



UNIVERSITAT DE  
BARCELONA

## Detection and classification of somatic structural variants, and its application in the study of neuronal development

Mercè Planas Fèlix

**ADVERTIMENT.** La consulta d'aquesta tesi queda condicionada a l'acceptació de les següents condicions d'ús: La difusió d'aquesta tesi per mitjà del servei TDX ([www.tdx.cat](http://www.tdx.cat)) i a través del Dipòsit Digital de la UB ([diposit.ub.edu](http://diposit.ub.edu)) ha estat autoritzada pels titulars dels drets de propietat intel·lectual únicament per a usos privats emmarcats en activitats d'investigació i docència. No s'autoritza la seva reproducció amb finalitats de lucre ni la seva difusió i posada a disposició des d'un lloc aliè al servei TDX ni al Dipòsit Digital de la UB. No s'autoritza la presentació del seu contingut en una finestra o marc aliè a TDX o al Dipòsit Digital de la UB (framing). Aquesta reserva de drets afecta tant al resum de presentació de la tesi com als seus continguts. En la utilització o cita de parts de la tesi és obligat indicar el nom de la persona autora.

**ADVERTENCIA.** La consulta de esta tesis queda condicionada a la aceptación de las siguientes condiciones de uso: La difusión de esta tesis por medio del servicio TDR ([www.tdx.cat](http://www.tdx.cat)) y a través del Repositorio Digital de la UB ([diposit.ub.edu](http://diposit.ub.edu)) ha sido autorizada por los titulares de los derechos de propiedad intelectual únicamente para usos privados enmarcados en actividades de investigación y docencia. No se autoriza su reproducción con finalidades de lucro ni su difusión y puesta a disposición desde un sitio ajeno al servicio TDR o al Repositorio Digital de la UB. No se autoriza la presentación de su contenido en una ventana o marco ajeno a TDR o al Repositorio Digital de la UB (framing). Esta reserva de derechos afecta tanto al resumen de presentación de la tesis como a sus contenidos. En la utilización o cita de partes de la tesis es obligado indicar el nombre de la persona autora.

**WARNING.** On having consulted this thesis you're accepting the following use conditions: Spreading this thesis by the TDX ([www.tdx.cat](http://www.tdx.cat)) service and by the UB Digital Repository ([diposit.ub.edu](http://diposit.ub.edu)) has been authorized by the titular of the intellectual property rights only for private uses placed in investigation and teaching activities. Reproduction with lucrative aims is not authorized nor its spreading and availability from a site foreign to the TDX service or to the UB Digital Repository. Introducing its content in a window or frame foreign to the TDX service or to the UB Digital Repository is not authorized (framing). Those rights affect to the presentation summary of the thesis as well as to its contents. In the using or citation of parts of the thesis it's obliged to indicate the name of the author.

# Detection and classification of somatic variants, and its application in the study of neuronal development

Ph.D. Thesis

**Mercè Planas Fèlix**

JUNE 2020



UNIVERSITAT DE  
BARCELONA

Facultat de Biologia  
Programa de Biomedicina (codi HDK05)  
Línia de recerca 101114 – Bioinformàtica

Memòria presentada per Mercè Planas Fèlix per optar al grau de doctor/a per  
la Universitat de Barcelona

## **Detection and classification of somatic structural variants, and its application in the study of neuronal development**

DOCTORANDA:

Mercè Planas Fèlix

DIRECTOR:

Dr. David Torrents Arenales

TUTOR:

Dr. Josep Lluís Gelpi Buchaca



**Barcelona  
Supercomputing  
Center**  
Centro Nacional de Supercomputación



This thesis was supported by

“La Caixa” – Severo Ochoa International Ph.D. Programme





# Agraïments

Sóc conscient de que aquesta tesi esta signada al meu nom, però no puc pas concebre la idea de la seva existència sense el recolzament i la presencia de tota la gent que m'ha recolzat durant tot aquest llarg camí. Tinc tot el meu convenciment que sense valtros aquesta tesis mai hauria hagués estat possible, gràcies per haver format part d'aquesta etapa que aquí culmina. En especial però, m'agradaria agrair a aquelles persones que considero que han estat claus en aquesta etapa.

En primer lloc voldria agrair al meu Director de tesi el Dr. David Torrents Arenales per donar-me l'oportunitat de poder cursar la tesi aquí present en el seu grup durant aquests anys, així com La Caixa per la beca que em van donar per poder cursar el meu doctorat.

Tornant amb el David, desitjo agrair-li la confiança que diposita en les persones, com es permet veure més enllà de unes notes i creure en els valors de cadascun de naltrus, aquest tracte més humà i personal que ha tingut sempre amb tots els membres del nostre grup. No han estat poques les converses que he disfrutat no només de ciencia, sino també de motivació, de formes de pensar, tot i el nostre tarannà tan diferent, que m'emporto amb mi i que se que les tindre present al llarg de la meva carrera. Considero que en un món com és el de la ciencia hi hauria d'haver més directors que et donessin l'oportunitat de poder treballar d'allò que més t'entusiasma com ho ha fet ell amb tots natros.

En aquest grup és on he pogut conèixer no només grans treballadors sino també grans persones que he tingut el privilegi que han passat a formar part del meu cercle d'amistats.

Voldria agrair a tots valtros haver format un grup com el que tenim, però no vull perdre l'oportunitat de poder-ho fer de forma més personal amb alguns d'ells.

Al Txema, que no només va ser la primera persona que vaig conèixer del grup, sinó també el meu supervisor durant la meva etapa en metagenòmica. Treballar amb ell m'ha ensenyat a ser constant, i tenir-ho tot controlat, però sobretot a esclatar de riure inclús en els moments més tensos de treball. El seu esperit jovial, el seu humor i la seva franquesa sempre ens ha contagiats a tots. Em quedo en el record el soroll de la brompton al entrar a la sala i els esmorzars gourmets que ens donem tot parlant de la vida.

A la Silvia, la primera doctorant amb qui vaig compartir experiències del grup i que va tenir la santa paciència d'ensenyar-me R. Gràcies per deixar-me treballar al teu costat i ensenyar-me valors com la eficiència, i poder compartir la teva gran implicació en la ciència que has demostrat sempre dia rere dia. Però sobretot gràcies per l'amistat tant sincera i mordaç que m'has ofert al llarg de tots aquests anys i que continua viva a dia d'avui, i per donar-me sempre aquella empenta en els meus moments més indecisos.

A la Marta i l'Elias, amb els quals ha estat un plaer i privilegi poder compartir aquesta etapa d'inici a final tots tres plegats. Anar a treballar cada dia al BSC es feia menys feixuc i es convertia en una festa sabent que els trobaria en la meua illa per poder petar la xerrada. Valoro l'amistat i la germanor que s'ha creat i que m'han permès poder compartir i disfrutar tant dels bons com els mal moments, aquelles converses que semblava que arreglaven el món o inclús les que ens sortia el "monocle", les birres i els cafés a la FIB, els sopars, els viatges, les festes! Se que gràcies a ells recordaré aquesta etapa com una de les millors que hauré viscut, i que sense ells això no hauria estat mai com és ara.

A la Montse, per no només ser la nostre informàtica i ensenyar-nos a tots amb una paciència desmesurada, sinó també a adquirir autonomia en l'entorn més



computacional. Gràcies també per tot l'afecte que m'has demostrat durant aquesta etapa compartint tants càfes i xerrades plegades.

A la Romina que tot i no ser del nostre grup sempre ha estat allà per donar-nos un cop de mà sempre que el necessitavem, i ens ha alimentat amb els seus grans postres setmana rere setmana.

Fora del nostre grup destacar al Jordà, el meu company en l'elaboració de SMuFin2. Treballar al seu costat ha estat un plaer no només a nivell de coneixements, sinó per ensenyar-me que encara que dues persones vinguin d'ambients i formes de raonar molt diferents la comunicació i els resultats poden ser fantàstics.

Fora del meu entorn més laboral voldria destacar el gran paper de tots els meus amics, la meva petita "família" que m'han donat tot el seu recolzament i la confiança per continuar endavant en aquesta etapa. Alguns d'ells de l'etapa del màster com l'Alba i l'Oihane, que han estat crucials per continuar amb la meva carrera d'investigació i que han fet possible la tesis aquí present. Altres de la carrera o de la infància que sempre han i han celebrat tots els meus triomfs com si fossin els seus propis, gràcies per no deixar-me enfonsar i confiar en mi en els bons i mal moments. A l'Oriol que durant molts anys va haver de patir que la feina em tingués sempre tant segrestada però no va deixar mai de confiar que tard o d'hora tota aquesta feina donaria els seus fruits.

M'agradaria destacar sobretot a les meves amigues i companyes de pis Maria i Ari. Gràcies per tirar de mi, cuidar-me i aguantar-me sobretot en aquest últim tram de la tesis que ha suposat una muntanya russa d'emocions. Tinc molt clar que no estaria aquí de no ser per valtros, i no sabria com agrair-vos tot el

recolzament que m'heu donat per continuar amb una de les meves passions que és la investigació, amb valtros em sento com a casa.

Finalment a la meva família per sempre recolzarme i respectar-me en totes les decisions que he pres tot i no tenir molt clar a que em dedicava en alguns moments, gràcies pel vostre suport incondicional. En especial a la meva mare per haver-me ensenyat el vertader significat de la paraula resiliència. La seva forma d'afrontar qualsevol entrebanc, inclús les pitjors adversitats, amb el millor somriure, de no fer un pas enrere quan les coses van maldades, de la seva forma tant sincera i determinada de viure. Totes aquestes ensenyances m'han permès poder treure les forces necessàries per poder embarcar-me en la tesis i poder-la finalitzar, podent valorar tot el procés amb la millor de les actituds. Gràcies per haver estat el meu principal pilar.

Per últim, m'agradria també agrair agrair als membres del tribunal haver acceptat la invitació a l'acte de defensa de la meva tesi. És un plaer i un honor per mi poder comptar amb la seva presència i experiència en la defensa de la tesis.

-M-





Per la cosa més bonica  
de la capa de la terra.  
-M-



# Strategy and Trajectory of my thesis

I would like to first provide context to the path and trajectory of my thesis, as this is essential for its assessment.

My research was initially focused on metagenomics. Within this field, I actively collaborated within an existing research line of the group. In this study, we characterized and analyzed the metaregulome of three different environments: Acid mine, Whale Fall, and Waseca Farm, and their impact in the adaptation to particular variable physicochemical conditions. Our results highlighted the potential effects of gene regulation on the adaptation of bacteria through habitats, by distributing their regulatory potential among specific functions. My contribution there consisted in the analysis of transcription factor regulatory networks underlying bacterial adaptive changes, and in the drafting of the manuscript (Fernandez et al., 2014). Afterwards, and following the metagenomics line, I also participated in another project that aimed at assessing the impact of metformin, a common treatment of type 2, on the composition and dynamics of the gut microbiome of patients (Wu et al., 2017). This work was performed in collaboration with Dr. Josep Manuel Fernández from the Trueta Hospital in Girona. After this contribution, and considering the difficulty of accessing proper metagenomic data to continue this research line and answering questions regarding

microbiome composition in diseases, forced us to change the direction of the research line.

For this reason, we redirected the thesis trajectory towards the identification and characterization of large-scale structural variants, which was, at that time, an emerging research line in the group with more possibilities of publication. Our lab, at that time, was devoting efforts to the analysis of structural variation associated to complex diseases, and to cancer, through the development of a novel algorithm for the identification of somatic structural variants, SMuFin (Moncunill et al., 2014), which has been applied since then to large-scale projects such as the ICGC and the PCAWG ICGC/TCGA Pan-Cancer (Consortium, 2020).

In line with this, I first participated in a meta-analysis for type 2 diabetes based on the re-analysis of publicly available individual genetic data for up. There, I validated and interpreted small to medium-size insertions and deletions (Indels) that were identified via genotype imputation, using novel sequence-based reference panels, such as the UK10K (Consortium et al., 2015) and 1000 Genomes Project (Genomes Project et al., 2015). This study demonstrated the value of reanalyzing existing genetic datasets for GWAS through a deeper variant analysis, like expanding to indels and other structural variants.

In parallel to this study, I already started to work on the second version of the SMuFin algorithm. My initial goal was to improve the detection capabilities and scalability of the original algorithm, by adding novel features and, at the same time, enhancing the computational performance, addressing those scenarios where SMuFin was offering poor results. During one year I tried to work on top of the original code but this strategy was not fruitful due to the obfuscation of the code, and the impossibility to communicate with the original developer. Thereafter, we decided to develop a new algorithm from scratch, SmuFin2, capable of identifying more efficiently a larger spectrum of somatic genetic variants and, at the same time, improving the scalability of the



code. The SMuFin2 algorithm presented below is the result of the close collaboration with the Data-Centric Computing group from the Barcelona Supercomputing Center (BSC), particularly with postdoctoral fellow Jordà Polo, Ph.D. candidate Nicola Cadenelli, and the head of the group David Carrera.

Finally, during the development of SMuFin2, I had the opportunity of collaborating with the research group led by Dr. Alex Kentsis from the Memorial Sloan Kettering (NY) together with my groupmate and Ph.D. student Elías Rodríguez-Fos. This collaboration was a continuation of a previous study towards uncovering the role of *PGBD5*, a transposase-like gene, in the generation of medium-size genomic deletions in cancer (Henssen et al., 2017a). This last study intends to characterize the role of *PGBD5* in the generation of somatic structural variation during the development of neural tissues in the brain. Specifically, here I have applied different variant calling and interpretation strategies to define and describe the landscape of somatic variation in wild type and *Pgbd5* knock-out mice. We expect to finalise the publication, of which I will share a first authorship.

All the aforementioned articles can be consulted in the Publications section, with a brief note describing my particular contribution to them.



# Abstract

The identification and analysis of genomic variation across individuals has been central in biology, first through comparative genomics to answer evolutionary questions, and then in the context of biomedicine, where it is actually becoming central to the study of most diseases. Next generation sequence technologies are allowing the systematic analysis of thousands of different types of genetic variation, enhancing the identification of disease markers and the understanding of the molecular basis of disease. For the past years, there has been a burst of new methodology for genome analysis around diseases coming from hundreds of groups around the world. Specific computational methods and strategies are being designed and improved around the identification and interpretation of genomic variation. The identification and classification of different types of genomic variants in the context of biomedicine is a key and foundational step for the development of a personalized medicine.

This has been particularly central in the field of cancer genomics, which has based the research of the past ten to fifteen years in the sequencing of genomic DNA, and the identification and interpretation of (mostly) somatic and germline variation. Throughout these years, a large number of methods for variant detection have been developed with different action ranges. Despite all these developments, the identification of genomic variants has still room for improvement, not only at the level of sensitivity and specificity, but also at the computational level. Given the emergence of many initiatives for personalized

medicine around the world, and the expected number of genomes that will have to be analyzed within health care systems, we require robust algorithms, designed together with a matching implementation that will minimize the computational costs of the analysis. With this aim, during this thesis, I have pushed and designed and implemented an algorithm for the efficient processing of genomic data, in close collaboration with computer scientists of our center that defined the implementation, focusing on lowering the energy and the time of the analysis. This methodology, which relies on a reference free approach of read classification, has been protected with a patent, and is being used as the foundation for the development of SMuFin2, a more accurate and computationally efficient version of the initial SMuFin from 2014. We here show that our method is able to process whole genome sequences very fast and with a minimal energy consumption, compared with existing methods, and that has great potential for the identification of all ranges of variants, including insertions of non-human DNA. Further developments on SMuFin2 are needed to finally assess its full variant calling capabilities.

Despite their great importance and their clear role in the biology of the cell, somatic variation that occurs in healthy tissues has remained diffuse in their roles. In the case of development, some hypotheses have been proposed to explain the observed somatic DNA damage that occurs during brain development (e.g., replication stress). But the real impact and the underlying mechanisms of this somatic variation are not yet understood. In order to shed light on the type and potential functional impact of somatic variation in brain development, we established a new collaboration to identify, and describe somatic DNA rearrangements induced by Pgbd5 during brain development and adult state in 36 mice neural tissue samples. The detection of somatic variants in healthy tissues presents more challenges than in the cancer scenario, where a variant is present in a significant number of cells and is easier to detect. We have identified, classified and interpreted the landscape of

somatic variation in neural development and identified interesting differences between adult and embryonic variation load, and specific types of variants, as the potential result of the activity of these transposase-like genes.



# Abbreviations

## **BAM**

Binary Alignment Map

## **bp**

base pair

## **BSC**

Barcelona Supercomputing Center

## **BWA**

Burrows-Wheeler Aligner

## **CA**

Contig assembly

## **CESC**

Cervical Squamous Cell Carcinoma

## **CL**

Clustering

## **DSB**

Double Strand Breaks

**EBV**

Epstein–Barr virus

**EGFR**

Epidermal growth factor receptor

**FCD**

Focal cortical dysplasia

**GWAS**

Genome-Wide Association studies

**HBV**

Hepatitis B virus

**HCV**

Hepatitis C virus

**HME**

Hemimegalencephaly

**HPC**

High-performance computing

**HPV**

Human papillomavirus

**IARC**

International Agency for Research on Cancer



**ICGC**

International Cancer Genome Consortium

**Indel**

Insertion or deletion

**ITR**

Inverted terminal repeats

**KO**

Knockout

**LIHC**

Liver Hepatocellular carcinoma

**MSK**

Memorial Sloan Kettering

**NF1**

Neurofibromatosis 1

**NGS**

Next-generation sequencing

**NT**

Nucleotide

**PCAWG**

Pancancer Analysis of Whole Genomes

## **PGBD5**

*PiggyBac Transposable Element Derived 5*

## **SA**

Split-reads alignment

## **SMuFin**

Somatic Mutation Finder

## **SNV**

Single nucleotide variant

## **ST**

Statistical testing

## **SV**

Structural variant

## **PBMC**

Peripheral blood mononuclear cells

## **TCGA**

The Cancer Genome Atlas

## **UCEC**

Uterine Corpus Endometrial Carcinoma

## **UV**

Ultraviolet

**VAF**

Variant allele frequency Variant

**VCF**

Variant Call Format

**VEP**

Variant Effect Predictor

**WGS**

Whole-genome sequencing

**WHO**

World Health Organization

**WT**

Wild type

**2GS**

Second-generation sequencing



# Table of contents

## **Agraïments**

5

---

## **Strategy and Trajectory of my thesis**

13

---

## **Abstract**

17

---

## **Abbreviations**

21

---

## **Table of contents**

27

---

## **Introduction**

1 Somatic Genomic variation: Definition, incidence, and types 34

1.1 The importance of studying somatic mutations 36

1.2 Genome analysis in the era of Next-Generation Sequencing 40

1.2.1 Emerging limitations 42

2 Cancer 42

2.1 General causes of somatic mutations 43

27

2.2 National and International initiatives	45
2.3 Analysis of somatic variation in cancer	47
2.3.1 The calling pipeline	49
2.3.1.1 Reference-based methods	52
2.3.1.2 Reference free methods	53
2.3.1.2.1 Somatic Mutation Finder (SMuFIn)	57
2.3.2 Variant caller virus	58
3 Somatic variations in non-disease scenarios	59
3.1 Somatic variation in Neurodevelopmental diseases	61
3.1.1 Neurodevelopmental genome analysis in the era of NGS	62
3.1.2 Somatic mosaicism in the normal human brain	63
3.1.3 Somatic activity on brain development	63
4 Final considerations	64
	66

---

## Objectives

68

---

## Methods

1 Development of a computer-implemented and reference-free based strategy for identifying variants in nucleic acid sequences	70
------------------------------------------------------------------------------------------------------------------------------	----

72

1.1 Polymorphic k-mer strategy	73
1.1.1 Important terms	73
1.2 Calibrate algorithm	75
1.2.1 Construction of the in-silico chromosome 20	76
1.3 Analysis of the in silico chromosome 20 with the strategy of the invention	76
1.4 PCAWG data	77
1.4.1 Running SMuFin2 to analyze PCAWG data	78
1.4.2 Identification of viruses presence on sample	78
2 Landscape of somatic variation in neural development and the role of Pgbd5	79
2.1 Experimental outline	79
2.2 Analysis of sequenced data	80
2.2.1 Sequenced data alignment	80
2.2.2 Variant calling	80
2.2.2.1 Joining and filtering of variant calling results	81
2.2.3 Variant allele Frequency calculus	81
2.2.4 Identification and analysis of genes	82
2.2.5 Identification and analysis of genomic intervals	82
2.2.6 Detection of motifs	83
2.2.7 Study of genetic ontology	83
	29

---

## Results

1 Design of the algorithm of the Somatic Mutation Finder, version 2 (SMuFin2)	84
1.1 SMuFin2 Algorithm	86
1.1.1 Inputting two sets of nucleic acid reads	87
1.1.2 Quality filtering of the raw sequenced data	90
1.1.3 Generating a hash table structure	90
1.1.3.1 Generation and analysis of k-mers	91
1.1.3.2 Building a hashtable structure	91
1.1.4 Detecting candidate somatic variants with kmers	94
1.1.5 Clustering and filtering candidate somatic variants to build blocks with candidate variants	95
1.2 SMuFin2-algorithm implementation	98
1.3 Algorithm validation	102
1.3.1 Generation of an <i>in silico</i> test sample for initial validation	103
1.4 Identification of tumor-associated viruses	103
1.5 Cataloguing and annotating blocks	108
1.6 Output files	109
1.7 SMuFin2 first block execution	111
	113



1.7.1 Compile	113
2 Landscape of somatic variation in neural development and the role of Pgbd5	117
2.1 Identification of somatic variants within neural tissues	121
2.2 Comparative analysis of somatic variation between adult and embryo	123
2.3 Characterization of the deletions on wild-type mice	125
2.4 Study of Pgbd5 KO mice vs WT mice	128
2.5 Identification and analysis of genes and genomic intervals	129
2.6 Detection of motifs around breakpoints	134
2.7 Functional enrichment analysis of genes affected by pgbd5-dependent deletions	134

---

## Discussion

1 Development of a computer-implemented and reference-free strategy for identifying variants in nucleic acid sequences	136
2 Landscape of somatic variation in neural development and the role of Pgbd5	146

---

## Conclusions

152

---

## Supplementary material

154

---

## Publications

Collaboration 1	170
Collaboration 2	171
Collaboration 3	184
Patent	197
	212

---

## **References**

265



# INTRODUCTION

The identification and analysis of genomic variation across individuals has been central in biology, first through comparative genomics to answer evolutionary questions, and then in the context of biomedicine, where it is actually becoming central to the study of most diseases. Next-generation sequence technologies (NGS) are allowing the systematic analysis of thousands of different types of genetic variation, enhancing the identification of disease markers and the understanding of the molecular basis of disease. For the past years, there has been a burst of new methodology for genome analysis around diseases coming from hundreds of groups around the world. Specific computational methods and strategies are being designed and improved around the identification and interpretation of genomic variation. This covers from Genome-Wide Association studies (GWAS) that aim at identifying risk polymorphic variants for complex diseases, to the analysis of rare germline and somatic mutations associated with rare diseases and cancer, respectively.

Therefore, the identification of different types of genomic variants in the context of biomedicine is a key and foundational step for the development of personalized Medicine. Diagnosis, Prognosis and treatment protocols are starting to be designed around specific genomic changes, which makes their identification crucial. The way genomic variation can influence cellular function and cause disease is very heterogeneous, depending on the location and the type of variation. Therefore, it is very important to find variants, but it is also very important to be able to classify and interpret them.

This thesis has contributed to the generation of genome analysis based strategy to identify somatic variants, as well as to the characterization of the landscape of somatic variants in healthy neural tissues during Neural development mice.

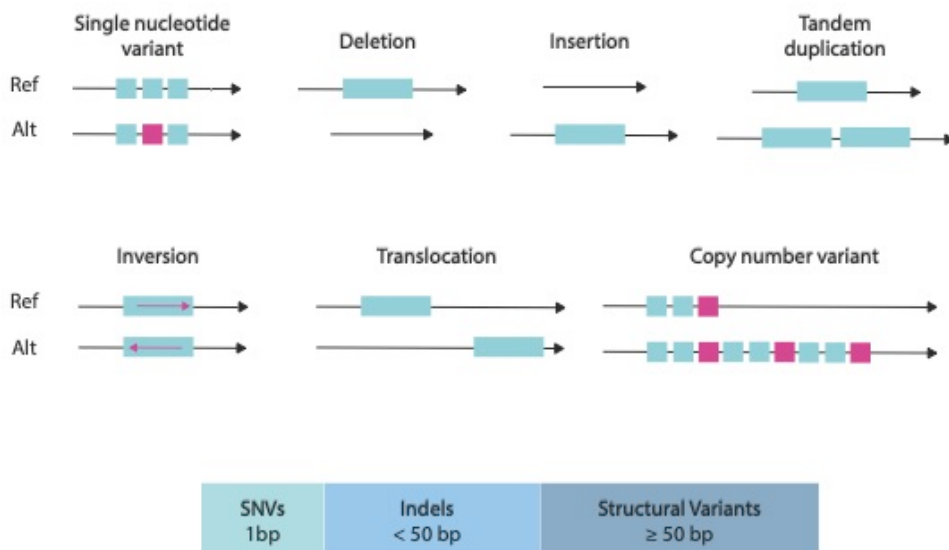
# 1. Somatic Genomic variation: Definition, incidence, and types

In general terms, somatic genomic variation are defined as changes in the genome of somatic (non-germline) cells, and therefore are not transmitted onto the offspring. In contrast to germline variation, which occurs in germline cells and are passed to the offspring, somatic variants have not been so much studied, due to the technical difficulties at the identification and characterization level. With the appearance of the NGS and the possibility to systematically identify and classify somatic variants from tumors, their study has dramatically increased, being now one of the hot research topics in biomedicine. Among others, one of the differences between somatic variants compared to germline variants is the mutation rate. In the case of Single Nucleotide Variant (SNV), it has been observed that the somatic mutation rate is  $2.8 \times 10^{-7}$  and  $4.4 \times 10^{-7}$  per base pair per generation for humans and mice, respectively; and in the case of germline mutation rate is  $1.2 \times 10^{-8}$  and  $5.3 \times 10^{-9}$  mutations per base pair (Milholland et al., 2017). This means that the somatic mutation rate is one order of magnitude higher than the germinal rate for these species.

Although somatic variants comprise all types of possible changes in the genome and can be classified in many different ways and using different criteria, the main classification derives from the methodology developed and used for their detection (Escaramis et al., 2015; Weischenfeldt et al., 2013). As shown in Figure 1, for example, some variants can be balanced, i.e. with no loss or gain of genetic material, such as SNV, inversions, and translocations, while others are considered unbalanced, when a part of the genome is duplicated or lost, like deletions and duplications.

In general this classification starts to be very challenging when considering large and complex chromosomal rearrangements. Within the community, variants are normally classified based on their length, which also agrees with the general detection range of available methodology. Although there is no rule that clearly defines the division between "small" and "large" categories, 50 base pairs (bp), is the cutoff currently considered in most studies (Guan and Sung, 2016; Sudmant et al., 2015; Tattini et al., 2015). Small variants include single nucleotide changes (SNVs), as well as short insertions or deletions (indels), and large variants, also known as chromosomal rearrangements or structural variation (SV), that include, from large deletions and insertions of DNA, to many types of complex variation, like multiple chromosomal rearrangements, transposition, copy of DNA, among others (Yi and Ju, 2018). Structural variants are defined by their breakpoints, which correspond to the points where the rearrangement occurs (Quinlan and Hall, 2012). Originally SVs were defined as insertions, deletions, and inversions higher than 1kb; with the arrival of the sequencing of the human genome, this varied to the current size and type (Alkan et al., 2011; Lupski, 2007).

In summary, the standards for classifying somatic variants, which are necessary for the comparison and globalization of genomic research, are normally based on the length of the change and defined by current variant identification methodology. A more important functional classification of variants is growing within the community, in order to understand the functional impact of the genomic change and translate this knowledge into the understanding of the underlying process, which in the case of disease can ultimately be translated into the development of clinical protocols.



**Figure 1. Diversity of genetic variation.** Depending on the size of the DNA sequence of the variant, we can differentiate between SNVs , Indels and SV.

An important characteristic of variants is the frequency, at which they are represented in a given sample. This, not only affects the possibilities of detection, as variants that are less represented (in less cells) of the sample are more challenging to find, but also is informative of the level of cellular mosaicism within that sample. Please, note that all somatic variants are expected and assumed to be heterozygous, involving less alleles and more difficulties for detection, compared to germline homozygous changes. The Variant Allele Frequency (VAF) is the parameter that measures and quantifies the relative abundance of a given mutated allele within the whole population of different alleles. This parameter follows the formula:

$$f_i = \frac{r_{mut,i}}{r_{mut,i} + r_{ref,i}}.$$



Where  $r_{mut,i}$  are the reads containing the variant, and  $r_{ref,i}$  the reads of the reference allele.

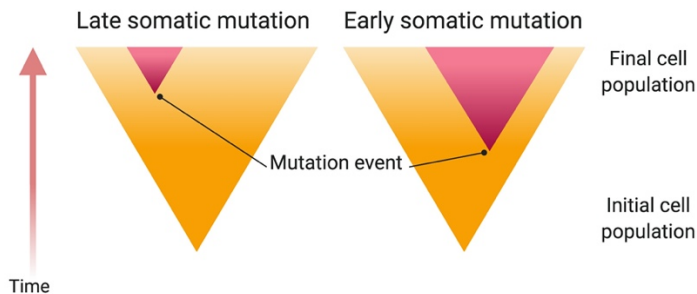
The VAF of a somatic mutation is conditioned by two factors: prevalence and heterogeneity (Figure 2).

Prevalence refers to how widespread the variation is, which depends mainly on how early the mutation occurs during the development. On the other hand, heterogeneity refers to the tissue from which the sample was subtracted for further sequencing.

For example, if a mutation occurs at the early stage of first cell division, and each cell produces the same number of offspring, the expected VAF value in an unbiased sample is around 0.25. On the other hand, if this mutation occurs uniquely in a post-mitosis cell, the VAF would be reduced to an infinitesimal value (Dou et al., 2018). In general terms, variations will have a higher VAF if they have occurred earlier compared to those that have happened later.

Assuming clonal cellular growth, as in tumors for example, the VAF of a variant depends on when that variation happens, relative to the clonal expansion of cells. Genomic variations occurring in the first stages of tumor (clonal) growth will theoretically be present in all of the derived cells, and the VAF would be of 0,5 (assuming heterozygous states of somatic variants), with a cell fraction of 1. This is typical of variants that drive and trigger tumor formation. On the contrary, variants that appear at later stages of the clonal expansion, will have lower representation within the sample, and cover lower cell fractions. These variants with lower VAF values are more difficult to detect (they are represented by less sequencing reads) and are becoming very important to understand the evolution and progression of tumors, as they represent and indicate the existence of different clones that might have different reactions to treatment, and often determine the fate of the patient. The detection of different levels of variants, according to their VAF values,

depends on the sequencing coverage of the sample, that is, the number of times that a single base is represented across all sequencing reads. The sequencing of samples at coverage levels of 30x, or higher, ensures the possibilities of detecting founder and prevalent variants (high VAFs), as well as those that are represented in lower levels, normally up to levels of VAFs of 0,2, or even lower depending on the sensitivity of the analysis methods.



**Figure 2. An early mutation produces a higher proportion of mutant cells in the growing population than a later mutation.** An earlier mutation (right) produces a larger population of mutant cells than a later mutation (left). Depending on when a mutation occurs, the size of the affected cell population differs. (Based on: An Introduction to Genetic Analysis. 7th edition; Griffiths AJF, Miller JH, Suzuki DT, et al.; 2000). Created with Biorender.com

## 1.1 The importance of studying somatic mutations

Somatic mutations accumulate relentlessly in our cells as we age. The concept of accumulation of somatic changes was first proposed more than 50 years ago, and it was associated with aging and even death (Szilard, 1959). Although the majority of somatic variation have no functional consequences and accumulate passively in cells. On the other hand, some somatic mutations can affect functional genomic regions, and have functional and cellular consequences, even leading to disease. Actually, the classification of variants can also follow functional criteria, as in the clinical context, where somatic variants are divided into: (i) those that confer a selective advantage to the cell, increasing survival or proliferation (so-called "driver" mutations, in the context of cancer), (ii)

those that are selectively neutral (iii) and those that are a disadvantageous, often leading to cell death (Martincorena et al., 2018). Important efforts are made in order to infer the potential functional impact of each of the variants detected in a certain study. These predictions are made with different types of programs that cross the location and the type of a variant, with the annotation of functional features in the same genomic region, such as genes, regulatory regions, epigenetic marks, among others. This functional classification of variants are currently used in the clinics for prioritizing those changes that possibly have a diagnosis or treatment value.

The field, in which somatic variants have a key role, and where they have been mostly studied is cancer genomics (see below), for which variation is assumed to be responsible for more than 90% of tumors (Martincorena and Campbell, 2015). In 1914, Boveri (Boveri, 2008) proposed in his book two important concepts between somatic mutations and cancer: control of cell proliferation (*proliferation* as the default state of cells) and carcinogenesis (chromosomal aberrations/mutations) as the cause of cancer.

Despite this clear implication in cancer, somatic variants are also known to be involved in other types of pathologies, like those related to other clonal based cell expansions, like in the hematopoietic system, such as Neurofibromatosis 1 (NF1) (Kehrer-Sawatzki et al., 2004), atrial fibrillation (Gollob et al., 2006), and the Alport syndrome (Krol et al., 2008), or autoimmune disease. It is also known that somatic mutations can play an essential role in some neurological diseases, including autism spectrum disorders, epilepsy, and intellectual disability (Poduri et al., 2013). For some of these diseases, the presence of somatic mutations, even in a small fraction (10%) within specific cell types can trigger the disease (Lee et al., 2012). Finally, somatic variation has also been assigned to physiological (non-pathological) processes (Michikawa et al., 1999). Despite this, there are very few studies tackling somatic variation in non-disease scenarios (see section 3).

## **1.2 Genome analysis in the era of Next-Generation Sequencing**

The NGS emerged at the end of the twentieth century, as a new and revolutionary sequencing approach to overcome the limitations of Sanger-based sequencing technology. The impact of NGS in biomedicine is so enormous that it has revolutionized basal experimental designs, changing the paradigm behind biomedical research. The possibilities of including whole genome sequencing in most research studies in biomedicine, have changed the basic underlying strategy to identify biomarkers (genes) (ENCODE Consortium, 2004) associated with diseases. Before the NGS era, disease genes were identified using a function-to-genetic approach, where first, a candidate gene was hypothesized to be associated with a particular disease based on its function, and validated on DNA for a particular number of patients. But now, the possibilities of high-throughput sequencing of genomic DNA allows us to directly evaluate which are the variants (or genes) recurrently identified within large cohorts of patients, and statistically associated with the disease (Mardis, 2008; van Dijk et al., 2014). The identification of the statistically significant correlation of a particular variant with a specific trait using genome information of large disease and control cohorts is the common and underlying principle behind all modern genomic-based studies in biomedicine. This approach has allowed the identification of genomic biomarkers at an unprecedented rate over the past years, setting up the basis for a personalized medicine, where the genomic profiles of patients will be considered for diagnosis, prognosis and treatment of disease.

### **1.2.1 Emerging limitations**

The drop in prices of high throughput sequencing and the increasing access to basic computing facilities has allowed, even to small and medium laboratories, to become data generators (Marx, 2013). This still increasing generation of

biomedical data (mostly from sequencing, but also from other data types) has also placed bottlenecks on different parts of the study, mostly at the analysis side. As the ambition and scope of current biomedical projects increase, the need of large computing infrastructures are becoming limiting factors and are actually driving, together with the access to the data, the organization of current biomedical data-centric research. Data security and privacy are particularly important when handling sensitive data, such as patient clinical and genomic information (Datta et al., 2016).

But these new bottlenecks and emerging challenges of data-centric strategies are not only found in the control of access, transfer, or management of data, but also, as mentioned, in the subsequent analysis of the data. The analysis of large datasets, not only requires large and powerful computing environments, but also a proper combination of algorithms and implementations that ensure an efficient processing of the data. For example, with such large data volumes, the scalability of a program is a crucial factor. This is the reason why bioinformaticians require close collaborations with computer scientists, in order to match the proper algorithm with an efficient implementation (Mattmann, 2013).

## **2. Cancer**

Cancer is currently one of the major research topics in biomedicine, due to the great burden that represents at medical and social level. In 2018, According to the Cancer Research UK ( <https://www.cancerresearchuk.org> ), 17 million new cases had been reported, and among them, 9.6 million of patients died during 2018. The incidence of cancer is heterogeneous around the world and depends on environmental factors (mutagens) and on the genetic background of each individual. This genetic background can determine the offset and the progression of the tumor, which in part explains the different incidence of different types of cancer within different populations and even between

genders. Economic development, social factors, and lifestyle are also factors involved in the incidence and treatment of cancer. In the case of men, for example, the most frequent tumor type is lung cancer, that presents the highest incidence of death, closely followed by liver and stomach cancer that present high mortality, and prostate and colorectal cancer, that present a severe incidence. For women, the most frequent cancer type is the one affecting breast, followed by lung cancer, as to mortality, and by colorectal cancer, as to the level of incidence (Bray et al., 2018). The rapid evolution of sequencing technologies, along with the increasing possibilities of genomic analysis, has changed the way tumors are nowadays classified and diagnosed, and is setting up the basis for a genomic and personalized oncology. The possibility of a deep genomic and molecular characterization of tumors are now allowing the gradual incorporation of more efficient and targeted treatment protocols, with the final aim of substituting traditional aggressive treatments based on chemotherapy and radiotherapy.

At molecular level, tumors emerge from a deregulation or malfunction of genes that are involved in the growth and death of the cell, usually through somatic alterations in its genome. In particular, the loss of function of tumor suppressor genes, and the gain of function for oncogenes can trigger the formation of a tumor, as an uncontrolled growth of cells (Zhang et al., 2018; Zia et al., 2012). Tumor suppressor genes, such as *TP53* (Varley et al., 1997), *PTEN* (Stambolic et al., 1998), *BRCA 1*, *BRCA 2* (Roy et al., 2011), are genes that regulate the cell during cell division and replication. If a mutation in a tumor suppressor gene results in a loss or reduction of its function, in combination with other genetic variations, this could lead to the cell growing abnormally and to cancer.

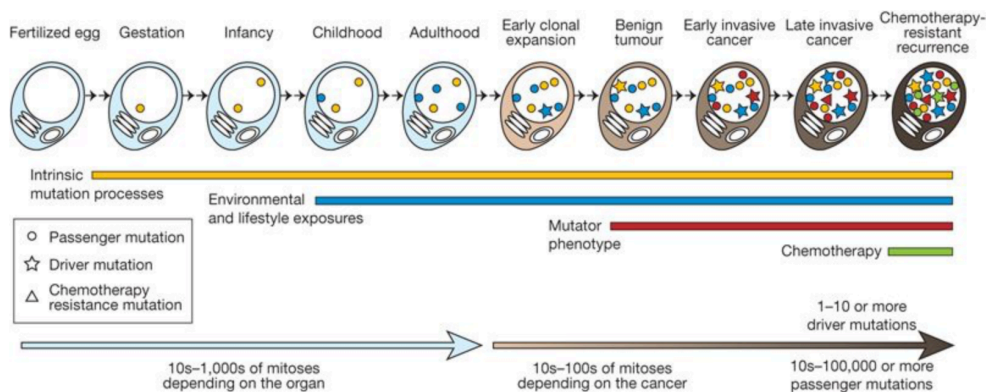
On the other hand, oncogenes represent the opposite side of cell growth control, where genes are involved in abnormal cell proliferation as a result of genetic alterations that either enhance gene expression or lead to uncontrolled

activity of the proteins encoded by the oncogene some examples are *RAS* (Lievre et al., 2008), *MYC* (Chen et al., 2018), and *ERK* (Koutsioumpa et al., 2018).

The number of somatic mutations among tumors varies according to the tissue and the molecular background of the cancer cells. This number usually ranges from 1.000 to 20.000 mutations, covering from single point mutations to large chromosomal rearrangements (Lawrence et al., 2013; Vogelstein et al., 2013). Several studies concluded that both endogenous and exogenous factors can contribute in different ways to the generation of somatic variation and the offset of tumors (Alexandrov et al., 2013; Alexandrov and Stratton, 2014; Martincorena and Campbell, 2015).

## **2.1 General causes of somatic mutations**

A large fraction of somatic genomic variation appears during DNA replication, and derives from errors during the cell division that are not repaired. Some forms of DNA alterations that can lead to somatic variants are caused by endogenous factors, such as reactive oxygen species, aldehydes, and by exogenous factors, such as chemicals (like those from tobacco smoking), ultraviolet (UV) light, and ionizing radiation (Figure 3). Other sources of somatic genomic variation involve the infection of viruses, as well as endogenous retrotransposition events, which can trigger chromosomal alteration and alteration of gene function (Talbot and Crawford, 2004). A well-known example is the human papillomavirus or Hepatitis B virus (HBV), which is involved in the origin of some types of cancer like Cervical Squamous Cell Carcinoma (CESC), Liver Hepatocellular carcinoma (LIHC), and Uterine Corpus Endometrial Carcinoma (UCEC).

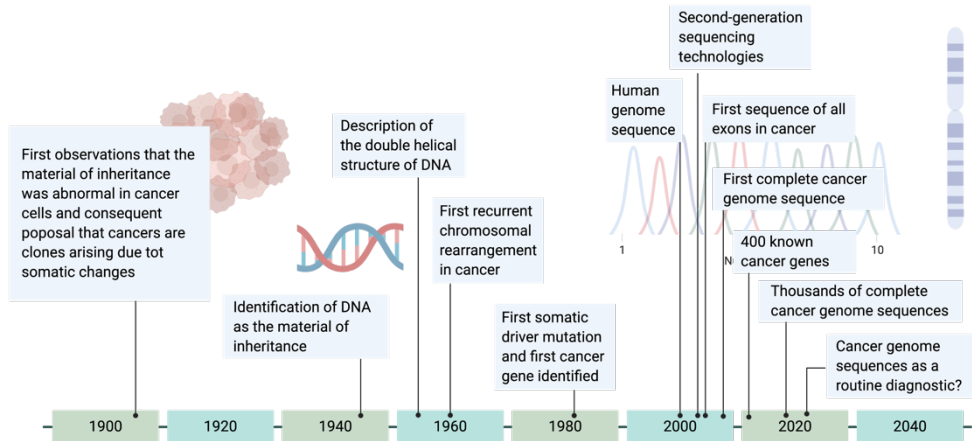


**Figure 3. The descendant line of the mitotic cell divisions of a fertilized egg represented in a single cell showing the processes that can contribute to acquiring somatic mutation at different stages.** Three different mutations are classified: passenger mutation (circles), driver mutation (stars), and Chemotherapy resistance mutation (triangle). Note that driver mutations tend to cause clonal expansion compared to the passenger mutations that do not affect the fitness of a clone but may be associated with clonal expansion. Another point to note is that in the field of relapse after chemotherapy, this phenomenon can be associated with resistance mutations before starting cancer therapy. Somatic mutations can be acquired during the normal cell lineage, due to cell division or by the effect of exogenous mutagens, or be generated by DNA repair defects during the development of cancer among other processes. Extracted from: (Stratton et al., 2009)

Over the last half century, the development of new technologies and analysis methodology has facilitated the systematic characterization and interpretation of cancer genomes at increasingly precise levels of resolution (Figure 4). Almost 30 years ago, the first cancer-related genetic mutation was discovered, a point mutation in the HRAS gene (Reddy et al., 1982), that changed a glycine to valine in codon 12. Although the functional impact of this mutation was originally not clear, many years of research have actually turned this gene as one of the "resistance marker" for the tumor response to anti-epidermal growth factor receptor (EGFR) therapies. This marker is used to determine the step to follow in targeting EGFR therapy in patients with colon or lung adenocarcinomas (Chin et al., 2011; Lievre et al., 2008) exemplifying the



potential and the importance of identifying and characterizing somatic variation in modern oncology.



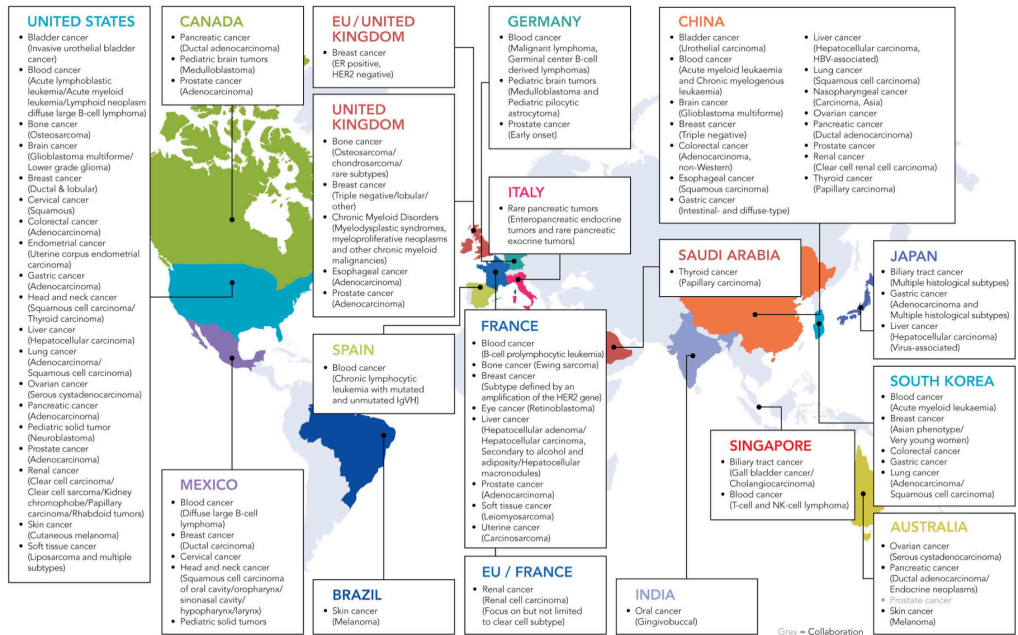
**Figure 4. Timeline showing the milestones in cancer genome research.** Based on (Stratton, 2011). Created with Biorender.com

There are different degrees of genome study, depending on the scope of the sequencing. Whole-genome sequencing (WGS) provides the sequence of the entire genome, in contrast to exome (only gene exons), or gene panels, which are the most frequent in hospitals and the way that genome sequencing is entering into current oncology protocols.

## 2.2 National and International initiatives

With the goal of understanding the role of genome variation in the biology of cancer, large efforts have taken place around the world. Some of these efforts in the form of large consortia. Among the most outstanding initiatives are the International Cancer Genome Consortium (ICGC) (<https://icgc.org>) (Figure 5) and The Cancer Genome Atlas (TCGA) (<https://www.cancer.gov/tcga>). These

two initiatives have pushed the field of cancer research, setting up the basis for current and future studies within the community. The first stages of these projects involved the identification of cancer driver genes, and the generation of general maps for somatic variation across different tumor types. The following phases involve the functional interpretation of these variants and the applicability to the clinics to achieve better and more personalized diagnosis and treatment protocols. In addition. Thanks to the expansion of these studies, and as a second step of this global characterization, samples were unified to gain statistical power and organized in PCAWG studies. For example, the TCGA pan cancer initiative, comprised the study of up to 11,000 tumor genomes, mostly exomes (Hoadley et al., 2018), the Pancancer Analysis of Whole Genomes (PCAWG) (<https://dcc.icgc.org/pcawg>) consortium was created with the goal of meta-analysing the genomic characteristics of the different types of tumors. This project, where our group has had a key role, has involved the collaboration of more than 1,300 researchers from 37 different countries, analyzing a total of more than 2,600 whole genomes, covering 38 different types of tumors . The results of the project were presented in February 2020, completing the most exhaustive study of the entire cancer genome to date. The results described in the different publications (Consortium, 2020; Cortes-Ciriano et al., 2020; Gerstung et al., 2020; Li et al., 2020), have helped to significantly improve the understanding of cancer and has provided new avenues for its diagnosis and treatment.



**Figure 5. Map representation of countries and tumor types involved in The International Cancer Genome Consortium ICGC.** Within the first part of the ICGC initiative, each one of the countries were committed to analyze and characterize the genomic variation associated to specific selected tumor types. Image extracted from: <https://icgc.org>

These initiatives have, not only played a major role in the technological revolution in genomics, but also in the area of international collaboration by generating a standard of norms to ensure that all data follows a quality criterion. This means that all the data generated presents the minimum overlap and redundancy, and thus, the overall value of the data increases. In addition, in cancer research, the strategies used by the consortium have become the standard format.

### 2.3 Analysis of somatic variation in cancer

The detection, classification and interpretation of somatic genomic variants has become an essential component in the study of cancer genomes. Practically, all the studies targeting tumor genomes follow a common strategy

or protocol, in which normal (normally from blood) and tumor genomes are sequenced from the same patient.

The first step of the analysis protocol (Figure 6) begins with the extraction of both normal and tumor samples from the same patient, whenever possible even from the same tissue, followed by the subsequent sequencing of the genome. This data is then analyzed to identify variations in the tumor sample, where the output is a potentially extensive list of somatic mutations. This list of variants is interpreted at the functional level to identify the genes that are affected and are part of the cancer biology. Subsequently, all those genes are analyzed at the molecular level to discover new drug targets and to design specific diagnosis and treatment protocols that will return to the patient.

The sequencing step can be targeted by different sequencing methods: WGS, Whole-exome sequencing (WES), or gene panels. Each has advantages and cons. Although panel tests and WGS offered similar diagnostic performance, WGS offered the benefit of reanalysis along the way to incorporate advances in knowledge. Until recently, only Multi-Gene Panel testing was used in clinical care, while WGS is already quite commonly applied in research. Even if substantial experience is needed for genomic interpretation of WGS (Cirino et al., 2017), it is expected that WGS is included in basic cancer analysis protocols soon.

Despite the technical and methodological challenges, these studies have also generated other more organizational adversities, as some steps of the analysis need to be conducted by different centres and communities. For example, the reception of the patient and the extraction of the corresponding samples happen in clinical environments, while the sequencing and analysis occur in sequencing and computing centres. More and more, the need of coordinating these efforts is driving the organization of large research environments, where sequencing and analysis technologies, are being developed close to clinical personalized medicine environments.

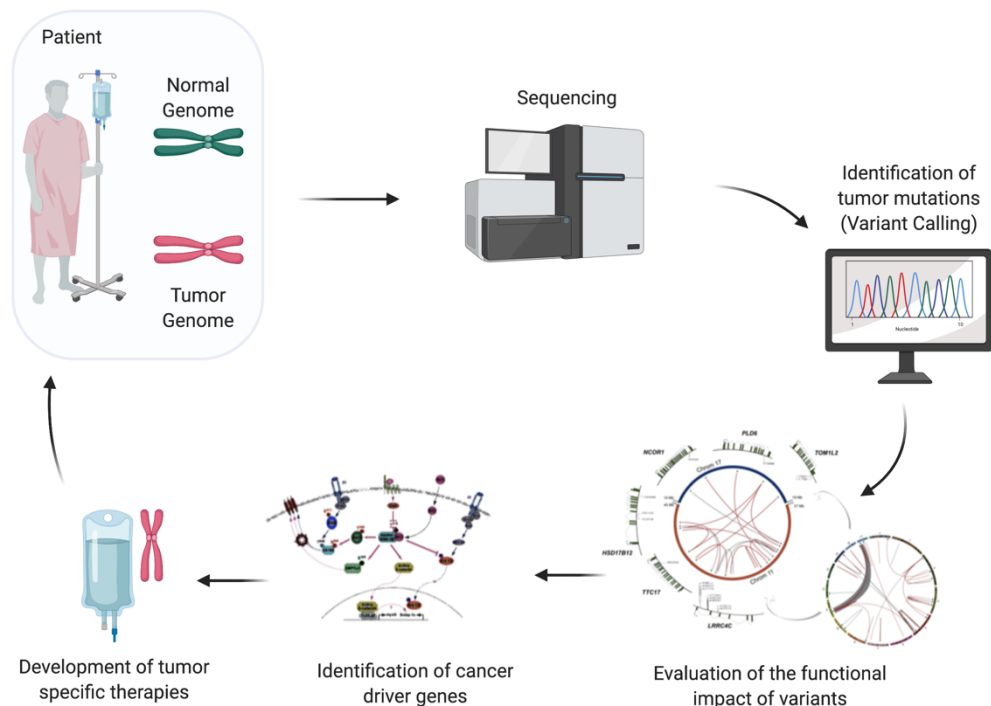


Figure 6. Common protocol for the identification of the genetic and molecular basis of tumorigenesis for the development of personalized therapies.

Throughout past years, several methods for variant detection have been developed with different scopes and capabilities. Most of the published methods are usually focused on the detection of a specific type of mutation, ranging from point mutations to structural variants (Chen et al., 2009; Cibulskis et al., 2013), while others are able to detect multiple classes of mutations in a single run (Rausch et al., 2012; Wala et al., 2018). These methods have, not only different detection scope, but also different levels of sensitivity (measures the proportion of actual positives that are correctly identified) and specificity (measures the proportion of true negatives that are correctly identified) (Guan and Sung, 2016), presenting their strengths and weaknesses

(Hurles et al., 2008) in the detection and the fact that it is the user who must decide which one best suits his needs. In the generation of pipelines to identify the full range of somatic mutation types in tumor samples, more than one program should be used to achieve the objective. Once the results of the different programs are obtained, filtering steps must be performed, which include information like: minimum coverage, mapping value, sequence quality, etc. for each of the mutations obtained. The selection of programs to generate this type of pipeline is a great challenge to achieve, both a good specificity and a good sensitivity for all kinds of variants. Notwithstanding, the VAF of somatic mutations in cancer usually has a higher value due to the selective advantage conferred by the mutations in cell proliferating. Accordingly, the vast majority of algorithms for specifics for cancer variant detection do not target low VAF values (Cibulskis et al., 2013). Therefore, the complete characterization of tumor genomes, as to their catalog of somatic variation, requires the development of complex and multimodular analysis pipelines gathering the results of different detection methods (Kosugi et al., 2019; Tattini et al., 2015), since no single method can detect each type of variant with high specificity and sensitivity.

### **2.3.1 The calling pipeline**

The underlying principle behind somatic variant identification programs relies on the search and detection of genomics changes present in the tumor sample, relative to the healthy one from the same patient. The vast majority of variant calling methods analyze all the tumoral and normal sequencing reads, aligned onto the reference genome. In contrast to that, a few methods, use alternative approaches based on the direct comparison of tumor and normal reads, and are therefore called reference-free methods.

### 2.3.1.1 Reference-based methods

Most of the available variant callers have been developed following the mapping based strategy, which is based on the accurate analysis of the reads aligned to the human reference genome, making this alignment step critical for the final sensitivity and specificity of the methods. In this direction, we can foresee some inherent limitations of reference-based methods. For example, this alignment process involves a high expenditure of resources and time. In addition, although there are several alternative alignment methods, like GEM (Marco-Sola et al., 2012), generally, this step is performed with the Burrows-Wheeler Aligner (BWA) program (Li and Durbin, 2009). The major difference between GEM and BWA is that GEM is five times faster than BWA execution in its default heuristic mode, giving a similar number of reads aligned. But the community uses BWA almost exclusively. This makes the alignment information the same for all, which means that all the analyses that need this information will share the same type of errors and therefore can be compared. At the same time, BWA provides a binary file (BAM file) containing information regarding the quality, structure and position of the read alignment, together with a list of all the reads that could not be mapped. The fact that BAMs are ready-to-be-used files, and that it conserves all the original read information, has made this file the current form of exchanging and storing genome sequencing data within databases.

But most importantly, the complex nature of the human genome represents a technical challenge for alignment accuracy. A human genome contains 3.2 billion bp, around 50-69% being repetitive sequences (de Koning et al., 2011), which includes transposable elements (i.e., LINES, SINES, and Long Terminal Repeats), low complexity regions (i.e., homopolymers), and pseudogenes. The complex nature of the human genomes presents significant challenges to achieve technical accuracy on alignments (Goldfeder et al., 2016). Larger insertions, deletions, and rearrangements within the genome are not represented in the reference genome and, therefore, adds additional

complexity to the alignment. Also, germinal variation does not only affect the accuracy of the alignment, as indeed, many of them are wrongly predicted as somatic mutations, increasing the number of false positives and lowering the specificity of the program. Finally, the impossibility to align reads with greater alignment complexity that cover large and complex genomic rearrangements (i.e. SVs), or that fall in polymorphic regions not included in the reference genome, will not be taken into account and will be disregarded as “unmapped reads”, which could indeed contain valid and important (Degner et al., 2009).

In addition to these limitations, it is worth mentioning that those methods specifically designed for the identification of large SVs, are usually also more inaccurate at the breakpoint of the variation. In these cases, the variation is located within a range of genome that is possible to align, becoming a restriction for further studies.

For the detection of point mutations or small indels, the alignment information is used within the read itself. In the case of SVs, the combination of information from mismatched reads is used (Figure 7; A1-A2). The four most common strategies are (Figure 7; B) (Guan and Sung, 2016):

1. Clustering (CL): All the discordant reads surrounding a region are grouped. Some of the callers employing this strategy are: VariationHunter(Hormozdiari et al., 2010), GASV (Cameron et al., 2019), and CLEVER.
2. Split-reads alignment (SA): is divided into two subcategories; (i) indirect case: align soft-clipped reads and one-end-anchored reads to locate the breakpoints that match. (ii) Direct case: refine the breakpoints identified by discordantly mapped reads. In the first subcategory we encounter the callers: CREST(Wang et al., 2011), ClipCrop(Suzuki et al., 2011), and Socrates (Schroder et al., 2014). In the second subcategory: Gustaf(Trappe et al., 2014), Prism(Jiang et al., 2012).

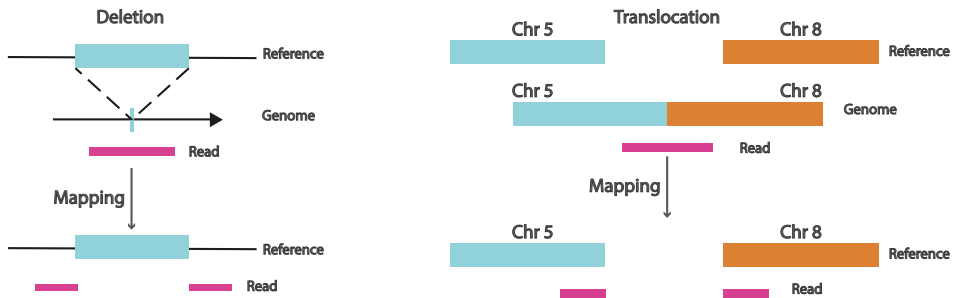


3. Contig Assembly (CA): anomalously mapped reads are de novo assembled to form longer consensus sequences (contigs) to identify the pairing breakpoints. Some of the callers using this strategy are TIGRA(Chen et al., 2014), and Cortex (Alekseyev and Pevzner, 2007).
4. Statistical testing (ST): use the local variations of reads depth, often used to detect copy-number variations. Breakdancer (Chen et al., 2009) is one of the variant callers that use this strategy.

Variant callers tend to use a combination of more than one of the above listed strategies for variant detection (Baker, 2012). A clear example is the variant caller DELLY (Rausch et al., 2012), that combines the detection and subsequent verification of mutations, using the information from discordant and one-end-anchored reads, and optionally from soft-clipped reads.

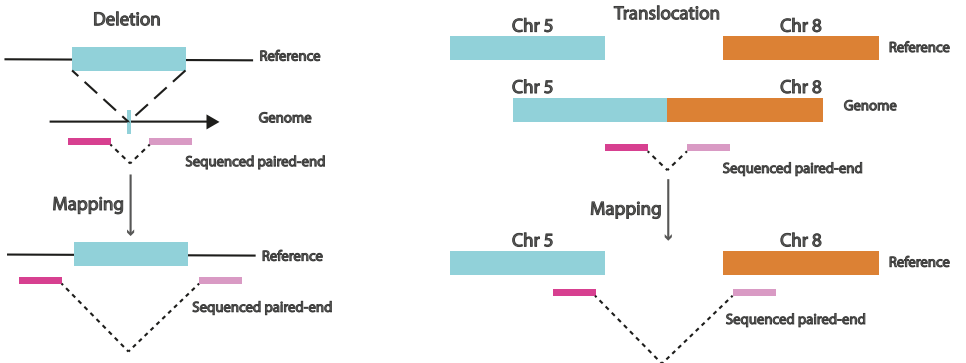
A1)

### Soft-clipped reads

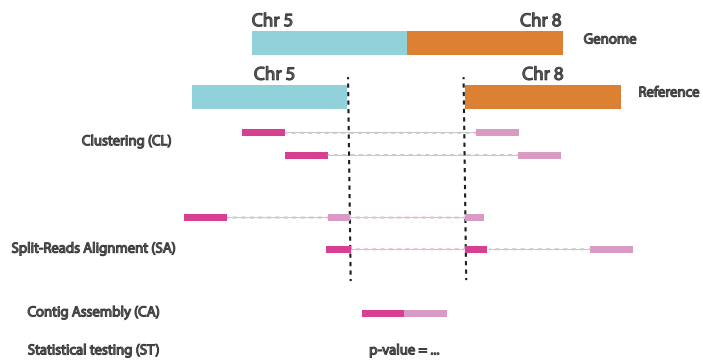


A2)

### Paired-end reads



B)



**Figure 7. A1 – A2) Types of mismatched reads used on SV calling in two different scenarios of variation: deletion and translocation.** A1) Soft-clipped reads; In Smith-Waterman alignment, the soft-clip readings are an unmatched fragment in a partially mapped read, within a sequence that is not aligned from the first residue to the last. Not to be confused with hard-clipping, they differ in that the subsequent clipping is not present in the alignment register. This clipped alignment is used to reconstruct those events that the read covers, as shown in the two scenarios: The read is mapped into two different fragments within the same chromosome with a separation between them due to the deletion caused. The read is mapped into two separate pieces of different chromosomes due to the translocation process. A2) Paired-end reads; The two reads of a paired-end are expected to be mapped into different strands of the same chromosome, and the distance between them will be consistent with the insertion size distribution. If SV callers detect pair reads mapped onto two different chromosomes, they will report a translocation or transposition event. If SV callers identify pair reads mapped with incorrect insertion sizes, they will indicate an insertion or deletion event; similarly for other types of SVs.

**B) SV calling techniques graphic representation of the methods:** Clustering (CL), Split-reads alignment (SA), Contig Assembly (CA), and Statistical testing (ST).

### 2.3.1.2 Reference free methods

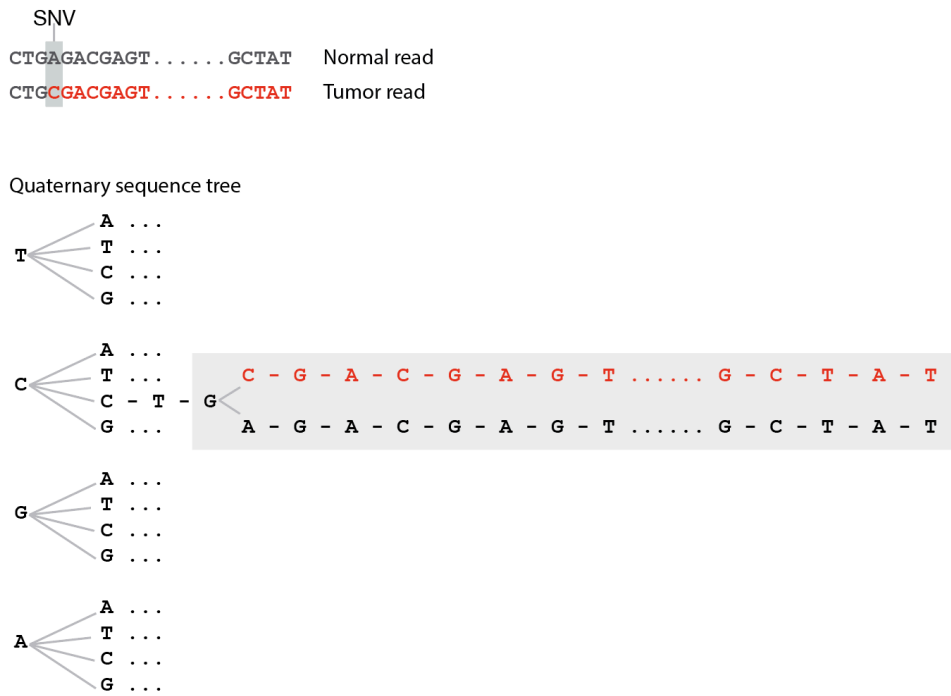
In order to overcome the above mentioned limitations of the mapping of reads that influence the rest of the analysis, a few alternative approximations have been developed using reference-free strategies. The following are some examples of these strategies: (i) the use of reference mapping combined with assembly-based methods (Chen et al., 2014); (ii) *de novo* assembly (Zhuang and Weng, 2015); and (iii) suffix tree approximations (Moncunill et al., 2014). While the first two strategies are based on the end-joining of reads in the tumor and normal genomes in order to identify discordant patterns, the latest is based on a suffix tree strategy, further developed in the next section.

### 2.3.1.2.1 Somatic Mutation Finder (SMuFin)

SMuFin is a reference free-method, generated within the group (Moncunill et al., 2014), that relies on the use of a quaternary sequence tree structures to compare directly tumor and normal reads and identify all types of genomic changes (except copy number changes) (Figure 8).

Some of the the advantages combined within SMuFiN for the detection of somatic variations include: (i) the direct comparison of normal and tumor readings without the need of using Binary Alignment Map (BAM) file alignment information; (ii) the single execution of the program to detect nearly all types of variants, including from SNV to SV, as well as inter- and intra-chromosomal translocations, inversions, insertions and deletions of any size; (iii) the detection of variants is reported at base pair resolution; and (iv) the accurate reconstruction of the region around variations in the tumor genome, including the sequence around the SVs breakpoints.

On the other hand, one of the limiting factors of using this approach for cancer genome analysis is the computational power that it requires. Quaternary sequence tree structures, also known as suffix-trees, are data structures that inherently demand blocking the access patterns to allow concurrent updates, thereby limiting the ability to implement these approaches efficiently in any high-performance parallel computing system for large scale analysis of genomes. This translates into a considerable memory requirement and makes it impossible to run the analysis of the genomes on any computer with a single node.



**Figure 8. Representation of quaternary sequence tree structure from the SMuFin algorithm representing an SNV.** The base containing the mutation and the bases following that generates a new branch on the quaternary sequence tree structure, are marked in red. We observe how from the SNV A --> C a new branch is generated in the tree structure. With this type of construction, it is possible to detect all those new branches that are candidates to contain a mutation. These branches will be later evaluated with different steps for its correct detection of variations.

**2.3.2 Variant caller virus**

Any research on cancer is incomplete without considering tumorigenic viruses. Several research groups are engaged in the search for therapeutic targets and novel vaccines to fight against these viruses (Sarid and Gao, 2011). The idea that viruses can cause cancer dates back more than a century (Javier and Butel, 2008). Nowadays, it has been unequivocally confirmed that several viruses are responsible for cancer in humans (Herrington et al., 2015; Moore and Chang, 2010). In fact, the World Health Organization (WHO) has estimated that 15.4% of all cancers are attributable to infections, 9.9% of which are linked

to viruses (Parkin, 2006; Plummer et al., 2016). The International Agency for Research on Cancer (IARC) classifies up to eleven pathogens as human carcinogens (Geisler et al., 2019). Moreover, it has been demonstrated that some viruses contribute to the biology of multi-step oncogenesis and are involved in many of the characteristics of cancer (Zapatka et al., 2020). To date, the four viruses that cause most of the infection-derived tumors, are the Human papillomavirus (HPV) (Munoz et al., 2006), that can cause cancer including anal, cervical, penile, throat, vaginal and vulvar; HBV (Bialecki and Di Bisceglie, 2005), is a leading cause of liver cancer; Hepatitis C virus (HCV) (Hermine et al., 2002) is a leading cause of liver cancer, and can cause non-Hodgkin's lymphoma; and Epstein-Barr virus (EBV) (Farrell, 2019), that can cause non-Hodgkin's lymphoma, and stomach cancer. Thanks to the appearance of the NGS, including the WGS and the RNA-seq, it has been possible to determine the position in which a virus is integrated within the tumor genomes (Duncavage et al., 2011). Accordingly, several analysis tools have been developed based on paired-end Illumina NGS data to tackle the detection of viruses in the tumor genomes, not only their presence but also their integration site: Capsid, VirusSeq(Chen et al., 2013), Virus- Finder (Gao et al., 2018), ViralFusionSeq(Li et al., 2013); VERSE (Wang et al., 2015), Virus-Clip (Ho et al., 2015) and Vy-PER(Forster et al., 2015) (Nguyen et al., 2018).

The strategy for the detection of the viruses behind each program varies but all of them are based on a standard scheme: the usage of alignment information of the reads that map in both genomes being analyzed, the human and the viral. For this reason, it is not only essential to pre-align the samples but to previously construct a new genome that contains the human reference genome and all the viral genomes to be identified. Despite the existence of these methods, the identification of viral copies remains a challenge, as normally viruses tend to integrate in repeat-rich genomic regions, and the sequencing reads covering internal parts of the virus are not considered and disregarded

as unmapped reads on the BAM file. This phenomenon gives rise to frequent false positives, and accordingly new methodologies have been created to eliminate false positive detection of virus integration events in next-generation sequencing data (Forster et al., 2015).

Accordingly, the study of the relationship of viruses and human cancer has opened new fronts for the development of novel strategies for preventing the infections that can evolve into carcinogenesis. In treatments like chemotherapy and radiation, the inability of the drugs to specifically target cancer cells instead of all types of cells, including healthy ones, and the toxicity that generates for the patients results in a significant drawback. Therefore, the new strategies are based on the presence of viral products in the tumor cells, as a target for guided therapies in which the tumor cells can be differentiated explicitly from normal ones. Therefore, the therapies that target the viral agent, generate immune responses to prevent infection, or kill infected or cancer cells, prove great potential due to their more effective and tolerable nature (Liao, 2006).

### **3. Somatic variation in non-disease scenarios**

After years in the shadows, recent studies changed the way of understanding somatic variations and their selection. The challenges associated with the study of somatic variation within healthy tissues, correspond to limitations in their detection, due to the high degree of tissue and cellular mosaicism in any targeted sample. It is now, when we can largely increase the sequencing coverage of genomes, when we can start studying the composition and potential role of somatic variation within physiological conditions. Therefore, proper sample collection protocols must be designed to ensure the success of this studies (Lupski et al., 2013).

#### **3.1 Somatic variation in Neurodevelopmental diseases**

The genetic variations involved in neurodevelopmental diseases were traditionally considered either to derive from the germline of one of the parents, or a *de novo* germline variation. Instead, more and more frequently a role for somatic variations in diseases other than cancer have been described, including neurodevelopmental diseases and other pathogenesis generated by *de novo* mutations that occur post-zygotically and, therefore, only targets a subset of the individual's cells.

During neurogenesis, where  $10^5$  neurons per minute are generated from an initial population of source cells (Workman et al., 2013), is when the human brain is most vulnerable to undergo somatic mutations. During the neurogenesis stage, neurons present a high mutation rate with about 5.1 point mutations per day (Bae et al., 2018), and some of these mutations may trigger neurological diseases. These diseases can be particularly sensitive to somatic mutations because even less than 10% of the cells carrying a mutation can affect phenotypes based on the distribution of these cells in the brain (Lee et al., 2012; Riviere et al., 2012). Furthermore, each neuron will continue accumulating somatic mutations linearly with age (Lodato et al., 2018), which could contribute to the development of neurodegenerative diseases (D'Gama and Walsh, 2018).

To understand the role of somatic variations in neurodevelopmental disease, two main factors must be taken into consideration: (i) the temporal moment and the progenitor cell in which the somatic mutation has appeared, and (ii) the effect that the mutation can produce in the cell (e.g., if the mutation is very harmful, the cell is selected against it and the mutation will not lead to disease) (D'Gama and Walsh, 2018). Numerous studies have been carried out around neurological development disorders with visible focal lesions generated by somatic variations, such as Focal cortical dysplasia (FCD) and hemimegalencephaly (HME) (Blumcke et al., 2011; Poduri et al., 2012). However, it is important to highlight that the same mutations have been also



studied in relation to diseases that do not present such visible lesions as the previous ones. This group includes cases of intellectual disability and autism spectrum disorder and epileptic encephalopathies (Lee et al., 2012; Poduri et al., 2012) and a wide range of neuropsychiatric diseases.

### **3.1.1 Neurodevelopmental genome analysis in the era of NGS**

Thanks to the appearance of NGS and single-sequencing techniques, the role of somatic variations in the development of the human brain has been understood in a more refined and extended way. These new technologies have allowed the scientific community to approach different hypotheses formulated thanks to the facilitation of a systematic analysis of all types of somatic mutations in both healthy and affected tissues.

The detection of these variations presents more difficulties than in any of the previously mentioned cancer scenarios, due to the low VAF expected for non-cancer somatic variation and the subsequent difficulty in detection. Therefore, in order to make a proper detection, the coverage of the sample being analyzed must be sufficiently high to ensure that the mutations are well represented (Jamuar et al., 2014; Lim et al., 2015) and are not considered sequencing errors during the analysis, as it usually happens with low VAF variants (D’Gama and Walsh, 2018).

### **3.1.2 Somatic mosaicism in the normal human brain**

During the study of somatic variations and their role in neurological diseases, it has always been of particular interest to understand whether these variations may actually play an important role in the development and physiology of brain cells. Recently, several studies that glimpse the role that somatic variations have in the generation of neuronal diversity have been published. In 2010 Muotri and colleagues (Muotri et al., 2010), demonstrated the impact of L1 insertions, initially considered as “junk DNA” in the human brain, since

these represent approximately 25% of our genome. The rate of insertion of retrotransposons can cause the inactivation of genes or the change of the expression. Although the rate of these is still a subject of debate, the impact they have on the development of the human brain remains an essential area of study. It is in this area, of the role that somatic variations play in neurodevelopment, that I focus the study of the second part of the thesis presented here.

### **3.1.3 Somatic activity on brain development**

During embryonic development, the brain undergoes rapid and sustained cell proliferation originating at the rostral end of the fetal neural tube. Brain cortical development is accomplished via a highly regulated sequence of neuroprogenitor cell division, migration, and differentiation. This neuroprogenitor cells divide asymmetrically, generating a progenitor stem cell and a neuron that migrates from the ventricles along the radial glia pathways to form a six-layered lamellar neocortex. Through this process, the neurons undergo DNA damage. This damage has been observed to reach a maximum between E11-E14.5 during development in mice and is mainly observed in postmitotic premigratory populations of the developing nervous system. It has been suggested that this DNA damage plays an influential role in the subsequent massive apoptotic event during development, an essential process for the elimination of overproduced neurons. Numerous hypotheses have been proposed to unravel the causes of somatic DNA damages during brain development (e.g., replication stress), nevertheless, it has not been established yet. Accordingly, in the present work we hypothesize that PGBD5, a transposase-like protein that presents nuclease activity, can produce double-stranded DNA breaks in neurons, contributing to the generation of somatic DNA changes, enabling the survival of the mutated cells during the subsequent apoptotic selection. In a recent study where our group contributed, it has been shown that an active nuclease *PiggyBac Transposable Element Derived 5*

*(PGBD5)* has the ability to generate somatic mutations in human cancer cells (Henssen et al., 2017a). PGBD5 expression is presently high and almost confined to the brain area, specifically the neurons of the cortex, hippocampus, and cerebellum. This fact raises an interesting and long-standing question about the somatic DNA rearrangements in brain cells. However, the function of PGBD5 remains elusive.

## 4. Final considerations

In brief, with the emergence of new sequencing technologies the study of somatic variants has experienced huge progress. In the area of cancer, it has led to the formation of research projects with a large number of samples to give a better picture of the role of mutations in the biology of tumor. These advances have come with numerous challenges; at the computational level to be able to analyze all those data, and at the level of analysis to formulate new methods to be able to make a more accurate detection of these mutations.

Albeit it is true that research associated with somatic variants in healthy tissues is increasingly abundant, the detection methods used present limitations and further studies are needed. The present thesis is aimed at overcoming part of these limitations, focusing first in the development of a reliable new algorithm for the detection of somatic variations with special emphasis in the scalability and implementation of the method for large dataset analysis, and second in the detection of somatic variants in healthy tissue directly related to neuronal development.



# OBJECTIVES

## **The main goals of this thesis are:**

- I. To design and implement a reference-free and scalable algorithm for the identification of somatic variants.
  
- II. To identify somatic DNA rearrangements induced by *Pgbd5* (the mouse ortholog of human *PGBD5*) during brain development and adult state, that might enhance the survival of the mutated cell to the subsequent apoptotic selection.

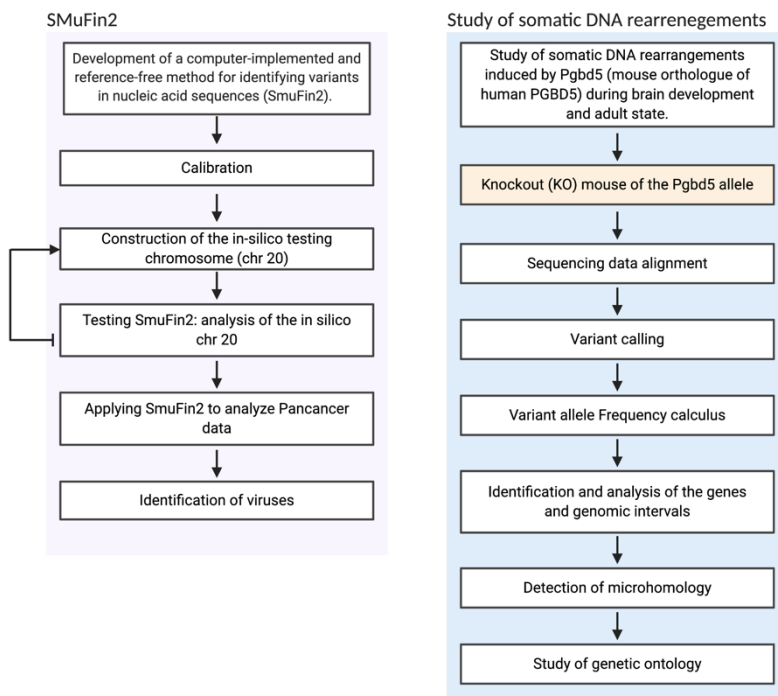
METHODS



The method section has been divided into two major blocks (Figure 9).

The first block corresponds to all the resources and methodology for the development of SMuFin2.

The second main block corresponds to the study of somatic DNA rearrangements induced by Pgbd5 (mouse orthologue of human PGBD5) during brain development and adult state.



**Figure 9. General workflow for the methods section.** The white boxes correspond to the work carried out in the center; the yellow boxes correspond to those carried out in external centers. created with Biorender.com

# 1. Development of a computer-implemented and reference-free based strategy for identifying variants in nucleic acid sequences

This strategy was conceived as a current reference-free algorithm-implementation and a redesign of the original SMuFin (Moncunill et al., 2014), a reference-free caller based on a suffix-tree strategy to identify somatic mutations on tumor genomes developed on Dr. Torrents group.

SMuFin2 is a reference-free detection strategy based on polymorphic k-mer strategy that allows both discovery of homozygous and heterozygous variation in genomes. That strategy also granted us the identification of most types of sequence genomes variation, from single nucleotide substitutions to large structural variants in a single run.

Polymorphic k-mer strategy is based on the sequential sub-selection of read regions, with a defined k-mer size, that will be compared to rely on the regions that contain variations. The core of the strategy relies on k-mers being managed as “dynamic entities”, this means that the k-mers suffer variations that make it possible for us to compare between two samples, and thus be able to hunt in a first pass the regions that are susceptible to contain a possible mutation.

In this way, we generate a reduced group of candidate reads, easier to treat, and analyze since all the reads without information related to any mutation have been eliminated.

## 1.1 Polymorphic k-mer strategy

SMuFin2 is based on a polymorphic k-mer strategy.

For a full understanding, we detail below the main terms, structures, we have been used to develop the strategy for the detection of variants.

### 1.1.1 Important terms

#### K-mer

In bioinformatics k-mers are all possible subsequences of length k contained within a biological sequence which have a length M. The total number of k-mers in a sequence of length M is  $M-k+1$ .

```

01234567890123456789012345678901234567890123456789012345678901234567890
GAAAACTAAGCTGAATTAGAAAGGAATAATGCTCATCGCA
GAAAACTAAGCTGAATTAGAAAGGAATAAT
AAAACTAAGCTGAATTAGAAAGGAATAATG
AAACTAAGCTGAATTAGAAAGGAATAATGC
AATAAGCTGAATTAGAAAGGAATAATGCT
ACTAAAGCTGAATTAGAAAGGAATAATGCTC
CTAAGCTGAATTAGAAAGGAATAATGCTCA
TAAGCTGAATTAGAAAGGAATAATGCTCAT
AAGCTGAATTAGAAAGGAATAATGCTCATC
AGCTGAATTAGAAAGGAATAATGCTCATCG
GCTGAATTAGAAAGGAATAATGCTCATCGC
CTGAATTAGAAAGGAATAATGCTCATCGCA

```

Read length = 40n  
k-mer length = 30n

#### Stem

The stem is a fragment of a k-mer. It can be a k-mer without a prefix, a k-mer without a suffix, a k-mer without an infinitive, or any combination of previous states.

Example:

K-mer with 30 nucleotides. To mark the base that was deleted from the original k-mer, we will use the character "-".

K-mer ID n°1  
 AAAACTAAGCTGAATTAGAAAGGAATAATG

K-mer ID n°1 without a prefix of length 1  
 -AAACTAAGCTGAATTAGAAAGGAATAATG

K-mer ID n°1 without a suffix of length 1  
 AAAACTAAGCTGAATTAGAAAGGAATAAT-

K-mer ID n°1 without suffix and prefix both of length 1  
 -AAACTAAGCTGAATTAGAAAGGAATAAT-

K-mer ID n°1 without an infix in position 3 of length 2  
 AA--CTAAGCTGAATTAGAAAGGAATAATG

## Inflection

The inflection in the method refers to all the possible fragments resulting from completing a k-mer stem, taking into account that we work with four different bases that are the nucleotides: A C T G.

If a stem only differs from the original k-mer in one base, we can obtain up to 4 different inflections, one of which will be identical to the original k-mer.

If instead of a single base, there were two bases, the total number of inflections would be 16 ( $4^2$ ).

E.g.

**-AAACTAAGCTGAATTAGAAAGGAATAATG**

We generated all its inflections:

AAACTAAGCTGAATTAGAAAGGAATAATA  
 AAAACTAAGCTGAATTAGAAAGGAATAATC  
 AAAACTAAGCTGAATTAGAAAGGAATAAAT  
 AAAACTAAGCTGAATTAGAAAGGAATAATG  
 CAAACTAAGCTGAATTAGAAAGGAATAATA  
 CAAACTAAGCTGAATTAGAAAGGAATAATC  
 CAAACTAAGCTGAATTAGAAAGGAATAAAT  
 CAAACTAAGCTGAATTAGAAAGGAATAATG  
 TAAACTAAGCTGAATTAGAAAGGAATAATA  
 TAAACTAAGCTGAATTAGAAAGGAATAATC  
 TAAACTAAGCTGAATTAGAAAGGAATAAAT  
 TAAACTAAGCTGAATTAGAAAGGAATAATG  
 GAAACTAAGCTGAATTAGAAAGGAATAATA  
 GAAACTAAGCTGAATTAGAAAGGAATAATC  
 GAAACTAAGCTGAATTAGAAAGGAATAAAT  
 GAAACTAAGCTGAATTAGAAAGGAATAATG

We behaved as the fourth inflection corresponds to k-mer Id n<sup>o</sup>1

### Partial inflection

The partial inflection is when we have a stem with a minimum of two positions removed from the original k-mer, and at least one of them is not extended to generate its inflections.

Using the stem generated in the previous section that had neither prefix nor starting suffix one.

`-AAACTAAGCTGAATTAGAAAGGAATAAT-`

We generate all the prefix inflections. To mark the position that has not been extended from the stem we will use the character ".".

```
·AAACTAAGCTGAATTAGAAAGGAATAATA  
·AAACTAAGCTGAATTAGAAAGGAATAATC  
·AAACTAAGCTGAATTAGAAAGGAATAATT  
·AAACTAAGCTGAATTAGAAAGGAATAATG
```

### Polymorphic k-mer

We refer with this term to that k-mer that from the stem of it can identify the totality of its inflections, as its partial inflections.

With this strategy, we use the k-mers to see all their variations through their inflections and to be able to detect with them those that may be related to a variation in the genome and to be able to catch all the information around it.

## **1.2 Calibrate algorithm**

In order to measure and calibrate the detection capabilities of the method algorithm, we executed it on a controlled system, consisting of modified sequences of chromosome 20. For testing purposes we selected a small chromosome (62.435.965 bp) that can be handled well on the calibration. We

use the same mutations as in the SMuFin method in order to make a direct comparison with it.

For each step of the method, different scenarios have been made to check its effectiveness. All the data has been extracted within intermediate house scripts in python only included on the developers' code.

### 1.2.1 Construction of the in-silico chromosome 20

A personalized chromosome 20 has been extracted from the hg19 reference genome downloaded from UCSC (with no repeat-masking) (<http://www.ucsc.edu>) and modified to match a randomly chosen human haplotype. Personalized chromosome 20 contains 148,639 variants consisting of 96,935 SNPs and 51,704 deletions. The catalog of somatic variants further added to this personalized chromosome and constituting the target of the invention, was composed of: 168 SNVs, 26 Indels, 20 SVs and 1 viral insertion of KI polyomavirus (extracted from: <http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?lvl=o&id=4234>),

ART Illumina (Altschul et al., 1990) has been used to in-silico sequencing, sequencing parameters like variation or read length has been extracted from Moo4 sample of mantle cell lymphoma (MCL) (Bea et al., 2013).

## 1.3 Analysis of the in silico chromosome 20 with the strategy of the invention

Using an internal pipeline, we extracted all the candidate blocks to contain a variant to calibrate the first block from SMuFin2.

Further details about the analysis of in-silico chromosome 20 can be found in the configuration file from this execution (Results chapter section 1.3.1).

## 1.4 PCAWG data

PCAWG BAMs files were obtained from The International Cancer Genome Consortium (ICGC)/The Cancer Genome Atlas (TCGA) Pan-Cancer Analysis of Whole Genomes (PCAWG) project, accessed through the ICGC data portal's Data Repository tool (<https://dcc.icgc.org>).

The PCAWG project enabled the study and characterization of the pattern of mutations of more than 2,700 cancer donors and 20 primary tumor sites.

The dataset is constituted by a total of 5,789 whole genomes of tumors and matched healthy tissue encompassing 39 tumor types. The tumor/normal pairs came from a total of 2,834 donors collected and sequenced by 48 sequencing projects across 14 jurisdictions from the International Cancer Genome Consortium.

Biorxiv preprint (Large-Scale Uniform Analysis of Cancer Whole Genomes in Multiple Computing Environments) (<https://www.biorxiv.org/content/10.1101/161638v1>) describes the generation of data and the phases of the uniform analysis of whole genomes where genomes are involved.

For the selection of samples, we considered all those of known tumor-associated viruses such as EBV, HBV, and several HPV types.

The selected studies where the samples belong and with which we carry out the tests are:

DCC Project Code ; Project Name ; Country

CESC-US ; Cervical Squamous Cell Carcinoma - TCGA, US ; US

LIHC-US ; Liver Hepatocellular carcinoma - TCGA, US ; US

UCEC-US ; Uterine Corpus Endometrial Carcinoma- TCGA, US ; US

The last access to all the data storage was in September 2018.

### **1.4.1 Running SMuFin2 to analyze PCAWG data**

Using an internal pipeline from SMuFin2, we extracted all the candidate blocks to contain a variant to select all those ones that could be involved in the integration of a virus on the tumoral sample.

Further details about the analysis of PCAWG data can be found in the configuration file from this execution in the Results chapter 1.4 Identification of tumor-associated viruses.

As a large part of this thesis was focused on the development of the algorithm of SMuFin2, detailed information about the algorithm can be found in the Results chapter 1.1 SMuFin2 Algorithm.

### **1.4.2 Identification of viruses presence on sample**

To identify the presence of viruses within the samples, a program of alignment with those sequences filtered by the method described in results, was used against a virus genome database.

The selected method was the command line version 2.6.0 of Basic Local Alignment Search Tool ; BLAST (Altschul et al., 1990).

The virus base that was used was downloaded from <https://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi> .

The last access to all the data storage was in May 2018.



## 2. Landscape of somatic variation in neural development and the role of *Pgbd5*

### 2.1 Experimental outline

To provide experimental context of the data used in this analysis Dr. Luz Jubierre Zapater from the group of Dr. Alex Kentsis, at Memorial Sloan Kettering, produced the knockout (KO) mouse model of the *Pgbd5* allele.

In summary they generated:

*Pgbd5* -/wild type (wt) females were crossed to *Pgbd5* -/wt males to obtain *Pgbd5* wt/wt and *Pgbd5* -/- (or KO) littermates.

For the identification of neuron-specific *Pgbd5* somatic DNA rearrangements in *Pgbd5*-wt and *Pgbd5*-KO mice models, was used Illumina high-coverage (80x) PCR-free paired-end genome-wide sequencing.

#### Adult brains

In the case of the detection of induced rearrangements by *Pgbd5* in adult mice, 3 *Pgbd5* wt and 3 *Pgbd5* KO littermates of 30 days of age were used. Just before the euthanasia, peripheral blood mononuclear cells (PBMC) was collected as a control for the experiment. As a case sample, three different neural tissues were collected: Olfactory bulb, Hippocampus, and Cerebellum. They extracted DNA using an Invitrogen DNA extraction kit (K1820-02) and quantified it using TapeStation Bioanalyzer. Genomics Core at Memorial Sloan Kettering made the library preparation and the Illumina sequencing.

## Embryo developing brain

In the case of the detection of induced rearrangements by *Pgbd5* in the developing brain, were used 3 *Pgbd5* wt and 3 *Pgbd5* KO E14 (14 days post-coitum) embryos from the same pregnancy. Just before the euthanasia, the embryos were extracted from the mother and spleen was collected as a control for the experiment.

As a case sample, three different neural tissues were collected: the Forebrain (the part that will give rise to the cortex among other structures), Midbrain (this part will give rise to the midbrain), and Hindbrain (this part give rise to the cerebellum and spine bulb). As adult brain samples, They extracted DNA using an Invitrogen DNA extraction kit (K1820-02) and quantified it using TapeStation Bioanalyzer. Genomics Core at Memorial Sloan Kettering made the library preparation and the Illumina sequencing.

## **2.2 Analysis of sequenced data**

### **2.2.1 Sequenced data alignment**

Once the data was generated we started by aligning the sequenced data to the mouse reference genome (GRCm38/mm10) downloaded from (<https://genome.ucsc.edu>) using Burrows-Wheeler Alignment (BWA) MEM algorithm (Li and Durbin, 2009). To improve the coverage for the detection, all the FASTQs corresponding to the same sample were merged in a single BAM file. We used bammarkduplicates to mark the duplicated reads. For the alignment summary metrics, we used Alfred v0.1.16 (Rausch et al., 2019). The last access to all the data storage was in January 2019.

### **2.2.2 Variant calling**

To perform the detection of the rearrangements produced by *Pgbd5*, we run three different variant callers: Pindel (Raine et al., 2015)(version 2.2.3) (Ye et al., 2009), and Delly (Rausch et al., 2012), detecting indels and structural variants,

and GATK (version 3.7) (McKenna et al., 2010) focusing on SNVs. In order to run the programs, we use the following reference files: the mouse reference genome (GRCm38/mm10), the simple repeats file from mouse, and the coding exons file from (GRCm38/mm10) downloaded from (<https://genome.ucsc.edu>).

The last access to all the data storage on (<https://genome.ucsc.edu>) was in January 2019

Each of the methods for the detection was run on pooled libraries (normal and tumor) using default settings except for the following parameters:

#### Delly2

Predictions obtained with Delly2 were considered with the following parameters: -c 0.05, -a 0.05 and -m 15.

#### **2.2.2.1 Joining and filtering of variant calling results**

Once we obtained the results, for each sample from each variant caller, we join the results from the different callers in order to increase our sensitivity and we filtered out the duplicates within and between callers, to avoid redundancy, considering a similarity window of 300bp. In case a mutation was found to be duplicated, we kept the one with the highest detection quality, VAF (Variant allele Frequency), or ultimately, we give more weight to the deletions. In the final step, in order to maintain specificity, we filter for those that had the default PASS quality filter for further analysis.

#### **2.2.3 Variant allele Frequency calculus**

VAF is the relative frequency of a variant at a particular locus, expressed as a percentage or fraction.

To calculate VAF, we divided the number of reads with the presence of the variant by the total number of reads of all the alleles.

$$f_i = \frac{r_{mut,i}}{r_{mut,i} + r_{ref,i}}.$$

Formula 1. Where  $r_{mut,i}$  are the reads containing the variant, and  $r_{ref,i}$  the reads as reference allele (normal) for the mutation  $i$ .

#### 2.2.4 Identification and analysis of genes

For the detection of genes affected by mutations we used the BEDTOOLS package (Quinlan, 2014) and the annotation of NCBI genes for mouse downloaded from (<https://genome.ucsc.edu>).

To study the effect of the different subsets of mutations we use the ENSEMBL Variant Effect Predictor (<https://www.ensembl.org/Tools/VEP>) (VEP) (McLaren et al., 2016). with default settings except for the following parameters:

Species: Mouse (mus musculus ; GRCm38.p6)

Transcript database: RefSeq transcripts

Filtering by the most severe consequence per variant. In the case of obtaining more than one result, we kept the one that was more deleterious.

#### 2.2.5 Identification and analysis of genomic intervals

For the detection of genomic intervals, the regions have been created dynamically through the list of mutations contained in each group, with a static window size of 3Mb. For each mutation entry we had in the file, we generated the window, and we observed how many mutations those windows covered.

We filter through those windows that contain a minimum of two mutations, and we remove those windows composed of subgroups of mutations that come from larger windows. This results in a single list for each group selected.

For the study of overlaps, we crossed the files of each group with the option Intersect from bedtools (Quinlan, 2014).

### 2.2.6 Detection of motifs

For the discovery and analysis of motifs around the breakpoints of the mutations that were affecting genes or highly mutated regions: (i) We reconstructed the sequence around both breakpoints of the selected deletions with a length of 20bp around each breakpoint, (ii) we executed the meme suite tool, specifically the tool MEME to discover possible motifs in each of the different subgroups (Bailey et al., 2009; Bailey and Elkan, 1994). With the following parameters: maximum number of motifs: 25, minimum width: 4bp, maximum width: 12bp

### 2.2.7 Study of genetic ontology

We perform the gene ontology analysis using the online tool: The Database for Annotation, Visualization and Integrated Discovery; DAVID (version 6.8) (<https://david.ncifcrf.gov>) (Huang da et al., 2009a, b) (Huang da et al., 2009b) in the set of genes that were uniquely associated (mutated) to each of the different subgroups of samples.  $P < 0.05$ , the threshold level for all gene ontology, was considered statistically significant.

RESULTS

Following the hierarchy of topics in the method section, the results have been split into two blocks; the first one is related to the development of SMuFin2, an standalone based strategy for the reference-free identification of somatic genomic variation, and the second is focused on the study of somatic DNA rearrangements induced by Pgbd5 during brain development and adult state.

# 1. Design of the algorithm of the Somatic Mutation Finder, version 2 (SMuFin2)

SMuFin2 is an algorithm designed for the identification and classification of somatic variation in cancer, and other normal-case genome pairs. SMuFin2 represents the second version of the original SMuFin program published in 2014, and was planned and designed in response to the limitations of the first version, mostly concerning the scalability and computing efficiency, limited by its underlying suffix-tree data structure. For the design of SMuFin2, we focused on computing efficiency and scalability, keeping the original qualities and capabilities of SMuFin: high sensibility and specificity, reference-free detection, detection of all variation types in a single execution, and base-pair resolution. From a close collaboration with Jordà Polo, from the David Carrera's group at the Computer Science Department at the BSC, we have ensured the combination of an efficient algorithm with a proper implementation and hardware integration. My specific activity has been centered in the design of that algorithm.

Following and adapting to the different computing needs across the general analysis of genomes, we have divided the algorithm into two blocks:

1. A first block that processes all raw data, and therefore is computationally (I/O) more intensive. This part starts by reading the raw genomic sequence data (thousands of millions of reads), to finally provide small sequence blocks that are candidates to contain a mutation. Due to the high computational requirements generated mainly by the genome lectures, this part was done in collaboration with David Carrera's group.



2. The second block consists of the detection and classification of the different variants within these blocks, and the subsequent alignment to the genome to provide their exact genomic coordinates. This step also includes the detection of the presence and insertion of non-human genetic material in the tumor sample

As the first block is suitable to be used for other types of sequence analysis, like for transcriptomics, it was decided to describe and protect this part through the submission of a patent (PATENT: A computer-implemented and reference-free method for identifying variants in nucleic acid sequences. NUM: WO 2018/007034). The patent was accepted by the European Patent Office and was published on Espacenet and Google patents. We are currently in the process of applying for the US patent.

The positive overall performance results of this first block, makes it potentially useful for the design of large scale infrastructures for genome analysis, in relation to the expected demand coming from Personalized Medicine initiatives.

Currently, the entire program and functionality of SmuFin2 is still not complete, as only the first block is finished and frozen. The second part is still under progress (see below).

## **1.1 SMuFin2 Algorithm**

The new algorithm is based on the direct comparison of genomic sequences coming from two genomes, normally tumor and normal from the same individual, in the case of cancer, to finally identify all the changes corresponding to somatic variation occurring in one of them. As explained below in more detail, this direct comparison is done by converting all the read sequences into k-mers, which are then scanned, searching for differences between the two genomes. As for the first version of SMuFin that used suffix-

tree organization of the data to identify tumor reads that had no counterpart in the normal, and therefore could potentially point to a somatic mutation, SMuFin2 uses the k-mer approximation with the same aim. We directly compare all the reads of tumor and normal sequences to identify candidate regions having a variation. From these reads, we then reconstruct, in the form of aligned sequence blocks, a specific candidate region of the genome with the corresponding reads of both normal and tumor genomic sequences that should contain the variation. In the second part of the algorithm, we analyze these sequence blocks to identify and classify the variation, to finally map it onto the reference genome to provide the type of variation and the exact genomic coordinates. In order to clarify the description of the algorithm, we have defined different steps, represented in Figure 10 with section reference from results chapter on each step.

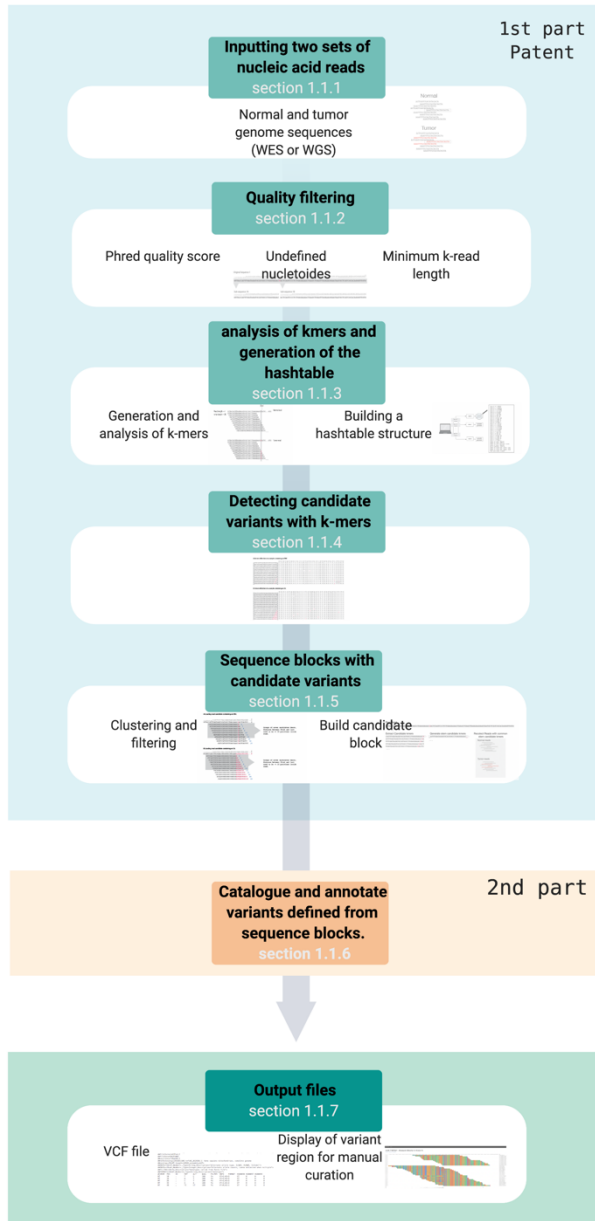


Figure 10. Schematic representation of SMuFin2's Algorithm with chapter section from results block for each step. SMuFin2 is divided into two blocks according computing needs across the general analysis of genomes, the first block was patented. SMuFin2's algorithm starts with sequenced data (FASTQ) or aligned data (BAM), finishing with a vcf file and a website of somatic variants detected on tumoral sample. created with Biorender.com

### 1.1.1 Inputting two sets of nucleic acid reads

In the first (1) step, SMuFin2 processes the input sequence files to start building a hashtable of k-mers. As input, the method accepts FASTQ files (Cock et al., 2010) and BAM files (Li et al., 2009), which contain all the reads with mapping coordinates onto the reference genome. This is important, as many databases and datasets of sequences are only in BAM format, from which one can easily extract all the reads and reconstruct the original FASTQ. For each read, we also extract the corresponding sequence quality score and the sequence identifier.

### 1.1.2 Quality filtering of the raw sequenced data

Preliminary, to start processing the reads, we perform the filtering of low quality and potentially erroneous reads. This is done, at different levels:

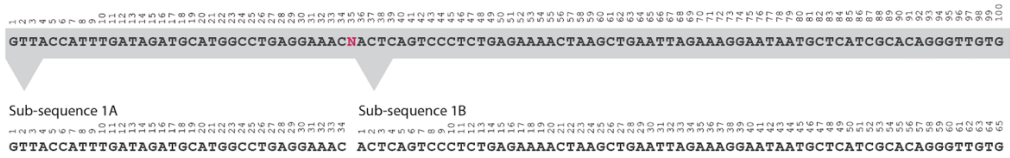
#### Phred quality score

A sequence is included if the input contains a minimum of bases with a Phred quality score (Ewing and Green, 1998; Ewing et al., 1998) greater than 20 (Q20), this means that the base call accuracy is 99%. For example, If this value is set to 80, it means that all those reads with more than 20% of their bases with a Phred quality lower than 20 will be eliminated. Therefore, we only keep those that at least 80% of their bases are of higher quality.

#### Undefined nucleotides

If the sequence contains undefined nucleotides, represented by "N", these will be eliminated, generating independent sub-sequences from the read at both sides of the N (Figure 11).

Original Sequence 1



**Figure 11. Sub-sequences generated from an original read containing an undefined nucleotide “N”.** Original Sequence 1 with a length of 100 nucleotides and 1 undefined nucleotide generates two sub-sequences: 1A with a length of 34 nucleotides, and 1B with a length of 65 nucleotides.

### Minimum read k-length

We discarded all the sequences whose length does not cover the k-mer size established to make the analysis, due to the fact that k-mers could not be generated for the detection of variants. Hence, we eliminate the sequences that come from the raw data as well as those resulting from the formation of sub-sequences explained in the previous section.

### 1.1.3 Generating a hash table structure

After the quality filtering step, the algorithm next, (3) generates a hashtable with all the k-mers from both samples. This step is divided into two major processes: (1) generate the k-mers and (2) build a hash table structure.

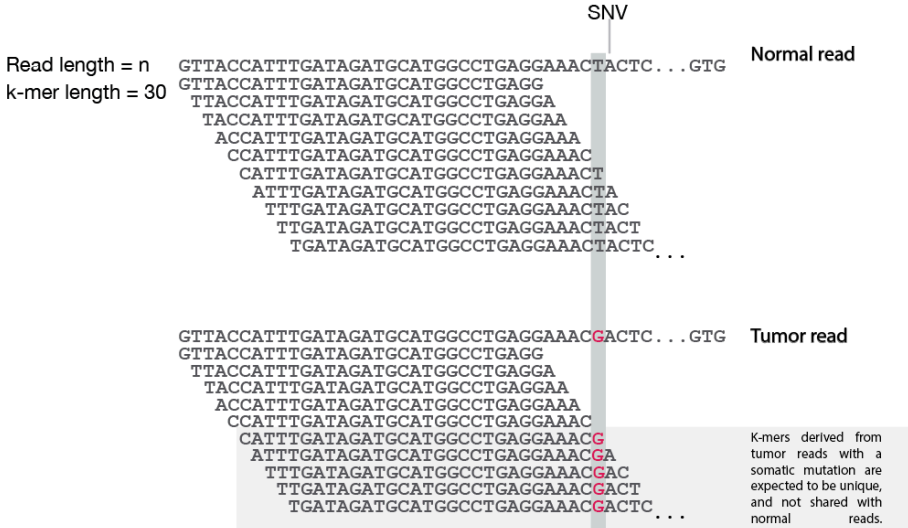
#### 1.1.3.1 Generation and analysis of k-mers

The purpose of the k-mer analysis is to be able to hunt and bring together, ideally, all the reads that correspond to the same region in the genome. For that reason, k-mer should be shared among the reads but unique along the genome in order not to gather information from different regions. On the basis of (Paszkievicz and Studholme, 2010) the approximate minimum sequence length that would allow the reconstruction of a whole genome is around 30nt. Therefore, taking this into account, and considering that a short k-mer would allow us to better explore the sequence space through the reads, we chose 30nt,

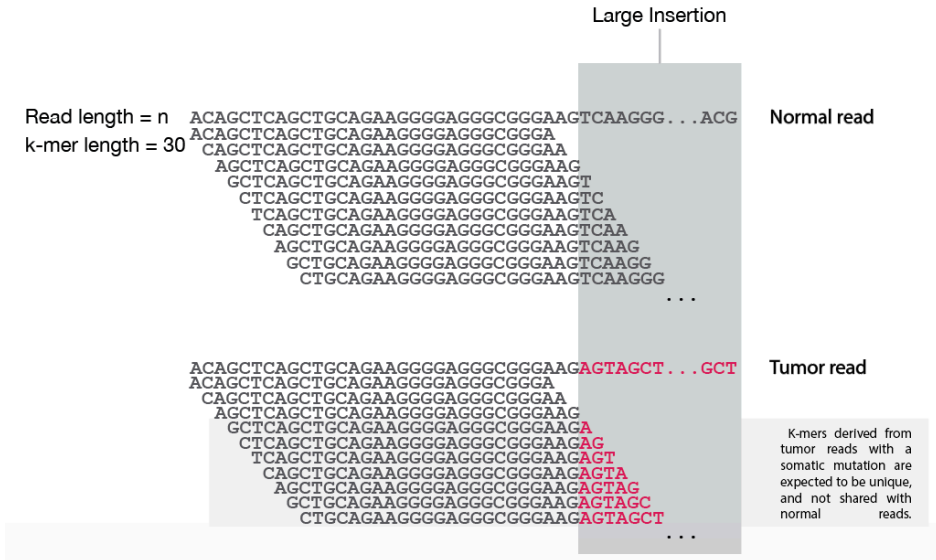
as the default value for the k-mer. Even Though we recommend this value, the program accepts other sizes of k-mers [28-32] to adjust to different genomes and situations.

After defining the k-mer size, we next counted the number of k-mers across all sequencing reads. As shown in section 1.1.1 k-mer from methods section, for each of the reads, we start from the beginning of the sequence moving base by base to annotate and count all possible 3onucleotide (nt) long k-mers. In this way, we make sure we cover the whole extent of the read, to capture any possible variation that can be found in it. Thanks to this procedure, we can generate kmers that will be common between the normal and tumoral samples and kmers that we will only find in the tumoural sample, which will be those susceptible to contain a variation (Figure 12).

**A) Generation of k-mers in a normal and tumor sample containing an SNV.**



**B) Generation of k-mers in a normal and tumor sample containing an SV.**



**Figure 12. Representation of k-mers generation on mutated scenarios.** For each scenario A and B , we have a read from the normal sample, and a read from the tumor sample that contains a variation that is marked by red characters. We observe how the generation of k-mers covers the whole sequence, and both common reads between the two samples and unique reads containing the mutation are generated.

### 1.1.3.2 Building a hashtable structure

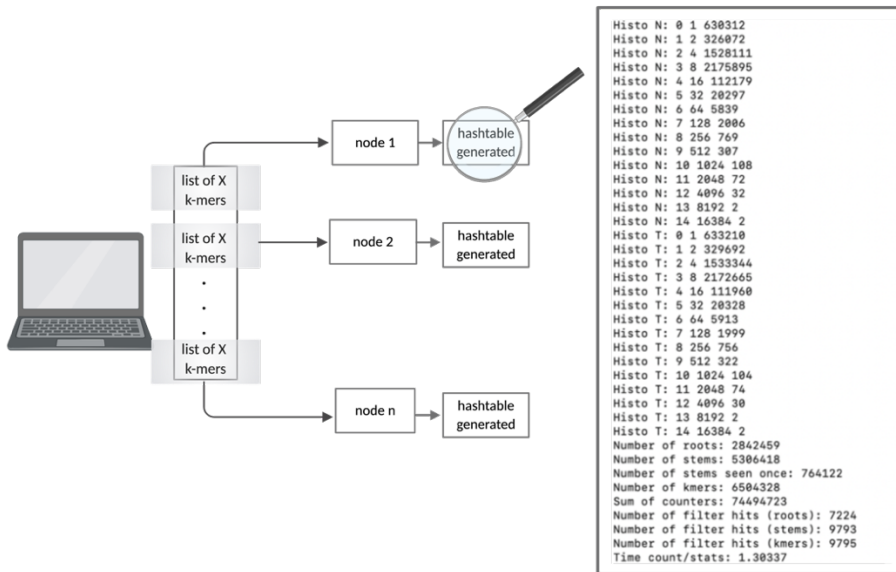
After analyzing and counting all the k-mers present in our datasets, we designed a form of having all this data accessible to be able to make different types of queries. For that, we stored this information in a hashtable structure that allows us easy and quick access to the data with reduced memory requirements, despite the large number of k-mers generated from both normal and tumour genome sequences. For each k-mer stored with their frequency on normal and tumor samples, information related to the read and the position in it is also . To provide the possibility of performing different queries to the hashtable, for each k-mer, we also store the information corresponding to the reverse complement of each particular k-mer.

To detect the mutations, the first step is to generate the stem with prefix one and suffix one for each of our kmers. Please, see section 1.1.1 of Methods for a better understanding of the generation of the stem through the kmer and its prefixes and suffixes. The objective is to find the beginning and end of the mutation. The stem is shared between the mutated and non-muted reads, and the inflection belongs to the beginning and /or end of the variation between the two, which allows us to collect all the reads of both samples for the same region and reconstruct it. By creating the inflection at the beginning and end of the kmer, we can cover the entire sequence and detect the mutation throughout the read.

Once all the k-mers of both samples are computed, a summary of the hashtable is generated (Figure 13) that includes, for example, the data for the creation of a histogram of k-mers frequency, k-mer counters, between others. In addition to directly pointing to candidate reads having the mutation, and to bring together normal and tumor reads of the same genomic region, we can also evaluate other useful information regarding, for example, the quality of the



sample. For instance, knowing the number of k-mers only seen once within a sample gives an estimation of the rate of sequencing errors within the sample.



**Figure 13. Storage and generation of hashtable.** Sample hashtable summary generated by each node where the generated k-mers are processed. created with Biorender.com

### 1.1.4 Detecting candidate somatic variants with k-mers

Once the hashtable is generated, with all the k-mers, their recurrence within the normal and tumor samples and information about the corresponding reads associated, the next step is to select those kmers that are likely to contain a variant. For this, we expect to find k-mers with different counts. For example, k-mers that have been found at the same rate in both samples, are expected to cover identical regions within both genomes, and therefore are not expected to contain mutations. On the other side, k-mers that have been found in tumor samples, but not in normal samples, are expected to cover somatic variants, and are actually the target of our analysis.



With this criterion, we make sure that:

- The mutation is covered in both directions (forward & reverse) and by a minimum of tumor reads; variables  $Y_1 - Y_3$ .
- The normal sample can contain a maximum of readings from the tumor sample, due to contamination; variables  $Y_2$  and  $Y_4$ .
- We detect those k-mers that, with their inflection, there are differences between normal and tumoural.
- We detect those k-mers whose own stem contains a variation

These variables can be modified by the user in the configuration file in the following fields:

max-normal-count-a = \_\_MAX\_NC\_A\_\_

min-tumor-count-a = \_\_MIN\_TC\_A\_\_

max-normal-count-b = \_\_MAX\_NC\_B\_\_

min-tumor-count-b = \_\_MIN\_TC\_B\_\_

A and B represent an arbitrary direction, since, as mentioned above, we do not know the real direction compared to the reference genome.

At this stage, besides selecting the candidate breakpoints with k-mers, we also keep additional information that will be necessary for the next steps: (i) the selection of all relative reads, (ii) the position of the k-mers within the reads, and (iii) a map of the k-mers.

The selection of the relative reads is based on the stem and the checking when any inflection meets the criteria described above, in this way we get both the reads that contain the mutation and the reads that pass through the same region without the mutation.

### 1.1.5 Clustering and filtering candidate somatic variants to build blocks with candidate variants

To this step we arrive with all the tumor-specific or tumor-enriched k-mers potentially having a variant, and the information of the read they belong to. From these data, the method next identifies and extracts the matching tumor reads, together with the normal reads ideally corresponding to the same genomic region. These tumor and normal reads are then piled-up to form a sequence block that will be analyzed more in detail, to define the final variant. For this, we first make a selection of the so-called leading reads, which are defined as those reads that cover the mutations as efficiently as possible to collect all the required information. This read is called “leading read”.

The criteria to select the leading reads is that it has to contain a minimum of (Y5) candidate k-mers, and (ii) the distance between these minimum k-mers must not be further than (Y6) nucleotides (Figure 15).

**A) Leading read candidate containing an SNV.**

```

GTTACCATTTGATAGATGCATGGCCTGAGGAAACGACTCAGTCC...G
CATTGATAGATGCATGGCCTGAGGAAACG 6
ATTGATAGATGCATGGCCTGAGGAAACGA 7
TTTATAGATGCATGGCCTGAGGAAACGAC 8
TTGATAGATGCATGGCCTGAGGAAACGACT 9
TGATAGATGCATGGCCTGAGGAAACGACTC 10
GATAGATGCATGGCCTGAGGAAACGACTCA 11
ATAGATGCATGGCCTGAGGAAACGACTCAG 12
TAGATGCATGGCCTGAGGAAACGACTCAGT 13
AGATGCATGGCCTGAGGAAACGACTCAGTC 14
GATGCATGGCCTGAGGAAACGACTCAGTCC 15
  
```

Groups of seven candidates k-mers. Distance between first and last need to be  $\leq 10$  positions inside read.

**B) Leading read candidate containing an SV.**

```

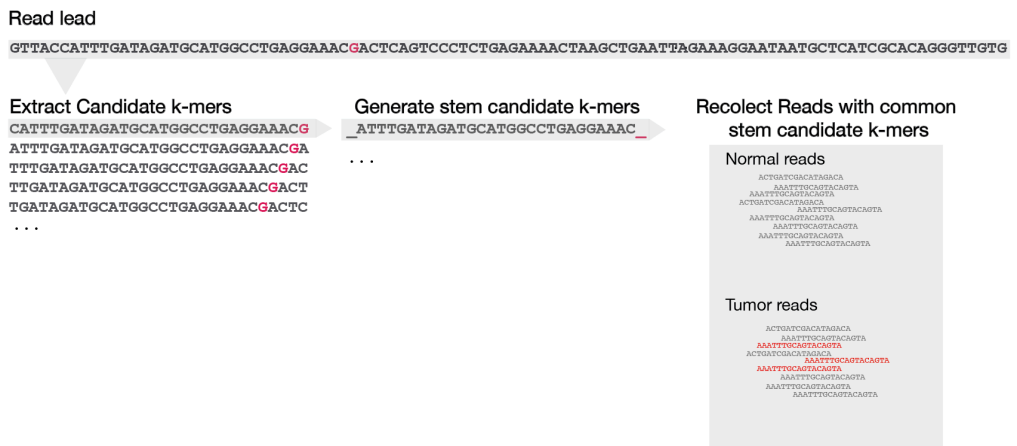
ACAGCTCAGTGCAGAAAGGGAGGGCGGGAAGAGTAGCTCCCC...T
GCTCAGCTGCAGAAAGGGAGGGCGGGAAGA 4
CTCAGCTGCAGAAAGGGAGGGCGGGAAGAG 5
TCAGCTGCAGAAAGGGAGGGCGGGAAGAGT 6
CAGCTGCAGAAAGGGAGGGCGGGAAGAGTA 7
AGCTGCAGAAAGGGAGGGCGGGAAGAGTAG 8
GCTGCAGAAAGGGAGGGCGGGAAGAGTAGC 9
CTGCAGAAAGGGAGGGCGGGAAGAGTAGCT 10
TGCAGAAAGGGAGGGCGGGAAGAGTAGCTC 11
GCAGAAAGGGAGGGCGGGAAGAGTAGCTCC 12
CAGAAAGGGAGGGCGGGAAGAGTAGCTCCC 13
AGAAAGGGAGGGCGGGAAGAGTAGCTCCCC 14
GAAGGGAGGGCGGGAAGAGTAGCTCCCC 15
  
```

Groups of seven candidates k-mers. Distance between first and last need to be  $\leq 10$  positions inside read.

**Figure 15. Leading reads for SNV and SV scenarios.** Example of positive read leader selection with the variables  $Y5 = 7$  and  $Y6 = 10$ .

With this criterion, we ensure that the leading read has created a minimum number of candidate k-mers in a range of nucleotides, a sign that the mutation is well covered.

Next, taking each leader as a seed, we start adding other reads that share the same stem, and are therefore expected to derive, a priori, from the same genomic region. Using sequence information (stem), we also fetch the corresponding reads coming from the normal sample, and construct the block. We will also use the stem in reverse complement in order to reconstruct the region in both directions, forward and reverse (Figure 16).



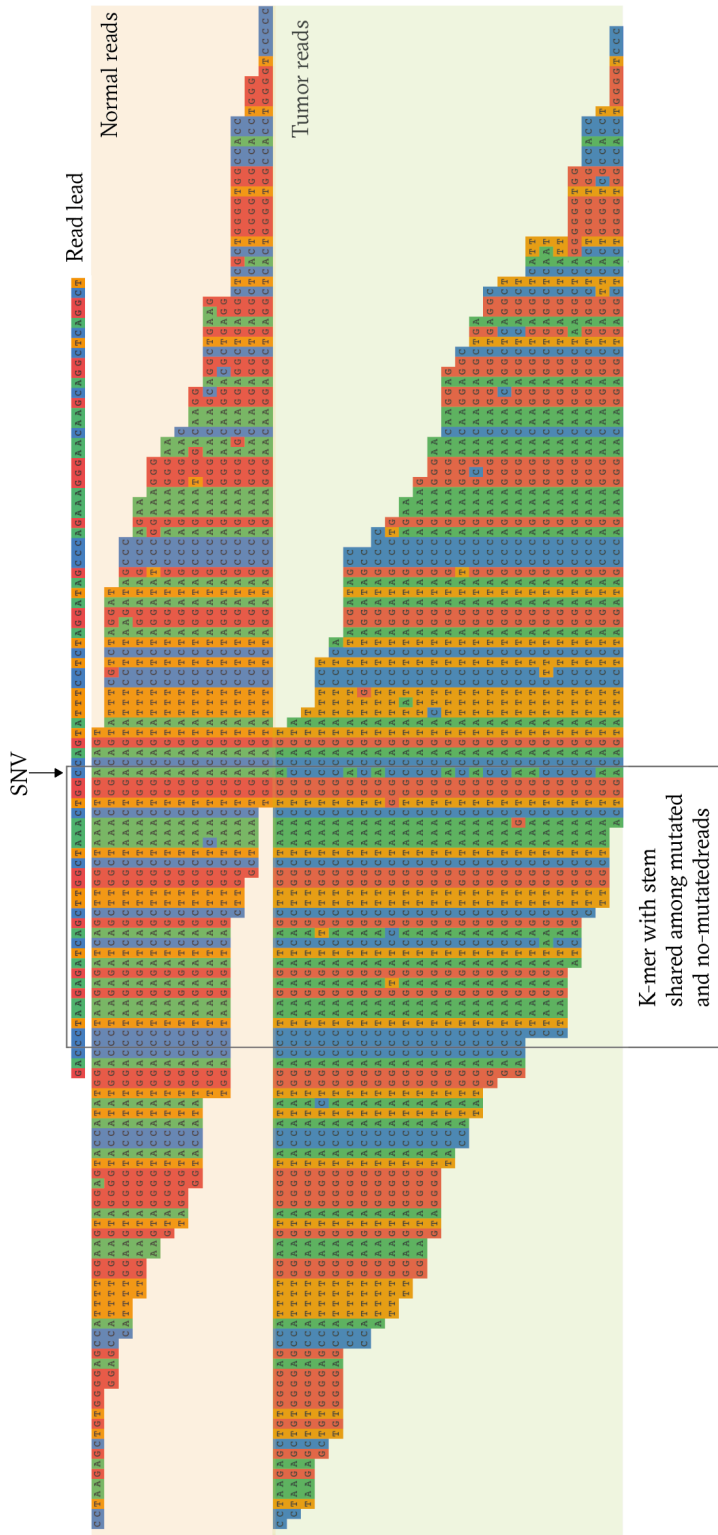
**Figure 16. Reconstruction of block.** From a read lead, we extract all the candidate k-mers that it contains, generating the stem of each one of them to be able to collect all the reads both mutated and not mutated of the normal and tumoral samples and to build the block that contains the somatic mutation.

With this approach a single candidate mutation can generate more than one block. This is because more than one read covering the mutation can become a read leader. The next step is to remove redundancy, by filtering out those blocks that contain the same reads, or that are already included in another block.

Then we pile-up the reads taking the stem from the read leader as an anchor to add the other reads. For each sequence collected through the read leader's k-mers, we look for the root that has been called and that is common with the

read leader sequence. Then, each read is successively positioned so that its inflection or partial inflection coincides with the central read. For the reverse direction of the block we use the same read leader but in a complementary reverse direction so that the k-mers match and could be pile-up.

The results of this section correspond to a collection of blocks reproducing approximately 150bp-long genomic regions (with original read size of 100bp), if original read size is 100bp, that are expected to contain a somatic variation. These blocks, finally are expected to contain, for each candidate region, reads covering both alleles of the normal sample, reads covering the same region of the non-mutated tumor allele, and reads covering the mutated allele with the mutation (see Figure 17).



**Figure 17. Representation of a candidate block in this block, the presence of an SNV is observed in the tumor sample with the change of A/C. In the first position, we have the leader read that has created the block. Read lead comes from the tumor sample and contains the somatic variation for which it has built. Read lead also includes the stem that is shared with all reads, both mutated and non-mutated. Through this k-mer, all the reads have been stacked to reconstruct the candidate region.**

## 1.2 SMuFin2–algorithm implementation

SMuFin was deployed on 16 nodes of MareNostrum 3, where, per each patient, it costs around 10 hours and 56 kWh to complete a single analysis. With improvements on algorithm, accelerators, and NVM used as main memory extension, SMuFin2 can be executed on one single enterprise-node with 512GB of main memory, and process a 30x coverage genome pair in 9 hours and as few as 4.3 kWh, which means a 13.1x improvement. Nevertheless, we were able to run SMuFin2 on a desktop machine only by adopting NVMe as an alternative to main memory. Running SMuFin2 in an affordable node with a 6-core i7 and only 32 GB of main memory, required 22.4 hours, a significant slowdown, but in consuming only 2.4 kWh, a 23.3x improvement over the original deployment.

If we compare the single enterprise node against the desktop machine this last one supposes only  $\frac{1}{4}$  of the cost, and it requires approximately half of energy for each execution. As a result, a cluster of multiple desktop machines costs half as much as a cluster of servers, and consumes half the energy while maintaining similar performance. These results (Figure 18) demonstrate that hardware/software co-design allows significant reduction in the total cost of ownership of data intensive genomics methods, facilitating their adoption in large genome repositories.

More detailed information can be found in Dr. Cadenelli thesis, as this was the result of a collaboration, and on papers (Cadenelli et al., 2017 ; Cadenelli et al., 2019).



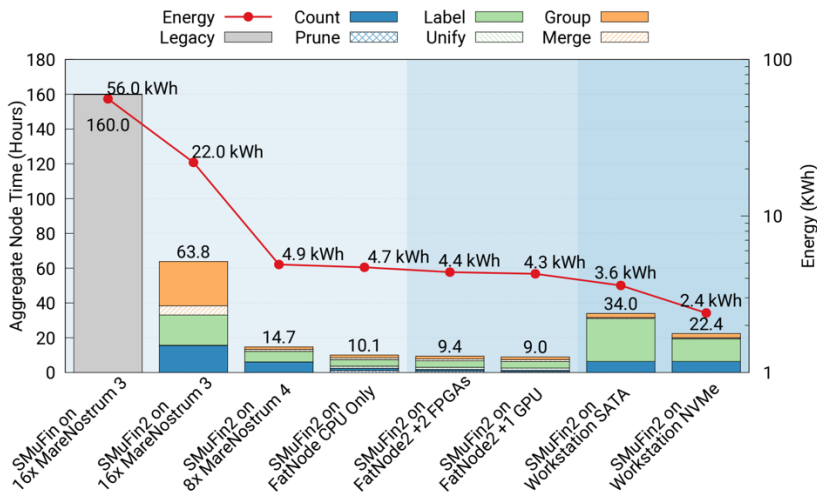


Figure 18. Aggregate node time and energy-to-solution of all SMuFin2 versions on the different hardware configurations.

### 1.3 Algorithm validation

In summary, at this stage, we have developed a comprehensive methodology for the processing and classification of sequence reads, according to their potential coverage of somatic variants, and using a k-mer-based methodology. The algorithm, and its implementation in an High-performance computing (HPC) environment, allows a complete processing of entire whole genome sequences very quickly, and at low computational cost. Driven by the novelty of this methodology, and because of its potential use for other specific sequence analysis, provided a different and adjusted processing of the sequence blocks, we decided to patent this algorithm (EP16178577.9). To validate the potential of this algorithm to identify somatic variants, and to develop the second part of the entire program, we performed extensive assessments as described below.

#### 1.3.1 Generation of an *in silico* test sample for initial validation

To validate the technique and verify that the expected prediction of the mutation was met, we perform calibration in parallel to the development of

the method. To verify that the algorithm works as expected in each step, we need to create controlled reference set containing all the elements to be analyzed.

In the case of the variant caller, an in-silico sample is used to control the totality of mutations and to know exactly all the sequences with or without mutation that pass through it. It is not recommended to use an in-vivo sample because it is not known with certainty the total mutations that it contains.

The sequences, both mutated and non-mutated, covering the mutations, are marked and selected. In the steps where the program executes a filter, which are: Detecting candidate breakpoints with k-mers, and Clustering and filtering candidate breakpoints to build candidate blocks, we extract all the reads that have passed the criterion, and in this way, we can check if a mutation is represented. Finally, in the block alignment step, those blocks that contain mutation sequences are marked. This marking allows us to check later in the final block what the mutations look like.

To emulate the steps taken by the algorithm, we re-create all potential/possible scenarios, most of them manually. This allowed knowing firsthand the failures and improvements that could be generated. The calibration sets of this step consisted of small regions of the genome (between 200-500 bp) that contained a unique mutation.

Once the algorithm was validated on different types of mutations, we proceeded to build a chromosome in-silico to exclusively validate the method against it.

For the creation of the in-silico (see methods section 1.2.1), we used the program ART-Illumina that allows us, on the one hand, to simulate the sequencing of a sample and on the other hand, thanks to a secondary file we know all the reads that pass through each position with the information of the

sequencing errors they contain. We selected the chr20 since, by size, it is more manageable to control along with all the steps of the program. In the in-silico, we added a profile of germinal mutations in both normal and tumor samples. In the tumoural sample, a wide range of mutations was chosen, from SNV to SV . This added variation was composed of: 168 SNVs, 26 Indels, 20 SVs, and 1 viral insertion

These mutations were the same used for the SMuFin test so that we could compare it with its predecessor. Besides, a virus fragment was inserted, namely Ki polyomavirus, at position 56.398.700. Detailed information about the potential of SMuFin2 for the identification of tumor-associated virus can be found in section 1.4.

The test with the first block from the algorithm was done with the variables on config.file:

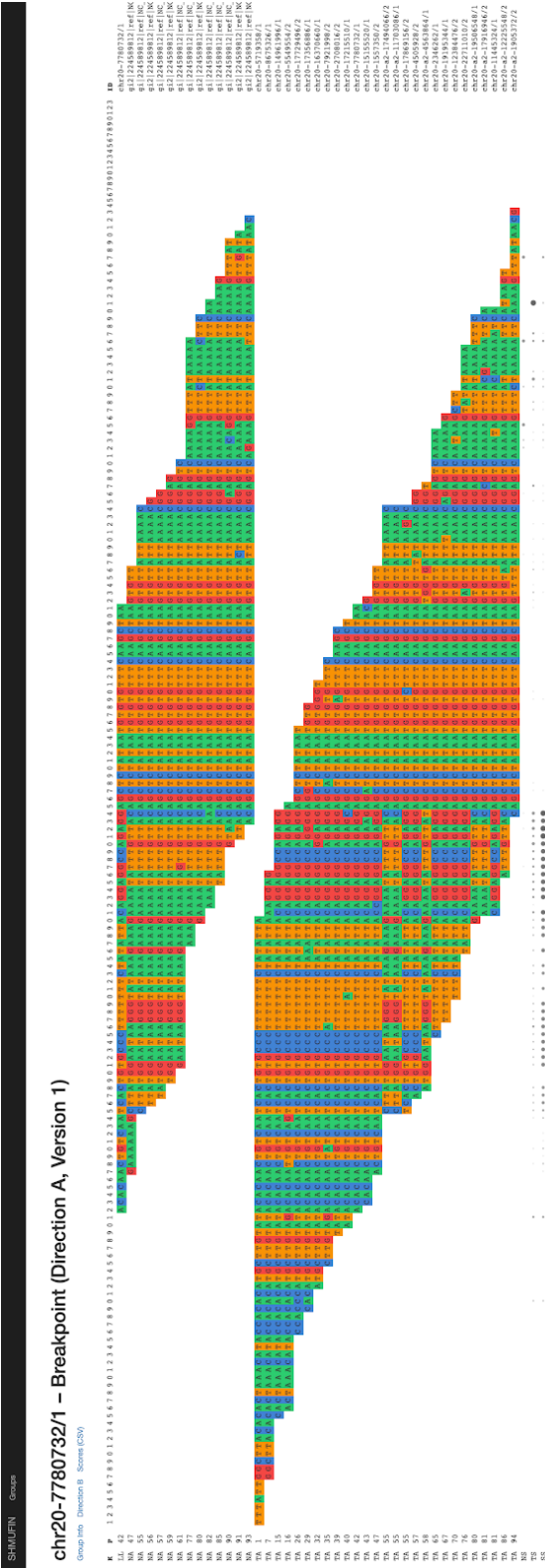
```
KMER_LENGTH=28
MAX_NORMAL_COUNT_A=1
MIN_TUMOR_COUNT_A=4
MAX_NORMAL_COUNT_B=1
MIN_TUMOR_COUNT_B=1
WINDOW_MIN=7
WINDOW_LEN=10
```

The results for this test were: 100% for SNVs, 100% for Small SVs, 100% for Large SVs, and 100% for virus Insertion (Table 1). In the table, we can check the results in the two checkpoints mentioned above, after the filter by candidate k-mers, and after the cluster to build the block.

	Mutations	After filtering	%sensitivity	After clustering	%sensitivity
SNV	168	168	100%	168	100%
Small SVs	26	26	100%	26	100%
Large SVs	20	20	100%	20	100%
Virus	1	1	100%	1	100%

**Table 1. Results from chr20 test.** In each filtering step we evaluate the number of detected mutations classified by type.

It is remarkable the capacity of detecting large structural variants with such a high sensitivity, even the insertion of a virus (Figure 19).



**Figure 19. Detection of Ki polyomavirus insertion.** Screenshot of Interactive output generated by SMuFin2 with the block aligned in one of the directions. In the upper part, we find the reads of the normal sample, in the lower part, the reads that come from the tumor sample. In this second block, we can see the reads that contain the insertion of the virus. The insertion point can also be seen with the variability value points that are in the last rows of the alignment.

## 1.4 Identification of tumor-associated viruses

To test the virus detection role of the first block from SMuFin2, we explored the WGS of PCAWG Consortium (Consortium, 2020) generated by the ICGC and the Cancer Genome Atlas projects.

For the selection of samples, we considered those where infection-related causes of cancer were estimated to be viral, such as Hepatitis B virus (Bialecki and Di Bisceglie, 2005), and several Human papillomaviruses (Munoz et al., 2006) types.

Therefore, we randomly selected some patients from the following studies: (DCC Project Code; Project Name; Country)

CEC-US ; Cervical Squamous Cell Carcinoma - TCGA, US ; US

LIHC-US ; Liver Hepatocellular Carcinoma - TCGA, US ; US

UCEC-US ; Uterine Corpus Endometrial Carcinoma- TCGA, US ; US

We ran SMuFin2 first block with the following variables:

```
KMER_LENGTH=32
MAX_NORMAL_COUNT_A=1
MIN_TUMOR_COUNT_A=3
MAX_NORMAL_COUNT_B=1
MIN_TUMOR_COUNT_B=2
WINDOW_MIN=7
WINDOW_LEN=10
```

Thanks to the summary information obtained in the hashtables, we were able to verify that in some patients, the normal sample had lower coverage than the tumor sample. We were also able to extract that the sequencing errors were low in the sample, and therefore we continued with the k-mer at a start of 32.

Our objective was to detect the presence of viruses in the samples, and therefore, we focused on the groups that did not contain reads from the normal sample. Following the detection strategy explained in the section 1.1.6 .

### Detection of HBV

In the patients of the LIHC-US project, we found positive results in virus detection. The most frequently detected virus was HBV. We also observed cases where the presence of HERV-K117 was detected.

### Detection of HPV

On patient samples from the CESC-US and UCEC-US projects, we found results favorable to the presence of viruses. The virus that appears most frequently is HPV16 in both samples. In the case of the UCEC-US project, the presence of the HPV18 virus was also found.

The results of the virus presence obtained are in agreement with those presented in the recent paper published by the working group of pathogens of the PCAWG consortium (Zapatka et al., 2020).

## **1.5 Cataloguing and annotating blocks**

This second part of the algorithm consists of the detection of the different variants within the blocks and the subsequent alignment against the genome identifying its exact position.

It is an interim strategy that we are currently implementing. This strategy is based on the detection of somatic mutations from SNV to SV. One point to note is that it addresses not only the detection but also the insertion point of non-human genetic material such as viruses.

Once we have all the candidate breakpoint blocks aligned from the previous step (section 1.1.5), we proceed to the detection and identification of variants for each of them.

We use the alignment information to observe the differences in each group: tumor, non-tumor, and both.

The first step consists of joining both directions, forward and reverse, of each block in one, to obtain a greater coverage of the region. For each position in the block, a value is given according to the variability within that position. As a result, we get a representation of all the variability within the block. This variability is calculated for the three different groups: tumor, non-tumor, and both.

These alignment scores are compared recursively to identify differences in both samples, tumor and non-tumor. With these scores, we first evaluate a consensus for each sample, to avoid false positives and misalignments.

We then look for all variants that are completely included within the comparand block. These variations will be SNV and small SV, which will consist of: insertions, deletions, and inversions. All blocks that do not meet this criterion will be candidates to contain a large SV, which means that the block only covers one of the breakpoints of possible large insertions, deletions, inversions, or intra- or inter-chromosomal translocations.

Once all types of variations are defined, we move on to identify the coordinate of the mutation. We generate a consensus of the normal block, for which we have stored the position where the mutation occurs and aligned it against the reference genome. Using the consensus, we obtain a sequence with a longer length than the original read, which allows a better alignment. In addition, we avoid possible alignment problems due to the presence of the mutation that we are questioning, as it usually happens in those non-reference-free methods that are based on references.



For the blocks containing an SV, the tumor consensus carrying the mutation will also be generated and aligned against the reference genome to know the chromosome that cause the SV and the coordinates of the variation. This same process of tumor consensus will be done by mapping this time against non-human databases to locate which viruses are inserted and the exact insertion position.

For the detection of non-human insertions in the genome, the user can choose against which database he wants to perform the alignment of the groups susceptible to contain an insertion of non-human material. In the test runs (section 1.4) a database of all viruses described in the methods section 1.4.2 was used. This guarantees that the user can make a more general or more specific search according to his criteria.

As mentioned before, the method is also able to detect the non-human sequences present in the sample. To do this, we rely on the knowledge that the sequences that come from viruses that are not homologous with the normal sample generate groups that only contain tumor reads. Following this criterion, we use the blocks that only contain tumor sequences, and we make a consensus and map them against a selected database.

Knowing which viruses are in the sample beforehand helps us to determine when we find a breakpoint with a virus insertion as the sequence is shorter to go towards a more targeted search.

## **1.6 Output files**

The results of the program are presented in two formats: Variant Call Format (VCF), a standard file of the detection methods, and an interactive web page.

The VCF will provide the eight mandatory columns:

- 1; CHROM        The chromosome on which the mutation is being called.
- 2; POS    The position of the mutation on the reference genome.
- 3; ID    The identifier of the variation
- 4; REF    The reference base (or bases in the case of a small deletion) at the given position of the non-mutated sequence.
- 5; ALT    The mutated base or bases at this position.
- 6; QUAL    A quality score associated with the mutation.
- 7; FILTER        A flag indicating the filters the mutation has passed.
- 8; INFO    An extensible list of key-value pairs (fields) describing the variation

The website consists of:

- Home page with a list of all the blocks lined up.
- Each block/group will be given the following information that has composed it:
  - Overview consisting of Number of k-mers (A), Number of k-mers (B), K-mers distance (A), K-mers distance (B), Number of reads (N), Number of reads (T) , and Number of reads (N+T)
  - Lead read
  - List of K-mers in direction A with respective counters for normal an tumoral sample
  - List of K-mers in direction B with respective counters for normal an tumoral sample
  - List of Normal reads ; ID + sequence

- List of Tumoral reads ; ID + sequence
- Alignment for the direction A
- Alignment for the direction B
- On the block alignment page the user can interact with the data:
  - Rearranging the alignment by type of read or by alignment position
  - Marking with color by base type, by read type or without color.

## 1.7 SMuFin2 first block execution

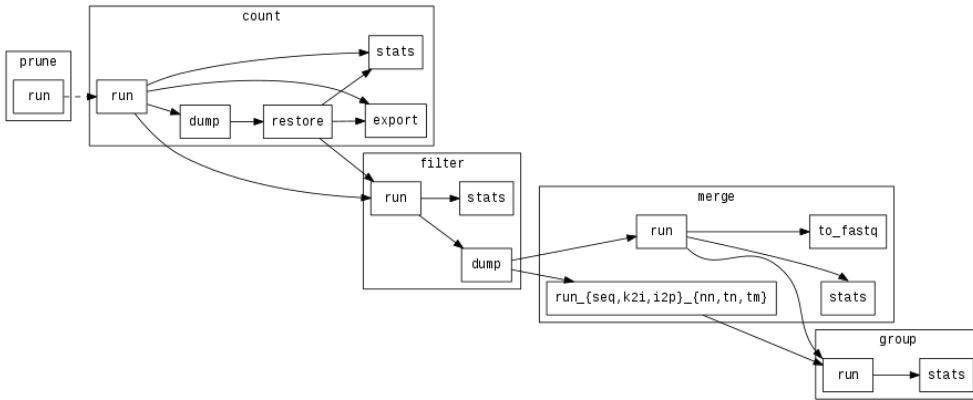
SMuFin2 has been conceived as a reconfigurable set of checkpointable stages (Figure 20), developed C++ , and Python programming language.

Depending on where the hardware is running, SMuFin2 supports different modes of execution to suit it: from scale-out executions in large data centers to scale-up solutions that take advantage of accelerators and storage-class memory in a single machine.

### 1.7.1 Compile

Compiling the first block from SMuFin2 requires make, a compiler such as gcc with C++11 support ( $\geq 4.8$ ), and the following libraries:

- sparsehash ( $\geq 2.0$ )
- boost ( $\geq 1.55$ ): Property trees and string algorithms
- ConcurrentQueue and ReaderWriterQueue: MPMC and SPSC queues
- libbf ( $\leq v0.1$ -beta): Bloom filters
- RocksDB ( $\geq 4.9$ ): Key-value store for flash storage
- htlib: Parse BAM files



**Figure 20. SMuFin2 execution command graph.** The graph shows the dependencies between SMuFin2 commands and all checkpointable stages: (i) prune, (ii) count, (iii) filter, (iv) merge and (v) group.

## Run

A configuration file (Supp Figure 1) with all the necessary variables and paths is required to run the method.

## Commands

The argument passed to the `--exec` flag or the configuration option in `core.exec`, must be a list of stage commands separated by semicolons. The commands are prepared with a stage name followed by a colon and chained in a comma-separated list. For example, `count:run,dump` or `count:restore;filter:run,dump`.

Note that the commands must follow a specific order, and some stages cannot be executed without running the previous stages first.

The following list contains all available steps and commands represented on Figure 16:

- prune
  - run: generates a bloom filter of stems that have been observed in the input more than once; optional stage that can be run first to save memory during count.

- count: build frequency table.
  - run: counts frequency of normal and tumoral k-mers in input sequence, ignoring k-mers whose stem is only seen once; counters hold values up to  $2^{16}$ .
  - dump: serialize k-mer frequency as sparsehash tables indexed by stem, for checkpointing and/or later analysis.
  - restore: unserialize dumped frequency tables from disk.
  - stats: display frequency stats, including size of different tables, and histograms for normal and tumoral counts.
  - export: serialize frequencies as plain CSV table files containing k-mers along with normal and tumoral counters; rows can be limited to k-mers that meet certain criteria through configuration options export-(Altschul et al., 1990).
- filter: select breakpoint candidates and build indexes.
  - run: build filter normal and tumoral (mutated and non-mutated) indexes containing candidate reads, along with their IDs and positions of candidate k-mers.
  - dump: finalize writing filter indexes to disk; when using RocksDB indexes, force a compaction.
  - stats: display sizes of the different filters.
- merge: combine multiple filter indexes.
  - run: read and combine filter indexes from different partitions into a single, unified index in RocksDB. Merges all possible indexes, sequentially one at a time.
  - run\_{seq,k2i,i2p}\_{nn,tn,tm}: read and combine specific filter indexes from different partitions into a single RocksDB instance.

- stats: display sizes of the merged filters.
- to\_fastq: convert indexed reads to FASTQ format.
- group: match candidates that belong to the same region.
  - run: window-based group leader selection and retrieval of related reads.
  - stats: display number of groups generated by each thread.

## 2. Landscape of somatic variation in neural development and the role of *Pgbd5*

The next project is an ongoing study. The results reported here are preliminary, and the results of Dr. Kentsis group, collaborating in this project, are not published yet.

In brief, this part of the thesis aims to answer the hypothesis that *gbd5* is active as a nuclease and produces DNA double-strand breaks (DSB). This activity produces DNA rearrangements in neurons during brain development, which allows them to survive subsequent apoptotic selection. The interest on *Pgbd5* is preceded by (Henssen et al., 2017a) that unveils the role of DNA transposase *PGBD5* that, by acting as a nuclease in human cells, underlies cell transformation by inducing site-specific genomic rearrangements. We seek to compare Knockout (KO) and wild-type (WT) individuals to determine *Pgbd5*-induced somatic rearrangements in neural tissues. To understand the role of *Pgbd5*, we divided the project into two parts: 1) the characterisation of somatic variation for neural tissues, and 2) the study of those variations associated with *Pgbd5*.

This project required collaboration between multidisciplinary teams. On one side, a wet lab focused on KO mice, led by Dr.Kentsis lab from the Memorial Sloan Kettering (MSK) New York (NY); and a dry lab, focused on the computational side, led by Dr.Torrents lab from Barcelona Supercomputing Center (BSC). My contribution here is the characterization of the landscape of somatic variation in neural tissues of adult and embryonic mice, and the contribution of *Pgbd5* (mouse orthologue of human *PGBD5*) during brain

development and adult state, together with my groupmate and Ph.D. student Elías Rodríguez-Fos. In particular, my contribution was on the data processing of FASTQ data received and the posterior detection and classification of somatic variants on neural samples; the preparation of the data for the characterization of the different type of variants on adult and embryo mice; and the identification of genes and genomic intervals specific related in each highlighted group.

The detection of somatic variants on non-tumoral samples was a methodological challenge. This new scenario, compared to the pattern of somatic mutations in cancer, is presented as a variation with a non-clonal profile, and a somatic variation with a lower VAF. These characteristics made its detection more complicated when using the conventional methods designed for a tumor mutation profile. Initially, when the study started we didn't know if we could detect variations, stating that the detection of somatic variants was something fundamental to the project.

To determine whether *Pgbd5* could be responsible for somatic rearrangements on neural samples, Dr.Kentsis group produced the KO mouse model of the *Pgbd5* allele. Then *Pgbd5* <sup>-</sup>/wt females and *Pgbd5* <sup>-</sup>/wt males were crossed to obtain *Pgbd5* wt/wt and *Pgbd5* <sup>-</sup>/<sup>-</sup> (or KO) littermates. For the identification of neuron-specific *Pgbd5* somatic DNA rearrangements in *Pgbd5*-wt and *Pgbd5*-KO mice models, Illumina high-coverage (80x) PCR-free paired-end genome-wide sequencing was used.

Two groups of mice were studied in parallel, adult and embryo, to analyze the effect of *Pgbd5* at different growth stages. We expected to study the effect in its origins on embryo samples to unravel the causes of DNA damages during

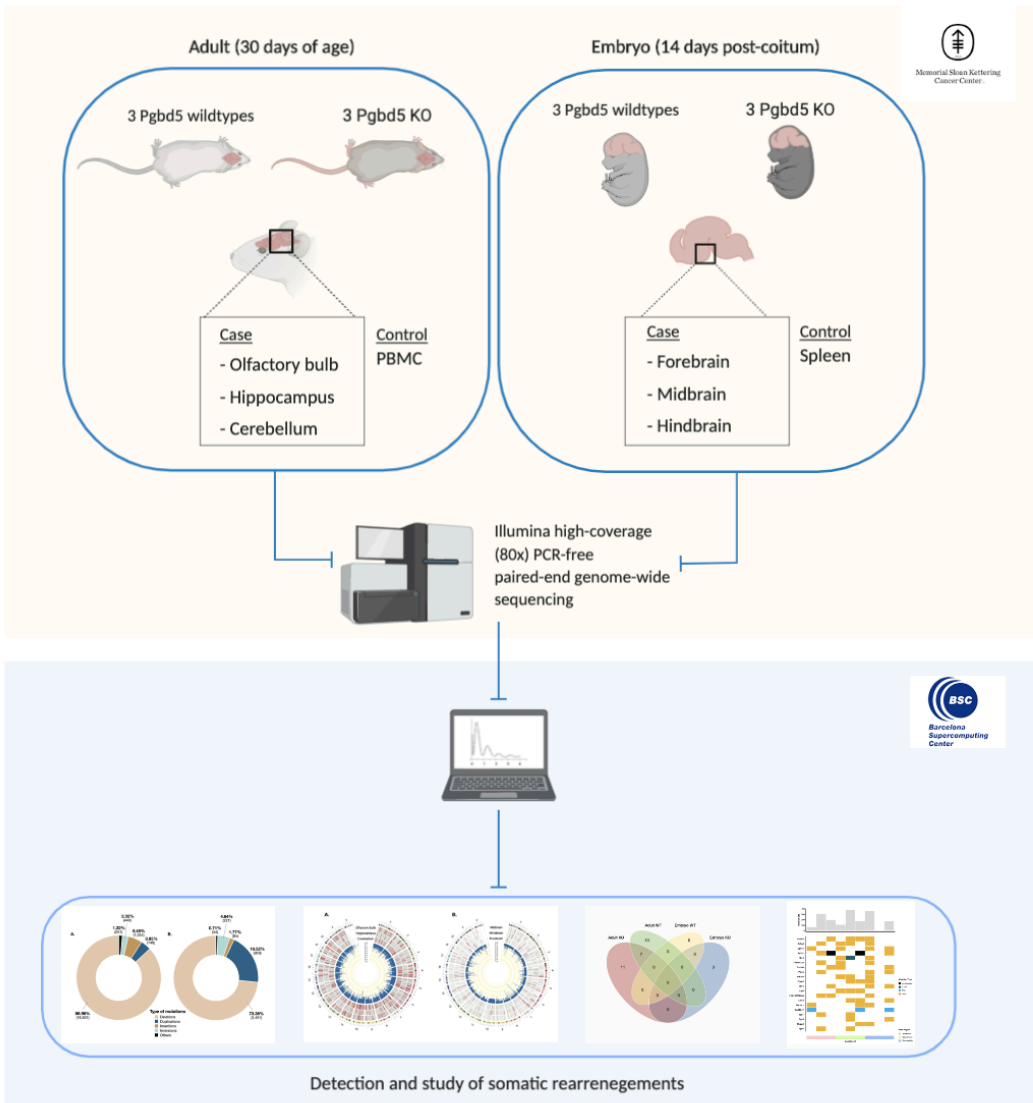


brain development, and follow the mutation to the adulthood to analyze how these rearrangements have turned out:

Adults brain: The number of samples was: 3 *Pgbd5* wild types and 3 *Pgbd5* KO littermates of 30 days of age. Just before the euthanasia, Peripheral blood mononuclear cells (PBMC) were collected as a control for the experiment. PBMCs are non-neural cells that undergo RAG1 recombination in the Immunoglobulin locus, serving as a perfect quality control for the subsequent analyses. As a case sample, three different neural tissue samples were collected: Olfactory bulb, Hippocampus, and Cerebellum.

Embryo developing brain: The number of samples was: 3 *Pgbd5* wild types and 3 *Pgbd5* KO E14 (14 days post-coitum) embryos from the same pregnancy. Just before the euthanasia, the embryos were extracted from the mother, and spleen was collected as a control for the experiment. The spleen is a hematopoietic organ during embryogenesis, and together with the liver, is where lymphocytes mature. As with PBMCs in adults, the spleen serves as a suitable control for this experiment. Structural differences exist between E14 developing and adult brains. As a case sample, three different neural tissue samples were collected: the Forebrain (the part that will give rise to the cortex among other structures), Midbrain (this part will give rise to the midbrain), and Hindbrain (this part give rise to the cerebellum and spine bulb).

In total, we studied rearrangements in 36 samples of neuronal tissue, 18 for each adult and embryonic group. Nine of them belong to the KO group and the other nine to the WT group.

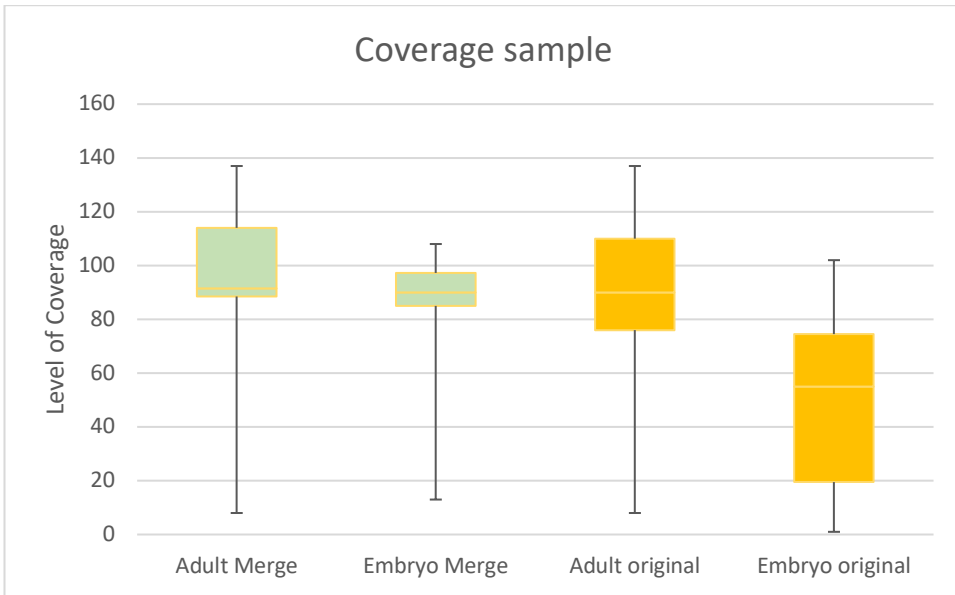


**Figure 21. General workflow for the study of somatic DNA rearrangements induced by Pgbd5 during brain development and adult state.** The first section represents the selection of the mouse samples obtained by Dr. Alex Kentsis' group, a total of 36 case samples and 12 controls that were analyzed during the study. The second represents the detection, and subsequent analysis of the variants carried out in our group, which include: the identification of somatic variants within non-tumoral tissues, the study of Pgbd5 KO mice vs. WT mice, the characterization of the deletions on wild-type mice, the identification and analysis of genes and genomic intervals, and the study of genetic ontology. Created with Biorender.com

## 2.1 Identification of somatic variants within neural tissues

Our aim in this part was to identify somatic variants. This was a particularly challenging task because, having experience in detecting variants in cancer tumor samples, we did not know a priori whether the methods used in that area would work for detection in normal tissues. The first question was if we were able to identify somatic variations from non-clonal tissue growth. Our first approach was to increase the coverage for the proper detection of somatic variants. This parameter is essential because the role that we hypothesized for *Pgbd5* would involve rearrangements that diversify the genomic content, leading to polyclonal rearrangements with reduced allele fractions, albeit sharing features that may lead to recurrent alignments. The higher the coverage rate, the higher the detection probability of these variants at such low fractions.

As we showed in Figure 21 we obtained the FASTQ data from the 48 samples sequenced. The first step in the data analysis pipeline was the alignment of the obtained sequences to the reference genome using BWA (Li and Durbin, 2009). To improve the coverage for the detection, all the FASTQs corresponding to the same sample were merged in a single BAM file.



**Figure 22. Boxplot for level of coverage sample before and after the merge step.** Notice the increment on level of coverage on samples before merge, particularly on embryo samples.

Figure 22 shows that in the case of the embryo samples, merging the data considerably increased the level of coverage from a median of 55 to an average coverage of 90. In the case of the adult samples, this increase was not seen in such an acute way, that increased only from an average of 90 to 91.5. It is worth emphasizing those samples that had very little initial coverage and that represented a greater challenge for the detection of variants. In the case of the embryos, we found an initial sample with only a coverage 1 that was increased to 13 thanks to the merge. In the case of the adults, the sample with the lowest coverage was a sample with only a coverage 8, since no more samples were available, its level of coverage could not be increased.

At the time of this study, the majority of variant callers variant callers are dedicated to the detection of somatic mutations in the clonal profile of cancer. Finally to perform the landscape of somatic variation in neural development and the role of *Pgbd5*, we chose three methods that allow the detection of

mutations at a very low VAF: Pindel, and Delly<sup>2</sup>, focusing on indels and large structural variants, and GATK focusing on SNVs. To perform the mutation detection, we follow the profiles previously indicated in the methods section 2.2.2. In this study, we did not include SMuFin because It did not produce convincing results during trial analysis, as it is quite conservative and disregards variants supported by a few reads. In one of the tests performed, specifically on a sample of Olfactory bulb from a healthy mouse, SMuFin detected only a Large SVs, zero small SV, and 90 SNVs.

In order to increase sensitivity, we joined the results for each sample obtained from the different callers We filtered out the duplicates within and between callers to avoid redundancy, considering a similarity window of 300bp. In case a mutation was found to be duplicated, we kept the one with the highest detection quality, VAF, or ultimately, we gave more weight to the deletions. In the final step, in order to maintain specificity, we selected those variations that had the default PASS quality filter for further analysis.

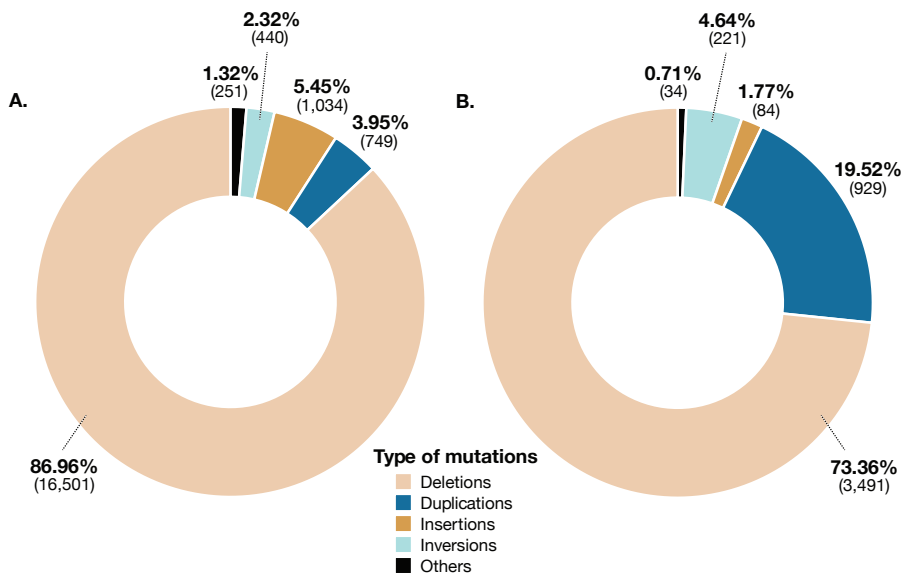
Altogether, we confirmed the ability to identify somatic variants on neural tissues despite the above-mentioned factors.

## **2.2 Comparative analysis of somatic variation between adult and embryo**

With the current results we had a first overview of the landscape of somatic mutations in neural tissues in embryo and adult samples (SNVs are currently being detected and will be included next in the study). Based on the observed variation we found a total of 32,190 somatic mutations in adults (18,194 on WT sample; 13,996 on KO sample) and 9,795 on embryos, (3,816 on WT sample; 5,979 on KO samples). As shown in the pie-chart (Figure 21), deletions were the most represented variants on adults wild-type mice with (86.96%) , followed by insertions (5.45%), duplications (3.95%), inversions (2.32%), and others (1.32%). In the case of embryos wild-type mice, deletions

were the most represented variants with (73.36%), followed by duplications (19.52%), inversions (4.64%), insertions (1.77%) and others (0.71%).

Given the predominant profile of the deletions in both adult and embryonic samples in the WT group, and based on previous studies (Henssen et al., 2017b) we focused our attention on deletions.

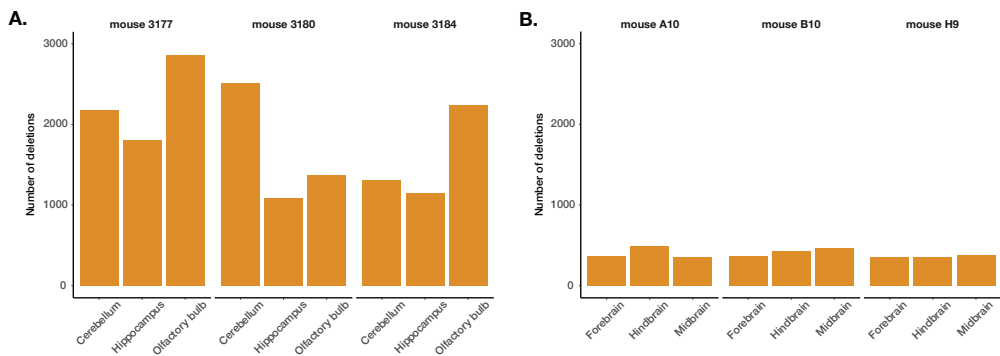


**Figure 23. Landscape of mutations in wild-type mice.** A) Representation of somatic variants detected on Adults wild-type mice. B) Representation of somatic variants detected on Embryos wild-type mice. Of note, the most prevalent type of intrachromosomal mutations acquired in WT mice were deletions, representing 86.39% of mutations in embryos, and 86.97% in adults

### 2.3. Characterization of the deletions on wild-type mice

To check if there was any deviation that could affect the total count of the samples and to analyze the differences between tissues and their possible relationship with *Pgbd5* activity, we studied the distribution of the deletions through the different tissues from neuronal samples.

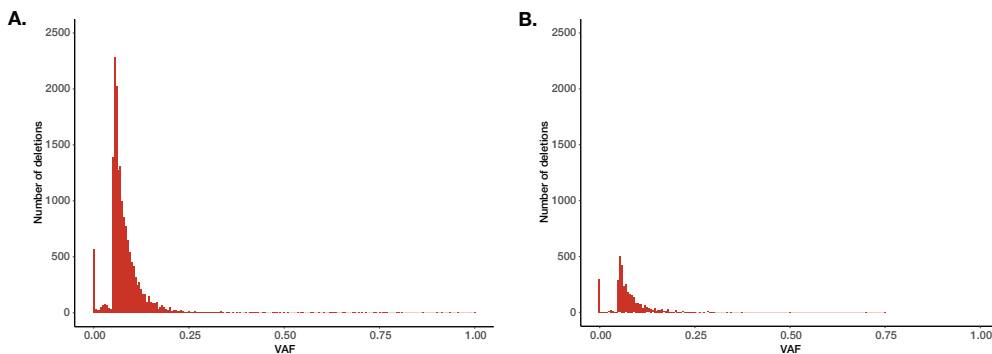
As the first exploration on wild-type mice, we observed that the number of deletions was not dependent on tissue type in either adult or embryo samples (Figure 24). In conclusion, we can indicate that no sample or mouse could deviate our statistics since the distribution is presented relatively homogeneously; and that a priori there is no significant difference between tissues due to the possible activity of *Pgbd5*.



**Figure 24. Number of deletions for each tissue on A) Adult sample and B) Embryo samples.** The X-axis is divided at the top by the mouse sample, and below it lists the three tissues for each one: Cerebellum tissue, hippocampus and olfactory bulb in the case of Adults, and Anterior brain, Posterior brain and Midbrain in the case of Embryos. On the X axis it represents the number of deletions, which ranges from 0 to 3,000 in both bar graphs, A and B.

Then, we examined the VAF of the deletions to confirm if the high detection observed in the samples could be correlated with a higher VAF compared to the rest of the variants. In the analysis of the VAF profile (Figure 25), we observed that the vast majority have a VAF of less than 0.5. This profile, which is similar

in both groups, is an indicator that the mutations occur in a low fraction. Furthermore, the profile of mutations that we expected to find in our hypothesis was corroborated by these low results in the VAF.



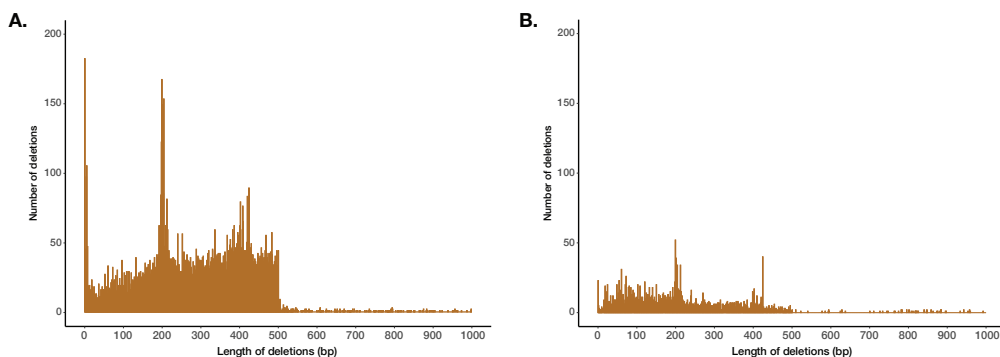
**Figure 25. VAF profile for the total of deletions on A) Adult sample and B) embryo sample.** X- axis represents the VAF that goes from 0 to 1, and the Y axis represents the number of deletions, that ranges from 0 to 2,500 in both barplots, A and B.

We further investigated the length distribution to determine if the deletions revealed a particular pattern in this area. The results (Figure 26) showed how the vast majority of the deletions were around 0-500bp. Within this range, two peaks of deletions were observed in the case of adults in the fields around 0-25 bp and 200 bp; this last peak was also present in the case of embryos. Both samples, with a higher degree in adults due to the accumulation of mutations along time, showed a drastic drop in the number of deletions around 500bp.

Based on our experience in the field of detecting somatic mutations in cancer, we presumed that some peaks may be due to methodological reasons (i.e., library size or read length), as some information is used by most callers to make predictions of mutations. Therefore we have compared our results with the ones reported by the same variant callers from Chronic lymphocytic leukemia (CLL) samples, that reported approximately a similar library size and

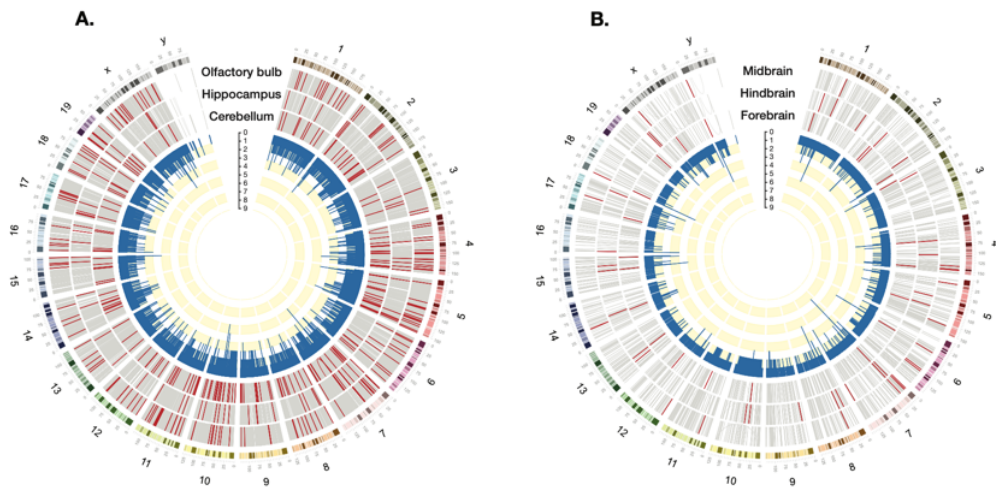


read length. A similar peak pattern was also observed in CLL samples. Therefore we confirmed their presence for methodological reasons.



**Figure 26. Length distribution for the total of deletions on A) Adult sample and B) embryo sample.** X- axis represents the length of deletions in pairs of bases (bp) that goes from 0 to 1,000; and the Y axis represents the number of deletions, that ranges from 0 to 200 in both barplots, A and B.

To detect if there was any over-mutated region in the genome, in addition to the distribution of the start of the deletions, we also studied the recurrence of the deletions across the samples (Figure 27). We could observe how there is a considerable accumulation of mutations showing a more significant recurrence in the sample of adults than in embryos. The highest value is six samples of the nine totals in the case of adults, while in embryos, we found a case that is up to seven samples out of the nine.



**Figure 27. Circular genome representation (Krzywinski et al., 2009) of the recurrency of deletions for each tissue on A) Adult sample and B) Embryo samples.** The red lines mark the deletions that we find in each of the three tissues that we analyze for each wild-type mouse. The blue histogram represents the total number of deletions detected across the entire genome in the different tissues, a total of nine. We can see how the number of variations is higher in adults than in embryos, as well as the recurrence of specific deletions, without reaching an event that is in the total of all tissues.

## 2.4 Study of *Pgbd5* KO mice vs WT mice

Following what we know about *Pgbd5* related to deletion variants we expected to find deletions that integrate two motifs on each point of breakpoint. For this reason we focused our analysis on deletions higher than 24bp.

Centring our attention on the length of deletions, we observed a significant difference between the distribution of the length of deletions in WT vs. KO mice, in adults and embryos. In the case of adults (Supp. Figure 2), we worked with a total of 27,856 mutations between the two KO and WT groups. We observed that there was a significant difference in the window range of 50-400 bp. This difference leads to an increase in the number of mutations in

favour of the WT samples. In the case of the embryos (Supp. Figure 3), the number of mutations studied was 8,313, which is lower than in the adult scenario. In this group, we found a significant difference in the window range of 150 to 550 bp. Unlike the adults, this difference was unexpectedly in favor of the KO samples.

In summary, through this analysis, we found significant differences in the total length distribution of deletions between WT and KO groups, and an enrichment of deletions with lengths around 200-300 bp following what was previously known about PGBD5-related deletions.

For the subsequent analysis of these deletions we decided to choose a length of deletion that would encompass these significant results for all the samples. Thus, we selected the consensus length of 150-400bp for both adult and embryo. With this criterion, we obtained: 9,149 mutations on adult wild-type, 7,440 mutations on adult KO, 1,625 mutations on embryo wild-type, and 3,172 embryo KO.

## **2.5 Identification and analysis of genes and genomic intervals**

In an attempt to elucidate the role of the selected rearrangements, we investigated the genes and genomic intervals that are somatically rearranged in wild-type versus knockout tissues.

On the side of genes, we first crossed the selected mutations with genomic annotations - NCBI genes (see methods section 2.2.4), using the intersect command from the bedtools suite. Furthermore, we studied those genes that are found exclusively in WT to determine if DNA rearrangement affects genes that occur recurrently in independent individuals and diverse brain regions. To do this, we filtered down genes that are rearranged in at least two out of

the three individual mice and loci that are rearranged in at least two out of the three brain regions. Through this analysis, we found genes that are exclusively mutated in WT and KO, for both groups , on the three scenarios (Figure 28).

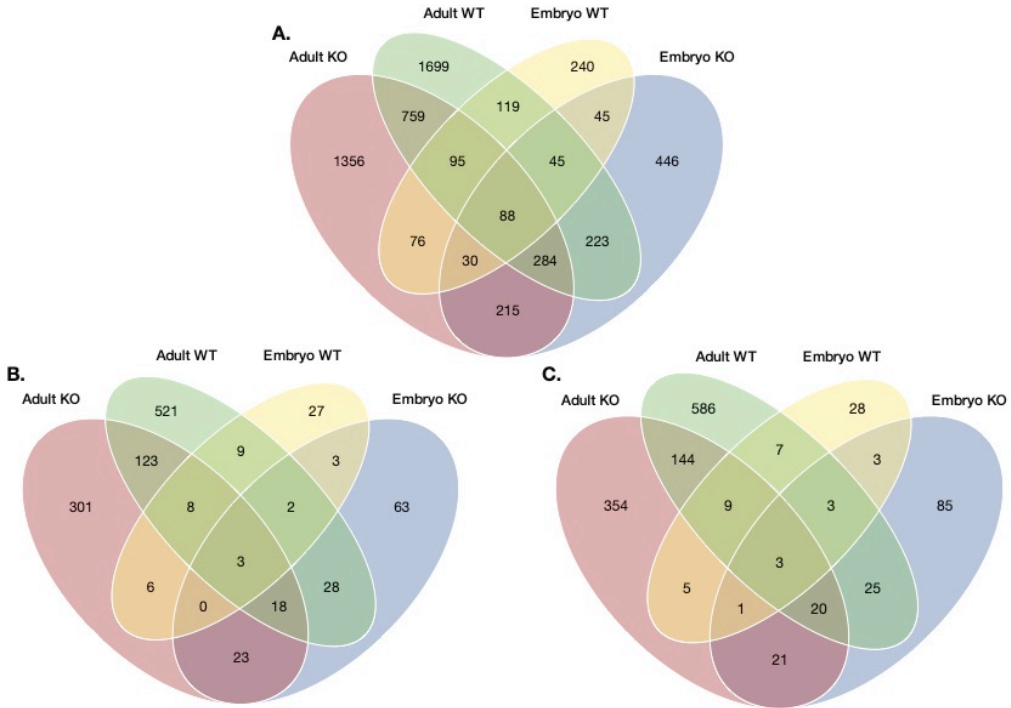
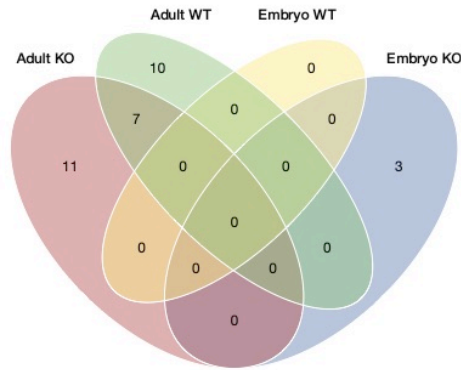


Figure 28. Venn diagram plots of mutated genes in A) Total samples , B) At least two mice ,and C) At least two tissues.

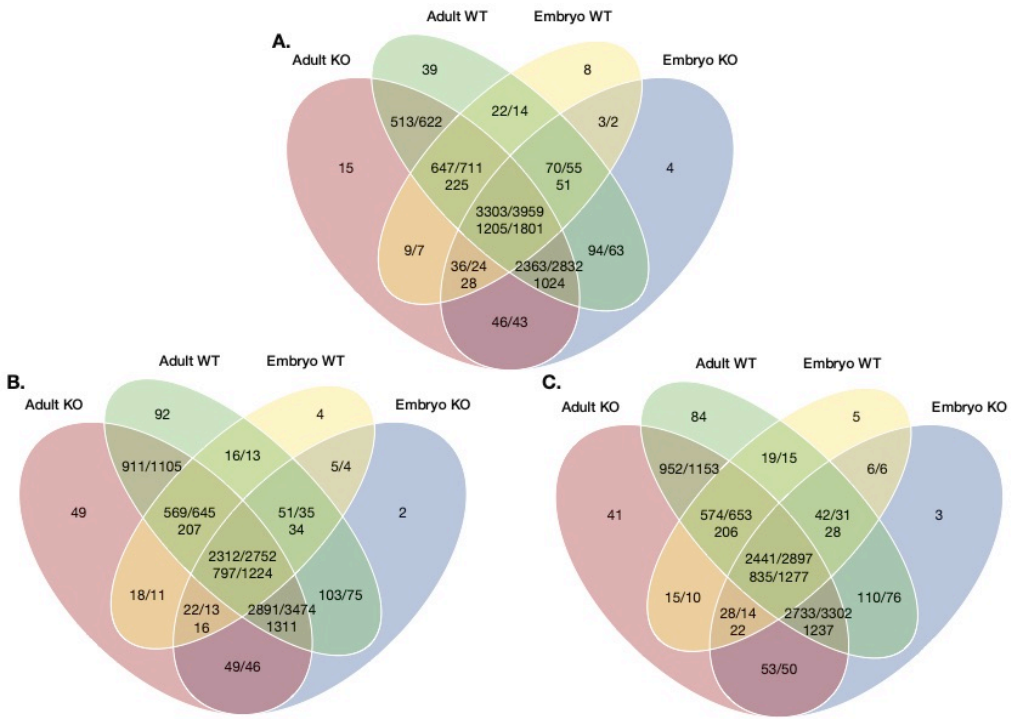
We focus the attention on those genes that have at least one rearrangement affecting their exons (Figure 29).

Only 31 mutated genes were found affecting coding regions. Among the ten genes with exonic rearrangements present in exclusivity on WT samples , we found four that were related to cell development (*Dazl*), nervous system development (*Ntn2*, *Traf3ip1*) and embryo development (*Prkra*).



**Figure 29. Venn diagram plot for mutated genes only affecting on coding regions**

On the side of genomics intervals, following the same procedure as in the previous section with the genes, we studied those genes that were exclusively found in WT. We also studied if its DNA rearrangement affecting genes occur recurrently in independent individuals and diverse brain regions (Figure 30). Through this analysis, we found genomic intervals exclusively mutated in WT and also KO, for both groups, and three scenarios. The regions have been created dynamically through the list of mutations contained in each group, with a static window size of 3Mb. For each mutation entry we had in the file we generated the window, and we observed how many mutations those windows covered.



**Figure 30. Venn diagram plots of mutated regions in A) Total samples , B) At least two mice and, C) At least two tissues.** The results for each group should be read clockwise. Example: graph A group Adult KO + Adult WT, have the values 513/622. This means that the result 513 belongs to the group Adult Ko and the 622 to the group Adult WT.

Below we turned our attention on those unique groups, both genes and of region intervals, for WT and KO within the adult and embryonic groups. With this, we observed not only the differentiation between WT and KO within each group, but also the unique characteristics of the embryonic groups that had continuity and were present in the adult group too.

The first objective was to look at the overall picture of how both genes and regions were distributed or linked to a particular tissue or mouse. The heatmap graphs were divided by tissue and mouse sample; thanks to the mutation rate described at the top of the graphs, we could verify that for both cases, genes

and regions, there was no clear relationship either by the tissue or by a particular mouse. It is worth mentioning that we didn't get any result for one of the adult KO mice olfactory bulb sample due to low coverage.

We then centered our attention on the top 20 genes and genomic regions of each group and studied their effect thanks to the variant effect predictor.

First of all, in the study of the genes (Supp. Figure 4 ; Supp. Figure 5 ; Supp. Figure 6 ; Supp. Figure 7) , we checked that when comparing the Adult and embryonic samples, both for KO and WT, there is no gene that is prolonged between the groups. In this scenario, we also analyzed the effect of the mutation thanks to the effect predictor of the We were able to observe that most of the mutations were in intronic areas. These also didn't cover any relationship with a particular tissue or mouse. Secondly, we studied the genomic region intervals (Supp. Figure 8 ; Supp. Figure 9 ; Supp. Figure 10 ; Supp. Figure 11). A key point to take into account for the evaluation of the results is that the coordinates are not the same across the groups due to the practice for the formation of the regions, discussed above. When we observe the regions highlighted in adults and in embryos, for each KO and WT group, we observed that in this case there are regions that are shared between them. In the case of the KO group, there are a total of 4 regions composed by: chr5:18,000,000- 21,000,000, chr8:20,000,000-24,000,000 , chr3:5,200,000-8,200,000 ,and chr16:54,000,000-57,000,000 . In the case of the WT group we see that they are a total of 5 regions composed of: chr9:28,000,000-32,000,000; chr5:6,216,000-9,216,000, chr3:140,000,000-143,000,000 , chr11:8,500,000-11,500,000 ,and chr5: 6,200,000-9,200,000.

Accordingly, we developed a functional analysis with the variant effect predictor to study regions scenario. The results showed that most of the regions were in intronic areas, and also not cover any relationship with a particular tissue or mouse.

## **2.6 Detection of motifs around breakpoints**

As mentioned in the previous study, deletions related to *Pgbd5* revealed significant enrichment of PGBD5-specific signal sequences motifs at the breakpoints of PGBD5-induced tumor structural variants. When we studied if the mutations that affected the above mentioned genes show any different sequence motifs at their breakpoints, no significant results were found.

## **2.7 Functional enrichment analysis of genes affected by *pgbd5*-dependent deletions**

To test if rearranged mouse brain genomic loci exhibited distinct *Pgbd5*-dependent and/or independent functional associations, we took the genes that are only mutated from each group and performed a functional analysis. Through this analysis, we found that there is a neuronal development gene ontology related to WT mutated genes in contrast with KO mutated genes, this scenario is only represented on adult samples.





DISCUSSION

Following the flow of the methods and results sections, the discussion has been also split into two blocks; the first one is related to the development of SMuFin2, an standalone reference-free strategy for the identification of somatic genomic variation; and the second is focused on the study of somatic DNA rearrangements induced by Pgbd5 during brain development and adult state. Following the same structure, this thesis contributes to science in two different ways: one corresponds to the generation and improvements of approaches for genome analysis, which can help the community to have a more accurate methodology for the incorporation and interpretation of genomic analysis in basic research and clinical practice. ; and the application of these methodologies to characterize and understand the landscape and the role of somatic variation in neuronal development in mammals.

# 1. Development of a computer-implemented and reference-free strategy for identifying variants in nucleic acid sequences

As described in the introduction, the identification and interpretation of the broad spectrum of potential variation within genomes, affecting from single nucleotides to large chromosomal rearrangements, requires a myriad of different variant callers. As each variation type and size impose different searching strategies through the sequencing reads. To this end, this PhD project proposes SMuFin2 algorithm as a comprehensive and highly scalable solution with the following advantages: (i) identifies a wider range of genetic variants in a single execution; (ii) achieves base-pair resolution; (iii) and detects the presence and insertions of non-human sequences, such as viruses.

This PhD includes the design and implementation of a computationally highly efficient reference-free strategy (described in Patent: A computer-implemented and reference-free method for identifying variants in nucleic acid sequences. NUM: WO 2018/007034) to process NGS reads from two different samples (states) to identify and isolate all potential changes between them in the form of aligned read blocks of approximately 250 nt long. This constitutes a highly efficient and independent module, where different algorithms can be plugged providing different types of outputs. For example, also as part of this thesis, a preliminary module has been added for the identification of somatic variants from the comparison of normal and tumor genomes. These two modules together constitute SMuFin2, which demonstrates great potential for the accurate identification of all types of

somatic variation, including the exact insertion position for cancer-related viruses. But the versatility that this strategy confers opens path to additional implementations in scenarios besides the identification of somatic genetic variants, by processing these blocks differently. For example, for the analysis of RNAseq data, a researcher would need to change the internal thresholds, and add a module to answer specific questions regarding the differential expression of genes, and/or the detection of different forms of splicing. This NGS analysis platform has easy access to tune and change the different thresholds and decisions made during the process, making the adaptation to other needs more feasible.

Of note, the development of this NGS read analysis platform and SMuFin2 represent an efficient example of close collaboration between computer and life-science groups. With the goal of developing an application with the intention of maximizing the efficiencies of algorithm and implementation, Jordà Polo (from the group of David Carrera, BSC) was devoted to the computational efficiency of the process, whereas my specific contribution to this project has been the design of the algorithm, and its validation in the form of SMuFin2 against real cancer and normal genome sequences. This close collaboration has proven to be crucial to address the development of applications that require a close interplay between implementation and algorithm.

As to computer efficiency, this analysis platform excels in minimizing time and energy consumption. Scalability is one of the main limiting factors in genomic research considering the current amount of petabytes of data that we are already facing up and even further, with the expected wide deployment of sequencing to millions of genomes that we will experience in the following decade. At this point scalability becomes from a highlight to a necessity in the life-science community, which is not addressed by many commonly used

programs for genome analysis. Consistent with this notion, SMuFin2 performance in a median sample patient with 30x coverage, using 8 nodes of Marenstrum 4, requires 14.7 CPU hours and 4.9 kWh consumption (section 1.3 SMuFin2 implementation). Variant identification in the complete PCAWG dataset (2,856 pairs of healthy-tumor samples) was performed on the group using 16 nodes of Marenstrum 3 and entailed a 425,280 CPU hours and 148,848 kWh consumption. The exact same analysis but implemented with the novel SMuFin2 algorithm generated in this project would have encompassed a total expense of 39,072.6 CPU hours and 13,024.2 kWh. Thereafter, SMuFin2 represents a giant leap in computational scalability by providing a 12-folds reduction in computational resources and, as SMuFin2, constitutes the most appropriate tool for genetic variant identification in large-scale consortiums. To facilitate the broad deployment of SMuFin2, we considered alternative scenarios besides cloud or HPC computing. As a result, it can be implemented and executed in a workstation computer, which is a relevant framework taking into account privacy policies in genomic data. Of note, anticipating the current trend to promote clinical translation of genomic research, our workstation version could be implemented in small laboratories or clinics. The computational efficiency of the methods used in variant identification is in response to the growing demand for genome analysis, a crucial factor not found in all current variant callers.

The strategy behind SMuFin2 is expected to contribute with additional advantages for the identification and classification of somatic variants, compared with other available variant callers. On one hand, SMuFin2, although it accepts aligned BAM files as input, uses the reads directly, as in FASTQ files. In this way, the method is not subjected to any pre-aligned file with an specific format or program performance (in this case BWA for the BAM file), which makes it less dependent on changes of these alignment

programs, as it happened with the most recent of the BWA program, with changes that required adjustments in all reference-dependent variant callers. In addition, this pre-alignment step involves further limitations: (i) it requires time and computing resources. On average, BWA takes approximately 2 days to align a 30x NGS sequenced whole genome, using 4 CPUs (Yung et al., 2017), (ii) the state-of-art and most widely tool used for pre-alignment is BWA, and hereby all variant callers, despite of their particular features, suffer from the same alignment errors, limiting the availability of alternative strategies to combine with, and to increase the efficiency of approaches that use multiple variant callers. (iii) the alignment of reads containing somatic or polymorphic variants relative to the reference genome, are expected to align with low quality or to be misaligned, depending on the type and size of the variant. In particular, reads with structural variants are poorly mapped by BWA, compared with reads covering one single nucleotide variant. (iv) Following the previous point, non-human DNA sequences (viruses, for example) present on the sequenced genome, will not be aligned to the reference genome, and will not be considered in the study, unless is specifically captured. On the other hand, reference-free methods do not depend on prior pre-alignment steps and therefore are not subjected to these limitations.

In order to gain detection and classification resolution, the strategy behind SMuFin2 pays particular attention to the alignment region where the variant is represented. Having a complete and properly aligned catalogue of the reads around a given variant is key for its identification and proper classification (Figure 17). Alignment is performed considering the consensus sequence of the block against that of a single read. This gain in alignment quality occurs, not only for the blocks that contain mutations in regions that belong to the reference human genome, but also for those not present in the reference genome, mainly corresponding to viral genomes. Of note, considering the large

genetic variability in viruses, this advantageous strategy sets SMuFin2 apart from its competitors.

Indeed, SMuFin2 shows a high potential performance in the identification of virus presence and insertions as our preliminary analysis shows (see 2.1 of the results section). In both *in-silico* genome and in Pan-Cancer data, SMuFin2 was capable of delineating the inclusion of a virus genome but also provided the actual insertion point in the same single execution on an in-silico model. To detect viruses, SMuFin2 does not use reference virus genomes as an additional step. Insertion detection is done directly, and then those blocks that do not align with the human reference genome are aligned against the virus database of the user's selection. This approach avoids two of the most common issues reported in virus identification tools: (i) the additional alignment of the entire sample against a reference genome containing the virus sequence, or the creation of a new reference genome with all the viruses, as it is required for VERSE (Wang et al., 2015) ; and (ii) the viruses in the real sample have to be quite similar to the reference sequence of the viruses, and given their high variability, they will not always be able to be detected, this also includes the possibility of detecting new viruses that are not contained in the database.

Based on the last big study published by the working group of pathogens within the PCAWG (Zapatka et al., 2020), the strategy used was based on two major steps: (1) Virus Detection, using three independently developed pathogen detection pipelines that rely on the: 'Computational Pathogen Sequence Identification' (CaPSID) (Borozan et al., 2012), 'Pathogen Discovery Pipeline' (P-DiP) b (<https://github.com/mzapatka/pdip>) and 'Searching for Pathogens' (SEPATH) to generate a large compendium of viral associations across 38 cancer types.; (2) Virus integration sites analysis: A subset of viral candidates identified to be present in tumor samples by the CaPSID analysis pipeline was selected for the detection of viral integration events using the



VERSE algorithm. SMuFin2' functionality would cover and provide additional information within these two steps, incorporating new ones. For example, SMuFin2 can use the original sample without having to first filter for those reads misaligned with the reference genome. Importantly, in general, SMuFin2 would fulfill the two steps of the strategy in a single execution, by evaluating the hash table of SMuFin2, we can directly detect reads (k-mers) in the tumor sample that have no match on the normal genome, likely corresponding to non-human sequences. This allows us to potentially identify the entire viral sequence that has integrated adn, by adding additional features, be able to assemble those fragments into complete integrated (viral) sequences. On the other hand, manual inspection of the results of the preliminary module for block processing, show that the identification of reads that contain human and non-human k-mers allow are informative of the exact insertion point within the genome. Therefore, SMuFin2 is potentially useful to complement current efforts towards the characterization of viral integrations in tumor genomes and to investigate their role in tumor formation and progression.

The organization of the data within SMuFin2, in the form of hashtables, results in highly efficient storing of the k-mers and their corresponding reads. This format enables us to save other read features along with the read id. The analysis of the structure of these hashtables, beyond providing direct variant candidates in the form of blocks, can also provide additional information regarding the sample. A prevalent scenario is that researchers have an imprecise knowledge about the coverage and sequencing errors, as these are strongly depending on the quality of the biological sample. This information is usually obtained using external software that is based on a pre-alignment step of the sample, such as BWA or Alfred (used in the mouse project, see Section 2.1). By calculating the frequency of occurrence of k-mers across the entire hashtable, SMuFin2 is able to have a quite precise estimate of

the real coverage of the sample. Moreover, we can estimate the sequence error rate of a particular sample by counting the fraction of unique k-mers, which is expected, in most cases to correspond to sequencing errors. This estimation will also be dependent on the sample, as tumor samples with a high degree of heterogeneity can have a larger fraction of variants corresponding to low cell fractions and can be taken as sequencing errors. These observations not only give a much comprehensive overview of the input data being analysed, but they are also crucial to apply the most appropriate filters to precisely identify the genetic variants that we are interested in.

Potential users of SMuFin2 cover a broad range of expertise in computational analysis. We addressed this by generating a versatile set of output results that should attain the needs of our diverse candidate users. On one hand, we provide output files in the formats accepted and used within the community, e.g.VCF format. Thus, the results can be easily compared with the rest of the variant callers available in the field. As a very useful feature, SMuFin2 is able to also generate an equivalent output file in html format, reconstructing the region that contains the somatic variant, as well as other types of sequence information, counters, etc. that enabled its particular identification. This provides an intuitive and alternative way of evaluating the quality and confidence of the variant call, which is particularly relevant in clinical contexts.

In summary, SMuFin2 excels in enlarging the landscape of genetic variants identified without computationally compromising the viability and costs of the analysis, that suits for current and upcoming large volumes of sequencing data within biomedical research and, slowly, also within the clinical practice. In this last environment, the low cost, as to time and computing resources provided by SMuFin2 is also crucial to give a rapid clinical response to the patient. Furthermore, Introducing a new strategy for the detection of somatic variants makes SMuFin2 well suited for evaluation for greater accuracy for

overlapping calls. The recall of overlapping calls varies depending on the combinations of the specific algorithms and not the combinations of the methods used in the algorithms (Kosugi et al., 2019).

## 2. Landscape of somatic variation in neural development and the role of *Pgbd5*

Non-cancer-related somatic mutations that occur during development may affect cell proliferation, as is the role of cancer, or may alter cell function without causing a proliferative or any other nocive effect. Recent literature suggests that somatic mutations might also occur during brain development without resulting in a disease status. This poses the question of whether this somatic genetic diversity could foster functional diversity among brain cell-types. Although it is known that somatic variation can play a role on neuronal development no extensive analysis has been done to characterise the landscape of somatic variation in mammal neural tissue. In this thesis we have covered part of this area by studying the landscape of variations in mammals on neuronal development, in order to understand, in particular, those arising from the activity of the *PGBD5* gene.

The characterisation of somatic variants in neural tissues is particularly challenging, when compared to the tumor genomes. This is due to the expected high tissue heterogeneity, and the consequent low coverage (VAF) of somatic variants, which are expected to be represented in low fraction of cells. By combining extensive sequencing and deep variant calling analysis we have partially overcome these limitations.

We first ensured that this methodology was able to detect different types of somatic variants from healthy, non-tumoral, tissue. Although our methodology does not capture a large fraction of somatic variants, represented by undetectable low cell fractions, we were able to identify an extensive

catalog of somatic variants, likely representative of the entire landscape of variation within this tissue. The variant calling and classification results on wt samples, as representative of the landscape of somatic variation in neural tissues in mice, show a wide range of types of variants, covering from indels, to large structural variants (analysis for SNVs is, at this moment, in progress). From the comparison of the different fractions of variant types, adult and embryo show deletions, as the most predominant type of variant with (86.96%) and (73.36%) of all the detected variation, respectively. The types of variants are also different in the two groups. In the case of adults, we detected insertions (5.45%), duplications (3.95%), and inversions (2.32%); while in embryos, we have duplications (19.52%), inversions (4.64%), and insertions (1.77%). It is important to note that the percentage of duplications that we find in the embryo samples is 5 times greater than what we find in the adult samples, being the number of mutations higher in embryos (929) than in adults (749) when we have previously mentioned that the number of total mutations was higher in the last ones. These results start giving us an overview of the landscape of somatic variation in neural tissues, and the differences between embryos and adults. For example, the number of mutations was three-fold higher in adult tissue samples (32.190) than in embryos (9.795). This phenomenon may be due to the accumulation of somatic variations over time that the adult tissues have undergone. Further studies are needed to evaluate whether these differences imply that the mechanisms for genome remodeling during embryogenesis are different from those present in later stages of the organism. Alternatively, this difference can also derive from our detection possibilities, i.e. to the different forms of clonal expansion of different cell types, accumulating specific forms of somatic variation. After having an overview of the general landscape of somatic variation in neural tissues, and the potential differences between different stages of the organism, our next goal was to determine which fraction of this variation was due to the activity

of *Pgbd5* gene, using the comparison between the *Pgbd5*-Knock-out and wt samples.

We focused on the deletions for the comparison of wild-type and knockout samples to determine the role of *Pgbd5* in the catalog of somatic variation. This decision was due to the predominant profile presented in both adult and embryonic samples in the WT group, and the previous study related to *Pgbd* activity 5 (Henssen et al., 2017b) of the type of variant associated with its activity. As the first exploration on wild-type mice, we observed that a priori there is no significant difference between tissues or samples (Figure 22). With this, we conclude that the deletions are not associated with a specific tissue, and we corroborate that no mouse specimen causes a deviation of these results. Regarding the VAF profile (Figure 23), it was similar in both groups. The majority of the deletions were below the value of 0.25, which corresponds to the pattern we expected from somatic variability in non-tumoral tissues. We then investigated the length distribution and concluded that the profile in both groups was similar, with a drastic drop in the number of deletions of 500bp (Figure 24). This last, we could confirm that it was due to methodological causes because a similar peak pattern was also observed in CLL samples that reported approximately a similar library size and read length. On the other hand, we did not find any region of the genome that was highly recurrent (Figure 25).

On the Study of *Pgbd5* KO mice vs WT mice, we observed statistically differences in the total distribution of the length of deletions between WT and KO. Surprisingly, these differences were opposite in the groups. While in adults the increase in variation was in favor of WT (Supp. Figure 2), in embryos it was just the opposite and was in favor of the KO group (Supp. Figure 3), in a similar windows length. Also, we found an enrichment in deletions of around 200-300 bp, in concordance with preliminary data available for *PGBD5*-related

deletions (Henssen et al., 2017b) For the subsequent analysis of these deletions we decided to select a consensus length of deletion of 150-400bp for both adult and embryo.

With this group of significant deletions, we identified those genes affected by the variations and those genomic intervals with more presence of variations for each one of the groups (Adult-KO, Adult-WT, Embryo-KO, and Embryo-WT). We were able to identify those that were exclusive to each group, as well as to analyze those that were common among the groups and see their dynamics. Note that of the total only 31 mutated genes were found affecting coding regions. Among the ten genes with exonic rearrangements present in exclusivity on WT samples, we found four that were related to cell development (*Dazl*), nervous system development (*Ntng2*, *Traf3ip1*) and embryo development (*Prkra*). Even follow up analysis was carried out to identify the mutated regions and genes, no correlation was found among the deletion profiles and any particular tissue or mouse.

Once the deletions of interest were selected, further studies focused on the functional analysis to better understand how the underlying biological processes correlated to the mutations. Accordingly, we selected the genes that were only mutated from each group (Adult-KO, Adult-WT, Embryo-KO, and Embryo-WT) (Figure 26). We found that there is a neuronal development gene ontology related to the genes that are mutated in the WT samples in contrast with the genes mutated for the KO. Interestingly, this scenario is only represented on adult samples and not in embryos. This phenomenon may be due to the fact that when embryo samples are collected, they are in the primary development stage, where the cells are actively dividing and the amount of cells increases drastically. Thus, many of the cells carrying the mutations observed in the adult mice are actually eliminated on embryos during this stage. It is interesting to highlight that the genetic classification that

accompanies the production of neuronal cells during embryogenesis is part of the neuronal selection process.

Taking the whole analysis as a unit, we cannot decide on either of the two models intended for such a physiological process. Model 1; the mechanism would be analogous to RAG1/2-mediated rearrangements of the immunoglobulin receptor genes, where Pgbd5 would induce deletions or inversions of exons in one or few genes, leading to the production of new exon-exon junctions. Model 2; Pgbd5 would act on a diverse set of genetic loci, including genes or intergenic regulatory sequences, which would be identified by the presence of conserved sequence features. The second model is analogous to Barbara McClintock's Activator-Dissociator transposition mechanism, and may indeed involve specific mouse sequences that are mobilized or rearranged by Pgbd5 (Comfort, 2001) (McClintock B, 1947). The genomic loci, genes and SV features may be diverse, but would all share a common set of DNA sequence substrates, such as for example transposon inverted terminal repeats (ITRs).

In summary, the objective of the detection and analysis of somatic DNA rearrangements induced by Pgbd5 has been fulfilled with promising results. Dr. Kentsis group is currently re-sequencing a list of candidate genes and regions in more than ten new samples to confirm the findings and gain statistical power. In particular, next steps will include the design of a custom Nimblegen hybrid capture probe set to re-sequence the samples at high depth, as well as to continue working with RNA-seq results to see if there is any correlation with the above described rearrangements.

Considering the promising results that we obtained with the deletions, we decided to expand the spectrum of mutations to be analyzed and not just concentrate on the deletions. Another question that remains unclear is why



the mice brain exhibits more somatic rearrangements than control tissue (blood/liver). To answer this question we will consider running the same variant callers but inverting the cases by controls, which would allow us to verify if the neuronal tissue presents a higher number of mutations than the blood samples. This could be due to two hypotheses: i) methodological because we cannot detect it due to its low cell-fraction, or ii) biological, because one of the tissues has less amount of variations. Normal preliminary results for blood sample deletions indicate that we have been able to corroborate that the number of deletions identified so far is lower than in neural tissues.

At the same time, by carrying out a characterization of all the samples (cases and controls), we will be able to confirm that all the mutations that we find in the samples of neural tissue are exclusive to this tissue, and we have no false positives in our results. In this approach, the procedure we will start by characterizing all the tissues. Following this, we will validate if the variations detected in the neuronal samples do not contain false positives. Once we have the variant landscape, we will reanalyze all types of variants with the same pipeline shown in the results section to identify all those variations that may be significantly associated with Pgbd5 activity.

CONCLUSIONS

## **Development of a computer-implemented and reference-free strategy for identifying variants in nucleic acid sequences**

(i) A reference-free and scalable algorithm, called SMuFin2, has been developed for the identification of somatic variants in tumoral samples.

(ii) SMuFin2, a reference-free based strategy, is scalable and highly efficient. It can be implemented into the normal research activity of cancer genomics, in contrast to what was believed due to the high computing demand generated by current sizes of data and expecting larger datasets.

(iii) The preliminary tests done with SMuFin2 have shown the potential to detect a wide range of somatic variation, including insertions of non-human DNA on tumoral samples.

## **Landscape of somatic variation in neural development and the role of *Pgbd5***

(i) The application of current tools for the identification of somatic variation in cancer can be applied to study somatic physiological modifications in neuronal tissues.

(ii) WT and *Pgbd5* KO present a difference in the total distribution of the deletion lengths, increasing in number of mutations in adult mice and embryos, respectively.

(iii) The study of genetic ontology on selected genes shows a neuronal development gene ontology related to WT on Adult samples.

(iv) Taken together we identify a *Pgbd5* dependent somatic activity in different neural tissues.

SUPPLEMENTARY  
MATERIAL

-----  
[core]

# Length of k-mers, which is the size of the substrings that reads will be  
# split into in order to be analyzed. The recommended value is in the range of  
# [28, 32]; currently only k-mers of up to length 32 are supported.

k = `__K-MER_LENGTH__`

# Partitioning is the scaling mechanism that allows distributing the  
# computation, splitting data into different chunks that can be processed  
# independently. Partitions can be adapted to run sequentially or in parallel,  
# in a single or multiple machines. While increasing the number of partitions  
# lowers the peak amount of memory, it also increases the amount of duplicate  
# data during the filter stage, which in turn may incur in slower merging.  
#

# «num-partitions» represents the total number of partitions, while «pid» is  
# the current partition that will be processed in a particular execution, in  
# the range [0, num-partitions).

num-partitions = `__NUM_PARTITIONS__`

pid = 0

num-loaders = 8

num-storers = 16

num-filters = 28

num-mergers = 16

num-groupers = 16

# Input format for normal and tumoral samples. Two formats are available:

# - fastq: gzipped FASTQ files (recommended)

# - bam: aligned BAM files with corresponding BAI index (experimental).

input-format = bam

# Paths to normal and tumoral input files. For multiple files, wildcard

# expansion is supported, e.g. «file-\*.fq.gz».

#filter quality check

input-normal = /path/to/normal/input/files

input-tumor = /path/to/tumor/input/files

output = /path/to/output/dir

data = ./data

check-quality = false

[prune]

# Desired false-positive (FP) probability for both bloom filters, «all» and  
# «allowed». Lower FP rates involve a higher number of hash functions to be  
# calculated, which translates into additional computation to create and  
# access the bloom filters. Note that this is only used as a performance  
# tradeoff; a lower rate doesn't have any impact in the results since later  
# filters address and discard FPs.

false-positive-rate = 0.05

# Number of expected items in the bloom filters. Higher capacity translates

```
# into additional memory. The «all» bloom filter should be approximately an
# order of magnitude larger than the «allowed» bloom filter.
all-size = 100000000000
allowed-size = 10000000000
```

### [count]

```
# The count cache keeps track of k-mers that are seen only once so as to not
# to
# include them in the tables, reducing the overall memory footprint. It's
# enabled by default and recommended when running standalone counts, but it
# can be disabled when running with prune.
enable-cache = false
```

```
# Total number of expected items in the cache and table; generally speaking,
# the cache contains stems seen once or more, while the table contains stems
# seen more than once (so it's smaller). Sizes may need to be adjusted
# depending on the input so as to not to over or under-provision the memory.
# E.g. an input with ~4,250 million 80bp reads with a coverage of 60x
# requires a cache of size 106240000000 and a table of size 128000000000.
cache-size = 106240000000
table-size = 128000000000
```

```
# Limit exported rows to a particular subset of k-mers that match the following
# minimum/maximum frequencies. That is, either the normal count or the
# tumoral
# count of the k-mer is strictly greater than «export-min» and less than
# «export-max» (both excluded).
export-min = 0
export-max = 131072
max-conversions = 4
output = __OUTPUT_COUNT__
prefilter = true
```

```
# Format used to store filtering indexes. Two kinds of formats are supported:
# - plain: In-memory hashtables that are dumped to disk as simple
# space-separated plain text files.
# - rocks: RocksDB-backed databases, optimized for writing, then compacted for
# later stages.
index-format = rocks
```

```
# Number of indexes built for each partition. Increasing the number of indexes
# can speed up the filter stage, potentially slowing down the merge stage.
# Should be smaller or equal to the number of filter threads,
# «core.num-filters».
num-indexes = 2
```

```
# Candidate k-mer filtering/selection based on imbalanced absolute counts with
# the following criteria: at most «max-normal-count» normal k-mers, and at
# least «min-tumor-count».
max-normal-count-a = __MAX_NC_A__
min-tumor-count-a = __MIN_TC_A__
max-normal-count-b = __MAX_NC_B__
min-tumor-count-b = __MIN_TC_B__
```

```
# Maximum number of reads per k-mer; k-mers seen in more than «max-reads»
# different reads are discarded when building the filter indexes.
max-reads = 2000
output = __OUTPUT_FILTER__
```

```
[merge]
output = __OUTPUT_MERGE__
```

```
[group]
# Groups are generated after finding "leader" reads using a window-based
# technique. A read becomes a "leader" if it contains at least «window-min»
# candidate k-mers in a window of «window-len» bases.
window-min = __WINDOW_MIN__
window-len = __WINDOW_LEN__
```

```
# Maximum number of reads per k-mer; reads from k-mers with more than
# «max-reads» reads are dropped from the groups file and marked as such in
the
# results. Note that only reads are dropped, no k-mers will be removed. This
# value should be lower than «filter.max-reads».
max-reads = 2000
```

```
# Estimate number of candidate lead reads, which identify groups. For best
# performance, this value should be slightly higher than the actual number of
# candidate leads.
leads-size = 12800000
output = __OUTPUT_GROUP__
```

```
[rocks]
# Number of RocksDB background threads. High priority threads flush
memtables
# to disk, while low priority threads compact sstables.
num-threads-high = 10
num-threads-low = 10
block-cache-size = 12884901888
block-size = 16384
```

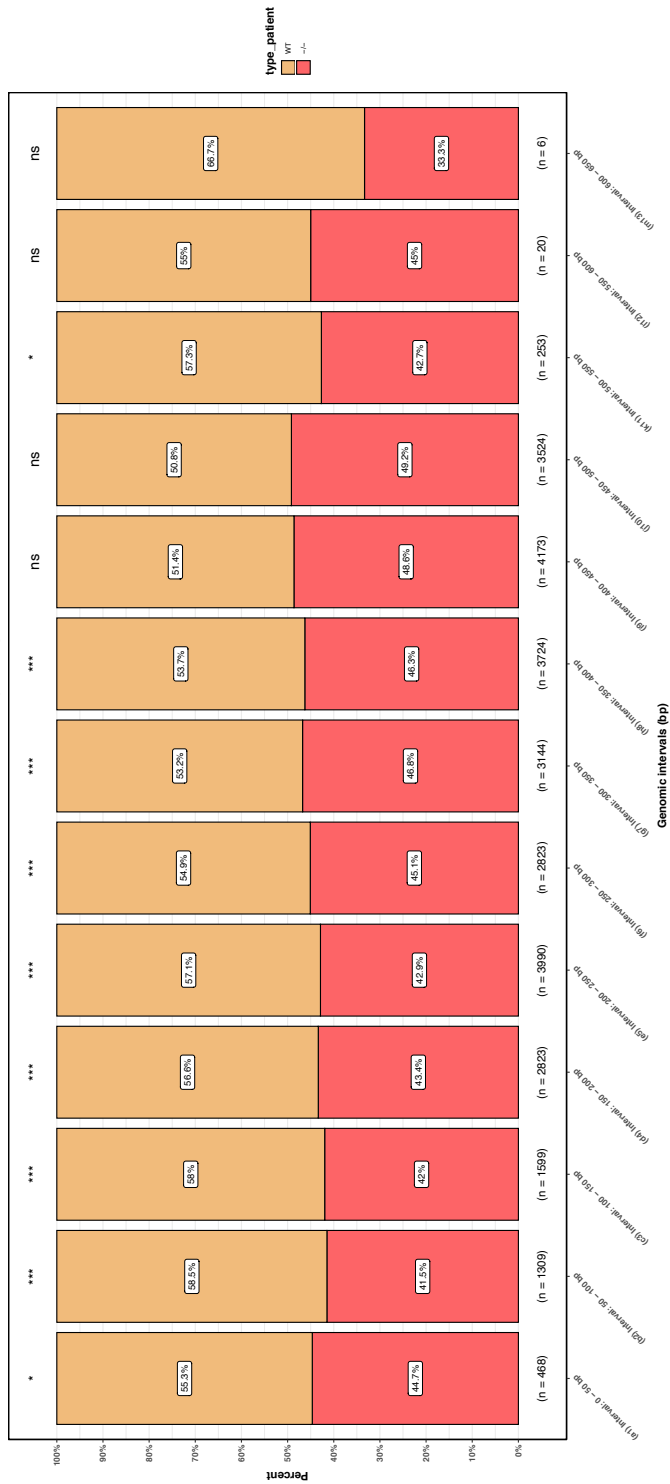
```
# vim: ft=dosini
```

-----  
**Supp Figure 1. SMuFin2 configuration file.** All the paths and variables are configured in advance to facilitate the user.



Comparison of number of deletions per length between WT and  $\Delta$ -ADULTS. Delly2+Pindel join PASS

$\chi^2(12) = 72.66, p < 0.001, V_{Cramer} = 0.05, CI_{95\%} [0.01, 0.02], n = 27856$

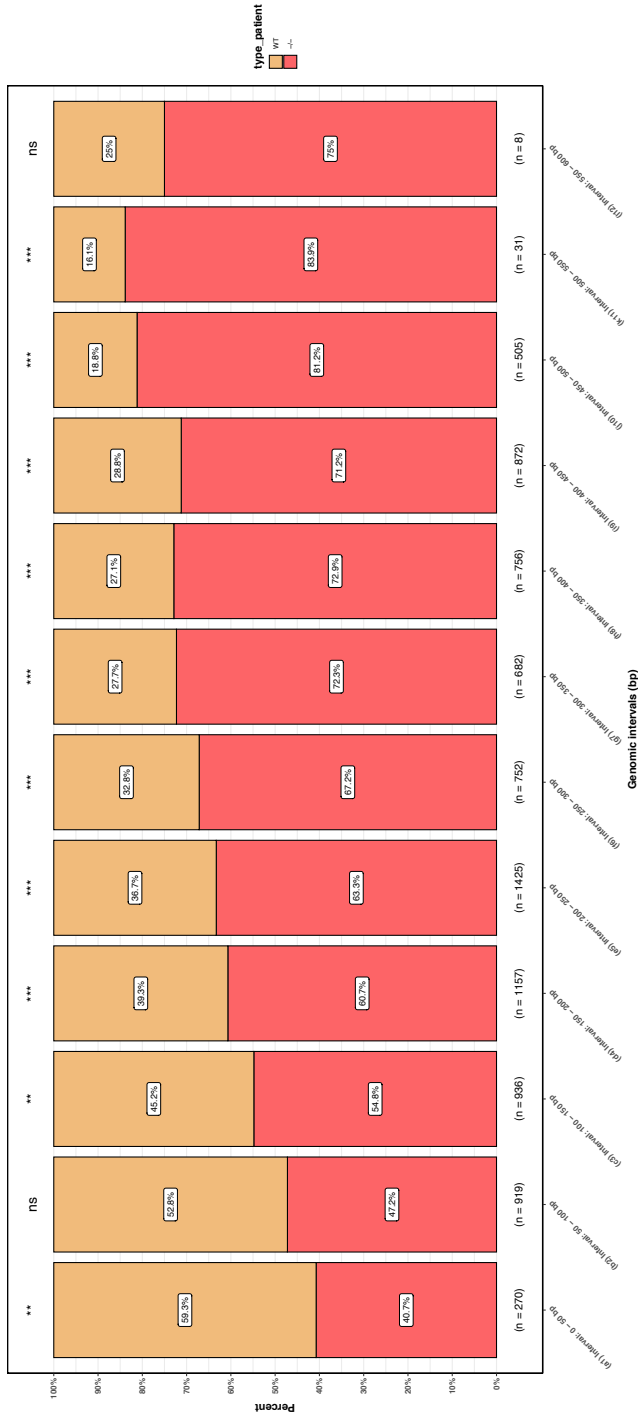


\*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ , ns = not significant

Supp. Figure 2. Comparison of number of deletions per length between WT and KO Adults samples.

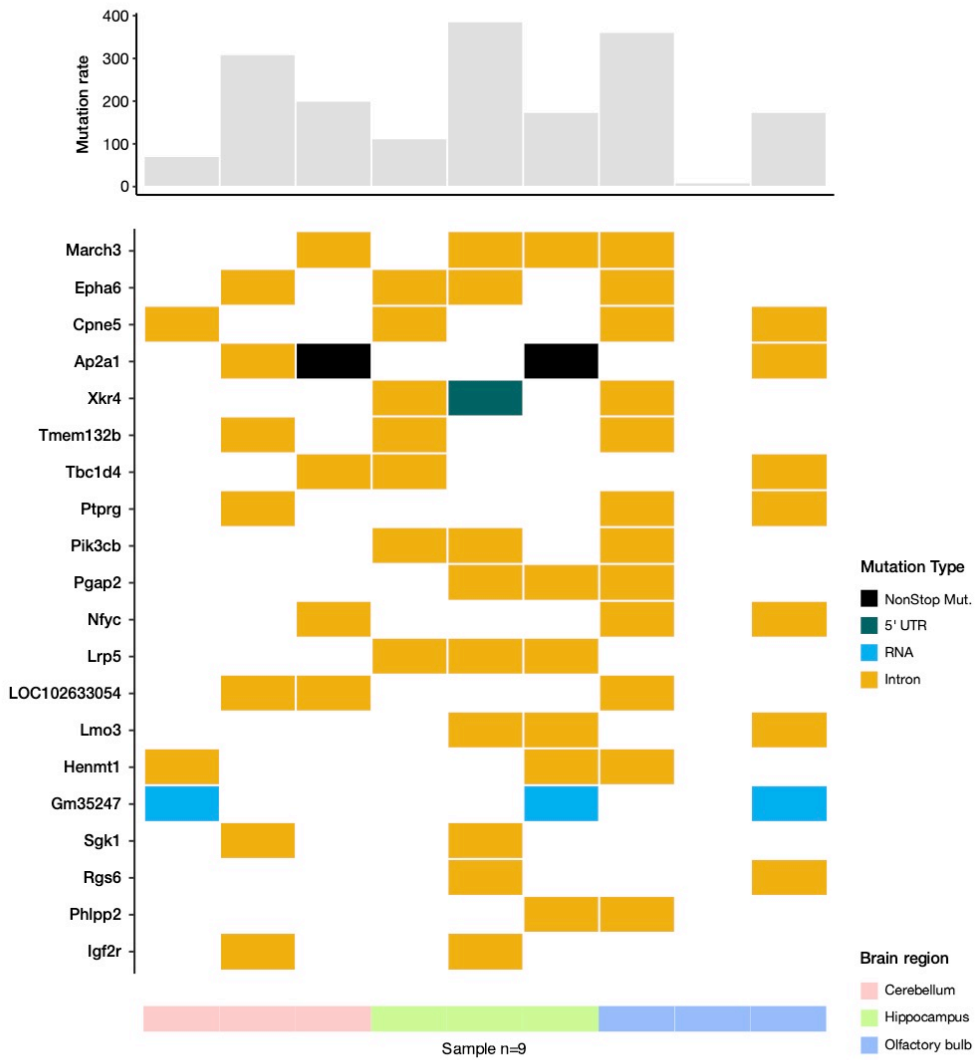
Comparison of number of deletions per length between WT and  $\Delta$ -, EMBRYONARY, Delyz+Pindel join PASS

$\chi^2(1) = 351.89, p < 0.001, V_{\text{Cramer}} = 0.21, C_{\text{adj}} [0.06, 0.07], n = 8313$

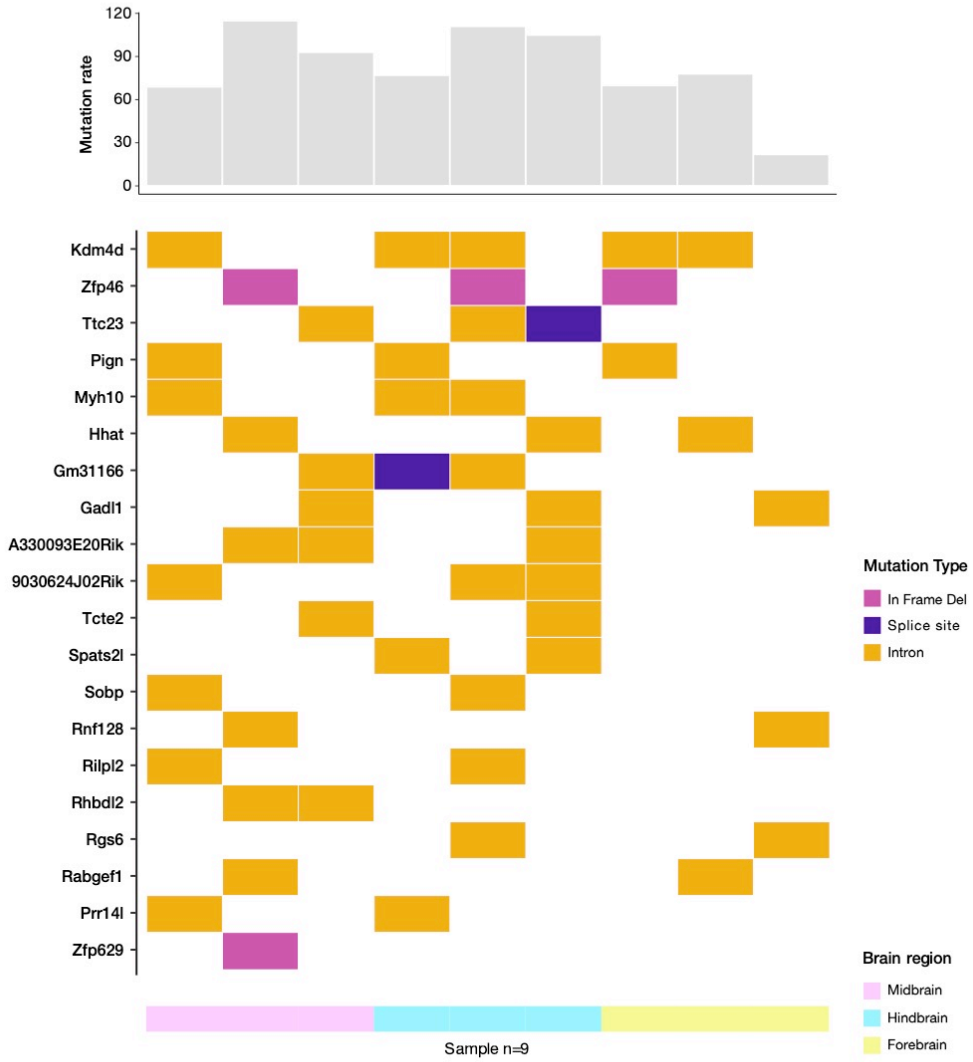


Number of nulling[95%] = 14346 samples - 14346 nulling[95%] = 14346

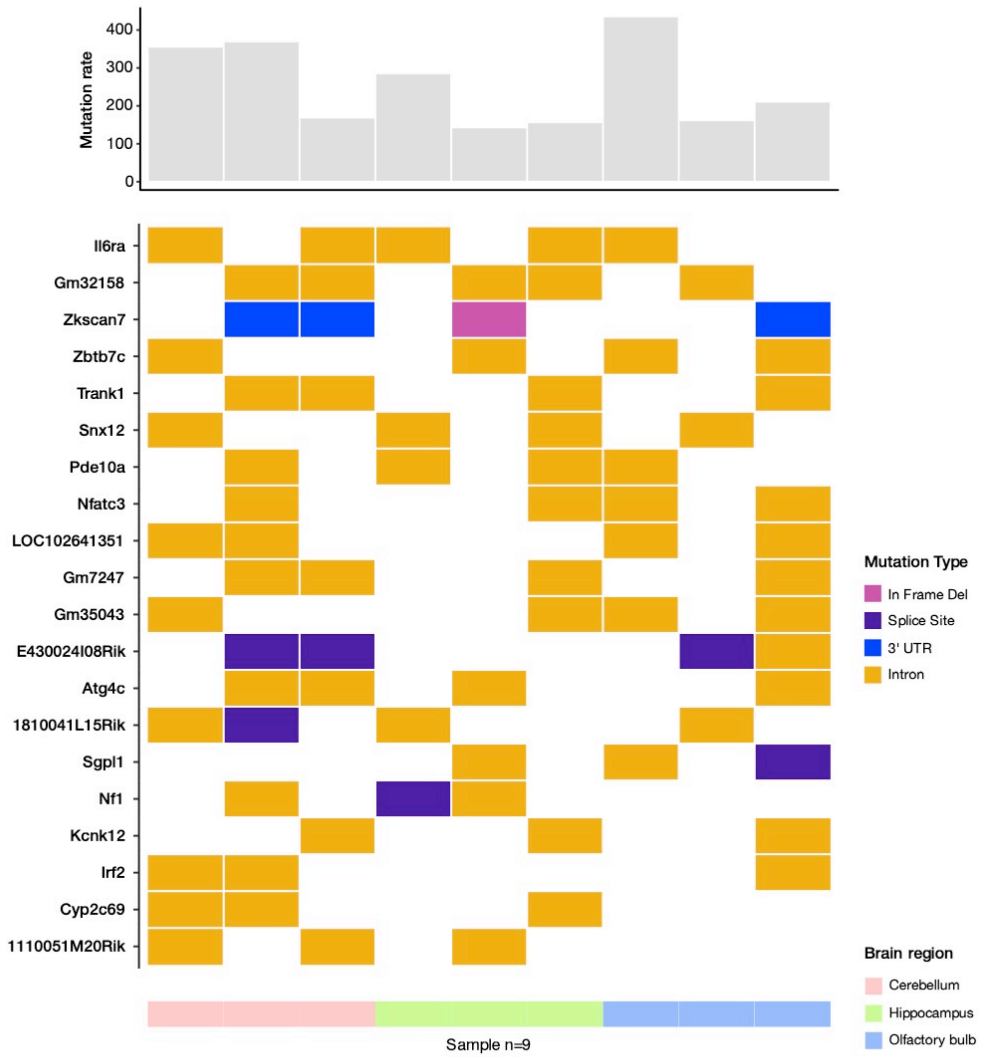
Supp. Figure 3. Comparison of number of deletions per length between WT and KO embryonic samples.



**Supp. Figure 4. Mutation rate and Heatmap for Top20 genes mutated only in Adult KO.**



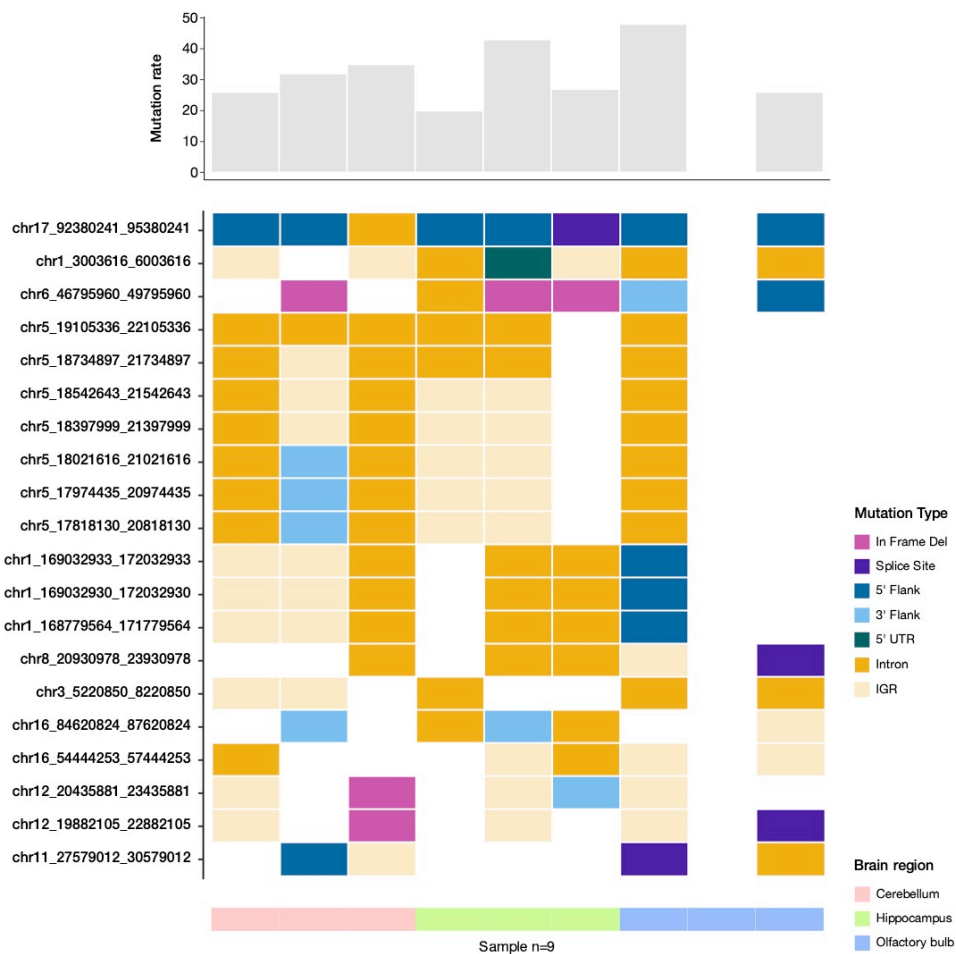
**Supp. Figure 5. Mutation rate and Heatmap for Top20 genes mutated only in Embryo KO.**



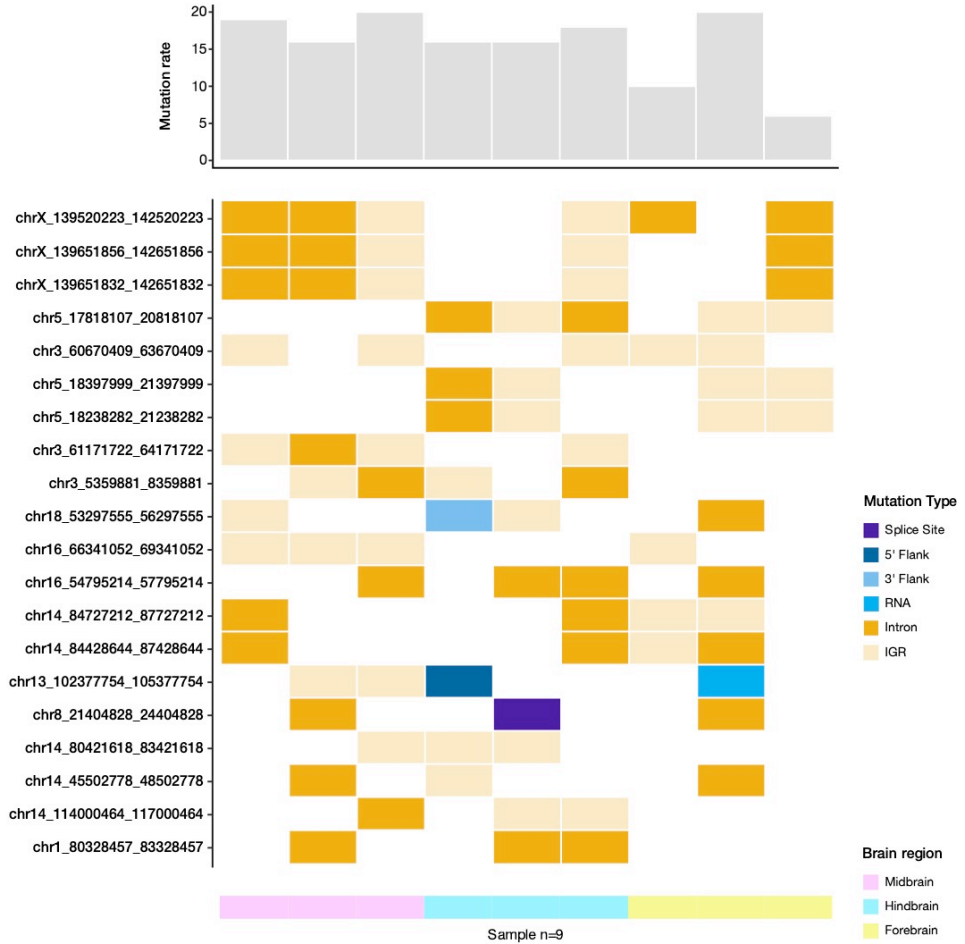
**Supp. Figure 6. Mutation rate and Heatmap for Top20 genes mutated only in Adult WT.**



**Supp. Figure 7. Mutation rate and Heatmap for Top20 genes mutated only in Embryo WT.**

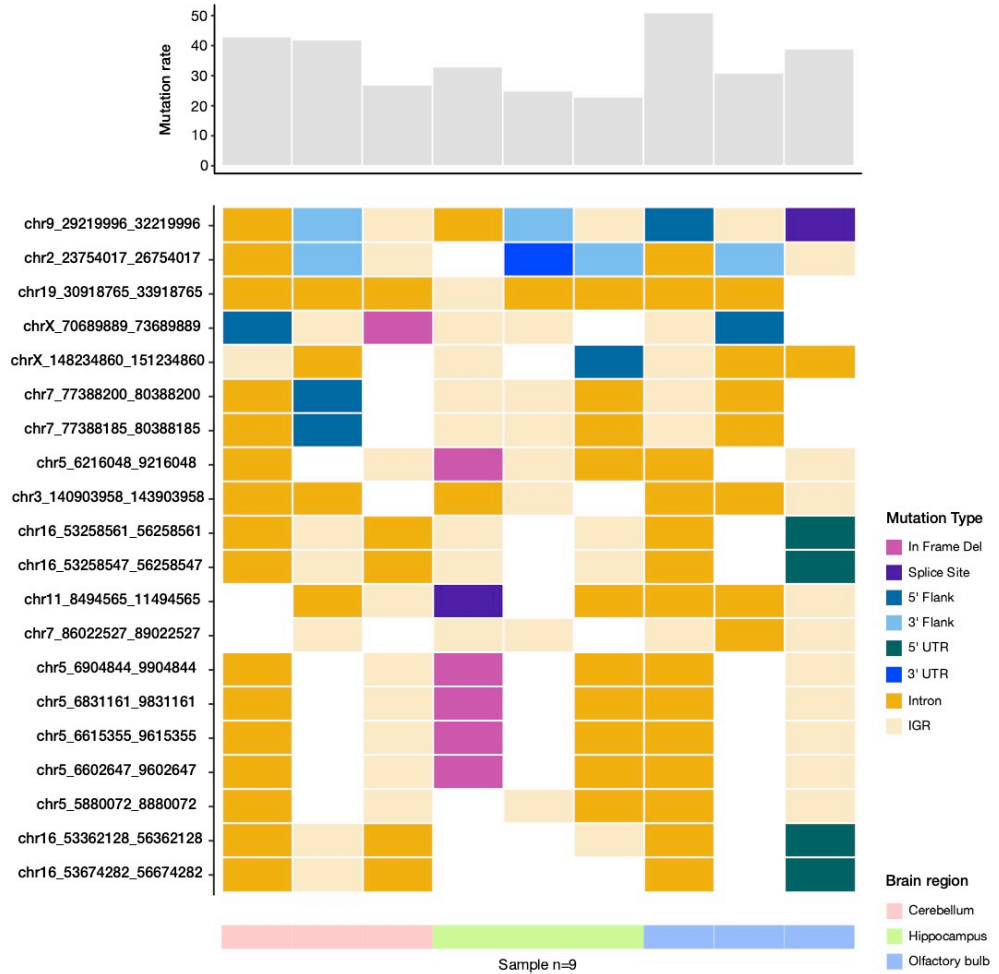


**Supp. Figure 8. Mutation rate and Heatmap for Top20 mutated regions only in Adult KO.**

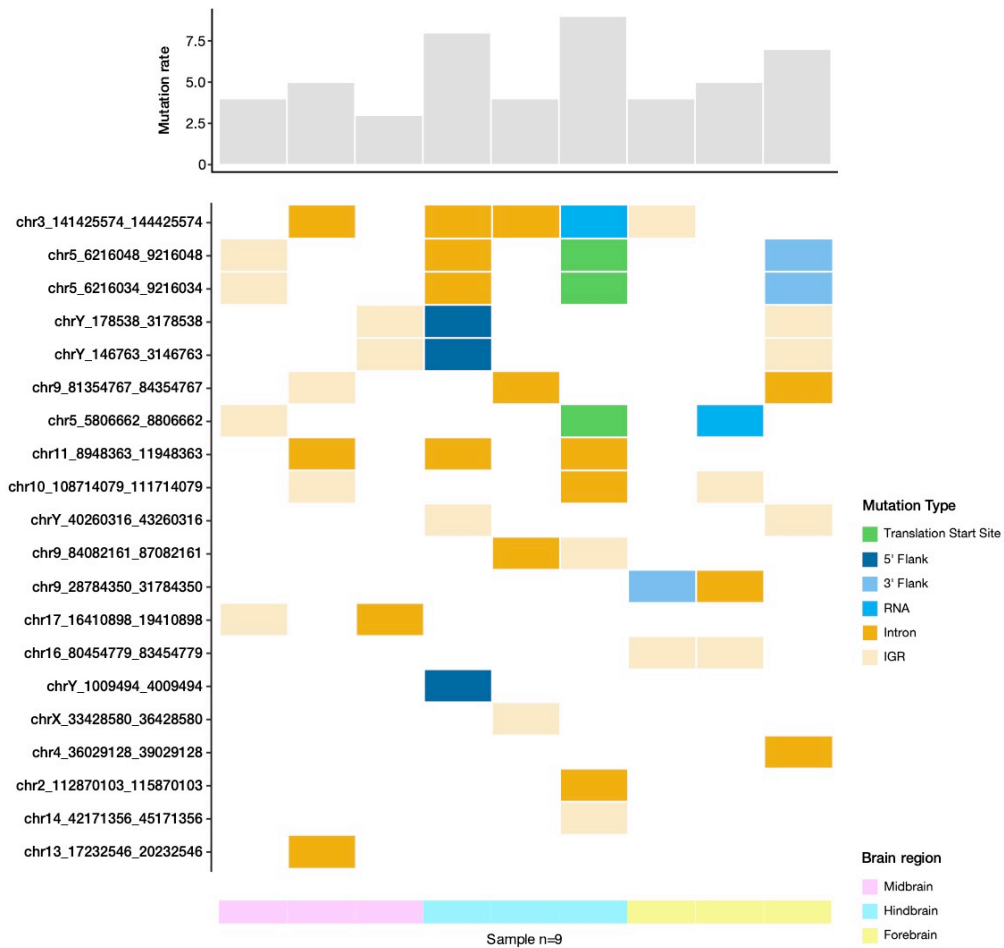


Supp. Figure 9. Mutation rate and Heatmap for Top20 mutated regions only in Embryo KO.





**Supp. Figure 10. Mutation rate and Heatmap for Top20 mutated regions only in Adult WT.**



**Supp. Figure 11. Mutation rate and Heatmap for Top20 mutated regions only in Embryo WT.**



PUBLICATIONS

## COLLABORATION 1

Title: Adaptation to environmental factors shapes the organization of regulatory regions in microbial communities

Authors: Leyden Fernandez, Josep M Mercader, **Mercè Planas-Fèlix**, David Torrents

Journal: BMC Genomics

Impact factor: 3.986

Citations: 9

Contribution: Ph.D. Candidate Mercè Planas-Fèlix contribution to this study involved the regulatory network analysis of transcription factor binding sites, and has been involved in drafting the manuscript.

RESEARCH ARTICLE

Open Access

# Adaptation to environmental factors shapes the organization of regulatory regions in microbial communities

Leyden Fernandez<sup>1</sup>, Josep M Mercader<sup>1</sup>, Mercè Planas-Fèlix<sup>1</sup> and David Torrents<sup>1,2\*</sup>

## Abstract

**Background:** It has been shown in a number of metagenomic studies that the addition and removal of specific genes have allowed microbiomes to adapt to specific environmental conditions by losing and gaining specific functions. But it is not known whether and how the regulation of gene expression also contributes to adaptation.

**Results:** We have here characterized and analyzed the metaregulome of three different environments, as well as their impact in the adaptation to particular variable physico-chemical conditions. For this, we have developed a computational protocol to extract regulatory regions and their corresponding transcription factors binding sites directly from metagenomic reads and applied it to three well known environments: Acid Mine, Whale Fall, and Waseca Farm. Taking the density of regulatory sites in promoters as a measure of the potential and complexity of gene regulation, we found it to be quantitatively the same in all three environments, despite their different physico-chemical conditions and species composition. However, we found that each environment distributes their regulatory potential differently across their functional space. Among the functions with highest regulatory potential in each niche, we found significant enrichment of processes related to sensing and buffering external variable factors specific to each environment, like for example, the availability of co-factors in deep sea, of oligosaccharides in soil and the regulation of pH in the acid mine.

**Conclusions:** These results highlight the potential impact of gene regulation in the adaptation of bacteria to the different habitats through the distribution of their regulatory potential among specific functions, and point to critical environmental factors that challenge the growth of any microbial community.

**Keywords:** Adaptation, Environment, Gene regulation, Metagenomes

## Background

Metagenomic studies generate a massive amount of sequence information of communities of organisms living in different physicochemical conditions. This allows, for the first time, to search for the molecular and genetic basis of adaptation through the comparison and the study of genomes of different species sharing the same environment, and of similar species living in different conditions. The comparative studies of the potential protein content in many of these datasets have already provided interesting examples of specific functions that correlate with specific

characteristics of the environment. For example, in the search of functional fingerprints related to specific habitats, a comparative analysis between soil, and deep and superficial aquatic environments found abundant orthologous groups specific of these particular habitats [1]. In this case, the examination of higher order processes reveals differences in energy production between these three niches, such as starch and sucrose metabolism in soil or photosynthesis in oligotrophic surface waters [1,2].

More recently, metagenomic studies have gone beyond the sequencing of DNA and the counting of genes, and have incorporated techniques and protocols to detect, measure and analyze their transcriptome. While the sequencing of metagenomes provides an overview of the genes present in specific environments that can potentially play a role in adaptation, the analysis of expression

\* Correspondence: david.torrents@bsc.es

<sup>1</sup>Joint IRB-BSC program on Computational Biology, BSC, Jordi Girona, 29, 08034 Barcelona, Spain

<sup>2</sup>Institució Catalana de Recerca i Estudis Avançats (ICREA) Pg Lluís Companys 23, Barcelona 08010, USA



provides a more precise picture of what functions are expressed and active in a particular moment of the environment. Even though the techniques for mRNA isolation and sequencing from metagenomic samples are still not able to provide comprehensive pictures of expression profiles, there have been important progresses in this direction and some interesting findings. For example, one of the first studies of metatranscriptome, despite it covered a small fraction of the expressed genes, identified specific biological processes active in bacterioplankton communities that could be correlated with either marine or freshwater conditions [3]. As the coverage and accuracy of these analysis increased (mostly by including next generation sequencing techniques), more active processes have been linked to variable environmental conditions. For instance, an expression time-series performed on microbial communities living in surface oceanic showed that processes of energy production were active in hours with light, while anabolic housekeeping processes were predominant during the night [4]. Despite the underlying methodology behind, metatranscriptomics still needs to overcome several challenges [5]. But the rapid progress in this field is promising and we will soon have the opportunity of building accurate expression profiles and compare them across environments, as well as exploring the interaction of processes of different organisms within specific environments.

In the present study we have conducted a novel approach that complements and bridges metagenomic and metranscriptomic concepts. The rationale behind this study relies on the hypothesis that the regulation of the expression of those biological functions that confer adaptation to variable environmental conditions will show higher complexity, i.e. they will have complex regulatory regions.

Previous studies [6,7] have shown that genes with complex regulation requirements show higher number of transcription factor binding sites (TFBSs) in their upstream cis-regulatory regions compared to housekeeping genes. For example, stress-response genes in yeast need a precise regulation of their expressions patterns to adapt to drastic changes of environmental conditions and also show a significantly higher number of different TFBSs in their upstream regulatory regions. Beyond the extensive analysis of the regulatory characteristics of particular functions [8], up to now, there are not global approaches and studies on how the regulatory potential of entire microbial communities is influenced and organized in natural habitats.

In particular, and using the same rationale, we have measured and compared the complexity of gene regulation in bacteria and archaea living in environments with distinct underlying physico-chemical conditions. For that purpose, we searched within each of the environments

for specific functional signatures predicted to have high regulatory potential. These are correlated with specific and also dynamic physico-chemical stress factors of each of the niches. The functional significance of the differences detected highlights the existence of adaptation strategies that rely on the regulatory potential of regions that control the expression of specific fitness genes.

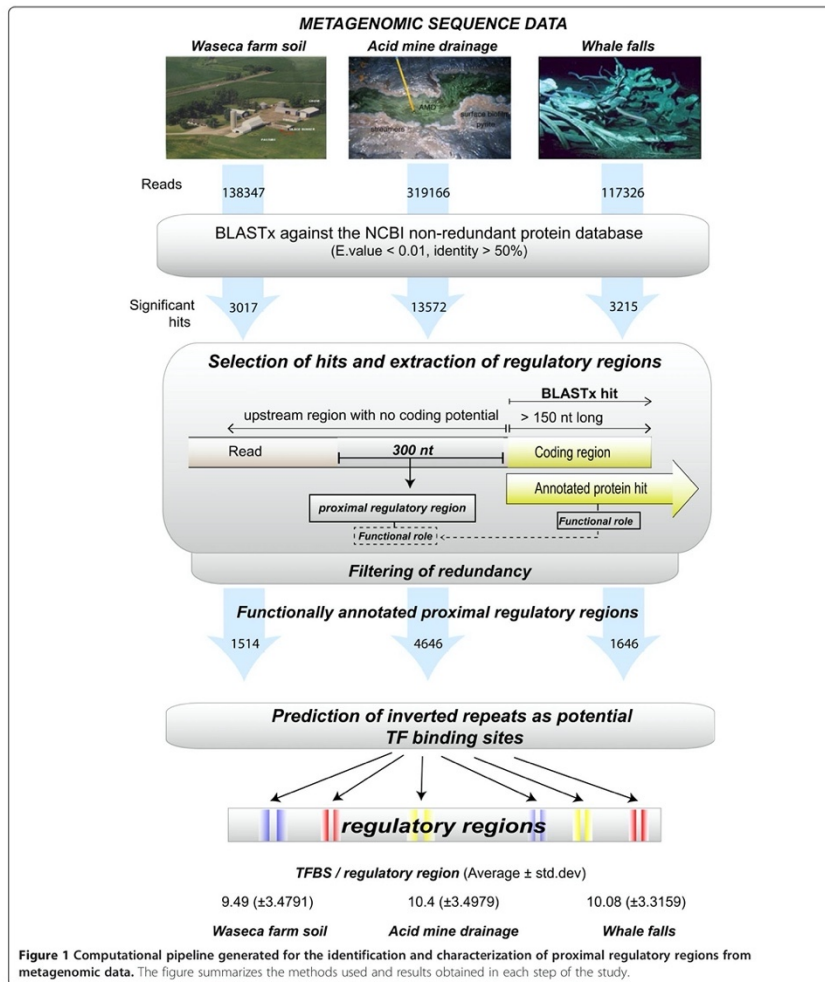
## Results and discussion

With the ultimate goal of identifying and characterizing the extend, to which environmental factors influence the organization of the regulatory potential of particular microbial communities, we have studied and compared the regulome of three fundamentally different ecological niches using whole metagenomic data. We next provide details on the major results and findings of this study: (1) The development of a new pipeline for the identification and prediction of proximal regulatory regions and their TFBS from metagenomic data; (2) and the generation of a collection of regulatory regions from three well studied and reference metagenomic samples (Whale Fall, Waseca Farm and Acid Mine). The comparative analysis of this data has shown that, while (3) the overall distribution of TFBS on promoters is the same across environments, their distribution across their functional space is significantly different, as (4) promoters with higher number of TFBS tend to regulate environment specific functions, and (5) a fraction of these are environment specific and can be linked to characteristic external physicochemical factors (Additional file 1: Figure S1).

### Identification and classification of proximal regulatory regions from metagenomic data

We first characterized and analyzed the gene regulatory space from metagenomic data obtained from three well-known sequenced environments with clearly different physico-chemical properties: Whale Fall Community, Acid Mine and Waseca County Farm Soil [1]. For that, we started by identifying and defining gene regulatory regions to later characterize them, as to their levels of TF binding, i.e. their regulatory potential. For the design of a search strategy, we followed two major considerations: first (1) avoiding biases in favor of most abundant and well-known bacteria (and closely related species), as well as, (2) ensuring an equal coverage through all the phyla detectable in those samples. As a result, we developed a pipeline that consists of two major steps: (1) first the identification of proximal regulatory regions and then, within each of them, (2) the prediction of potential regulatory transcription factor binding sites. The complete pipeline is detailed in the Methods and summarized graphically in Figure 1.

Through extensive homology searches, our procedure identified putative proximal regulatory regions in Waseca



Farm Soil, in Acid Mine Drainage, and in the Whale Falls Sample (a complete catalogue of these regions can be found in Additional files 2, 3 and 4). A first and basic taxonomical analysis of these sequences shows that these

promoters cover all phyla (Additional file 1: Figure S2) that were previously described in these environments [1].

Next, we estimated the level of regulatory potential for each of these promoter regions through the prediction



of their transcription factor binding sites. In order to minimize possible biases favoring promoters from well-studied bacteria (or close species), we did not consider TFBSs prediction strategies that rely on the homology mapping of described TFBSs. Instead, we used a *de novo* prediction protocol that relies on the identification of palindromic repeats [9], which have been previously determined as preferred binding sites for transcription regulators in bacteria and archaea [9-14]. Because this method was originally developed for the analysis of single genomes [9] and, although it has been applied to a wide variety of bacterial sequences and studies [14-16], we needed to adapt it to cope with the heterogeneity and redundancy of metagenomic data by including some modification in the scoring system.

#### Evaluation of predicted promoters and TFBSs

Like any other *de novo* prediction method in sequence analysis, we have to initially assume the presence of false positive TFBS models among correct predictions. To assess for the reliability of all of our predictions and to put our strategy and results into the context of our goals and of the current knowledge about regulatory regions in prokaryotes, we performed different quantitative and qualitative comparisons with available independent data and methodologies.

From a quantitative point of view, we (1) first observed that the global average of 10 TFBS per promoter (with 0 as minimum and 25 as maximum values) that we identify from all three environments is in agreement with previous estimates obtained with different bacterial species and methodologies. For example, using genome comparative analysis, an average of 11-13 TFBS motifs per promoter was found for *Shewanella* [17]. In addition, a study of the transcription regulatory network of *E. Coli K12* predicted up to 16 sites per promoter [18], and up to 20 through the identification of half-sites motifs [19]. (2) We also evaluated the performance of our methodology by comparing our results with those obtained with an independent method, MotifClick, that predicts cis-regulatory regions using a graph-based polynomial-time algorithm [20]. After running both predictors over intergenic *E. Coli* regions, we observed that the densities of TFBS resulting from one or the other strategy showed high correlation values ( $\rho = 0.52$ ,  $p\text{-value} < 2.2 \times 10^{-6}$ ; (Additional file 1: Figure S3).

From a qualitative point of view, we first (1) assessed the biological significance of our predictions by carrying out a randomization test consisting in applying the same prediction pipeline to our collection of promoters with their nucleotide sequence completely shuffled, i.e. with no biological information. We observed that the distributions of the number of motifs per promoter were significantly different between the real and the randomized

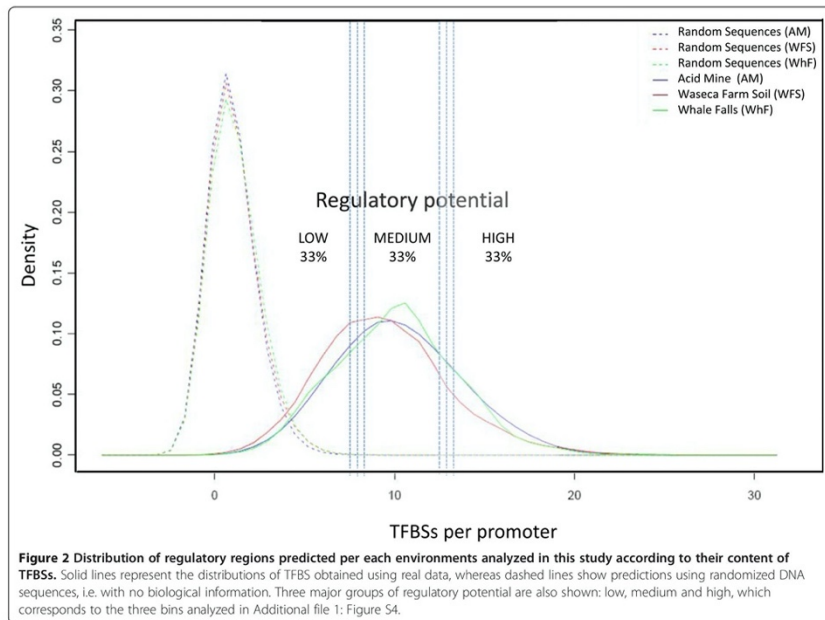
sample (Figure 2). (2) Furthermore, we screened for coincidences between our predicted TFBSs and those reported in the RegPrecise database [21], which consist on manually curated site reconstructions in various bacteria genomes. This comparison showed that 28% of our predicted binding sites include, at least, one possible binding sequence of the matrices for each of the 38 TFs included in RegPrecise (Additional file 5). (3) Finally, we also searched for a particular type of false predictions, which consist on regulatory palindromic repeats with no binding potential, named Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) [22]. The results that we obtained using the CRISPRFinder web tool [23] showed a negligible amount of these regions (less than 1% of our set of promoters), which were subsequently removed from the analysis.

In summary, all these evaluation tests suggest that our set of promoters is both, quantitatively and qualitatively reliable, as they show a significant fraction of reported TFBS, and a small portion of false positives. But, most importantly, the presence of this small fraction of false positives is not expected to affect our final conclusions, as these come from comparisons within and between environments and do not rely on absolute TFBSs counts.

#### Functional organization of regulatory potentials within each environment

We then studied how microbial communities living in these environments organize and distribute their regulatory potential through the different biological functions and to which extend this could obey to specific adaptation needs. It is interesting to observe that, whereas the range of density of predicted sites per promoter is wide within each of the environments, the overall distribution and the averages are similar: 9.98 ( $\pm 3.29$ ), 9.58 ( $\pm 3.49$ ) and 10.28 ( $\pm 3.35$ ) for Acid Mine Drainage, Waseca Farm Soil and Whale Falls samples, respectively (Figure 2). This indicates that, although these three environments present (1) different sequence coverage, (2) different physicochemical characteristics and (3) different species composition, the overall regulatory potential, as to the total number of different TFBS, and their distribution across the promoters follow a similar pattern.

To go beyond simple counts and to explore whether or not this regulatory potential is distributed equally through all the functions of each of the metagenomes, we first identified the functions under the control of our collection of proximal regulatory regions. For this, we assigned to each promoter the functional category (from SEED database) [24] of the corresponding downstream coding region using MEGAN [25] see (Additional files 6, 7 and 8, for a complete list of functions and TFBS densities). We first investigated whether the regulatory potential is organized differently over the functional space



of each of the environments. For this, we ranked all promoters of each sample according to their TFBS density and count, for each density group, how many associated functions are specific of that particular environment, or co-occur in one or in the other two samples. This analysis showed significant differences between promoters. Interestingly, the functions under the control of promoters with high number of TFBSs show significantly less co-occurrences between environments, than those regulated by promoters with lower regulatory potential. The fact that promoters with high density of TFBS are enriched in environment specific functions provides the first hint that processes that require complex regulation might provide adaptation to environment specific variable external factors. (Additional file 1: Figure S4). We expect that a large fraction of functions that showed a higher co-occurrence among environments likely correspond to housekeeping roles.

To study this further, we next investigated which functions are specifically enriched among the highly regulated ones in each of the environments. For this, we zoomed into the fraction of the 33% highest regulated

promoters (i.e. with more than 12 TFBSs/promoter) and subdivided it further into subgroups covering the 1, 5, 10, 20, 30 and 40 top percentages of TFBS density, to finally analyze the functional enrichments within each of them. This analysis highlighted different enriched functions in each of the environments (see Additional file 1: Figure S5 (Acid Mine), S6 (Wasaca Soil), S7 (Whale Falls)). These enriched functions cover different types of processes, the majority of them involved in sensing and buffering external factors, such as, receptors and transporters in Acid Mine and stress response systems in Whale Falls.

#### Potential environment-gene regulation relationships

In order to finally highlight potential points of interaction between highly regulated functions that could provide adaptation to variable conditions specific to each of the environments, we first selected for each habitat, those functions that show stronger enrichment, i.e. with  $p$ value < 0.05, among the top 1, 5, 10 and 20% groups and with clear orthologous functions in the other two samples. This subgroup of functions include (virulence,

cell cycle, carbohydrates metabolism, stress response and cofactors metabolism), which we then compared among environments and evaluated their relationship with the niche specific variable factors. For this, we carried out extensive literature searches on different biochemical mechanisms of adaptation guided by these functions and the characteristics of the environment. Despite the limited information about the environment physico-chemical factors characteristic of available metagenomic studies, we propose in the following sections potential adaptive scenarios by correlating highly regulated functions with known variable external factors in each of the environments.

#### Waseca Farm Soil

In Waseca Farm Soil, carbohydrates metabolism related functions appear as highly regulated, more precisely di and oligosaccharides metabolism ( $p\text{value} = 1 \times 10^{-16}$ , within environment and adjusted  $p\text{value}$  (Bonferroni)  $= 9 \times 10^{-13}$  for Fisher's exact test between groups). This fact could be in concordance with the fluctuations in organic matter concentrations in the soil, such as, plant debris, which has also been previously proposed as an explanation for the presence of other carbohydrate metabolism functions specific of this environment [1]. This further agrees with the behavior observed in lower eukaryotes abundant in soil, like yeast, where high complexity in their transcriptional regulation were found upstream of genes that play a role in carbohydrates metabolism [26]; and with the fact that, in this niche, the upstream region of the FruR gene, a known TF that regulates carbohydrate metabolism, appears as highly regulated, with the highest number of predicted TFBS (Additional file 1: Figure S8).

#### Whale Falls samples

A different scenario is observed in Whale Fall where, even though each of the subsamples were collected in a specific moment of decomposition from two different whales and at different depths, they all share similar general physico-chemical patterns, predominating the drastic fluctuations of nutrient availability [1]. In agreement of what would be expected for microorganisms living in these kind of environments, most of the highly regulated functions that are enriched in whale falls samples are related to adaptation capabilities to starving periods (Figure 3). Particularly, we found TFBS rich promoters upstream of genes that are involved in cell cycle and growth, i.e. the control of basic macromolecular synthesis operon. This is in contrast to what happens in Waseca and Acid Mine, where the same functions present lower density of TFBSs (Figure 3).

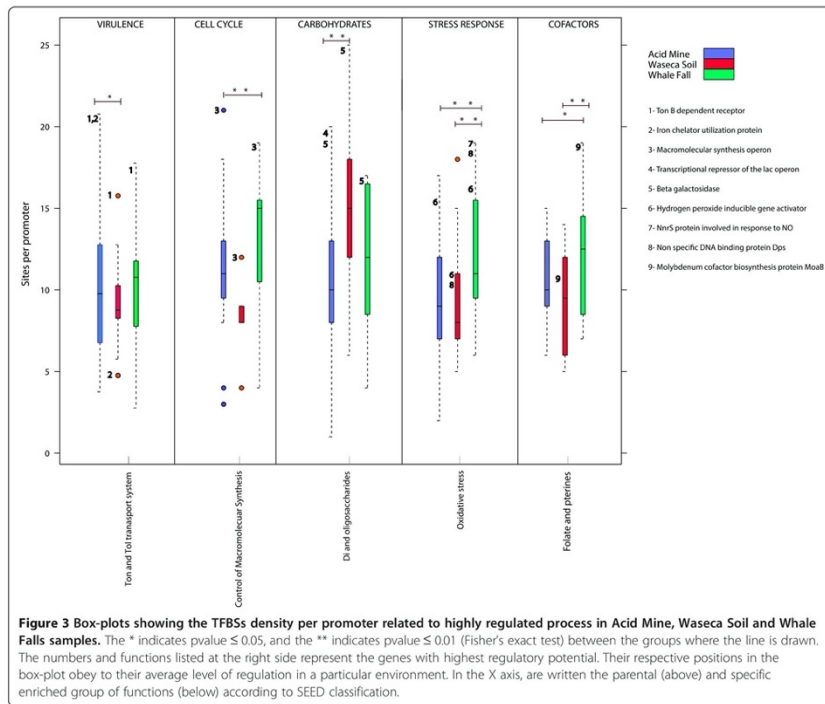
Moreover, bacterial communities living in cold water are also exposed to high concentrations of oxidant

reagents [27] causing an increase in the metabolic costs associated with the activation of antioxidant defenses. In fact, functions related with the response to oxidative stress appear specifically enriched in this environment compared to others. These functions comprise, for example, the hydrogen peroxide-inducible gene activator and a hem and copper containing membrane protein (NnrS), that needs to respond to external NO concentrations. Additionally, parts involved in the machinery that protects genomic DNA during prolonged non-growing phases [28], like the non-specific DNA binding protein (Dps), also appear as highly regulated in this niche.

It has been also pointed before, that the uptake and metabolisms of cofactors and amino acids are particularly variable in marine environments, essential to adapt to typical oceanic oligotrophic conditions [2]. In agreement with this, cofactor metabolism related functions are also enriched (adjusted  $p\text{value}$  (between groups)  $\leq 0.05$ ). In particular, we found enrichment for enzymes involved in the metabolism of molybdenum cofactors, pterin and folate (Figure 3). These findings were further confirmed by the overrepresentation of TyrR and ArgR binding sites in this niche, both known to be TFs involved in the control of amino acid transport for the synthesis of proteins (according to the RegPrecise database; see Additional file 1: Figure S8).

#### Acid Mine

The acid mine is characterized by extreme physico-chemical conditions, showing low pH records and fluctuating temperature, conductivity and rainfall (see Figure 4A) [29]. Among the functions with high regulatory potential that appear enriched in this niche are those known to play a role in the adaptation to changes in external osmolarity, typical of environments with variable distribution of rainfall across the year [30,31] (Figure 4A). It is worth mentioning the high regulatory potential of some genes related to the TonB transport system (Figures 3, 4B), which are also involved in avoiding toxicity by keeping metal homeostasis inside the cell [32], in particular of iron. The high regulatory potential of the TonB-dependent receptor and the iron chelator utilization protein (Figure 3) might provide homeostasis (i.e. plasticity) to acid mine bacteria living under variable ferric concentrations, which is further confirmed by the fact that a significant fraction of homeostasis-related promoters could be assigned to *Leptospirillum* (genus known to be adapted to low pH [33]) (Figure 4). In addition, we found overrepresentation of binding sites for LexA transcription factor in this niche (see Additional file 1: Figure S8), and, specifically in Ton and Tol transport systems related promoters (the sequence for LexA binding site is in Figure 4B, colored in red). LexA transcription factor is known to be involved in the response to DNA damage

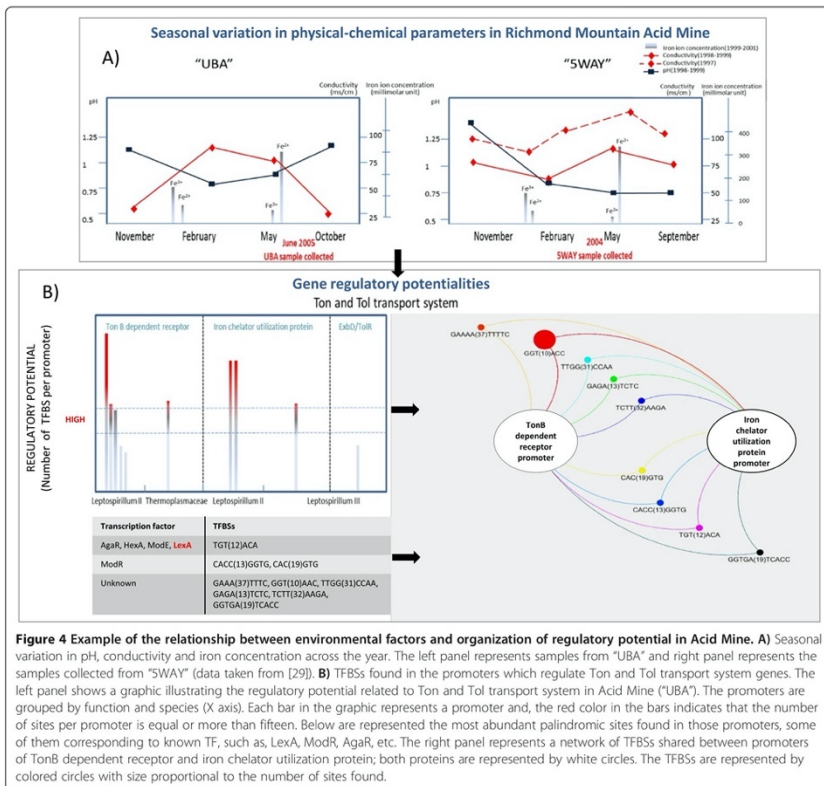


and external pH fluctuations [34]. In fact, when we evaluated the fraction of binding sites shared between two members of the Ton and Tol system (iron chelator utilization protein and TonB dependent receptor), we found a high number of coincidences for other sites besides LexA (i.e. sites for the transcription factors ModR and ModE involved in metal metabolism) (Figure 4B).

Taken together, the fact that highly regulated functions are not the same between the different environments agrees with previous metatranscriptome studies [3,4] and indicates that the organization of the regulatory potential between the functional space of each niche is different and influenced by the environmental physico-chemical conditions. This could reflect organism-environment interaction points where gene regulation should be able to provide enough plasticity to the functional network for the adaptation to variable external parameters.

## Conclusions

We have here studied how variable physico-chemical conditions of the environment can shape the regulome of microbial communities living there to provide adaptation. We have combined existing and novel methodologies and applied it to three environments (Acid Mine, Whale Fall and Waseca Farm Soil) to identify and characterize, for the first time, their regulatory space, i.e. proximal promoters and their corresponding TFBSs. Taking the density of TFBSs as a measure of the level of regulatory potential, we first observed that, despite the differences of the living conditions of each of the environments studied here, their distribution of the regulatory potential, at quantitative level, appears to be nearly identical. However, when we went beyond simple counts we observed that the associated cellular functions in different groups related to the regulatory potential tend to be environment specific. This supports our hypothesis and expectation that point to



a role of gene regulation in the adaptation of organisms to particular and variable external factors. Also in this direction, we have found specific functional enrichments among highly regulated functions in each of the metagenomas, suggesting potential interaction points between gene regulation and dynamic environmental conditions. In particular, we have identified points of interaction between signatures of significant functional enrichment and specific characteristics of the marine and terrestrial environments. These results highlight the impact of gene regulation in the adaptation of microbes to their habitat. Beyond contributing to the general understanding of how wild bacterial communities interact with the environment, our methodology can also be used to identify potential

external factors to which bacteria are particularly sensitive in order to design efficient communities for therapeutic, or ecological needs.

## Methods

### Datasets

Metagenomic samples (i.e. Sanger sequencing reads) were downloaded from the Camera Database [35]. In particular, (1) samples of whale falls were obtained from three independent libraries named Whale falls: CAM\_SMPL\_WHALEFALLBONE (Whale fall carcass bone, W. Antarctic Peninsula Shelf), CAM\_SMPL\_WHALEFALLMAT (Whale fall carcass microbial mat, Santa Cruz Basin), and CAM\_SMPL\_WHALEFALLRIB

(Whale fall carcass rib bone, Santa Cruz Basin) in the database. These three sets differ in the depth of the sampling and come from two different whale samples. (2) The Acid Mine dataset is formed by 5-Way (CG) Acid Mine Drainage Biofilm Metagenome and UBA Acid Mine Drainage Biofilm Metagenome reads. The first corresponds to a low-complexity microbial biofilm growing hundreds of feet underground within a pyrite (FeS<sub>2</sub>) ore body. The UBA biofilm was subaerial, collected from the base of a ~2 m high pile of pyrite sediment. (3) The third environment corresponds to a surface soil (0–10 cm) collected from a Waseca County farm in Minnesota.

#### Promoter identification

The prediction and classification of regulatory regions from metagenomic data relies on the extraction of DNA regions upstream of coding genes detected through homology searches directly from the sequencing reads. For this protocol we selected conservative filters to ensure the reliability of the putative promoters found. Simplified in Figure 1, our protocol consisted in: (1) filtering out reads shorter than 800 base pairs. This filter keeps up to 90% of all reads and ensures both, the detection of the coding region and the extraction of the putative promoter from the reads; (2) detection of reads with coding potential through the comparison of all the sequences of each metagenome with all bacterial and archaeal annotated proteins (NCBI; <http://www.ncbi.nlm.nih.gov/Ftp/>), using BLASTx (default parameters [36]), and selecting those reads with a match to a known protein over, at least, 150 amino acids and with more than 50% of sequence identity; (3) filtering out those positive reads that did not contain at least 300 nucleotide of non-coding sequence upstream of the region matching in BLASTx. This filter enriches our sampling in regions with regulatory potential by avoiding internal genes of operons, which are expected to have short upstream regions with no regulatory potential. Finally, from the remaining accepted reads (13572, 3017, and 3215, for Acid Mine Drainage, Waseca Farm Soil and Whale Falls Samples, respectively) we extracted 300 nucleotides upstream of the coding region as putative promoter sequence. We expect that the 300 base pairs criteria will affect equally all bacteria and environments and will not favor bacteria with largest genomes, as this length has been also described for *Pelagibacter ubique*, the free living bacteria with the smallest genome known [37]. Moreover, fixing this length also avoids short intergenic regions within operons, as their regulatory role is not yet well understood.

To avoid other possible biases favoring common species in these environments and to make possible comparative and qualitative analyses between them, we also removed the redundancy within these collections of putative promoter sequences using a cutoff of 98% of

sequence identity. We also removed those reads that correspond to eukaryotic DNA, mostly from plant species in the Waseca sample, identified using MEGAN [25]. To discard the inclusion of (parts of) ncRNA genes into the collection of promoters, we applied a second filter to remove ncRNAs that target untranslated 5' portions of mRNAs by using Rfam [38] and also we did a second prediction of coding region in our set of putative promoters using the software Prodigal [39] that allows the identification of genes even if the specie is unknown.

#### Prediction of transcription factor binding sites

We next searched for sequence motifs with binding potential within the putative promoters identified before. For this, we used a *de novo* prediction method that is based on the identification of palindromic repeats separated by a spacer DNA region. In particular, we used the most recent adaptations of the method [14] originally described by Li and coworkers [9].

In order to identify putative cis regulatory elements, we screened each promoter sequence for  $W_1NW_2$ , DNA motifs, where  $W_1$  and  $W_2$  are 3–5 nucleotide long palindromic sequences separated by  $N$  (0–30) arbitrary bases. This method relies on the fact that prokaryotic TFBSs are usually palindromes between 12 and 30 base pairs, which may facilitate the dimerization and binding of TFs [12].

To assign a probabilistic values to all motifs found, we first calculated the probability of observing  $n(D)$  copies of a dimer  $D$  by chance, by pooling all the promoters and calculating its expected frequency from the formula,

$$y(D) = \text{Leff}(D) \frac{n(W_1)}{\text{Leff}(W_1)} \frac{n(W_2)}{\text{Leff}(W_2)} \quad (1)$$

where  $n(W_1)$  and  $n(W_2)$  are the total number of occurrences of  $W_1$  and  $W_2$  in the whole data set (all three environments together) and  $\text{Leff}(D) = \sum_r (L(r) - L(D) + 1)$  is the number of independent positions in the data where a motif  $D$  of length  $L(D)$  can be found. The summation is over all the occurrences among 11,614 promoters identified, each with a length  $L(r)$  (i.e. the estimated distance between coding regions). Finally, a  $P$ -value is assigned to each of the motifs assuming that the background follows a Poisson distribution:

$$P = \sum_{n \geq n(D)} \frac{y^n(D)}{n!} e^{-y(D)} \quad (2)$$

and is considered significant if  $P < 1/N_{\text{motifs}}$  where  $N_{\text{motifs}}$  is the total number of positive motifs found. As  $W_1$  is the reverse complement of  $W_2$  (palindrome), the cutoff on  $P$  is corrected by the total number of palindromic dimers found [9,14].

In order to identify environment specific enrichment of our know TFBSs (i.e. those present in the RegPrecise

database), we run a Kruskal Wallis test to compare the density of each particular known TFBS among all three environments. The density of known TFBS per metagenomes is calculated as follows:

$$D(x) = \frac{\sum_{i=0}^N TFBS}{N * Tbp} \quad (3)$$

where  $D(x)$  is the density of TFBSs per metagenome,  $N$  represents the number of promoters found in the  $x$  metagenome and  $Tbp$  is the number of base pairs per promoter (300 base pairs). The complete list of overrepresented TFBSs found in our selected promoter set are shown in Additional files 6, 7, 8, 9, 10 and 11, for Acid Mine, Waseca Farm Soil and Whale Falls samples, respectively.

#### Method validation

For the randomization test on TFBS prediction, we run the corresponding searching methodology on predicted promoter regions after shuffling their sequence using a 20 nt window to ensure the minimum variance of local nucleotide composition.

For the comparison with the MotifClick method [20] we first downloaded intergenic regions from the *Escherichia coli* K12-W3110 genome from IMG database (<https://img.jgi.doe.gov>). We ran MotifClick (motif length = 14 nt) over these regions, specifically 300 nucleotide upstream annotated TSS and recorded the number of positive predictions per promoter. These values were then compared with the results provided by our method applied on the same set of *E. Coli* regulatory regions (Additional file 1).

#### Statistical procedure for the functional analysis

Functional assignment for all the data was performed by MEGAN software [25] using the output of BLAST searches of our reads against databases of known bacterial proteins. Through this comparison we could identify up to 1646 (Whale Falls Samples), 4646 (Acid Mine Drainage) and 1514 (Minnesota Farm Soil) gene upstream segments with functional assignment. In order to roughly study up to which level low, medium and high regulated functions are shared among environment we have run a Spearman test for independence using R, for the rectangular plot and correspondence analysis we use the plot function included in R graphics (<http://www.r-project.org/>) (see Additional file 1: Figure S4).

In addition, functional enrichment analysis was done by first ranking all promoters as to their number of predicted TFBSs. Then, for each of the groups of interest, we ran a Fisher's exact test for count data to see whether particular functions within each group (top 1%, 5%, 10%, 20%) were specifically enriched versus the total

distribution of functions. For this, we have used "all intermediate" functional levels according to MEGAN classification. Heat maps for all function within environment were obtained using package ggplot2 for R (Additional file 1). Then, we retained significant cases based on two criteria 1) functions whose  $p < 0.05$  within environment and 2) functions with orthologous in the other three environments. Those selected groups were compared again, this time among environments, for this analysis we ran a Fisher's exact test to see whether functional enrichment within environment were maintained among them.

#### Additional files

**Additional file 1: Figure S1.** Shows the overview of the general results of this study. **Figure S2.** shows the comparative analysis of the taxa obtained with MEGAN on our promoter regions compared with that obtained previously using 16S rRNA information from the same samples in Waseca soil (a), Whale falls (b), and Acid mine (c). **Figure S3.** represents the correlation analysis between the TFBSs predictions per promoter using the method explained in this paper versus MotifClick predictions. **Figure S4.** illustrates a global view of the relationship between regulatory potential and the level of co-occurring functions within each of the environments. **Figure S5.** Results of the functional enrichment analysis for Acid Mine using the predefined bins. **Figure S6.** Results of the functional enrichment analysis for Waseca Farm using predefined bins. **Figure S7.** Results of the functional enrichment analysis for Whale Falls using predefined bins. **Figure S8.** shows the relative abundances of our TFBS prediction that matched known TFBS.

**Additional file 2:** List of promoters selected after applying the methodology described in Figure 1 on Waseca Farm Soil data.

**Additional file 3:** List of the promoters selected after applying the methodology described in Figure 1 on Acid Mine data.

**Additional file 4:** List of the promoters selected after applying the methodology described in Figure 1 on Whale Fall Samples data.

**Additional file 5:** Table listing the number of TFBSs per genomes found after applying our method versus the number of sites described in Regprecise database.

**Additional file 6:** Table in CSV format listing the number of TFBSs identified for each promoter and the function assigned to the corresponding downstream coding region in Acid Mine.

**Additional file 7:** Table in CSV format listing the number of TFBSs identified for each promoter and the function assigned to the corresponding downstream coding region in Waseca Soil.

**Additional file 8:** Table in CSV format listing the number of TFBSs identified for each promoter and the function assigned to the corresponding downstream coding region in Whale Falls.

**Additional file 9:** A list (CSV MS-DOS format) of overrepresented TFBSs per promoter found in Acid Mine, Waseca Soils and Whale falls, respectively. The abbreviated nomenclature used for the binding sites is the following: N, W, Sequence, where N is the number of variable nucleotides. W is the number of nucleotides defining the inverted repeat. Sequence is the actual sequence of the site. Example: 10 3 ATC, corresponds to: ATCNNNNNNNNNGAT.

**Additional file 10:** A list (CSV MS-DOS format) of overrepresented TFBSs per promoter found in Acid Mine, Waseca Soils and Whale falls, respectively. The abbreviated nomenclature used for the binding sites is the following: N, W, Sequence, where N is the number of variable nucleotides. W is the number of nucleotides defining the inverted repeat. Sequence is the actual sequence of the site. Example: 10 3 ATC, corresponds to: ATCNNNNNNNNNGAT.

**Additional file 11: A list (CSV MS-DOS format) of overrepresented TFBSs per promoter found in Acid Mine, Waseca Soils and Whale falls, respectively.** The abbreviated nomenclature used for the binding sites is the following: N, W, Sequence, where N is the number of variable nucleotides. W is the number of nucleotides defining the inverted repeat. Sequence is the actual sequence of the site. Example: 10 3 ATC, corresponds to: ATCNNNNNNNNNGAT.

#### Competing interests

The authors declare that they have no competing interests.

#### Authors' contributions

DT conceived this study. LF and DT designed the methodology. LF analyzed the data and performs the literature searches. JMM and LF participated in the statistical analysis. LF and MP performed the regulatory network analysis of TFBSs. JMM, DT, MP and LF have been involved in drafting the manuscript. All authors read and approved the final version of the manuscript.

#### Acknowledgments

This work was supported by a grant from Ministerio de Economía y Competitividad through the project BES-2008-005973. Josep M. Mercader was supported by Sara Borrell Fellowship from the Instituto Carlos III. This work has been supported by the grant SEV-2011-00067 of Severo Ochoa Program, awarded by the Spanish Government. Merce Planas is funded by the Obra Social Fundación la Caixa under the Severo Ochoa 2013 program.

Received: 4 February 2014 Accepted: 24 September 2014

Published: 8 October 2014

#### References

1. Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, Bork P, Hugenholtz P, Rubin EM: **Comparative metagenomics of microbial communities.** *Science* 2005, **308**:554–557.
2. Gianoulis TA, Raes J, Patel PV, Bjornson R, Korbel JO, Letunic L, Yamada T, Paccanaro A, Jensen LJ, Snyder M, Bork P, Gerstein MB: **Quantifying environmental adaptation of metabolic pathways in metagenomics.** *Proc Natl Acad Sci U S A* 2009, **106**:1374–1379.
3. Poretsky RS, Bano N, Buchan A, LeCleir G, Klei Kemper J, Pickering M, Pate WM, Moran MA, Hollibaugh JT: **Analysis of microbial gene transcripts in environmental samples.** *Appl Environ Microbiol* 2005, **71**:4121–4126.
4. Poretsky RS, Hewson L, Sun S, Allen AE, Zehr JP, Moran MA: **Comparative day/night metatranscriptomic analysis of microbial communities in the North Pacific subtropical gyre.** *Environ Microbiol* 2009, **11**:1358–1375.
5. Gifford SM, Sharma S, Rinta-Kanto JM, Moran MA: **Quantitative analysis of a deeply sequenced marine microbial metatranscriptome.** *ISME J* 2011, **5**:461–472.
6. Farre D, Bellora N, Mularoni L, Messegueur X, Alba MM: **Housekeeping genes tend to show reduced upstream sequence conservation.** *Genome Biol* 2007, **8**:R140.
7. Lin Z, Wu WS, Liang H, Woo Y, Li WH: **The spatial distribution of cis regulatory elements in yeast promoters and its implications for transcriptional regulation.** *BMC Genomics* 2010, **11**:581.
8. Merino E, Jensen RA, Yanofsky C: **Evolution of bacterial trp operons and their regulation.** *Curr Opin Microbiol* 2008, **11**:78–86.
9. Li H, Rhodius V, Gross C, Siggia ED: **Identification of the binding sites of regulatory proteins in bacterial genomes.** *Proc Natl Acad Sci U S A* 2002, **99**:11772–11777.
10. McCue L, Thompson W, Carmack C, Ryan MP, Liu JS, Derbyshire V, Lawrence CE: **Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes.** *Nucleic Acids Res* 2001, **29**:774–782.
11. Huffman JL, Brennan RG: **Prokaryotic transcription regulators: more than just the helix-turn-helix motif.** *Curr Opin Struct Biol* 2002, **12**:98–106.
12. Rodionov DA: **Comparative genomic reconstruction of transcriptional regulatory networks in bacteria.** *Chem Rev* 2007, **107**:3467–3497.
13. Gelfand MS, Koonin EV, Mironov AA: **Prediction of transcription regulatory sites in Archaea by a comparative genomic approach.** *Nucleic Acids Res* 2000, **28**:695–705.
14. Laing E, Sidhu K, Hubbard SJ: **Predicted transcription factor binding sites as predictors of operons in Escherichia coli and Streptomyces coelicolor.** *BMC Genomics* 2008, **9**:79.
15. Iqbal M, Mast Y, Amin R, Hodgson DA, Consortium S, Wohlleben W, Burroughs NJ: **Extracting regulator activity profiles by integration of de novo motifs and expression data: characterizing key regulators of nutrient depletion responses in Streptomyces coelicolor.** *Nucleic Acids Res* 2012, **40**:5227–5239.
16. Li L: **GADEM: a genetic algorithm guided formation of spaced dyads coupled with an EM algorithm for motif discovery.** *J Comput Biol* 2009, **16**:317–329.
17. Liu J, Xu X, Stormo GD: **The cis-regulatory map of Shewanella genomes.** *Nucleic Acids Res* 2008, **36**:5376–5390.
18. Sun J, Tuncay K, Haidar AA, Ensmann L, Stanley F, Trelnski M, Ortoleva P: **Transcriptional regulatory network discovery via multiple method integration: application to e. coli K12.** *Algorithms Mol Biol* 2007, **2**:2.
19. Leuze MR, Karpinets TV, Syed MH, Bellaev AS, Ueberbacher EC: **Binding Motifs in Bacterial Gene Promoters Modulate Transcriptional Effects of Global Regulators CRP and Arca.** *Gene Regul Syst Bio* 2012, **6**:93–107.
20. Zhang S, Li S, Niu M, Pham PT, Su Z: **MotifClick: prediction of cis-regulatory binding sites via merging cliques.** *BMC Bioinformatics* 2011, **12**:238.
21. Novichkov PS, Laikova ON, Novichkova ES, Gelfand MS, Arkin AP, Dubchak I, Rodionov DA: **RegPrecise: a database of curated genomic inferences of transcriptional regulatory interactions in prokaryotes.** *Nucleic Acids Res* 2010, **38**:D111–118.
22. Grissa I, Vergnaud G, Pourcel C: **The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats.** *BMC Bioinformatics* 2007, **8**:172.
23. Grissa I, Vergnaud G, Pourcel C: **CRISPRfinder: a web tool to identify clustered regularly interspaced short palindromic repeats.** *Nucleic Acids Res* 2007, **35**:W52–57.
24. Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, de Crecy-Lagard V, Diaz N, Disz T, Edwards R, Fonstein M, Frank ED, Gerdes S, Glass EM, Goesmann A, Hanson A, Iwata-Reuyl D, Jensen R, Jamshidi N, Krause L, Kubal M, Larsen N, Linke B, McHardy AC, Meyer F, Neuweger H, Olsen G, Olson R, Osterman A, Portnoy V: **The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes.** *Nucleic Acids Res* 2005, **33**:5691–5702.
25. Huson DH, Auch AF, Qi J, Schuster SC: **MEGAN analysis of metagenomic data.** *Genome Res* 2007, **17**:377–386.
26. Chin CS, Chuang JH, Li H: **Genome-wide regulatory complexity in yeast promoters: separation of functionally conserved and neutral sequence.** *Genome Res* 2005, **15**:205–213.
27. Abele D, Puntarulo S: **Formation of reactive species and induction of antioxidant defence systems in polar and temperate marine invertebrates and fish.** *Comp Biochem Physiol A Mol Integr Physiol* 2004, **138**:405–415.
28. Storz G, Imlay JA: **Oxidative stress.** *Curr Opin Microbiol* 1999, **2**:188–194.
29. Edwards KJ, Gihring TM, Banfield JF: **Seasonal variations in microbial populations and environmental conditions in an extreme acid mine drainage environment.** *Appl Environ Microbiol* 1999, **65**:3627–3632.
30. Albers S-V, Koning SM, Konings WN, Driessen AJM: **Insights into ABC Transport in Archaea.** *J Bioenerg Biomembr* 2004, **36**:5–15.
31. Kempf B, Bremer E: **Uptake and synthesis of compatible solutes as microbial stress responses to high-osmolality environments.** *Arch Microbiol* 1998, **170**:319–330.
32. Osorio H, Martinez V, Nieto PA, Holmes DS, Quatrini R: **Microbial iron management mechanisms in extremely acidic environments: comparative genomics evidence for diversity and versatility.** *BMC Microbiol* 2008, **8**:203.
33. Schrenk MO, Edwards KJ, Goodman RM, Hamers RJ, Banfield JF: **Distribution of Thiobacillus ferrooxidans and Leptospirillum ferrooxidans: Implications for generation of acid mine drainage.** *Science* 1998, **279**:1519–1522.
34. Guazzaroni ME, Morgante V, Mirete S, Gonzalez-Pastor JE: **Novel acid resistance genes from the metagenome of the Tinto River, an extremely acidic environment.** *Environ Microbiol* 2013, **15**:1088–1102.
35. Sun S, Chen J, Li W, Altintas I, Lin A, Peltier S, Stocks K, Allen EE, Ellisman M, Grethe J, Wooley J: **Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource.** *Nucleic Acids Res* 2011, **39**:D546–551.



36. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**:403–410.
37. Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D, Bibbs L, Eads J, Richardson TH, Noordewier M, Rappé MS, Short JM, Carrington JC, Mathur EJ: **Genome streamlining in a cosmopolitan oceanic bacterium.** *Science* 2005, **309**:1242–1245.
38. Burge SW, Daub J, Eberhardt R, Tate J, Barquist L, Nawrocki EP, Eddy SR, Gardner PP, Bateman A: **Rfam 11.0: 10 years of RNA families.** *Nucleic Acids Res* 2013, **41**:D226–D232.
39. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ: **Prodigal: prokaryotic gene recognition and translation initiation site identification.** *BMC Bioinformatics* 2010, **11**:119.

doi:10.1186/1471-2164-15-877

**Cite this article as:** Fernandez et al.: Adaptation to environmental factors shapes the organization of regulatory regions in microbial communities. *BMC Genomics* 2014 **15**:877.

**Submit your next manuscript to BioMed Central  
and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)



## COLLABORATION 2

Title: Metformin alters the gut microbiome of individuals with treatment-naive type 2 diabetes, contributing to the therapeutic effects of the drug

Authors: Hao Wu, Eduardo Esteve, Valentina Tremaroli, Muhammad Tanweer Khan, Robert Caesar, Louise Mannerås-Holm, Marcus Ståhlman, Lisa M Olsson, Matteo Serino, **Mercè Planas-Fèlix**, Gemma Xifra, Josep M Mercader, David Torrents, Rémy Burcelin, Wifredo Ricart, Rosie Perkins, José Manuel Fernández-Real, Fredrik Bäckhed

Journal: Nature medicine

Impact factor: 32.621

Citations:392

Contribution: Ph.D. Candidate Mercè Planas-Fèlix contribution to this study involved the study and analysis of 16s data from patients. This involved, the design and subsequent analysis of the study with the 16s samples, the results of which led to the new focus of the article.

# Metformin alters the gut microbiome of individuals with treatment-naive type 2 diabetes, contributing to the therapeutic effects of the drug

Hao Wu<sup>1,12</sup>, Eduardo Esteve<sup>2-4,12</sup>, Valentina Tremaroli<sup>1</sup>, Muhammad Tanweer Khan<sup>1</sup>, Robert Caesar<sup>1</sup>, Louise Mannerås-Holm<sup>1</sup>, Marcus Ståhlman<sup>1</sup>, Lisa M Olsson<sup>1</sup>, Matteo Serino<sup>5</sup>, Mercè Planas-Fèlix<sup>6</sup>, Gemma Xifra<sup>2-4</sup>, Josep M Mercader<sup>6</sup>, David Torrents<sup>6,7</sup>, Rémy Burcelin<sup>8,9</sup>, Wifredo Ricart<sup>2-4</sup>, Rosie Perkins<sup>1</sup>, José Manuel Fernández-Real<sup>2-4</sup> & Fredrik Bäckhed<sup>1,10,11</sup>

Metformin is widely used in the treatment of type 2 diabetes (T2D), but its mechanism of action is poorly defined. Recent evidence implicates the gut microbiota as a site of metformin action. In a double-blind study, we randomized individuals with treatment-naive T2D to placebo or metformin for 4 months and showed that metformin had strong effects on the gut microbiome. These results were verified in a subset of the placebo group that switched to metformin 6 months after the start of the trial. Transfer of fecal samples (obtained before and 4 months after treatment) from metformin-treated donors to germ-free mice showed that glucose tolerance was improved in mice that received metformin-altered microbiota. By directly investigating metformin–microbiota interactions in a gut simulator, we showed that metformin affected pathways with common biological functions in species from two different phyla, and many of the metformin-regulated genes in these species encoded metalloproteins or metal transporters. Our findings provide support for the notion that altered gut microbiota mediates some of metformin's antidiabetic effects.

Metformin is the most prescribed pharmacotherapy for the treatment of individuals with type 2 diabetes (T2D) because of its relative safety, low cost, and beneficial effects on blood glucose and cardiovascular mortality<sup>1,2</sup>. However, its mechanism of action remains unclear. Although metformin is generally considered to mediate its antihyperglycemic effects by suppressing hepatic glucose output through the activation of AMP-activated protein kinase (AMPK)-dependent<sup>3-5</sup> and AMPK-independent pathways<sup>6-8</sup> in the liver, accumulating evidence indicates that it might also act through pathways in the gut<sup>9,10</sup>. For example, its glucose-lowering effect is more pronounced when given orally than when administered intravenously<sup>11</sup>. In addition, a study comparing metformin formulations with reduced and normal plasma exposure provided evidence to indicate that the lower bowel is a major site of action for metformin<sup>12</sup>. Furthermore, recent studies in both rodents<sup>13-15</sup> and humans<sup>16-18</sup> suggest that gut microbial changes might contribute to the antidiabetic effect of metformin. So far, however, it is not known how metformin affects the gut microbiota of individuals with treatment-naive T2D, nor how metformin interacts with gut bacteria.

Here we performed a randomized, placebo-controlled, double-blind study in individuals with newly diagnosed T2D on a calorie-restricted diet, and we combined metagenomics and targeted metabolomics to investigate the effect of metformin on the composition and function of the gut microbiota. We also transferred human fecal samples to germ-free mice to study the effects of metformin-altered microbiota on host glucose metabolism, and we used an *in vitro* gut simulator to investigate metformin–microbiota interactions directly.

## RESULTS

### Metformin alters the gut microbiota composition

To investigate how metformin affects the composition of the gut microbiota, we randomized treatment-naive individuals with recently diagnosed T2D to receive either placebo ( $n = 18$ ) or 1,700 mg/d of metformin ( $n = 22$ ) for 4 months in a double-blind study. Clinical characteristics of these individuals before and after treatment are presented in **Table 1**. Both groups were recommended to consume a calorie-restricted diet for the 4-month study period (**Supplementary Table 1**); calorie intake was reduced by a median of 342 kcal/d, and

<sup>1</sup>Department of Molecular and Clinical Medicine, Wallenberg Laboratory, Institute of Medicine, University of Gothenburg, Gothenburg, Sweden. <sup>2</sup>Department of Diabetes, Endocrinology and Nutrition, Institut d'Investigació Biomèdica de Girona, Hospital Josep Trueta, Girona, Spain. <sup>3</sup>Departament de Medicina, Facultat de Medicina, University of Girona, Girona, Spain. <sup>4</sup>Centro de Investigación Biomédica en Red de Fisiopatología de la Obesidad y Nutrición (CIBEROBN), Instituto de Salud Carlos III, Madrid, Spain. <sup>5</sup>IRSD, Université de Toulouse, INSERM, INRA, ENVT, UPS, Toulouse, France. <sup>6</sup>Barcelona Supercomputing Center (BSC), Joint BSC-CRG-IRB Research Program in Computational Biology, Barcelona, Spain. <sup>7</sup>Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain. <sup>8</sup>Institut National de la Santé et de la Recherche Médicale (INSERM), Toulouse, France. <sup>9</sup>Université Paul Sabatier (UPS), Unité Mixte de Recherche 1048, Institut de Maladies Métaboliques et Cardiovasculaires, Toulouse, France. <sup>10</sup>Sahlgrenska University Hospital, Gothenburg, Sweden. <sup>11</sup>Novo Nordisk Foundation Center for Basic Metabolic Research, Section for Metabolic Receptology and Endocrinology, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark. <sup>12</sup>These authors contributed equally to this work. Correspondence should be addressed to J.M.F.-R. (jmfreal@idibgi.org) or F.B. (fredrik.backhed@wlab.gu.se).

Received 27 June 2016; accepted 19 April 2017; published online 22 May 2017; doi:10.1038/nm.4345

**Table 1** Clinical characteristics for the 40 individuals with T2D enrolled in this study

	Placebo group (n = 18)			Metformin group (n = 22)		
	P0	P2	P4	M0	M2	M4
Age (years)	54.9 ± 1.9	–	–	52.6 ± 2.0	–	–
Sex (male/female)	9/9	–	–	8/14	–	–
Weight (kg)	85.4 ± 5.6	82.2 ± 5.6 <sup>+</sup>	81.5 ± 5.4 <sup>#</sup>	96.5 ± 4.1	92.9 ± 4.0	91.4 ± 3.9
Waist circumference (cm)	106.1 ± 4.4	101.7 ± 4.3 <sup>*</sup>	101.9 ± 3.6	111.5 ± 2.7	108.3 ± 2.9 <sup>*</sup>	108.7 ± 2.9
HOMA	8.0 ± 1.5	8.9 ± 1.6	8.1 ± 1.8	8.3 ± 1.2	6.2 ± 0.9	6.0 ± 0.8 <sup>*</sup>
Total cholesterol (mg/dl)	205.8 ± 8.8	197.8 ± 7.7	190.7 ± 6.9 <sup>*</sup>	206.0 ± 7.4	196.6 ± 7.2	198.8 ± 7.5
HDL-C (mg/dl)	46.9 ± 3.4	45.0 ± 3.0	46.8 ± 3.1	48.4 ± 2.7	55.2 ± 6.1	51.1 ± 3.0 <sup>*</sup>
LDL-C (mg/dl)	126.8 ± 6.6	124.8 ± 5.7	118.1 ± 6.2 <sup>*</sup>	129.4 ± 6.4	117.4 ± 6.2 <sup>*</sup>	121.5 ± 6.8
Triglycerides (mg/dl)	151.9 ± 18.7	155.9 ± 15.2	129.3 ± 12.5	129.0 ± 17.8	139.6 ± 11.6	135.9 ± 12.7
ALT (U/liter)	33.2 ± 7.2	24.3 ± 2.9	22.3 ± 2.1	35.5 ± 3.5	28.0 ± 1.8 <sup>*</sup>	32.8 ± 3.2
GGT (U/liter)	38.4 ± 5.4	28.2 ± 2.1 <sup>+</sup>	26.3 ± 2.0 <sup>+</sup>	44.0 ± 6.0	31.3 ± 3.2 <sup>+</sup>	34.1 ± 3.9 <sup>*</sup>
CRP (mg/dl)	0.4 ± 0.1	0.6 ± 0.1	0.5 ± 0.1	0.4 ± 0.1	0.4 ± 0.1	0.4 ± 0.1
Statin treatment (n)	3	–	–	4	–	–
Antihypertensive treatment (n)	2	–	–	3	–	–

ALT, alanine transaminase; CRP, C-reactive protein; GGT,  $\gamma$ -glutamyl transferase; HDL-C, high-density lipoprotein cholesterol; HOMA, homeostatic model assessment; LDL-C, low-density lipoprotein cholesterol. <sup>\*</sup> $P < 0.05$ ; <sup>+</sup> $P < 0.01$ ; <sup>#</sup> $P < 0.001$  versus P0 or M0. Wilcoxon signed-rank test; data are shown as means  $\pm$  s.e.m.

no significant differences were seen between the groups ( $P = 0.90$ ). A subset of the placebo group switched to receive metformin (850 or 1,700 mg/d;  $n = 13$ ) 6 months after the start of the study; to validate our findings from the randomized study, we analyzed samples from this group after a further 6 months.

As expected given the reduced caloric intake, body-mass index (BMI) decreased significantly in both the placebo and metformin groups over the initial 4-month study period (Fig. 1a). However, significant decreases in % hemoglobin A1c (HbA1c) and fasting blood glucose were observed only in the group randomized to metformin treatment (Fig. 1b,c). BMI did not decrease further in the switched subgroup after 6 months on metformin (Fig. 1a), but %HbA1c and fasting blood glucose were significantly reduced by metformin in this subgroup (Fig. 1b,c).

To characterize the effects of metformin on the gut microbiome, we performed whole-genome shotgun sequencing of 131 fecal samples. On average, we obtained 38 million paired-end reads for each sample (ranging from 15 million to 116 million; Supplementary Table 2). The taxonomy and gene profiles were estimated by mapping the high-quality reads to nonredundant genome and gene catalogs implemented in the metagenomic data-utilization and analysis (MEDUSA) pipeline<sup>19</sup>, respectively. Only one bacterial strain was altered over the 4-month study period in the placebo group (Fig. 1d), despite the reduction in BMI. By contrast, metformin treatment for 2 and 4 months resulted in significant alterations in the relative abundance of 81 and 86 bacterial strains, respectively, most of which belonged to  $\gamma$ -proteobacteria (for example, *Escherichia coli*) and Firmicutes (Fig. 1d and Supplementary Table 3; false-discovery rate (FDR)  $< 0.05$ ). At the genera level, we observed an increase of *Escherichia* and a decrease of *Intestinibacter* in the metformin-treated group (Supplementary Table 3). Notably, the microbial changes observed after 2 and 4 months of metformin treatment in our randomized study correlated with the microbial changes observed in the switched subgroup after 6 months on metformin (Fig. 1e). We also observed a metformin-induced increase in *Bifidobacterium* in this subgroup (Supplementary Table 3).

Earlier studies have shown an association between metformin and the abundance of *Akkermansia muciniphila*<sup>13–15,18</sup> and between *A. muciniphila* and improved metabolic features in mice<sup>13,20,21</sup> and humans<sup>22</sup>. In a targeted analysis of our metagenome data, we showed

increased abundance of *A. muciniphila* in individuals who received metformin for 4 months (Supplementary Fig. 1). However, we did not observe any significant correlations between %HbA1c and *A. muciniphila* abundance in our cohort ( $P > 0.1$ , Supplementary Fig. 1).

To investigate how different gut bacteria interact with each other, we performed a coabundance network analysis. We showed that 2 months of metformin treatment promoted an increased number of positive connections among microbial genera, especially those within Proteobacteria and Firmicutes (Fig. 1f). We also identified a few interphylum connections, such as between *Shewanella* (Proteobacteria) and *Blautia* (Firmicutes), a short-chain fatty acid (SCFA)-producing genus<sup>23</sup>.

To test the effect of metformin on microbial growth, we mapped whole-genome shotgun reads to the genomes of common strains in the human gut to determine the ratio between DNA copy number near the replication origin and DNA copy number near the terminus (termed the peak-to-trough ratio, PTR) of bacterial genomes<sup>24</sup>. After correction for FDR, we found that the PTR of only one bacterial species (*Bifidobacterium adolescentis*) was significantly increased by metformin (Fig. 2a). In agreement, the PTR of *B. adolescentis* was also increased in the switched subgroup after 6 months on metformin (Fig. 2a). Furthermore, in our cohort, we observed a negative correlation between the PTR of *B. adolescentis* and %HbA1c (Spearman coefficient  $\rho = -0.28$ ,  $P < 0.01$ ). Consistent with this observation, *in vitro* analysis showed that metformin directly promoted the growth of *B. adolescentis* in pure cultures (Fig. 2b). We also showed that metformin directly promoted the growth of *A. muciniphila*, but not of *E. coli*, in pure cultures (Fig. 2c,d).

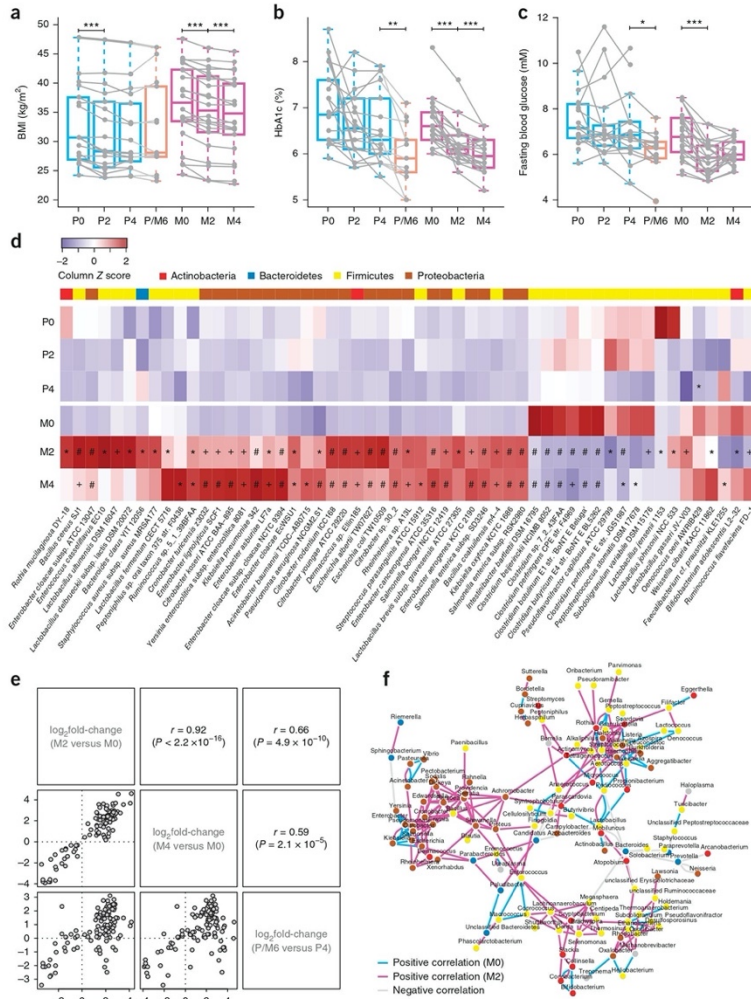
#### Metformin-altered microbiota improves glucose tolerance

To investigate whether metformin-altered microbiota could contribute to the glucose-lowering effect of metformin, we transferred fecal samples from three metformin-treated participants (before and 4 months after metformin, here termed M0 and M4 microbiota) to germ-free mice. All three of the metformin recipients responded similarly to metformin in terms of reduced %HbA1c, as compared to baseline, after 2 and 4 months on metformin. The mice were fed a high-fat diet for 1 week before and during colonization for 18 d.

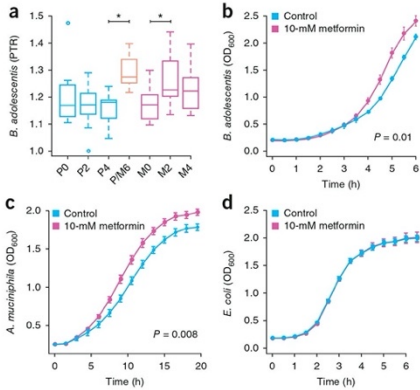
We did not observe any differences in body weight, body fat, or fasting insulin between mice that received M4 and M0 microbiota (Fig. 3a,b and Supplementary Fig. 2a–c). However, we found improvements

ARTICLES

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.



**Figure 1** Metformin treatment promotes rapid changes in the composition of the gut microbiota. (a–c) Boxplots (with median) showing BMI, %HbA1c, and fasting blood glucose before treatment (P0 and M0) and after 2 and 4 months in individuals with T2D randomized to placebo (P2 and P4;  $n = 18$ ) or metformin (M2 and M4;  $n = 22$ ), and 6 months after metformin in a subgroup that switched from placebo to metformin after the randomized study period (P/M6;  $n = 13$ ). Wilcoxon signed-rank test; \* $P < 0.05$ ; \*\* $P < 0.01$ ; \*\*\* $P < 0.001$ . (d) Heat map showing changes in the abundance of bacterial strains after placebo or metformin treatment (only strains with >50 reads mapped are shown). Wald test; \*FDR < 0.05; \*FDR < 0.01; #FDR < 0.001. (e) Pearson correlations between microbial changes observed at M2 as compared to M0; M4 as compared to M0; and P/M6 as compared to P4. (f) Genus–genus coabundance network before (M0) and after 2 months of metformin treatment (M2) in individuals with T2D. The edges indicate Spearman correlations of >0.6 or <-0.6 between genera present in at least 80% samples.

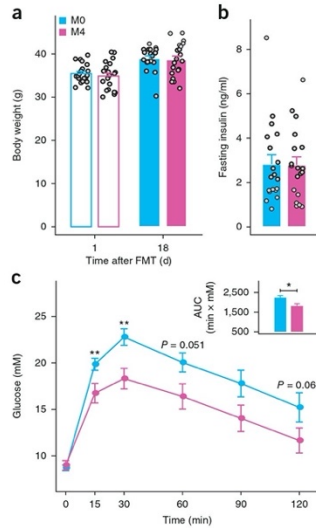


**Figure 2** Metformin treatment promotes the growth of gut bacteria. (a) Boxplots (with median) showing *B. adolescentis* growth as estimated by peak-to-trough ratio (PTR) before treatment (P0 and M0) and after 2 and 4 months in individuals with T2D randomized to placebo (P2 and P4;  $n = 18$ ) or metformin (M2 and M4;  $n = 22$ ) and 6 months after metformin in a subgroup that switched from placebo to metformin after the randomized study period (P/M6;  $n = 13$ ). Wilcoxon signed-rank test; \*FDR < 0.05. (b–d) Growth of *B. adolescentis*, *A. muciniphila*, and *E. coli* as single cultures in the presence or absence of 10-mM metformin (with six technical replicates).  $P$  values were determined by two-way analysis of variance (ANOVA) with repeated measurements. Data are shown as means  $\pm$  s.e.m.

in glucose tolerance in mice that received M4 microbiota, as compared to those that received M0 microbiota, from two of the three donors (Supplementary Fig. 2d–f) and when combining the results from all three transfer experiments (Fig. 3c).

#### Metformin promotes functional shifts in the gut microbiome

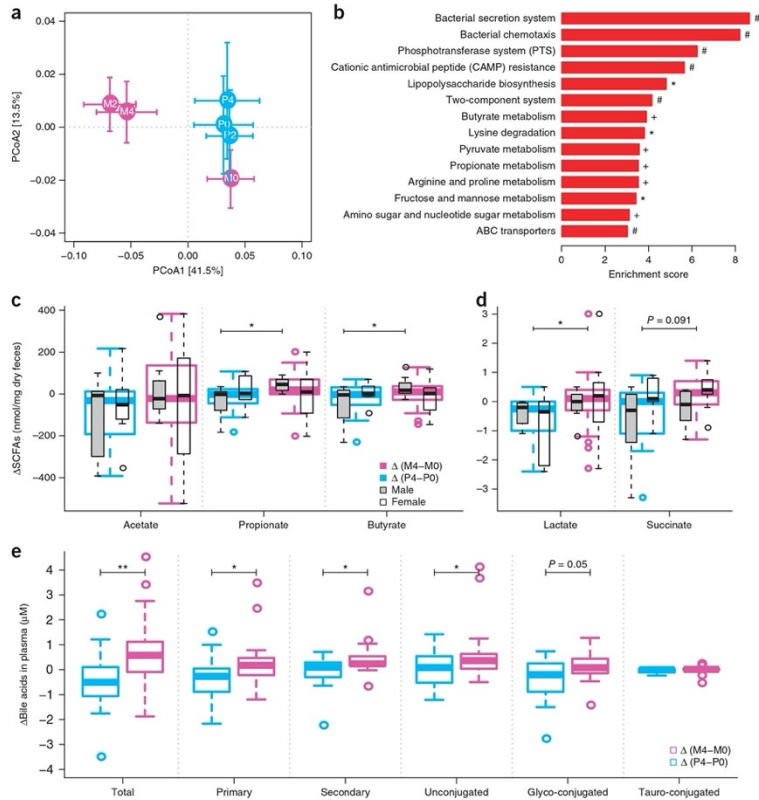
To further investigate functional changes in the gut microbiome after metformin treatment, we annotated genes to Kyoto encyclopedia of genes and genomes (KEGG) orthology (KO)<sup>25</sup>. Only two KOs were significantly altered over the 4-month study period in the placebo group (FDR < 0.05). By contrast, 626 and 473 KOs were increased, whereas 130 and 69 KOs were decreased after 2 and 4 months of metformin, respectively (Supplementary Fig. 3 and Supplementary Table 4; FDR < 0.05), and most of the shifts were consistent between the two sampling times (Supplementary Fig. 3). Principal coordinate analysis (PCoA) of the relative abundance of all of the significantly altered KOs revealed similar gene functions in the placebo group at all time points and the metformin group at baseline (i.e., before treatment), but we observed significant shifts after metformin treatment for 2 months and 4 months (Fig. 4a). Pathway-enrichment analysis revealed that metformin treatment was linked mainly to the enrichment of genes for bacterial environmental responses (for example, bacterial secretion system, two component system, and ATP-binding cassette (ABC) transporters), drug resistance (bacterial chemotaxis and cationic antimicrobial peptide resistance), central carbohydrate metabolism (phosphotransferase system, pyruvate, butyrate, and propionate metabolism), amino acid metabolism, and lipopolysaccharide (LPS) biosynthesis (Fig. 4b and Supplementary Table 4; FDR < 0.05).



**Figure 3** Metformin-altered microbiota improves glucose tolerance. (a) Body weight of mice 1 d and 18 d after colonization with fecal microbiota obtained from three individuals with T2D before (M0) and 4 months after metformin treatment (M4). (b) Fasting plasma insulin concentrations measured in the same mice used in a 18 d after colonization. (c) Plasma glucose concentrations measured in the same mice used in a during an intraperitoneal glucose-tolerance test 18 d after colonization. Data are shown as means  $\pm$  s.e.m. and are the combined results of three independent transfer experiments (shown individually in Supplementary Fig. 2). M0:  $n = 20$  mice; M4;  $n = 21$ . Wilcoxon rank-sum test; \* $P < 0.05$ ; \*\* $P < 0.01$ . FMT, fecal microbiota transplantation.

Although it is not clear how alterations in the gut microbiota promote beneficial effects in the host, a potential mechanism includes increased production of SCFAs, primarily acetate, propionate, and butyrate, and other organic acids<sup>26,27</sup>. We therefore performed targeted metabolomics to investigate whether the observed enrichment in genes for SCFA metabolism in the gut microbiome following metformin treatment was paralleled by an increased production of SCFAs. We observed significantly larger increases in fecal propionate and butyrate concentrations in the metformin group, as compared to the placebo group, after 4 months of treatment in men; however, no differences were observed when results from men and women were combined (Fig. 4c). We also observed significantly larger increases in fecal concentrations of lactate and a trend toward a larger increase in fecal concentrations of succinate in the metformin group, as compared to the placebo group, after 4 months of treatment (Fig. 4d).

The gut microbiota is also known to be a major regulator of bile acid metabolism<sup>28</sup>, which may contribute to its effects on host metabolism. Furthermore, a few studies have indicated a potential role of metformin in altering the bile acid profile<sup>29,30</sup>, but this link is not well established. Here we investigated the effect of metformin treatment



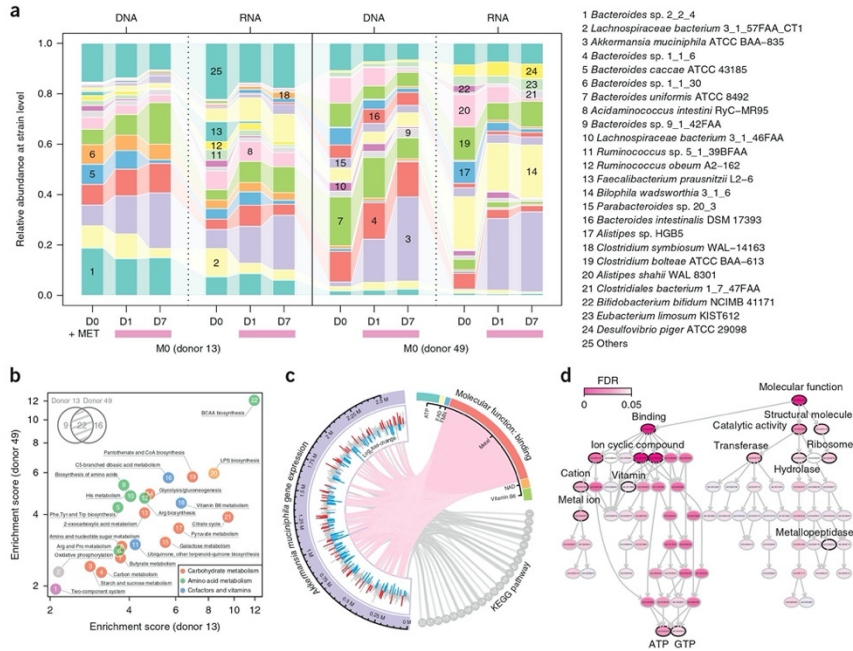
**Figure 4** Metformin treatment promotes functional shifts in the gut microbiota. **(a)** Principal coordinate analysis (PCoA) of all KOs that are significantly altered from baseline (P0 and M0) after 2 and 4 months in individuals with T2D randomized to placebo (P2 and P4;  $n = 18$ ) or metformin (M2 and M4;  $n = 22$ ). Adonis test based on 5,000 permutations;  $P_{M0 \text{ vs. } M2} = 0.00040$ ;  $P_{M0 \text{ vs. } M4} = 0.0062$ . Data are shown as means  $\pm$  s.e.m. **(b)** Pathway-enrichment analysis of all significantly altered KOs. Hypergeometric test; \*FDR < 0.05; \*FDR < 0.01; \*FDR < 0.001. **(c–e)** Boxplots (with median) showing changes from baseline (P0 and M0) for fecal concentrations of SCFAs **(c)**, fecal concentrations of lactate and succinate **(d)**, and plasma concentrations of bile acids **(e)** after 4 months in individuals with T2D randomized to placebo (P4;  $n = 18$ ) or metformin (M4;  $n = 22$ ). Wilcoxon rank-sum test; \* $P < 0.05$ ; \*\* $P < 0.01$ .

for 4 months on fecal and plasma bile acid composition. No substantial changes in fecal bile acids were detected following metformin treatment (**Supplementary Fig. 4a**). However, we observed significantly larger increases in plasma bile acid concentrations (total, primary, secondary, and unconjugated) in the metformin group, as compared to the placebo group, after 4 months of treatment (**Fig. 4e**). By using a targeted metagenomic analysis, we showed an increased abundance of *bsh* genes encoding bile salt hydrolases, after 2 months on metformin (**Supplementary Fig. 4b**). These enzymes are produced by the gut microbiota and catalyze the deconjugation of glycine- or taurine-conjugated bile acids, and thus increases in *bsh* could contribute to the

increased concentrations of unconjugated bile acids. Furthermore, we found a significant negative correlation between the concentrations of unconjugated bile acids and %HbA1c ( $\rho = -0.27$ ,  $P < 0.05$ ), which suggests a possible link between the modulation of bile acid composition and the therapeutic effect of metformin.

#### Direct effects of metformin on the gut microbiota

To directly investigate how metformin affects the gut microbiota, we cultured fecal samples (obtained before metformin treatment from two participants, donors 13 and 49) in two separate gut-simulator experiments, and exposed the samples to a constant flow of metformin



**Figure 5** Direct metformin-microbiota interactions identified using an *in vitro* gut simulator. **(a)** Microbial compositional profiling of samples taken from the *in vitro* gut simulator before (D0) and after 1 d and 7 d (D1 and D7) of metformin exposure (+MET). Fecal material from two participants (donors 13 and 49) was used to inoculate the gut simulator in two independent experiments (i.e., the fecal material from the two samples was not pooled). The top ten bacterial strains detected in the samples at each time point are shown. **(b)** Metformin-enriched pathways identified by both metagenomic and metatranscriptomic analyses of samples taken from the *in vitro* gut simulator. Hypergeometric test; FDR < 0.05. Only the 22 pathways affected by metformin exposure in both experiments (inset) are shown. **(c)** Circos plot of KEGG pathway and molecular-function annotations for metformin-regulated genes in *A. muciniphila*. The circular bar plot on the left side indicates metformin-regulated genes along the genome. Red, significantly increased gene expression; blue, significantly decreased; gray, no statistical differences; Wald test; FDR < 0.1. The metformin-regulated genes involved in the 22 pathways shown in **(b)** (gray links) and coding for proteins with metal-binding abilities (pink links) are highlighted. **(d)** Gene ontology (GO) enrichment of all metformin-regulated genes in *A. muciniphila*; GO terms with direct hierarchical parent-child relationships are linked by arrows. Hypergeometric test; FDR < 0.05.

(10 mM) for 1 week. We then profiled the microbiomes by whole-genome shotgun sequencing at both the DNA and RNA level.

Compositional profiling revealed that metformin exposure significantly altered the DNA and RNA abundance of 24 bacterial strains when culturing the feces of donor 13 but of only 4 for the feces of donor 49 (Supplementary Table 5; FDR < 0.05). Donor-specific effects of metformin exposure included, for example, increased RNA abundance of *Bifidobila wadsworthia* (donor 13) and increased DNA abundance of *Lachnospiraceae* bacterium (donor 49) (Fig. 5a and Supplementary Table 5). *A. muciniphila* was the only taxon that increased in both DNA and RNA abundance in response to metformin in both samples, and it was also the taxon that increased the most in abundance (Fig. 5a and Supplementary Table 5).

Functional profiling of the combined metagenome and metatranscriptome showed that metformin exposure significantly altered the

abundance of 686 and 909 KOs in samples from donors 13 and 49, respectively (Supplementary Table 6; Wald test, FDR < 0.05). In total, 31 and 38 pathways were enriched after metformin exposure in samples from donors 13 and 49, respectively; of these, 22 enriched pathways were common to both samples (Fig. 5b and Supplementary Table 6; hypergeometric test, FDR < 0.05). Six pathways that were enriched in the metformin-treated samples in the *in vivo* metagenome analysis (see Fig. 4b)—including those for genes involved in LPS synthesis, butyrate and pyruvate metabolism, and two-component systems—were also shown to be enriched by metformin in both gut-simulator experiments (Fig. 5b). In addition, the *in vitro* analysis revealed enrichment of metabolic pathways linked to the metabolism of cofactors and vitamins (Fig. 5b). These results show that although metformin exerts donor-specific taxonomic effects, it induces overlapping microbial functional changes in samples from both donors.



## ARTICLES

Finally, we performed in-depth transcriptome analyses (using the *in vitro* cultured fecal sample from donor 13) to investigate direct interactions between metformin and individual bacterial species. We first examined the RNA reads that mapped to the gene catalog of *A. muciniphila* (the taxon with the overall highest abundance in this fecal sample; Fig. 5a). We found that nearly 10% (207/2138) of the protein-coding genes in *A. muciniphila* were significantly regulated by metformin; of these, 65% were downregulated by metformin (Supplementary Table 7; FDR < 0.1). Furthermore, 78 of the 207 metformin-regulated genes could be annotated to KOs, and of these, 41 genes mapped to the 22 metformin-enriched pathways common to cultured fecal samples from both donors (Fig. 5c and Supplementary Table 7). By manual annotation, we found that the protein products of 108/207 metformin-regulated genes required cofactors or coenzymes such as ATP, FAD, FMN, metal, NAD, and vitamin B6 (Fig. 5c and Supplementary Table 7); most of the remaining genes (63/99) have not been characterized (Supplementary Table 7). Of particular interest, 81 of the 108 metformin-regulated annotated genes encoded metalloprotein or metal transporters (Fig. 5c and Supplementary Table 7). Gene ontology (GO) analysis of metformin-regulated genes in *A. muciniphila* confirmed that their gene products were enriched in proteins that bind to metal ions in addition to several other cofactors and coenzymes, as well as transferase, hydrolase, ligase, and protein components of ribosomes (Fig. 5d). To address whether those observations were specific to *A. muciniphila*, we also analyzed *B. wadsworthia* (the second most abundant taxon in this cultured fecal sample after metformin treatment; Fig. 5a). According to protein-homology detection, only 14 metformin-regulated genes were orthologous between these two bacteria; however, most annotated metformin-regulated genes in *B. wadsworthia* also encoded metalloproteins (Supplementary Table 7).

### DISCUSSION

In this study, we performed a randomized, placebo-controlled, double-blind study in individuals with newly diagnosed T2D on a calorie-restricted diet and showed that metformin, but not calorie restriction, had rapid effects on the composition and function of the gut microbiota in parallel with the reduction of %HbA1c and fasting blood glucose concentrations. Transfer of the microbiota to germ-free mice showed that the metformin-altered microbiota could improve glucose metabolism. Furthermore, transcriptome analyses of feces cultured with metformin *in vitro* in a gut simulator showed that metformin had direct effects on the microbiota and regulated the expression of genes encoding metalloproteins in the gut bacteria.

By using paired samples in our prospective human study, we reduced the effect of interindividual variations, a common issue in previous studies investigating the effect of metformin on the microbiota<sup>16,17,31</sup>. A further strength of our cohort is that these individuals had been newly diagnosed with T2D and thus were not taking other medications for T2D, and only a small number in each group were taking statins or antihypertensive therapy. We also monitored the dietary intake of the participants before and 4 months after treatment and showed that calorie restriction did not affect the gut microbiome to any great extent in our study; it should be noted that the calorie reduction reported was mild relative to that in earlier studies showing profound changes in the gut microbiome in response to dietary intervention<sup>22,32</sup>. The design of our study thus enabled us to minimize the effect of major confounding factors known to have an impact on the gut microbiome.

By performing whole-genome shotgun sequencing of fecal samples, we observed dramatic shifts in the composition of the gut microbiota after

2 and 4 months on metformin in individuals with newly diagnosed T2D. Notably, these changes were similar to those observed after 6 months on metformin in a placebo subgroup that switched to metformin 6 months after the study start. In particular, we observed significant changes in *Escherichia* and *Intestinibacter* abundance across all sampling points in the metformin-treated group, a finding that is in agreement with results reported in a cross-sectional study that compared metformin-treated and untreated groups of people with T2D<sup>17</sup>. Growth of *E. coli* in an *in vitro* analysis was not affected by metformin. Thus, the effects of metformin on the abundance of *Escherichia* spp. are likely indirect, and possibly, a result of modified bacteria–bacteria interactions or of other physiological and/or environmental changes within the gut upon metformin treatment. We showed that metformin promoted the growth of *B. adolescentis* both *in vivo* (after 2 months in the main study and after 6 months in the switched subgroup, as measured by PTR) and *in vitro* using pure cultures, and also that it increased the abundance of *Bifidobacterium* in the switched subgroup. Supplementation with *B. adolescentis* in a rodent model of the metabolic syndrome has previously been shown to increase insulin sensitivity<sup>33</sup>. In our cohort, we also observed a negative correlation between the PTR of *B. adolescentis* and %HbA1c, which suggests that increased growth of this bacterial species could potentially contribute to the antidiabetic effect of metformin.

To observe direct metformin–microbiota interactions, we incubated fecal samples from treatment-naïve participants with metformin in a gut simulator. In this system, we did not observe any significant changes in *E. coli* or *B. adolescentis*. In fact, the only taxon that increased in response to metformin (at both DNA and RNA levels and in samples from two separate donors) was *A. muciniphila*. Metformin has previously been shown to increase the abundance of *A. muciniphila* in rodents on a high-fat diet<sup>13–15</sup>, and this increased abundance has been linked to improved glucose metabolism<sup>13,20,21</sup>. However, evidence for a link between metformin and *A. muciniphila* is less clear in humans<sup>16,17</sup>. We did observe a significant increase in the abundance of *A. muciniphila* over time in the individuals who received metformin for 4 months, but only when we used a targeted analysis. Similarly, a recent study that screened mucin-degrading and butyrate-producing bacteria showed increased abundance of *A. muciniphila* in humans taking metformin<sup>18</sup>. In agreement with these observations, we showed that metformin increased the growth of *A. muciniphila* *in vitro* using pure cultures. However, it is likely that the growth of this taxon is affected in humans *in vivo* by factors that differ between individuals, such as fiber<sup>34</sup> and polyphenol availability<sup>35,36</sup>, immune responses<sup>37,38</sup>, and age<sup>39,40</sup>. Furthermore, in our study, we did not observe any significant correlations between %HbA1c and *A. muciniphila* abundance, and therefore, cannot conclude that *A. muciniphila* is a major contributor to the beneficial effects of metformin in our human cohort.

By comparing results from the *in vivo* metagenomics analysis and the *in vitro* metagenomics and metatranscriptomics analyses, we noted that metformin promoted consistent shifts in microbial functions, including LPS biosynthesis and SCFA metabolism. Increased LPS biosynthesis might reflect the increased abundance of Gram-negative bacteria such as Proteobacteria, but it was not associated with increased systemic inflammation, because C-reactive protein was unaltered (Table 1). Similarly, enrichment of the LPS biosynthesis pathway without an increase in inflammation has been observed both in humans after bariatric surgery<sup>41</sup> and in prebiotic-treated mice on a high-fat diet<sup>42</sup>. Increased SCFA metabolism in response to metformin has also been predicted in earlier metagenomics

analyses in humans<sup>17,18</sup>, and in agreement, our targeted metabolomics analysis showed that metformin significantly increased butyrate and propionate in men.

In-depth transcriptome analysis of the effects of metformin on two distantly related bacterial species in the gut simulator showed that most of the metformin-regulated genes encoded metalloproteins or metal transporters. It is unlikely that the transcriptional responses of this subset of genes were due to a growth-induced increase in total transcripts because the majority of these genes were downregulated by metformin. Interestingly, some metals are known to contribute to T2D pathophysiology<sup>43</sup>, and it has been known for many years that metformin binds to metals<sup>44</sup>. Furthermore, a recent study showed that the effects of metformin on a mammalian liver cell line are dependent on the metal-binding properties of this drug<sup>44</sup>. However, our study, to the best of our knowledge, is among the first to indicate a link between metformin and metal-binding proteins produced by the gut microbiota.

Fecal transfer to germ-free mice resulted in improved glucose tolerance in recipients of metformin-altered microbiota from two out of three donors, with overall substantial improvement of glucose metabolism, thus indicating that metformin-adapted microbiota could contribute to the beneficial effects of metformin on glucose homeostasis. It is not clear why the responses to the metformin-altered microbiota differed between the donors, given that all the donors showed improved glucose tolerance after both 2 and 4 months of metformin treatment. However, there are large interindividual differences in gut-microbiota composition in humans, and the lack of response of microbiota from one donor might be attributable to an incomplete transfer of key species, as has previously been described<sup>45</sup>. Furthermore, the different diets of the recipient mice and the human donors (i.e., high fat as opposed to calorie restriction) would likely exacerbate differences in the gut-microbiota composition between the donors and recipients. The taxa that are successfully transferred to recipient mice will therefore be dependent not only on the composition of the donor gut microbiota, but also on how well the taxa respond to different macronutrients.

There is increasing evidence to indicate that SCFAs and bile acids have a role in the regulation of glucose homeostasis<sup>26,27,46,47</sup>, and we recently reported that microbiota-produced succinate could improve glucose metabolism by activating intestinal gluconeogenesis in mice<sup>48</sup>. Here we observed metformin-induced alterations in these microbially regulated metabolites, which suggests that they might be partly responsible for the stronger glucose-lowering effect that has been observed when metformin is administered orally as compared with intravenous injection<sup>11</sup>. It should be noted that we cannot conclude that the major mechanism of action of metformin is through the microbiota. For example, a recent study in mice showed that the phosphorylation of acetyl-CoA carboxylases (ACC) 1 and 2 by AMPK is required to observe the insulin-sensitizing effects of metformin<sup>5</sup>, demonstrating the importance of AMPK/ACC signaling. However, it is possible that the gut microbiota might also act through ACCs, given that we previously showed that diet-induced obesity involved cross-talk between the microbiota, AMPK, and downstream ACC2 phosphorylation<sup>49</sup>.

In summary, our work shows that metformin interacts with different gut bacteria, possibly through the regulation of metal homeostasis. However, additional studies combining untargeted metabolomics and metaproteomics are essential to identify further microbial metabolites or proteins and to determine how they interact with the host targets in improving host metabolism.

## METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

*Note: Any Supplementary Information and Source Data files are available in the online version of the paper.*

## ACKNOWLEDGMENTS

We thank C. Arvidsson, S. Nordin-Larsson, C. Wennberg, and U. Enqvist for superb mouse husbandry. The administrative and technical help of J.M. Moreno Navarrete, E. Huertos, M. Sabater, and O. Rovira is also acknowledged. The strain *Akkermansia muciniphila* DSM22959 was kindly provided by W. de Vos (Wageningen University and Helsinki University). The strain *Bifidobacterium adolescentis* L2-32 was kindly provided by K. Scott (The Rowett Institute of Nutrition and Health, University of Aberdeen). Whole-genome shotgun sequencing was performed at the Genomics Core Facility at the Sahlgrenska Academy, University of Gothenburg. The computations for metagenomics analyses were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX). This study was supported by the Swedish Diabetes Foundation; Swedish Research Council; Swedish Heart Lung Foundation; Torsten Söderberg's Foundation; Göran Gustafsson's Foundation; Inga Britt and Arne Lundberg's Foundation; Swedish Foundation for Strategic Research; Knut and Alice Wallenberg Foundation; the Novo Nordisk Foundation; the regional agreement on medical training and clinical research (ALF) between Region Västra Götaland and Sahlgrenska University Hospital; the Ministerio de Economía y Competitividad (PI11-00214 and PI15/01934); and FEDER funds. CIBEROBN Fisiopatología de la Obesidad y Nutrición is an initiative from the Instituto de Salud Carlos III from Spain. M.P.-F. is funded by the Obra Social Fundación la Caixa fellowship under the Severo Ochoa 2013 program. J.M.M. was supported by the Sara Borrell Fellowship from the Instituto Carlos III, EFSD/Lilly Research Fellowship and Beatrice de Pinós Fellowship from the Agency for Management of University and Research Grants (AGAUR). E.B. is a recipient of ERC Consolidator Grant (European Research Council, Consolidator grant 615362—METABASE).

## AUTHOR CONTRIBUTIONS

E.B., J.M.F.-R., V.T., and R.B. conceived and designed the study. E.E., M.P.-F., G.X., J.M.M., D.T., W.R., and J.M.F.-R. recruited cohort individuals and performed the clinical study. H.W., V.T., and E.B. conducted the bioinformatics study, analyzed all results unless otherwise indicated. H.W., V.T., R.P., and E.B. wrote the paper. M.T.K. performed the *in vitro* gut simulator and bacterial growth experiments. R.C. and L.M.-H. performed and analyzed the fecal microbiota transplantation experiments. M. Ståhlman performed the metabolomics experiments. M. Serino and V.T. extracted the bacterial DNA and discussed the results. L.M.O. and V.T. extracted the bacterial RNA and coordinated the metagenomics and metatranscriptomics sequencing. All authors commented on the manuscript.

## COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

- Nathan, D.M. *et al.* Medical management of hyperglycemia in type 2 diabetes: a consensus algorithm for the initiation and adjustment of therapy: a consensus statement of the American Diabetes Association and the European Association for the Study of Diabetes. *Diabetes Care* **32**, 193–203 (2009).
- Pernicova, I. & Korbonits, M. Metformin—mode of action and clinical implications for diabetes and cancer. *Nat. Rev. Endocrinol.* **10**, 143–156 (2014).
- Zhou, G. *et al.* Role of AMP-activated protein kinase in mechanism of metformin action. *J. Clin. Invest.* **108**, 1167–1174 (2001).
- Shaw, R.J. *et al.* The kinase LKB1 mediates glucose homeostasis in liver and therapeutic effects of metformin. *Science* **310**, 1642–1646 (2005).
- Fullerton, M.D. *et al.* Single phosphorylation sites in *Acc1* and *Acc2* regulate lipid homeostasis and the insulin-sensitizing effects of metformin. *Nat. Med.* **19**, 1649–1654 (2013).
- Foretz, M. *et al.* Metformin inhibits hepatic gluconeogenesis in mice independently of the LKB1/AMPK pathway via a decrease in hepatic energy state. *J. Clin. Invest.* **120**, 2355–2369 (2010).
- Madriraju, A.K. *et al.* Metformin suppresses gluconeogenesis by inhibiting mitochondrial glycerolphosphate dehydrogenase. *Nature* **510**, 542–546 (2014).
- Miller, R.A. *et al.* Biguanides suppress hepatic glucagon signalling by decreasing production of cyclic AMP. *Nature* **494**, 256–260 (2013).

## ARTICLES

9. McCreight, L.J., Bailey, C.J. & Pearson, E.R. Metformin and the gastrointestinal tract. *Diabetologia* **59**, 426–435 (2016).
10. Duca, F.A. *et al.* Metformin activates a duodenal Ampk-dependent pathway to lower hepatic glucose production in rats. *Nat. Med.* **21**, 506–511 (2015).
11. Stepensky, D., Friedman, M., Raz, I. & Hoffman, A. Pharmacokinetic-pharmacodynamic analysis of the glucose-lowering effect of metformin in diabetic rats reveals first-pass pharmacodynamic effect. *Drug Metab. Dispos.* **30**, 861–868 (2002).
12. Buse, J.B. *et al.* The primary glucose-lowering effect of metformin resides in the gut, not the circulation. Results from short-term pharmacokinetic and 12-week dose-ranging studies. *Diabetes Care* **39**, 198–205 (2016).
13. Shin, N.R. *et al.* An increase in the *Akkermansia* spp. population induced by metformin treatment improves glucose homeostasis in diet-induced obese mice. *Gut* **63**, 727–735 (2014).
14. Zhang, X. *et al.* Modulation of gut microbiota by berberine and metformin during the treatment of high-fat diet-induced obesity in rats. *Sci. Rep.* **5**, 14405 (2015).
15. Lee, H. & Ko, G. Effect of metformin on metabolic improvement and gut microbiota. *Appl. Environ. Microbiol.* **80**, 5935–5943 (2014).
16. Karlsson, F.H. *et al.* Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* **498**, 99–103 (2013).
17. Forslund, K. *et al.* Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. *Nature* **528**, 262–266 (2015).
18. de la Cuesta-Zuluaga, J. *et al.* Metformin is associated with higher relative abundance of mucin-degrading *Akkermansia muciniphila* and several short-chain fatty acid-producing microbiota in the gut. *Diabetes Care* **40**, 54–62 (2017).
19. Karlsson, F.H., Nookaew, I. & Nielsen, J. Metagenomic data utilization and analysis (MEDUSA) and construction of a global gut microbial gene catalogue. *PLoS Comput. Biol.* **10**, e1003706 (2014).
20. Everard, A. *et al.* Cross-talk between *Akkermansia muciniphila* and intestinal epithelium controls diet-induced obesity. *Proc. Natl. Acad. Sci. USA* **110**, 9066–9071 (2013).
21. Plovier, H. *et al.* A purified membrane protein from *Akkermansia muciniphila* or the pasteurized bacterium improves metabolism in obese and diabetic mice. *Nat. Med.* **23**, 107–113 (2017).
22. Dao, M.C. *et al.* *Akkermansia muciniphila* and improved metabolic health during a dietary intervention in obesity: relationship with gut microbiome richness and ecology. *Gut* **65**, 426–436 (2016).
23. Park, S.K., Kim, M.S., Roh, S.W. & Bae, J.W. *Blautia stercoris* sp. nov., isolated from human faeces. *Int. J. Syst. Evol. Microbiol.* **62**, 776–779 (2012).
24. Korem, T. *et al.* Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. *Science* **349**, 1101–1106 (2015).
25. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
26. Wong, J.M., de Souza, R., Kendall, C.W., Emam, A. & Jenkins, D.J. Colonic health: fermentation and short chain fatty acids. *J. Clin. Gastroenterol.* **40**, 235–243 (2006).
27. Koh, A., De Vadder, F., Kovatcheva-Datchary, P. & Bäckhed, F. From dietary fiber to host physiology: short-chain fatty acids as key bacterial metabolites. *Cell* **165**, 1332–1345 (2016).
28. Ridlon, J.M., Harris, S.C., Bhowmik, S., Kang, D.J. & Hylemon, P.B. Consequences of bile salt biotransformations by intestinal bacteria. *Gut Microbes* **7**, 22–39 (2016).
29. Casparly, W.F. *et al.* Alteration of bile acid metabolism and vitamin-B12-absorption in diabetics on biguanides. *Diabetologia* **13**, 187–193 (1977).
30. Scarpello, J.H., Hodgson, E. & Howlett, H.C. Effect of metformin on bile salt circulation and intestinal motility in type 2 diabetes mellitus. *Diabet. Med.* **15**, 651–656 (1998).
31. Zhernakova, A. *et al.* Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* **352**, 565–569 (2016).
32. Collard, A. *et al.* Dietary intervention impact on gut microbial gene richness. *Nature* **500**, 585–588 (2013).
33. Chen, J., Wang, R., Li, X.F. & Wang, R.L. *Bifidobacterium adolescentis* supplementation ameliorates visceral fat accumulation and insulin sensitivity in an experimental model of the metabolic syndrome. *Br. J. Nutr.* **107**, 1429–1434 (2012).
34. Desai, M.S. *et al.* A dietary fiber-deprived gut microbiota degrades the colonic mucus barrier and enhances pathogen susceptibility. *Cell* **167**, 1339–1353.e21 (2016).
35. Roopchand, D.E. *et al.* Dietary polyphenols promote growth of the gut bacterium *Akkermansia muciniphila* and attenuate high-fat diet-induced metabolic syndrome. *Diabetes* **64**, 2847–2858 (2015).
36. Anhe, F.F. *et al.* A polyphenol-rich cranberry extract protects from diet-induced obesity, insulin resistance and intestinal inflammation in association with increased *Akkermansia* spp. population in the gut microbiota of mice. *Gut* **64**, 872–883 (2015).
37. Greer, R.L. *et al.* *Akkermansia muciniphila* mediates negative effects of IFN $\gamma$  on glucose metabolism. *Nat. Commun.* **7**, 13329 (2016).
38. Zhang, H., Sparks, J.B., Karyala, S.V., Settlage, R. & Luo, X.M. Host adaptive immunity alters gut microbiota. *ISME J.* **9**, 770–781 (2015).
39. Collado, M.C., Derrien, M., Isolauri, E., de Vos, W.M. & Salminen, S. Intestinal integrity and *Akkermansia muciniphila*, a mucin-degrading member of the intestinal microbiota present in infants, adults, and the elderly. *Appl. Environ. Microbiol.* **73**, 7767–7770 (2007).
40. Kong, F. *et al.* Gut microbiota signatures of longevity. *Curr. Biol.* **26**, R832–R833 (2016).
41. Tremaroli, V. *et al.* Roux-en-Y gastric bypass and vertical banded gastroplasty induce long-term changes on the human gut microbiome contributing to fat mass regulation. *Cell Metab.* **22**, 228–238 (2015).
42. Everard, A. *et al.* Microbiome of prebiotic-treated mice reveals novel targets involved in host response during obesity. *ISME J.* **8**, 2116–2130 (2014).
43. Fernández-Real, J.M. & Manco, M. Effects of iron overload on chronic metabolic diseases. *Lancet Diabetes Endocrinol.* **2**, 513–526 (2014).
44. Logie, L. *et al.* Cellular responses to the metal-binding properties of metformin. *Diabetes* **61**, 1423–1433 (2012).
45. Wahlström, A. *et al.* Induction of farnesoid X receptor signaling in germ-free mice colonized with a human microbiota. *J. Lipid Res.* **58**, 412–419 (2017).
46. Wahlström, A., Sayin, S.I., Marschall, H.U. & Bäckhed, F. Intestinal crosstalk between bile acids and microbiota and its impact on host metabolism. *Cell Metab.* **24**, 41–50 (2016).
47. Schaap, F.G., Trauner, M. & Jansen, P.L. Bile acid receptors as targets for drug development. *Nat. Rev. Gastroenterol. Hepatol.* **11**, 55–67 (2014).
48. De Vadder, F. *et al.* Microbiota-produced succinate improves glucose homeostasis via intestinal gluconeogenesis. *Cell Metab.* **24**, 151–157 (2016).
49. Bäckhed, F., Manchester, J.K., Semenkovich, C.F. & Gordon, J.I. Mechanisms underlying the resistance to diet-induced obesity in germ-free mice. *Proc. Natl. Acad. Sci. USA* **104**, 979–984 (2007).

## ONLINE METHODS

**Clinical study design.** 40 individuals with T2D were recruited and randomized (using a computational random generator Aleator) to treatment with metformin ( $n = 22$ ) or placebo ( $n = 18$ ) for 4 months. Metformin (Acyfabrik, Madrid, Spain) was started at a dose of 425 mg/d and increased progressively during the first week to reach 1,700 mg/d (in three doses). We instructed both groups to maintain a reduction in daily caloric intake of 500 kcal during the entire treatment. We recommended a hypocaloric diet containing 25 kcal/kg or 20 kcal/kg and used a validated food-frequency questionnaire<sup>50</sup>. The composition of the diet was 15% protein, 30% fat (<10% saturated fat), 55% carbohydrates, and 20–25 g dietary fiber (Supplementary Table 1). Lifestyle changes were also suggested, including regular physical activity (150 min/week). We collected both fecal and plasma samples at baseline and at 2 and 4 months after treatment. A subgroup of those on placebo switched to metformin treatment after 6 months of dietary intervention (850 or 1,700 mg/d,  $n = 13$ ); we obtained fecal and plasma samples from these individuals after a further 6 months. People in this group were not randomly selected, but had agreed to be treated with metformin. Compliance and side effects were monitored at each visit.

Inclusion criteria were: (i) aged between 18 and 65 years; (ii) T2D diagnosis in the previous 6 months, as defined by the American Diabetes Association Criteria<sup>51</sup>; (iii) absence of systemic and metabolic disease other than T2D, and absence of infection within the previous month; (iv) absence of diet or medication that might interfere with glucose homeostasis, such as glucocorticoids or antibiotics in the previous 3 months; and (v) HbA1c lower than 9%.

Exclusion criteria were: (i) clinically significant major systemic disease, including malignancy; (ii) clinical evidence of hemoglobinopathies or anemia; (iii) history of drug or alcohol abuse, defined as >80 g/d in men and >40 g/d in women; (iv) acute major cardiovascular event in the previous 6 months; (v) acute illnesses or current evidence of acute or chronic inflammatory or infective disease; and (vi) mental illness rendering the participants unable to understand the nature, scope, and possible consequences of the study.

All individuals gave written informed consent. The experimental protocol was approved by the Ethics Committee and the Committee for Clinical Investigation of the Hospital Universitari Dr. Josep Trueta (Girona, Spain). We certify that all applicable institutional regulations concerning the ethical use of information and samples from human volunteers were followed during this research. Complete clinical trial registration is deposited in the EU clinical trials register (EudraCT number 2010-022394-34).

**Extraction of fecal genomic DNA and whole-genome shotgun sequencing.** Fecal genomic DNA was extracted from 100 mg of frozen stools using the QIAamp DNA mini stool kit (Qiagen, Courtabouef, France) following repeated bead-beating (6,500 r.p.m.,  $3 \times 30$  s). The DNA was extracted from 131 fecal samples, obtained from the participants at three different time points during the study ( $n = 118$ ; two fecal samples were not obtained) and from 13 participants additionally sampled 6 months after switch to metformin treatment. DNA fragments of approximately 300 bp were sequenced on an Illumina NextSeq 500 instrument (150 bp; paired-end) at Genomics Core Facility at the Sahlgrenska Academy, University of Gothenburg.

**Metagenomics analyses.** We obtained a total of 941 Gb of raw paired-end reads. The taxonomic and KO composition was obtained by using an updated version of the MEDUSA pipeline<sup>19</sup>, in which the raw reads were trimmed by FASTX ([http://hamnonlab.cshl.edu/fastx\\_toolkit/](http://hamnonlab.cshl.edu/fastx_toolkit/)); with a quality threshold of 20 bp and minimum length of 35 bp, filtered to remove human reads (version hg19), and then mapped to bacterial gene and genome catalogs using Bowtie2 (ref. 52). An additional filter was applied during the mapping process containing reads with at least 95% identity to obtain high-quality reads. The mean mapping rates for the genome and gene catalogs were 36.6% and 64.6%, respectively (Supplementary Table 2). The obtained taxonomic composition and KO profile matrix were further analyzed by DESeq2 package<sup>53</sup>. Pathway-enrichment analyses are based on KEGG annotation<sup>54</sup> and hypergeometric test using goseq<sup>54</sup>. The beta diversity and PCoA analysis were calculated on the basis of the relative abundance of all significant KOs (further transformed by the square root to reduce the influence of dominant KOs, as suggested previously<sup>55</sup>) using phyloseq (version 1.12.2)<sup>56</sup>. Genus-to-genus

coabundance network analysis was based on Spearman correlation. Only genera present in at least 80% of samples were used for correlation analysis, and only connections with a rho value larger than 0.6 or smaller than -0.6 were used for network building and visualization, on the basis of igraph<sup>57</sup>. For the estimation of PTRs, metagenomic reads were mapped to a local genome database containing ~200 common human bacterial strains, using the software PTRC<sup>24</sup>. For the analysis of bile salt hydrolase (*bsh*) genes, a local gene database containing all available *bsh* genes was constructed by blasting<sup>58</sup> against NCBI reference genes<sup>59</sup>, MEDUSA gene catalog<sup>19</sup>, and the integrated gene catalog for human microbiome (IGC)<sup>60</sup> using 16 randomly selected seed *bsh* genes (gi: 169212173, 47121626, 488267184, 489835719, 491501450, 491807128, 499725619, 503743756, 524844235, 558633790, 654788256, 753801014, 759977951, 814507153, 823277295 and 933135484). The local *bsh* gene database was then used for targeted reads screening on the basis of the MEDUSA pipeline<sup>19</sup>.

**Targeted metabolomics analyses.** Fecal SCFAs were measured using gas chromatography coupled to mass spectrometry detection (GC-MS), as described previously<sup>61</sup>. In brief, approximately 50–250 mg of feces were mixed with internal standards, added to glass vials and freeze-dried. All samples were then acidified with HCl, and SCFAs were extracted with two rounds of diethyl ether extraction. The organic supernatant was collected, the derivatization agent *N*-tert-butylidimethylsilyl-*N*-methyltrifluoroacetamide (Sigma-Aldrich, Stockholm, Sweden) was added, and samples were incubated overnight. SCFAs were quantified with a 7090A gas chromatograph coupled to a 5975C mass spectrometer (Agilent Technologies 5975C, Santa Clara, CA). SCFA standards were obtained from Sigma-Aldrich (Stockholm, Sweden).

Bile acids were analyzed using ultra-performance liquid chromatography coupled to tandem mass spectrometry (UPLC-MS/MS), as described before<sup>41</sup>. Briefly, bile acids from plasma were extracted using protein precipitation with ten volumes of methanol containing internal standards. After mixing and centrifugation, the samples were evaporated and reconstituted in 200  $\mu$ l of methanol:water (1:1) for analysis. For feces, about 50 mg of stool samples were placed in a 2-ml polypropylene tube together with six ceramic beads (3 mm; Retsch GmbH, Haan, Germany) and 500  $\mu$ l of internal standard containing methanol. Stools were homogenized and centrifuged, and the supernatant was diluted ten times in methanol:water (1:1) before analysis. Bile acids were separated using a Kinetex C18 column (2.1  $\times$  100 mm with 1.7- $\mu$ m particles) (Phenomenex, Torrance, CA, USA) kept at 60 °C. The mobile phases consisted of water with 7.5-mM ammonium acetate and 0.019% formic acid (pH 4.5) as mobile phase A, and acetonitrile with 0.1% formic acid as phase B. A QTRAP 5500 instrument (Sciex, Toronto, Canada) was used for detection using multiple-reaction monitoring in negative mode. Bile acid standards were obtained from Sigma-Aldrich (Stockholm, Sweden), CDN Isotopes (Quebec, Canada), and Toronto Research Chemicals (Downsview, Ontario, Canada).

**Animal procedures.** Animal procedures were approved by the Gothenburg Animal Ethics Committee. For the fecal-microbiota transplant experiments, we used 10- to 12-week-old male Swiss Webster germ-free mice. Mice were kept in individually ventilated cages (ISOcage N System, Tecniplast, Buguggiate, Italy) with a maximum of five mice per cage. Water was given *ad libitum*. 500 mg of frozen stools obtained at baseline (M0) and 4 months after metformin treatment (M4) from three individuals were suspended in 5 ml of reduced PBS buffer containing 0.2 g/liter Na<sub>2</sub>S and 0.5 g/liter cysteine as reducing agents. The three donors were chosen from the metformin arm of the randomized clinical study who showed a reduction in %HbA1c after 2 and 4 months of metformin treatment, and the individual stool samples were not pooled. The germ-free mice were randomized into two groups and colonized by oral gavage with 200  $\mu$ l of M0/M4 fecal slurry from each donor. The mice were fed an irradiated high-fat diet (40% kcal fat, TD09683, Harlan Teklad) for 1 week before and during the 18 d of colonization. Body composition was determined with an EchoMRI instrument (EchoMRI) 1 d after colonization and at the end of the experiment. Insulin was measured with a kit from Crystal Chem (Downers Grove, IL) according to the manufacturer's protocol, and an intraperitoneal glucose-tolerance test was performed at the end of the experiment, as previously described<sup>62</sup>. The investigators were not blinded to the group allocation. No mice were excluded from this study.

**In vitro bacterial growth experiments.** Precultures of *B. adolescentis* L2-32 and *E. coli* were inoculated anaerobically in a Coy chamber (5% hydrogen, 10% carbon dioxide, and 85% nitrogen) as single colonies in 7 ml of brain-heart infusion (BHI) medium containing (in g/liter) yeast extract (5), cellobiose (1), maltose (1), cysteine (0.5), and hemin (0.01). For *A. muciniphila*, modified BHI broth was used, into which cysteine (0.05%) and type II mucin (1%) were added. Before inoculation, the modified medium was filtered through a 0.22- $\mu$ m filter. After incubation for 14 h, each preculture was inoculated in freshly prepared BHI broth or modified BHI broth with or without metformin in a 24-well plate or 96-well microplate at a concentration (v/v) of 0.5% *E. coli* or 1% *B. adolescentis* in a volume of 2.5 ml and 2% *A. muciniphila* in a volume of 300  $\mu$ l. The effect of metformin on bacterial-growth kinetics was analyzed in a CLARIOstar microplate reader equipped with atmospheric control unit (BMG Labtech) by following the optical density (OD<sub>600</sub>). The atmospheric oxygen concentration was reduced to 0.1% and was maintained with nitrogen as ground gas. The growth-curve data over 10 h for *E. coli* and *B. adolescentis* and 30 h for *A. muciniphila* were analyzed using MARS data-analysis software (BMG Labtech).

**In vitro gut simulator.** A simulated human intestinal redox model (SHRIM) was used to explore the effect of metformin on a stabilized gut microbial community *in vitro*. SHRIM is a two-chamber fermenter with an anaerobic luminal chamber (250 ml) and an oxygen feeder (100 ml), which are separated by a Nafion Membrane N115 (DuPont, USA; diameter 2.5 cm) and continuously purged with nitrogen and oxygen, respectively. The luminal chamber was continuously stirred at 250 r.p.m. and kept at 37 °C. The oxygen feeder contained 100-mM potassium phosphate buffer. The luminal chamber was seeded with 250 ml of feed containing: (in g/liter) arabinogalactan (1.0), pectin (2.0), xylose (1.5), starch (3.0), glucose (0.4), yeast extract (3.0), peptone (1.0), mucin type II (4.0), and cysteine (0.5). To simulate digestion processes, the feed was acidified to around pH 2 with 6-M HCl, and neutralized with simulated pancreatic juice to a pH of around 6.9. The simulated pancreatic juice contained: (in g/liter) NaHCO<sub>3</sub> (12.5), Oxgall bile salts (6.0), and pancreatin (0.9). The feed and pancreatic juice mix (70:30), referred to as SHRIM feed, was kept anaerobic by continuously purging with nitrogen<sup>65</sup>. The SHRIM feed was fed continuously to the luminal chamber at a rate giving a retention time of around 24 h, and pH was maintained between 6.9 and 6.6 with pH controller and dosing Pump (Black stone BL7912, Hanna Instruments, UK).

The SHRIM system was inoculated with an aliquot of the M0 fecal sample from each donor individually. A preculture was prepared anaerobically in a Coy chamber (5% hydrogen, 10% carbon dioxide, and 85% nitrogen) by adding 2% fecal material to 5 ml of BHI broth as described above. The preculture was incubated for 3 h at 37 °C, and 2% of the pre-culture was seeded into the luminal compartment of the SHRIM.

**Analysis of metagenome and metatranscriptome of the microbial community in the *in vitro* gut simulator.** After 1 week of stabilization, the microbial community was challenged with 10-mM metformin continuously, and samples (2 × 1 ml) were taken at baseline (time zero) and then daily for 1 week. After centrifugation at 16,000 r.p.m. for 2 min at 4 °C, the cellular pellet was suspended in 1 ml of Tris-EDTA buffer (10-mM Tris, 1-mM EDTA pH 7.5), and 500- $\mu$ l aliquots were used for DNA and RNA extractions. Total DNA was extracted by repeated bead-beating, as previously described<sup>64</sup>. Total RNA was extracted according to the Macaloid isolation protocol using the Phase Lock Gel Heavy tubes (5 Prime GmbH) and the RNeasy mini kit with on-column DNaseI treatment (Qiagen) for purification, as previously described<sup>65,66</sup>.

Whole-genome shotgun sequencing was performed both on isolated DNA and RNA from the gut simulator at baseline and after 1 d and 7 d of metformin treatment on Illumina NextSeq 500 instrument at Genomics Core Facility at the Sahlgrenska Academy, University of Gothenburg. An average of 21.6 million paired-end 150-bp DNA reads and 35.2 million paired-end 75-bp RNA reads from both donors were generated for metagenome and metatranscriptome analysis, respectively. Libraries for metagenome sequencing were prepared as mentioned above. Libraries for metatranscriptome sequencing were prepared from rRNA-depleted total RNA using the TruSeq Stranded Total RNA Library Preparation kit (Illumina). rRNA was depleted using the Ribo-Zero RNA Removal Kit for Gram-positive and Gram-negative bacteria (Illumina).

Then the MEDUSA pipeline<sup>19</sup> was used to obtain the taxonomical composition and functional KO profiles, as described for the metagenomics analysis. In addition, to analyze the gene expression profile of *A. muciniphila* and *B. wadsworthia*, a local gene database containing only genes from those two bacteria was downloaded from NCBI (accession number NC\_010655.1 and NZ\_LADCP00000002, respectively). The circo plot was produced using R package circlize<sup>67</sup>. GO enrichment was performed using R package STRINGdb (version 1.10.0)<sup>68</sup>.

**Statistical analysis.** All statistical analyses were performed in the R environment<sup>69</sup>. A power of 97% was obtained using pwr package<sup>70</sup> for this study, on the basis of 22 individuals, with paired design, 5% significance, and an estimated effect size of 0.87 for metformin in improving fasting blood glucose<sup>71</sup>. Because the primary aim of our randomized controlled study was to investigate the effect of metformin on the composition and function of the gut microbiota, we did not perform a power calculation for the placebo relative to metformin groups, because the effect size of metformin together with a calorie-restricted diet on the microbiota was previously unknown. For animal tests, sample size was chosen on the basis of our earlier experience and no statistical test was used to predetermine sample size.

Wald test with paired design implemented in DESeq2 (ref. 53) was used for differential abundance analyses for all count data (in the case of both metagenomics and metatranscriptomics). The Spearman's rank-order correlation was used to determine the strength and direction of the monotonic relationships between two variables unless strong collinearity was observed, in which case the Pearson product-moment correlation was calculated. Multivariate analysis with the Adonis test was performed on the basis of 5,000 permutations, using vegan<sup>72</sup>. Statistical testing for bacterial growth rates was examined by two-way ANOVA with repeated measures based on six technical replicates. Changes in SCFA concentrations between the metformin and placebo groups were compared by linear regression adjusted for BMI, gender, and fiber intake; the same procedure was used for bile acids, except the data were adjusted for total calorie intake instead of fiber intake. Otherwise, two-tailed Wilcoxon rank-sum tests or Wilcoxon signed-rank tests were used throughout the study, depending on whether the samples were paired. Raw *P* values were adjusted by the Benjamini-Hochberg method<sup>73</sup> with a false discovery rate of 5%, unless indicated otherwise.

**Data availability.** Sequence data are available for download from the Sequence Read Archive with accession number PRJNA361402.

50. Vioque, J. *et al.* Reproducibility and validity of a food frequency questionnaire among pregnant women in a Mediterranean area. *Nutr. J.* **12**, 26 (2013).
51. American Diabetes Association. Diagnosis and classification of diabetes mellitus. *Diabetes Care* **33** (Suppl. 1), S62-S69 (2010).
52. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357-359 (2012).
53. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
54. Young, M.D., Wakefield, M.J., Smyth, G.K. & Oshlack, A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* **11**, R14 (2010).
55. Kindt, R. & Coe, R. *Tree diversity analysis: A manual and software for common statistical methods for ecological and biodiversity studies* (World Agroforestry Centre, Nairobi, Kenya, 2005).
56. McMurdie, P.J. & Holmes, S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* **8**, e61217 (2013).
57. Csárdi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal, Complex Syst.* **1695**, 1695 (2006).
58. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410 (1990).
59. Tatusova, T., Ciuflo, S., Fedorov, B., O'Neill, K. & Tolstoy, I. RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Res.* **42**, D553-D559 (2014).
60. Li, J. *et al.* An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32**, 834-841 (2014).
61. Wichmann, A. *et al.* Microbial modulation of energy availability in the colon regulates intestinal transit. *Cell Host Microbe* **14**, 582-590 (2013).
62. Lee, Y.S. *et al.* Insulin-like peptide 5 is a microbially regulated peptide that promotes hepatic glucose production. *Mol. Metab.* **5**, 263-270 (2016).
63. Van den Abbeele, P. *et al.* Microbial community development in a dynamic gut model is reproducible, colon region specific, and selective for *Bacteroidetes* and *Clostridium* cluster IX. *Appl. Environ. Microbiol.* **76**, 5237-5246 (2010).

64. Salonen, A. *et al.* Comparative analysis of fecal DNA extraction methods with phylogenetic microarray: effective recovery of bacterial and archaeal DNA using mechanical cell lysis. *J. Microbiol. Methods* **81**, 127–134 (2010).
65. Murphy, N.R. & Halliwig, R.J. Improved nucleic acid organic extraction through use of a unique gel barrier material. *Biotechniques* **21**, 934–936, 938–939 (1996).
66. Zoetendal, E.G. *et al.* Isolation of RNA from bacterial samples of the human gastrointestinal tract. *Nat. Protoc.* **1**, 954–959 (2006).
67. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize Implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811–2812 (2014).
68. Franceschini, A. *et al.* STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* **41**, D808–D815 (2013).
69. Team, R.C.R. A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org/> (2015).
70. Champely, S. pwr: Basic Functions for Power Analysis. R package version 1.1-3. <http://CRAN.R-project.org/package=pwr> (2015).
71. Leucht, S., Helfer, B., Gartlehner, G. & Davis, J.M. How effective are common medications: a perspective based on meta-analyses of major drugs. *BMC Med.* **13**, 253 (2015).
72. Oksanen, J. *et al.* vegan: Community Ecology Package. R package version 2.2-1 <http://CRAN.R-project.org/package=vegan> (2015).
73. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).

## COLLABORATION 3

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: Nature Communications

Impact factor: 11.878

Citations:27

Contribution: Ph.D. Candidate Mercè Planas-Fèlix contribution to this study involved structural variant analyses. This involved the hand check and validation of each of the structural variants detected, in the BAM files of the selected patients. She also participated in the generation of the data to be analyzed for the pathways analysis.


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](mailto:mercader@broadinstitute.org)) or to D.T. (email: [david.torrents@bsc.es](mailto:david.torrents@bsc.es))  
#A full list of authors and their affiliations appears at the end of the paper

NATURE COMMUNICATIONS | (2018)9:321

| DOI: 10.1038/s41467-017-02380-9 | [www.nature.com/naturecommunications](http://www.nature.com/naturecommunications)

1



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<sup>1</sup>. 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<sup>2</sup>, reaching good-quality imputation of low frequency (minor allele frequency [MAF]:  $0.01 \leq \text{MAF} < 0.05$ ) and rare variants ( $\text{MAF} < 0.01$ ), 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<sup>3</sup>. The reanalysis of publicly available data also enables interrogation of this chromosome.

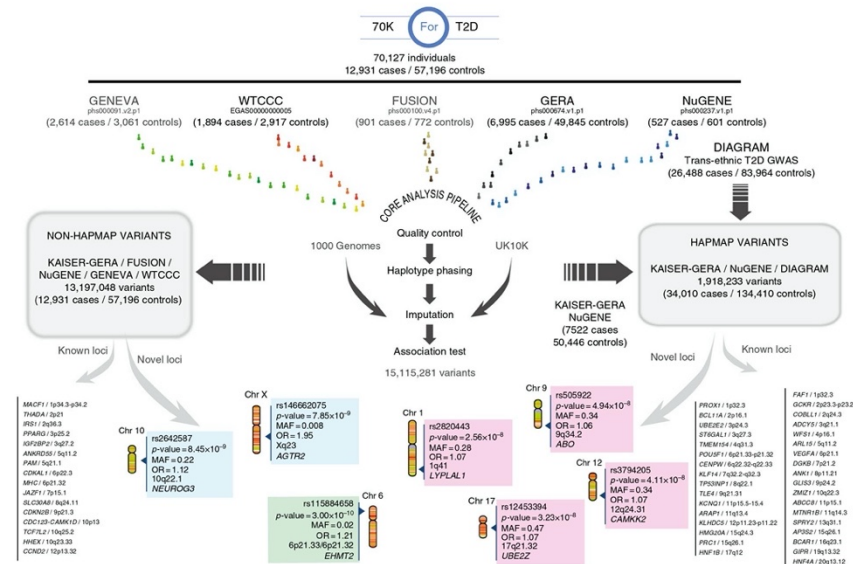
We hypothesized that a unified reanalysis of multiple publicly available data sets, applying homogeneous standardized quality control (QC), genotype imputation, and association methods, as well as novel and denser sequence-based reference panels for imputation would provide new insights into the genetics 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<sup>4</sup>. These studies have allowed the identification of more than 100 independent loci, most of them driven by common variants, with a few exceptions<sup>5</sup>. 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<sup>6</sup>, 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<sup>2</sup>, 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)<sup>7</sup> and UK10K<sup>2</sup> 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  $\geq 0.7$ ,  $\text{MAF} \geq 0.001$ , and  $I^2$  heterogeneity score  $< 0.75$ ), across 12,931 T2D cases and 57,196 controls. Of these, 6,845,408 variants were common ( $\text{MAF} \geq 0.05$ ), 3,100,848 were low-frequency ( $0.01 \leq \text{MAF} < 0.05$ ), and 5,169,025 were rare ( $0.001 \leq \text{MAF} < 0.01$ ). 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<sup>8</sup> (26,488 cases and 83,964 controls), selecting 1,918,233 common variants ( $\text{MAF} \geq 0.05$ ), 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 ( $\text{MAF} < 0.05$ ) 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  $p \leq 1 \times 10^{-4}$  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<sup>9</sup>.

**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  $p \leq 1 \times 10^{-5}$  using DEPICT<sup>9</sup> (Methods). This analysis showed enrichment of genes expressed in pancreas (ranked first in tissue enrichment analysis,  $p = 7.8 \times 10^{-4}$ ,  $\text{FDR} < 0.05$ , Supplementary Data 2) and cellular response to insulin stimulus (ranked second in gene-set enrichment analysis,  $p = 3.9 \times 10^{-8}$ ,  $\text{FDR} = 0.05$ ,



**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  $\leq 1 \times 10^{-4}$  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<sup>10</sup>.

**Fine-mapping and functional characterization of T2D loci.** The three association strategies allowed us to identify 57 genome-wide significant associated loci ( $p \leq 5 \times 10^{-8}$ ), 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 *PAM5* 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 ( $\rho = 0.92$ ,  $p = 1 \times 10^{-248}$ , 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<sup>11</sup> 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 ( $p \leq 5.3 \times 10^{-4}$  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<sup>12</sup>, we found that 12 of the 57 variants within its 99% credible set correspond to indels with genome-wide significance ( $5.6 \times 10^{-16} < p < 2.4 \times 10^{-15}$ ), which collectively represented 18.4% posterior probability of being causal.

**Table 1 Novel T2D-associated loci**

Novel Locus	Chr	rsID--Risk Allele	OR (95% CI) P-value			MAF
			Stage1 Discovery Meta-analysis	Stage2 Replication Meta-analysis	Stage1 + Stage2 Combined Meta-analysis	
<i>LYPLALI/ZC3H1B</i> (1q41)	1	rs2820443-T	1.08 (1.04–1.13) 2.94 × 10 <sup>-4</sup> a	1.06 (1.03–1.09) 2.10 × 10 <sup>-5</sup> b	1.07 (1.04–1.09) 2.56 × 10 <sup>-8</sup> c	0.28
<i>EHMT2</i> (6p21.33–p21.32)	6	rs115884658-A	1.34 (1.18–1.53) 1.00 × 10 <sup>-5</sup> a	1.17 (1.09–1.26) 2.90 × 10 <sup>-6</sup> c, d	1.21 (1.14–1.29) 3.00 × 10 <sup>-10</sup> c	0.02
<i>ABO</i> (9q34.2)	9	rs505922-C	1.07 (1.03–1.11) 6.93 × 10 <sup>-4</sup> a	1.06 (1.03–1.09) 1.90 × 10 <sup>-5</sup> b	1.06 (1.04–1.09) 4.94 × 10 <sup>-8</sup> c	0.34
<i>NEUROG3</i> (10q22.1)	10	rs2642587-G	1.12 (1.08–1.16) 8.45 × 10 <sup>-9</sup> e	-	-	0.22
<i>CAMKK2</i> (12q24.31)	12	rs3794205-G	1.09 (1.05–1.14) 4.18 × 10 <sup>-5</sup> a	1.06 (1.03–1.09) 1.60 × 10 <sup>-4</sup> b	1.07 (1.04–1.10) 4.11 × 10 <sup>-8</sup> c	0.32
<i>CALCOCO2/ATP5G1/UBE2Z/SNF8/GIP</i> (17q21.32)	17	rs12453394-A	1.08 (1.04–1.12) 7.86 × 10 <sup>-5</sup> a	1.07 (1.03–1.11) 9.60 × 10 <sup>-5</sup> b	1.07 (1.05–1.10) 3.23 × 10 <sup>-8</sup> c	0.47
<i>AGTR2</i> (Xq23)	X	rs146662075-T	3.09 (2.06–4.60) 3.24 × 10 <sup>-8</sup> f	1.57 (1.19–2.07) 1.42 × 10 <sup>-3</sup> g	1.95 (1.56–2.45) 7.85 × 10 <sup>-9</sup>	0.008

Chr chromosome, OR odds ratio, MAF minor allele frequency  
<sup>a</sup>Imputed based public GWAS discovery meta-analysis (NuGene + GERA cohort, 7,522 cases and 50,446 controls)  
<sup>b</sup>Transancestry DIAGRAM Consortium (26,488 cases and 83,964 controls)<sup>9</sup> Meta P-value estimated using a weighted Z-score method due to unavailable SE information from Stage 2 replication cohorts<sup>12</sup> T2D Diabetes Genetic Portal (Exome-Chip + Exome Sequencing, 35,789 cases and 56,738 controls)<sup>10</sup> Full imputed based public GWAS meta-analysis (NuGene + GERA cohort + GENEVA + FUSION + WTCCC, 12,931 cases and 57,196 controls)  
<sup>c</sup>70Kfor T2D Men Cohort (GERA cohort + GENEVA + FUSION, 5,277 cases and 15,702 controls older than 55 years)  
<sup>d</sup>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<sup>-1</sup>, 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<sup>13</sup> (Supplementary Data 7), and the Combined Annotation-Dependent Depletion (CADD)<sup>14</sup> and LINSIGHT<sup>15</sup> 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<sup>16</sup> and RNA-sequencing gene expression data from pancreatic islets<sup>17</sup>.

**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 (*LYPLALI-ZC3H1B*, 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<sup>18–23</sup>. Among the genes in this locus, *LYPLALI*, which encodes for lysophospholipase-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<sup>24</sup>.

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<sup>25</sup>, and rs14659 ( $r^2$  with top = 1) is associated with an increased risk for cardiometabolic disorders<sup>26</sup>. One of the variants within the credible set was the single base-pair frame-shift deletion defining the blood group O<sup>27</sup>. In concordance with previous results that linked O blood type with a lower risk of developing T2D<sup>28</sup>, the frame-shift deletion determining the blood group type O was associated with

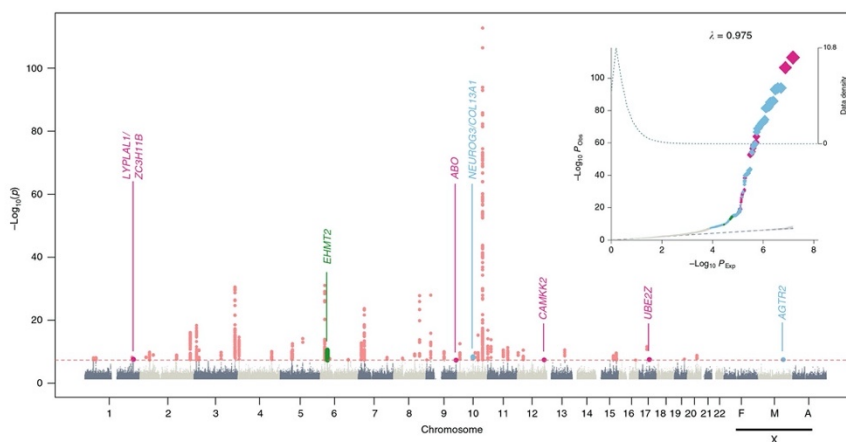
a protective effect for T2D in our study (rs8176719,  $p = 3.4 \times 10^{-4}$ , OR = 0.95 [0.91–0.98]). In addition, several variants within this credible set are associated with the expression of the *ABO* gene in multiple tissues including skeletal muscle, adipose tissue, and pancreatic islets (Supplementary 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<sup>29</sup>. 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<sup>30</sup>.

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<sup>31</sup>. 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<sup>32</sup>, 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<sup>33</sup>. Variants in the *GIP* receptor (*GIPR*) have been previously associated with insulin response to oral glucose challenge and beta-cell function<sup>34</sup>, proposing *GIP* as a plausible candidate effector gene of this locus<sup>35</sup>.

**A new T2D signal driven by a low-frequency variant.** Furthermore, we selected all low-frequency (0.01 ≤ MAF < 0.05) variants with  $p \leq 1 \times 10^{-4}$  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  $\lambda$  is a measure of the genomic inflation and corresponds to the observed median  $\chi^2$  test statistic divided by the median expected  $\chi^2$  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<sup>6</sup> 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 *EHM2* 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). *EHM2* is involved in the mediation of FOXO1 translocation induced by insulin<sup>36</sup>. Since this variant is less than 1 Mb away from *HLA-DQA1*, a locus reported to be associated with T2D<sup>37</sup>, we performed a series of reciprocal conditional analyses and excluded the possibility that our analysis was capturing previously reported T2D<sup>8, 37</sup> or T1D<sup>38–40</sup> signals (Supplementary Data 11). Beyond this missense *EHM2* 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 (multi-tissue 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 ( $p(\text{exon 1}) = 1.4 \times 10^{-19}$ ,  $p(\text{exon 2}) = 1.9 \times 10^{-13}$ , 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<sup>41</sup>. 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<sup>42</sup>. 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 ( $R^2_{\text{UK10K,males}} = 0.94$ ,  $R^2_{\text{UK10K,females}} = 0.66$ ,  $R^2_{\text{1000G,males}} = 0.62$ , and  $R^2_{\text{1000G,females}} = 0.43$ ; 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<sup>6</sup>, INTERACT<sup>43</sup>, Partners Biobank<sup>44</sup>, and UK Biobank<sup>45</sup>), 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,  $p = 1.2 \times 10^{-3}$ ; Supplementary Fig. 9A), and revealed a strong degree of heterogeneity (heterogeneity  $p_{\text{het}} = 0.004$ ), 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<sup>46</sup>. 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  $7.8 \text{ mmol l}^{-1}$ , a threshold employed to identify impaired glucose tolerance (prediabetes)<sup>47</sup>, 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<sup>48–50</sup>. From the analysis of available epigenomic data sets<sup>51</sup>, 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.

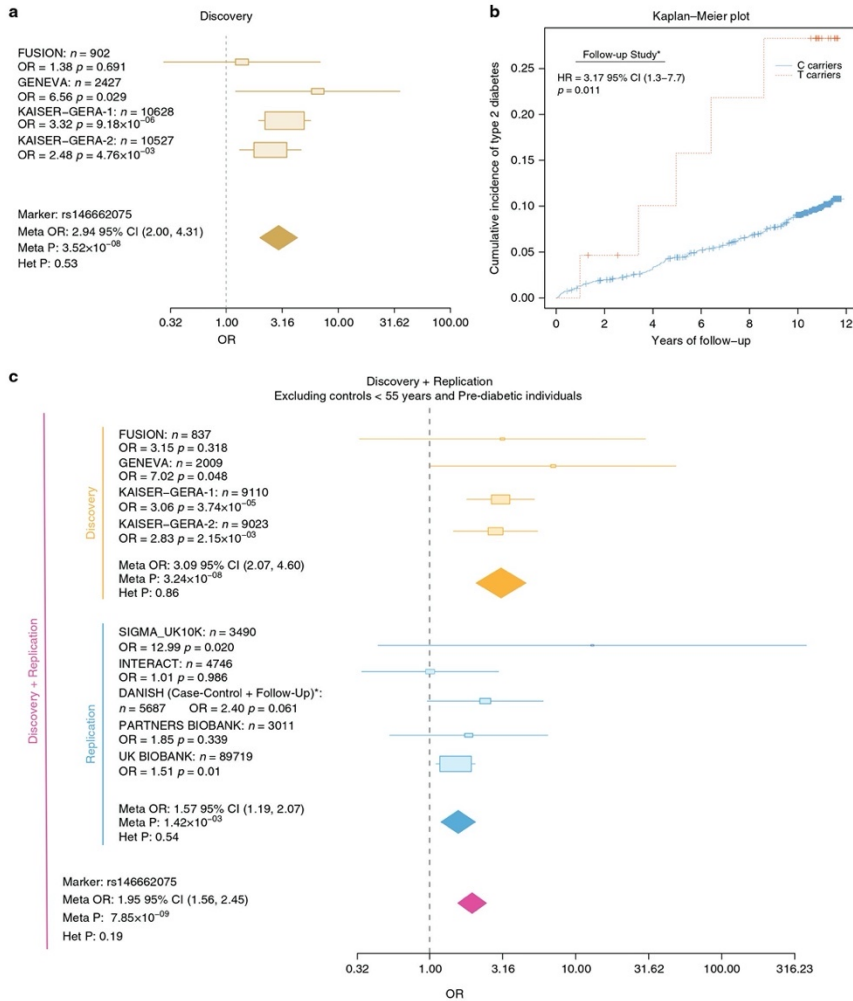
## Discussion

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<sup>52, 53</sup>.

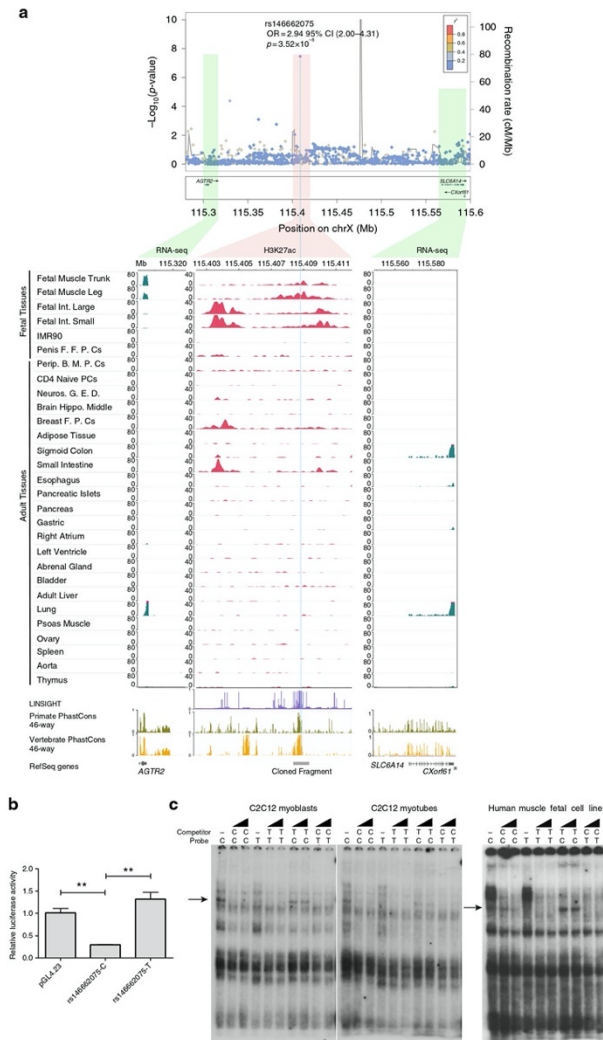
Our study complemented other efforts that also aim at unraveling the genetics behind T2D through the generation of new



**Fig. 3** Discovery and replication of rs146662075 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<sup>-1</sup>, and controls with family history of T2D in both the discovery and replication cohorts when available

genetic data<sup>5, 54</sup>. 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<sup>52</sup>. 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<sup>48</sup>. *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<sup>55</sup> and in humans<sup>56</sup> prevents the onset of T2D, and restores normoglycemia<sup>57, 58</sup>.

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<sup>4, 59</sup>, 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)<sup>40</sup> data available through the European Genotype Archive (EGA, <https://www.ebi.ac.uk/ega/studies/EGAS0000000028>). 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<sup>60</sup>. 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<sup>61</sup> to ameliorate the computational burden required for genotype imputation through IMPUTE2<sup>62</sup>. 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<sup>63</sup> and the UK10K<sup>64</sup>.

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  $\geq 0.05$ , (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  $r \geq 0.125$  from which we removed the individual with the highest proportion of missingness), (iii) missing call rates per sample  $\geq 0.02$ , 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 SNPTTEST, 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)<sup>63</sup> 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  $\geq 0.7$ , MAF  $\geq 0.001$ , 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<sup>65</sup>, 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<sup>8</sup>, 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  $-\log_{10}$  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)<sup>6</sup>, 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)<sup>9</sup> 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](http://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:  $-\text{clump-p1 } 5e-8$ ,  $-\text{clump-p2 } 1e-5$ ,  $-\text{clump-r2 } 0.6$ , and  $-\text{clump-kb } 250$ . 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<sup>9</sup>. 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<sup>64</sup>. 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 (CI–C5) were obtained from<sup>10</sup>.

**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 WTCOC 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<sup>11</sup> (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<sup>11</sup>. We estimated the approximate Bayes' factor (ABF) for each variant as

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

where

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

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

The  $\beta$  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

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

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  $\beta$  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)<sup>15</sup>. 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<sup>14</sup> (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<sup>15</sup>. 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<sup>16</sup>. 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<sup>17</sup> 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 1 Mb 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<sup>8,38–40</sup>. 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 *HNFA1* locus (rs1169288, rs1800574, and chr12:121440833:D)<sup>54</sup>, which is summarized in Supplementary Data 12. Moreover, this approach confirmed known secondary signals in the 9p21 locus<sup>65</sup> 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<sup>43</sup>, the Slim Initiative in Genomic Medicine for the Americas (SIGMA) consortium GWAS data set<sup>6</sup>, the Partners HealthCare Biobank (Partners Biobank) data set<sup>44</sup>, and the UK Biobank cohort<sup>65</sup>. 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™ UK BiLEVE Axiom™ Array and the Applied Biosystems™ UK Biobank Axiom™ 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 SNPTTEST 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)<sup>66</sup> 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<sup>-1</sup> 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<sup>31</sup>. 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<sup>67</sup>, 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 ( $\lambda_{local}$ ).

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 MACS2 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.

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 MACS2 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)<sup>68</sup>. 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<sup>69</sup>. Double-stranded oligonucleotides containing either the common or rare variants of rs146662075 were labeled using dCTP ( $\alpha$ -32P) (Perkin Elmer). Oligonucleotide sequences are as follows (SNP location is underlined): probe-C-F: 5'-gatCTTGAACACcGAGGGGAAAAT-3' and R5'-gatATTTCCTCCCTcGTGTTCAAA-3' and probe-T-F: 5'-gatTTTGAACACcGAGGGGAAAAT-3' and R: 5'-gatATTTCCTCCCTcGTGTTCAAA-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'-GCTAGCATATGGAGGTGATTGTG-3' and R: 5'-GGCACTTCCTTCCTCGGTAGA-3' oligonucleotides and cloned into pENTR/D-TOPO (Invitrogen). Allelic variant rs146662075T was introduced by site-directed mutagenesis using the following primers: F: 5'-CCTTTTTTACTTTGAACACcGAGGGGAAAATcATGCTTGGC-3' and R: 5'-GCCAAGCATGATTTCCCTCAGTGTTCAAAAGTAAAAAAGG-3'. Enhancer sequences were shuttled into pGL4.23[luc2/min]P 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/min]P 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/](http://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

- Welter, D. et al. The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).
- Huang, J. et al. Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat. Commun.* **6**, 8111 (2015).
- 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).
- Flannick, J. & Florez, J. C. Type 2 diabetes: genetic data sharing to advance complex disease research. *Nat. Rev. Genet.* **17**, 535–549 (2016).
- 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).
- Fuchsberger, C. et al. The genetic architecture of type 2 diabetes. *Nature* **536**, 41–47 (2016).
- Abecasis, G. R. et al. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
- 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).
- Pers, T. H. et al. Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.* **6**, 5890 (2015).
- Pasquali, L. et al. Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat. Genet.* **46**, 136–143 (2014).
- 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).
- 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).
- McLaren, W. et al. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069–2070 (2010).
- Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
- 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).
- Mele, M. et al. Human genomics. The human transcriptome across tissues and individuals. *Science* **348**, 660–665 (2015).
- 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).
- 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).
- 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).
- 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).
- 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).
- 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).
- 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).
- Lei, X., Callaway, M., Zhou, H., Yang, Y. & Chen, W. Obesity associated Lyp1 gene is regulated in diet induced obesity but not required for adipocyte differentiation. *Mol. Cell. Endocrinol.* **411**, 207–213 (2015).
- 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).
- the CARDIoGRAMplusC4D Consortium. A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet.* **47**, 1121–1130 (2015).
- 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).
- 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).
- Gradwohl, G., Dierich, A., LeMeur, M. & Guillemin, 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).
- Rubio-Cabezas, O. et al. Permanent neonatal diabetes and enteric anodocrinosis associated with biallelic mutations in *NEUROG3*. *Diabetes* **60**, 1349–1353 (2011).
- 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).
- 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).
- 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).
- 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).
- Lyssenko, V. et al. Pleiotropic effects of *GIP* on islet function involve osteopontin. *Diabetes* **60**, 2424–2433 (2011).
- 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 Dev.* **20**, 340–357 (2015).
- 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).
- 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).
- Hakonarson, H. et al. A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene. *Nature* **448**, 591–594 (2007).
- 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).
- 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).
- Voight, B. F. et al. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat. Genet.* **42**, 579–589 (2010).
- Langenberg, C. et al. Gene-lifestyle interaction and type 2 diabetes: the EPIC interact case-cohort study. *PLoS Med.* **11**, e1001647 (2014).
- 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).
- Bycroft, C. et al. Genome-wide genetic data on ~500,000 UK Biobank participants. Preprint at *bioRxiv* <https://doi.org/10.1101/166298> (2017).
- Beccerra, M. B. & Beccerra, B. J. Disparities in age at diabetes diagnosis among Asian Americans: Implications for early preventive measures. *Prev. Chronic Dis.* **12**, E146 (2015).
- Bartoli, E., Fra, G. P. & Carnevale Schianca, G. P. The oral glucose tolerance test (OGTT) revisited. *Eur. J. Intern. Med.* **22**, 8–12 (2011).
- 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).
- 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).
- 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).
- Roadmap Epigenomics, C. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
- Molke, I. et al. A common greenlandic *TBC1D4* variant confers muscle insulin resistance and type 2 diabetes. *Nature* **512**, 190–193 (2014).
- 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).
- 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).
- Frantz, E. D., Crespo-Mascarenhas, C., Barreto-Vianna, A. R., Aguilá, 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).
- 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).
- 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-Salari, 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. Creighton, 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 EFSF/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](http://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](http://www.uk10k.org). Funding for UK10K was provided by the Wellcome Trust under award WT091310. We acknowledge PRACE for awarding us to access MareNostrum super-computer, 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-Félix is funded by the Obra Social Fundacion 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 Skłodowska-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](http://www.metabol.ku.dk)). This research has been conducted using the UK Biobank Resource (application number 14803). 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.R., 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.

 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018

Sílvia Bonàs-Guarch<sup>1</sup>, Marta Guindo-Martínez<sup>1</sup>, Irene Miguel-Escalada<sup>2,3,4</sup>, Niels Grarup<sup>5</sup>, David Sebastian<sup>3,6,7</sup>, Elias Rodríguez-Fos<sup>1</sup>, Friman Sánchez<sup>1,8</sup>, Mercè Planas-Félix<sup>1</sup>, Paula Cortes-Sánchez<sup>1</sup>, Santi González<sup>1</sup>, Pascal Timshel<sup>5,9</sup>, Tune H. Pers<sup>5,9,10,11</sup>, Claire C. Morgan<sup>4</sup>, Ignasi Moran<sup>4</sup>, Goutham Atla<sup>2,3,4</sup>, Juan R. González<sup>12,13,14</sup>, Montserrat Puiggros<sup>1</sup>, Jonathan Marti<sup>8</sup>, Ehm A. Andersson<sup>5</sup>, Carlos Díaz<sup>8</sup>, Rosa M. Badia<sup>8,15</sup>, Miriam Udler<sup>16,17</sup>, Aaron Leong<sup>17,18</sup>, Varindepal Kaur<sup>17</sup>, Jason Flannick<sup>16,17,19</sup>, Torben Jørgensen<sup>20,21,22</sup>, Allan Linneberg<sup>20,23,24</sup>, Marit E. Jørgensen<sup>25,26</sup>, Daniel R. Witte<sup>27,28</sup>, Cramer Christensen<sup>29</sup>, Ivan Brandslund<sup>30,31</sup>, Emil V. Appel<sup>5</sup>, Robert A. Scott<sup>32</sup>, Jian'an Luan<sup>32</sup>,

Claudia Langenberg<sup>3,2</sup>, Nicholas J. Wareham<sup>3,2</sup>, Oluf Pedersen<sup>5</sup>, Antonio Zorzano<sup>3,6,7</sup>, Jose C Florez<sup>16,17,33</sup>,  
Torben Hansen<sup>5,3,4</sup>, Jorge Ferrer<sup>2,3,4</sup>, Josep Maria Mercader<sup>1,16,17</sup> & David Torrents<sup>1,35</sup>

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

## PATENT

Application number: EP16178577.9

Publication number: WO 2018/007034

Publication date: January 11th, 2018

Title: A computer-implemented and reference-free method for identifying variants in nucleic acid sequences

Applicant: Barcelona Supercomputing Center – Centro Nacional de Supercomputación, Institució Catalana de Recerca i Estudis Avançats, Universitat Politècnica de Catalunya.

Inventor: David Carrera Perez, Jordà Polo, Nicola Cadenelli, David Torrents Arenales, **Mercè Planas**

The patent successfully passed the European assessment with the number EP17714441.7.

A computer-implemented and reference-free method for identifying variants in nucleic acid sequences

5 The present invention relates to a computer implemented method for the identification and characterization of sequence variants in nucleic acids. In particular, this method is able to quickly and accurately identify most types of sequence genome variations with a potential association to a disease, that is, from single nucleotide substitutions to large structural variants. This method may have multiple and direct applications in genomics-based diagnosis, prognosis and therapy.

10 The invention further relates to a computer program and to systems suitable for performing such a method. The computer program may be designed to be lock-less and scalable, thereby allowing for high performance implementations on parallel execution environments such as specialized hardware accelerators.

#### BACKGROUND ART

20 The genetic basis of disease is increasingly becoming more accessible thanks to the emergence of the Next Generation Sequencing (NGS) platforms, which have extremely reduced the costs and increased the throughput of genomic sequencing. For the first time in history, personalized medicine is close to becoming a reality through the analysis of each patient's genome.

25 A wide range of genome variation of cells and individuals has been identified to be the direct cause, or a predisposition to genetic diseases: from single nucleotide variants (SNVs if they are somatic, and SNPs if they are polymorphic in the population), to structural variants (SVs), which can correspond to deletions, insertions, inversions, translocations and copy number variations (CNVs), ranging from a few nucleotides to large genomic regions, including complete chromosome arms. These variations can exist between patients and also emerge among cells of the same patient. The unveiling of changes in the genome is driving discoveries such as the Philadelphia translocation between chromosomes 9 and 22, whose presence

implies the development of chronic myelogenous leukemia (CML) and its identification allows the development and selection of last-generation therapies.

- 5 The ideal exploitation of genomic sequencing should involve the accurate identification of all variants, in order to derive a correct diagnosis and to select the best therapy. For clinical purposes, it is important that this computational process be carried out within an effective timeframe. But a simple sequencing experiment typically yields thousands of millions of reads  
10 per genome, which have to be stored and analysed. The task is severely hindered by a variety of factors such as PCR-amplification and sequencing errors, limitations intrinsically linked to the size of the reads, biases in the sequencing techniques employed, the inherent repetitive and dynamic nature of the genomic sequences, and others. As a consequence, the analysis of  
15 genomes with diagnostic and therapeutic purposes is still a great challenge, both in the design of efficient algorithms and at the level of computing performance.

- Modern medicine will rely on the identification of genetic markers for precision  
20 diagnosis and for the application of more specific therapies. Cancer is one of the most active diagnostic and therapeutic areas where genetic analysis is being applied. Having access to all somatic variation accumulated in a tumor cell is now allowing the study of the genetic causes of the tumor and the development of new clinical protocols that are already starting to be applied  
25 in some clinical centres, and that will be soon a reality for all modern healthcare systems around the world. This is why, the identification of tumor variants is key in research and soon, also in medical care. The variants responsible for the origin and progression of tumors are currently searched using a common scheme that involves the sequencing of both tumor and  
30 normal genome samples of the same patient, and the subsequent scan and identification of the differences between them. Most of the available methods rely on an initial step, where all the normal and tumor sequence reads are aligned to a reference genome to then identify the changes present in the tumor compared with the normal and reference genomes. Despite these  
35 methods have provided a great number of disease-associated variation so far, they still entail intrinsic limitations associated to the need of a



prealignment to a reference genome, affecting their performance and accuracy. More precisely, the reference-based identification of somatic variation in cancer genomes has currently the following sources of errors and limitations: (i) the initial alignment step, on which all the methods rely, is time consuming and particularly error prone with the tumor reads that carry the sequence variation, which are the most relevant for the analysis. It has been proven that many of these reads that carry changes and differences in their sequence are difficult or even impossible to align to the reference unmutated genome. The absence and the misplacement of tumor reads in the final alignment drastically affect all existing downstream methods for variant searching and calling. Although a number of alternative methods exist, this alignment step is generally performed with the same program (Li, H., et.al. "Fast and accurate short read alignment with Burrows-Wheeler transform" *Bioinformatics* 2009, vol. 25, pp. 1754-1760), which implies that nearly all analyses done nowadays share the same type of errors derived from this mapping of reads. (ii) The usage of a reference genome also involves the interference with millions of inherited variants (germline, i.e. not somatic) that affect both, the accuracy at the level of read mapping and the actual identification of the target somatic fraction (normally comprising only between 2 and 10 thousand variants). A considerable number of these germline variants are then frequently mispredicted as somatic changes, increasing the rate of false positives and decreasing, consequently, the final reliability and applicability of the results.

On top of the limitations and errors inherent to the generation and dependency of this initial alignment, the subsequent analysis, where somatic changes are finally identified, also implies a number of restrictions and complications. For example, despite the great deal of possibilities in terms of available methods, not a single one of them is able to identify a wide range of somatic variation, but instead, each is limited to the detection of a particular size and type of mutation (Medvedev P. "Computational methods for discovering structural variation with next-generation sequencing" *Nat. Methods Suppl.* 2009, vol. 6, S13-S20) There are programs that use this alignment to detect only SNVs and others that only identify SVs, among which, each one is able to detect a particular variant size. For instance, some methods identify insertions or deletions that comprise a few nucleotides (from

2 to a few dozens), others detect medium size SVs (from a few dozens of nucleotides to a few hundreds), and a small fraction of them, are designed for the identification of larger SVs (Ding, L., et.al. "Expanding the computational toolbox for mining cancer genomes" Nat. Rev. Genet. 2014, vol. 15, pp. 556-570). As the detection complexity increases with the detection range, methods designed for the identification of large SVs are also more imprecise in defining the exact location in the genome and the type of change, which are often necessary for being able to derive the functional consequences of the mutation. These programs often only report regions where SVs might be located.

In order to overcome these limitations, to date, alternative approximations have been developed. The term reference-free becoming more popular and has recently been used to describe a wide range of fundamentally different underlying strategies. For example, these methodologies are using fundamentally different unrelated strategies covering, from the use of reference mapping plus assembly-based (Chen K. "TIGRA: A targeted Iterative Graph Routing Assembler for Breakpoint Assembly" Genome Res. 2016, vol. 24, pp.310-317), *de novo* assembly (Zhuang, J. "Local sequence assembly reveals a high-resolution profile of somatic structural variations in 97 cancer genomes" Nucleic Acid Research 2015, vol. 43, pp.8146-8156), to suffix tree approximations (Moncunill V., et al., "Comprehensive characterization of complex structural variations by directly comparing genome sequence reads" Nature Biotech. 2014, vol. 32, pp. 1106-1112). The first two examples are based on the end-joining of reads in the tumor and normal genomes in order to identify discordant patterns. Although these assembly can also suffer the mapping-derived limitations, they have other major limitations associated to the underlying mechanism of the assembly, mostly when using NGS reads, as the overlapping regions needed to extend over the read size are often too small to be position-specific in the genome. Among other reference-free approximations reported to date, it can be highlighted a suffix tree-based method (SMUFIN) that compares in a tree-like structure all tumor and normal reads, to then extract discordant branches as candidate positions for variation. Although this particular way of analysing reads may directly overcome many of the limitations mentioned, it still lacks possibilities for detecting non-human sequences and is limited by the size of

the tree, which grows in memory demands as sequencing coverage grows. Additionally, and in contrast to the approach followed in the method of the present invention, suffix trees are data structures that inherently require locking access patterns to allow for concurrent updates to take place, therefore limiting the ability to efficiently implement these approaches in high performance parallel computing systems. These two fundamentally different approaches of analysing sequence reads have also limitations of scalability, as the design of the code has not considered alternative ways of adapting to specific and more efficient hardware architectures. In fact, all the limitations mentioned hinder the incorporation of this type of genomic analysis into identification of somatic mutations applied to the clinical practice, which calls for much faster and more accurate computational methods. In addition, current methods for somatic variant calling still miss an important fraction of large SVs, which are relevant for the diagnosis and treatment of diseases.

Similar approximations have also been extended to deal with other type of problems in molecular biology. For example, some reference free methods have been developed to quantify the abundance of RNA isoforms from RNA-seq data (Patro R., et. al. "Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms" Nature Biotech. 2014, vol. 32, pp. 462-464), or to identify evolutionary-driven substitutions in homozygosis, using *de novo* assembly of plant genomes (Nordstrom K.J.V, et.al. "Mutation identification by direct comparison of whole-genome sequencing data from mutant and wild-type individuals using k-mers" Nature Biotech. 2013, vol. 31, pp. 325-331).

Clearly quick and robust comparative methods, able to detect all kinds of SVs differentiating two states (normal vs. pathological, undifferentiated vs. differentiated, etc.) with high sensitivity, specificity, speed and scalability are still needed, as well as systems and computer programs suitable for performing such methods.

#### SUMMARY OF THE INVENTION

In contrast to what is found in the *prior art*, inventors have come up with a computer-implemented method for identifying nucleic acid variants between

two genomic states that does not depend on the alignment of reads of either state to a reference genome, or on the construction of sequence-based suffix trees. Using a different underlying mechanism, this method is, on its own, able to accurately identify all types of variants (heterozygous and  
5 homozygous), from single nucleotide variations to large structural variants at base pair resolution with unprecedented performance at the level of variant detection and execution possibilities. Importantly, the method is not restricted to the identification of a certain type of variant (SNVs, insertions, inversions, etc.) nor does it only perform for variants of a certain size, as is the case for  
10 many of the methods found in the art. Because of its underlying principle, based on the use of k-mers and hashtables independently of a reference genome, all the limitations that apply to all other existing methods (outlined above) are overcome. This translates to a method that is not only more robust, but is also more thorough and much faster than the methods  
15 described to date.

This invention entails a reference-free detection method that allows discovering homozygous and heterozygous variation in genomes using a polymorphic k-mer strategy consisting in the sequential sub-selection of read  
20 regions that will be compared in different ways to finally isolate variant-containing regions. One of the key elements of this computer-implemented method is the way k-mers are handled, as they are taken as “dynamic entities” by taking their stems and using the latter to explore for their inflections and partial inflections (see below). Compared to the other  
25 reference-free methods described above, the invention relies on a fundamentally different way of dealing with the reads since, instead of constructing an assembly, or a complete suffix-tree with all normal and tumor reads, it uses a particular k-mer strategy (see below) to fish reads with potential variation and discards, in one pass, the vast majority of reads with  
30 no information. This allows inventors to quickly filter and retain a subset of reads representing all the variants that are now computationally easy to treat and analyse.

Of note, the use of k-mers for direct comparison of genomes has only been  
35 explored in simple scenarios with a small scope, data, and requirements (see for instance Nordstrom K., *ibid*). In general, the use of k-mers has some

limitations due to their strict nature, and the way k-mers are distributed in genomes, requiring large amounts of computing resources if the identification of unique features is sought. Inventors address these limitations by using a more flexible approach: polymorphic k-mers, which in addition to k-mers, also  
 5 identify variations (inflections) of k-mers with similar patterns (stems, see below). Unlike regular, fixed-length k-mers, polymorphic k-mers enable the identification of unique features, and at the same time provide the means to gather and group related sequences even if they are not strictly the same. This element is key, as will be seen in the examples found below.

10 Thus, a first aspect of the present invention is a computer-implemented method for identifying nucleic acid variants between two genomic states comprising the steps of: A) Inputting 2 sets of nucleic acid reads, which are sequences retrieved from a nucleotide sequencing method, wherein the first  
 15 set of reads corresponds to cells representing a first test state, and the second set of reads corresponds to cells representing a second control state; B) Filtering the reads, wherein the filtering comprises: B1) Keeping only the reads with at least a percentage X1 of their bases with a Phred quality score higher than 20, being X1 equal to or above 90%; B2) Splitting the reads  
 20 with an undefined nucleotide, giving one sequence before, and one sequence after the undefined nucleotide, the latter being discarded; and B3) Discarding the sequence reads with less than X2 bases, wherein X2 is from 25 to 50; C) Generating a hashtable structure comprising: C1) Generating a number of N-X2+1 new reads for each read of sequence length N, wherein the new N-X2+1 reads correspond to all k-mers with length X2 nucleotides; and C2)  
 25 Building a hashtable structure, which comprises all the k-mers generated in step C1) and further comprises the number of times each k-mer is observed in the two sets of reads corresponding to first and second states. D) Detecting candidate variants in the sequence between first state and  
 30 second state, wherein a k-mer of the hashtable structure is taken as a candidate breakpoint, which represents a variant between the first and second states, if it fulfills all the following requirements: D1) At least one inflection based on a k-mer's stem must have at least X3 reads with the same variation between first and second states, being X3 at least 2; D2) The  
 35 percentage of first state reads in second state reads is not over a threshold X4, to account for possible contamination of control state reads with test state

reads, wherein X4 is at least 5%. E) Clustering and filtering test and control reads derived from all candidate breakpoints accepted in step D to build blocks, by carrying out the steps: E1) Retrieving reads which contain the stem of at least one k-mer that represents the candidate breakpoint selected in step D); E2) The reads of step E1) with at least X5 k-mer variants within a window of X6 nucleotides are taken as leading reads, wherein X5 is at least 7 and X6 is at least 10; E3) Reads whose k-mers share at least one stem with a leading read are merged to give a block; and E4) If the nucleic acid whose variant is being identified is a double stranded DNA, then both forward and reverse variants are taken into account when building the block. F) Aligning blocks taking their leading reads as a reference: F1) For each read in the block, take the leading read's stem and find the longest inflection or partial inflection between the read and the leading read. F2) Successively position each read so that its matching inflection or partial inflection is aligned against the leading read.

The performance and speed of this method make it more suitable for clinical applications (such as genomic analysis of cancer cells) than the alternative solutions, which result complex and time-consuming. As it is shown in the data below, the method has a superior performance even to last-generation methods such as the one published in Moncunill (*ibid*), which is also reference-free.

A second aspect of the present invention is a computer program product comprising program instructions for causing a computer system to perform the computer-implemented method for identifying nucleic acid variants between two genomic states of the first aspect of the invention.

The computer program product may be embodied on a storage medium (for example, a CD-ROM, a DVD, a USB drive, on a computer memory or on a read-only memory) or carried on a carrier signal (for example, on an electrical or optical carrier signal).

The computer program may be in the form of source code, object code, a code intermediate source and object code such as in partially compiled form, or in any other form suitable for use in the implementation of the processes

according to the invention. The carrier may be any entity or device capable of carrying the computer program.

5 For example, the carrier may comprise a storage medium, such as a ROM, for example a CD ROM or a semiconductor ROM, or a magnetic recording medium, for example a floppy disc or hard disk. Further, the carrier may be a transmissible carrier such as an electrical or optical signal, which may be conveyed via electrical or optical cable or by radio or other means.

10 When the computer program is embodied in a signal that may be conveyed directly by a cable or other device or means, the carrier may be constituted by such cable or other device or means.

15 Alternatively, the carrier may be an integrated circuit in which the computer program is embedded, the integrated circuit being adapted for performing, or for use in the performance of, the relevant methods.

A third aspect of the invention is a system for identifying nucleic acid variants between two genomic states, the system comprising:

- 20 A) Computer/Electronic means for inputting 2 sets of nucleic acid reads, which are sequences retrieved from a nucleotide sequencing method, wherein the first set of reads corresponds to cells representing a first test state, and the second set of reads corresponds to cells representing a second control state;
- 25 B) Computer/Electronic means for filtering the reads, wherein the filtering comprises: B1) Keeping only the reads with at least a percentage X1 of their bases with a Phred quality score higher than 20, being X1 equal to or above 90%; B2) Splitting the reads with an undefined nucleotide, giving one sequence before, and one sequence after the undefined nucleotide, the latter
- 30 being discarded; and B3) Discarding the sequence reads with less than X2 bases, wherein X2 is from 25 to 50; C) Computer/Electronic means for generating a hashtable structure comprising: C1) Generating a number of N-X2+1 new reads for each read of sequence length N, wherein the new N-X2+1 reads correspond to all k-mers with length X2 nucleotides; and
- 35 C2) Building a hashtable structure, which comprises all the k-mers generated in step C1) and further comprises the number of times each k-mer is observed

in the two sets of reads corresponding to first and second states.

- D) Computer/Electronic means for detecting variants in the sequence between first state and second state, wherein a k-mer of the hashtable structure is taken as a candidate breakpoint, which represents a variant
- 5 between the first and second states, if it fulfills all the following requirements:
- D1) At least one inflection based on a k-mer's stem must have at least X3 reads with the same variation between first and second states, being X3 at least 2; D2) The percentage of first state reads in second state reads is not over a threshold X4, to account for possible contamination of control state
- 10 reads with test state reads, wherein X4 is at least 5%. E) Computer/Electronic means for clustering and filtering test and control reads derived from all candidate breakpoints accepted in step D to build blocks, by carrying out the steps: E1) Retrieving reads which contain the stem of at least one k-mer that represents the candidate breakpoint selected in step D); E2) The reads of
- 15 step E1) with at least X5 k-mer variants within a window of X6 nucleotides are taken as leading reads, wherein X5 is at least 7 and X6 is at least 10; E3) Reads whose k-mers share at least one stem with a leading read are merged to give a block; and E4) If the nucleic acid whose variant is being identified is a double stranded DNA, then both forward and reverse variants are taken into
- 20 account when building the block. F) Computer/Electronic means for aligning blocks taking their leading reads as a reference: F1) For each read in the block, take the leading read's stem and find the longest inflection or partial inflection between the read and the leading read. F2) successively position each read so that its matching inflection or partial inflection is aligned against
- 25 the leading read.

The electronic/computer means may be used interchangeably, that is, a part of the described means may be electronic means and the other part may be computer means, or all described means may be electronic means or all

30 described means may be computer means. Examples of an apparatus comprising only electronic means may be a CPLD (Complex Programmable Logic Device), a FPGA (Field Programmable Gate Array) or an ASIC (Application-Specific Integrated Circuit).

- 35 A fourth aspect of the invention is a computer system comprising a processor and a memory, wherein the memory stores computer executable instructions



that, when executed by the processor, cause the system to perform the method for identifying nucleic acid variants between two genomic states. In some examples, the computer system may further comprise a hardware accelerator.

5

#### BRIEF DESCRIPTION OF THE DRAWINGS

FIG1. Data flow for the proposed configuration of the system. 1.1 Quality Check; 1.2 Base Conversion; 1.3 Reduce; 1.4 Hash; 1.5 Orchestable Data  
10 Movement; 1.6 Store Data into Associative Structure; 1.7 Load Sequences and Quality Markers; 1.8 Network; 1.9 Disk(s).

FIG2. Filtering capabilities of the method at different stages. -o- is Total Reads; -x- is Supporting Reads; -l- is Identifiable Mutations.

15

FIG3. Visualization of a breaking-block with an insertion of virus

#### DETAILED DESCRIPTION OF THE INVENTION

20 All terms as used herein, unless otherwise stated, shall be understood in their ordinary meaning as known in the art. Other more specific definitions for certain terms as used in the present application are as set forth below and are intended to apply uniformly throughout the description and claims unless an otherwise expressly set out definition provides a broader definition.

25

It must be noted that for clarity reasons, a variety of nucleotide sequences are given in the definitions and the examples found below. These nucleotide sequences are made up for these examples and they do not refer to any real nucleotide sequence of any organism. They are only listed so that the reader  
30 of the present application understands the terms used such as k-mer, stem, inflection, partial inflection, etc. They are absolutely unrelated to the invention being disclosed herewith, which has nothing to do with any particular nucleotide sequence.

35 The terms "computational method" and "computer implemented method" are taken here to mean the same and are used interchangeably. Therefore,

“computational method” and “computer implemented method” are taken as synonyms.

5 The terms “Next Generation Sequencing” (NGS), “deep sequencing”, “ultra-deep sequencing”, “high throughput sequencing” are all used interchangeably and refer to the technology platforms currently being used as standard to enable the sequencing of genomes (“sequencing methods”) with high speed and contained cost, such as the Roche/454, Illumina/Solexa, Life/APG and Pacific Biosciences platforms.

10

The term “base” and the term “nucleotide” are herein used interchangeably, and refer to the monomers (subunits) which are repeated in a nucleic acid such as DNA or RNA, giving its sequence or primary structure.

15 The term “reference genome” as used herein refers to the complete nucleic acid sequence representing the whole genome of a species normally accepted by the wide community. Since the reference genome is usually assembled from the sequencing of DNA from a number of donors, it does not accurately represent the set of genes of any one single individual. Instead, a reference genome provides a mosaic of different DNA sequences from each donor. But, at general levels, the reference genome provides a good approximation of the DNA of any single individual. However, in genomic regions with high allelic diversity, the reference genome may differ significantly from any one single individual. For example, GRCh37, the  
20 Genome Reference Consortium human genome (build 37) is derived from thirteen anonymous volunteers from New York. Reference genomes are typically used as a guide on which new genomes are built and aligned, enabling their assembly and comparison.

25

30 The term “forward strand” as used herein refers to a nucleic acid sequence read from 5’ terminal to 3’ terminal ends. The term “reverse strand” refers to the nucleic acid sequence which is complementary to the forward strand.

The term “nucleic acid variant” or simply “variant” as used herein refers to a  
35 difference in sequence between two genomic states. A variant can be a single nucleotide variant (SNV) if the difference between the two genomes (or two

states of the same genome) is only due to the change of a single nucleotide. All other variants, among them insertions, deletions, inversions, duplications, translocations and others are termed structural variants (SV). The latter can have many sizes, from two bases up to entire pieces of a chromosome.

5

The term “genomic state” as used herein can refer to two different genomes derived from two different individuals, or two genomes derived from two different cells of the same individual. In the second case, the two different cells can be a normal vs. a pathological cell, an undifferentiated vs. a differentiated cell, a cell which has been exposed to a certain external factor vs. an unexposed cell, etc.

10

The term “mapping” as used herein refers to aligning blocks of the first and second state to a reference genome.

15

The term “read” as used herein refers to a fragment of nucleic acid that is sequenced in its entirety. The nucleic acid might be DNA, RNA, or even chemically altered nucleic acids. The initial step in a high throughput sequencing run is the random fragmentation of a genome into millions of partly overlapping fragments called reads, which are usually amplified by Polymerase Chain Reaction and sequenced using a variety of techniques that are platform-dependent. The lengths of the reads can also vary depending on the platform, and are usually on the order of a few dozens to a few hundreds of nucleotides. The partly overlapping reads must be assembled if a complete picture of the genome is to be built.

20

25

The term “depth of coverage” as used herein refers to the number of times a nucleotide is read during the sequencing process. Deep sequencing means that the total number of reads is many times larger than the length of the sequence under study. Standard depth of coverage currently range from 30x to 100x for whole genomes, meaning that each position in the genome is represented from 30 to 100 times. Coverage similarly designates the average number of reads representing a given nucleotide in the reconstructed sequence.

30

35

Depth of coverage can be calculated from the length of the original genome

(G), the number of reads ( $N$ ), and the average read length ( $L$ ) as  $N \cdot L / G$ .

The term “undefined nucleotide” as used herein refers to a certain position inside a sequenced read that could not be determined during the sequencing process, that is, a position for which the sequencing experiment has not unambiguously resolved whether it is occupied by an adenine (A), guanine (G), cytosine (C) or thymine (T), and therefore its nature is unknown. Undefined nucleotides in reads are filtered out (removed) in the method of the invention, generating two or more fragments of defined sequence if the undefined nucleotides are removed from inner positions of the read.

The term “Phred quality score” as used herein refers to the quality score given to each nucleotide base call in a sequenced read. The Phred score is a property given to each sequenced nucleotide and it is logarithmically related to the base-calling error probability. A Phred score of 10 assigned to a certain nucleotide in a sequenced read means that there is a 90% probability that the base call is correct, a Phred score of 20 means that there is a 99% probability that the base call is correct, and a Phred score of 30 means that there is a 99.9% probability that the base call is correct.

The term “assembling” as used herein refers to grouping all the first state reads, and separately second state reads that share the same variant.

The term “hashtable” as used herein refers to a data structure used in computing that allows (by applying a hash function) assigning and mapping hashes (values) to strings of data, that is, it associates a series of hashes (values) to a series of strings in pairs, such that the association of hash-string is established. The addition, removal and modification of pairs is easily achieved by computational means, as well as the lookup and accession of strings thanks to their respective hashes (values).

The term “hash” as used herein refers to the value given by the hash function and linked to a certain string. This value allows computationally storing, retrieving, deleting and sorting strings in a very efficient manner.

The term “hash function” as used herein refers to the function that is used for

linking hashes (values given by the hash function) to strings of data given as input.

5 The term “k-mer” as used herein refers to all possible substrings of length  $k$  that are contained in a string. In genomics, all  $k$ -mers of a nucleic acid read are all the possible sub-sequences within the original read which have a length  $k$ . The amount of  $k$ -mers in a read of length  $M$  is  $M-k+1$ .

10 The term “polymorphic k-mer” as used herein refers to a  $k$ -mer that also identifies inflections and partial inflections of the  $k$ -mer’s stem. By polymorphic  $k$ -mer it is here understood the way the method of the invention handles  $k$ -mers, that is, the way they are used to derive stems and the latter to search for, manipulate and fish inflections and partial inflections.

15 The term “stem” as used herein refers to a fragment of a  $k$ -mer of length  $k$  with  $S$  defined bases, where  $S < k$ , and  $k-S$  omitted (undefined) bases. The stem fragment can either be a  $k$ -mer without a prefix, a  $k$ -mer without a suffix, a  $k$ -mer without an infix, or any combination thereof. Stem fragments without infix and/or prefix are consecutive, while stems without infixes can be non-  
 20 consecutive. In a stem, the character “-“ denotes a base that is omitted from the  $k$ -mer. Examples of the 30-mer SEQ ID NO: 1  
 CACGGCAGCTGAGTCAACAGGTTCTCCCA:  
 SEQ ID NO:2  
 CACGGCAGCTGAGTCAACAGGTTCTCCC- (omission of suffix of length 1)  
 25 SEQ ID NO:3  
 -ACGGCAGCTGAGTCAACAGGTTCTCCC- (omission of prefix of length 1, and suffix of length 1)  
 SEQ ID NO:4  
 CACG--AGCTGAGTCAACAGGTTCTCCCA (omission of infix of length 2  
 30 starting at position 5)

The term “prefix” as used herein, refers to the first part of a sequencing read, that is, from position 1 to a given position depending on the context. This term is used here, as it is used in a grammatical context referring to words.

35 The term “suffix” as used herein, refers to the last part of a sequencing read,

that is, from the last position to a given position depending on the context. This term is used here, as it is used in a grammatical context referring to words.

- 5 The term “infix” as used herein, refers to a part of a read positioned in the middle of the sequence.

The term “inflection” as used herein refers to a fragment of length  $k$  that can be derived from extending a stem of length  $k-1$ ,  $k-2$ ,  $k-3$ , etc. of a  $k$ -mer of length  $k$ . E.g. A stem of length  $k-1$  can be used to derive 4 inflections of length  $k$  since there is a single unknown position, and 4 different bases ( $4^1=4$ ). A stem of length  $k-2$  can be used to derive 16 inflections of length  $k$  ( $4^2=16$ ). Following the example given above, for the stem (SEQ ID NO:5): “CACGGCAGCTGAGTCAACAGGTTCTTCCC-“ (omission of suffix of length

10

15 1) the inflections would be (SEQ ID NO:6 to SEQ ID NO:9):

CACGGCAGCTGAGTCAACAGGTTCTTCCCA  
 CACGGCAGCTGAGTCAACAGGTTCTTCCCC  
 CACGGCAGCTGAGTCAACAGGTTCTTCCCT  
 CACGGCAGCTGAGTCAACAGGTTCTTCCCG

20 and further,

inflections based on stem (SEQ ID NO:10)

“-ACGGCAGCTGAGTCAACAGGTTCTTCCC-“ would be (SEQ ID NO:11-SEQ ID NO:26):

AACGGCAGCTGAGTCAACAGGTTCTTCCCA  
 25 AACGGCAGCTGAGTCAACAGGTTCTTCCCC  
 AACGGCAGCTGAGTCAACAGGTTCTTCCCT  
 AACGGCAGCTGAGTCAACAGGTTCTTCCCG  
 CACGGCAGCTGAGTCAACAGGTTCTTCCCA  
 CACGGCAGCTGAGTCAACAGGTTCTTCCCC  
 30 CACGGCAGCTGAGTCAACAGGTTCTTCCCT  
 CACGGCAGCTGAGTCAACAGGTTCTTCCCG  
 TACGGCAGCTGAGTCAACAGGTTCTTCCCA  
 TACGGCAGCTGAGTCAACAGGTTCTTCCCC  
 TACGGCAGCTGAGTCAACAGGTTCTTCCCT  
 35 TACGGCAGCTGAGTCAACAGGTTCTTCCCG  
 GACGGCAGCTGAGTCAACAGGTTCTTCCCA

GACGGCAGCTGAGTCAACAGGTTCTTCCCC  
 GACGGCAGCTGAGTCAACAGGTTCTTCCCT  
 GACGGCAGCTGAGTCAACAGGTTCTTCCCG

- 5 The term “partial inflection” as used herein refers to a fragment with P defined bases that can be derived from extending a stem of S defined bases of a k-mer of length k, and where  $S \leq P < k$ . In a partial inflection, the «.» character denotes a non-extended position of its stem. Partial inflections must have at least one non-extended position. Only omitted bases («-») can be marked as  
 10 non-extended.

Following the example given above:

Partial inflections based on stem (SEQ ID NO:27)

“-ACGGCAGCTGAGTCAACAGGTTCTTCCC-“ would be (SEQ ID NO:28 –  
 SEQ ID NO:36):

- 15 .ACGGCAGCTGAGTCAACAGGTTCTTCCC.  
 AACGGCAGCTGAGTCAACAGGTTCTTCCC.  
 CACGGCAGCTGAGTCAACAGGTTCTTCCC.  
 TACGGCAGCTGAGTCAACAGGTTCTTCCC.  
 GACGGCAGCTGAGTCAACAGGTTCTTCCC.  
 20 .ACGGCAGCTGAGTCAACAGGTTCTTCCCA  
 .ACGGCAGCTGAGTCAACAGGTTCTTCCCC  
 .ACGGCAGCTGAGTCAACAGGTTCTTCCCT  
 .ACGGCAGCTGAGTCAACAGGTTCTTCCCG

- 25 The term “breakpoint” as used herein refers to the the nucleotide position where the sequence changes, that is sequence immediately flanking a sequence variant. For SVs, a breakpoint is the point where the DNA broke in the second state and appears as a change in the first state compared to the second control state. In other words, where the continuity of the sequence of  
 30 the control second state breaks (changes) in the first state.

The term “leading read” as used herein refers to a complete sequenced read that contains at least one k-mer that is a candidate breakpoint (variant).

- 35 Normally, in the case of heterozygous variation, only the reads derived from the altered allele contain the mutation or variant.

The term “block” refers to a leading read along with all reads derived from the sequencing of all four alleles involved (two coming from the first state genome and two coming from the second state genome) covering the same region as the leading read.

5

The term “similarity score” as used herein refers to numbers that help to identify how different sets of aligned sequences are, and can be used as part of the proposed method to measure the quality of an aligned block. Similarity scores can be vertical or horizontal. The former measures, for every position in a sequence, how many bases in the set of aligned sequences are different than the mode base. The latter measures, for every sequence in the set, how many positions of the sequence are different to the mode/consensus sequence. Similarity scores can be measured for different sets of sequences, e.g. the set of control sequences, the set of test sequences, or the set containing both.

15

The term “ambiguous path” as used herein refers to multiple possible sequence solutions in a given tree. It is referred here as the opposite of unique and unambiguous path or sequence.

20

The terms defined above are used in the following example for increasing their clarity and conciseness:

Imagine the read to be input:

25

A) (SEQ ID NO: 37)  
CACGGCAGCTGAGTCAACAGGTTCTTCCCAGGAGCGGACGGCGGTGGC  
CAGCAGGCACGTG (its length N=61).

After the quality filtering (step B) of the method), the next step of the method C) would be to generate a hashtable with all the k-mers of length X2. If X2 is taken to be 30, then, there should be  $N-X2+1$  30-mers, that is,  $61-30+1=32$  (SEQ ID NO:38 – SEQ ID NO:69):

30

CACGGCAGCTGAGTCAACAGGTTCTTCCCAGGAGCGGACGGCGGT GGCCAGCAGGCACGTG
CACGGCAGCTGAGTCAACAGGTTCTTCCCA



```

ACGGCAGCTGAGTCAACAGGTTCTTCCCAG
CGGCAGCTGAGTCAACAGGTTCTTCCCAGG
GGCAGCTGAGTCAACAGGTTCTTCCCAGGA
GCAGCTGAGTCAACAGGTTCTTCCCAGGAG
CAGCTGAGTCAACAGGTTCTTCCCAGGAGC
AGCTGAGTCAACAGGTTCTTCCCAGGAGCG
GCTGAGTCAACAGGTTCTTCCCAGGAGCGG
CTGAGTCAACAGGTTCTTCCCAGGAGCGGA
TGAGTCAACAGGTTCTTCCCAGGAGCGGAC
GAGTCAACAGGTTCTTCCCAGGAGCGGACG
AGTCAACAGGTTCTTCCCAGGAGCGGACCG
GTCAACAGGTTCTTCCCAGGAGCGGACGGC
TCAACAGGTTCTTCCCAGGAGCGGACGGCG
CAACAGGTTCTTCCCAGGAGCGGACGGCGG
AACAGGTTCTTCCCAGGAGCGGACGGCGGT
ACAGGTTCTTCCCAGGAGCGGACGGCGGTG
CAGGTTCTTCCCAGGAGCGGACGGCGGTGG
AGGTTCTTCCCAGGAGCGGACGGCGGTGGC
GGTTCTTCCCAGGAGCGGACGGCGGTGGCC
GTTCTTCCCAGGAGCGGACGGCGGTGGCCA
TTCTTCCCAGGAGCGGACGGCGGTGGCCAG
TCTTCCCAGGAGCGGACGGCGGTGGCCAGC
CTTCCCAGGAGCGGACGGCGGTGGCCAGCA
TTCCCAGGAGCGGACGGCGGTGGCCAGCAG
TCCCAGGAGCGGACGGCGGTGGCCAGCAGG
CCCAGGAGCGGACGGCGGTGGCCAGCAGGC
CCAGGAGCGGACGGCGGTGGCCAGCAGGCA
CAGGAGCGGACGGCGGTGGCCAGCAGGCAC
AGGAGCGGACGGCGGTGGCCAGCAGGCACG
GGAGCGGACGGCGGTGGCCAGCAGGCACGT
GAGCGGACGGCGGTGGCCAGCAGGCACGTG

```

The hashtable to be generated with all k-mers and their number of times they are observed in first state and second state, would look like (SEQ ID NO:70 – SEQ ID NO:75):

K-mer	Normal	Tumor
ACGGCAGCTGAGTCAACAGGTTCTTCCCAG	0	1
CGGCAGCTGAGTCAACAGGTTCTTCCCAGG	0	1
GGCAGCTGAGTCAACAGGTTCTTCCCAGGA	0	1
GCAGCTGAGTCAACAGGTTCTTCCCAGGAG	0	1
CAGCTGAGTCAACAGGTTCTTCCCAGGAGC	0	1
AGCTGAGTCAACAGGTTCTTCCCAGGAGCG	0	1
...		

D) The next step in the method would be to detect variants between the first and second states. The first step would be to derive a stem for each one of the k-mers (SEQ ID NO:76 – SEQ ID NO:108):

5

CACGGCAGCTGAGTCAACAGGTTCTTCCCAGGAGCGGACGGCGGT GGCCAGCAGGCACGTG
CACGGCAGCTGAGTCAACAGGTTCTTCCC- ACGGCAGCTGAGTCAACAGGTTCTTCCCA- CGGCAGCTGAGTCAACAGGTTCTTCCCAG- GGCAGCTGAGTCAACAGGTTCTTCCCAGG- GCAGCTGAGTCAACAGGTTCTTCCCAGGA- CAGCTGAGTCAACAGGTTCTTCCCAGGAG- AGCTGAGTCAACAGGTTCTTCCCAGGAGC- GCTGAGTCAACAGGTTCTTCCCAGGAGCG- CTGAGTCAACAGGTTCTTCCCAGGAGCGG- TGAGTCAACAGGTTCTTCCCAGGAGCGGA- GAGTCAACAGGTTCTTCCCAGGAGCGGAC- AGTCAACAGGTTCTTCCCAGGAGCGGACG- GTCAACAGGTTCTTCCCAGGAGCGGACGG- TCAACAGGTTCTTCCCAGGAGCGGACGGC- CAACAGGTTCTTCCCAGGAGCGGACGGCG- AACAGGTTCTTCCCAGGAGCGGACGGCGG- ACAGGTTCTTCCCAGGAGCGGACGGCGGT- CAGGTTCTTCCCAGGAGCGGACGGCGGTG- AGGTTCTTCCCAGGAGCGGACGGCGGTGG-

```

GGTTCTTCCCAGGAGCGGACGGCGGTGGC-
GTTCTTCCCAGGAGCGGACGGCGGTGGCC-
TTCTTCCCAGGAGCGGACGGCGGTGGCCA-
TCTTCCCAGGAGCGGACGGCGGTGGCCAG-
CTTCCCAGGAGCGGACGGCGGTGGCCAGC-
TTCCCAGGAGCGGACGGCGGTGGCCAGCA-
TCCCAGGAGCGGACGGCGGTGGCCAGCAG-
CCCAGGAGCGGACGGCGGTGGCCAGCAGG-
CCAGGAGCGGACGGCGGTGGCCAGCAGGC-
CAGGAGCGGACGGCGGTGGCCAGCAGGCA-
AGGAGCGGACGGCGGTGGCCAGCAGGCAC-
GGAGCGGACGGCGGTGGCCAGCAGGCACG-
GAGCGGACGGCGGTGGCCAGCAGGCACGT-

```

For each stem, inflections are to be generated. For instance, the second stem found in the table above would give the following 4 inflections (SEQ ID NO:109 – SEQ ID NO:113):

ACGGCAGCTGAGTCAACAGGTTCT TCCCA-	ACGGCAGCTGAGTCAACAGG TTCTTCCCAA
	ACGGCAGCTGAGTCAACAGG TTCTTCCCAC
	ACGGCAGCTGAGTCAACAGG TTCTTCCCAT
	ACGGCAGCTGAGTCAACAGG TTCTTCCCAG

5

Find each one of the inflections in the hashtable:

K-mer	Normal	Tumor
ACGGCAGCTGAGTCAACAGGTTCTTCCCAA	0	1
ACGGCAGCTGAGTCAACAGGTTCTTCCCAC	0	4
ACGGCAGCTGAGTCAACAGGTTCTTCCCAT	0	1
ACGGCAGCTGAGTCAACAGGTTCTTCCCAG	0	1

If one inflection meets the requirements (has at least X3 reads with the same variation between the first and second states, and the amount of first state

reads in second state reads is not over X4), select k-mer as candidate breakpoint:

K-mer	Normal	Tumor
ACGGCAGCTGAGTCAACAGGTTCTTCCCAA	0	1
ACGGCAGCTGAGTCAACAGGTTCTTCCCAC	0	4
ACGGCAGCTGAGTCAACAGGTTCTTCCCAT	0	1
ACGGCAGCTGAGTCAACAGGTTCTTCCCAG	0	1

- 5 E) Clustering and filtering test and control reads derived from all candidate breakpoints:

For each candidate breakpoint selected in step D), retrieve reads which contain the stem of the k-mer (SEQ ID NO:114 – SEQ ID NO:116):

Reads for candidate breakpoint based on k-mer AGTCAACAGGTTCTTCCCAGGAGCGGACGC
CACGGCAGCTGAGTCAACAGGTTCTTCCCAGGAGCGGACGGCGG TGGCCAGCAGGCACGTG
AGCAGGCACGTGACGGCAGCTGCAGTCAACAGGTTCTTCCCAGG ACGGGACGCCGGTGGCC

10

Then, for each read, we will end up with all k-mers that are candidate breakpoints, and their positions in the read (SEQ ID NO:117 – SEQ ID NO:126):

Read #1
CACGGCAGCTGAGTCAACAGGTTCTTCCCAGGAGCGGACGGCGG TGGCCAGCAGGCACGTG
ACGGCAGCTGAGTCAACAGGTTCTTCCCAG (1) GGCAGCTGAGTCAACAGGTTCTTCCCAGGA (3) GCAGCTGAGTCAACAGGTTCTTCCCAGGAG (4) CAGCTGAGTCAACAGGTTCTTCCCAGGAGC (5)

AGCTGAGTCAACAGGTTCTTCCCAGGAGCG (6)
GCTGAGTCAACAGGTTCTTCCCAGGAGCGG (7)
TGAGTCAACAGGTTCTTCCCAGGAGCGGAC (9)
GAGTCAACAGGTTCTTCCCAGGAGCGGACG (10)
AGTCAACAGGTTCTTCCCAGGAGCGGACGG (11)

(SEQ ID NO:127 – SEQ ID NO:128)

Read #2

AGCAGGCACGTGACGGCAGCTGCAGTCAACAGGTTCTTCCCAGG AGCGGACGCCGGTGGCC
-------------------------------------------------------------------

AGTCAACAGGTTCTTCCCAGGAGCGGACGC (23)
-------------------------------------

Reads with (X5) 7 k-mers containing candidate breakpoints within a window of (X6) 10 nucleotides are taken as leading reads (SEQ ID NO:129).

Leading reads

CACGGCAGCTGAGTCAACAGGTTCTTCCCAGGAGCGGACGGCGG TGGCCAGCAGGCACGTG
-------------------------------------------------------------------

5

Generate stems from leading reads (SEQ ID NO:130 – SEQ ID NO:138):

CACGGCAGCTGAGTCAACAGGTTCTTCCCAGGAGCGGACGGCGG TGGCCAGCAGGCACGTG
-------------------------------------------------------------------

ACGGCAGCTGAGTCAACAGGTTCTTCCA- GGCAGCTGAGTCAACAGGTTCTTCCCAGG- GCAGCTGAGTCAACAGGTTCTTCCCAGGA- CAGCTGAGTCAACAGGTTCTTCCCAGGAG- AGCTGAGTCAACAGGTTCTTCCCAGGAGC- GCTGAGTCAACAGGTTCTTCCCAGGAGCG- TGAGTCAACAGGTTCTTCCCAGGAGCGGA- GAGTCAACAGGTTCTTCCCAGGAGCGGAC- AGTCAACAGGTTCTTCCCAGGAGCGGACG-
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Build blocks by finding other candidate reads that share stems with leading

reads (SEQ ID NO: 139 – SEQ ID NO:141):

Leading read #1
CACGGCAGCTGAGTCAACAGGTTCTTCCCAGGAGCGGACGGCGG TGGCCAGCAGGCACGTG
Additional reads
AGCAGGCACGTGACGGCAGCTGCAGTCAACAGGTTCTTCCCAGG AGCGGACGCCGGTGGCC
Shared stems
AGTCAACAGGTTCTTCCCAGGAGCGGACG-

5 By executing the steps outlined above in the right order, the application of the computer-implemented method of the invention allows to easily cut the huge numbers of reads given in the input down to a much reduced number where most of the true positives are found, as will be seen in the experimental data found in the examples section (below).

10 As is revealed in Paszkiewicz K., et.al. "De novo assembly of short sequence reads" Brief Bioinform. 2010, vol. 11, pp. 475-472, the approximate minimum length that NGS reads must have in order to be able to reconstruct a genome is around 30. Bearing in mind the latter, a minimum length of approximately  
15 30 bases was taken to be the minimum length of a productive read. 30 is a value that was found to be a viable cutoff for variable X2 in the definition of the method of the invention, although slightly smaller values might be viable as well.

20 As it has been cited above, the first aspect of the present invention is a computer-implemented method for identifying nucleic acid variants between two genomic states comprising the steps of: A) Inputting 2 sets of nucleic acid reads, which are sequences retrieved from a nucleotide sequencing method, wherein the first set of reads corresponds to cells representing a first test  
25 state, and the second set of reads corresponds to cells representing a second control state;B) Filtering the reads, wherein the filtering comprises: B1)

Keeping only the reads with at least a percentage X1 of their bases with a Phred quality score higher than 20, being X1 equal to or above 90%; B2) Splitting the reads with an undefined nucleotide, giving one sequence before, and one sequence after the undefined nucleotide, the latter being discarded;

5 and B3) Discarding the sequence reads with less than X2 bases, wherein X2 is from 25 to 50; C) Generating a hashtable structure comprising: C1) Generating a number of N-X2+1 new reads for each read of sequence length N, wherein the new N-X2+1 reads correspond to all k-mers with length X2 nucleotides; and C2) Building a hashtable structure, which comprises all the

10 k-mers generated in step C1) and further comprises the number of times each k-mer is observed in the two sets of reads corresponding to first and second states. D) Detecting candidate variants in the sequence between first state and second state, wherein a k-mer of the hashtable structure is taken as a candidate breakpoint, which represents a variant between the first and

15 second states, if it fulfills all the following requirements: D1) At least one inflection based on a k-mer's stem must have at least X3 reads with the same variation between first and second states, being X3 at least 2; 2) The percentage of first state reads in second state reads is not over a threshold X4, to account for possible contamination of control state reads with test state

20 reads, wherein X4 is at least 5%. E) Clustering and filtering test and control reads derived from all candidate breakpoints accepted in step D to build blocks, by carrying out the steps: E1) Retrieving reads which contain the stem of at least one k-mer that represents the candidate breakpoint selected in step D); E2) The reads of step E1) with at least X5 k-mer variants within a

25 window of X6 nucleotides are taken as leading reads, wherein X5 is at least 7 and X6 is at least 10; E3) Reads whose k-mers share at least one stem with a leading read are merged to give a block; and E4) If the nucleic acid whose variant is being identified is a double stranded DNA, then both forward and reverse variants are taken into account when building the block. F) Aligning

30 blocks taking their leading reads as a reference: F1) For each read in the block, take the leading read's stem and find the longest inflection or partial inflection between the read and the leading read. F2) Successively position each read so that its matching inflection or partial inflection is aligned against the leading read.

35

In a particular embodiment of the first aspect of the invention, the computer-

implemented method further comprises the step following F2: F3) F3)  
Obtaining first state scores, second state scores, and global similarity scores  
of each position in the block by measuring a ratio of most frequent nucleotide  
in that position relative to the total number of nucleotides.

5

In a particular embodiment of the first aspect of the invention, the computer-  
implemented method further comprises the step: G) Cataloguing and  
annotating blocks according to the following: G1) If blocks between the first  
and second states only differ in one substituted nucleotide, the variant is  
10 catalogued as containing a single nucleotide variant and the single nucleotide  
variant is annotated; G2) If blocks between the first and second states differ  
in more than one nucleotide but the whole difference in sequence is  
contained within the block, the variant is catalogued as a small structural  
variant, and the small structural variant is annotated; and G3) If blocks  
15 between the first and second states differ in more than one nucleotide and the  
whole difference in sequence is not contained within the block, the variant is  
catalogued as a large structural variant, and the boundaries of all large  
structural variants are extended by retrieving blocks overlapping at least X2  
nucleotides in an iterative process which ends when the extended sequence  
20 reaches 200 nucleotides or when an ambiguous path is found.

In a particular embodiment of the first aspect of the invention, the computer-  
implemented method further comprises the step: H) Filtering of the blocks,  
according to the following: H1) The percentage of second state reads in first  
25 state reads is not over a threshold X7, to account for possible contamination  
of test state reads with control state reads, wherein X7 is at least 20%;

In a particular embodiment of the first aspect of the invention, the method  
further comprises optionally mapping second state blocks, and subsequently  
30 mapping first state blocks, on a reference genome.

In a particular embodiment of the first aspect of the invention, optionally in  
combination with any embodiment above or below, X1 is equal or above 95%.

35 In a particular embodiment of the first aspect of the invention, optionally in  
combination with any embodiment above or below, X1 is equal or above 99%.



In a particular embodiment of the first aspect of the invention, optionally in combination with any embodiment above or below, X2 is from 25 to 40.

- 5 In a particular embodiment of the first aspect of the invention, optionally in combination with any embodiment above or below, X2 is from 30 to 35.

In a particular embodiment of the first aspect of the invention, optionally in combination with any embodiment above or below, X2 is from 30 to 32.

10

In a particular embodiment of the first aspect of the invention, optionally in combination with any embodiment above or below, X2 is equal to 30.

- 15 In a particular embodiment of the first aspect of the invention, optionally in combination with any embodiment above or below, X3 is equal or above 4.

In a particular embodiment of the first aspect of the invention, optionally in combination with any embodiment above or below, X3 is equal or above 6.

- 20 In a particular embodiment of the first aspect of the invention, optionally in combination with any embodiment above or below, X3 is equal or above 8.

- 25 In a particular embodiment of the first aspect of the invention, X3 is directly proportional to the depth of coverage in the sequencing experiment. This means that, the deeper the coverage, the more restrictive (higher) is the value for X3.

- 30 In a particular embodiment of the first aspect of the invention, optionally in combination with any embodiment above or below, X4 is between 5-10%.

In a particular embodiment of the first aspect of the invention, optionally in combination with any embodiment above or below, X4 is between 5-7%.

- 35 In a particular embodiment of the first aspect of the invention, optionally in combination with any embodiment above or below, X4 is 5%.

X4 is expressed as a percentage, reflecting the maximum accepted ratio of first state (test) reads vs. second state (control) reads for each of the k-mers, and represents the levels of contamination expected for each of the samples (usually in the direction of tumor cells within normal samples). This value  
5 should be set by the user accordingly. Setting up a low value for X4 ensures high specificity but might result in a lower sensitivity, whereas the selection of high values, might result in the accumulation of false positives.

10 In a particular embodiment of the first aspect of the invention, optionally in combination with any embodiment above or below, X5 is from 10 to 15.

In a particular embodiment of the first aspect of the invention, optionally in combination with any embodiment above or below, X5 is from 12 to 14.

15 In a particular embodiment of the first aspect of the invention, optionally in combination with any embodiment above or below, X6 is from 12 to 25.

20 In a particular embodiment of the first aspect of the invention, optionally in combination with any embodiment above or below, X6 is from 12 to 20.

In a particular embodiment of the first aspect of the invention, optionally in combination with any embodiment above or below, X6 is from 12 to 15.

25 In a particular embodiment of the first aspect of the invention, optionally in combination with any embodiment above or below, X7 is between 20-25%.

In a particular embodiment of the first aspect of the invention, optionally in combination with any embodiment above or below, X7 is 20%.

30 In a particular embodiment of the first aspect of the invention, optionally in combination with any embodiment above or below, X5 is from 10 to 15 and the threshold X6 is from 12 to 20.

35 In a particular embodiment of the first aspect of the invention, optionally in combination with any embodiment above or below, X5 is from 12 to 14 and the threshold X6 is from 12 to 15.

In a particular embodiment of the first aspect of the invention, the first set of reads corresponds to pathological cells of a patient, and the second set of reads corresponds to non-pathological cells of the same patient;

5

In another particular embodiment of the first aspect of the invention, the first set of reads corresponds to cancer cells of a patient, and the second set of reads corresponds to non-cancer cells of the same patient.

10 In another particular embodiment of the first aspect of the invention, the first set of reads corresponds to virus-infected cells of a patient, and the second set of reads corresponds to non-infected cells of the same patient.

15 In another particular embodiment of the first aspect of the invention, the first set of reads and the second set of reads correspond to the same cell of the same patient in two different developmental stages.

20 In another particular embodiment of the first aspect of the invention, the first set of reads corresponds to cells of a patient which have been exposed to a drug, and the second set of reads corresponds to cells of the same patient which have not been exposed to a drug.

25 In a particular embodiment of the first aspect of the invention, the first set of reads corresponds to cells of a tissue, and the second set of reads corresponds to cells of another tissue of the same or a different individual.

30 Although the present invention has been described in detail for purpose of illustration, it is understood that such detail is solely for that purpose, and variations can be made therein by those skilled in the art without departing from the scope of the invention.

35 Thus, while the preferred embodiments of the methods and of the systems have been described in reference to the environment in which they were developed, they are merely illustrative of the principles of the invention. Other embodiments and configurations may be devised without departing from the scope of the appended claims.

Further, although the embodiments of the invention comprise processes performed in computer systems, the invention also extends to computer systems and to computer programs (which may be embodied on a storage medium and/or carried on a carrier signal) adapted for putting the invention into practice.

Accordingly, the invention also provides a computer program product comprising program instructions for causing a computer system to perform the method for identifying nucleic acid variants between two genomic states as defined above.

In a preferred embodiment, the computer program product is embodied on a storage medium.

In another preferred embodiment, the computer program product is carried on a carrier signal. The carrier may be any entity or device capable of carrying the program.

As it has been cited above, a fourth aspect of the invention is a computer system comprising a processor and a memory, wherein the memory stores computer executable instructions that, when executed by the processor, cause the system to perform the method for identifying nucleic acid variants between two genomic states.

In a preferred embodiment of the fourth aspect of the invention, the system may further comprise a hardware accelerator, which may be in some examples an FPGA or a GPU.

Throughout the description and claims the word "comprise" and variations of the word, are not intended to exclude other technical features, additives, components, or steps. Furthermore, the word "comprise" and its variations encompasses the term "consisting of". Additional objects, advantages and features of the invention will become apparent to those skilled in the art upon examination of the description or may be learned by practice of the invention. The following examples are provided by way of illustration, and they are not

intended to be limiting of the present invention. Furthermore, the present invention covers all possible combinations of particular and preferred embodiments described herein.

## 5 EXAMPLES

Examples of using the method of the invention for detecting characterizing sequence variants in nucleic acid sequences are given below.

- 10 In the *in silico* tests it is revealed that the method of the invention is capable of identifying SNVs and SVs of all sorts and sizes. Remarkably, the method of the invention is even proven to be capable of identifying novel non-human insertions. In one of the examples found below, the method is remarkably capable of detecting the insertion of a virus where, other methods (including
- 15 the one disclosed in Moncunill et al., *ibid*) fail.

### Material and methods

An implementation of the computer-implemented method

- 20 The general structure and the complete variant identification and characterization carried out by the method of the invention comprise the steps outlined below:

#### 25 A) Input data.

As input, the method takes high quality sequences data directly from FASTQ files of tumor and non-tumor control cells samples of the same individual. Alternatively, it is also able to accept BAM files, from which it extracts all the sequencing reads. Tumor sample corresponds to the first state and non-tumor control sample correspond to the second state.

30

#### B) Filtering the data.

- When inputting the data, the user can define a cut-off so that reads having over a certain threshold of their bases with a Phred quality score <math>q20</math> are discarded. X1=90 has been found to be especially suited for the purposes tested. This means that only reads with at least 90% of their bases with a
- 35

Phred quality score higher than 20 are kept. In the case of the presence of undefined base pairs ("N"), these are removed and the original sequence is split forming new shorter reads, which are considered only if they are longer than  $X2$  base pairs.

5

In order to lower the amount of space needed to store k-mers, they are converted into integers by mapping each base of the k-mer to a 2-bit code. For instance, a k-mer of length 32 represented as a sequence of characters takes 256 bits, but after the conversion it is turned into a single value of 64 bits.

10

In a particular embodiment of the computer-implemented method, wherein hardware accelerator(s) may be used, reads that don't meet the aforementioned quality criteria are discarded by means of marking their k-mers as discardable, which are then ignored in the subsequent steps of the pipeline. Marking k-mers as discardable instead of deleting them immediately enables a faster pipeline without conditional execution.

15

C) Generating a hashtable structure.

20

After the quality filtering step of the method, the next step of the method is to generate a hashtable with all the k-mers of length  $X2$  using all high quality tumor and non-tumor control reads (see Table 1 below for a simplified version). If  $X2$  is taken to be 28, then, there should be  $N-X2+1$ , that is,  $100-28+1=73$  resultant k-mers for a 100-nucleotide read. Each of k-mers generated is inserted into the hashtable and their number of times they are observed in tumor (test state) and non-tumor (second control state) cells.

25

The mapping of k-mers to their observed frequencies in the input is, generally speaking, one to one, meaning each entry of the associative data structure contains a pair of frequencies. In order to find the position where to store data into an associative structure a hash function is used to compute an index into a position, from which the desired value can be stored and retrieved. Any function that guarantees a homogeneous distribution of the results can be used as hash function.

30

35

Table 1

(SEQ ID NO:142 – SEQ ID NO:145)

kmer (length X2)	Frequencies
ACTGACTGACTGACTGACTGACTGAA	(0, 1)
ACTGACTGACTGACTGACTGACTGAC	(1, 0)
ACTGACTGACTGACTGACTGACTGAG	(23, 2)
ACTGACTGACTGACTGACTGACTGAT	(9, 10)
...	

15 Each item of the hashtable consist of its k-mer (in nucleotide string or encoded key format), along with its pair frequencies in the first state and second state sets of reads (Table 1).

20 In particular embodiments of this aspect, each entry of the associative data structure may contain frequencies for more than one k-mer. This is accomplished by means of indexing stems instead of k-mers. Table 2 below depicts such an example, where stems are basically the original k-mer truncating the last base, which is then included as part of the list of frequencies. Indexing stems instead of k-mers improves the locality of the data, reducing the number of lookup queries.

stem (length X2-1)	Frequencies
ACTGACTGACTGACTGACTGACTGACTGA	A: (0, 1) C: (1, 0) G: (23, 2) T: (9, 10)
ACTGACTGACTGACTGACTGACTGACTGC	
...	

25

Table 2 (SEQ ID NO:146 and SEQ ID NO:147)

30 In a particular embodiment, the hash function operates over the encoded key as described in step B instead of the k-mer containing a string of nucleotides. In order to lower the number of updates to the associative data structure, a

particular embodiment of this aspect involves generating an additional structure to store the set of k-mers seen only once, meaning the main data structure is only updated for k-mers with a frequency of 2 or more.

- 5 In another embodiment, a lower number of updates to the main data structure is achieved by generating partial data structures containing frequencies for a subset of the input, which are then merged into the main associative data structure.
- 10 D) Detecting k-mers containing variants once all the reads are derived on k-mers and loaded into the hashtable structure, the next step consists in identifying all tumor specific reads. Inventors expect that variants generate new and distinct sequences in the tumor genome compared to the non-mutated control genome.
- 15 If one inflection meets the requirements: has at least 4 (X3) k-mers with the same variation between tumor cells and maximum 1 (X4) k-mer non-tumor control cells, then this k-mer is select as a candidate breakpoint.
- 20 The detection of entries of the hashtable containing k-mers with variants between the first and second state is accomplished by reading the filtered input again, and selecting all reads that contain k-mers whose frequencies meet certain criteria.
- 25 In addition to selecting candidate variants, in this step the implementation also selects additional information needed during later steps of the pipeline, namely: 1) selection of relative reads; 2) position of candidate k-mers for each read; and 3) map of k-mers to reads. In particular, the selection of relative reads is done based on stems, and checking whether any inflection matches
- 30 the criteria described in the previous paragraph.
- E) Clustering and filtering test and control reads  
For each candidate breakpoint selected in step D), retrieve reads which contain the stem of the k-mer. Then for each read, we will end up with all k-
- 35 mers that are candidate breakpoints, and their positions in the read. Reads with 7 (X5) k-mers containing candidate breakpoints within a window of 10



(X6) nucleotides are taken as leading reads. This process is in order to find enough support information for each breakpoint detected (if not it is removed from the candidate list). Reads whose k-mers share at least one stem with a leading read are merged to give a block, and if the nucleic acid whose variant is being identified is a double stranded DNA, then both forward and reverse variants are taken into account when building the block.

F) Aligning blocks taking their leading reads as a reference.

For each read in the block, inventors take the leading read's stem and find the longest inflection or partial inflection between the read and the leading read. Successively position each read so that its matching inflection or partial inflection is aligned against the leading read.

Ideally each block represents a region in the genome containing the mutated and the non-mutated version. In order to classify and characterize the type of variation identified, the method takes into account the align score for tumor block, non-tumor block and the sum up of both. Then the method extracts the consensus mutated and normal sequences from these blocks. The corresponding normal consensus sequence can be used at the end of the procedure and mapped onto a reference genome to obtain the coordinates of the variant.

Optionally the method also can include step G

G) Cataloguing and annotating blocks

Once all possible breakpoint blocks are defined, the next step consists in identifying and classifying the variation included there. At this point the method uses the aligning score to observe the differences in each group: tumor, non-tumor and both. For each position on the block supported by a read it puts a value depending on the similarity, so finally it has a representation of all the variability on the block. These aligning scores are recursively compared to identify differences between tumor and non-tumor samples. A first evaluation will search a consensus in non-tumor block in order to avoid false positives and wrong alignments from different regions, and the same way on tumor blocks. The next step searches for all the small variants, which consist on those that are completely included within the block (SNV and small SVs: insertions, deletions and inversions). All the blocks that

do not match this criterion are then considered candidates for large SVs, i.e. likely to cover break points of intra or interchromosomal transitions, part of large deletions, insertions or inversions.

- 5 After small and large somatic variants are defined, the method of the invention identifies the coordinates of the changes by mapping onto a reference genome the normal consensus sequences corresponding to each of the variants, avoiding potential mapping conflicts derived from the presence of the variant, as usually happens when using reference-based
- 10 approaches. Sequences mapping (with the same score) to several positions in the genome are discarded. The same process can be done by mapping onto virus databases to locate which are the viruses inserted in the genome being analyzed.
- 15 Construction of the in silico chromosome 20  
In order to measure and calibrate the detection capabilities (sensitivity) of the method of the invention, inventors executed it on a controlled system, consisting of modified sequences of the chromosome 20.
- 20 A personalized chromosome 20 was extracted from the hg19 reference genome downloaded from UCSC (with no repeat-masking) (<http://www.ucsc.edu>) and modified to match a randomly chosen human haplotype. This chromosome contains 148,639 variants consisting of 96,935 SNPs and 51,704 deletions. The catalogue of somatic variants further added to this personalized chromosome and constituting the target of the invention, includes random 168 SNVs, 12 random deletions, 18 random insertion, 6
- 25 inversions and 1 insertion of KI polyomavirus (extracted from: <http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?lvl=0&id=423445>)
- 30 In silico sequencing was simulated using ART Illumina (Huang W. et.al. "ART: a next-generation sequencing read simulator" *Bioinformatics* 2012, vol. 28, pp. 593-594). For this inventors, like on the previous in silico, first generated a profile using the M0004 sample to extract parameters, like sequencing
- 35 variation or read length.

Analysis of the in silico chromosome 20 with the method of the invention

The method of the invention was run with using the next variables  $X1=90$ ,  $X2=28$ ,  $X3=4$ ,  $X4=1$ ,  $X5=7$  and  $X6=10$  considered as default at the moment of the invention.

5

For the sake of clarity and simplicity, inventors have set up  $X4$  as 1 in this example of the in silico chromosome 20.

A System Configuration

10

The invention also comprised an efficient hardware configuration for the system focused on the execution of the computer program performing the method.

Due to the characteristics of the method, which involve a highly parallelizable input without dependencies and huge amounts of intermediate data, the input

15

sources are split into chunks of  $C1$  sequences and processed in a pipelined fashion. Thus, each step of the method is executed by a different component of the computer system, in parallel, over different chunks of input data. This strategy is more efficient in terms of execution time since it attempts to maximize the utilization of the system, and also overlaps data movement

20

times.

When accelerators are capable of streaming data while computing, it is possible to add additional steps to the pipeline, eg. data is being sent to and received from the accelerators. When using accelerators that have

25

interconnections unable to offer bidirectional concurrent transfers, data movement must be serialized and this could become the bottleneck of the entire pipeline. Generally, the bigger the number of sequences per chunk ( $C1$ ), the higher the achievable bandwidth on data transfers from host system to accelerators and vice versa is. On the other hand, this number is limited by

30

the memory capacity of the accelerators. Moreover, encoding the keys may require to sorting all keys and some parallel sorting algorithms require a number of items that's a power 2, meaning extra padding might be required in order to reach the next power of 2. When this happens, the number of reads sent to each accelerator must be accordingly chosen in order to minimize the

35

number of padding elements.

The hardware components of the computer system consisted of:

A) Input Source

As input source, the processing pipeline reads high quality sequencing data  
 5 from local disks or from network resources. An example of a network  
 resource, is a sequencing machine connected to the system. In this case the  
 system receives the input reads directly from the sequencing machine and  
 starts elaborating at the same time as the sequencing machine is still  
 working.

10

B) Host System, and C) Accelerators

In order to compute and elaborate data, the invention comprises processors  
 and accelerators for example, hardware in the form of PCI cards or network  
 resources, for offloading part of the computation. The amount of work to  
 15 offload to the accelerators depends on the interconnection between  
 accelerator and host system as well on the processing power and memory of  
 both processors and accelerators.

Ideally, processors will read the input sources that might need to be  
 uncompressed and filtered in order to extract only sequence reads and  
 20 quality markers. Once enough sequences to fill a chunk are loaded in  
 memory, sequences and relative quality markers are sent to the accelerators.  
 In turn, accelerators will: filter reads, convert all read fragments to their  
 encoded key representation; reduce k-mers producing the <key, count> pairs;  
 and eventually, hashing the keys to obtain the tuple <key, count, hash>. Once  
 25 the data is transferred back to the system's main memory, processors will  
 consume it updating the counters stored in the associative structure.  
 Optionally, when accelerators have network interconnections and inputs are  
 streamed from network resources, accelerators can read input directly from  
 the source without requiring any work from the main system, offloading even  
 30 more work. When the capabilities of the accelerators are limited, some of the  
 processing steps can be carried out by the processors instead. For example,  
 if transferring the tuple <key, count, hash> from accelerators to the host  
 system becomes the bottleneck of the processing pipeline, it's wise to migrate  
 the hashing step to the processors. In this way, it is reduced to 3/5 the  
 35 amount of data that is transferred from the accelerators to the host system.

If the steps offloaded to the accelerators constitutes the bottleneck and having a higher number of updates to the hash table does not have an impact on the performance, the key conversion step can be disabled in order to increase the global throughput.

5

If more than one accelerator is available, data is split in different parts and each one of the parts is sent to a different accelerator. In order to efficiently offload the computation among multiple accelerators, the criterion used to split the data may take into account the available memory of each accelerator and may be as balanced as possible to the throughput offered by each accelerator. Which, for example, is: if two accelerators are available and one of them offers a throughput that is double of the other then the former accelerator should process an amount of data that is about 2/3 of the total, meanwhile the latter accelerator is about 1/3. If accelerators are different among each other and it's clear that different accelerators might suit better for different steps; then, the steps of the method are split among the available accelerator assigning each step to the accelerator that suits better. In this case, the output of one accelerator becomes the input of another and if no direct interconnection is available between the accelerators the host system must intervene to orchestrate data transfers. On the other hand, if no accelerators are available all the computation must be carried out by the main processors.

10

15

20

#### D) Main Memory, and E) Memory Expansion Cards

25

In addition to staging intermediate data from a one step of the pipeline to the next one, main memory is also used to store a partial version of the associative structure. This in-memory structure is used to store k-mers seen just few times, while those seen more than 11 times are stored into a permanent associative structure in the memory expansions card.

30

In order to store the persistent associative structure containing all the useful information from normal (CNR) and tumoral (CTR) sample, memory expansion cards are used. When multiple cards are available inventors can use each card to store part of the associative structure allowing to split read requests among the cards increasing the possible number of in-flight requests proportionally with the number of the cards. Examples of an apparatus that

35

can be used as memory expansion cards may be NVMe cards.

FIG. 1 shows the data flow from the input source, first to host system to load the reads, and then to accelerators for quality checking and base conversion.

- 5 Afterwards, data is reduced again by the accelerators, which also generate the hash before data is loaded back into memory.

### Results

- 10 As explained above, to assess the performance of the method of the invention, inventors measured the fraction of somatic variants detected (sensitivity).

In silico validation with chromosome 20

- 15 This sample was created exclusively in order to validate the method against an in silico sample with an insertion of a virus, also the sample includes different somatic mutations.

- Inventors first observed that the calling of somatic SNVs and SVs is optimal with a sensitivity of 96.4% for SNVs, 96.2% for Small SVs, 100% for Large SVs and 100% for virus Insertion. It is remarkable the capacity of detecting large structural variants with such a high sensitivity (Table 3)
- 20

	Mutations	After filtering	% detection	After clustering	% detection
Point mutations	168	167	99.4%	162	96.4%
Small SVs (indels)	26	26	100.0%	25	96.2%
Large SVs	10	10	100.0%	10	100.0%
Virus	1	1	100.0%	1	100.0%

Table 3. Assessment variant calling for chromosome 20

25

FIG. 2 provides an overview of the filtering capabilities of the invention, along with its sensitivity, at different stages of the execution. Step D) (filter) reduces the number of input reads to less than 10% of the input, while still keeping 99% of mutations identifiable, and high number of reads supporting those

mutations (72%). Step E) (clustering) further decreases the size of relevant reads, while still keeping very high number of identifiable mutations.

Besides that, it is also capable of detecting the insertion of the KI polyomavirus. The computer-implemented method of the invention allows finding viral integration events with better accuracy and recall than the available alternative methods which are based on pre-alignment steps of the reads onto a reference genome. (FIG. 3).

In the case of normal reads, inventors observed the genome sequence without any kind of structural variation. In contrast, on tumor reads (with heterozygosity A1-A2) inventors were able to detect the KI polyomavirus insertion on chromosome 20 at position 56398701.

In silico validation of the chromosome 20 insertion with the method disclosed in Moncunill et. al (ibid):

This method, with the default parameters, performed poorly in this test. Its results on large structural variants does not show any evidence of the detection of KI polyomavirus insertion on chromosome 20 at position 56398701, it was just able to describe variants from the own genome. Therefore, the method of the present invention was shown to be superior to the reference-free suffix-tree- based method described in Moncunill et al.

Taking into account all the results obtained at this point, the method of the invention has shown the capacity to detect all kinds of SNVs and SVs with great sensitivity without restrictions with the size. Also it is capable to detect the insertions of a virus in the genome with a base-pair resolution in comparison with the available alternative methods, which are based on suffix trees.

#### REFERENCES CITED IN THE APPLICATION

Li, H., et.al. "Fast and accurate short read alignment with Burrows-Wheeler transform" *Bioinformatics* 2009, vol. 25, pp. 1754-1760

Medvedev P. "Computational methods for discovering structural variation with

- next-generation sequencing" Nat. Methods Suppl. 2009, vol. 6, S13-S20
- Ding, L., et.al. "Expanding the computational toolbox for mining cancer genomes" Nat. Rev. Genet. 2014, vol. 15, pp. 556-570
- 5 Chen K. "TIGRA: A targeted Iterative Graph Routing Assembler for Breakpoint Assembly" Genome Res. 2016, vol. 24, pp.310-317.
- Zhuang, J. "Local sequence assembly reveals a high-resolution profile of somatic structural variations in 97 cancer genomes" Nucleic Acid Research
- 10 2015, vol. 43, pp.8146-8156
- Moncunill V., et al., "Comprehensive characterization of complex structural variations by directly comparing genome sequence reads" Nature Biotech. 2014, vol. 32, pp. 1106-1112
- 15 Patro R., et. al. "Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms" Nature Biotech. 2014, vol. 32, pp. 462-464
- 20 Nordstrom K.J.V, et.al. "Mutation identification by direct comparison of whole-genome sequencing data from mutant and wild-type individuals using k-mers" Nature Biotech. 2013, vol. 31, pp. 325-331
- Paszkiwicz K., et.al. "De novo assembly of short sequence reads" Brief Bioinform. 2010, vol. 11, pp. 475-472
- 25 Huang W. et.al. "ART: a next-generation sequencing read simulator" Bioinformatics 2012, vol. 28, pp. 593-594



## CLAIMS

1. A computer-implemented method for identifying of nucleic acid variants between two genomic states comprising the steps of:

5

A) Inputting 2 sets of nucleic acid reads, which are sequences retrieved from a nucleotide sequencing method, wherein the first set of reads corresponds to cells representing a first test state, and the second set of reads corresponds to cells representing a second control state;

10

B) Filtering the reads, wherein the filtering comprises:

B1) Keeping only the reads with at least a percentage X1 of their bases with a Phred quality score higher than 20, being X1 equal to or above 90%;

15

B2) Splitting the reads with an undefined nucleotide, giving one sequence before, and one sequence after the undefined nucleotide, the latter being discarded; and

B3) Discarding the sequence reads with less than X2 bases, wherein X2 is from 25 to 50;

20

C) Generating a hashtable structure comprising:

C1) Generating a number of  $N-X2+1$  new reads for each read of sequence length N, wherein the new  $N-X2+1$  reads correspond to all k-mers with length X2 nucleotides; and

25

C2) Building a hashtable structure, which comprises all the k-mers generated in step C1) and further comprises the number of times each k-mer is observed in the two sets of reads corresponding to first and second states.

D) Detecting candidate variants in the sequence between first state and second state, wherein a k-mer of the hashtable structure is taken as a candidate breakpoint, which represents a variant between the first and second states, if it fulfills all the following requirements:

30

D1) At least one inflection based on a k-mer's stem must have at least X3 reads with the same variation between first and second states, being X3 at least 2;

35

D2) The percentage of first state reads in second state reads is not

over a threshold  $X_4$ , to account for possible contamination of control state reads with test state reads, wherein  $X_4$  is at least 5%.

- E) Clustering and filtering test and control reads derived from all candidate breakpoints accepted in step D to build blocks, by carrying out the steps:
- 5 E1) Retrieving reads which contain the stem of at least one k-mer that represents the candidate breakpoint selected in step D);
- E2) The reads of step E1) with at least  $X_5$  k-mer variants within a window of  $X_6$  nucleotides are taken as leading reads, wherein  $X_5$  is at least 7  
10 and  $X_6$  is at least 10;
- E3) Reads whose k-mers share at least one stem with a leading read are merged to give a block; and
- E4) If the nucleic acid whose variant is being identified is a double stranded DNA, then both forward and reverse variants are taken into account  
15 when building the block.
- F) Aligning blocks taking their leading reads as a reference:
- F1) For each read in the block, take the leading read's stem and find the longest inflection or partial inflection between the read and the leading read.  
20 F2) Successively position each read so that its matching inflection or partial inflection is aligned against the leading read.
2. The computer-implemented method according to claim 1, further comprising the step following F2:
- 25 F3) Obtaining first state scores, second state scores, and global similarity scores of each position in the block by measuring a ratio of most frequent nucleotide in that position relative to the total number of nucleotides.
3. The computer-implemented method according to any one of claims 1-2,  
30 further comprising the step of:
- G) Cataloguing and annotating blocks according to the following:
- G1) If blocks between the first and second states only differ in one substituted nucleotide, the variant is catalogued as containing a single  
35 nucleotide variant and the single nucleotide variant is annotated;
- G2) If blocks between the first and second states differ in more than

one nucleotide but the whole difference in sequence is contained within the block, the variant is catalogued as a small structural variant, and the small structural variant is annotated; and

5 G3) If blocks between the first and second states differ in more than one nucleotide and the whole difference in sequence is not contained within the block, the variant is catalogued as a large structural variant, and the boundaries of all large structural variants are extended by retrieving blocks overlapping at least X2 nucleotides in an iterative process which ends when the extended sequence reaches 200 nucleotides or when an ambiguous path  
10 is found.

4. The computer-implemented method according to any one of the claims 1-3, further comprising the step of:

H) Filtering of the blocks, according to the following:

15 H1) The percentage of second state reads in first state reads is not over a threshold X7, to account for possible contamination of test state reads with control state reads, wherein X7 is at least 20%;

20 5. The computer-implemented method according to any of the claims 1 to 4 further comprising optionally mapping second state blocks, and subsequently mapping first state blocks, on a reference genome.

6. The computer-implemented method according to any one of claims 1 to 5, wherein X2 is from 25 to 40.

25 7. The computer-implemented method according to any one of claims 1 to 6, wherein X3 is equal to or above 3.

30 8. The computer-implemented method according to any one of claims 1 to 7, wherein threshold X4 is 5%.

9. The computer-implemented method according to any one of claims 1 to 8, wherein the threshold X5 is from 10 to 15 and the threshold X6 is from 12 to  
35 20.

10. The computer-implemented method according to any one of claims 1 to 9,

wherein the first set of reads corresponds to pathological cells of a patient and the second set of reads corresponds to non-pathological cells of the same patient.

- 5 11. A computer program product comprising program instructions for causing a computer system to perform the method for identifying nucleic acid variants between two genomic states as defined in any of claims 1 to 10.
12. The computer program product according to claim 11 embodied on a  
10 storage medium.
13. The computer program product according to claim 11 carried on a carrier signal.
- 15 14. A system for identifying nucleic acid variants between two genomic states comprising the steps of:
- A) Computer/Electronic means for inputting 2 sets of nucleic acid reads, which are sequences retrieved from a nucleotide sequencing method,  
20 wherein the first set of reads corresponds to cells representing a first test state, and the second set of reads corresponds to cells representing a second control state;
- B) Computer/Electronic means for filtering the reads, wherein the filtering  
25 comprises:
- B1) Keeping only the reads with at least a percentage X1 of their bases with a Phred quality score higher than 20, being X1 equal to or above 90%;
- B2) Splitting the reads with an undefined nucleotide, giving one  
30 sequence before, and one sequence after the undefined nucleotide, the latter being discarded; and
- B3) Discarding the sequence reads with less than X2 bases, wherein X2 is from 25 to 50;
- 35 C) Computer/Electronic means for generating a hashtable structure comprising:

C1) Generating a number of  $N-X2+1$  new reads for each read of sequence length  $N$ , wherein the new  $N-X2+1$  reads correspond to all k-mers with length  $X2$  nucleotides; and

5 C2) Building a hashtable structure, which comprises all the k-mers generated in step C1) and further comprises the number of times each k-mer is observed in the two sets of reads corresponding to first and second states.

D) Computer/Electronic means for detecting variants in the sequence between first state and second state, wherein a k-mer of the hashtable  
10 structure is taken as a candidate breakpoint, which represents a variant between the first and second states, if it fulfills all the following requirements:

D1) At least one inflection based on a k-mer's stem must have at least  $X3$  reads with the same variation between first and second states, being  $X3$  at least 2;

15 D2) The percentage of first state reads in second state reads is not over a threshold  $X4$ , to account for possible contamination of control state reads with test state reads, wherein  $X4$  is at least 5%

E) Computer/Electronic means for clustering and filtering test and control  
20 reads derived from all candidate breakpoints accepted in step D to build blocks, by carrying out the steps:

E1) Retrieving reads which contain the stem of at least one k-mer that represents the candidate breakpoint selected in step D);

25 E2) The reads of step E1) with at least  $X5$  k-mer variants within a window of  $X6$  nucleotides are taken as leading reads, wherein  $X5$  is at least 7 and  $X6$  is at least 10;

E3) Reads whose k-mers share at least one stem with a leading read are merged to give a block; and

30 E4) If the nucleic acid whose variant is being identified is a double stranded DNA, then both forward and reverse variants are taken into account when building the block.

F) Computer/Electronic means for aligning blocks taking their leading reads as a reference:

35 F1) For each read in the block, take the leading read's stem and find the longest inflection or partial inflection between the read and the leading read.

F2) Successively position each read so that its matching inflection or partial inflection is aligned against the leading read.

5 15. A computer system comprising a processor and a memory, wherein the memory stores computer executable instructions that, when executed by the processor, cause the system to perform the method for identifying nucleic acid variants between two genomic states as defined by any of claims 1 to 10.

10

15

20

25

30

35

**ABSTRACT**

A computer-implemented and reference-free method for identifying variants in nucleic acid sequences

There is provided a computer-implemented method for identifying of nucleic acid variants between two cells, such as a normal cell vs. a pathological cell of a patient, or a cell at two different stages of development. The method is alignment-free, as it does not depend on the use of a reference genome, and is based on the generation and comparison of polymorphic k-mers derived from the nucleotide sequence reads of both biological states. The invention accurately identifies all sorts of genetic variants, ranging from single nucleotide substitutions (SNVs) to large structural variants with great sensitivity and specificity. As a major novelty, it also identifies non-human insertions, such as those derived from retroviruses. Altogether, this invention allows the integration with specific hardware architectures in order to speed up the executions to an unprecedented level.

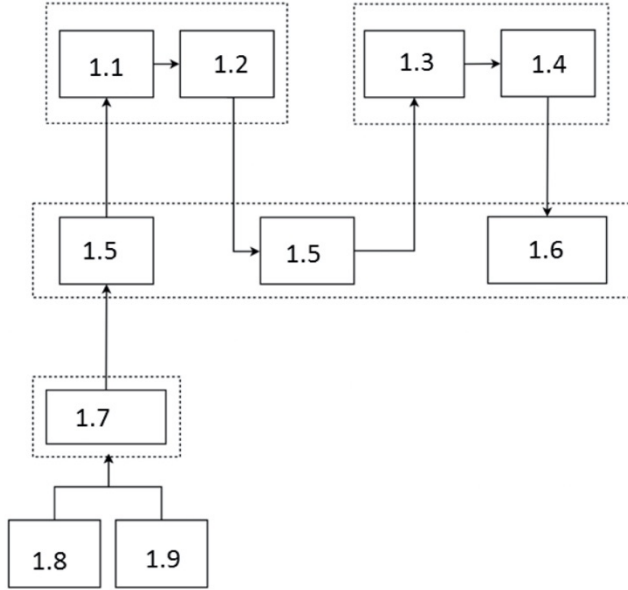


FIG. 1



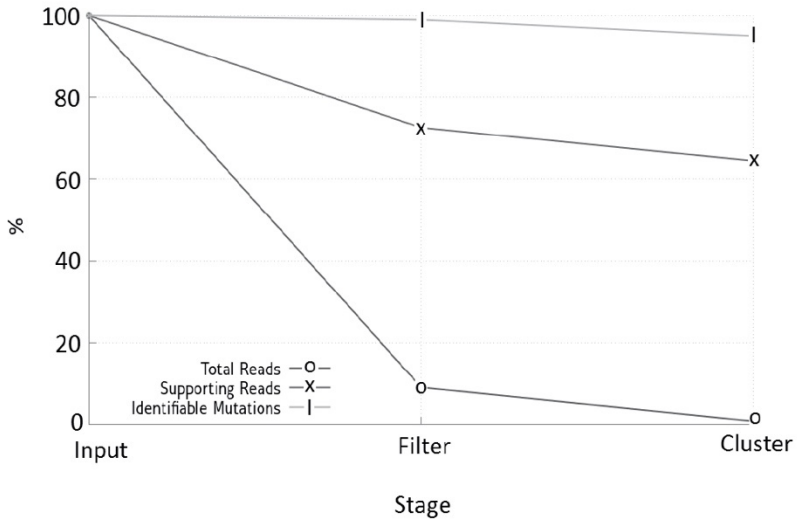


FIG.2

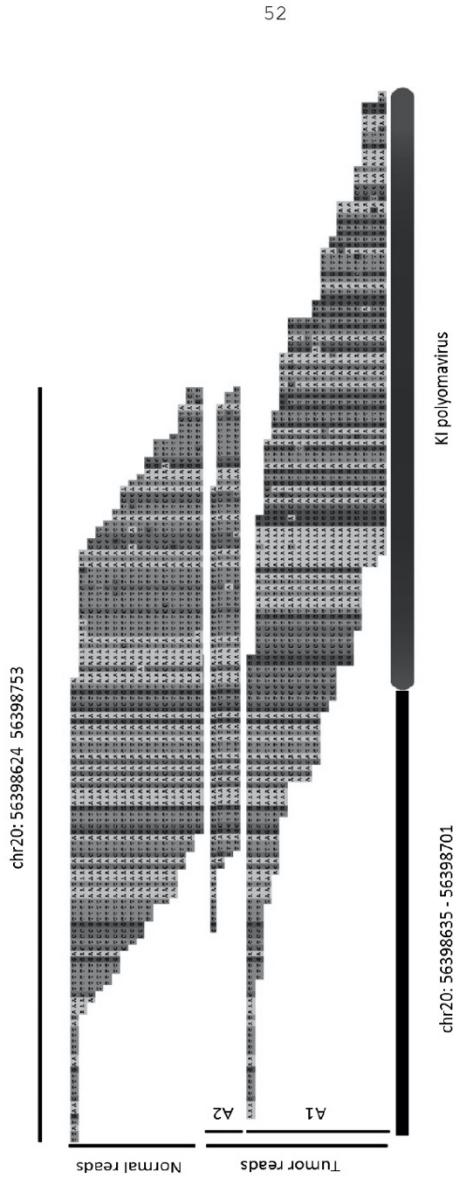


FIG. 3

# REFERENCES

- Alekseyev, M.A., and Pevzner, P.A. (2007). Colored de Bruijn graphs and the genome halving problem. *IEEE/ACM Trans Comput Biol Bioinform* *4*, 98-107.
- Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Borresen-Dale, A.L., *et al.* (2013). Signatures of mutational processes in human cancer. *Nature* *500*, 415-421.
- Alexandrov, L.B., and Stratton, M.R. (2014). Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Curr Opin Genet Dev* *24*, 52-60.
- Alkan, C., Coe, B.P., and Eichler, E.E. (2011). Genome structural variation discovery and genotyping. *Nat Rev Genet* *12*, 363-376.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. *J Mol Biol* *215*, 403-410.
- Bae, T., Tomasini, L., Mariani, J., Zhou, B., Roychowdhury, T., Franjic, D., Pletikos, M., Pattni, R., Chen, B.J., Venturini, E., *et al.* (2018). Different mutational rates and mechanisms in human cells at pregastrulation and neurogenesis. *Science* *359*, 550-555.
- Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* *37*, W202-208.
- Bailey, T.L., and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* *2*, 28-36.
- Baker, M. (2012). Structural variation: the genome's hidden architecture. *Nat Methods* *9*, 133-137.
- Bea, S., Valdes-Mas, R., Navarro, A., Salaverria, I., Martin-Garcia, D., Jares, P., Gine, E., Pinyol, M., Royo, C., Nadeu, F., *et al.* (2013). Landscape of somatic mutations and clonal evolution in mantle cell lymphoma. *Proc Natl Acad Sci U S A* *110*, 18250-18255.
- Bialecki, E.S., and Di Bisceglie, A.M. (2005). Clinical presentation and natural course of hepatocellular carcinoma. *Eur J Gastroenterol Hepatol* *17*, 485-489.
- Blumcke, I., Thom, M., Aronica, E., Armstrong, D.D., Vinters, H.V., Palmmini, A., Jacques, T.S., Avanzini, G., Barkovich, A.J., Battaglia, G., *et al.* (2011). The clinicopathologic spectrum of focal cortical dysplasias: a consensus classification proposed by an ad hoc Task Force of the ILAE Diagnostic Methods Commission. *Epilepsia* *52*, 158-174.
- Borozan, I., Wilson, S., Blanchette, P., Laflamme, P., Watt, S.N., Krzyzanowski, P.M., Sircoulomb, F., Rottapel, R., Branton, P.E., and Ferretti, V. (2012). CaPSID: a

bioinformatics platform for computational pathogen sequence identification in human genomes and transcriptomes. *BMC Bioinformatics* *13*, 206.

Boveri, T. (2008). Concerning the origin of malignant tumours by Theodor Boveri. Translated and annotated by Henry Harris. *J Cell Sci* *121 Suppl 1*, 1-84.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* *68*, 394-424.

Cameron, D.L., Di Stefano, L., and Papenfuss, A.T. (2019). Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. *Nat Commun* *10*, 3240.

Chen, H., Liu, H., and Qing, G. (2018). Targeting oncogenic Myc as a strategy for cancer treatment. *Signal Transduct Target Ther* *3*, 5.

Chen, K., Chen, L., Fan, X., Wallis, J., Ding, L., and Weinstock, G. (2014). TIGRA: a targeted iterative graph routing assembler for breakpoint assembly. *Genome Res* *24*, 310-317.

Chen, K., Wallis, J.W., McLellan, M.D., Larson, D.E., Kalicki, J.M., Pohl, C.S., McGrath, S.D., Wendl, M.C., Zhang, Q., Locke, D.P., *et al.* (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods* *6*, 677-681.

Chen, Y., Yao, H., Thompson, E.J., Tannir, N.M., Weinstein, J.N., and Su, X. (2013). VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue. *Bioinformatics* *29*, 266-267.

Chin, L., Andersen, J.N., and Futreal, P.A. (2011). Cancer genomics: from discovery science to personalized medicine. *Nat Med* *17*, 297-303.

Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* *31*, 213-219.

Cirino, A.L., Lakdawala, N.K., McDonough, B., Conner, L., Adler, D., Weinfeld, M., O'Gara, P., Rehm, H.L., Machini, K., Lebo, M., *et al.* (2017). A Comparison of Whole Genome Sequencing to Multigene Panel Testing in Hypertrophic Cardiomyopathy Patients. *Circ Cardiovasc Genet* *10*.

Cock, P.J., Fields, C.J., Goto, N., Heuer, M.L., and Rice, P.M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res* *38*, 1767-1771.

Comfort, N.C. (2001). From controlling elements to transposons: Barbara McClintock and the Nobel Prize. *Endeavour* *25*, 127-130.

Consortium, E.P. (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* *306*, 636-640.

Consortium, I.T.P.-C.A.o.W.G. (2020). Pan-cancer analysis of whole genomes. *Nature* *578*, 82-93.

Consortium, U.K., Walter, K., Min, J.L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J.R., Xu, C., Futema, M., *et al.* (2015). The UK10K project identifies rare variants in health and disease. *Nature* *526*, 82-90.

Cortes-Ciriano, I., Lee, J.J., Xi, R., Jain, D., Jung, Y.L., Yang, L., Gordenin, D., Klimczak, L.J., Zhang, C.Z., Pellman, D.S., *et al.* (2020). Comprehensive analysis of

chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nat Genet* *52*, 331-341.

D’Gama, A.M., and Walsh, C.A. (2018). Somatic mosaicism and neurodevelopmental disease. *Nat Neurosci* *21*, 1504-1514.

Datta, S., Bettinger, K., and Snyder, M. (2016). Corrigendum: Secure cloud computing for genomic data. *Nat Biotechnol* *34*, 1072.

de Koning, A.P., Gu, W., Castoe, T.A., Batzer, M.A., and Pollock, D.D. (2011). Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet* *7*, e1002384.

Degner, J.F., Marioni, J.C., Pai, A.A., Pickrell, J.K., Nkadori, E., Gilad, Y., and Pritchard, J.K. (2009). Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* *25*, 3207-3212.

Dou, Y., Gold, H.D., Luquette, L.J., and Park, P.J. (2018). Detecting Somatic Mutations in Normal Cells. *Trends Genet* *34*, 545-557.

Duncavage, E.J., Magrini, V., Becker, N., Armstrong, J.R., Demeter, R.T., Wylie, T., Abel, H.J., and Pfeifer, J.D. (2011). Hybrid capture and next-generation sequencing identify viral integration sites from formalin-fixed, paraffin-embedded tissue. *J Mol Diagn* *13*, 325-333.

Escaramis, G., Docampo, E., and Rabionet, R. (2015). A decade of structural variants: description, history and methods to detect structural variation. *Brief Funct Genomics* *14*, 305-314.

Ewing, B., and Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* *8*, 186-194.

Ewing, B., Hillier, L., Wendl, M.C., and Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* *8*, 175-185.

Farrell, P.J. (2019). Epstein-Barr Virus and Cancer. *Annu Rev Pathol* *14*, 29-53.

Fernandez, L., Mercader, J.M., Planas-Felix, M., and Torrents, D. (2014). Adaptation to environmental factors shapes the organization of regulatory regions in microbial communities. *BMC Genomics* *15*, 877.

Forster, M., Szymczak, S., Ellinghaus, D., Hemmrich, G., Ruhlemann, M., Kraemer, L., Mucha, S., Wienbrandt, L., Stanulla, M., Group, U.F.O.S.C.w.I.-B.S., *et al.* (2015). Vy-PER: eliminating false positive detection of virus integration events in next generation sequencing data. *Sci Rep* *5*, 11534.

Gao, S., Hu, X., Xu, F., Gao, C., Xiong, K., Zhao, X., Chen, H., Zhao, S., Wang, M., Fu, D., *et al.* (2018). BS-virus-finder: virus integration calling using bisulfite sequencing data. *Gigascience* *7*, 1-7.

Geisler, J., Touma, J., Rahbar, A., Soderberg-Naucler, C., and Vetvik, K. (2019). A Review of the Potential Role of Human Cytomegalovirus (HCMV) Infections in Breast Cancer Carcinogenesis and Abnormal Immunity. *Cancers (Basel)* *11*.

Genomes Project, C., Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., *et al.* (2015). A global reference for human genetic variation. *Nature* *526*, 68-74.

Gerstung, M., Jolly, C., Leshchiner, I., Dentre, S.C., Gonzalez, S., Rosebrock, D., Mitchell, T.J., Rubanova, Y., Anur, P., Yu, K., *et al.* (2020). The evolutionary history of 2,658 cancers. *Nature* *578*, 122-128.

Goldfeder, R.L., Priest, J.R., Zook, J.M., Grove, M.E., Waggott, D., Wheeler, M.T., Salit, M., and Ashley, E.A. (2016). Medical implications of technical accuracy in genome sequencing. *Genome Med* 8, 24.

Gollob, M.H., Jones, D.L., Krahn, A.D., Danis, L., Gong, X.Q., Shao, Q., Liu, X., Veinot, J.P., Tang, A.S., Stewart, A.F., *et al.* (2006). Somatic mutations in the connexin 40 gene (GJA5) in atrial fibrillation. *N Engl J Med* 354, 2677-2688.

Guan, P., and Sung, W.K. (2016). Structural variation detection using next-generation sequencing data: A comparative technical review. *Methods* 102, 36-49.

Henssen, A.G., Koche, R., Zhuang, J., Jiang, E., Reed, C., Eisenberg, A., Still, E., MacArthur, I.C., Rodriguez-Fos, E., Gonzalez, S., *et al.* (2017a). Erratum: PGBD5 promotes site-specific oncogenic mutations in human tumors. *Nat Genet* 49, 1558.

Henssen, A.G., Koche, R., Zhuang, J., Jiang, E., Reed, C., Eisenberg, A., Still, E., MacArthur, I.C., Rodriguez-Fos, E., Gonzalez, S., *et al.* (2017b). PGBD5 promotes site-specific oncogenic mutations in human tumors. *Nat Genet* 49, 1005-1014.

Hermine, O., Lefrere, F., Bronowicki, J.P., Mariette, X., Jondeau, K., Eclache-Saudreau, V., Delmas, B., Valensi, F., Cacoub, P., Brechot, C., *et al.* (2002). Regression of splenic lymphoma with villous lymphocytes after treatment of hepatitis C virus infection. *N Engl J Med* 347, 89-94.

Herrington, C.S., Coates, P.J., and Duprex, W.P. (2015). Viruses and disease: emerging concepts for prevention, diagnosis and treatment. *J Pathol* 235, 149-152.

Ho, D.W., Sze, K.M., and Ng, I.O. (2015). Virus-Clip: a fast and memory-efficient viral integration site detection tool at single-base resolution with annotation capability. *Oncotarget* 6, 20959-20963.

Hoadley, K.A., Yau, C., Hinoue, T., Wolf, D.M., Lazar, A.J., Drill, E., Shen, R., Taylor, A.M., Cherniack, A.D., Thorsson, V., *et al.* (2018). Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* 173, 291-304 e296.

Hormozdiari, F., Hajirasouliha, I., Dao, P., Hach, F., Yorukoglu, D., Alkan, C., Eichler, E.E., and Sahinalp, S.C. (2010). Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics* 26, i350-357.

Huang da, W., Sherman, B.T., and Lempicki, R.A. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 37, 1-13.

Huang da, W., Sherman, B.T., and Lempicki, R.A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 4, 44-57.

Hurles, M.E., Dermitzakis, E.T., and Tyler-Smith, C. (2008). The functional impact of structural variation in humans. *Trends Genet* 24, 238-245.

Jamuar, S.S., Lam, A.T., Kircher, M., D'Gama, A.M., Wang, J., Barry, B.J., Zhang, X., Hill, R.S., Partlow, J.N., Rozzo, A., *et al.* (2014). Somatic mutations in cerebral cortical malformations. *N Engl J Med* 371, 733-743.

Javier, R.T., and Butel, J.S. (2008). The history of tumor virology. *Cancer Res* 68, 7693-7706.

- Jiang, Y., Wang, Y., and Brudno, M. (2012). PRISM: pair-read informed split-read mapping for base-pair level detection of insertion, deletion and structural variants. *Bioinformatics* 28, 2576-2583.
- Kehrer-Sawatzki, H., Kluwe, L., Sandig, C., Kohn, M., Wimmer, K., Krammer, U., Peyrl, A., Jenne, D.E., Hansmann, I., and Mautner, V.F. (2004). High frequency of mosaicism among patients with neurofibromatosis type 1 (NF1) with microdeletions caused by somatic recombination of the JJAZ1 gene. *Am J Hum Genet* 75, 410-423.
- Kosugi, S., Momozawa, Y., Liu, X., Terao, C., Kubo, M., and Kamatani, Y. (2019). Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol* 20, 117.
- Koutsoumpa, M., Chen, H.W., O'Brien, N., Koinis, F., Mahurkar-Joshi, S., Vorvis, C., Soroosh, A., Luo, T., Issakhanian, S., Pantuck, A.J., et al. (2018). MKAD-21 Suppresses the Oncogenic Activity of the miR-21/PPP2R2A/ERK Molecular Network in Bladder Cancer. *Mol Cancer Ther* 17, 1430-1440.
- Krol, R.P., Nozu, K., Nakanishi, K., Iijima, K., Takeshima, Y., Fu, X.J., Nozu, Y., Kaito, H., Kanda, K., Matsuo, M., et al. (2008). Somatic mosaicism for a mutation of the COL4A5 gene is a cause of mild phenotype male Alport syndrome. *Nephrol Dial Transplant* 23, 2525-2530.
- Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. *Genome Res* 19, 1639-1645.
- Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499, 214-218.
- Lee, J.H., Huynh, M., Silhavy, J.L., Kim, S., Dixon-Salazar, T., Heiberg, A., Scott, E., Bafna, V., Hill, K.J., Collazo, A., et al. (2012). De novo somatic mutations in components of the PI3K-AKT3-mTOR pathway cause hemimegalencephaly. *Nat Genet* 44, 941-945.
- Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754-1760.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078-2079.
- Li, J.W., Wan, R., Yu, C.S., Co, N.N., Wong, N., and Chan, T.F. (2013). ViralFusionSeq: accurately discover viral integration events and reconstruct fusion transcripts at single-base resolution. *Bioinformatics* 29, 649-651.
- Li, Y., Roberts, N.D., Wala, J.A., Shapira, O., Schumacher, S.E., Kumar, K., Khurana, E., Waszak, S., Korbel, J.O., Haber, J.E., et al. (2020). Patterns of somatic structural variation in human cancer genomes. *Nature* 578, 112-121.
- Liao, J.B. (2006). Viruses and human cancer. *Yale J Biol Med* 79, 115-122.
- Lievre, A., Bachet, J.B., Boige, V., Cayre, A., Le Corre, D., Buc, E., Ychou, M., Bouche, O., Landi, B., Louvet, C., et al. (2008). KRAS mutations as an independent prognostic factor in patients with advanced colorectal cancer treated with cetuximab. *J Clin Oncol* 26, 374-379.

Lim, J.S., Kim, W.I., Kang, H.C., Kim, S.H., Park, A.H., Park, E.K., Cho, Y.W., Kim, S., Kim, H.M., Kim, J.A., *et al.* (2015). Brain somatic mutations in MTOR cause focal cortical dysplasia type II leading to intractable epilepsy. *Nat Med* *21*, 395-400.

Lodato, M.A., Rodin, R.E., Bohrsen, C.L., Coulter, M.E., Barton, A.R., Kwon, M., Sherman, M.A., Vitzthum, C.M., Luquette, L.J., Yandava, C.N., *et al.* (2018). Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science* *359*, 555-559.

Lupski, J.R. (2007). Structural variation in the human genome. *N Engl J Med* *356*, 1169-1171.

Lupski, J.R., Gonzaga-Jauregui, C., Yang, Y., Bainbridge, M.N., Jhangiani, S., Buhay, C.J., Kovar, C.L., Wang, M., Hawes, A.C., Reid, J.G., *et al.* (2013). Exome sequencing resolves apparent incidental findings and reveals further complexity of SH3TC2 variant alleles causing Charcot-Marie-Tooth neuropathy. *Genome Med* *5*, 57.

Marco-Sola, S., Sammeth, M., Guigo, R., and Ribeca, P. (2012). The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat Methods* *9*, 1185-1188.

Mardis, E.R. (2008). The impact of next-generation sequencing technology on genetics. *Trends Genet* *24*, 133-141.

Martincorena, I., and Campbell, P.J. (2015). Somatic mutation in cancer and normal cells. *Science* *349*, 1483-1489.

Martincorena, I., Raine, K.M., Gerstung, M., Dawson, K.J., Haase, K., Van Loo, P., Davies, H., Stratton, M.R., and Campbell, P.J. (2018). Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* *173*, 1823.

Marx, V. (2013). Biology: The big challenges of big data. *Nature* *498*, 255-260.

Mattmann, C.A. (2013). Computing: A vision for data science. *Nature* *493*, 473-475.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernysky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., *et al.* (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* *20*, 1297-1303.

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biol* *17*, 122.

Michikawa, Y., Mazzucchelli, F., Bresolin, N., Scarlato, G., and Attardi, G. (1999). Aging-dependent large accumulation of point mutations in the human mtDNA control region for replication. *Science* *286*, 774-779.

Milholland, B., Dong, X., Zhang, L., Hao, X., Suh, Y., and Vijg, J. (2017). Differences between germline and somatic mutation rates in humans and mice. *Nat Commun* *8*, 15183.

Moncunill, V., Gonzalez, S., Bea, S., Andrieux, L.O., Salaverria, I., Royo, C., Martinez, L., Puiggros, M., Segura-Wang, M., Stutz, A.M., *et al.* (2014). Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads. *Nat Biotechnol* *32*, 1106-1112.

Moore, P.S., and Chang, Y. (2010). Why do viruses cause cancer? Highlights of the first century of human tumour virology. *Nat Rev Cancer* *10*, 878-889.

Munoz, N., Castellsague, X., de Gonzalez, A.B., and Gissmann, L. (2006). Chapter 1: HPV in the etiology of human cancer. *Vaccine* *24 Suppl 3*, S3/1-10.



- Muotri, A.R., Marchetto, M.C., Coufal, N.G., Oefner, R., Yeo, G., Nakashima, K., and Gage, F.H. (2010). L1 retrotransposition in neurons is modulated by MeCP2. *Nature* *468*, 443-446.
- Nguyen, N.D., Deshpande, V., Luebeck, J., Mischel, P.S., and Bafna, V. (2018). ViFi: accurate detection of viral integration and mRNA fusion reveals indiscriminate and unregulated transcription in proximal genomic regions in cervical cancer. *Nucleic Acids Res* *46*, 3309-3325.
- Parkin, D.M. (2006). The global health burden of infection-associated cancers in the year 2002. *Int J Cancer* *118*, 3030-3044.
- Paszkievicz, K., and Studholme, D.J. (2010). De novo assembly of short sequence reads. *Brief Bioinform* *11*, 457-472.
- Plummer, M., de Martel, C., Vignat, J., Ferlay, J., Bray, F., and Franceschi, S. (2016). Global burden of cancers attributable to infections in 2012: a synthetic analysis. *Lancet Glob Health* *4*, e609-616.
- Poduri, A., Evrony, G.D., Cai, X., Elhosary, P.C., Beroukhim, R., Lehtinen, M.K., Hills, L.B., Heinzen, E.L., Hill, A., Hill, R.S., *et al.* (2012). Somatic activation of AKT3 causes hemispheric developmental brain malformations. *Neuron* *74*, 41-48.
- Poduri, A., Evrony, G.D., Cai, X., and Walsh, C.A. (2013). Somatic mutation, genomic variation, and neurological disease. *Science* *341*, 1237758.
- Quinlan, A.R. (2014). BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr Protoc Bioinformatics* *47*, 11 12 11-34.
- Quinlan, A.R., and Hall, I.M. (2012). Characterizing complex structural variation in germline and somatic genomes. *Trends Genet* *28*, 43-53.
- Raine, K.M., Hinton, J., Butler, A.P., Teague, J.W., Davies, H., Tarpey, P., Nik-Zainal, S., and Campbell, P.J. (2015). cgpPindel: Identifying Somatically Acquired Insertion and Deletion Events from Paired End Sequencing. *Curr Protoc Bioinformatics* *52*, 15 17 11-15 17 12.
- Rausch, T., Hsi-Yang Fritz, M., Korbel, J.O., and Benes, V. (2019). Alfred: interactive multi-sample BAM alignment statistics, feature counting and feature annotation for long- and short-read sequencing. *Bioinformatics* *35*, 2489-2491.
- Rausch, T., Zichner, T., Schlattl, A., Stutz, A.M., Benes, V., and Korbel, J.O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* *28*, i333-i339.
- Reddy, E.P., Reynolds, R.K., Santos, E., and Barbacid, M. (1982). A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene. *Nature* *300*, 149-152.
- Riviere, J.B., Mirzaa, G.M., O'Roak, B.J., Beddaoui, M., Alcantara, D., Conway, R.L., St-Onge, J., Schwartzentruber, J.A., Gripp, K.W., Nikkel, S.M., *et al.* (2012). De novo germline and postzygotic mutations in AKT3, PIK3R2 and PIK3CA cause a spectrum of related megalencephaly syndromes. *Nat Genet* *44*, 934-940.
- Roy, R., Chun, J., and Powell, S.N. (2011). BRCA1 and BRCA2: different roles in a common pathway of genome protection. *Nat Rev Cancer* *12*, 68-78.
- Sarid, R., and Gao, S.J. (2011). Viruses and human cancer: from detection to causality. *Cancer Lett* *305*, 218-227.
- Schroder, J., Hsu, A., Boyle, S.E., Macintyre, G., Cmero, M., Tothill, R.W., Johnstone, R.W., Shackleton, M., and Papenfuss, A.T. (2014). Socrates: identification of

genomic rearrangements in tumour genomes by re-aligning soft clipped reads. *Bioinformatics* *30*, 1064-1072.

Stambolic, V., Suzuki, A., de la Pompa, J.L., Brothers, G.M., Mirtsos, C., Sasaki, T., Ruland, J., Penninger, J.M., Siderovski, D.P., and Mak, T.W. (1998). Negative regulation of PKB/Akt-dependent cell survival by the tumor suppressor PTEN. *Cell* *95*, 29-39.

Stratton, M.R. (2011). Exploring the genomes of cancer cells: progress and promise. *Science* *331*, 1553-1558.

Stratton, M.R., Campbell, P.J., and Futreal, P.A. (2009). The cancer genome. *Nature* *458*, 719-724.

Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M.H., *et al.* (2015). An integrated map of structural variation in 2,504 human genomes. *Nature* *526*, 75-81.

Suzuki, S., Yasuda, T., Shiraishi, Y., Miyano, S., and Nagasaki, M. (2011). ClipCrop: a tool for detecting structural variations with single-base resolution using soft-clipping information. *BMC Bioinformatics* *12 Suppl 14*, S7.

Szilard, L. (1959). On the Nature of the Aging Process. *Proc Natl Acad Sci U S A* *45*, 30-45.

Talbot, S.J., and Crawford, D.H. (2004). Viruses and tumours--an update. *Eur J Cancer* *40*, 1998-2005.

Tattini, L., D'Aurizio, R., and Magi, A. (2015). Detection of Genomic Structural Variants from Next-Generation Sequencing Data. *Front Bioeng Biotechnol* *3*, 92.

Trappe, K., Emde, A.K., Ehrlich, H.C., and Reinert, K. (2014). Gustaf: Detecting and correctly classifying SVs in the NGS twilight zone. *Bioinformatics* *30*, 3484-3490.

van Dijk, E.L., Auger, H., Jaszczyszyn, Y., and Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends Genet* *30*, 418-426.

Varley, J.M., Evans, D.G., and Birch, J.M. (1997). Li-Fraumeni syndrome--a molecular and clinical review. *Br J Cancer* *76*, 1-14.

Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., Jr., and Kinzler, K.W. (2013). Cancer genome landscapes. *Science* *339*, 1546-1558.

Wala, J.A., Bandopadhyay, P., Greenwald, N.F., O'Rourke, R., Sharpe, T., Stewart, C., Schumacher, S., Li, Y., Weischenfeldt, J., Yao, X., *et al.* (2018). SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Res* *28*, 581-591.

Wang, J., Mullighan, C.G., Easton, J., Roberts, S., Heatley, S.L., Ma, J., Rusch, M.C., Chen, K., Harris, C.C., Ding, L., *et al.* (2011). CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nat Methods* *8*, 652-654.

Wang, Q., Jia, P., and Zhao, Z. (2015). VERSE: a novel approach to detect virus integration in host genomes through reference genome customization. *Genome Med* *7*, 2.

Weischenfeldt, J., Symmons, O., Spitz, F., and Korbel, J.O. (2013). Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat Rev Genet* *14*, 125-138.

Workman, A.D., Charvet, C.J., Clancy, B., Darlington, R.B., and Finlay, B.L. (2013). Modeling transformations of neurodevelopmental sequences across mammalian species. *J Neurosci* *33*, 7368-7383.

- Wu, H., Esteve, E., Tremaroli, V., Khan, M.T., Caesar, R., Manneras-Holm, L., Stahlman, M., Olsson, L.M., Serino, M., Planas-Felix, M., *et al.* (2017). Metformin alters the gut microbiome of individuals with treatment-naive type 2 diabetes, contributing to the therapeutic effects of the drug. *Nat Med* 23, 850-858.
- Ye, K., Schulz, M.H., Long, Q., Apweiler, R., and Ning, Z. (2009). 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.
- Yi, K., and Ju, Y.S. (2018). Patterns and mechanisms of structural variations in human cancer. *Exp Mol Med* 50, 98.
- Yung, C.K., O'Connor, B.D., Yakneen, S., Zhang, J., Ellrott, K., Kleinheinz, K., Miyoshi, N., Raine, K.M., Royo, R., Saksena, G.B., *et al.* (2017). Large-Scale Uniform Analysis of Cancer Whole Genomes in Multiple Computing Environments. *bioRxiv*, 161638.
- Zapatka, M., Borozan, I., Brewer, D.S., Iskar, M., Grundhoff, A., Alawi, M., Desai, N., Sultmann, H., Moch, H., Pathogens, P., *et al.* (2020). The landscape of viral associations in human cancers. *Nat Genet* 52, 320-330.
- Zhang, Y., Yang, L., Kucherlapati, M., Chen, F., Hadjipanayis, A., Pantazi, A., Bristow, C.A., Lee, E.A., Mahadeshwar, H.S., Tang, J., *et al.* (2018). A Pan-Cancer Compendium of Genes Deregulated by Somatic Genomic Rearrangement across More Than 1,400 Cases. *Cell Rep* 24, 515-527.
- Zhuang, J., and Weng, Z. (2015). Local sequence assembly reveals a high-resolution profile of somatic structural variations in 97 cancer genomes. *Nucleic Acids Res* 43, 8146-8156.
- Zia, Z., Thurley, P.D., Pollock, J.G., DeNunzio, M., Bungay, P., and Whitaker, S.C. (2012). The diagnosis and endovascular management of superior mesenteric artery (SMA) branch pseudoaneurysms after appendectomy. *Vasc Endovascular Surg* 46, 54-57.
- Cadenelli, N., Polo, J., and Carrera, D (2017) Accelerating K-mer frequency counting with GPU and non-volatile memory. *IEEE 19th International Conference on High Performance Computing and Communications; IEEE 15th International Conference on Smart City; IEEE 3rd International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*
- Cadenelli, N., Jaksić, Z., Polo J., and Carrera D (2019) Considerations in using OpenCL on GPUs and FPGAs for throughput-oriented genomics workloads. *Future Generation Computer Systems*



