.75% / 29.94%
69.88%
0.02% / 0.09%
0.09% 0%
0.09% 45.45%
31.81%
Tandem-duplications (n=22)
g
MYCN amp.
no MYCN amp.
Length Log2(kb)
6
5
4
3
Cell line
Patient
h
1000
10
Length (kb)
P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
P11
P12
P13
P14
P15
P16
P17
P18
P19
P20
P21
i
1000
10
Length (kb)
Kelly
IMR5
LAN-1
SH-EP
CHP-212
NGP
NB69
GI-MEN
SY5Y
SK-N-AS
SK-N-FI
SK-N-SH
j
Estimated density
0.003
0.002
0.001
0.000
400
800
1200
Circle length (bp)
k
Estimated density
0.3
0.2
0.1
0.0
3
10
30
100
300
Insertion length density at circle breakpoint (bp)
l
Count
30000
20000
10000
0
0
10
20
Homology length at circle breakpoint (bp)

Supplementary Figure 4. Combining whole-genome sequencing with Circle-seq enables the characterization of extrachromosomal circular DNAs.

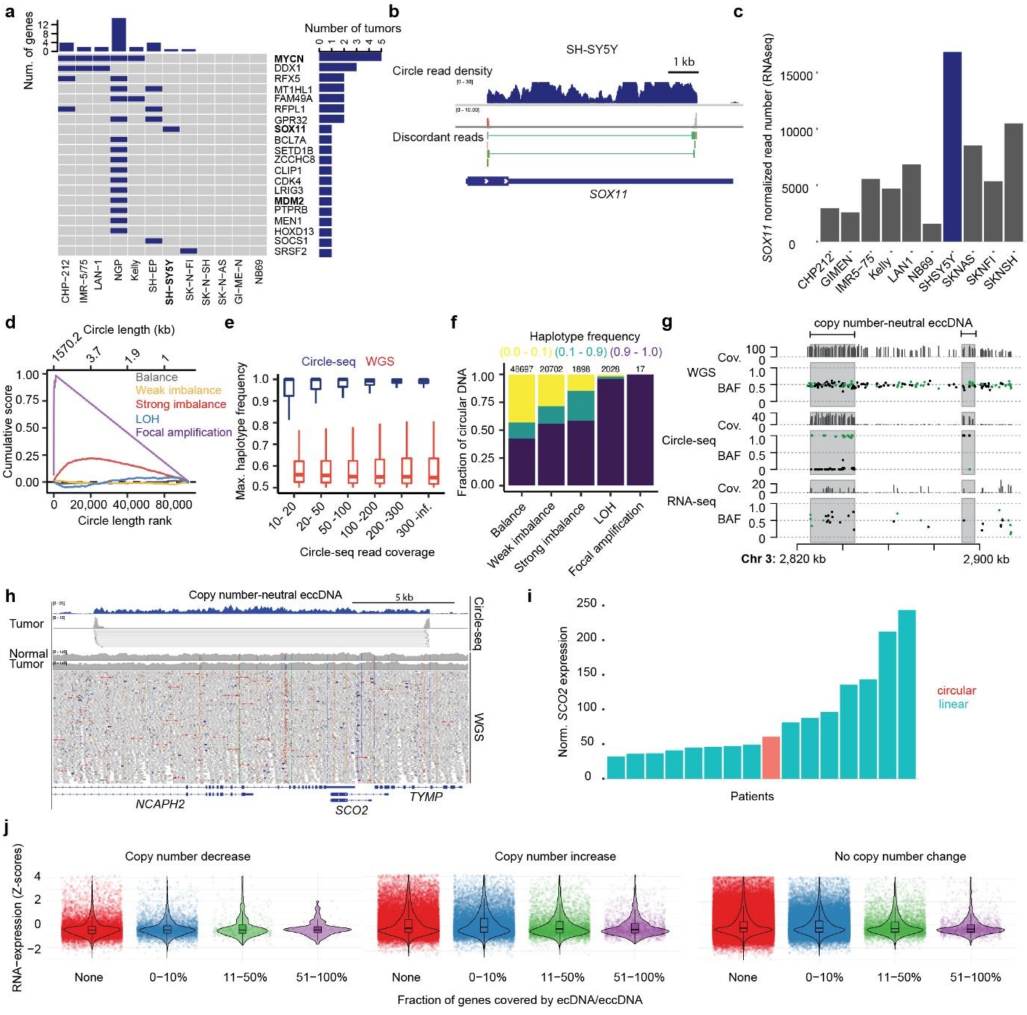
(**a**), Circular DNAs detected by Nanopore long-read sequencing (*N*=19,947) and with fully closed circles by junction-spanning reads (*N*=12,985; 65%) in 21 patient samples. (**b**), Exemplary genome track at site of extrachromosomal circularization with long reads spanning an entire circular DNA multiple times, physically confirming its circular structure. (**c**), Fraction of unmappable Nanopore long reads, which are mapped to telomeric and centromeric sequences after *de novo* assembly. Intersection between all circular DNAs (**d**), ecDNAs (**e**) and eccDNAs (**f**) detected using Circleseq compared to circular DNAs inferred from whole-genome sequencing and regions recognized as tandem-duplications by 5 variant callers in 16 neuroblastoma patients. Size distribution of extrachromosomal circular DNAs identified using Circle-seq in 21 primary neuroblastomas (**g+h**) and 12 neuroblastoma cell lines (**g+i**). (**j**), Size distribution of small extrachromosomal circles less than 1,500bp in length (*N*=59,560) with basepair-accurate breakpoint reconstruction using Circle-seq in 17 primary neuroblastomas. Density estimate using a Gaussian kernel with standard deviation set to 3. (**k**), Length distribution of sequence insertions (*N*=2,145) at basepairaccurate reconstructions of circle junctions in 17 primary neuroblastomas. Density estimate using a Gaussian kernel with standard deviation set to 1. (**l**), Length distribution of homologous sequences (microhomologies) at base-pair accurate reconstructions of circle junctions in 17 primary neuroblastomas (*N*=76,220).

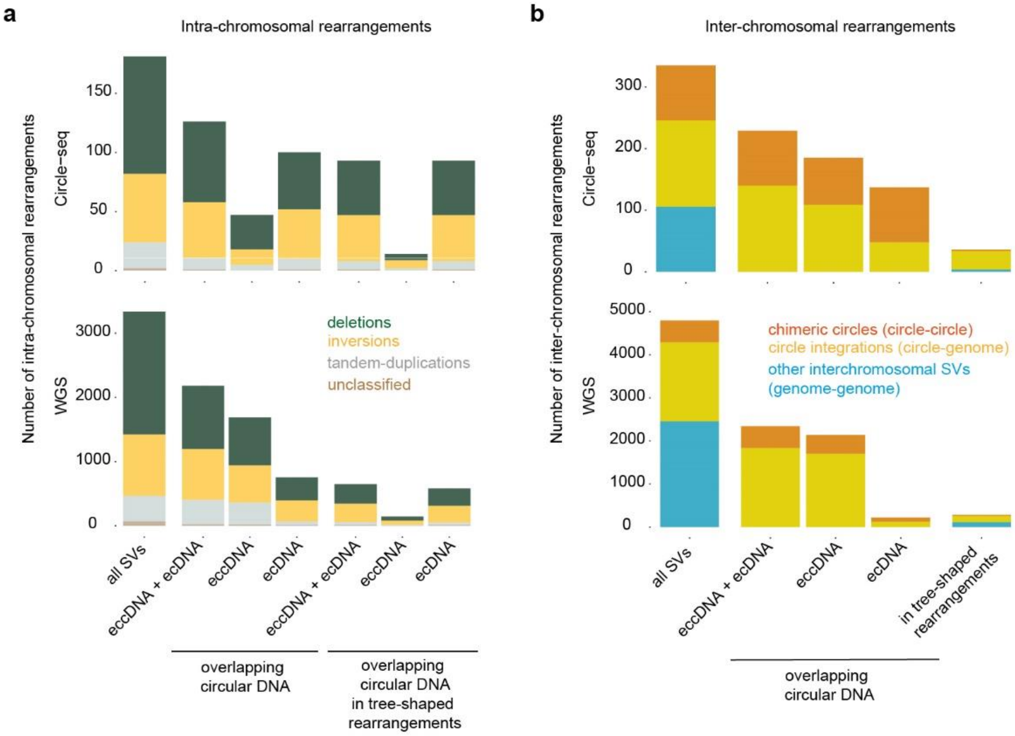
(previous page)

Supplementary Figure 5. Copy number-neutral extrachromosomal circular DNAs are not associated with changes in gene expression.

(**a**), Cancer-relevant genes (rows) circularized in neuroblastoma cell lines (columns) as detected using Circle-seq (*N*=12 cell lines). (**b**), Circle read density and genome track at *SOX11* gene. (**c**), mRNA expression of *SOX11* in 12 neuroblastoma cell lines. (**d**), Cumulative enrichment score of genomic copy number states within a ranked list of extrachromosomal circle sizes (*N*=73,342). Large extrachromosomal circular DNAs are significantly associated with focal copy number amplifications (pink, P_{emp}=0, empirical nominal one-sided *P*-value of absolute maximum cumulative scores from 10,000 random permutations of copy number scores. None of the random absolute maximum cumulative scores was greater than the observed score of 0.98; LOH: Loss of heterozygosity). (**e**), Maximum haplotype frequency in Circle-seq compared to WGS at different Circle-seq coverages of 73,342 circles across samples (Box indicates first, second and third quartile. Whiskers extend to lowest and highest value at max. 1.5 × interquartile distance from first and third quartile). (**f**), Fraction of extrachromosomal circular DNA with distinct haplotype preferences depending on their copy number status. (**g**), Genome track at the site of a copy number neutral extrachromosomal circular DNA showing the coverage of haplotype specific, phased, reads from Circle-seq, whole genome sequencing (WGS) and RNA sequencing. (**h**), Genome track at site of a copy number neutral extrachromosomal circular DNA affecting *SCO2*. (***i***), Normalized *SCO2* RNA expression in a subset of patient tumors. (**j**), Degree of gene circularization and gene copy number differences (increase, *N*=26,374; decrease, *N*=5,656; no change, *N*=58,862) compared to RNA expression differences (z-scores).

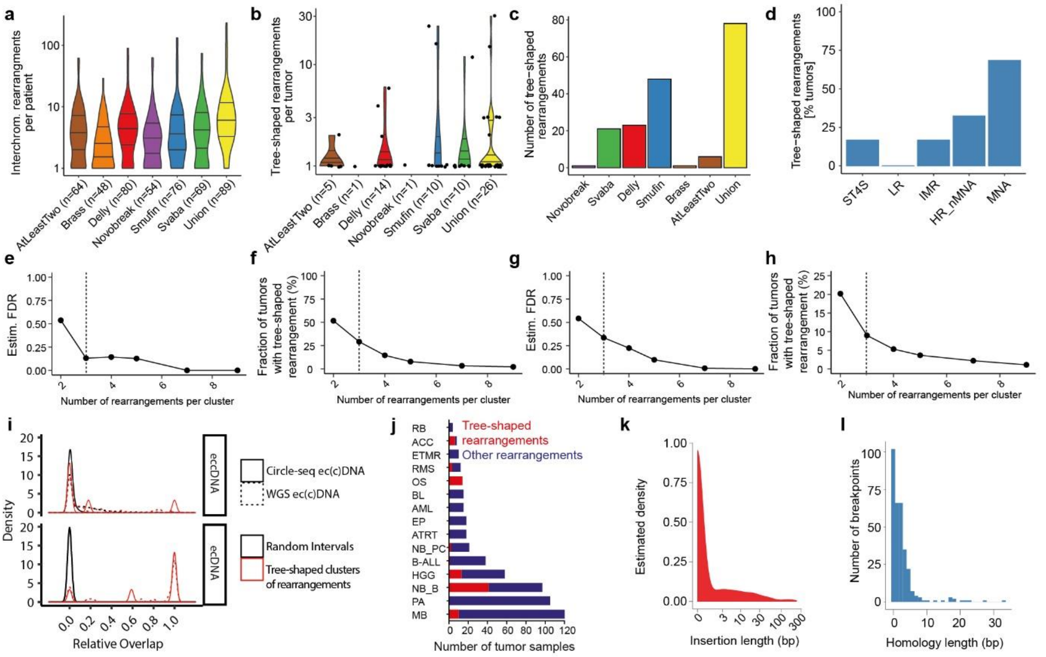
(next page)


a
Num. of genes
Number of tumors
MYCN
DDX1
RFX5
MT1HL1
FAM49A
RFPL1
GPR32
SOX11
BCL7A
SETD1B
ZCCHC8
CLIP1
CDK4
LRIG3
MDM2
PTPRB
MEN1
HOXD13
SOCS1
SRSF2
CHP-212
IMR-5/75
LAN-1
NGP
Kelly
SH-EP
SH-SY5Y
SK-N-FI
SK-N-SH
SK-N-AS
GI-ME-N
NB69
b
SH-SY5Y
1 kb
Circle read density
Discordant reads
SOX11
c
SOX11 normalized read number (RNAseq)
CHP212
GIMEN
IMR5-75
Kelly
LAN1
NB69
SHSY5Y
SKNAS
SKNFI
SKNSH
d
Circle length (kb)
Cumulative score
Circle length rank
Balance
Weak imbalance
Strong imbalance
LOH
Focal amplification
e
Circle-seq
WGS
Max. haplotype frequency
Circle-seq read coverage
f
Haplotype frequency
(0.0 - 0.1)
(0.1 - 0.9)
(0.9 - 1.0)
Fraction of circular DNA
48697
20702
1898
2026
17
Balance
Weak imbalance
Strong imbalance
LOH
Focal amplification
g
copy number-neutral eccDNA
Cov.
WGS
BAF
Circle-seq
RNA-seq
Chr 3: 2,820 kb
2,900 kb
h
Copy number-neutral eccDNA
5 kb
Tumor
Normal
Tumor
Circle-seq
WGS
NCAPH2
SCO2
TYMP
i
Norm. SCO2 expression
circular
linear
Patients
j
Copy number decrease
Copy number increase
No copy number change
RNA-expression (Z-scores)
None
0–10%
11–50%
51–100%
Fraction of genes covered by ecDNA/eccDNA



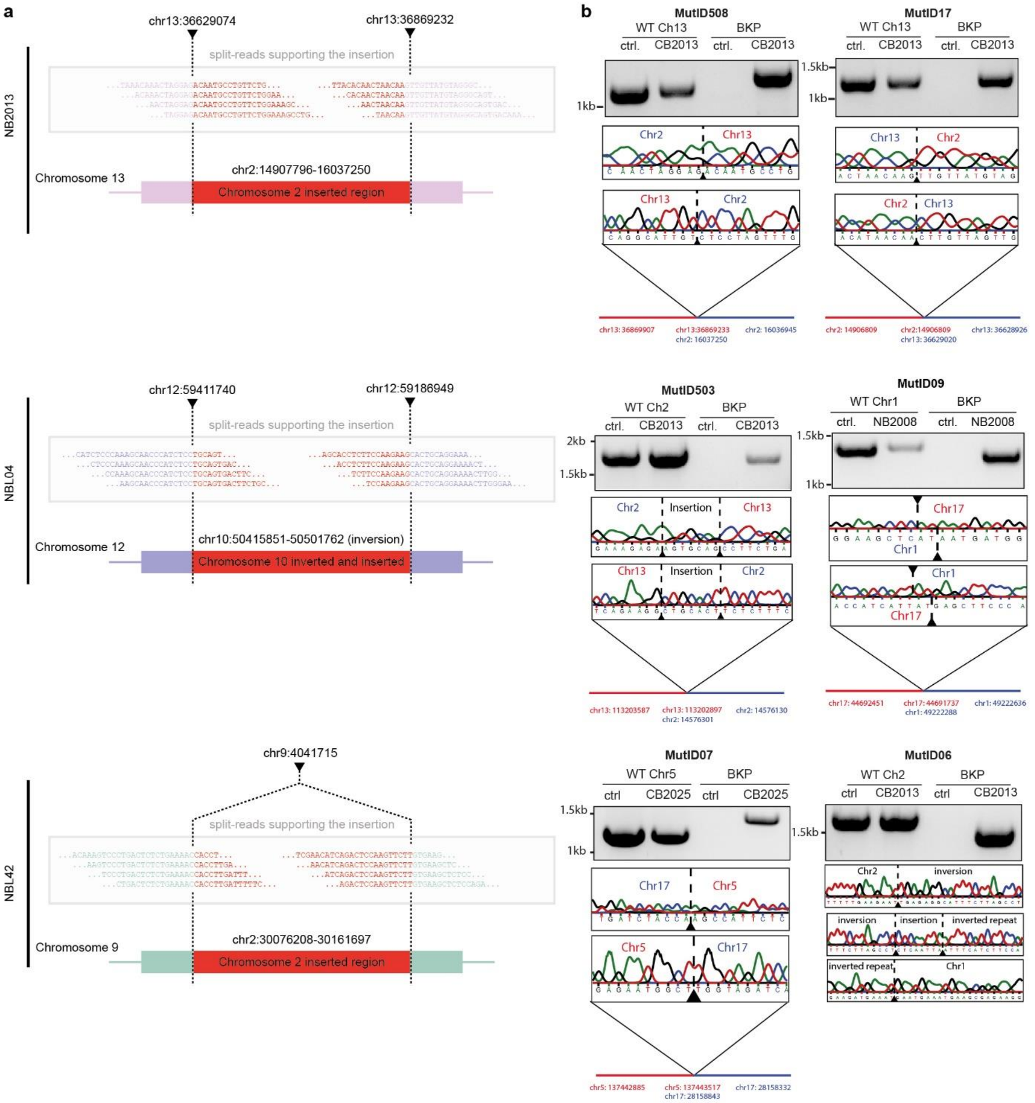
Supplementary Figure 6. The majority of somatic structural rearrangements in neuroblastoma involve extrachromosomal circular DNA.

Intra- (**a**) and inter-chromosomal (**b**) somatic rearrangements and their association with extrachromosomal circular DNAs (ecDNAs compared to eccDNAs) as detected by Circle-seq (*N*=16 tumors) and WGS (*N*=93 tumors) define two classes of circle-associated rearrangements, chimeric circles (connecting extrachromosomal circular DNA to extrachromosomal circular DNA, circle-circle) and circle integrations (connecting extrachromosomal circular DNA to chromosomal linear DNA, circle-genome).



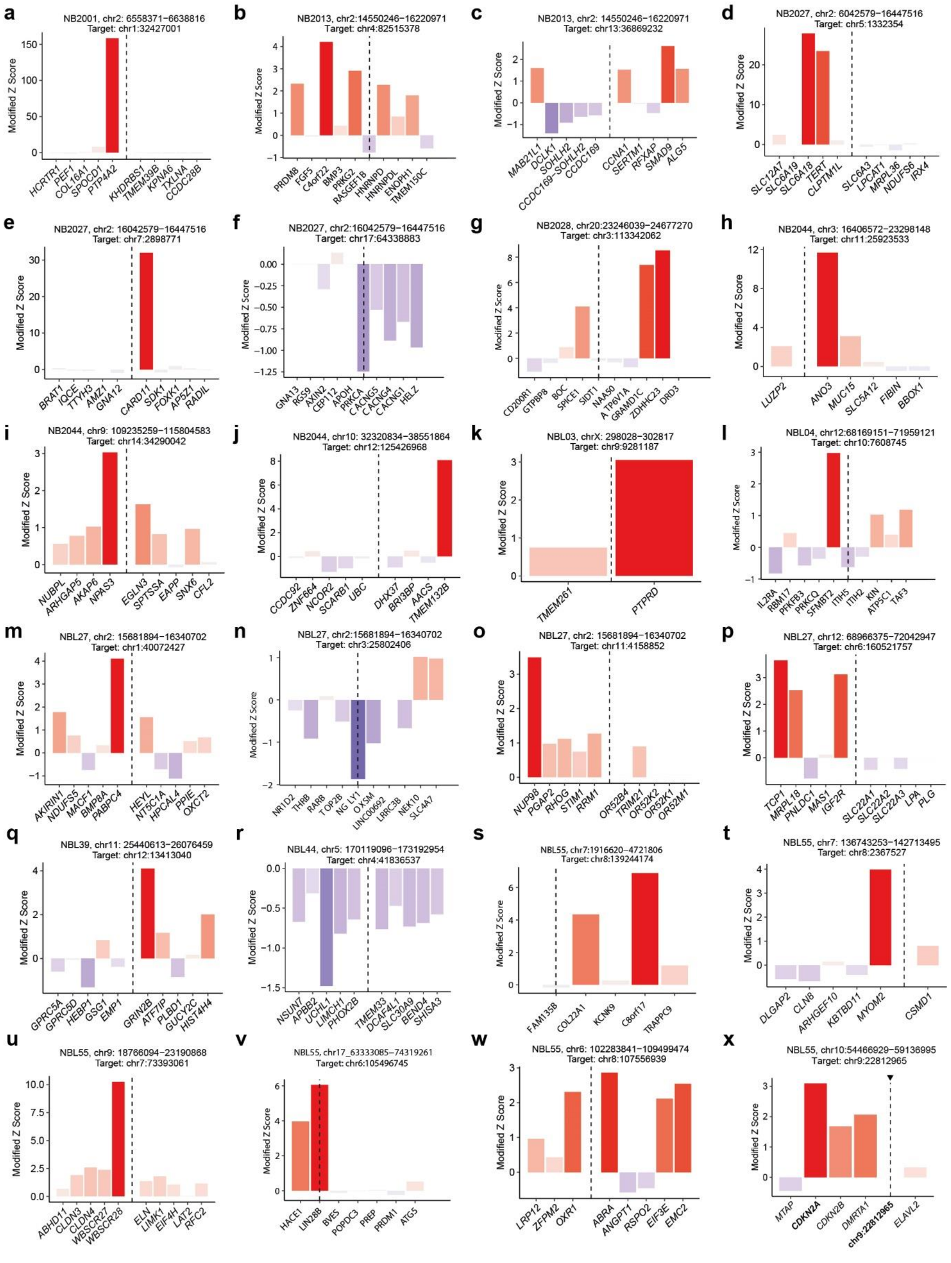
Supplementary Figure 7. Clustered tree-shaped circle-associated rearrangements can be detected in neuroblastoma genomes.

(**a**), Number of inter-chromosomal rearrangements per tumor detected by each structural variant caller, at least two variant callers (AtLeastTwo) compared to the union of rearrangements detected by all variant calling algorithms (Bars indicate maximum, minimum and quartile boundaries). (**b**), Number of tree-shaped rearrangement patterns per tumor detected in 91 neuroblastoma genomes by five different variant callers (Bars indicate maximum, minimum and quartile boundaries). (**c**), Total number of tree-shaped rearrangements detected in 91 neuroblastoma genomes by five different variant callers. (**d**), Fraction of neuroblastomas with at least one tree-shaped rearrangement pattern in different neuroblastoma subgroups (ST4S = stadium 4S, LR = low risk, IMR = intermediate risk, HR_nMNA = high-risk non-*MYCN*-amplified, MNA = *MYCN*-amplified). (**e**), Estimated false discovery rate (FDR) for the detection of tree-shaped rearrangement patterns based on the number of rearrangements set as a threshold to define such a pattern in the neuroblastoma cohort (*N*=91). (**f**), Fraction of neuroblastoma genomes with tree-shaped rearrangement patterns depending on the number of rearrangements set as a threshold to detect such a pattern in the neuroblastoma cohort (*N*=91). (**g**), Estimated FDR for the detection of tree-shaped rearrangement patterns in a publicly available dataset from 546 pediatric cancers comprising 15 types. (**h**), Fraction of pediatric cancer genomes with tree-shaped rearrangement patterns depending on the number of rearrangements set as a threshold to detect such a pattern in 546 pediatric cancers. (**i**), Relative overlap of circular DNAs (ecDNA and eccDNA) with tree-shaped clustered rearrangements compared to overlap of randomized regions as measured using Circle-seq compared to WGS. All overlaps except eccDNA x Circle-seq are significantly above chance (empirical p-values based on 2000 randomized datasets, one-tailed test, Benjamini-Hochberg-corrected; *N*=6 cluster regions for Circle-seq data, *N*=78 cluster regions for WGS data; *P*=1.0 for Circle-seq eccDNA, *P*=9.995e-4 for Circle-seq ecDNA, *P*=0.0227 for WGS eccDNA and *P*=9.995e-4 for WGS ecDNA). (**j**), Frequency of tree-shaped rearrangements indicative of circleassociated rearrangements in our cohort of 91 neuroblastomas (NB_B) and a publicly available dataset from 546 pediatric cancers comprising 15 types (RB = retinoblastoma, ACC = adrenocortical carcinoma, ETMR = embryonal tumor with multilayered rosettes, RMS = rhabdomyosarcoma, OS = osteosarcoma, BL = Burkitt lymphoma, AML = acute myeloid 9 leukemia, EP = ependymoma, ATRT = atypical teratoid rhabdoid tumor, NB_PC = neuroblastoma Heidelberg cohort, B-ALL = B-cell acute lymphoblastic leukemia, HGG = highgrade glioma, PA = pilocytic astrocytoma, MB = medulloblastoma). (**k**), Length distribution of sequence insertions at accurately reconstructable breakpoints (N=320) of rearrangement breakpoints. 14.5% of breakpoints showed small insertions of at least 5bp. (**l**), Length distribution of homologous sequences at accurately reconstructable rearrangement breakpoints (*N*=320). Microhomologies of at least 5 bp are found at 10.0% of rearrangement breakpoint junctions. Mean homology length was found to be 5-times longer for real SV than for a set of randomly permuted breakpoint pairs (both groups *N*=320; Group means 2.47bp vs. 0.49bp; two-sided unequal variances *t*-test, *t*=-8.64, df=342.12, *P*=2.2e-16).



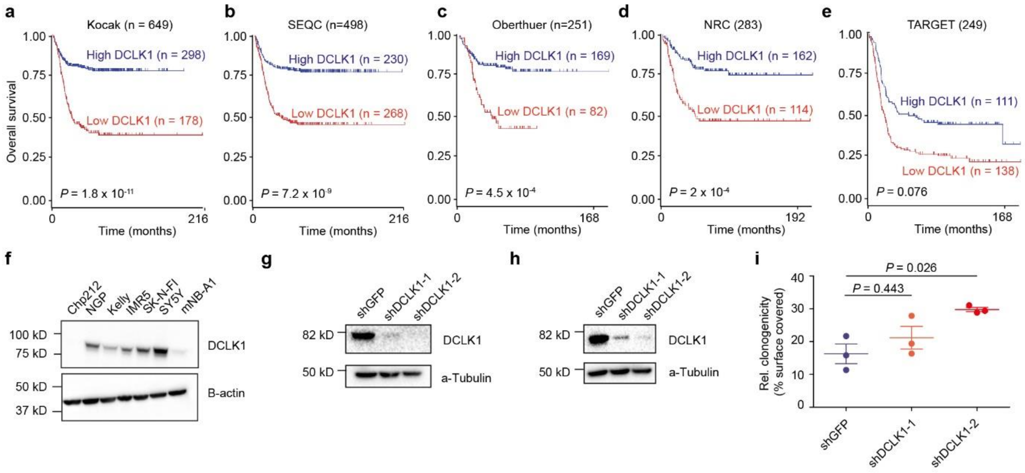
Supplementary Figure 8. Validation of circle-associated rearrangements by allele-specific PCR and Sanger sequencing.

(**a**), Split read support for variant breakpoints of exemplary circleassociated rearrangements. (**b**), Validation of circle integration as assessed using variant and wild-type allele-specific PCR in matched tumor and normal primary patient specimens (repeated at least three independent times).



Supplementary Figure 9. Cancer-driving lesions can emerge out of circle-associated treeshaped rearrangement clusters.

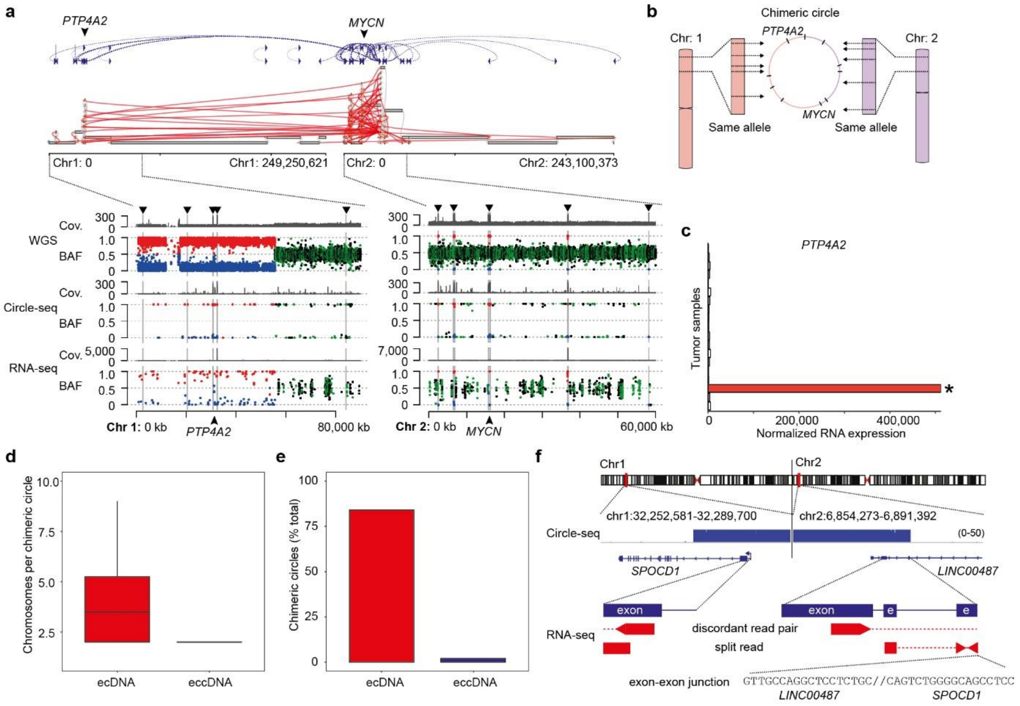
Modified z-scores for the mRNA expression of a subset of breakpoint-neighboring genes at different genomic loci in different neuroblastomas affected by circle-associated tree-shaped rearrangements. The genomic interval indicates the rearrangement cluster. Targets indicate the breakpoint partner connected to the rearrangement cluster (**a-z**).



Supplementary Figure 10. . Cancer-driving lesions with clinical relevance can emerge out of circle integration.

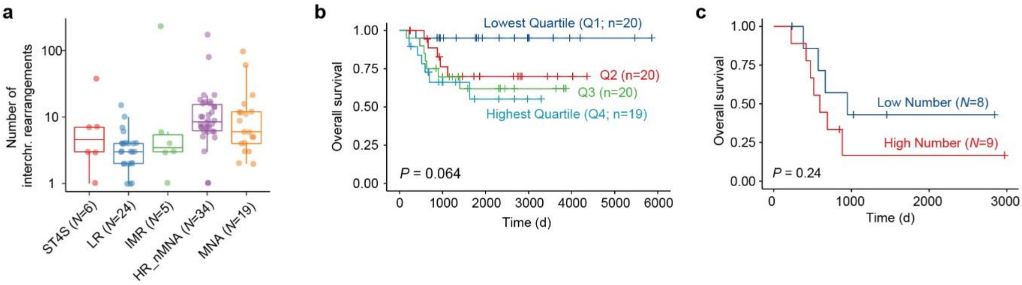
Low *DCLK1* mRNA expression in neuroblastomas correlates with adverse clinical outcome in five independently published gene expression datasets (statistical difference was calculated using a two-sided log rank test corrected after Bonferroni) (**a-e**). (**f**), DCLK1 protein expression in a panel of 7 neuroblastoma cell lines as measured using western immunoblotting (measured at least three times). DCLK1 protein expression as measured using western immunoblotting after shRNA-mediated knock-down of *DCLK1* in IMR5 (**g**) and Kelly cells (**h**) (repeated independently three times). (**i**), Surface area of tissue culture plate covered by cell colonies after shRNA-mediated *DCLK1* knock-down compared to shGFP expressing cells (Center line indicates mean, error bars represent standard deviation of three independent cell culture plates, ANOVA test between groups followed by post-hoc pairwise comparisons using Tukey's Honestly Significant Difference test, *P*=0.443 for shGFP vs. shDCLK1-1, *P*=0.026 for shGFP vs. shDCLK1-2 and *P*=0.136 for shDCLK1-1 vs. shDCLK1-2).

]



Supplementary Figure 11. . Extrachromosomal chimeric circles can lead to co-amplification and overexpression of oncogenes and expression of aberrant circle-specific fusion transcripts.

(**a**), Genome track with genomic copy number alterations at chromosome 1 (top left) and chromosome 2 (top right) connected through formation of a chimeric extrachromosomal circle (top, blue lines). Coverage and B-allelic frequency of reads from regions connected between chromosome 1 and 2 as detected by whole genome sequencing, Circle-seq and RNA sequencing (bottom). (**b**), Schematic of chimeric circle formation depicted in (**a**). (**c**), Normalized gene expression (mRNA) for protein tyrosine phosphatase type IVA, member 2 (*PTP4A2*) in 21 neuroblastomas (tumor affected by chimeric circle shown in (**a**) is marked by asterisk). (**d**), Number of chromosomes included per chimeric extrachromosomal circular DNA (*N*=19 ecDNA; *N*=3,514, eccDNA). (**e**), Fraction of extrachromosomal circular DNA that are of chimeric structure (*N*=16/19, ecDNA; *N*=3,514/167,793, eccDNA). (**f**), RNA sequencing split reads indicating the expression of an aberrant fusion transcript on the chimeric extrachromosomal circular DNA.



Supplementary Figure 12. Higher number of inter-chromosomal rearrangements does not distinguish clinically distinct subgroups of *MYCN*-amplified neuroblastoma.

(**a**), Number of inter-chromosomal rearrangements in neuroblastoma from different clinical risk groups (Tukey-style boxplots with box encompassing the second and third quartile, whiskers include data points within 1.75 times the interquartile range.). (**b**), Kaplan Meier analysis of patient survival comparing patients with neuroblastomas affected by different numbers of somatic inter-chromosomal rearrangements (Q1-4: quartile of numbers of inter-chromosomal rearrangements; two-sided log rank test, *P*=0.064). (**c**), Kaplan Meier analysis comparing patient survival with *MYCN*-amplified neuroblastomas (*N*=17) and high numbers of interchromosomal rearrangements to neuroblastomas with low numbers of inter-chromosomal rearrangements (Low number of rearrangements are defined as numbers below the median; two-sided log-rank test, *P*=0.24).

Appendix 3

Appendix publication 1

Enhancer hijacking determines intra- and extrachromosomal circular *MYCN* amplicon architecture in neuroblastoma

Enhancer hijacking determines intra- and extrachromosomal circular *MYCN* amplicon architecture in neuroblastoma

Konstantin Helmsauer[1*], Maria Valieva[2,3,4*], Salaheddine Ali[2,3,4*], Rocio Chamorro Gonzalez[1], Robert Schöpflin[2,3], Claudia Röefzaad[1], Yi Bei[1], Heathcliff Dorado Garcia[1], Elias Rodriguez-Fos[5], Montserrat Puiggròs[5], Katharina Kasack[6], Kerstin Haase[1], Luis P. Kuschel[7], Philipp Euskirchen[7,8], Verena Heinrich[9], Michael Robson[2,3,4], Carolina Rosswog[10], Jörn Tödling[1], Annabell Szymansky[1], Falk Hertwig[1], Matthias Fischer[10], David Torrents[5,11], Angelika Eggert[1], Johannes H. Schulte[1,6,9], Stefan Mundlos[2,3,4*], Anton G. Henssen[1,6,8,12*], Richard P. Koche[8,13*]

[1]Department of Pediatric Oncology and Hematology, Charité – Universitätsmedizin Berlin, Germany
[2]RG Development & Disease, Max Planck Institute for Molecular Genetics, Berlin, Germany
[3]Institute for Medical Genetics, Charité – Universitätsmedizin Berlin, Germany
[4]Berlin-Brandenburg Center for Regenerative Therapies (BCRT), Charité – Universitätsmedizin Berlin, Germany
[5]Barcelona Supercomputing Center, Joint BSC-CRG-IRB Research Program in Computational Biology, Barcelona, Spain
[6]German Cancer Consortium (DKTK), partner site Berlin, and German Cancer Research Center DKFZ, Heidelberg, Germany
[7]Department of Neurology, Charité – Universitätsmedizin Berlin, Germany
[8]Berlin Institute of Health, Berlin, Germany
[9]Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Berlin, Germany
[10]Department of Experimental Pediatric Oncology, University Children's Hospital of Cologne, and Center for Molecular Medicine Cologne (CMMC), University of Cologne, Germany
[11]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain
[12]Experimental and Clinical Research Center (ECRC) of the MDC and Charité, Berlin, Germany
[13]Center for Epigenetics Research, Memorial Sloan Kettering Cancer Center, New York, New York, USA

[*]These authors contributed equally to this work.
Correspondence should be addressed to A.G.H. (henssenlab@gmail.com) and R.P.K. (kocher@mskcc.org).

Abstract

MYCN amplification drives one in six cases of neuroblastoma. The supernumerary gene copies are commonly found on highly rearranged, extrachromosomal circular DNA. The exact amplicon structure has not been described thus far and the functional relevance of its rearrangements is unknown. Here, we analyzed the *MYCN* amplicon structure and its chromatin landscape. This revealed two distinct classes of amplicons which explain the regulatory requirements for *MYCN* overexpression. The first class always co-amplified a proximal enhancer driven by the noradrenergic core regulatory circuit (CRC). The second class of *MYCN* amplicons was characterized by high structural complexity, lacked key local enhancers, and instead contained distal chromosomal fragments, which harbored CRC-driven enhancers. Thus, ectopic enhancer hijacking can compensate for the loss of local gene regulatory elements and explains a large component of the structural diversity observed in *MYCN* amplification.

Introduction

Oncogene amplification is a hallmark of cancer genomes. It leads to excessive proto-oncogene overexpression and is a key driver of oncogenesis. The supernumerary gene copies come in two forms, i. self-repeating arrays on a chromosome (homogeneously staining regions, HSR) and ii. many individual circular DNA molecules (extrachromosomal DNA, ecDNA, alias double minute chromosomes, dmin)[1]. EcDNA can arise during genome reshuffling events like chromothripsis and are subsequently amplified[2,3]. This partially explains why such circular DNAs can consist of several coding and non-coding distant parts of one or more chromosomes[4]. Over time, amplified DNA acquires additional internal rearrangements as well as coding mutations, which can confer adaptive advantages such as resistance to targeted therapy[5-7]. EcDNA re-integration into chromosomes can lead to intrachromosomal amplification as HSRs[8,9] and act as a general driver of genome remodeling[10]. Our knowledge of the functional relevance of non-coding regions co-amplified on ecDNA, however, is currently limited.

MYCN amplification is a prototypical example of a cancer-driving amplification. The developmental transcription factor was identified as the most commonly amplified gene in a recent pediatric pan-cancer study[11]. Its most prominent role is in neuroblastoma, a pediatric malignancy of the sympathetic nervous system. *MYCN* amplification characterizes one in six cases and confers dismal prognosis[12]. In contrast to long-term survival of more than 80% for non-amplified cases, 5-year overall survival is as low as 32% for *MYCN*-amplified neuroblastoma[12]. In these cases, *MYCN* amplification is likely an early driver of neuroblastoma formation. Accordingly, *MYCN* overexpression is sufficient to induce neuroblastic tumor formation in mice[13,14]. Despite its central role in neuroblastoma biology, the epigenetic regulation of *MYCN* is incompletely understood.

Recently, studies have identified a core regulatory circuit (CRC) including half a dozen transcription factors that drive a subset of neuroblastoma with noradrenergic cell identity, including most *MYCN*-amplified cases[15-18]. The epigenetic landscape around *MYCN* is less well described. In part, this is due to the structural complexity of *MYCN* amplicons and difficulties in the interpretation of epigenomic data in the presence of copy number variation. Recent evidence has emerged suggesting that local enhancers may be required for proto-oncogene expression on amplicons[19]. Here, we sought out to identify key regulatory elements near *MYCN* in neuroblastoma by integrating short- and long-read genomic and epigenomic data from neuroblastoma cell lines and primary tumors. We investigated the activity of regulatory elements in the context of *MYCN* amplification and characterized the relationship between amplicon structure and epigenetic regulation.

Results

Local CRC-driven enhancers contribute to MYCN expression in neuroblastoma

In order to identify candidate regulatory elements near *MYCN*, we examined public H3K27ac chromatin immunoprecipitation and sequencing (ChIP-seq) and RNA sequencing (RNA-seq) data from 25 neuroblastoma cell lines[15]. ChIP-seq data for amplified genomic regions are characterized by a very low signal-to-noise ratio, which has complicated their interpretation in the past[16]. We therefore focused our analysis on 12 cell lines lacking *MYCN*-amplifications but expressing *MYCN* at different levels, allowing for the identification of *MYCN*-driving enhancers in neuroblastoma. Comparison of composite H3K27ac signals of *MYCN*-expressing vs. non-expressing cell lines identified at least 5 putative enhancer elements (e1-e5) that were exclusively present in the vicinity of *MYCN* in cells expressing *MYCN*, thus likely contributing to *MYCN* regulation (Fig. 1, Supplementary Fig. 1a). Consistent with differential RNA expression, a strong differential H3K27ac peak was identified spanning the *MYCN* promotor and gene body (MYCNp; Fig. 1). The identified enhancers were not active in developmental precursor cells such as embryonic stem cells, neuroectodermal cells or neural crest cells (Supplementary Fig. 1b), suggesting these enhancers were specific for later stages of sympathetic nervous system development or neuroblastoma. Transcription factor ChIP-seq in *MYCN*-expressing cells confirmed that four of the enhancers (e1, e2, e4, e5) were bound by each of three noradrenergic neuroblastoma core regulatory circuit factors (PHOX2B, HAND2, GATA3; Fig. 1b). All but enhancer e3 harbored binding motifs for the remaining members of the core regulatory circuit (ISL1, TBX2, ASCL1; Supplementary Fig. 1c) for which ChIP-seq data were unavailable. Additionally, all enhancers contained binding motifs for TEAD4, a transcription factor implicated in a positive feedback loop with MYCN in *MYCN*-amplified neuroblastoma[20]. Two of the enhancers (e1, e2) also harbored canonical E-boxes, suggesting binding of MYCN at its own enhancers (Supplementary Fig. 1c). Thus, a common set of CRC-driven enhancers is found specifically in *MYCN* expressing neuroblastoma cells, indicating that *MYCN* expression is regulated by the CRC.

Local enhancer co-amplification explains asymmetric MYCN amplicon distribution

MYCN is expressed at the highest levels in neuroblastomas with *MYCN* amplifications (Supplementary Fig. 1d). It is unclear, however, to what extent enhancers are required for sustained MYCN expression on *MYCN*-containing amplicons. To address this, we mapped amplified genomic regions in a meta-dataset of copy-number variation in 240 *MYCN*-amplified neuroblastomas[21]. This revealed an asymmetric pattern of *MYCN* amplification (Fig. 2a, Supplementary Fig. 2). Intriguingly, a 290kb region downstream of *MYCN* was co-amplified in more than 90% of neuroblastomas, suggesting that *MYCN* amplicon boundaries were not randomly distributed, which is in line with recent reports in a smaller tumor cohort[19] Notably, the consensus amplicon boundaries did not overlap with common fragile sites (Supplementary Fig. 2g), challenging a previous association found in ten neuroblastoma cell lines[8]. Regions of increased chromosomal instability alone are therefore unlikely to explain amplicon boundaries. Intriguingly, several *MYCN*-specific enhancers were found to be commonly co-amplified (Fig. 2b). The distal *MYCN*-specific CRC-driven enhancer, e4, was part of the consensus amplicon region in 90% of cases. Randomizing amplicon boundaries around *MYCN* showed that e4 co-amplification was significantly enriched on *MYCN* amplicons (empirical *P*=0.0003). Co-amplification frequency quickly dropped downstream of e4, suggesting that *MYCN*-specific, CRC-driven enhancers are a determinant of *MYCN* amplicon structure and may be required for *MYCN* expression, even in the context of high-level amplification.

Distal CRC-driven super enhancers are significantly co-amplified with MYCN in neuroblastoma

We and others have previously described chimeric *MYCN* amplicons[10] containing distal chromosomal fragments. We therefore systematically inspected *MYCN*-distal regions on chromosome 2 for signs of co-amplification. Distinct regions were statistically enriched for co-amplification with *MYCN* (Fig. 2c). In line with previous reports[22], significant co-amplification of 19 protein-coding genes, including known neuroblastoma drivers such as *ODC1*, *GREB1* and *ALK* occurred in *MYCN*-amplified neuroblastoma. Intriguingly, co-amplification of distal CRC-driven super enhancers (SE) occurred in 23.3% of samples. Seven specific CRC-driven SEs were significantly co-amplified more often than expected by chance. Most of these SEs were found in gene-rich regions, precluding to determine whether genes or regulatory elements were driving co-amplification. One significantly co-amplified CRC-driven SE, however, was found in a gene-poor region in 2p25.2, where most co-amplified segments did not overlap protein-coding genes (Fig. 2c). This raised the question whether hijacking of such distal regulatory elements may explain co-amplification with *MYCN*.

Enhancers remain functional on MYCN amplicons

Based on our amplicon boundary analysis, two classes of *MYCN* amplicons could be distinguished in neuroblastoma, i. amplicons containing local MYCN-specific enhancers, including e4, (here referred to as class I amplicons; Fig. 3a) and ii. amplicons lacking local *MYCN*-specific enhancers, and at least lacking e4 (referred to as class II amplicons; Fig. 3b). To determine whether co-amplified enhancers were active, we acquired genomic (long- and short-read whole genome sequencing) and epigenomic (ATAC-seq and H3K4me1 and H3K27ac ChIP-seq) data for two neuroblastoma cell lines with class I amplicons (Kelly and NGP) and two neuroblastoma cell lines with class II amplicons (IMR-5/75 and CHP-212). Notably, H3K27ac signal-to-noise ratio was lower on *MYCN* amplicons than in non-amplified regions. While the fraction of reads in peaks on the amplicon did not clearly differ between the amplicon and randomly drawn genomic regions, we observed more peaks than for non-amplified regions (Supplementary Fig. 3). These peaks were characterized by a lower relative signal compared to the amplicon background signal, indicating a larger variety of active regulatory regions on different *MYCN* amplicons. Using Nanopore long read-based *de novo* assembly, we reconstructed the *MYCN* neighborhood, confirming that *MYCN* and e4 were not only co-amplified in class I amplicons, but also lacked large rearrangements, which could preclude enhancer-promoter interaction (Supplementary Fig. 4-5). Enhancer e4 was characterized by increased chromatin accessibility and active enhancer histone marks as determined by ATAC-seq, H3K4me1 and H3K27ac ChIP-seq (Fig. 3c). Importantly, 4C chromatin conformation capture analysis showed that e4 spatially interacted with the *MYCN* promotor on the amplicon (Fig. 3c). Thus, e4 presents as a functional enhancer and appears to contribute to *MYCN* expression even in the context of class I *MYCN* amplification.

Super enhancer hijacking compensates for the loss of local enhancers on chimeric intra- and extrachromosomal circular MYCN amplicons

In contrast to class I amplicons, class II amplicons did not include local enhancers, raising the possibility of alternative routes of *MYCN* regulation. The lack of a strong local regulatory element on class II amplicons and our observation of frequent co-amplification of distal SE (Fig. 2c), led us to hypothesize that ectopic enhancers might be recruited to enable *MYCN* expression in class II amplicons. In line with our hypothesis, primary neuroblastomas with class II amplicons were more likely to harbor complex amplifications containing more than one fragment (66.7% vs. 35.7%, Fisher's Exact Test *P*=0.003; Fig. 3e). All class II amplicons co-amplified at least one CRC-driven super enhancer element distal of *MYCN*. Some enhancers were recurrently found on class II amplicons, including an enhancer 1.2Mb downstream of *MYCN* that was co-amplified in 20.8% (5/24) of *MYCN*-amplified neuroblastomas, 2.1-fold higher than expected for randomized amplicons that include *MYCN* but not e4 (Fig. 3f). Thus, class II *MYCN* amplicons are of high chimeric structural complexity allowing for the replacement of local enhancers through hijacking of distal CRC-driven enhancers.

To determine the structure and epigenetic regulation of class II amplicons in detail, we inspected long-read based *de novo* assemblies and short read-based reconstructions of IMR-5/75 and CHP-212 *MYCN* amplicons. IMR-5/75 was characterized by a linear HSR class II *MYCN* amplicon, not including e3-e5 (Fig. 3b). Inspection of the IMR-5/75 *MYCN* amplicon structure revealed that the amplicon consisted of six distant genomic regions, which were joined together to form a large and complex chimeric amplicon (Fig. 4a-d). In line with enhancer hijacking, an intronic segment of *ALK* containing a large super enhancer, marked by H3K27ac modification and chromatin accessibility, was juxtaposed with *MYCN* on the chimeric amplicon. Similar to e4, this enhancer was bound by adrenergic CRC factors in non-amplified cells (Supplementary Fig. 6a). Notably, a CTCF-bound putative insulator was added to the amplicon by yet another distal fragment (Fig. 4a-c, Supplementary Fig. 6a). In CHP-212, *MYCN* is amplified on extrachromosomal circular DNA, as confirmed by fluorescence in situ hybridization (Supplementary Fig. 7). Both *de novo* assembly and short-read based reconstruction of the amplicon confirmed the circular *MYCN* amplicon structure independently (Fig. 4f-h). Similar to IMR-5/75, distal fragments containing CRC-driven SEs and putative CTCF-bound insulators were joined to the *MYCN* neighborhood (Fig. 4e-g, Supplementary Fig. 6b).

To analyze the interaction profile in circular and linear amplicons we performed Hi-C and mapped the reads to the reconstructed amplicon (Fig. 4c, g). This analysis supported the genomic sequencing-based reconstruction of the amplicon, recapitulating the order and orientation of the joined fragments and confirmed that the ectopic enhancers spatially interacted with *MYCN*. Notably, high-frequency interactions in the corners of the maps opposite to the main diagonal, confirmed the circularity of CHP-212 amplicon and the presence of tandem amplification in IMR-5/75. In IMR-5/75 and CHP-212, we observed insulated TADs, boundaries and loops as in the rest of the genome. Due to the rearrangements in CHP-212, the *MYCN* gene became part of a neo-TAD consisting of a sub-TAD that originated from the wild type genome as an intact unit, and a second sub-TAD that resulted from the fusion and co-amplification of the first region with another region from a distal part of chromosome 2 (chr2:12.6-12.8Mb), containing multiple CRC-driven SEs (Fig. 4g, Supplementary Fig. 6b). Since the fused segments are now part of one TAD and not separated by a boundary, *MYCN* interaction with the SEs in this region becomes possible. A similar situation was observed for the linear amplicon. In IMR-5/75, Hi-C showed frequent contacts between *MYCN* and SEs from the genomic regions juxtaposed to *MYCN***,** containing intronic parts of *ALK* (Fig. 4c, Supplementary Fig. 6a). The map also reflected the high complexity and genomic heterogeneity of the IMR-5/75 amplicon. Nevertheless, the TAD structure, boundaries and loops were clearly visible on the reconstructed Hi-C map. Thus, hijacking of ectopic enhancers and insulators can compensate for the loss of endogenous regulatory elements on intra- and extrachromosomal circular *MYCN* amplicons via the formation of neo-TADs, which may explain the higher structural complexity of *MYCN* amplicons lacking endogenous enhancers.

Nanopore long-read DNA sequencing can be used for parallel assessment of MYCN amplicon structure and epigenetic regulation

In addition to allowing the alignment-free *de novo* assembly of the *MYCN* amplicon in several samples (Fig. 4b-d, f-h, Supplementary Fig. 4-5), Nanopore sequencing also allows for the direct measurement of DNA methylation without the need for bisulfite conversion (Fig. 5a)[23]. While DNA methylation at regulatory elements is often associated with repression, a trough in DNA methylation may indicate a transcription factor binding event, a poised or active gene regulatory element, or a CTCF-occupied insulator element (Fig. 5b). In theory, Nanopore sequencing and assembly might allow for the simultaneous inference of both structure and regulatory landscape (Fig. 5b). Prior to evaluating the *MYCN* amplicons, the DNA methylation landscape of highly expressed and inactive genes demonstrated the expected distribution of decreased methylation at active promoters and increased methylation within active gene bodies (Fig. 5c). In order to assess the DNA methylation status of putative regulatory elements near

MYCN, we first used the amplicon-enriched ATAC-seq peaks to classify relevant motif signatures (Fig. 5d). While *MYCN* was surrounded by the expected CRC-driven regulatory elements at the overlapping core enhancers as well as some CTCF sites, both their number and location varied, indicative of sample-specific sites of regulation. Indeed, DNA methylation decreased in accordance with sites specific to a given sample (Fig. 5e), opening up the possibility of using these data to infer regulatory elements in patient samples when no orthogonal epigenomic data are available.

Class II MYCN amplicons clinically phenocopy class I amplicons

MYCN-amplified neuroblastoma is characterized by significant clinical heterogeneity, which cannot entirely be explained genetically. Whether the structure of the *MYCN* amplicon itself could account for some of this variation is currently unknown. In line with previous reports[22], higher counts of amplified fragments were associated with a more malignant clinical phenotype (Fig. 6a). Co-amplification of *ODC1,* a gene located 5.5Mb upstream of *MYCN* and co-amplified in 9% (21/240) of *MYCN*-amplified neuroblastomas (Fig. 2c), defined an ultra-high risk genetical subgroup of *MYCN*-amplified neuroblastoma (HR 2.3 (1.4-3.7), Log-rank test *P*=0.001; Fig. 6b). Similarly, *ALK* co-amplification, present in in 5% (12/240) of *MYCN*-amplified tumors, was also associated with adverse clinical outcome (HR 1.8 (0.94-3.4), Log-rank test *P*=0.073; Fig. 6c). In contrast, differences in the *MYCN* amplicon enhancer structure, i.e. class II amplification, did not confer prognostic differences (HR 1.3 (0.78-2.1), Log rank test *P*=0.34; Fig. 6d). We therefore conclude that chimeric co-amplification of proto-oncogenes partly explain the malignant phenotype of neuroblastomas with complex *MYCN* amplicons, whereas enhancer hijacking in class II amplicons does not change clinical behavior, fully phenocopying class I *MYCN* amplicons.

Discussion

Here, we show that neuroblastoma-specific CRC-driven enhancers contribute to *MYCN* amplicon structure in neuroblastoma and retain the classic features of active enhancers after genomic amplification. While most *MYCN* amplicons contain local enhancers, ectopic enhancers are regularly incorporated into chimeric amplicons lacking local enhancers, leading to enhancer hijacking.

A large subset of neuroblastomas was recently found to be driven by a small set of transcription factors that form a self-sustaining core regulatory circuit, defined by their high expression and presence of super-enhancers[15-18]. In how far *MYCN* itself is directly regulated by CRC factors was previously unclear, particularly due to the challenging interpretation of epigenomic data on amplicons[16]. Our results provide empiric evidence that *MYCN* is driven by CRC factors even in the context of *MYCN* amplification. This is in line with and can mechanistically explain the previous observation that genetic depletion of CRC factors represses *MYCN* expression even in *MYCN*-amplified cells[16]. The finding that ectopic enhancers driven by the CRC are juxtaposed to *MYCN* on amplicons that lack local enhancers further strengthens the relevance of the CRC in *MYCN* regulation.

In line with our observation of local enhancer co-amplification, Morton et al. recently described that local enhancers are significantly co-amplified with other proto-oncogenes in other cancer entities[19]. They showed that experimentally interfering with local *EGFR* enhancers in *EGFR*-amplified glioblastoma impaired oncogene expression and cell viability in *EGFR*-amplified as well as non-amplified cases. In line with our findings, a region overlapping e4 was identified to be significantly co-amplified in *MYCN*-amplified neuroblastomas, corresponding to class I amplicons observed in our cohort. In contrast to Morton et al., who suggest that the inclusion of local enhancers is necessary for proto-oncogene expression on amplicons, we show that exceptions to this rule occur in a significant subset of *MYCN* amplified neuroblastomas. In such cases, amplicons are of highly complex chimeric structure enabling the reshuffling of ectopic enhancers and insulators to form neo-TADs that can compensate for disrupted local neighborhoods through enhancer hijacking.

More generally, we show that TADs also form in ecDNA, in line with recent findings by Wu et al.[24]. We extend this observation to homogeneously staining regions, which form extremely expanded stretches of chromatin in interphase nuclei and lose chromosomal territoriality[25]. Gene activation by enhancer adoption requires the fusion of distant DNA fragments and the formation of new chromatin domains, called neo-TADs[26]. This fusion requires a convergent directionality of CTCF sites in order to form a new boundary. Only in this case, aberrant gene activation is possible[27].

Reconstruction of amplicons has previously relied on combining structural breakpoint coordinates to infer the underlying structure. This regularly resulted in ambiguous amplicon reconstructions, which had to be addressed by secondary data such as Chromium linked reads or optical mapping[4,6,24]. We demonstrate the feasibility of long-read *de novo* assembly for the reconstruction of amplified genomic neighborhoods. *De novo* assembly was able to reconstruct entire ecDNA molecules and confirm the tandem duplicating nature of homogeneously staining regions. Integrating *de novo* assembly with methylation data from Nanopore sequencing reads will likely benefit further studies of other proto-oncogene-containing amplicons by enabling the characterization of the interplay between structure and regulation in highly rearranged cancer genomes.

Functional studies have shown that both *ODC1* and *ALK* are highly relevant in neuroblastoma [28,29]. Co-amplification with *MYCN* has been reported before[22], but to our knowledge the clinical relevance of co-amplification had not been determined so far. Similar to our previous observations of *PTP4A2* co-amplification on chimeric ecDNA[10], we demonstrate here that proto-oncogenes reside side-by-side on the same extrachromosomal circular DNAs, sometimes even sharing the same regulatory neighborhood. It is tempting to speculate that this structural coupling of genes could confer MYCN-independent but *MYCN*-amplicon-specific, collateral therapeutic vulnerabilities in *MYCN*-amplified tumors.

We conclude that the structure of genomic amplifications can be explained by selective pressure not only on oncogenic coding elements, but also on non-coding regulatory elements. CRC-driven enhancers are required for successful *MYCN* amplification and remain functional throughout this process. Even though the majority of amplicons contain endogenous enhancers, these can be replaced by ectopic CRC-driven elements that are juxtaposed to the oncogene through complex chimeric amplicon formation. We envision that our findings also extend to oncogene amplifications in other cancers and will help identify functionally relevant loci amongst the diverse array of complex aberrations that drive cancer.

Materials and Methods

Cell lines

Neuroblastoma cell lines (CHP-212, IMR-5/75, NGP, Kelly) were a gift from from Carol J. Thiele, obtained from the German Collection of Microorganisms and Cell Cultures or obtained from the American Type Culture Collection. Cell line identity was verified by STR genotyping (Genetica DNA Laboratories, Burlington, NC and IDEXX BioResearch, Westbrook, ME) and absence of *Mycoplasma sp.* contamination was determined with a Lonza MycoAlert system (Lonza Group Ltd., Basel, CH). All cell lines were cultured in RPMI-1640 medium (Thermo Fisher Scientific, Inc., Waltham, MA) with 1% Penicillin/Streptomycin, and 10% FCS.

RNA-seq

Public RNA-seq data was downloaded from Gene Expression Omnibus (GSE90683)[15]. FASTQ files were quality controlled (FASTQC 0.11.8) and adapters were trimmed (BBMap 38.58). We mapped reads to GRCh37 (STAR 2.7.1 with default parameters), counted them per gene

(Ensembl release 75, featureCounts from Subread package 1.6.4) and normalized for library size and composition (sizeFactors from DESeq2 1.22.2).

ChIP-seq

As reported before[27], cells were digested with Trypsin–EDTA 0.05% (Gibco) for 10 min at 37 °C. The cells were mixed with 10% FCS–PBS, and a single-cell suspension was obtained using a 40-µm cell strainer [30]. After centrifugation, cells were resuspended in 10% FCS-PBS again and fixed in 1% paraformaldehyde (PFA) for 10 min at room temperature and reaction quenched with 2.5M glycine (Merck) on ice and centrifuged at 400g for 8min. Pelleted cells were then resuspended in lysis buffer (50 mM Tris, pH 7.5; 150 mM NaCl; 5 mM EDTA; 0.5% NP-40; 1.15% Triton X-100; protease inhibitors (Roche)), and nuclei were pelleted again by centrifugation at 750g for 5min. For sonication, nuclei were resuspended in sonication buffer (10 mM Tris–HCl, pH 8.0; 100 mM NaCl; 1 mM EDTA; 0.5 mM EGTA; 0,1% Na-deoxycholate; 0.5% *N*-lauroylsarcosine; protease inhibitors (Roche complete)). Chromatin was sheared using a Bioruptor until reaching a fragment size of 200–500 base pairs (bp). Lysates were clarified from sonicated nuclei, and protein–DNA complexes were immunoprecipitated overnight at 4 °C with the respective antibody. A total of 10–15 µg chromatin was used for each replicate of histone ChIP and 20-25µg of transcription factor ChIP. Anti-H3K27ac (Diagenode; c15410037; A1657D), anti-H3K4m1 (Abcam; ab8895; Lot A1657D), anti-RAD21 (Abcam; ab992; Lot GR221348-8) and anti-CTCF (Active Motif; 613111; Lot 34614003) antibodies were used. Sequencing libraries were prepared using standard Nextera adapters (Illumina) according to the supplier's recommendations. 25 million reads per sample were sequenced on HiSeq 2500 sequencer (Illumina) in 50bp single read mode.

Additional public ChIP-seq FASTQ files were downloaded from Gene Expression Omnibus (GSE90683, GSE24447 and GSE28874)[15,31]. FASTQ files were quality controlled (FASTQC 0.11.8) and adapters were trimmed (BBMap 38.58). Reads were then aligned to hg19 (BWA-MEM 0.7.15 with default parameters) and duplicate reads removed (Picard 2.20.4). We generated BigWig tracks by extending reads to 200bp for single-end libraries and extending to fragment size for paired-end libraries, filtering by ENCODE DAC blacklist and normalizing to counts per million in 10bp bins (Deeptools 3.3.0). Peaks were called using MACS2 (2.1.2) with default parameters. Super enhancers were called for H3K27ac data using LILY (https://github.com/BoevaLab/LILY) with default parameters. ChIP-seq data was quality controlled using RSC and NSC (Phantompeakqualtools 1.2.1).

ATAC-seq

ATAC-seq samples were processed as reported in Buenrostro et al[32]. 5x10^5 cells were used per sample. For sequencing, libraries were generated using Illumina/Nextera adapters and size selected (100–1000bp) with AMPure Beads (Beckman Coulter). Approximately 100 million 75bp paired-end reads were acquired per sample on the HiSeq 2500 system (Illumina). Additional public ATAC-seq FASTQ files were downloaded from Gene Expression Omnibus (GSE80154)[33]. Adapter trimming, alignment and duplicate removal as for ChIP-seq. We generated BigWig tracks by extending paired-end reads to fragment size, filtering by the ENCODE DAC blacklist and normalizing to counts per million in 10bp bins (Deeptools 3.3.0). Peaks were called using MACS2 (2.1.2) with default parameters.

Hi-C

3C libraries for Hi-C and 4C were prepared from confluent neuroblastoma cells according to the cell culture section above. Hi-C experiments were performed as duplicates. Cells were washed twice with PBS and digested with Trypsin–EDTA 0.05% (Gibco) for 10 min at 37 °C. To obtain a single cell suspension, cells were pipetted through a 40-µm cell strainer [30].

After centrifugation at 300g for 5min, cell pellets were resuspended with 10% FCS and fixed by adding an equal volume of 4% formaldehyde (Sigma-Aldrich) and mixed for 10 min at room temperature while shaking. Fixation was quenched using 1.425 M glycine (Merck) on ice and immediately centrifuged at 400g for 8 min. Pelleted cells were then resuspended in lysis buffer (50 mM Tris, pH 7.5; 150 mM NaCl; 5 mM EDTA; 0.5% NP-40; 1.15% Triton X-100; protease inhibitors (Roche)), and nuclei were pelleted again by centrifugation at 750g for 5min.

The pellet was washed with 1x DpnII buffer, resuspended in 50µl 0.5% SDS and incubated for 10min at 62°C. After that 145µl water and 25µl 10% Triton (Sigma) was added to quench the SDS. After a 37°C incubation, 25µl DpnII buffer and 100U DpnII was added. The digestion reaction was incubated for 2h at 37°C, after 1h another 10U were added. After the digestion, DpnII was inactivated at 65°C for 20min.

The digested sticky ends were filled up with 10mM dNTPs (without dATP) and 0.4mM biotin-14-dATP (Life Technologies) and 40U DNA Pol I, Large Klenow (New England BioLabs, Inc. (NEB), Ipswich, MA) at 37°C for 90min. Biotinylated blunt ends were then ligated using a ligation reaction (663µl water, 120µl 10X NEB T4 DNA ligase buffer (NEB), 100µl 10% Triton X-100 (Sigma), 12µl 10mg/ml BSA and 2400U of T4 DNA liagse (NEB)) overnight at 16°C with slow rotation.

The 3C library was then sheared using a Covaris sonicator (duty cycle: 10%; intensity: 5; cycles per burst: 200; time: 6 cycles of 60 s each; set mode: frequency sweeping; temperature: 4–7 °C). After sonication, religated DNA was pulled down using 150µl of 10mg/ml Dynabeads Streptavidin T1 beads (Thermo Fisher) according to the supplier's recommendation. Sheared and pulled down DNA was treated using a 100µl end-repair reaction (25mM dNTPs, 50U NEB PNK T4 Enzyme, 12U NEB T4 DNA polymerase, 5U NEB DNA pol I, Large (Klenow) Fragment, 10X NEB T4 DNA ligase buffer with 10mM ATP) and incubated for 30min at 37°C. Universal sequencing adaptor were added using the NEBnext Ultra DNA Library Kit (NEB) according to the supplier's recommendation. Samples were sequenced with Ilumina Hi-Seq technology according to standard protocols and 75bp PE mode. 200 million reads were generated for IMR-5/75, 5 million reads per sample were generated for all other cell lines.

FASTQ files were processed using the Juicer pipeline v1.5.6, CPU version[34], which was set up with BWA v0.7.17[35] to map short reads to reference genome hg19, from which haplotype sequences were removed and to which the sequence of Epstein-Stein-Barr Virus (NC_007605.1) was added. Replicates were processed individually. Mapped and filtered reads were merged afterwards. A threshold of MAPQ≥30 was applied for the generation of Hi-C maps with Juicer tools v1.7.5[34]. Knight-Ruiz normalization of Hi-C signal was used for Hi-C maps. Virtual 4C signal for the *MYCN* locus was generated by the mean Knight-Ruiz-normalized Hi-C signal across three 5kb bins (chr2:16,075,000-16,085,000).

4C-seq

4C-seq libraries were generated as described before[26], using a starting material of $5x10^6$ – $1x10^7$ cells. The fixation and lysis were performed as described in the Hi-C section. For the *MYCN* promotor viewpoint, 1.6 µg DNA was amplified by PCR (Primer 1 5'-GCAGAATCGCCTCCG-3', Primer 2 5'-CCTGGCTCTGCTTCCTAG-3'). For the viewpoint, 4bp cutters were used. DpnII (NEB) was used as first cutter and Csp6I (NEB) as second cutter. All samples were sequenced with the HiSeq 2500 (Illumina) technology according to standard protocols and with 8 million reads per sample.

Reads were pre-processed, filtered for artefacts and mapped to the reference genome GRCh37 using BWA-MEM as described earlier[26]. After removing the viewpoint fragment as well as 1.5 kb up- and downstream of the viewpoint the raw read counts were normalized per million mapped reads (RPM) and a window of 10 fragments was chosen to smooth the profile.

Whole-genome sequencing

Cells were harvested and DNA was extracted using the NucleoSpin Tissue kit (Macherey-Nagel GmbH & Co. KG, Düren, Germany). Libraries for whole genome sequencing were prepared with the NEBNext Ultra II FS DNA Library Prep Kit for Illumina (New England BioLabs, Inc., Ipswich, MA). Libraries were sequenced on a NovaSeq S1 flow cell (Illumina, Inc., San Diego, CA) with 2x150bp paired-end reads. Quality control, adapter trimming, alignment, duplicate removal as for ChIP-seq data. Copy number variation was called (Control-FREEC[36] 11.4 with default parameters). Structural variants were called using SvABA[37] (1.1.1) in germline mode and discarding regions in a blacklist provided by SvABA (https://data.broadinstitute.org/snowman/svaba_exclusions.bed).

Nanopore Sequencing

Cells were harvested and high molecular weight DNA was extracted using the MagAttract HMW DNA Kit (Qiagen N.V., Venlo, Netherlands). Size selection was performed to remove fragments <10 kilobases (kb) using the Circulomics SRE kit (Circulomics Inc., Baltimore, MD). DNA content was measured with a Qubit 3.0 Fluorometer (Thermo Fisher) and sample quality control was performed using a 4200 TapeStation System (Agilent Technologies, Inc., Santa Clara, CA). Libraries were prepared using the Ligation Sequencing Kit (SQK-LSK109, Oxford Nanopore Technologies Ltd., Oxford, UK) and sequenced on a R9.4.1 MinION flowcell (FLO-MIN106, Oxford Nanopore Technologies Ltd., Oxford, UK). Quality control was performed using NanoPlot 1.0.0. For the NGP cell line, DNA was extracted with the NucleoSpin Tissue kit (Macherey-Nagel GmbH & Co. KG, Düren, Germany) and libraries prepared using the ONT Rapid Kit (SQK-RBK004, Oxford Nanopore Technologies Ltd., Oxford, UK). Guppy 2.3.7 (Oxford Nanopore Technologies Ltd., Oxford, UK) was used for basecalling with default parameters. For de novo assembly, Flye 2.4.2[38] was run in metagenomics assembly mode on the unfiltered FASTQ files with an estimated genome size of 1Gb. Contigs were mapped back to hg19 using minimap2 2.16 with parameter -ax asm5. Assembly results were visualized with Bandage 0.8.1 (https://rrwick.github.io/Bandage) and Ribbon (no version available, https://github.com/MariaNattestad/Ribbon). CpG methylation was called from the unfiltered raw FAST5 files using Megalodon 0.1.0 (Oxford Nanopore Technologies Ltd., Oxford, UK).

Fluorescence in situ hybridization

Cells were grown to 200,000 per well in six-well plates and metaphase-arrested using Colcemid (20µl/2ml; Roche #10295892001) for 30min-3h, trypsinized, centrifuged (1000rpm/10min) washed and pelleted. 5ml 0.4% KCl (4°C; Roth #6781.1) was added to the pellet and incubated for 10min. 1ml KCl and 1ml MeOH/acetic acid 3:1 (Roth #4627.2, #KK62.1) was added drop-wise. 2/5/5ml of MeOH/acetic acid were added in between centrifugation steps (1000 rpm/10min) respectively. Suspension was dropped on a slide from a height of 40cm. Slides were washed with PBS (Gibco, #70011036) and digested for 10min in 0,04% pepsin solution in 0,001N HCl. Slides were washed in 0.5x SSC, dehydrated with 70%/80%/100% EtOH (3min each) and air-dried. 10µl of the probe (Vysis LSI N-MYC; #07J72-001; Lot #472123; Abbott Laboratories, Abbott Park, IL) were added and coverslips fixed on the slide. Slides were incubated at 75°C for 10min and at 37°C over night. The coverslip was removed and the slide washed in 0.4xSSC/0.3% IGEPAL (CA-630, #18896, Sigma-Aldrich Inc.) for 3min at 60°C and 2xSSC/0.1% IGEPAL for 3min at RT. 5µl DAPI (Vectashield, #H-1200, Vector) was added. A coverslip was added and fixed with nail polish.

Enhancer calling

MYCN-expressing cell lines were defined as cell lines with sizeFactor normalized expression of 100 or above based. We identified enhancer candidate regions in a ±500kb window around *MYCN*. We focused on regions with a H3K27ac peak in the majority of *MYCN*-expressing, non-*MYCN*-amplified cell lines, i.e. three or more. If the gap between two such regions was less than 2kb, they were joined. These regions were then ranked by the maximum difference in H3K27ac signal fold change between non-amplified, *MYCN*-expressing and non-expressing cell lines. We chose the five highest-ranking regions as candidate regulatory elements. Enhancer regions were screened for transcription factor binding sequences from the JASPAR2018 (http://jaspar2018.genereg.net/) and JASPAR2020 (http://jaspar2020.genereg.net/) database using the TFBSTools (1.20.0) function matchPWM with min.score='85%'. CRC-driven super enhancers were defined as all regions with a LILY-defined super enhancer in *MYCN*-expressing, non-*MYCN*-amplified cell lines that overlapped with a GATA3, HAND2 or PHOX2B peak in CLB-GA.

Analysis of copy number data

Public data was downloaded. Samples that were described as *MYCN*-amplified in the metadata but did not show *MYCN* amplification in the copy number profile were excluded. In order to generate an aggregate copy number profile, the genome was binned in 10kb bins and number of samples with overlapping amplifications was counted per bin. Randomized copy number profiles were generated by randomly sampling one of the original copy number profiles on chromosome 2 and randomly shifting it such that *MYCN* is still fully included within an amplified segment. For class I-specific shuffling, e4 had to be included as well; for class II-specific shuffling, e4 was never included on the randomly shifted amplicon. Empirical *P*-values for significant co-amplification were derived by creating 10,000 randomized datasets with each amplicon randomly shifted and comparing the observed co-amplification frequency to the distribution of co-amplification frequencies in the randomized data. Empirical *P*-values were always one-sided and adjusted for multiple comparisons using the Benjamini-Hochberg procedure.

Amplicon reconstruction

All unfiltered SvABA structural variant calls were filtered to exclude regions from the ENCODE blacklist[39] and small rearrangements of 1kb or less. As we were only aiming at the rearrangements common to all amplicons, we only considered breakpoints with more than 50 variant-support reads ('allele depth'). gGnome[40] was used to represent these data as a genome graph with nodes being breakpoint-free genomic intervals and edges being rearrangements ('alternate edge') or connections in the reference genomes ('reference edge'). We considered only nodes with high copy number, i.e. with a mean whole-genome sequencing coverage of at least 10-fold the median coverage of chromosome 2. Then, reference edges were removed if its corresponding alternate edge was among the 25% highest allele-depth edges. The resulting graph was then searched for the circular, *MYCN*-containing walk that included the highest number of nodes without using any node twice. We used gTrack (https://github.com/mskilab/gTrack) for visualization. For custom Hi-C maps of reconstructed amplicon sequences of CHP-212 and IMR-5-75, respectively, the corresponding regions from chromosome 2 were copied, ordered, oriented and compiled according to the results from the amplicon reconstruction and added to the reference genome. Additionally, these copied regions were masked with 'N' at the original locations on chromosome 2 to allow a proper mapping of reads to the amplicon sequence. The contribution of Hi-C di-tags from these regions on chromosome 2 to the amplicon Hi-C map is expected be minor, because the copy number of amplicons is much higher than the number of wild type alleles. Juicebox v1.11.08 was used to visualize Hi-C maps with a bin size of 5 kb and Knight-Ruiz normalization[41-43].

Data availability

Copy number data for high-risk neuroblastoma were downloaded from https://github.com/padpuydt/copynumber_HR_NB/. Sequencing data supporting the findings of this manuscript is available at the Gene Expression Omnibus under accessions GSE90683, GSE80152, GSE24447 and GSE28874. Sequencing data for primary neuroblastoma samples is available at the European Genome-Phenome archive under accessions EGAS00001001308 and EGAS00001004022. Corresponding BigWig und narrowPeak files can be downloaded from https://data.cyverse.org/dav-anon/iplant/home/konstantin/helmsaueretal/. An accompanying UCSC genome browser track hub is provided for ChIP-seq and ATAC-seq data visualization (https://de.cyverse.org/dl/d/27AA17DA-F24C-4BF4-904C-62B539A47DCC/hub.txt). All other data is available from the corresponding authors upon reasonable request.

Code availability

Code is available at https://github.com/henssenlab/MYCNAmplicon.

References

1 Turner, K. M. *et al.* Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. *Nature* **543**, 122-125, doi:10.1038/nature21356 (2017).

2 Zhang, C. Z. *et al.* Chromothripsis from DNA damage in micronuclei. *Nature* **522**, 179-184, doi:10.1038/nature14493 (2015).

3 Ly, P. *et al.* Chromosome segregation errors generate a diverse spectrum of simple and complex genomic rearrangements. *Nat Genet* **51**, 705-715, doi:10.1038/s41588-019-0360-8 (2019).

4 Deshpande, V. *et al.* Exploring the landscape of focal amplifications in cancer using AmpliconArchitect. *Nat Commun* **10**, 392, doi:10.1038/s41467-018-08200-y (2019).

5 Nathanson, D. A. *et al.* Targeted therapy resistance mediated by dynamic regulation of extrachromosomal mutant EGFR DNA. *Science* **343**, 72-76, doi:10.1126/science.1241328 (2014).

6 Xu, K. *et al.* Structure and evolution of double minutes in diagnosis and relapse brain tumors. *Acta Neuropathol* **137**, 123-137, doi:10.1007/s00401-018-1912-1 (2019).

7 deCarvalho, A. C. *et al.* Discordant inheritance of chromosomal and extrachromosomal DNA elements contributes to dynamic disease evolution in glioblastoma. *Nat Genet* **50**, 708-717, doi:10.1038/s41588-018-0105-0 (2018).

8 Storlazzi, C. T. *et al.* Gene amplification as double minutes or homogeneously staining regions in solid tumors: origin and structure. *Genome Res* **20**, 1198-1206, doi:10.1101/gr.106252.110 (2010).

9 Wahl, G. M. The Importance of Circular DNA in Mammalian Gene Amplification. *Cancer Res* **49**, 1333-1340 (1989).

10 Koche, R. P. *et al.* Extrachromosomal circular DNA drives oncogenic genome remodeling in neuroblastoma. *Nature Genetics*, doi:10.1038/s41588-019-0547-z (2019).

11 Gröbner, S. N. *et al.* The landscape of genomic alterations across childhood cancers. *Nature* **555**, 321-327, doi:10.1038/nature25480 (2018).

12 Cohn, S. L. *et al.* The International Neuroblastoma Risk Group (INRG) classification system: an INRG Task Force report. *J Clin Oncol* **27**, 289-297, doi:10.1200/JCO.2008.16.6785 (2009).

13 Weiss, W. A., Aldape, K., Mohapatra, G., Feuerstein, B. G. & Bishop, J. M. Targeted expression of MYCN causes neuroblastoma in transgenic mice. *The EMBO Journal* **16**, 2985–2995 (1997).

14 Althoff, K. *et al.* A Cre-conditional MYCN-driven neuroblastoma mouse model as an improved tool for preclinical studies. *Oncogene* **34**, 3357-3368, doi:10.1038/onc.2014.269 (2015).

15 Boeva, V. *et al.* Heterogeneity of neuroblastoma cell identity defined by transcriptional circuitries. *Nat Genet* **49**, 1408-1413, doi:10.1038/ng.3921 (2017).

16 Durbin, A. D. *et al.* Selective gene dependencies in MYCN-amplified neuroblastoma include the core transcriptional regulatory circuitry. *Nat Genet* **50**, 1240-1246, doi:10.1038/s41588-018-0191-z (2018).

17 Decaesteker, B. *et al.* TBX2 is a neuroblastoma core regulatory circuitry component enhancing MYCN/FOXM1 reactivation of DREAM targets. *Nat Commun* **9**, 4866, doi:10.1038/s41467-018-06699-9 (2018).

18 Wang, L. *et al.* ASCL1 is a MYCN- and LMO1-dependent member of the adrenergic neuroblastoma core regulatory circuitry. *Nat Commun* **10**, 5622, doi:10.1038/s41467-019-13515-5 (2019).

19 Morton, A. R. *et al.* Functional Enhancers Shape Extrachromosomal Oncogene Amplifications. *Cell*, doi:10.1016/j.cell.2019.10.039 (2019).

20 Rajbhandari, P. *et al.* Cross-Cohort Analysis Identifies a TEAD4-MYCN Positive Feedback Loop as the Core Regulatory Element of High-Risk Neuroblastoma. *Cancer Discov* **8**, 582-599, doi:10.1158/2159-8290.CD-16-0861 (2018).

21 Depuydt, P. *et al.* Meta-mining of copy number profiles of high-risk neuroblastoma tumors. *Sci Data* **5**, 180240, doi:10.1038/sdata.2018.240 (2018).

22 Depuydt, P. *et al.* Genomic Amplifications and Distal 6q Loss: Novel Markers for Poor Survival in High-risk Neuroblastoma Patients. *J Natl Cancer Inst* **110**, 1084-1093, doi:10.1093/jnci/djy022 (2018).

23 Simpson, J. T. *et al.* Detecting DNA cytosine methylation using nanopore sequencing. *Nat Methods* **14**, 407-410, doi:10.1038/nmeth.4184 (2017).

24 Wu, S. *et al.* Circular ecDNA promotes accessible chromatin and high oncogene expression. *Nature*, doi:10.1038/s41586-019-1763-5 (2019).

25 Solovei, I. *et al.* Topology of double minutes (dmins) and homogeneously staining regions (HSRs) in nuclei of human neuroblastoma cell lines. *Genes Chromosomes Cancer* **29**, 297-308, doi:10.1002/1098-2264(2000)9999:9999<::aid-gcc1046>3.0.co;2-h (2000).

26 Franke, M. *et al.* Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* **538**, 265-269, doi:10.1038/nature19800 (2016).

27 Despang, A. *et al.* Functional dissection of the Sox9-Kcnj2 locus identifies nonessential and instructive roles of TAD architecture. *Nat Genet* **51**, 1263-1271, doi:10.1038/s41588-019-0466-z (2019).

28 Hogarty, M. D. *et al.* ODC1 is a critical determinant of MYCN oncogenesis and a therapeutic target in neuroblastoma. *Cancer Res* **68**, 9735-9745, doi:10.1158/0008-5472.CAN-07-6866 (2008).

29 Gamble, L. D. *et al.* Inhibition of polyamine synthesis and uptake reduces tumor progression and prolongs survival in mouse models of neuroblastoma. *Sci Transl Med* **11**, doi:10.1126/scitranslmed.aau1099 (2019).

30 Nikolaev, S. *et al.* Extrachromosomal driver mutations in glioblastoma and low-grade glioma. *Nat Commun* **5**, 5690, doi:10.1038/ncomms6690 (2014).

31 Rada-Iglesias, A. *et al.* Epigenomic annotation of enhancers predicts transcriptional regulators of human neural crest. *Cell Stem Cell* **11**, 633-648, doi:10.1016/j.stem.2012.07.006 (2012).

32 Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr Protoc Mol Biol* **109**, 21 29 21-21 29 29, doi:10.1002/0471142727.mb2129s109 (2015).

33 Zeid, R. *et al.* Enhancer invasion shapes MYCN-dependent transcriptional amplification in neuroblastoma. *Nat Genet* **50**, 515-523, doi:10.1038/s41588-018-0044-9 (2018).

34 Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**, 95-98, doi:10.1016/j.cels.2016.07.002 (2016).

35 Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-595, doi:10.1093/bioinformatics/btp698 (2010).

36 Boeva, V. *et al.* Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics* **28**, 423-425, doi:10.1093/bioinformatics/btr670 (2012).

37 Wala, J. A. *et al.* SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Res* **28**, 581-591, doi:10.1101/gr.221028.117 (2018).

38 Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* **37**, 540-546, doi:10.1038/s41587-019-0072-8 (2019).

39 Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci Rep* **9**, 9354, doi:10.1038/s41598-019-45839-z (2019).

40 Hadi, K. *et al.* Novel patterns of complex structural variation revealed across thousands of cancer genome graphs. *bioRxiv*, doi:10.1101/836296 (2019).

41 Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-1680, doi:10.1016/j.cell.2014.11.021 (2014).

42 Durand, N. C. *et al.* Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst* **3**, 99-101, doi:10.1016/j.cels.2015.07.012 (2016).

43 Knight, P. A. & Ruiz, D. A fast algorithm for matrix balancing. *IMA J Numer Anal* **33**, 1928–1047, doi:10.1093/imanum/drs019 (2013).

Acknowledgments

We thank the patients and their parents for granting access to the tumor specimen and clinical information that were analyzed in this study. We are grateful to Yingqian Zhan, Natalia Munoz Perez, Jennifer von Stebut, Victor Bardinet and Celine Chen for critical discussions. R.P.K is supported by the Berlin Institute of Health visiting professorship program. A.G.H. is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 398299703, the Wilhelm Sander Stiftung a participant in the BIH-Charité Clinical Scientist Program funded by the Charité – Universitätsmedizin Berlin and the Berlin Institute of Health. A.G.H. and K.H. are supported by Berliner Krebsgesellschaft e.V. K.H. is supported by Boehringer Ingelheim Funds. This work was also supported by the TransTumVar project – PN013600. We thank B. Hero, H. Düren, N. Hemstedt of the neuroblastoma biobank and neuroblastoma trial registry

(University Children's Hospital Cologne) of the German Society of Pediatric Oncology and Hematology (GPOH) for providing samples and clinical data.

Author Contributions

All authors contributed to the study design and collection and interpretation of the data. M.V. and S.A. acquired ChIP-seq, ATAC-seq, 4C and Hi-C data. R.C., L.P.K. and P.E. acquired nanopore sequencing data. K.K. acquired Illumina whole genome sequencing data. C.Rö. and C.Ro. performed FISH experiments. R.S., V.H. and K.H. analyzed 4C and Hi-C data. K.H. analyzed ChIP-seq, ATAC-seq and RNA-seq data. E.F., M.P., J.T. and K.H. analyzed Illumina whole-genome sequencing data. K.H. and R.P.K. analyzed nanopore sequencing data. M.F., F.H., A.K.-S. and J.H.S. collected and prepared patient samples. M.V., S.A., R.C., Y.B., H.D.G. and K.Ha. performed experiments and analyzed data. K.Ha., M.R., D.T. and J.H.S. contributed to study design. K.H., M.V., S.A., S.M., A.G.H. and R.P.K. led the study design, performed data analysis and wrote the manuscript, to which all authors contributed.

Competing interests

The authors have no competing interests to declare.

Materials & Correspondence

Request for materials can be made to A.G.H.

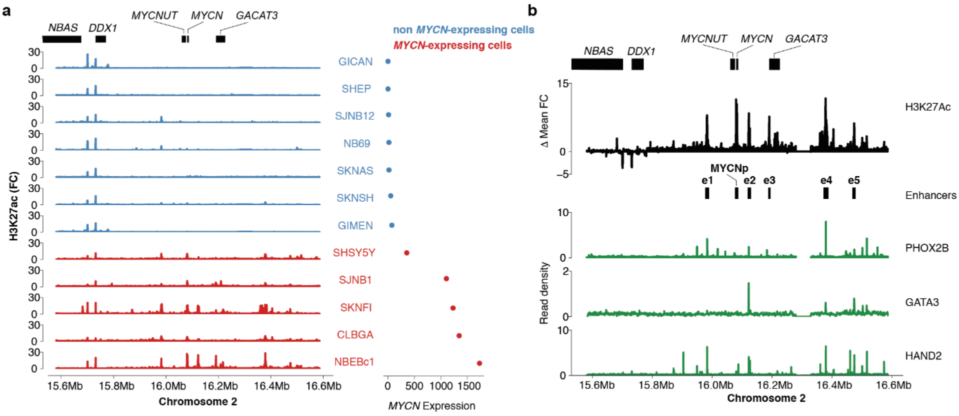


Figure 1. Five enhancers are specifically found in *MYCN*-expressing neuroblastoma cells. **a,** H3K27ac ChIP-seq fold change over input (left) and size-factor normalized *MYCN* expression as determined from RNA-seq for 12 non-MYCN-amplified neuroblastoma cell lines (*MYCN*-expressing, red; non *MYCN*-expressing, blue). **b,** Aggregated H3K27ac signal of *MYCN*-expressing compared to non-expressing cells (top; MYCNp, *MYCN* promotor; e1-e5, *MYCN*-specific enhancers). PHOX2B, GATA3 and HAND2 core regulatory circuit transcription factor ChIP-seq in a *MYCN*-expressing neuroblastoma cell line (green, CLB-GA).

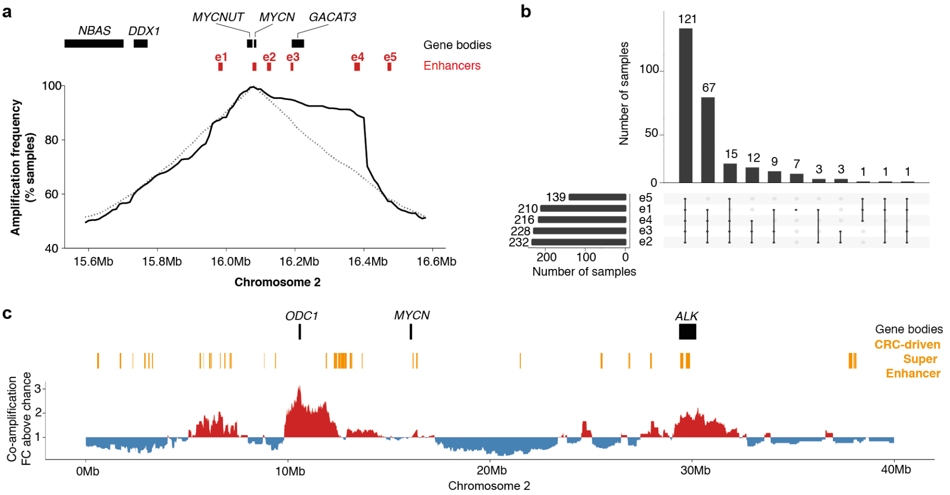


Fig. 2. *MYCN*-specific enhancer e4 is significantly co-amplified with *MYCN* and retains functional enhancer characteristics after amplification. **a,** Co-amplification frequency of the immediate *MYCN* neighborhood measured using copy number profiles from 240 *MYCN*-amplified neuroblastomas (solid line) compared to the expected co-amplification frequencies for randomized *MYCN*-containing amplicons (dashed line). **b,** Upset plot showing the co-amplification patterns of all five *MYCN*-specific local enhancers identified in neuroblastoma. **c,** Enrichment for co-amplification with *MYCN* of genomic regions on 2p (red, co-amplification more frequent than expected by chance; blue, co-amplification less frequent than expected by chance).

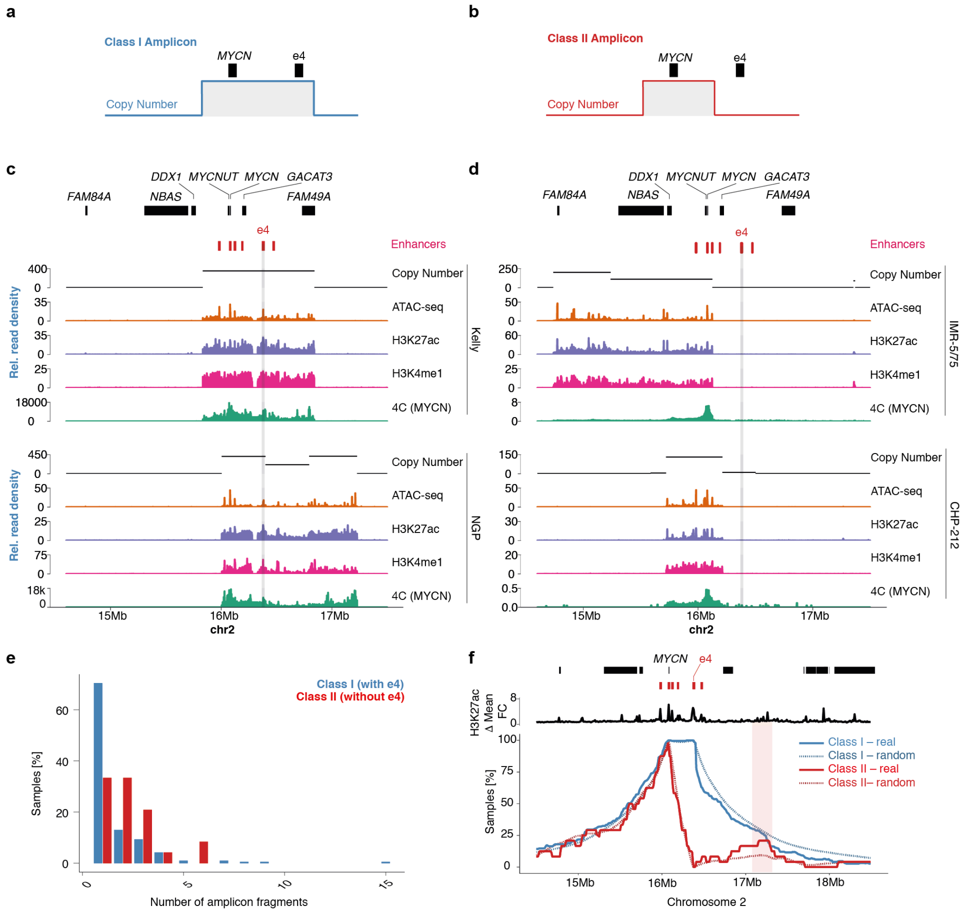


Fig. 3. Two classes of *MYCN* amplicons can be identified in neuroblastoma. Schematic representation of class I (**a**) and class II (**b**) *MYCN* amplicons. **c,** Copy number profile, ATAC-seq, H3K27ac ChIP-seq, H3K4me1 ChIP-seq and 4C (*MYCN* promotor as the viewpoint) for two neuroblastoma cell lines with class I amplicons, co-amplifying the e4. **d,** Copy number profile, ATAC-seq, H3K27ac ChIP-seq, H3K4me1 ChIP-seq and 4C (*MYCN* promotor as the viewpoint) for two neuroblastoma cell lines class II amplicons, not co-amplifying e4. **e,** Number of non-contiguous amplified fragments in class I vs. class II *MYCN* amplicons. **d** Amplicon boundary frequency relative to gene and enhancer positions in class I (blue) vs. class II (red) amplicons compared to random amplicon boundary frequencies (dotted lines).

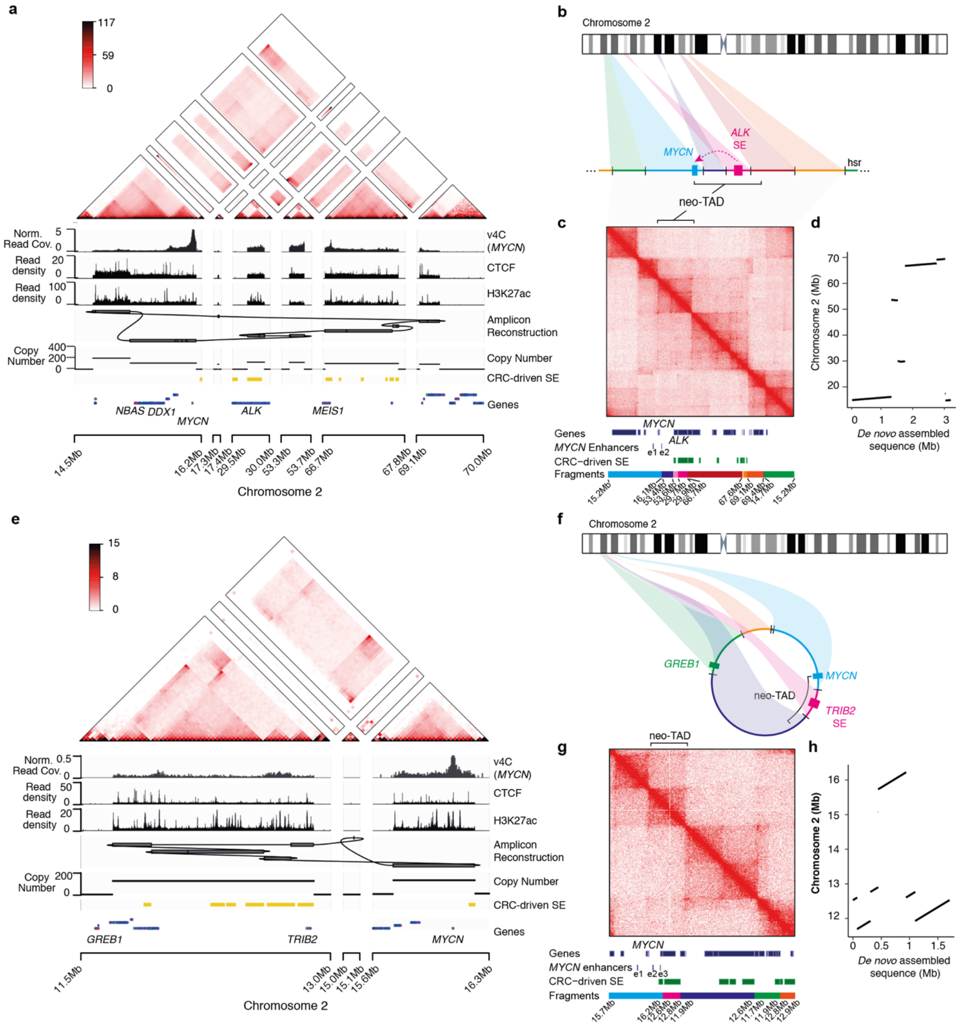
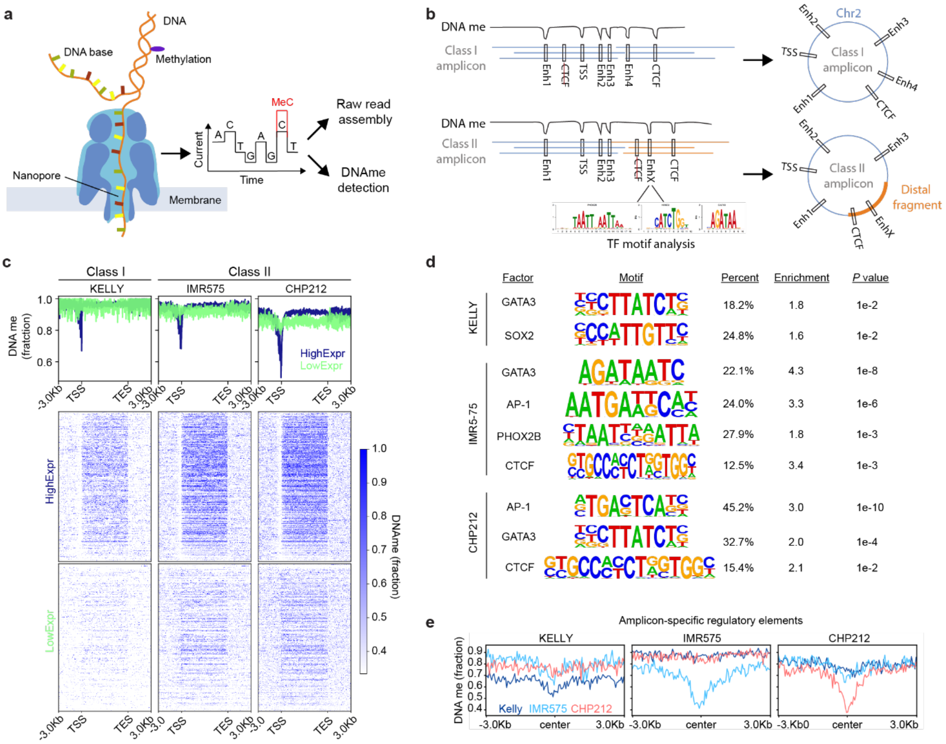


Figure 4. Reconstruction and epigenetic markup of class II intra- and extrachromosomal circular *MYCN* amplicons in neuroblastoma cells. a Short-read based reconstruction and epigenomic characterization of the *MYCN* amplicon in IMR-5/75 cells. Top to bottom: Hi-C map (color indicating Knight-Ruiz normalized read counts in 25kb bins), virtual 4C (MYCN viewpoint, v4C), CTCF ChIP-seq, H3K27Ac ChIP-seq, Amplicon reconstruction, copy number profile, super enhancer locations (yellow), gene positions (blue) (scale). **b** Schematic representation of the class II amplicon described in (**a**), showing ectopic enhancers and insulator reshuffling leading to locally disrupted regulatory neighborhoods on the HSR. **c** Alignment of Hi-C reads to the reconstructed *MYCN* amplicon in IMR-5/75 and positions of genes, local *MYCN* enhancers and CRC-driven super enhancers on the amplicon. **d** Mapping of the long

read sequencing-based *de novo* assembly of the *MYCN* amplicon in IMR5/75 on chromosome 2. **e** Short-read based reconstruction and epigenomic characterization of the *MYCN* amplicon in CHP-212 cells. Top to bottom: Hi-C map (color indicating Knight-Ruiz normalized read counts in 25kb bins), virtual 4C (MYCN viewpoint, v4C), CTCF ChIP-seq, H3K27Ac ChIP-seq, Amplicon reconstruction, copy number profile, super enhancer locations (yellow), gene positions (blue). **f**, Schematic representation of the class II amplicon described in (**e**), showing ectopic enhancers and insulator reshuffling leading to locally disrupted regulatory neighborhoods on extrachromosomal circular DNA. **g** Alignment of Hi-C reads to the reconstructed *MYCN* amplicon in CHP-212 and positions of genes, local *MYCN* enhancers and CRC-driven super enhancers on the amplicon. **h** Mapping of the long read sequencing-based *de novo* assembly of the *MYCN* amplicon in CHP-212 on chromosome 2.



	Factor	Motif	Percent	Enrichment	*P* value
KELLY	GATA3	TCCTTATCTC	18.2%	1.8	1e-2
	SOX2	CCCATTGTTC	24.8%	1.6	1e-2
IMR5-75	GATA3	AGATAATC	22.1%	4.3	1e-8
	AP-1	AATGATTCAT	24.0%	3.3	1e-6
	PHOX2B	CTAATTAAATTA	27.9%	1.8	1e-3
	CTCF	GTGCCACCTGGTGGC	12.5%	3.4	1e-3
CHP212	AP-1	ATGACTCATC	45.2%	3.0	1e-10
	GATA3	TCCTTATCTC	32.7%	2.0	1e-4
	CTCF	GTGCCACCTGGTGGC	15.4%	2.1	1e-2

Figure 5. Nanopore long read sequencing allows for the simultaneous characterization of amplicon structure and DNA methylation. a, Schematic of experimental approach. **b,** Schematic representation of how Nanopore sequencing facilitates *de novo* amplicon assembly and can be used to simultaneously to detect regulatory elements through DNA methylation analysis. **c,** Composite DNA methylation signal detected using Nanopore sequencing over genes expressed at high (blue) vs. low (green) levels. **d,** Motif analysis based on accessibility in regulatory elements co-amplified on *MYCN* amplicons. **e,** Amplicon-specific methylation pattern detected in three neuroblastoma cell lines using Nanopore sequencing-based DNA methylation analysis.

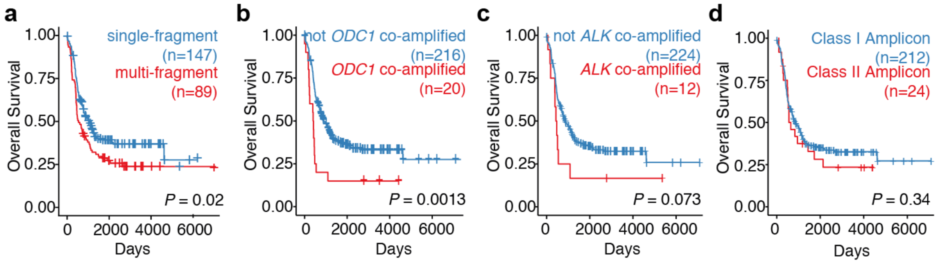


Figure 6. Class II amplicons clinically phenocopy class I amplicons. Kaplan Meier survival analysis of patients with *MYCN*-amplified neuroblastoma, comparing single-fragment vs. non-contiguously amplified *MYCN* amplicons (**a**), co-amplification of *ODC1* vs. no co-amplification (**b**), co-amplification of *ALK* vs. no co-amplification, and class I amplicons vs. class II amplicons (**d**; *N*=236 *MYCN*-amplified neuroblastomas; *P*- value based on two-sided log rank test).

Appendix 4

Appendix publication 2

Re-analysis of public genetic data reveals a rare X-chromosomal variant associated with type 2 diabetes

nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-017-02380-9 OPEN

Corrected: Publisher correction

Re-analysis of public genetic data reveals a rare X-chromosomal variant associated with type 2 diabetes

Sílvia Bonàs-Guarch et al.[#]

The reanalysis of existing GWAS data represents a powerful and cost-effective opportunity to gain insights into the genetics of complex diseases. By reanalyzing publicly available type 2 diabetes (T2D) genome-wide association studies (GWAS) data for 70,127 subjects, we identify seven novel associated regions, five driven by common variants (*LYPLAL1*, *NEUROG3*, *CAMKK2*, *ABO*, and *GIP* genes), one by a low-frequency (*EHMT2*), and one driven by a rare variant in chromosome Xq23, rs146662075, associated with a twofold increased risk for T2D in males. rs146662075 is located within an active enhancer associated with the expression of Angiotensin II Receptor type 2 gene (*AGTR2*), a modulator of insulin sensitivity, and exhibits allelic specific activity in muscle cells. Beyond providing insights into the genetics and pathophysiology of T2D, these results also underscore the value of reanalyzing publicly available data using novel genetic resources and analytical approaches.

Correspondence and requests for materials should be addressed to J.M.M. (email: mercader@broadinstitute.org) or to D.T. (email: david.torrents@bsc.es)
#A full list of authors and their affliations appears at the end of the paper

During the last decade, hundreds of genome-wide association studies (GWAS) have been performed with the aim of providing a better understanding of the biology of complex diseases, improving their risk prediction, and ultimately discovering novel therapeutic targets[1]. However, the majority of the published GWAS have only reported primary findings, which generally explain a small fraction of the estimated heritability. To examine the missing heritability, most strategies involve generating new genetic and clinical data. Very rarely are new studies based on the revision and reanalysis of existing genetic data by applying more powerful analytic techniques and resources after the primary GWAS findings are published. These cost-effective reanalysis strategies are now possible, given emerging (1) data-sharing initiatives with large amounts of primary genetic data for multiple human genetic diseases, as well as (2) new and improved GWAS methodologies and resources. Notably, genotype imputation with novel sequence-based reference panels can now substantially increase the genetic resolution of GWASs from previously genotyped data sets[2], reaching good-quality imputation of low frequency (minor allele frequency [MAF]: $0.01 \le \mathrm{MAF} < 0.05$) and rare variants ($\mathrm{MAF} < 0.01$), increasing the power to identify novel associations, and fine map the known ones. Moreover, the availability of publicly available primary genetic data allows the homogeneous integration of multiple data sets from different origins providing more accurate meta-analysis results, particularly at the low ranges of allele frequency. Finally, the vast majority of reported GWAS analyses omits the X chromosome, despite representing 5% of the genome and coding for more than 1,500 genes[3]. The reanalysis of publicly available data also enables interrogation of this chromosome.

We hypothesized that a unified reanalysis of multiple publicly available data sets, applying homogeneous standardized quality control (QC), genotype imputation, and association methods, as well as novel and denser sequence-based reference panels for imputation would provide new insights into the genetics and the pathophysiology of complex diseases. To test this hypothesis, we focused this study on type 2 diabetes (T2D), one of the most prevalent complex diseases for which many GWAS have been performed during the past decade[4]. These studies have allowed the identification of more than 100 independent loci, most of them driven by common variants, with a few exceptions[5]. Despite these efforts, there is still a large fraction of genetic heritability hidden in the data, and the role of low-frequency variants, although recently proposed to be minor[6], has still not been fully explored. The availability of large T2D genetic data sets in combination with larger and more comprehensive genetic variation reference panels[2], provides the opportunity to impute a significantly increased fraction of low-frequency and rare variants, and to study their contribution to the risk of developing this disease. This strategy also allows us to fine map known associated loci, increasing the chances of finding causal variants and understanding their functional impact. We therefore gathered publicly available T2D GWAS cohorts with European ancestry, comprising a total of 13,857 T2D cases and 62,126 controls, to which we first applied harmonization and quality control protocols covering the whole genome (including the X chromosome). We then performed imputation using 1000 Genomes Project (1000G)[7] and UK10K[2] reference panels, followed by association testing. By using this strategy, we identified novel associated regions driven by common, low-frequency and rare variants, fine mapped and functionally annotated the existing and novel ones, allowing us to describe a regulatory mechanism disrupted by a novel rare and large-effect variant identified at the X chromosome.

Results

Overall analysis strategy. As shown in Fig. 1, we first gathered all T2D case-control GWAS individual-level data that were available through the EGA and dbGaP databases (i.e., Gene Environment-Association Studies [GENEVA], Wellcome Trust Case Control Consortium [WTCCC], Finland–United States Investigation of NIDDM Genetics [FUSION], Resource for Genetic Epidemiology Research on Aging [GERA], and Northwestern NuGENE project [NuGENE]). We harmonized these cohorts, applied standardized quality control procedures, and filtered out low-quality variants and samples (Methods and Supplementary Notes). After this process, a total of 70,127 subjects (70KforT2D, 12,931 cases, and 57,196 controls, Supplementary Data 1) were retained for downstream analysis. Each of these cohorts was then imputed to the 1000G and UK10K reference panels using an integrative method, which selected the results from the reference panel that provided the highest accuracy for each variant, according to IMPUTE2 info score (Methods). Finally, the results from each of these cohorts were meta-analyzed (Fig. 1), obtaining a total of 15,115,281 variants with good imputation quality (IMPUTE2 info score ≥ 0.7, $\mathrm{MAF} \ge 0.001$, and I^2 heterogeneity score < 0.75), across 12,931 T2D cases and 57,196 controls. Of these, 6,845,408 variants were common ($\mathrm{MAF} \ge 0.05$), 3,100,848 were low-frequency ($0.01 \le \mathrm{MAF} < 0.05$), and 5,169,025 were rare ($0.001 \le \mathrm{MAF} < 0.01$). Merging the imputation results derived from the two reference panels substantially improved the number of good-quality imputed variants, particularly within the low-frequency and rare spectrum, compared to the imputation results obtained with each of the panels separately. For example, a set of 5,169,025 rare variants with good quality was obtained after integrating 1000G and UK10K results, while only 2,878,263 rare variants were imputed with 1000G and 4,066,210 with UK10K (Supplementary Fig. 1A). This strategy also allowed us to impute 1,357,753 indels with good quality (Supplementary Fig. 1B).

To take full advantage of publicly available genetic data, we used three main meta-analytic approaches to adapt to the three most common strategies for genetic data sharing: individual-level genotypes, summary statistics, and single-case queries through the Type 2 Diabetes Knowledge Portal (T2D Portal) (http://www.type2diabetesgenetics.org/). We first meta-analyzed all summary statistics results from the DIAGRAM trans-ancestry meta-analysis[8] (26,488 cases and 83,964 controls), selecting 1,918,233 common variants (MAF ≥ 0.05), mostly imputed from HapMap, with the corresponding fraction of non-overlapping samples in our 70KforT2D set, i.e. the GERA and the NuGENE cohorts, comprising a total of 7,522 cases and 50,446 controls (Fig. 1, Supplementary Data 1). Second, the remaining variants (13,197,048), which included mainly non-HapMap variants ($\mathrm{MAF} < 0.05$) or variants not tested above, were meta-analyzed using all five cohorts from the 70KforT2D resource (Supplementary Data 1). Finally, low-frequency variants located in coding regions and with $p \le 1 \times 10^{-4}$ were meta-analyzed using the non-overlapping fraction of samples with the data from the T2D Portal through the interrogation of exome array and whole-exome sequence data from ~80,000 and ~17,000 individuals, respectively[6].

Pathway and functional enrichment analysis. To explore whether our results recapitulate the pathophysiology of T2D, we performed gene-set enrichment analysis with all the variants with $p \le 1 \times 10^{-5}$ using DEPICT[9] (Methods). This analysis showed enrichment of genes expressed in pancreas (ranked first in tissue enrichment analysis, $p = 7.8 \times 10^{-4}$, FDR < 0.05, Supplementary Data 2) and cellular response to insulin stimulus (ranked second in gene-set enrichment analysis, $p = 3.9 \times 10^{-8}$, FDR = 0.05,

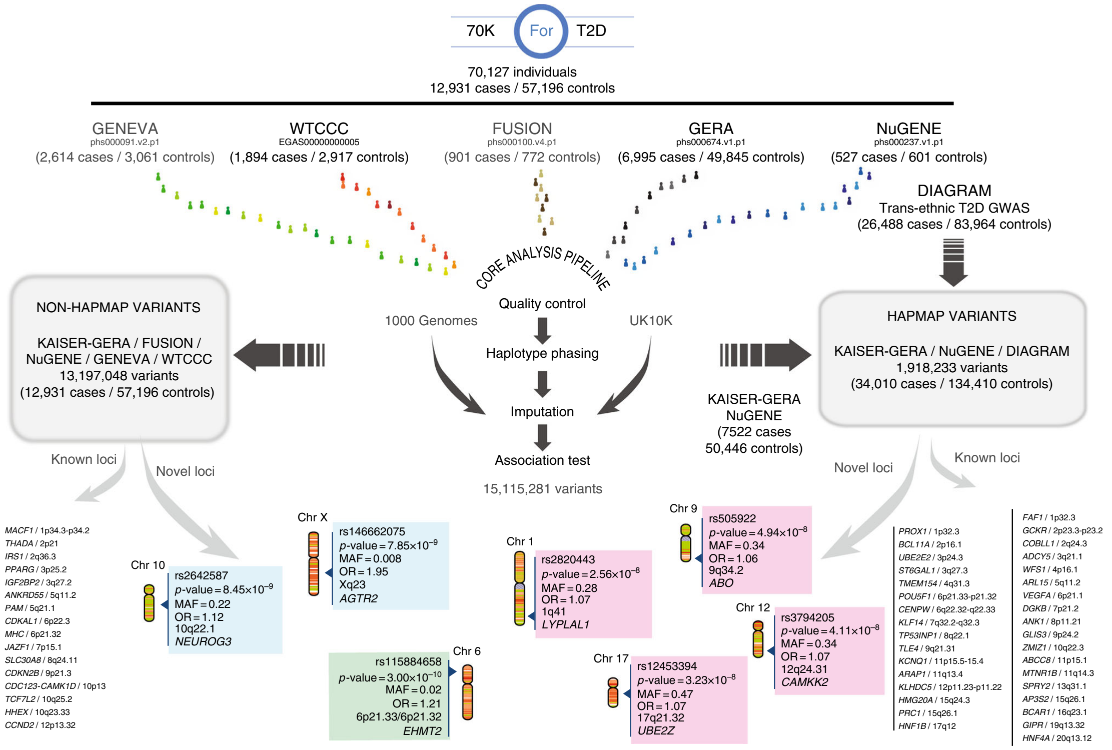


Fig. 1 Discovery and replication strategy. Publicly available GWAS datasets representing a total of 12,931 cases and 57,196 controls (70KforT2D) were first quality controlled, phased, and imputed, using 1000G and UK10K separately. For those variants that were present in the DIAGRAM trans-ethnic meta-analysis, we used the summary statistics to meta-analyze our results with the cohorts that had no overlap with any of the cohorts included in the DIAGRAM trans-ethnic meta-analysis. With this first meta-analysis, we discovered four novel loci (within magenta panels). For the rest of the variants, we meta-analyzed all the 70KforT2D data sets, which resulted in two novel loci (in blue panels). All the variants that were coding and showed a p-value of $\leq 1\times10^{-4}$ were tested for replication by interrogating the summary statistics in the Type 2 Diabetes Knowledge Portal (T2D Portal) (http://www.type2diabetesgenetics.org/). This uncovered a novel low-frequency variant in the *EHMT2* gene (highlighted with a green panel)

Supplementary Data 3, Supplementary Fig. 2, Supplementary Fig. 3), in concordance with the current knowledge of the molecular basis of T2D.

In addition, variant set enrichment analysis of the T2D-associated credible sets across regulatory elements defined in isolated human pancreatic islets showed a significant enrichment for active regulatory enhancers (Supplementary Fig. 4), suggesting that causal SNPs within associated regions have a regulatory function, as previously reported[10].

Fine-mapping and functional characterization of T2D loci. The three association strategies allowed us to identify 57 genome-wide significant associated loci ($p \leq 5\times10^{-8}$), of which seven were not previously reported as associated with T2D (Table 1). The remaining 50 loci have been previously reported and included, for example, two low-frequency variants recently discovered in Europeans, one located within one of the *CCND2* introns (rs76895963), and a missense variant within the *PAM*[5] gene. Furthermore, we confirmed that the magnitude and direction of the effect of all the associated variants ($p \leq 0.001$) were highly consistent with those reported previously ($\rho = 0.92$, $p = 1\times10^{-248}$, Supplementary Fig. 5). In addition, the direction of effect was consistent with all 139 previously reported variants, except three that were discovered in east and south Asian populations (Supplementary Data 4).

The high coverage of genetic variation ascertained in this study allowed us to fine-map known and novel loci, providing more candidate causal variants for downstream functional interpretations. We constructed 99% credible variant sets[11] for each of these loci, i.e. the subset of variants that have, in aggregate, 99% probability of containing the true causal variant for all 57 loci (Supplementary Data 5). As an important improvement over previous T2D genetic studies, we identified small structural variants within the credible sets, consisting mostly of insertions and deletions between 1 and 1,975 nucleotides. In fact, out of the 8,348 variants included within the credible sets for these loci, 927 (11.1%) were indels, of which 105 were genome-wide significant (Supplementary Data 6). Interestingly, by integrating imputed results from 1000G and UK10K reference panels, we gained up to 41% of indels, which were only identified by either one of the two reference panels, confirming the advantage of integrating the results from both reference panels. Interestingly, 15 of the 71 previously reported loci that we replicated ($p \leq 5.3\times10^{-4}$ after correcting for multiple testing) have an indel as the top variant, highlighting the potential role of this type of variation in the susceptibility for T2D. For example, within the *IGF2BP2* intron, a well-established and functionally validated locus for T2D[12], we found that 12 of the 57 variants within its 99% credible set correspond to indels with genome-wide significance ($5.6\times10^{-16} < p < 2.4\times10^{-15}$), which collectively represented 18.4% posterior probability of being causal.

Table 1 Novel T2D-associated loci

			OR (95% CI) *P*-value			
Novel Locus	**Chr**	**rsID--Risk Allele**	**Stage1 Discovery Meta-analysis**	**Stage2 Replication Meta-analysis**	**Stage1 + Stage2 Combined Meta-analysis**	**MAF**
LYPLAL1/ZC3H11B (1q41)	1	rs2820443-T	1.08 (1.04–1.13) 2.94×10^{-4} [a]	1.06 (1.03–1.09) 2.10×10^{-5} [b]	1.07 (1.04–1.09) 2.56×10^{-8} [c]	0.28
EHMT2 (6p21.33–p21.32)	6	rs115884658-A	1.34 (1.18–1.53) 1.00×10^{-5} [a]	1.17 (1.09–1.26) 2.90×10^{-6} [c, d]	1.21 (1.14–1.29) 3.00×10^{-10} [c]	0.02
ABO (9q34.2)	9	rs505922-C	1.07 (1.03–1.11) 6.93×10^{-4} [a]	1.06 (1.03–1.09) 1.90×10^{-5} [b]	1.06 (1.04–1.09) 4.94×10^{-8} [c]	0.34
NEUROG3 (10q22.1)	10	rs2642587-G	1.12 (1.08–1.16) 8.45×10^{-9} [e]	-	-	0.22
CAMKK2 (12q24.31)	12	rs3794205-G	1.09 (1.05–1.14) 4.18×10^{-5} [a]	1.06 (1.03–1.09) 1.60×10^{-4} [b]	1.07 (1.04–1.10) 4.11×10^{-8} [c]	0.32
CALCOCO2/ATP5G1/ UBE2Z/SNF8/GIP (17q21.32)	17	rs12453394-A	1.08 (1.04–1.12) 7.86×10^{-5} [a]	1.07 (1.03–1.11) 9.60×10^{-5} [b]	1.07 (1.05–1.10) 3.23×10^{-8} [c]	0.47
AGTR2 (Xq23)	X	rs146662075-T	3.09 (2.06–4.60) 3.24×10^{-8} [f]	1.57 (1.19–2.07) 1.42×10^{-3} [g]	1.95 (1.56–2.45) 7.85×10^{-9}	0.008

Chr chromosome, *OR* odds ratio, *MAF* minor allele frequency

[a] Imputed based public GWAS discovery meta-analysis (NuGENE + GERA cohort, 7,522 cases and 50,446 controls)

[b] Transancestry DIAGRAM Consortium (26,488 cases and 83,964 controls)[c] Meta *P*-value estimated using a weighted *Z*-score method due to unavailable SE information from Stage 2 replication cohorts[d] T2D Diabetes Genetic Portal (Exome-Chip + Exome Sequencing, 35,789 cases and 56,738 controls)[e] Full imputed based public GWAS meta-analysis (NuGENE + GERA cohort + GENEVA + FUSION + WTCCC, 12,931 cases and 57,196 controls)

[f] 70KforT2D Men Cohort (GERA cohort + GENEVA + FUSION, 5,277 cases and 15,702 controls older than 55 years)

[g] Replication Men Cohort SIGMA UK10K imputation + InterAct + Danish Cohort (case control and follow-up) + Partners Biobank + UK Biobank (18,370 cases and 88,283 controls older than 55 years and OGTT > 7.8 mmol l^{-1}, when available)

To prioritize causal variants within all the identified associated loci, we annotated their corresponding credible sets using the Variant Effector Predictor (VEP) for coding variants[13] (Supplementary Data 7), and the Combined Annotation-Dependent Depletion (CADD)[14] and LINSIGHT[15] tools for non-coding variation (Supplementary Data 8 and 9). In addition, we tested the effect of all variants on expression across multiple tissues by interrogating GTEx[16] and RNA-sequencing gene expression data from pancreatic islets[17].

Novel T2D-associated loci driven by common variants. Beyond the detailed characterization of the known T2D-associated regions, we also identified seven novel loci, among which, five were driven by common variants with modest effect sizes (1.06 < OR < 1.12; Table 1, Fig. 2, Supplementary Fig. 6 and 7).

Within the first novel T2D-associated locus in chromosome 1q41 (*LYPLAL1-ZC3H11B*, rs2820443, OR = 1.07 [1.04–1.09], $p = 2.6 \times 10^{-8}$), several variants have been previously associated with waist-to-hip ratio, height, visceral adipose fat in females, adiponectin levels, fasting insulin, and non-alcoholic fatty liver disease[18–23]. Among the genes in this locus, *LYPLAL1*, which encodes for lysophospholypase-like 1, appears to be the most likely effector gene, as it has been found to be downregulated in mouse models of diet-induced obesity and upregulated during adipogenesis[24].

Second, a novel locus at chromosome 9q34.2 region (*ABO*, rs505922, OR = 1.06 [1.04–1.09], $p = 4.9 \times 10^{-8}$) includes several variants that have been previously associated with other metabolic traits. For example, the variant rs651007, in linkage disequilibrium (LD) with rs505922 ($r^2 = 0.507$), has been shown to be associated with fasting glucose[25], and rs514659 (r^2 with top = 1) is associated with an increased risk for cardiometabolic disorders[26]. One of the variants within the credible set was the single base-pair frame-shift deletion defining the blood group O[27]. In concordance with previous results that linked O blood type with a lower risk of developing T2D[28], the frame-shift deletion determining the blood group type O was associated with a protective effect for T2D in our study (rs8176719, $p = 3.4 \times 10^{-4}$, OR = 0.95 [0.91–0.98]). In addition, several variants within this credible set are associated with the expression of the *ABO* gene in multiple tissues including skeletal muscle, adipose tissue, and pancreatic islets (Supplementary Data 9, Supplementary Data 10).

Third, a novel locus at chromosome 10q22.1 locus (*NEUROG3/COL13A1/RPL5P26*, rs2642587, OR = 1.12 [1.08–1.16], $p = 8.4 \times 10^{-9}$) includes *NEUROG3* (Neurogenin3), which is an essential regulator of pancreatic endocrine cell differentiation[29]. Mutations in this gene have been reported to cause permanent neonatal diabetes, but a role of this gene in T2D has not been yet reported[30].

The lead common variant of the fourth novel locus at chromosome 12q24.31 (rs3794205, OR = 1.07 [1.04–1.10], $p = 4.1 \times 10^{-8}$) lies within an intron of the *CAMKK2* gene, previously implicated in cytokine-induced beta-cell death[31]. However, other variants within the corresponding credible set could also be causal, such as a missense variant within the *P2RX7*, a gene previously associated with glucose homeostasis in humans and mice[32], or another variant (rs11065504, r^2 with lead variant = 0.81) found to be associated with the regulation of the *P2RX4* gene in tibial artery and in whole blood, according to GTEx (Supplementary Data 9).

The fifth novel locus driven by common variants is located within 17q21.32 (rs12453394, OR = 1.07 [1.05–1.10], $p = 3.23 \times 10^{-8}$). It includes three missense variants located within the *CALCOCO2*, *SNF8*, and *GIP* genes. *GIP* encodes for glucose-dependent insulinotropic peptide, a hormonal mediator of enteral regulation of insulin secretion[33]. Variants in the GIP receptor (GIPR) have been previously associated with insulin response to oral glucose challenge and beta-cell function[34], proposing *GIP* as a plausible candidate effector gene of this locus[35].

A new T2D signal driven by a low-frequency variant. Furthermore, we selected all low-frequency ($0.01 \le \text{MAF} < 0.05$) variants with $p \le 1 \times 10^{-4}$ in the 70KforT2D meta-analysis that

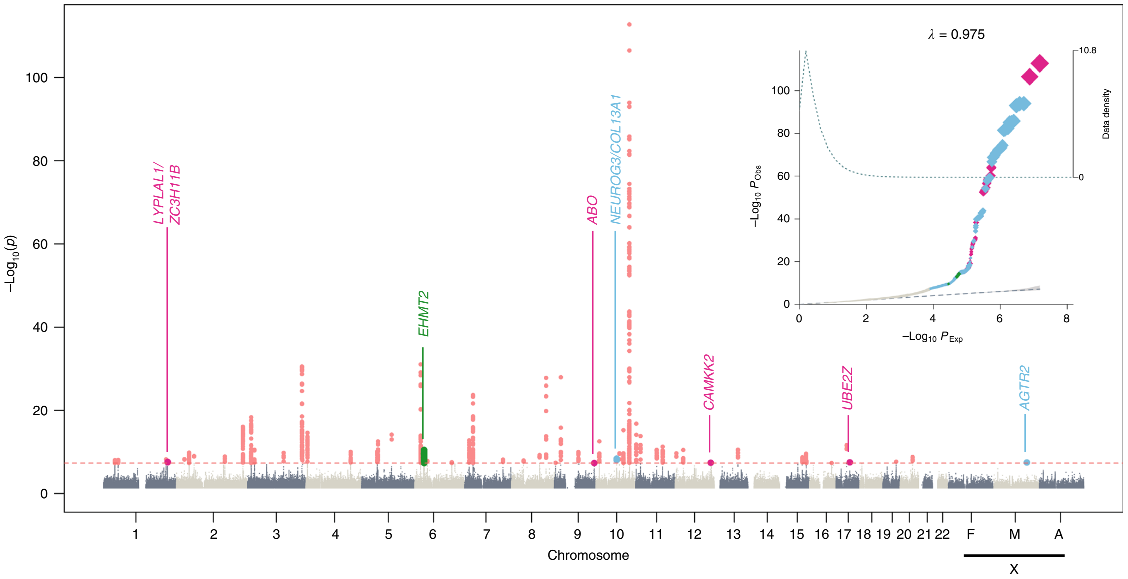


Fig. 2 Manhattan and quantile–quantile plot (QQ-plot) of the discovery and replication genome-wide meta-analysis. The upper corner represents the QQ-plot. Expected −log$_{10}$ *p*-values under the null hypothesis are represented in the *x* axis, while observed −log$_{10}$ *p*-values are represented in the *y* axis. Observed *p*-values were obtained according to the suitable replication dataset used (as shown in Fig. 1) and were depicted using different colors. HapMap variants were meta-analyzed using the trans-ethnic summary statistics from the DIAGRAM study and our meta-analysis based on the Genetic Epidemiology Research on Aging (GERA) cohort and the northwestern NuGENE project, and that resulted in novel associations depicted in magenta. The rest of non-HapMap variants meta-analyzed using the full 70KforT2D cohort are represented in gray, and the fraction of novel GWAS-significant variants is highlighted in light blue. Coding low-frequency variants meta-analyzed using the 70KforT2D and the T2D Portal data that resulted in novel GWAS-significant associations are depicted in green. The shaded area of the QQ-plot indicates the 95% confidence interval under the null and a density function of the distribution of the *p*-values was plotted using a dashed line. The *λ* is a measure of the genomic inflation and corresponds to the observed median *χ*2 test statistic divided by the median expected χ^2 test statistic under the null hypothesis. The Manhattan plot, representing the −log$_{10}$ *p*-values, was colored as explained in the QQ-plot. All known GWAS-significant associated variants within known T2D genes are also depicted in red. X chromosome results for females (F), males (M), and all individuals (A) are also included

were annotated as altering protein sequences, according to VEP. This resulted in 15 coding variants that were meta-analyzed with exome array and whole-exome sequencing data from a total of ~97,000 individuals[6] after excluding the overlapping cohorts between the different data sets. This analysis highlighted a novel genome-wide association driven by a low-frequency missense variant (Ser58Phe) within the *EHMT2* gene at chromosome 6p21.33 (rs115884658, OR = 1.21 [1.14–1.29], $p = 3.00 \times 10^{-10}$; Fig. 2, Supplementary Figures 6 and 7). *EHMT2* is involved in the mediation of FOXO1 translocation induced by insulin[36]. Since this variant is less than 1 Mb away from *HLA-DQA1*, a locus reported to be associated with T2D[37,] we performed a series of reciprocal conditional analyses and excluded the possibility that our analysis was capturing previously reported T2D[8, 37] or T1D[38–40] signals (Supplementary Data 11). Beyond this missense *EHMT2* variant, other low-frequency variants within the corresponding credible set may also be causal. For example, rs115333512 (r^2 with lead variant = 0.28) is associated with the expression of *CLIC1* in several tissues according to GTEx (multitissue meta-analysis $p = 8.9 \times 10^{-16}$, Supplementary Data 9). In addition, this same variant is associated with the expression of the first and second exon of the *CLIC1* mRNA in pancreatic islet donors (*p*(exon 1) = 1.4×10^{-19}, *p*(exon 2) = 1.9×10^{-13}, Supplementary Data 10). Interestingly, *CLIC1* has been reported as a direct target of metformin by mediating the antiproliferative effect of this drug in human glioblastoma[41]. All these findings support *CLIC1,* as an additional possible effector transcript, likely driven by rs115333512.

A novel rare X chromosome variant associated with T2D. Similar to other complex diseases, the majority of published large-scale T2D GWAS studies have omitted the analysis of the X chromosome, with the notable exception of the identification of a T2D-associated region near the *DUSP9* gene in 2010[42]. To fill this gap, we tested the X chromosome genetic variation for association with T2D. To account for heterogeneity of the effects and for the differences in imputation performance between males and females, the association was stratified by sex and tested separately, and then meta-analyzed. This analysis was able to replicate the *DUSP9* locus, not only through the known rs5945326 variant (OR = 1.15, $p = 0.049$), but also through a three-nucleotide deletion located within a region with several promoter marks in liver (rs61503151 [GCCA/G], OR = 1.25, $p = 3.5 \times 10^{-4}$), and in high LD with the first reported variant ($r^2 = 0.62$). Conditional analyses showed that the originally reported variant was no longer significant (OR = 1.01, $p = 0.94$) when conditioning on the newly identified variant, rs61503151. On the other hand, when conditioning on the previously reported variant, rs5945326, the effect of the newly identified indel remained significant and with a larger effect size (OR = 1.33, $p = 0.003$), placing this deletion, as a more likely candidate causal variant for this locus (Supplementary Data 14).

In addition, we identified a novel genome-wide significant signal in males at the Xq23 locus driven by a rare variant (rs146662075, MAF = 0.008, OR = 2.94 [2.00–4.31], $p = 3.5 \times 10^{-8}$; Fig. 3a). Two other variants in LD with the top variant, rs139246371 (chrX:115329804, OR = 1.65, $p = 3.5 \times 10^{-5}$, $r^2 =$

0.37 with the top variant) and rs6603744 (chrX:115823966, OR = 1.28, $p = 1.7 \times 10^{-4}$, $r^2 = 0.1$ with the top variant), comprised the 99% credible set and supported the association. We tested in detail the accuracy of the imputation for the rs146662075 variant by comparing the imputed results from the same individuals genotyped by two different platforms (Methods) and found that the imputation was highly accurate in males only when using UK10K, but not in females, nor when using 1000G ($R^2_{[\mathrm{UK10K,males}]} = 0.94$, $R^2_{[\mathrm{UK10K,females}]} = 0.66$, $R^2_{[\mathrm{1000G,males}]} = 0.62$, and $R^2_{[\mathrm{1000G,females}]} = 0.43$; Supplementary Fig. 8). Whether this association is specific to men, or whether it also affects female carriers, remains to be clarified with datasets that allow accurate imputation on females, or with direct genotyping or sequencing.

To further validate and replicate this association, we next analyzed four independent data sets (SIGMA[6], INTERACT[43], Partners Biobank[44], and UK Biobank[45]), by performing imputation with the UK10K reference panel. In addition, a fifth cohort was genotyped de novo for the rs146662075 variant in several Danish sample sets. The initial meta-analysis, including the five replication data sets did not reach genome-wide significance (OR = 1.57, $p = 1.2 \times 10^{-5}$; Supplementary Fig. 9A), and revealed a strong degree of heterogeneity (heterogeneity $p_{\mathrm{het}} = 0.004$), which appeared to be driven by the replication cohorts.

As a complementary replication analysis, within one of the case-control studies, there was a nested prospective cohort study, the Inter99, which consisted of 1,652 nondiabetic male subjects genotyped for rs146662075, of which 158 developed T2D after 11 years of follow-up. Analysis of incident diabetes in this cohort confirmed the association with the same allele, as previously seen in the case-control studies, with carriers of the rare T allele having increased risk of developing incident diabetes, compared to the C carriers (Cox-proportional hazards ratio (HR) = 3.17 [1.3–7.7], $p = 0.011$, Fig. 3b). Nearly 30% of carriers of the T risk allele developed incident T2D during 11 years of follow-up, compared to only 10% of noncarriers.

To understand the strong degree of heterogeneity observed after adding the replication datasets, we compared the clinical and demographic characteristics of the discovery and replication cohorts, and found that the majority of the replication datasets contained control subjects that were significantly younger than 55 years, the average age at the onset of T2D reported in this study and in Caucasian populations[46]. This was particularly clear for the Danish cohort (age controls [95%CI] = 46.9 [46.6–47.2] vs. age cases [95%CI] = 60.7 [60.4–61.0]) and for INTERACT (age controls [95%CI] = 51.7 [51.4–52.1] vs. age cases [95%CI] = 54.8 [54.6–55.1]; Supplementary Fig. 10). Given the supporting results with the Inter99 prospective cohort, we performed an additional analysis using a stricter definition of controls, to minimize the presence of prediabetics or individuals that may further develop diabetes after reaching the average age at the onset. For this, we applied two additional exclusion criteria: (i) subjects younger than 55 years and (ii), when possible, excluding individuals with measured 2-h plasma glucose values during oral glucose tolerance test (OGTT) above 7.8 mmol l^{-1}, a threshold employed to identify impaired glucose tolerance (prediabetes)[47], or controls with family history of T2D, both being strong risk factors for developing T2D. While the application of the first filter alone did not yield genome-wide significant results (Supplementary Fig. 9B), upon excluding individuals with prediabetes or a family history of T2D, the replication results were significant and consistent with the initial discovery results (OR = 1.57 [1.19–2.07], $p = 0.0014$). The combined analysis of the discovery and replication cohorts resulted in genome-wide significance, confirming the association of rs146662075 with T2D (OR = 1.95 [1.56–2.45], $p = 7.8 \times 10^{-9}$, Fig. 3c).

Allele-specific enhancer activity of the rs146662075 variant. We next explored the possible molecular mechanism behind this association, by using different genomic resources and experimental approaches. The credible set of this region contained three variants, with the leading SNP alone (rs146662075), showing 78% posterior probability of being causal (Supplementary Fig. 7, Supplementary Data 5), as well as the highest CADD (scaled *C*-score = 15.68; Supplementary Data 8), and LINSIGHT score (Supplementary Data 9). rs146662075 lies within a chromosomal region enriched in regulatory (DNase I) and active enhancer (H3K27ac) marks, between the *AGTR2* (at 103 kb) and the *SLC6A14* (at 150 kb) genes. The closest gene *AGTR2*, which encodes for the angiotensin II receptor type 2, has been previously associated with insulin secretion and resistance[48–50]. From the analysis of available epigenomic data sets[51], we found no evidences of H3K27ac or other enhancer regulatory marks in human pancreatic islets; whereas a significant association was observed between the presence of H3K27ac enhancer marks and the expression of *AGTR2* across multiple tissues (Fisher test $p = 4.45 \times 10^{-3}$), showing the highest signal of both H3K27ac and *AGTR2* RNA-seq expression, but not with other genes from the same topologically associated domain (TAD), in fetal muscle (Fig. 4a; Supplementary Figure 11).

We next studied whether the region encompassing the rs146662075 variant could act as a transcriptional enhancer and whether its activity was allele-specific. For this, we linked the DNA region with either the T (risk) or the C (non-risk) allele, to a minimal promoter and performed luciferase assays in a mouse myoblast cell line. The luciferase analysis showed an average 4.4-fold increased activity for the disease-associated T allele, compared to the expression measured with the common C allele, suggesting an activating function of the T allele, or a repressive function of the C allele (Fig. 4b). Consistent with these findings, electrophoretic mobility shift assays using nuclear protein extracts from mouse myoblast cell lines, differentiated myotubes, and human fetal muscle cell line, revealed sequence-specific binding activity of the C allele, but not the rare T allele (Fig. 4c). Overall, these data indicate that the risk T allele prevents the binding of a nuclear protein that is associated with decreased activity of an *AGTR2*-linked enhancer.

Discussion

Through harmonizing and reanalyzing publicly available T2D GWAS data, and performing genotype imputation with two whole-genome sequence-based reference panels, we are able to perform deeper exploration of the genetic architecture of T2D. This strategy allowed us to impute and test for association with T2D more than 15 million of high-quality imputed variants, including low-frequency, rare, and small insertions and deletions, across chromosomes 1–22 and X.

The reanalysis of these data confirmed a large fraction of already-known T2D loci, and identified novel potential causal variants by fine mapping and functionally annotating each locus.

This reanalysis also allowed us to identify seven novel associations, five driven by common variants in or near *LYPLAL1*, *NEUROG3*, *CAMKK2*, *ABO*, and *GIP;* a low-frequency variant in *EHMT2*, and a rare variant in the X chromosome. This rare variant identified in Xq23 chromosome was located near the *AGTR2* gene, and showed nearly twofold increased risk for T2D in males, which represents, to our knowledge, the largest effect size identified so far in Europeans, and a magnitude similar to other variants with large effects identified in other populations[52, 53].

Our study complemented other efforts that also aim at unraveling the genetics behind T2D through the generation of new

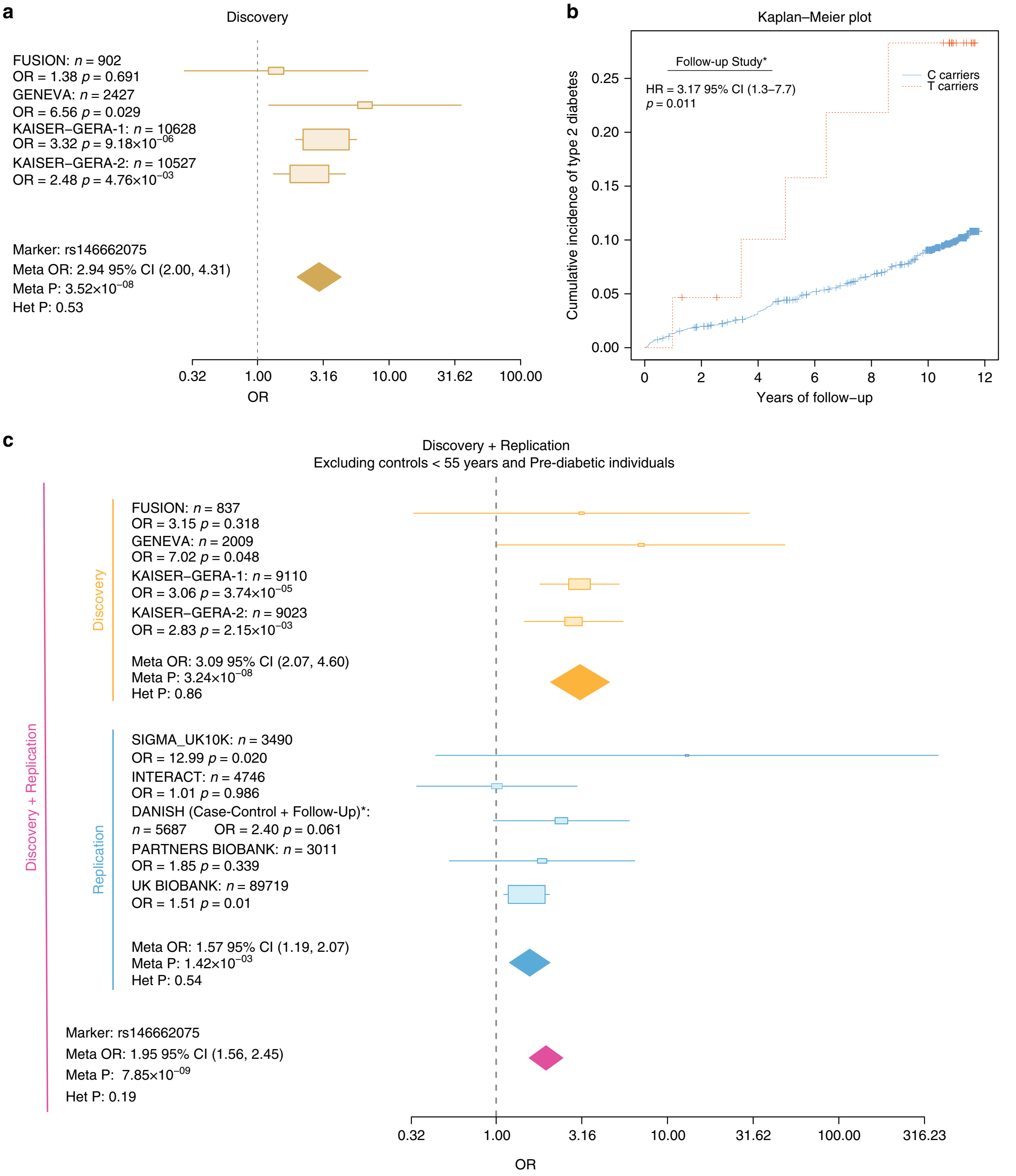
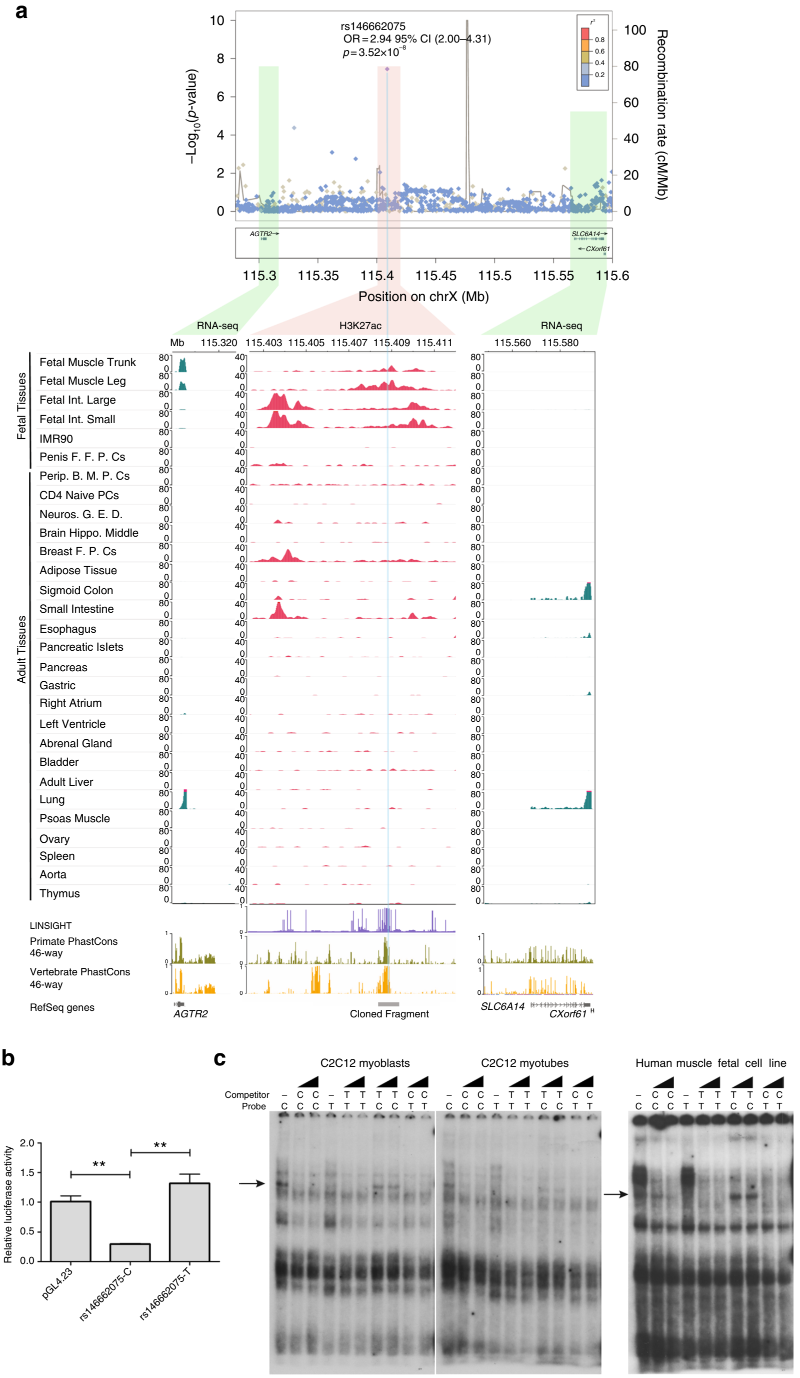


Fig. 3 Discovery and replication of rs14666075 association signal. **a** Forest plot of the discovery of rs146662075 variant. Cohort-specific odds ratios are denoted by boxes proportional to the size of the cohort and 95% CI error bars. The combined OR estimated for all the data sets is represented by a diamond, where the diamond width corresponds to 95% CI bounds. The *p*-value for the meta-analysis (Meta P) and for the heterogeneity (Het P) of odds ratio is shown. **b** Kaplan–Meier plot showing the cumulative incidence of T2D for a 11 years follow-up. The red line represents the T carriers and in light blue, C carriers are represented ($n = 1,652$, cases = 158). **c** Forest plot after excluding controls younger than 55 years, OGTT >7.8 mmol l^{-1}, and controls with family history of T2D in both the discovery and replication cohorts when available

genetic data[6, 54]. For example, we provided for the first time a comprehensive coverage of structural variants, which point to previously unobserved candidate causal variants in known and novel loci, as well as a comprehensive coverage of the X chromosome through sequence-based imputation.

This study also highlights the importance of a strict classification of both cases and controls, in order to identify rare variants associated with disease. Our initial discovery of the Xq23 locus was only replicated when the control group was restricted to T2D-free individuals who were older than 55 years (average age



at the onset of T2D), had normal glucose tolerance, and no family history of T2D. This is in line with previous results obtained for a T2D population-specific variant found in Inuit within the *TBC1D4* gene, which was only significant when using OGTT as criteria for classifying cases and controls, but not when using HbA1c[52]. Our observation that 30% of the rs146662075 risk allele carriers developed T2D over 11 years of follow-up, compared to 10% of noncarriers, further supports the association of this variant and suggests that an early identification of these subjects through genotyping may be useful to tailor pharmacological or lifestyle intervention to prevent or delay the onset of T2D.

Using binding and gene-reporter analyses, we demonstrated a functional role of this variant and proposed a possible mechanism behind the pathophysiology of T2D in T risk allele carriers, in which this rare variant could favor a gain of function of *AGTR2*, previously associated with insulin resistance[48]. *AGTR2* appears, therefore, as a potential therapeutic target for this disease, which would be in line with previous studies showing that the blockade of the renin–angiotensin system in mice[55] and in humans[56] prevents the onset of T2D, and restores normoglycemia[57, 58].

Overall, beyond our significant contribution toward expanding the number of genetic associations with T2D, our study also highlights the potential of the reanalysis of public data, as a complement to large studies that use newly generated data. This study informs the open debate in favor of data sharing and democratization initiatives[4, 59], for investigating the genetics and pathophysiology of complex diseases, which may lead to new preventive and therapeutic applications.

Methods

Quality filtering for imputed variants. In order to assess genotype imputation quality and to determine an accurate post-imputation quality filter, we made use of the Wellcome Trust Case Control Consortium (WTCCC)[40] data available through the European Genotype Archive (EGA, https://www.ebi.ac.uk/ega/studies/EGAS00000000028). The genotyping data and the subjects included in the following tests were filtered according to the guidelines provided by the WTCCC, whose criteria of exclusion are in line with standard quality filters for GWAS[60]. We used the 1958 British Birth cohort (~3,000 samples, 58C) that was genotyped by Affymetrix v6.0 and Illumina 1.2M chips. After applying the quality-filtering criteria, 2,706 and 2,699 subjects from the Affymetrix and Illumina data, respectively, were available for the 58C samples, leaving an intersection of 2,509 individuals genotyped by both platforms. After variant quality filtering and excluding all the variants with minor allele frequency (MAF) below 0.01, 717,556, and 892,516 variants remained for 58C Affymetrix and Illumina platforms, respectively.

We used a two-step genotype imputation approach based on prephasing the study genotypes into full haplotypes with SHAPEIT2[61] to ameliorate the computational burden required for genotype imputation through IMPUTE2[62]. We used the GTOOL software (http://www.well.ox.ac.uk/~cfreeman/software/gwas/gtool.html, version 0.7.5) to homogenize strand annotation by merging the imputed results obtained from each set of genotyped data. To ensure that there were no strand orientation issues, we excluded all C/G and A/T SNPs. To perform genotype imputation, we used two sequence-based reference panels: the 1000G Phase1 (June 2014) release[7] and the UK10K[2].

We evaluated genotype imputation for each reference panel considering 2,509 58C individuals that were genotyped by both independent genotyping platforms. Four scenarios were considered: (a) fraction of variants originally genotyped (GT) by both Illumina (IL) and Affymetrix (Affy) platforms (both GT), (b) variants genotyped by Affy, but not present in IL array (Affy GT), (c) variants genotyped by IL, but not present in the Affy array (IL GT), and (d) variants not typed in IL nor in the Affy arrays, and therefore, imputed from IL and Affy data sets (d). This last scenario comprised the largest fraction of variants.

As the individuals typed (and imputed) using Affy and IL SNPs as backbones were the same, we expected no statistical differences when comparing the allele and genotype frequencies with any of the variants. The quality of the imputed variants was evaluated using the allelic dosage R^2 correlation coefficient, between the genotype dosages estimated when imputing using Affy or IL as the backbone. The Affy GT and IL GT SNPs were used to evaluate the correspondence between the allelic dosage R^2 scores and the IMPUTE2 info scores for the imputed genotypes. The linear model, between the allelic dosage R^2 and the IMPUTE2-info, was used to set an info score threshold of 0.7, which corresponds to an allelic dosage R^2 of 0.5. The correlation between R^2 and info score was uniform across all reference panels and platforms.

The 70KforT2D resource. We collected genetic individual-level data for T2D case/control studies from five independent datasets, Gene Environment-Association Studies initiative [GENEVA], Wellcome Trust Case Control Consortium [WTCCC], Finland–United States Investigation of NIDDM Genetics [FUSION], Resource for Genetic Epidemiology Research on Aging [GERA], and the Northwestern NUgene project [NuGENE] publicly available in the dbGaP (http://www.ncbi.nlm.nih.gov/gap) and EGA (https://www.ebi.ac.uk/ega/home) public repositories, comprising a total of 13,201 cases and 59,656 controls (for the description of each cohort, see Supplementary Note 1 and Supplementary Data 1).

Each dataset was independently harmonized and quality controlled with a three-step protocol, including two stages of SNP removal and an intermediate stage of sample exclusion. The exclusion criteria for variants were (i) missing call rate ≥ 0.05, (ii) significant deviation from Hardy–Weinberg equilibrium (HWE) $p \leq 1 \times 10^{-6}$ for controls and $p \leq 1 \times 10^{-20}$ for the entire cohort, (iii) significant differences in the proportion of missingness between cases and controls $p \leq 1 \times 10^{-6}$, and (iv) MAF < 0.01 (for the GERA cohort, we considered a MAF of 0.001). The exclusion criteria for samples were i) gender discordance between the reported and genetically predicted sex, ii) subject relatedness (pairs with $\pi \geq 0.125$ from which we removed the individual with the highest proportion of missingness), iii) missing call rates per sample ≥ 0.02, and iv) population structure showing more than four standard deviations within the distribution of the study population according to the first four principal components.

We performed genotype imputation independently for each cohort by prephasing the genotypes to whole haplotypes with SHAPEIT2 and then, we performed genotype imputation with IMPUTE2. We tested for association with additive logistic regression using SNPTEST, seven derived principal components sex, age, and body-mass index (BMI), except for WTCCC, for which age and BMI were not available (Supplementary Data 1). To maximize power and accuracy, we combined the association results from 1000G Phase1 integrated haplotypes (June, 2014)[7] and UK10K (http://www.uk10k.org/) reference panels by choosing for each variant, the reference panel that provided the best IMPUTE2 info score. For 1000G-based genotype imputation in chromosome X (chrX), we used the "v3. macGT1" release (August, 2012). For chrX, we restricted the analysis to non-pseudoautosomal (non-PAR) regions and stratified the association analysis by sex to account for hemizygosity for males, while for females, we followed an autosomal model. Also, we did not apply HWE filtering in the X chromosome variants. Finally, for the GERA cohort due to the large computational burden that comprises the whole genotype imputation process in such a large sample size, we randomly split this cohort into two homogeneous subsets of ~30,000 individuals each, in order to minimize the memory requirements.

We included variants with IMPUTE2 info score ≥ 0.7, MAF ≥ 0.001, and for autosomal variants, HWE controls $p > 1 \times 10^{-6}$. Further details about genotype imputation and covariate information used in association testing are summarized in Supplementary Data 1.

70KforT2D and inclusion of previous summary statistics data. We meta-analyzed the different sets from the 70KforT2D data set with METAL[63], using the inverse variance-weighted fixed effect model. We included variants with I^2 heterogeneity < 75. This filter was not applied to the final X chromosome data set, after meta-analyzing the results from males and females separately (which were already filtered by $I^2 < 75$).

For the meta-analysis with the DIAGRAM trans-ethnic study[8], we excluded from the whole 70KforT2D datasets those cohorts that overlapped with the DIAGRAM data. Therefore, we meta-analyzed the GERA and NuGENE cohorts (7,522 cases and 50,446 controls) from the 70KforT2D analysis with the trans-ethnic summary statistics results. As standard errors were not provided for the

Fig. 4 Functional characterization of rs146662075 association signal. **a** Signal plot for X chromosome region surrounding rs146662075. Each point represents a variant, with its *p*-value (on a −log10 scale, *y* axis) derived from the meta-analysis results from association testing in males. The *x* axis represents the genomic position (hg19). Below, representation of H3K27ac and RNA-seq in a subset of cell types is shown. The association between RNA-seq signals and H3K27ac marks suggests that *AGTR2* is the most likely regulated gene by the enhancer that harbors rs146662075. **b** The presence of the common allelic variant rs146662075-C reduces enhancer activity in luciferase assays performed in a mouse myoblast cell line. **c** Electrophoretic mobility shift assay in C2C12 myoblast cell lines, C2C12-differentiated myotubes, and human fetal myoblasts showed allele-specific binding of a ubiquitous nuclear complex. The arrows indicate the allele-specific binding event. Competition was carried out using 50- and 100-fold excess of the corresponding unlabeled probe

DIAGRAM trans-ethnic meta-analysis, we performed a sample size based meta-analysis, which converts the direction of the effect and the p-value into a Z-score. In addition, we also performed an inverse variance-weighted fixed effect meta-analysis to estimate the final effect sizes. This approach required the estimation of the beta and standard errors from the summary statistics (p-value and odds ratio).

For the meta-analysis of coding low-frequency variants with the Type 2 Diabetes Knowledge Portal (T2D Portal)[6], we included from the 70KforT2D data set the NuGENE and GERA cohorts (7,522 cases and 50,446 controls), to avoid overlapping samples. Like in the previous scenario, standard errors were not provided for the T2D Portal data and we used a sample size based meta-analysis with METAL. However, to estimate the effect sizes, we also calculated the standard errors from the p-values and odds ratios, and we performed an inverse variance-weighted fixed effect meta-analysis.

See further details about the cohorts in Supplementary Note 1.

Pathway and enrichment analysis. Summary statistics that resulted from the 70KforT2D meta-analysis were analyzed by Data-driven Expression-Prioritized Integration for Complex Traits (DEPICT)[9] to prioritize likely causal genes, to highlight enriched pathways, and to identify the most relevant tissues/cell types; DEPICT relies on publicly available gene sets (including molecular pathways) and leverages gene expression data from 77,840 gene expression arrays, to perform gene prioritization and gene-set enrichment based on predicted gene function and the so-called reconstituted gene sets. A reconstituted gene set contains a membership probability for each gene and conversely, each gene is functionally characterized by its membership probabilities across 14,461 reconstituted gene sets. As an input to DEPICT, we used all summary statistics from autosomal variants with $p < 1 \times 10^{-5}$ in the 70KforT2D meta-analysis. We used an updated version of DEPICT, which handled 1000G Phase1-integrated haplotypes (June 2014, www.broadinstitute.org/depict). DEPICT was run using 3,412 associated SNPs ($p < 1 \times 10^{-5}$), from which we identified independent SNPs using PLINK and the following parameters: --clump-p1 5e-8, --clump-p2 1e-5, --clump-r2 0.6, and --clump-kb 250. We used LD $r^2 > 0.5$ distance to define locus limits yielding 70 autosomal loci comprising 119 genes (note that this is not the same locus definition that we used elsewhere in the text). We ran DEPICT with default settings, i.e., using 500 permutations to adjust for bias and 50 replications to estimate false discovery rate (FDR). We used normalized expression data from 77,840 Affymetrix microarrays to reconstitute gene sets[9]. The resulting 14,461 reconstituted gene sets were tested for enrichment analysis. A total of 209 tissue or cell types expression data assembled from 37,427 Affymetrix U133 Plus 2.0 Array samples were used for enrichment in tissue/cell-type expression. DEPICT identified 103 reconstituted gene sets significantly enriched (FDR < 5%) for genes found among the 70 loci associated to T2D. We did not consider reconstituted sets in which genes of the original gene set were not nominally enriched (Wilcoxon rank-sum test), as these are expected to be enriched in the reconstituted gene set by design. The lack of enrichment makes the interpretation of the reconstituted gene set challenging because the label of the reconstituted gene set will not be accurate. Hence, the following reconstituted gene sets were removed from the results (Wilcoxon rank sum and P-values in parentheses): MP:0004247 gene set ($p = 0.73$), GO:0070491 gene set ($p = 0.14$), MP:0004086 gene set ($p = 0.17$), MP:0005491 gene set ($p = 0.54$), GO:0005159 gene set ($p = 0.04$), MP:0005666 gene set ($p = 0.05$), ENSG00000128641 gene set ($p =$ 0.02), MP:0006344 gene set ($p = 0.42$), MP:0004188 gene set ($p = 0.22$), MP:0002189 gene set ($p = 0.02$), MP:0000003 gene set ($p = 0.08$), ENSG00000116604 gene set ($p = 0.13$), GO:0005158 gene set ($p = 0.07$), and MP:0001715 gene set ($p = 0.01$). After applying the filters described above, there were 89 significantly enriched reconstituted gene sets. We used the affinity propagation tool to cluster related reconstituted gene sets (network diagram script available from https://github.com/perslab/DEPICT).

We also used the VSE R package to compute the enrichment or depletion of genetic variants comprised in the 57 credible sets listed in Supplementary Data 5 across regulatory genomic annotations, as described in[64]. Each GWAS lead variant from the final meta-analysis was considered as a tag SNP and variants from the corresponding 99% credible set (Supplementary Data 5) in LD with the tag SNP ($R^2 \geq 0.4$), as a cluster or associated variant set (AVS). In order to account for the size and structure of the AVS, a null distribution was built based on random permutations of the AVS. Each permuted variant set was matched to the original AVS, cluster by cluster using HapMap data by size and structure. This Matched Random Variant Set (MRVS) was calculated using 500 permutations. Significant enrichments or depletions were considered when the Bonferroni-adjusted p-value was < 0.01. Human islet regulatory elements (C1–C5) were obtained from[10].

Definition of 99% credible sets of GWAS-significant loci. For each genome-wide significant region locus, we identified the fraction of variants that have, in aggregate, 99% probability of containing the causal T2D-associated variant. By using our 70KforT2D meta-analysis based on imputed data (NuGENE, GERA, FUSION, GENEVA, and WTCCC data sets, comprising 12,231 cases and 57,196 controls), we defined the 99% credible set of variants for each locus with a Bayesian refinement approach[11] (we considered variants with an $R^2 > 0.1$ with their respective leading SNP).

Credible sets of variants are analogous to confidence intervals as we assume that the credible set for each associated region contains, with 99% probability, the true causal SNP if this has been genotyped or imputed. The credible set construction provides, for each variant placed within a certain associated locus, a posterior probability of being the causal one[11]. We estimated the approximate Bayes' factor (ABF) for each variant as

$$\mathrm{ABF} = \sqrt{1 - r}\, e^{(rz^2/2)},$$

where

$$r = \frac{0.04}{(\mathrm{SE}^2 + 0.04)},$$

$$z = \frac{\beta}{\mathrm{SE}}.$$

The β and the SE are the estimated effect size and the corresponding standard error resulting from testing for association under a logistic regression model. The posterior probability for each variant was obtained as

$$\text{Posterior Probability}_i = \frac{\mathrm{ABF}_i}{T},$$

where ABF_i corresponds to the approximate Bayes' factor for the marker i and T represents the sum of all the ABF values from the candidate variants enclosed in the interval being evaluated. This calculation assumes that the prior of the β corresponds to a Gaussian with mean 0 and variance 0.04, which is also the same prior commonly employed by SNPTEST, the program being used for calculating single-variant associations.

Finally, we ranked variants according to the ABF (in decreasing order) and from this ordered list, we calculated the cumulative posterior probability. We included variants in the 99% credible set of each region until the SNP that pushed the cumulative posterior probability of association over 0.99.

The 99% credible sets of variants for each of the 57 GWAS-significant regions are summarized in Supplementary Data 5.

Characterization of indels. We examined whether indels from the 99% credible sets were present or absent in the 1000G Phase1 or UK10K reference panels, and also checked whether they were present or not in the 1000G Phase3 reference panel. All the information has been summarized in Supplementary Data 6. We also visually inspected the aligned BAM files of the most relevant indels from both projects to discard that they could be alignment artifacts.

Functional annotation of the 99% credible set variants. To determine the effect of 99% credible set variants on genes, transcripts, and protein sequence, we used the variant effect predictor (VEP, GRCh37.p13 assembly)[13]. The VEP application determines the effect of variants (SNPs, insertions, deletions, CNVs, or structural variants) on genes, transcripts, proteins, and regulatory regions. We used as input the coordinates of variants within 99% credible sets and the corresponding alleles, to find out the affected genes and RefSeq transcripts and the consequence on the protein sequence by using the GRCh37.p13 assembly. We also manually checked all these annotations with the Exome Aggregation Consortium data set (ExAC, http://exac.broadinstitute.org) and the most updated VEP server based on the GRCh38.p7 assembly. All these annotations are provided in Supplementary Data 7.

We used combined annotation-dependent depletion (CADD) scoring function to prioritize functional, deleterious, and disease causal variants. We obtained the scaled C-score (PHRED-like scaled C-score ranking each variant with respect to all possible substitutions of the human genome) metric for each 99% credible set variant, as it highly ranks causal variants within individual genome sequences[14] (Supplementary Data 8). We also used the LINSIGHT score to prioritize functional variants, which measures the probability of negative selection on noncoding sites by combining a generalized linear model for functional genomic data with a probabilistic model of molecular evolution[15]. For each credible set variant, we retrieved the precomputed LINSIGHT score at that particular nucleotide site, as well as the mean LINSIGHT precomputed score for a region of 20 bp centered on each credible set variant, respectively (https://github.com/CshlSiepelLab/LINSIGHT). These metrics are summarized in Supplementary Data 9.

In order to prioritize functional regulatory variants, we used the V6 release from the GTEx data that provides gene-level expression quantifications and eQTL results based on the annotation with GENCODE v19. This release included 450 genotyped donors, 8,555 RNA-seq samples across 51 tissues, and two cell lines, which led to the identification of eQTLs across 44 tissues[16]. Moreover, RNA-seq data from human pancreatic islets from 89 deceased donors cataloged as eQTLs and exon use (sQTL) were also integrated with the GWAS data to prioritize candidate regulatory variants[17] but in pancreatic islets, which is a target tissue for T2D. Both analyses are summarized in Supplementary Data 10 and Supplementary Data 11, respectively.

Conditional analysis. To confirm the independence between novel loci and previously known T2D signals, we performed reciprocal conditional analyses (Supplementary Data 5, Supplementary Data 12, Supplementary Data 13, and Supplementary Data 14). We included the conditioning SNP as a covariate in the logistic regression model, assuming that every residual signal that arises corresponds to a secondary signal independent from this conditioning SNP. We applied this method to the *EHMT2* locus (less than 1Mb away from the *HLA* where T2D and T1D signals have been identified), to confirm that this association was independent of previously reported T2D signals and also to discard that this association is also driven by possible contamination of T1D diagnosed as T2D cases. We conditioned on the top variant identified in this study and the top variant from the 99% credible set analysis, but also on the top variants previously described for T2D and T1D[8, 38–40]. For this purpose, we used the full 70KforT2D resource (NuGENE, GERA, FUSION, GENEVA, and WTCCC cohorts imputed with 1000G and UK10K reference panels). Finally, all the results were meta-analyzed as explained in previous sections. These analyses are provided in Supplementary Data 13. This approach was also applied to confirm that the novel *CAMKK2* signal at rs3794205 is independent of known T2D signals at the *HNF1A* locus (rs1169288, rs1800574, and chr12:121440833:D)[54], which is summarized in Supplementary Data 12. Moreover, this approach confirmed known secondary signals in the 9p21 locus[65] which allowed us to build 99% credible sets based on the results from the conditional analyses (included in Supplementary Data 5), and allowed us to identify the most likely causal variant for the *DUSP9* locus (Supplementary Data 14).

Replication of the rare variant association at Xq23. To replicate the association of the rs146662075 variant, we performed genotype imputation with the UK10K reference panel in four independent data sets: the InterAct case-cohort study[43], the Slim Initiative in Genomic Medicine for the Americas (SIGMA) consortium GWAS data set[6], the Partners HealthCare Biobank (Partners Biobank) data set[44], and the UK Biobank cohort[45]. Phasing was performed with SHAPEIT2 and the IMPUTE2 software was used for genotype imputation.

The current UK Biobank data release did not contain imputed data for the X chromosome, for which phasing and imputation had to be analyzed in-house. The data release used comprises X chromosome QCed genotypes of 488,377 participants, which were assayed using two arrays sharing 95% of marker content (Applied Biosystems^{TM} UK BiLEVE Axiom^{TM} Array and the Applied Biosystems^{TM} UK Biobank Axiom^{TM} Array). We included samples and markers that were used as input for phasing by UK Biobank investigators. At the sample level, we also excluded women, individuals with missing call rate > 5% or showing gender discordance between the reported and the genetically predicted sex. At the variant level, we excluded markers with MAF < 0.1% and with missing call rate > 5%. The final set of 16,463 X chromosome markers and 222,725 male individuals was split into six subsets due to the huge computational burden that would require phasing into whole haplotypes the entire data set. We also excluded indels, variants with MAF < 1%, and variants showing deviation of Hardy–Weinberg equilibrium with $p < 1 \times 10^{-20}$ before the imputation step. In addition, from those pairs of relatives reported to be third degree or higher according to UK Biobank, we excluded from each pair the individual with the lowest call rate. We then tested the rs146662075 variant for association with type 2 diabetes using SNPTEST v2.5.1 and the threshold method. To avoid contamination from other types of diabetes mellitus, we excluded from the entire sample data set, individuals with ICD10 codes falling in any of these categories: E10 (insulin-dependent diabetes mellitus), E13 (other specified diabetes mellitus), and E14 (unspecified diabetes mellitus). Then, we designated as T2D cases those individuals with E11 (non-insulin-dependent diabetes mellitus) ICD10 codes, and the rest as controls. Moreover, we only kept as control subjects those individuals without reported family history of diabetes mellitus and older than 55 years, which is the average age at the onset of T2D.

We also genotyped de novo the rs146662075 variant with KASPar SNP genotyping system (LGC Genomics, Hoddeson, UK) in the Danish cohort, which comprises data from five sample sets (Supplementary Note 2 also for the genotyping and QC analysis for this variant).

We used Cox-proportional hazard regression models to assess the association of the variant with the risk of incident T2D in 1,652 nondiabetic male subjects genotyped in the Inter99 cohort (part of the Danish cohort) that were followed for 11 years on average. The follow-up analysis was restricted to male individuals younger than 45 years who were 56 years old after 11 years of follow-up. Individuals with self-reported diabetes at the baseline examination and individuals present in the Danish National diabetes registry before the baseline examination were also excluded. To include the follow-up study as a part of the replication cohorts, we used a meta-analysis method that accounts for overlapping samples (MAOS)[66], as we had to control for the sample overlap between the follow-up and the case-control study from the Danish samples.

See Supplementary Note 2 for a larger description of each of the five replication cohorts and how they have been processed.

We meta-analyzed the association results from these five replication data sets with the 70KforT2D data sets. In the final meta-analysis, we excluded whenever it was possible (a) controls younger than 55 years and (b) with OGTT > 7.8 mmol l^{-1} or with family history of T2D.

In silico functional characterization of rs146662075. This variant is located in an intergenic region, flanked by *AGTR2* and *SLC6A14* genes, and within several DNase I hypersensitive sites. We searched for regulatory marks (i.e., H3K4me1 and H3K27ac marks) through the HaploReg web server (http://archive.broadinstitute.org/mammals/haploreg/haploreg.php), in order to assess which type of regulatory element was associated with the rs146662075 variant.

To further evaluate the putative regulatory role of rs146662075, we used the WashU EpiGenome Browser (http://epigenomegateway.wustl.edu/browser/, last access on June 2016). We used the following public data hubs: (1) the reference human epigenomes from the Roadmap Epigenomics Consortium track hubs and (2) the Roadmap Epigenomics Integrative Analysis Hub. These data were released by the NIH Roadmap Epigenomics Mapping Consortium[51]. RNA-seq data were used to evaluate whether gene expression of any of the closest genes (*AGTR2* and *SLC6A14* genes, fixed scale at 80 RPKM) correlated with the presence of the H3K27ac enhancer marks (a more strict mark for active enhancers in contrast with H3K4me1[67], which were highlighted by the HaploReg search) at the rs146662075 location. For visualizing the H3K27ac marks around rs146662075, we focused on a region of 8 kb and we used a fixed scale at 40 $-\log_{10}$ Poisson p-value of the counts relative to the expected background count (λ_{local}).

The NIH Roadmap Epigenomics Consortium data from standardized epigenomes also allowed us to further interrogate which target gene within the same topologically associating domain (TAD) was more likely to be regulated by this rs146662075 enhancer. We used H3K27ac narrow peaks from 59 tissues called using MACSv2 with a p-value threshold of 0.01 from 98 consolidated epigenomes to seek for enhancer marks in a given tissue (the presence of H3K27ac peak). To assess gene expression for any of the putative target genes within TAD, we used the RPKM expression matrix for 57 consolidated epigenomes (http://egg2.wustl.edu/roadmap/data/byDataType/rna/) and gene expression quantifications for fetal muscle leg, fetal muscle trunk, and fetal stomach provided by ENCODE (https://www.encodeproject.org/). With this, we were able to test for each of the genes, the association between gene expression and enhancer activity in 31 tissues with a Fisher's exact test.

Allele-specific enhancer activity at rs146662075. The mouse C2C12 cell line (ATCC CRL-1772) was grown in DMEM medium supplemented with 10% FBS and was induced to differentiate in DMEM with 10% horse serum for 4 days.

The human fetal myoblast cell line was established by Prof. Giulio Cossu (Institute of Inflammation and Repair, University of Manchester)[68]. The authors played no role in the procurement of the tissue. Cells were cultured in DMEM medium supplemented with 10% fetal calf serum and was induced to differentiate in DMEM with 2% horse serum for 4 days.

To perform an electrophoretic mobility shift assay, nuclear extracts from mouse myoblast C2C12 cells and the human myoblast cell line (ATCC CRL-1772) were obtained as described before[69]. Double-stranded oligonucleotides containing either the common or rare variants of rs146662075 were labeled using dCTP [α-32P] (Perkin Elmer). Oligonucleotide sequences are as follows (SNP location is underlined): probe-C-F: 5′-gatcTTTGAACACcGAGGGGAAAAT-3′ and R:5′-gatcATTTTCCCCTCgGTGTTCAAA-3′ and probe-T-F: 5′- gatcTTTGAACACtGAGGGGAAAAT-3′ and R: 5′-gatcATTTTCCCCTCaGTGTTCAAA-3′. Assay specificity was assessed by preincubation of nuclear extracts with 50- and 100-fold excess of unlabeled wild-type or mutant probes, followed by electrophoresis on a 5% nondenaturing polyacrylamide gel. Findings were confirmed by repeating binding assays on separate days.

For evaluating if the activity of the rs146662075 enhancer was allele specific, we performed a luciferase assay. A region of 969 bp surrounding rs146662075 was amplified from human genomic DNA using F: 5′-GCTAGCATATGGAGGTGATTTGT-3′ and R: 5′-GGCACTTCCTTCTCTGGTAGA-3′ oligonucleotides and cloned into pENTR/D-TOPO (Invitrogen). Allelic variant rs146662075T was introduced by site-directed mutagenesis using the following primers: F: 5′-CCTTTTTTTACTTTGAACACTGAGGGGAAAATCATGCTTGGC-3′ and R: 5′-GCCAAGCATGATTTTCCCCTCAGTGTTCAAAGTAAAAAAAGG-3′. Enhancer sequences were shuttled into pGL4.23[luc2/minP] vector (Promega) adapted for Gateway cloning (pGL4.23-GW, **2**) using Gateway LR Clonase II Enzyme mix (Invitrogen). Correct cloning was confirmed both by Sanger sequencing and restriction digestion.

C2C12 (ATCC CRL-1772) and 293T (ATCC CRL-3216) cells were transfected in quadruplicates with 500 ng of pGL4.23-GW enhancer containing vectors and 0.2 ng of Renilla normalizer plasmid. Transfections were carried out in 24-well plates using Lipofectamine 2000 and Opti-MEM (Thermo Fisher Scientific) following the manufacturer's instructions. Luciferase activity was measured 48 h after transfection using Dual-Luciferase Reporter Assay System (Promega). Firefly luciferase activity was normalized to Renilla luciferase activity, and the results were expressed as a normalized ratio to the empty pGL4.23[luc2/minP] vector backbone. Experiments were repeated three times. Statistical significance was evaluated through a Student's t-test.

Data availability. The association results are available at the Type 2 Diabetes Knowledge portal (www.type2diabetesgenetics.org/) and the complete summary statistics are available for download at http://cg.bsc.es/70kfort2d/

Received: 13 April 2017 Accepted: 24 November 2017
Published online: 22 January 2018

References

1. Welter, D. et al. The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).
2. Huang, J. et al. Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat. Commun.* **6**, 8111 (2015).
3. Tukiainen, T. et al. Chromosome X-wide association study identifies loci for fasting insulin and height and evidence for incomplete dosage compensation. *PLoS Genet.* **10**, e1004127 (2014).
4. Flannick, J. & Florez, J. C. Type 2 diabetes: genetic data sharing to advance complex disease research. *Nat. Rev. Genet.* **17**, 535–549 (2016).
5. Steinthorsdottir, V. et al. Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. *Nat. Genet.* **46**, 294–298 (2014).
6. Fuchsberger, C. et al. The genetic architecture of type 2 diabetes. *Nature* **536**, 41–47 (2016).
7. Abecasis, G. R. et al. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
8. DIAbetes Genetics Replication and Meta-analysis (DIAGRAM) Consortium et al. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat. Genet.* **46**, 234–244 (2014).
9. Pers, T. H. et al. Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.* **6**, 5890 (2015).
10. Pasquali, L. et al. Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat. Genet.* **46**, 136–143 (2014).
11. Wellcome Trust Case Control Consortium et al. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat. Genet.* **44**, 1294–1301 (2012).
12. Diabetes Genetics Initiative of Broad Institute of Harvard and MIT et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316**, 1331–1336 (2007).
13. McLaren, W. et al. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069–2070 (2010).
14. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
15. Huang, Y. F., Gulko, B. & Siepel, A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat. Genet.* **49**, 618–624 (2017).
16. Mele, M. et al. Human genomics. The human transcriptome across tissues and individuals. *Science* **348**, 660–665 (2015).
17. Fadista, J. et al. Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. *Proc. Natl Acad. Sci. USA* **111**, 13924–13929 (2014).
18. Manning, A. K. et al. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat. Genet.* **44**, 659–669 (2012).
19. Randall, J. C. et al. Sex-stratified genome-wide association studies including 270,000 individuals show sexual dimorphism in genetic loci for anthropometric traits. *PLoS Genet.* **9**, e1003500 (2013).
20. Berndt, S. I. et al. Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nat. Genet.* **45**, 501–512 (2013).
21. Dastani, Z. et al. Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: a multi-ethnic meta-analysis of 45,891 individuals. *PLoS Genet.* **8**, e1002607 (2012).
22. Fox, C. S. et al. Genome-wide association for abdominal subcutaneous and visceral adipose reveals a novel locus for visceral fat in women. *PLoS. Genet.* **8**, e1002695 (2012).
23. Speliotes, E. K. et al. Genome-wide association analysis identifies variants associated with nonalcoholic fatty liver disease that have distinct effects on metabolic traits. *PLoS Genet.* **7**, e1001324 (2011).
24. Lei, X., Callaway, M., Zhou, H., Yang, Y. & Chen, W. Obesity associated Lyplal1 gene is regulated in diet induced obesity but not required for adipocyte differentiation. *Mol. Cell. Endocrinol.* **411**, 207–213 (2015).
25. Wessel, J. et al. Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility. *Nat. Commun.* **6**, 5897 (2015).
26. the CARDIoGRAMplusC4D Consortium. A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat Genet.* **47**, 1121–1130 (2015).
27. Yamamoto, F., Clausen, H., White, T., Marken, J. & Hakomori, S. Molecular genetic basis of the histo-blood group ABO system. *Nature* **345**, 229–233 (1990).
28. Fagherazzi, G., Gusto, G., Clavel-Chapelon, F., Balkau, B. & Bonnet, F. ABO and Rhesus blood groups and risk of type 2 diabetes: evidence from the large E3N cohort study. *Diabetologia* **58**, 519–522 (2015).
29. Gradwohl, G., Dierich, A., LeMeur, M. & Guillemot, F. Neurogenin3 is required for the development of the four endocrine cell lineages of the pancreas. *Proc. Natl Acad. Sci. USA* **97**, 1607–1611 (2000).
30. Rubio-Cabezas, O. et al. Permanent neonatal diabetes and enteric anendocrinosis associated with biallelic mutations in NEUROG3. *Diabetes* **60**, 1349–1353 (2011).
31. Beck, A. et al. An siRNA screen identifies transmembrane 7 superfamily member 3 (TM7SF3), a seven transmembrane orphan receptor, as an inhibitor of cytokine-induced death of pancreatic beta cells. *Diabetologia* **54**, 2845–2855 (2011).
32. Todd, J. N. et al. Variation in glucose homeostasis traits associated with P2RX7 polymorphisms in mice and humans. *J. Clin. Endocrinol. Metab.* **100**, E688–E696 (2015).
33. Hinke, S. A., Hellemans, K. & Schuit, F. C. Plasticity of the beta cell insulin secretory competence: preparing the pancreatic beta cell for the next meal. *J. Physiol.* **558**, 369–380 (2004).
34. Saxena, R. et al. Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge. *Nat. Genet.* **42**, 142–148 (2010).
35. Lyssenko, V. et al. Pleiotropic effects of GIP on islet function involve osteopontin. *Diabetes* **60**, 2424–2433 (2011).
36. Arai, T., Kano, F. & Murata, M. Translocation of forkhead box O1 to the nuclear periphery induces histone modifications that regulate transcriptional repression of PCK1 in HepG2 cells. *Genes. Cells* **20**, 340–357 (2015).
37. Cook, J. P. & Morris, A. P. Multi-ethnic genome-wide association study identifies novel locus for type 2 diabetes susceptibility. *Eur. J. Hum. Genet.* **24**, 1175–1180 (2016).
38. Barrett, J. C. et al. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat. Genet.* **41**, 703–707 (2009).
39. Hakonarson, H. et al. A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene. *Nature* **448**, 591–594 (2007).
40. Wellcome Trust Case Control Consortium et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
41. Gritti, M. et al. Metformin repositioning as antitumoral agent: selective antiproliferative effects in human glioblastoma stem cells, via inhibition of CLIC1-mediated ion current. *Oncotarget* **5**, 11252–11268 (2014).
42. Voight, B. F. et al. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat. Genet.* **42**, 579–589 (2010).
43. Langenberg, C. et al. Gene-lifestyle interaction and type 2 diabetes: the EPIC interact case-cohort study. *PLoS. Med.* **11**, e1001647 (2014).
44. Karlson, E. W., Boutin, N. T., Hoffnagle, A. G. & Allen, N. L. Building the partners healthcare biobank at partners personalized medicine: informed consent, return of research results, recruitment lessons and operational considerations. *J Pers Med* **6**, 2 (2016).
45. Bycroft, C. et al. Genome-wide genetic data on ~500,000 UK Biobank participants. Preprint at *bioRxiv* https://doi.org/10.1101/166298 (2017).
46. Becerra, M. B. & Becerra, B. J. Disparities in age at diabetes diagnosis among Asian Americans: Implications for early preventive measures. *Prev. Chronic Dis.* **12**, E146 (2015).
47. Bartoli, E., Fra, G. P. & Carnevale Schianca, G. P. The oral glucose tolerance test (OGTT) revisited. *Eur. J. Intern. Med.* **22**, 8–12 (2011).
48. Shao, C., Zucker, I. H. & Gao, L. Angiotensin type 2 receptor in pancreatic islets of adult rats: a novel insulinotropic mediator. *Am. J. Physiol. Endocrinol. Metab.* **305**, E1281–E1291 (2013).
49. Yvan-Charvet, L. et al. Deletion of the angiotensin type 2 receptor (AT2R) reduces adipose cell size and protects from diet-induced obesity and insulin resistance. *Diabetes* **54**, 991–999 (2005).
50. Liu, M., Jing, D., Wang, Y., Liu, Y. & Yin, S. Overexpression of angiotensin II type 2 receptor promotes apoptosis and impairs insulin secretion in rat insulinoma cells. *Mol. Cell. Biochem.* **400**, 233–244 (2015).
51. Roadmap Epigenomics, C. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
52. Moltke, I. et al. A common greenlandic TBC1D4 variant confers muscle insulin resistance and type 2 diabetes. *Nature* **512**, 190–193 (2014).
53. Sigma Type 2 Diabetes Consortium. et al. Association of a low-frequency variant in HNF1A with type 2 diabetes in a Latino population. *JAMA* **311**, 2305–2314 (2014).
54. Gaulton, K. J. et al. Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. *Nat. Genet.* **47**, 1415–1425 (2015).
55. Frantz, E. D., Crespo-Mascarenhas, C., Barreto-Vianna, A. R., Aguila, M. B. & Mandarim-de-Lacerda, C. A. Renin-angiotensin system blockers protect pancreatic islets against diet-induced obesity and insulin resistance in mice. *PLoS ONE* **8**, e67192 (2013).
56. Leung, P. S. Mechanisms of protective effects induced by blockade of the renin-angiotensin system: novel role of the pancreatic islet angiotensin-generating system in Type 2 diabetes. *Diabet. Med.* **24**, 110–116 (2007).
57. Geng, D. F., Jin, D. M., Wu, W., Liang, Y. D. & Wang, J. F. Angiotensin converting enzyme inhibitors for prevention of new-onset type 2 diabetes

mellitus: a meta-analysis of 72,128 patients. *Int. J. Cardiol.* **167**, 2605–2610 (2013).

58. Investigators, D. T. et al. Effect of ramipril on the incidence of diabetes. *N. Engl. J. Med.* **355**, 1551–1562 (2006).
59. The ups and downs of data sharing in science. *Nature* **534**, 435-436 (2016).
60. Anderson, C. A. et al. Data quality control in genetic case-control association studies. *Nat. Protoc.* **5**, 1564–1573 (2010).
61. Delaneau, O., Zagury, J. F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10**, 5–6 (2013).
62. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529 (2009).
63. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
64. Cowper-Sal lari, R. et al. Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nat. Genet.* **44**, 1191–1198 (2012).
65. Shea, J. et al. Comparing strategies to fine-map the association of common SNPs at chromosome 9p21 with type 2 diabetes and myocardial infarction. *Nat. Genet.* **43**, 801–805 (2011).
66. Lin, D. Y. & Sullivan, P. F. Meta-analysis of genome-wide association studies with overlapping subjects. *Am. J. Hum. Genet.* **85**, 862–872 (2009).
67. Creyghton, M. P. et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl Acad. Sci. USA* **107**, 21931–21936 (2010).
68. Cossu, G., Cicinelli, P., Fieri, C., Coletta, M. & Molinaro, M. Emergence of TPA-resistant ‘satellite’ cells during muscle histogenesis of human limb. *Exp. Cell. Res.* **160**, 403–411 (1985).
69. Boj, S. F., Parrizas, M., Maestro, M. A. & Ferrer, J. A transcription factor regulatory circuit in differentiated pancreatic cells. *Proc. Natl Acad. Sci. USA* **98**, 14481–14486 (2001).

Acknowledgements

This work has been sponsored by the grant SEV-2011-00067 of Severo Ochoa Program, awarded by the Spanish Government. This work was supported by an EFSD/Lilly research fellowship. Josep M. Mercader was supported by Sara Borrell Fellowship from the Instituto Carlos III and Beatriu de Pinós fellowship from the Agency for Management of University and Research Grants (AGAUR). Sílvia Bonàs was FI-DGR Fellowship from FI-DGR 2013 from Agència de Gestió d’Ajuts Universitaris i de Recerca (AGAUR, Generalitat de Catalunya). This study makes use of data generated by the WTCCC. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113. This study also makes use of data generated by the UK10K Consortium, derived from samples from UK10K COHORT IMPUTATION (EGAS00001000713). A full list of the investigators who contributed to the generation of the data is available in www.UK10K.org. Funding for UK10K was provided by the Wellcome Trust under award WT091310. We acknowledge PRACE for awarding us to access MareNostrum supercomputer, based in Spain at Barcelona. The technical support group, particularly Pablo Ródenas and Jorge Rodríguez, from the Barcelona Supercomputing Center is gratefully acknowledged. This project has received funding from the European Union’s Horizon 2020 research and innovation program under grant agreement No 667191. Mercè Planas-Fèlix is funded by the Obra Social Fundación la Caixa fellowship under the Severo Ochoa 2013 program. Work from Irene Miguel-Escalada, Ignasi Moran, Goutham Atla, and Jorge Ferrer was supported by the National Institute for Health Research (NIHR) Imperial Biomedical Research Centre, the Wellcome Trust (WT101033), Ministerio de Economía y Competitividad (BFU2014-54284-R) and Horizon 2020 (667191). Irene Miguel-Escalada has received funding from the European Union’s Horizon 2020 research and innovation program under the Marie Sklodowska–Curie grant agreement No 658145. We acknowledge Prof. Giulio Cossu (Institute of Inflammation and Repair, University of Manchester) for providing the muscle myoblast cell line. We also acknowledge the InterAct and SIGMA Type 2 Diabetes Consortia for access to the data to replicate the rs146662075 variant. A full list of the investigators of the SIGMA Type 2 Diabetes and the InterAct consortia is provided in Supplementary Notes 3 and 4. The Novo Nordisk Foundation Center for Basic Metabolic Research is an independent research center at the University of Copenhagen partially funded by an unrestricted donation from the Novo Nordisk Foundation (www.metabol.ku.dk). This research has been conducted using the UK Biobank Resource (application number 16803). We also acknowledge Bianca C. Porneala, MS for his technical assistance in the collection and curation of the genotype and phenotype data from Partners Biobank. We also thank Marcin von Grotthuss for their support for uploading the summary statistics data to the Type 2 Diabetes Genetic Portal (AMP-T2D portal). Finally, we thank all the Computational Genomics group at the BSC for their helpful discussions and valuable comments on the manuscript.

Author contributions

S.B-G., J.M.M., and D.T. conceived, planned, and performed the main analyses. S.B-G., J.M.M., and D.T. wrote the manuscript. M.G-M., F.S., P.C-S., M.P., C.D., and R.M.B. developed a framework for large-scale imputation analyses. E.R-F., P.T., and T.H.P. performed pathway analysis. I.M-E. performed the enrichment analysis. M.P-F. and S.G. performed structural variant analyses. N.G., J.R-G., J.M., E.A.A., M.U., A.L., V.K., J.F., T.J., A.L., M.E.J., D.R.W., C.C., I.B., E.V.A., R.A.S., J.L., C.L., N.J.W., O.P., J.C.F., and T.H. contributed with additional data and analyses. G.A., I.M., and C.C.M. performed additional bioinformatics analyses. D.S. and A.Z. contributed muscle cell lines. I.M-E. and J.F. performed luciferase and electrophoretic mobility shift assays. J.M.M. and D.T. designed and supervised the study. All authors reviewed and approved the final manuscript.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-017-02380-9.

Competing interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Sílvia Bonàs-Guarch[1], Marta Guindo-Martínez[1], Irene Miguel-Escalada[2,3,4], Niels Grarup[5], David Sebastian[3,6,7], Elias Rodriguez-Fos[1], Friman Sánchez[1,8], Mercè Planas-Fèlix[1], Paula Cortes-Sánchez[1], Santi González[1], Pascal Timshel[5,9], Tune H. Pers[5,9,10,11], Claire C. Morgan[4], Ignasi Moran[4], Goutham Atla[2,3,4], Juan R. González[12,13,14], Montserrat Puiggros[1], Jonathan Martí[8], Ehm A. Andersson[5], Carlos Díaz[8], Rosa M. Badia[8,15], Miriam Udler[16,17], Aaron Leong[17,18], Varindepal Kaur[17], Jason Flannick[16,17,19], Torben Jørgensen[20,21,22], Allan Linneberg[20,23,24], Marit E. Jørgensen[25,26], Daniel R. Witte[27,28], Cramer Christensen[29], Ivan Brandslund[30,31], Emil V. Appel[5], Robert A. Scott[32], Jian’an Luan[32], Claudia Langenberg[32], Nicholas J. Wareham[32], Oluf Pedersen[5], Antonio Zorzano[3,6,7], Jose C Florez[16,17,33], Torben Hansen[5,34], Jorge Ferrer[2,3,4], Josep Maria Mercader[1,16,17] & David Torrents[1,35]

[1]Barcelona Supercomputing Center (BSC), Joint BSC-CRG-IRB Research Program in Computational Biology, 08034 Barcelona, Spain. [2]Genomic Programming of Beta-cells Laboratory, Institut d'Investigacions August Pi i Sunyer (IDIBAPS), 08036 Barcelona, Spain. [3]Instituto de Salud Carlos III, Centro de Investigación Biomédica en Red de Diabetes y Enfermedades Metabólicas Asociadas (CIBERDEM), 28029 Madrid, Spain. [4]Section of Epigenomics and Disease, Department of Medicine, Imperial College London, London W12 0NN, UK. [5]The Novo Nordisk Foundation Center for Basic Metabolic Research, Section for Metabolic Genetics, Faculty of Health and Medical Sciences, University of Copenhagen, 2100 Copenhagen, Denmark. [6]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldiri Reixac 10-12, 08028 Barcelona, Spain. [7]Departament de Bioquímica i Biomedicina Molecular, Facultat de Biologia, Universitat de Barcelona, 08028 Barcelona, Spain. [8]Computer Sciences Department, Barcelona Supercomputing Center (BSC-CNS), 08034 Barcelona, Spain. [9]Department of Epidemiology Research, Statens Serum Institut, 2300 Copenhagen, Denmark. [10]Division of Endocrinology and Center for Basic and Translational Obesity Research, Boston Children's Hospital, Boston, MA 02116, USA. [11]Medical and Population Genetics Program, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. [12]ISGlobal, Centre for Research in Environmental Epidemiology (CREAL), 08003 Barcelona, Spain. [13]CIBER Epidemiología y Salud Pública (CIBERESP), 28029 Madrid, Spain. [14]Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain. [15]Artificial Intelligence Research Institute (IIIA), Spanish Council for Scientific Research (CSIC), 28006 Madrid, Spain. [16]Programs in Metabolism and Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA. [17]Diabetes Unit and Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA 02114, USA. [18]Division of General Internal Medicine, Massachusetts General Hospital, Boston, MA 02114, USA. [19]Department of Molecular Biology, Harvard Medical School, Boston, MA 02114, USA. [20]Research Centre for Prevention and Health, Capital Region of Denmark, DK-2600 Glostrup, Denmark. [21]Faculty of Health and Medical Sciences, University of Copenhagen, DK-2200 Copenhagen, Denmark. [22]Faculty of Medicine, University of Aalborg, DK-9220 Aalborg East, Denmark. [23]Department of Clinical Experimental Research, Rigshospitalet, Glostrup, 2100 Copenhagen, Denmark. [24]Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, DK-2200 Copenhagen, Denmark. [25]Steno Diabetes Center, 2820 Gentofte, Denmark. [26]National Institute of Public Health, Southern Denmark University, DK-5230 Odense M, Denmark. [27]Department of Public Health, Aarhus University, DK-8000 Aarhus C, Denmark. [28]Danish Diabetes Academy, DK-5000 Odense C, Denmark. [29]Medical department, Lillebaelt Hospital, 7100 Vejle, Denmark. [30]Department of Clinical Biochemistry, Lillebaelt Hospital, 7100 Vejle, Denmark. [31]Institute of Regional Health Research, University of Southern Denmark, DK-5230 Odense, Denmark. [32]MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, Cambridge Biomedical Campus, Cambridge CB2 0QQ, UK. [33]Department of Medicine, Harvard Medical School, Boston, MA 02115, USA. [34]Faculty of Health Sciences, University of Southern Denmark, DK-5230 Odense M, Denmark. [35]Institució Catalana de Recerca i Estudis Avançats (ICREA), 08010 Barcelona, Spain. Josep Maria Mercader and David Torrents jointly supervised this work.