PhD degree in Molecular Medicine (curriculum in Computational Biology)

European School of Molecular Medicine (SEMM),

University of Milan and University of Naples "Federico II"

Settore Disciplinare: MED/04

Computational frameworks for the identification of somatic and germline variants contributing to cancer predisposition and development

Giorgio Enrico Maria Melloni

IIT@SEMM, Milan

Matricola n. R10338

Supervisor: Prof. Piergiuseppe Pelicci

IEO, Milan

Added Supervisor: Dr. Laura Riva

IIT@SEMM, Milan

Anno accademico 2015-2016

TABLE OF CONTENTS

1 ABSTRACT 5

2 INTRODUCTION 6

2.1 CANCER AS AN EVOLUTIONARY PROCESS 6
2.2 ACCUMULATING DRIVER MUTATIONS 8
2.3 TUMOR HETEROGENEITY 10
2.4 CANCER GENOME LANDSCAPES 11
2.5 DRIVER VS PASSENGER: A PROBLEM OF MUTATION RATE 13
2.6 CANCER GENOMICS AND HUMAN GENETICS 14
2.7 CANCER GENOMICS IN THE NGS ERA 16

3 MATERIAL AND METHODS 17

3.1 DATA FORMAT 17
3.1.1 VCF format 17
3.1.2 MAF format 18
3.2 DATA RETRIEVAL 19
3.2.1 TCGA 19
3.2.2 ICGC 19
3.2.3 cBioPortal 20
3.2.4 COSMIC and CGC 20
3.2.5 ExAC 20
3.3 DATA PROCESSING AND MANIPULATION 21

4 RESULTS 22

4.1 DOTS-FINDER: A COMPREHENSIVE TOOL TO ASSESSING DRIVER GENES IN CANCER GENOMES 22
4.1.1 Abstract 22
4.1.2 Introduction 23
4.1.3 Implementation 25
4.1.3.1 Overview of DOTS-Finder 25
4.1.3.2 The Functional Step: finding tumor suppressor gene and oncogene candidates 30
4.1.3.3 The Frequentist step: assessing the possible drivers 35
4.1.4 Material and Methods 36
4.1.4.1 Availability 36
4.1.4.2 Input Format 36
4.1.4.3 Requirements 37
4.1.4.4 Mutation data 37
4.1.4.5 Databases 38
4.1.4.6 DOTS-Finder step by step 42
4.1.4.6.1 Setting the threshold for TSG-S and OG-S 48
4.1.5 Results 50
4.1.5.1 Application of DOTS-Finder to individual cancer types 50
4.1.5.2 Driver genes and tissue specificity 52
4.1.5.3 Breast carcinoma 53
4.1.5.4 Thyroid Carcinoma 55
4.1.5.5 Acute Myeloid Leukemia 56
4.1.5.6 Bladder Carcinoma 57
4.1.5.7 Atypical tumor suppressor genes and oncogenes 58
4.1.5.8 The importance of considering subsets of samples 62
4.1.5.9 Small sample size analysis. The --lax option 63
4.1.5.10 Comparison of DOTS-Finder to existing tools using Pan-Cancer12 data 65
4.1.5.11 Statistical power using a small number of cancer samples 68
4.1.6 Discussion 69
4.2 LOWMACA: EXPLOITING PROTEIN FAMILY ANALYSIS FOR THE IDENTIFICATION OF RARE DRIVER MUTATIONS IN CANCER 71
4.2.1 Abstract 71
4.2.2 Introduction 72
4.2.3 Materials and Methods 74
4.2.3.1 Software Implementation and Overview 74

4.2.3.2 Input Data ... 75
4.2.3.3 Alignment and Mapping ... 76
4.2.3.4 Statistical Testing ... 78
4.2.3.4.1 Testing the randomness of the global mutational profile ... 78
4.2.3.4.2 Testing for the identification of hotspots of mutation ... 79
4.2.3.5 LowMACA Output ... 79
4.2.4 Results ... 80
4.2.4.1 Ras superfamily analysis ... 82
4.2.4.2 Mutual exclusivity analysis ... 86
4.2.4.3 Comparison with functional impact tools ... 88
4.2.4.4 Analysis of driver genes: comparison with available tools ... 91
4.2.4.5 Analysis of silent mutations ... 97
4.2.5 Discussion ... 98
4.3 A KNOWLEDGE-BASED FRAMEWORK FOR THE DISCOVERY OF CANCER PREDISPOSING VARIANTS USING LARGE-SCALE SEQUENCING BREAST CANCER DATA. ... 101
4.3.1 Abstract ... 102
4.3.2 Introduction ... 103
4.3.3 Materials and Methods ... 106
4.3.3.1 Control Data ... 106
4.3.3.2 Case Data ... 106
4.3.3.3 Annotation Data ... 106
4.3.3.4 Data Preprocess ... 109
4.3.3.5 Statistical Analysis ... 110
4.3.3.5.1 Frequency and annotation based analysis ... 110
4.3.3.5.2 Loss-of-function gene-wise testing ... 110
4.3.3.5.3 Age-dependent polygenic model ... 111
4.3.4 Results ... 113
4.3.4.1 Pathogenic and Breast Cancer Related variants ... 113
4.3.4.2 Analysis of rare variants in target cancer genes ... 118
4.3.4.3 Analysis of loss-of-function genes ... 121
4.3.4.4 Polygenic age-dependent model ... 122
4.3.5 Discussion ... 129

5 DISCUSSION ... 131
5.1 DRIVER GENE DISCOVERY ... 133
5.2 ONCOGENES AND DRIVER MUTATIONS DISCOVERY ... 135
5.3 BRIDGING THE GAP BETWEEN GENETICS AND GENOMICS ... 139
5.4 CONCLUSION ... 142

6 REFERENCES ... 142

7 LIST OF ABBREVIATIONS ... 159

8 APPENDIX ... 161

1 Abstract

The most recent cancer classification from NIH includes ~200 types of tumor that originates from several tissue types (http://www.cancer.gov/types). Although macroscopic and microscopic characteristics varies significantly across subtypes, the starting point of every cancer is believed to be a single cell that acquires DNA somatic alterations that increases its fitness over the surrounding cells and makes it behave abnormally and proliferate uncontrollably. Somatic mutations are the consequence of many possible defective processes such as replication deficiencies, exposure to carcinogens, or DNA repair machinery faults. Mutation development is a random and mostly natural process that frequently happens in every cell of an individual. Only the acquisition of a series of subtype-specific alterations, including also larger aberrations such as translocations or deletions, can lead to the development of the disease and this is a long process for the majority of adult tumor types. However, genetic predisposition for certain cancer types is epidemiologically well established. In fact, several cancer predisposing genes where identified in the last 30 years with various technologies but they characterize only a small fraction of familial cases. This work will therefore cover two main steps of cancer genetics and genomics: the identification of the genes that somatically changes the behavior of a normal human cell to a cancer cell and the genetic variants that increase risk of cancer development. The use of publicly available datasets is common to all the three results sections that compose this work. In particular, we took advantage of several whole exome sequencing databases (WES) for the identification of both driver mutations and driver variants. In particular, the use of WES in cancer predisposition analysis represents one of the few attempts of performing such analysis on genome-wide sequencing germline data.

2 Introduction

The purpose of this work is to delineate a workflow to analyze and classify genes that are important for cancer progression both at the germinal level (cancer predisposing genes, CPGs) and somatic level (somatic driver genes, SDGs). We generally refer at variants as those hereditable germline Single Nucleotide Polymorphisms (SNPs) that happen in the normal DNA of the person and they can increase or not the risk of developing cancer. We call mutations, all those Single Nucleotide Variants (SNVs) that can be seen in the tumor but are not part of the original genome of the individual. To be more precise, we refer to the latter as somatic mutations since the term mutation is sometimes used to define pathogenic germline variants, especially in clinical settings (for example when referring to *BRCA1* mutants). In general, mutation, compared to variant, assumes a negative connotation when referring to pathogenic alterations. This introduction represents a brief historical summary of the main milestones in cancer research concerning DNA alterations in cancer. This history runs on a parallel track with the history of cell biology, as a lot of what we know now on how a human cell behave is nothing but a byproduct of cancer research.

2.1 Cancer as an evolutionary process

The somatic evolution of cancer is a theory that states that cancer is the effect of the accumulation of mutations over time from a single aberrant cell of origin that passes the mutations to its next generation. This cell of origin, ultimately evolve in a tumor via mutations that confer a selective growth advantage with respect to its surrounding cells. In 1902, the German biologist Theodor Boveri introduced for the first time the concept of chromosomal aberration as a possible cause for a cell to become malignant, by reviving some observations made by David von Hansemann in 1890 (Boveri, 2008). This

hypothesis came from the observation that in sea urchins, all chromosomes are necessary for a proper embryonic development. Quoting from his seminal work:

' We may therefore regard it as probable that individual chromosomes have different properties in vertebrates too, and it is this assumption that forms the basis of the tumour hypothesis I have put forward. A malignant tumour cell is – and here again I take up the ideas of Hansemann – a cell with a specific abnormal chromosome constitution '.

The idea of tumors as cells with chromosomal defects came as a sort of side note in his work, since the experiments carried out on sea urchin were aimed at demonstrating what is called the Boveri-Sutton hypothesis that chromosomes are responsible for mendelian inheritance. Therefore, it is noteworthy that the biology of cancer was born together with the biology of the cell. Boveri's ideas on oncogenesis, summarized in 1914 *Zur Frage der Entstehung maligner Tumoren* (*On the Origin of Malignant Tumors, Williams & Wilkins. Philadelphia, PA, USA, 1914)* were mostly speculative rather than experimental. Chromosomal inheritance was definitely ruled out by Thomas Morgan a year later and the term *somatic mutation* was introduced by Tyzzer approximately in the same period of time (Tyzzer, 1916). Unfortunately, Boveri died in 1915, without knowing that his seminal ideas were shaping cancer research for the next 100 years. In 1919, Whitman associated anaplasia, the condition of a cell that loses the morphological characteristics of mature cells, with the concept of somatic mutations (Whitman, 1919). He also sets the somatic mutation mechanism as the cause of uncontrolled proliferation and aberrant cell division, as postulated by Boveri himself. 25 years before the discovery of DNA as the molecule of inheritance and 50 years before the first experiments on cell cycle regulation, cancer was already seen as an evolutionary mechanism that starts from a single aberrant cell that proliferates and passes the mutations to the next generation of cells (clonal evolution). In Whitman words:

‘ This cell, the cancer cell, is thus a ‘new kind of cell’. In modern terminology it is, strictly and literally, a mutated cell. Since the process is, or at least may be, repeated itself from time to time, and here and there, in a tumor, it follows that the tumor cells themselves are by no means all alike in their biologic properties; that, on the contrary, an ever recurring process of mutation is taking place, with a tendency, however, to deviate more and more from the normal type. This explains why metastatic tumors, for example, are often more, but never less, malignant than the primary tumor, as well as other related phenomena of tumor growth ’.

2.2 Accumulating driver mutations

In the beginning of the 20th century, scientists referred at *mutations* as chromosomal aberrations, because DNA structure and function was still unknown. In this view, two fundamental concepts of somatic tumor evolution were still missing. First, the idea of accumulation of mutations over time was first observed by two statistical models by Nordling in 1953 and Armitage and Dolls in 1954 (Armitage and Doll, 1954; Nordling, 1953) that for the first time clearly stated that age is the main risk factor for cancer death and that multi-mutations (at least 7 in Nordling model) must occur to develop the disease. Secondly, what kind of mutations must occurs to develop the various subtypes of cancer was mostly unknown until 1971 (Knudson, 1971). Knudson observed that the heritable form of retinoblastoma occurred at a much earlier age than the non-heritable form and he explained this observation by speculating that at least two mutational events were necessary for the development of this cancer. Patients that present with the heritable form of retinoblastoma harbor a germline mutation since conception and require only one DNA mutation in a somatic cell to develop the cancer. In contrast, in the non-hereditary type of retinoblastoma, two DNA mutations need to occur in a somatic cell in order to initiate oncogenesis. It represented the first explanation of the mechanism of mutations in cancer that over 10 years later will be identified as *RB1*, the

first *tumor suppressor gene* (TSG) (Murphree and Benedict, 1984). TSGs are entities with three main characteristics:

- Their normal function is to prevent tumor formation by inhibiting cell cycle and ultimately tumor growth. Generally, they can induce apoptosis, promoting DNA repair or arrest the cell cycle
- The mutations affecting these genes must disrupt the normal function of the protein
- Mutations on tumor suppressors are generally recessive, in the sense that one single healthy copy of the gene is sufficient to maintain the normal behavior of the cell

The second category of genes that promote tumorigenesis is called *oncogene* (OG) and the first evidence of this class of genes and related mutations has been fully understood 10 years after the work of Knudson on retinoblastoma. Oncogenes are cancer genes that when mutated increase or modify their activity within the cell and promote cell growth and survival. In 1979, Bishop and Vamus discovered c-Src in chickens, a gene that when mutated resembles a viral form called v-Src that can be found in Rous Sarcoma Virus (Stehelin et al., 1976). Once the oncogene is transfected back into a chicken, it can lead to cancer. This discovery led to the idea of the viral infection as a tumor-promoting factor and also to the definition of the first *proto-oncogene*. Nevertheless, the first "natural occurring" oncogene can be seen as *HRAS* (Reddy et al., 1982), that has been demonstrated to have oncogenic potential by itself in NIH/3T3 cell line. With this last discovery, we can define all the main characteristics of oncogenes:

- The mutations affecting this class of genes are generally missense mutations, so that the resulting protein function changes but is not compromised

- The mutations affecting this class of genes alter the original function of the gene by changing it (shift-of-function mutations) or more commonly by enhancing it (gain-of-function)
- Oncogenes are generally dominant, in the sense that one single altered copy is sufficient for an oncogenic effect
- In some cases (like *RAS* family genes) one oncogene is sufficient to transform the normal cell into a neoplastic cell (Fasano et al., 1984)

This brief historical context of mutational theory of cancer can be summarized in a series of milestones as such:

1. Cancer is a somatic disease, originating from aberrant behavior of normal cells
2. The aberrant behavior is given by alterations in the DNA structure or content
3. Cancer is a multi-step process, given by the accumulation of mutations over time and therefore age is the main risk factor for carcinogenesis
4. There exists a genetic predisposition towards the development of such alterations that can be inherited and creates tumors with early onset
5. Cancer develops through two main forces: loss-of-function in tumor suppressor genes and gain-of-function in oncogenes

2.3 Tumor heterogeneity

What is still missing from these milestones is what is the concordance in terms of genomic makeup among tumor types and also among tumors within the same type. In other terms, what are the possible ways a cancer could develop? To answer this question, scientists were setting the basis of what is currently called *tumor heterogeneity*, so what are the intrinsic genomic differences between different tumor cells. This difference can be seen both between tumors (inter-tumor heterogeneity) and within tumors (intra-tumor heterogeneity). The first attempt at a definition of tumor heterogeneity is strictly correlated to clonal evolution theory (Nowell, 1976). In 1976, Nowell wrote:

‘ *The acquired genetic instability and associated selection process, most readily recognized cytogenetically, results in advanced human malignancies being highly individual karyotypically and biologically. Hence, each patient's cancer may require individual specific therapy, and even this may be thwarted by emergence of a genetically variant subline resistant to the treatment* ’. Nowell not only pinpointed that each tumor has its own history and biology, but poses the bases of the direction of cancer medicine of the last 10 years. In fact, what we now refer to *personalized or precision medicine* implies by definition that we need to understand the specific genetic makeup of each and everyone disease in order to tailor a specific treatment. This incredible heterogeneity is probably the main reason why finding effective treatments for cancer turns out to be still a major challenge in cancer research. Clonal evolution generates branches that compose a tumor made of various genomes. When a drug is designed to kill certain kind of cells, those with a specific genome, it leaves the possibility to other minor branches to win the battle for survival and recreate a tumor resistant to that drug.

2.4 Cancer genome landscapes

Although the concept of tumor heterogeneity was known since the 70s, it is only with the advent of genome-wide studies and Next Generation Sequencing (NGS) that scientists start to understand fully the landscapes of possible mutational patterns in various tumor types. In 2000, Perou and colleagues delineate the first example of tumor landscapes using microarray data, by bridging the gap between macroscopic subtypes seen by a pathologist via immunohistochemestry and genomic subtypes derived from gene expression (Perou et al., 2000). The authors were able to recapitulate both primary and metastatic machinery and ultimately give a “name”, in their words a *molecular portrait*, to the tumor of each of the 42 patients analyzed in their work. In fact, in the conclusion they stated:

‘ *A striking conclusion from these data concerns the stability, homogeneity and uniqueness of the 'molecular portraits' provided by the quantitative analysis of gene*

expression patterns. We infer that these portraits faithfully represent the 'tumour' itself, and not merely the particular tumour 'sample', because we could recognize the distinctive expression pattern of a tumour in independent samples. ‘

This first example used the most prominent genome-wide technique at that time, expression microarray. Somatic mutations detection was still extremely expensive and just a few targeted genes at a time could be analyzed using polymerase chain reaction (PCR)-based capillary sequencing techniques. The principle behind the identification of somatic mutations was, at its core, the same we use today with NGS: separately sequencing normal germline DNA and a tumor sample and call as somatic mutations every base that is present in the tumor but not in the germinal line. In 2007, Vogelstein’s group at the John Hopkins was able to delineate the first mutational landscape of two kinds of solid tumors, breast and colon (Wood et al., 2007). In this seminal work that looked at most of the coding genes in the human genome (20,857 transcripts from 18,191 genes), they found an average of 70 mutations per sample, approximately ten-times more than the estimation made by Nordling model in the 1953. Furthermore, they add three fundamental milestones to the somatic mutation theory:

1. The landscape of mutations in cancer is formed by few mountains and many hills. Mountains represent genes mutated in more than 10-20% of the samples while hills represent genes mutated at a frequency of 5% or lower.
2. Not all mutations are as important as the others. ~15/70 can be called *drivers* as they promote tumor growth and survival. The majority of them are simply *passengers*, so mutations that appear in the context of genomic instability of neoplastic cells and are simply dragged over the generations of cells being not under any selective pressure.
3. While mountains are certainly driver, each tumor has its own hills. Hills are linked between each other in common pathways.

2.5 Driver VS Passenger: a problem of mutation rate

Vogelstein's work on cancer genome landscapes represented in fact the last of a series of other seminal works probably opened by Vogelstein himself in 2004. In the review *Cancer Genes and the pathways they control* (Vogelstein and Kinzler, 2004), the problem of understanding the entire molecular landscapes of cancer finally emerged in its paramount importance:

' *There are at least three major challenges that will occupy most cancer researchers' time over the next 10 years. The first is the discovery of new genes that have a causal role in neoplasia, particularly those that initiate and conclude the process. The second is the delineation of the pathways through which these genes act and the basis for the varying actions in specific cell types. The third is the development of new ways to exploit this knowledge for the benefit of patients* '.

Before the early years of 2000, a precise estimate of how much of the genome was mutated in cancer was mostly based on mathematical models about mutation progression (Goldman and Yang, 1994; Yang et al., 2003). The word *driver* itself was used mostly to define oncogenes and tumor suppressors behaviors through an automotive metaphor that is still used in every basic course in cancer genomics today. Mutations in oncogenes are like cars with a stuck accelerator and mutations in tumor suppressors are like cars with a dysfunctional brake (Vogelstein and Kinzler, 2004). The concept of *passenger* in the somatic mutation theory was therefore linked to *driver* when the first PCR-based work on hundreds of genes started to be published in the attempt to correct an estimation of the mutation rate that was largely wrong. Milestone works in this sense are the ones by Greenman in 2006 and 2007, based on the analysis of 518 kinase genes on a cohort of 210 tumors (Greenman, 2006; Greenman et al., 2007). Mathematical models on somatic mutation processes could be applied and polished based on actual genome-wide data, taking into account previous modeling about phylogenetic (Goldman and Yang, 1994) and codon substitution rates and pathogenic effects (Yang et al., 2003). In the same

period, before the rise of the NGS era, another publication by Sjöblom *et al.* reached similar conclusions about the nature of driver and passengers through the analysis of ~13'000 genes in 11 breast and 11 colorectal cancers (Sjöblom et al., 2006). The works of Greenman and Sjöblom were compared in a Nature editorial in 2007 (Haber and Settleman, 2007), showing the poor overlap in terms of specific mutations on kinases, despite the approaches were in fact very similar to each other. The necessity of a larger sample to discern the entire repertoire of driver genes was evident and finally brought on by the efforts of the Human Cancer Genome Project, whose main repository, The Cancer Genome Atlas (TCGA), is used over the entire Results section.

2.6 Cancer genomics and human genetics

The idea of passenger mutations was not completely new 10 years ago. There was already huge evidence of what is called *hitchhiking* in genetics. Some of the variations seen in the human genome can increase their allele frequency or go extinct simply by being "close" (in genetics vocabulary, in linkage disequilibrium, LD) with alleles under selective pressure. The term genetic hitchhiking was coined in 1974 by Mainard Smith and Haigh (Smith and Haigh, 1974) and brought back to attention by Gillespie with the pseudo-hitchhiking model of genetic draft (Gillespie, 2000). This model made the fortune of the Genome-Wide Association Studies (GWAS) era that relies on the possibility to find regions of the genome associated with a disease by simply looking at few SNPs all along the genome. Under LD model, each SNP can also account for the entire surrounding region that is inherited together according to its LD-block. When a SNP is found associated with a trait, it is possible to calculate where the real pathogenic variant could reside and this SNP will represent a proxy for the unknown pathogenic variant. The need of a map of human variation to understand common and rare haplotypes in any kind of diseases was of paramount importance and HapMap and 1000 Genome Projects were both born, along with the Human Genome Project, to serve this purpose. Again, the

same urge of map of human variations was already discussed by Dulbecco 20 years earlier when talking about the future of cancer research after viral models for the discovery of oncogenes (Dulbecco, 1986):

‘ *If we wish to learn more about cancer we must concentrate on the cellular genome [...] we have two options: either to try to discover the genes important in malignancy by a piecemeal approach or to sequence the whole genome of a selected animal spieces. [...] In which species should this effort be made? If we wish to understand human cancer, it should be made in humans because the genetic control on cancer seems to be different in different species* ‘.

Research on cancer predisposition had a burst in interest during the early 90s with the discovery of BRCA1 and BRCA2 responsible for the early onset of certain breast and ovarian cancer. The risk for a carrier has been estimated to be at least 5-fold higher for breast cancer by the age of 70 (Chen and Parmigiani, 2007). In 1990, the group of Mary-Claire King at UC Berkley defined the region of susceptibility as being 17q21 (Hall et al., 1990) and after a rush lasted four years, BRCA1 was finally cloned by University of Utah and Myriad Genetics (Miki et al., 1994). It was believed that after the discovery of other highly penetrant (but way more rare) susceptibility genes like *PALB2*, the remaining missing familiarity toward cancer could be only explained by combinations of common variants that announced the advent of GWAS. While genome-wide somatic studies on cancer were still in preparation around the 2010, when NGS technology was taking place, large-scale GWAS studies were conducted on cancer. In particular, the Collaborative Oncological Gene-Environment Study (COGS), developed a unified array for the study of three major hormone dependent tumor types, breast, ovarian and prostate cancer, that encompassed over 200’000 individuals by the beginning of 2013 (Sakoda et al., 2013). The results of GWAS studies in cancer, despite the huge efforts and investments, are still controversial (Check Hayden, 2013; Visscher et al., 2012).

2.7 Cancer genomics in the NGS era

Although similar in principle, there is a substantial difference between passenger variants and passenger mutations. While genetic draft is a mechanical phenomenon, somatic mutations are way more unpredictable. The probability for a base to be altered in a tumor could depend by a plethora of different events, which include genomic instability, carcinogens effect, viral infection or simply the downstream effects of other driver mutations. So, since Laura Wood and Bert Vogelstein molecular landscapes, one of the main tasks of cancer genomics was to distinguish driver from passenger mutations and create a catalogue of cancer genes with the final aim of developing personalized treatments. The first attempt at a catalogue for cancer DNA similar to what 1000 Genomes represented for human DNA is COSMIC (Catalogue of Somatic Mutations In Cancer), established in 2008 under the Cancer Genome Project (CGP) (Forbes et al., 2008). In the same year, Timothy Ley and colleagues showed the results of the first whole genome sequencing of a leukemia sample (Ley et al., 2008) performed using one of the first Next Generation Sequencing (NGS) technology, the Roche 454 pyrosequencing. NGS technologies dramatically changed the way cancer genomics was perceived and studied, by allowing hundreds of sample data to be aggregated and analyzed at the same time and showing how deep the heterogeneity within and between tumors was. Furthermore, a complete new set of tools and standards in bioinformatics were developed to support this new biological big data era (Li et al., 2009). Other tumor types followed: the first breast cancer (Shah et al., 2009), then lung (Pleasance et al., 2010) and prostate (Berger et al., 2011), analyzed with the Illumina Genome Analyzer II. This technology was routinely used for the second phase of The Cancer Genome Atlas (TCGA) to collect mutation data. This consortium, born in 2006, started with the idea of characterizing 3 tumor types, glioblastoma multiforme, lung and ovarian cancer. In

2009, with the decrease in sequencing cost, a 5-years project started with the aim of characterizing 20-25 tumor types. Currently, over 30 tumor types are included.

This work will take advantage of these data, to further understand both cancer predisposition and cancer somatic development and ultimately divide what is driver from what is passenger in both germinal and tumor genomes.

3 Material and Methods

This work represents a collection of bioinformatics methodological approaches to tackle the identification of driver forces that lead to cancer risk and tumor formation in the DNA. Therefore, an in depth analysis of materials and methods was included in the results for each section, as they represents results per se. Nevertheless, there are a few common elements at the base of this works, mostly about data format, retrieval, and management.

3.1 Data Format

Each experimental approach in NGS bioinformatics has its own format for storing and analyzing data. While raw formats such as FASTQ and BAM files are common to all sequencing technologies, mutations and variants have their own specific representation.

3.1.1 VCF format

The Variant Calling Format (https://samtools.github.io/hts-specs/VCFv4.2.pdf) represents the first real shift in genome sequences storage from SNP-based arrays towards NGS technologies. In fact, it was born from the advent of the 1000 Genomes Project (Danecek et al., 2011) as a more appropriate format for large-scale genome sequencing compared to NCBI General Feature Format (GFF, http://gmod.org/wiki/GFF3), which is oriented to larger regions of the genome and

resembles the BED format, or PEDigree files (PED, https://www.cog-genomics.org/plink2/formats#ped) that still are the standard for genetics and genotype calls. At its core, it is composed by 4 columns, representing chromosome, position, reference and alternative allele, plus the genotype call of each sample included in the call. The main advantage over the aforementioned formats are: i) its ability to be directly connected to the reference genome (via the chromosome-position-reference system), ii) its ability to carry variant call measures such as depth of coverage along side with the genotype call itself and iii) compared to the PED format, an emphasis on variant/mutation (one for each row), that is necessary when working on millions of variant/mutation at a time. PED format put the emphasis on the subjects (one for each row), in a time where the number of variants mapped could be even less than the number of subjects genotyped. This shift from a sample-based to a variant-based format, allows an easy annotation of the file, with several optional field such as allele counts and ethnicity-wise allele counts introduced by the 1000 Genomes Project itself.

3.1.2 MAF format

The Mutation Annotation Format (https://wiki.nci.nih.gov/display/TCGA/Mutation+Annotation+Format+(MAF)+Specification), is a cancer-specific format for somatic mutations developed together with The Cancer Genome Atlas (TCGA) to show VCF files in a conveniently annotated version that include, by default, features such as gene name, mutation type (missense/indel/splice site etc.). The main difference with respect to VCF, is that the same mutation is repeated with a new row for every new samples that harbors it. This shift from a wide to long format facilitates readability and data manipulation (the number of columns does not change by adding new samples) at the cost of increasing file size. The number of somatic mutations is generally not too high (on average from

thousands to hundreds of thousands), even in whole-genome-sequencing, and file size is not an issue compared to a VCF storing germline variants with millions of records.

3.2 Data Retrieval

In the last 5 years, several freely available databases of sequencing data were released to the public. These resources represent an unprecedented opportunity for analyzing DNA sequencing data with thousands of patients with mutations and variants at exome and genome scale. The main resources used in the three works presented here are: TCGA, ICGC, cBioPortal, COSMIC/CGC and ExAC.

3.2.1 TCGA

The Cancer Genome Atlas (TCGA) is an international consortium based in the United States born in 2006 from the collaboration between the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI). It represents an unprecedented effort to integrate RNA, mRNA, DNA, copynumber and epigenomic data from thousands of patients in ~30 cancer types. Data from this resource were used in all the three chapters of the results. In general, somatic mutation calls are freely available in MAF format that is extremely convenient for the high level analysis of the first two chapters of the results. The third section of the results instead, made use of raw bam file from 673 breast cancer germline samples that were analyzed from scratch.

3.2.2 ICGC

The International Cancer Genome Consortium (ICGC) was born in 2008 to coordinate worldwide cancer sequencing projects of over 50 cancer types. Its main datacenter and secretariat is based in Toronto, Canada. In recent years, it also absorbed large part of the TCGA database and the format used for mutation is very similar to the MAF format used by TCGA. Data from this consortium were used throughout all the presented works.

3.2.3 cBioPortal

The cBioPortal for Cancer Genomics is an aggregation datacenter and analysis web-tool for cancer genomic analysis that gathers studies from 146 projects in over 50 tumor types at the time this work was written (Cerami et al., 2012). It includes almost every data from ICGC and TCGA plus other smaller studies. It represents one of most complete resource for exploratory analysis on cancer data and ships an R package called cgdsr for the fast retrieval of gene specific and clinical information. This package was wrapped and used inside our LowMACA R package to analyze pattern of mutations and is presented in the second section of the results of this work.

3.2.4 COSMIC and CGC

The Catalogue Of Somatic Mutation In Cancer (COSMIC) is a database created at the Wellcome Trust Sanger Institute in 2004 as an actual collection of somatic mutation in cancer that gather information from both sequencing and array based data to draw a map of the known somatic alterations of thousands of cancer samples (Forbes et al., 2008, 2011). From this database, the Cancer Gene Census (CGC) was created as a source of the established cancer genes both at predisposition and somatic driver level (Futreal et al., 2004). Both COSMIC and CGC were used in all the sections of the results as an annotation and reference set of established somatic mutation and driver genes for various comparisons.

3.2.5 ExAC

The Exome Aggregation Consortium (ExAC) is one of the largest databases of human variants that is currently freely available. It is a collection of 60'706 unrelated individuals from 15 different studies with exome sequencing data from their germline DNA. The database also includes data from germline of TCGA patients that represents over the 10% of the participants of the ExAC. These data were used in the third section of the

results as a control sets against breast cancer patients, after the removal of all cancer samples and samples not from European origin. In total, we used over 20'000 control individuals to characterize possible cancer predisposing variants (see section **4.3**).

3.3 Data Processing and Manipulation

Mutations and variants acquired a common representation with the advent of NGS. While the formats reached a sort of standardization over the last years, methodologies to obtain such data did not. In particular, mutation call is one of the most controversial points in cancer genomics, with many different algorithms developed and results that hardly agree with each other, both at somatic and germline level (Alioto et al., 2015; Bodini et al., 2014; Hwang et al., 2015). While sections 1 and 2 of the results make use of precomputed mutations from ICGC and TCGA, with pipelines standardized at least for each tumor type, in the third section we had to developed the entire pipeline of preprocess by ourselves. Among the many available, we chose the Genomic Analysis Tool Kit (GATK) to preprocess our data from alignment to variant call (DePristo et al., 2011; McKenna et al., 2010; Van der Auwera et al., 2013). In particular, a typical variant call pipeline is composed as such:

1. FASTQ filter for bad quality reads
2. Alignment using BWA (Li and Durbin, 2009)
3. Picard markduplicates (http://broadinstitute.github.io/picard)
4. Indel Local realignment (GATK)
5. Estimate systematic error and base quality (GATK)
6. HaplotypeCaller for variant calling (GATK)
7. Combine multiple VCF and GenotypeGVCF (GATK)
8. Variant Quality Score Recalibration for both SNPs and InDels (GATK)

At this point we obtain the final VCF that includes all the samples. Further preprocess includes: i) Split multiallelic sites (Tan et al., 2015) ii) Annotate using ANNOVAR

(Wang et al., 2010) iii) Adjust ANNOVAR output to obtain a MAF-like format. Heavy formatting of variants were carried on using vcftools/bcftools (Danecek et al., 2011) and vt (Tan et al., 2015), while all data munching and statistics made use of R and data.table package for speed up (https://cran.r-project.org/web/packages/data.table).

This brief summary gives an idea of the tools that can be used when parsing and analyze mutation data, but for a detailed explanation of the methodological set of each section, refer to the specific Materials and Methods paragraph in **4.1.4**, **4.2.3**, **4.3.3**.

4 Results

4.1 DOTS-Finder: a comprehensive tool to assessing driver genes in cancer genomes

This section is adapted from (Melloni et al., 2014) and it represents an attempt to create a comprehensive method for the identification of somatic driver genes. Following the seminal work from (Vogelstein et al., 2013), we developed a tool capable of detecting driver genes and separate them in the two main classes of driver genes, tumor suppressors and oncogenes, characterized by distinct patterns of mutation distribution.

4.1.1 Abstract

A key challenge in the analysis of cancer genomes is the identification of driver genes from the vast number of mutations present in a cohort of patients. DOTS-Finder is a new tool that allows the detection of driver genes through the sequential application of functional and frequentist approaches, and is specifically tailored to the analysis of few tumor samples. We have identified driver genes in the genomic data of 34 tumor types derived from existing exploratory projects such as The Cancer Genome Atlas and from studies investigating the usefulness of genomic information in the clinical settings. DOTS-Finder is available at https://cgsb.genomics.iit.it/wiki/projects/DOTS-Finder/.

4.1.2 Introduction

In the last few years, there has been an enormous increase in the amount of data regarding somatic mutations in various cancer types, thanks to technological advancements and reduction of sequencing costs. The massive sequencing of several cancer genomes has led to the identification of thousands of mutated genes. However, only a minority of the identified mutations has a true impact on the fitness of the cancer cells, in terms of conferring a selective growth advantage and leading to clonal expansion (*drivers*), while the others are simply *passengers*, namely, mutations that occur by genetic hitchhiking in an unstable environment and have no role in tumor progression.

Several statistical strategies have been developed to properly identify driver mutations and driver genes. These strategies can be roughly classified in four main categories: 'protein function', 'frequentist', 'pathway-oriented' and 'pattern-based' approaches. The 'protein function' approaches are based on the prediction of the functional impact of a specific mutation in the coding sequence of a protein (Reva et al., 2011; Shihab et al., 2013b; Sim et al., 2012). Although they do not permit the identification of driver genes, they can predict the effect of the mutation on the protein product. The 'frequentist' approaches evaluate the frequency of mutations in a gene compared with the background mutation-rate (Dees et al., 2012; Lawrence et al., 2013; Wood et al., 2007). The 'pathway-oriented' approaches are based on the analysis of the co-occurrence of mutations in a pathway-centered view (Bashashati et al., 2012; Ciriello et al., 2012; Leiserson et al., 2013; Vandin et al., 2012) and are usually focused on searching for driver genes belonging to the most significant mutated pathways. Lastly, the 'pattern-based' approaches identify driver genes by assessing the type of mutations (e.g. missense/truncating/silent) and their relative position on an amino acid map across many cancer samples (Davoli et al., 2013; Tamborero et al., 2013a; Tian, 2011; Vogelstein et al., 2013). They exploit the known structural properties of mutations in tumor

suppressor genes (TSG) and oncogenes (OG). Nevertheless, the identification of driver mutations in cancer remains a major challenge in computational biology and cancer genomics. Indeed, discovering driver mutations is one of the main goals of genome re-sequencing efforts, as the knowledge generated by exome-sequencing will translate from research to the clinic. The results of some of the cited tools are summarized in a recent database called DriverDB (Cheng et al., 2014) and also aggregated in one of the Pan Cancer analysis publications (Tamborero et al., 2013b). From their comparison, it is clear that all these approaches are complementary and only the integration of many of these strategies can improve the identification of driver genes.

Here, we present an innovative tool called DOTS-Finder (Driver Oncogene and Tumor Suppressor Finder) that integrates a novel pattern-based method with a protein function approach (functional step) and a frequentist method (frequentist step) to identify driver genes. In addition, it allows the classification of driver genes in TSGs or OGs. The software is freely available and has been designed to return robust results even with few tumor samples.

4.1.3 Implementation

4.1.3.1 Overview of DOTS-Finder

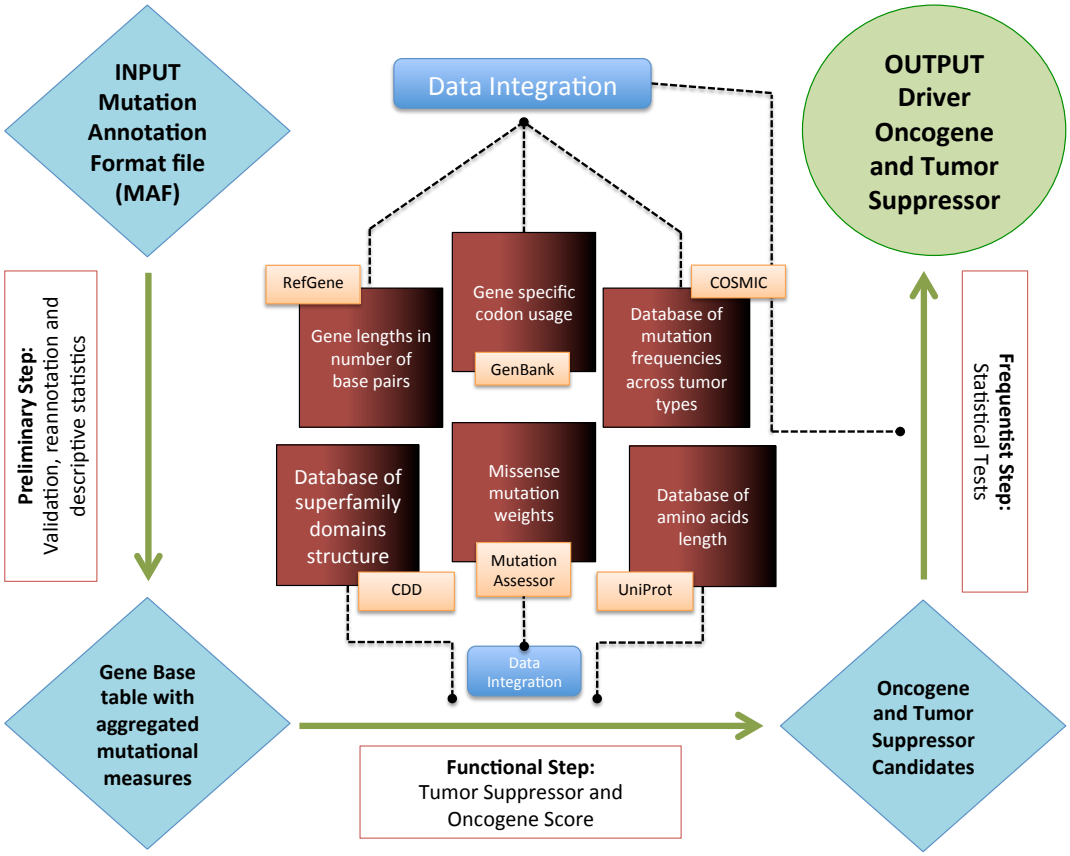

Figure 1 DOTS-Finder workflow. Illustration of the three main steps and the databases used to identify driver genes. Starting from the top left, a MAF file is taken as input. This file can encompass patients with any particular kind of tumor or any stratification of homogeneous samples under specific criteria (e.g. smoker patients with lung cancer, patients <50 years of age, etc.). The workflow includes the following three steps (green arrows): 1) Preliminary step: the dataset is filtered, reannotated and aggregated by gene (from top-left to bottom-left); 2) Functional step: TumorSuppressorGene – Score (TSG-S) and OncoGene – Score (OG-S) are calculated (from bottom-left to bottom-right) 3) Frequentist step: four statistical tests are run on genes that exceed the TSG-S and OG-S threshold (from bottom-right to top-right). The center panel (Data Integration) lists the external sources used by DOTS-Finder.

The DOTS-finder pipeline is illustrated in Figure 1. Our method can be applied to genes that are targeted by single nucleotide variants (SNVs) and small insertions and/or deletions (InDels). Given a set of mutations in an exome/genome sequence dataset, the output is a ranked list of genes that prioritizes the best candidate driver genes and classifies them as TSGs or OGs. The user can submit an input Mutation Annotation Format (MAF) file for a set of patients that can be grouped by different criteria. In the Preliminary step, the MAF file is reannotated and several descriptive statistics are

calculated. This produces a gene-based table with aggregated mutational measures. The next two main steps, a functional assessing procedure and a statistical confirming procedure, constitute the core of DOTS-Finder. In the former, putative candidate OGs and TSGs are identified by calculating a Tumor Suppressor Gene Score (**TSG-S**) and an OncoGene Score (**OG-S**), based on the type and location of the mutations occurring in each gene. These scores are inspired by the concepts expressed in a recent study by *Vogelstein et al. (Vogelstein et al., 2013).* The **TSG-S** is based on the ratio between truncating (i.e. inactivating) mutations and total number of mutations found in each gene, under the null hypothesis that this value is equal to the average truncating/total ratio of patients' exomes. The **OG-S** is based on the entropy of the pattern of missense SNVs and inframe insertions/deletions calculated using a Gaussian density model on the protein product. In the latter step, the statistical confirming procedure, the two lists of possible OGs and TSGs undergo four tests to assess whether the mutational pattern in each gene shows a statistically-defined evidence of positive selection based on the mutation rate and the number of non-silent mutations, calculating their statistical probability of being true driver mutations. After correction for false discovery rate, all the genes with a q-value<0.1 are identified as candidate driver OGs or TSGs. The user is free to modify this threshold.

DOTS-Finder is a comprehensive method that considers three main aspects of a mutated gene: it takes into consideration where the mutations are collectively found (pattern-based approach), what is the effect of mutations on protein products (protein-change approach), and what is the frequency of these mutations in the sample (frequentist approach). Our method is able to overcome many of the problems derived from the application of each individual approach. First of all, the prediction ability of frequentist approaches such as MutSigCV (Lawrence et al., 2013) relies on the estimation of the so-called background mutation rate (BMR). Nevertheless, a precise map of the BMR in the

whole genome is still unavailable and constitutes one of the unresolved challenges of cancer genomics. A plethora of genomic events, such as transcription and replication timing, are associated with the fact that part of the genome is more prone or less prone to mutation. In particular, experimental data of these two events showed a significant correlation with the probability of a mutational event (Lawrence et al., 2013). However, while these experiments should be context specific (tissue/patient specific), data on replication timing are hard to obtain for every patient and/or tissue. Finally, pure frequentist methods do not allow any classification of the type of aberrations in terms of gain or loss of function. A pattern-based approach can bypass the problem of achieving a correct BMR estimation by focusing on the position of the observed mutations and not on their frequency. Thus, the frequency simply becomes a statistical power boost and not the point of investigation. *Vogelstein et al. (Vogelstein et al., 2013)* provide a scheme to assess whether a gene can be considered an OG or a TSG, but a large amount of data are needed in order to evaluate rarely mutated genes. The Authors' approach, as well as the method developed in TUSON Explorer (Davoli et al., 2013) have been used to collectively evaluate general cancer genes across tumor types, however, when applied to single tumor type, they were found to lack the statistical power to recapitulate the overall results. In particular, with these methods, the discrete calculation of an OG test requires many mutations in the exact same hotspots to reach statistical significance. On the contrary, our approach, which takes into consideration the proximity of mutations by using the Gaussian smoothing, is able to identify also small deviations from a uniform distribution.

The main problem in assessing the value of our method is the absence of a gold standard in the identification of driver genes and the lack of benchmark studies. Indeed, the objects of our investigation are the driver genes of the different cancer types, which are still mostly unknown. However, to have an estimate of the prediction ability of DOTS-

Finder, we decided to compare the aggregated predictions for 12 cancer types with the results of a well-documented global analysis from the Pan-Cancer 12 (Tamborero et al., 2013b) (see section **4.1.5.10**). In this analysis, the Authors combined the outputs of several approaches and we were able to compare our tool with the single output from MutSig, MuSiC, ActiveDriver (Reimand et al., 2013), OncodriveFM (Gonzalez-Perez and Lopez-Bigas, 2012) and OncodriveClust (Tamborero et al., 2013a). We also related the predictions of each method with the Cancer Gene Census (CGC) database (Futreal et al., 2004), a manually curated collection of driver genes. Notably, DOTS-Finder emerged as the best available tool because of its sensitivity to find both known and new candidate driver genes.

Moreover, we have applied DOTS-finder to 34 tumor types and compared its output with the results of other approaches. Our approach shows results which are consistent with the literature for both high and low mutation rate cancers; DOTS-finder allows detection of new plausible driver candidates while excluding highly mutated genes not associated with cancer, the so-called "fishy genes", such as the Mucins, Titin and most of the olfactory receptors.

DOTS-Finder requires minimal input files, it is easy to use, and does not necessitate any programming skill or statistical knowledge. Indeed, we created a tool accessible to researchers in a wide range of fields. Compared with popular tools like MuSiC (Dees et al., 2012) and MutSigCV (Lawrence et al., 2013), we only require the availability of easily accessible MAF files. The users do not need to have bam files as in MuSiC, which are not publicly available or easily accessible. In addition, the users do not necessitate any proprietary software, as the source code is written in Python and contains some embedded R codes, which are two freely available languages. Since DOTS-Finder is released under the GNU GPLv3+ license, users are also free to modify the code and implement new features.

DOTS-Finder is an easy solution for investigating genomic information from existing exploratory projects like The Cancer Genome Atlas (TCGA), but it is especially useful to identify reliable driver candidates in small studies assessing the value of genomic information for clinical purposes, such as understanding and predicting chemoresistance or metastatic spread. Indeed, we performed a saturation analysis on the mutational data present in 238 bladder cancer patients using 9 subsampling fractions, and, as shown in section **4.1.5.11**, DOTS-Finder can perform statistically better than our best competitor, MutSigCV Version 1.4, in terms of number of drivers found and precision-recall balance in small sample datasets. Our tool could recapitulate up to 40% of the results of the entire dataset with just 5% (i.e. 12 patients) of the dataset. Thus, it can be used in the clinical research setting to help identifying driver genes that can assist patient stratification for prognosis and choice of treatment. We envisage that DOTS-Finder might facilitate the identification of candidate targets, which could be used to develop diagnostic, prognostic or therapeutic strategies, even in situations where the available data are scarce (e.g. rare tumors).

4.1.3.2 The Functional Step: finding tumor suppressor gene and oncogene candidates

On the basis of previous proposals (Tian, 2011; Vogelstein et al., 2013), we developed scores to assess if a gene in a given tumor could be considered either a TSG or an OG candidate. A TSG is characterized by loss of function mutations. Typically, these mutations are truncating and tend to destroy the protein product or make it non-functional. Frame shift mutations, SNVs creating a stop codon, non-synonymous mutations on the stop codon, translations in the start site, and splice site mutations are all considered of the *truncating* type. Ultimately, a TSG is characterized by truncating mutations in a non-specific pattern (Figure 1, Panel A).

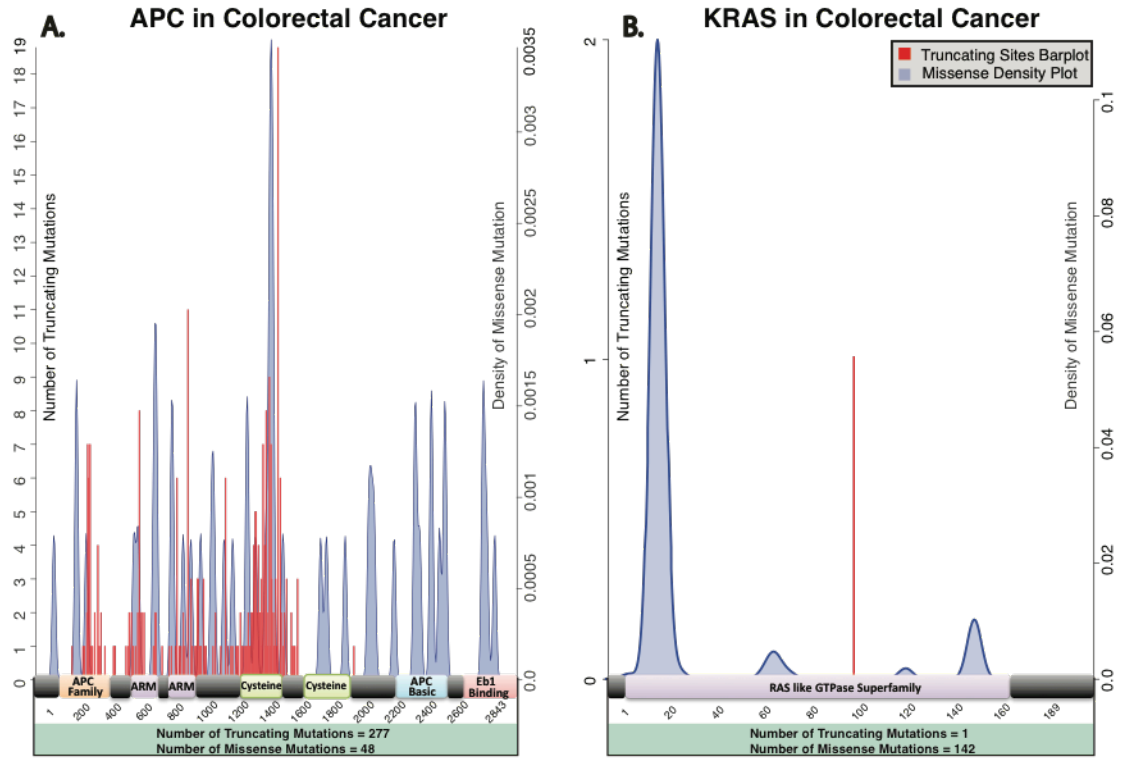

Figure 2 Mutational patterns of typical tumor suppressor genes and oncogenes. (A) Mutations of *APC* in Colorectal Cancer. This is the mutational landscape of a typical TSG with diffuse truncating mutations (in red) and a non-specific pattern of missense mutations (blue density plots). Truncating mutations cover 85% of all the non-synonymous mutations on *APC*. **(B)** Mutations of *KRAS* in Colorectal Cancer. This is the mutational landscape of a typical oncogene with significant clusters of mutations, which are present in specific hot spots of the protein ideogram (blue density plots). In particular, *KRAS* tends to mutate on amino acids 12 and 13 (119/143 mutations). The total numbers of truncating sites and missense mutations are indicated in the panels. The mutations are mapped on the corresponding canonical protein ideogram, therefore, not all the mutations can be represented (e.g. splice sites mutations are not included in the figure).

An OG, on the other hand, is characterized by gain or switch of function mutations that confer new properties on the protein product or simply enhance the existing ones. Hence, the typical mutations affecting an OG are missense mutations on key amino acids or on specific domains. We consider a "*missense* type" mutation all the non-synonymous SNVs that do not create a stop codon and occur outside start codons or stop codons, and all the insertions and deletions not altering the reading frame (Inframe InDels). These mutations have a particular pattern, as they are generally clustered in one or more regions along the protein (Figure 2, Panel B). For example, in leukemias, *IDH1* can bear different kind of mutations, but almost always at amino acid position 132 (Figure 3).

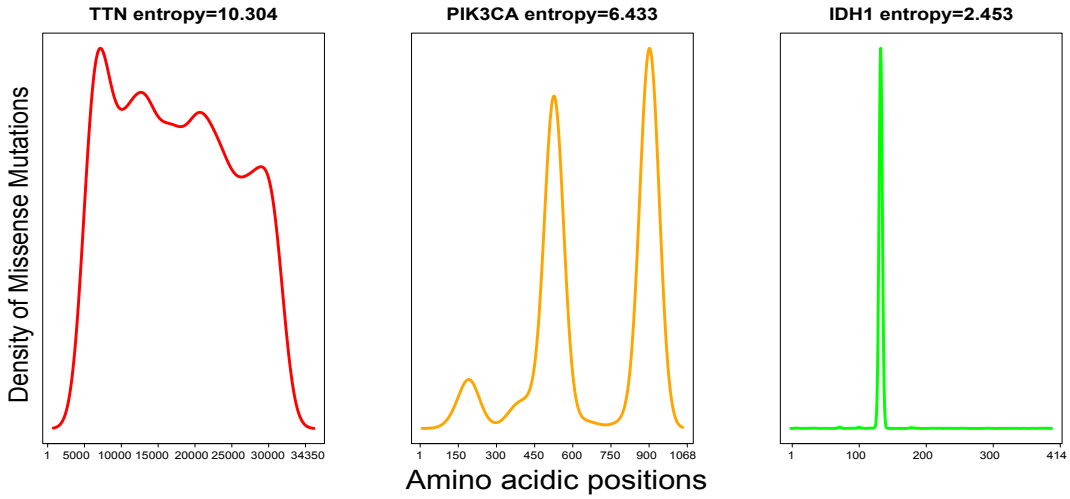

Figure 3 Density plot of genes tested for oncogenetic characteristic. Three known highly mutated genes are presented showing data from COSMIC. *TTN* is a notorious "giant gene" that is often found mutated because of its length. It does not show any particular clusterization around hotspots and the information entropy of its mutations is therefore very high. *PIK3CA* and *IDH1* retains visible clusters of mutations; three hotspots for the first one (entropy=6.433) and one unique hotspot on amino acid 132 for the second one (entropy=2.453)

The **Tumor Suppressor Gene Score** (**TSG-S**) evaluates whether a gene harbors an elevated number of truncating mutations compared with the total number of mutations present on that gene. Given 64 codons in the DNA and 9 possible SNVs per codon (3 nucleic acids × 3 possible changes) we have a total of 576 possible base changes. Only 23 of them can be considered truncating (~ 3.9% of all the SNVs, weighted for the actual human codon usage) against the 415 non-synonymous single base changes that lead to missense variations and 138 silent mutations. If we take into account all the InDels that

corrupt the reading frame of a gene, we can estimate, based on our sample data, that the ratio between truncating mutations and total number of mutations in cancer is approximately around 14%, with a standard deviation of 4. This percentage ranges from a minimum of 9% in glioblastoma (GBM) to a maximum of 25% in pancreatic adenocarcinoma (PAAD), with high intra-tumor variability among patients. This discrepancy indicates that some tumors are more prone than others to acquire and maintain truncating mutations (Figure 4).

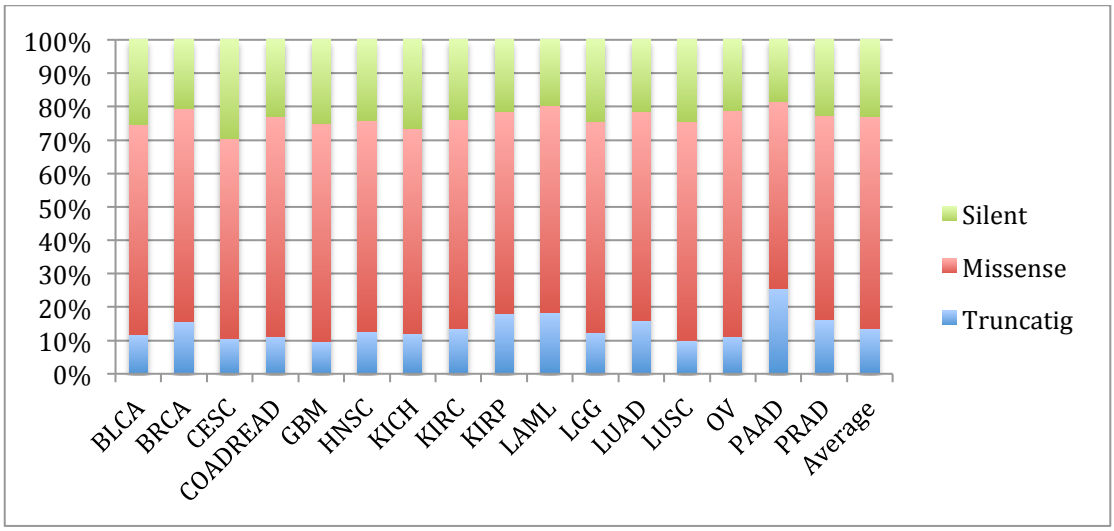

Figure 4 Distribution of mutations per type across cancer types. In this figure we calculated the average percentage of truncating, missense and silent mutations in the patients of 16 different cancer types from TCGA data. These percentages can vary considerably across tumor types but we can assess that on average, 14% of the mutations can be considered truncating, 21% silent and the vast majority, 65%, are missense.

The TSG-S is calculated using a binomial distribution under the null hypothesis that the ratio between truncating mutations and total number of mutations found in each gene is equal to the average truncating/total ratio in patients' exomes (Figure 5).

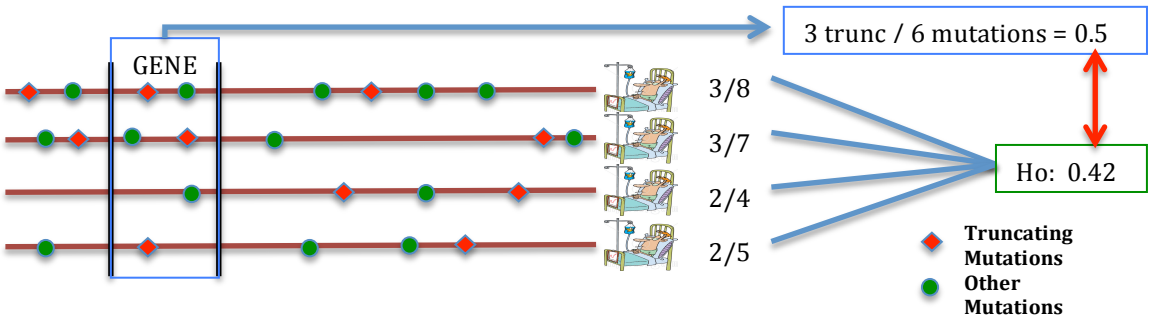

Figure 5 Calculation of TSG-Score. The TSG-S is calculated using the ratio between the number of truncating mutations and total mutations found in the gene. In the example, 3 mutations are truncating over a total of 6. This ratio is compared to the average ratio of truncating over total in the affected patients.

The calculation of this score is set in the specific cancer-patient environment where the gene is found mutated, following the idea that a truncating mutation in a sample with few other alterations weights more than a mutation in a hypermutated sample.

The **OncoGene Score (OG-S)** indicates whether a gene harbors an elevated number of missense mutations in certain regions of the gene. The Score is based on the Shannon's entropy of the pattern of missense SNVs and inframe insertions/deletions, calculated using a Gaussian density model on the protein product. Every mutation is weighted for the actual Functional Impact provided by Mutation Assessor (a 'protein function' method) (Reva et al., 2011) and compared with a random model estimated by a bootstrapping procedure. The score is able to catch the clusterization of mutations around significant hot spots in a gene.

We set a threshold for the two scores based on the analysis of the Catalogue Of Somatic Mutations In Cancer (COSMIC) (Forbes et al., 2008), using as positive control the CGC genes that encompass somatic point mutations. To evaluate the quality of our scores with regard to the classification in driver and non-driver, and avoid making assumptions on the behavior of driver genes, we adopted two strategies. First, we did not consider any *a priori* set of true non-driver genes (negative control) and, second, we did not divide the Cancer Gene Census in OGs and TSGs. As mentioned before, the **OG-S** and **TSG-S** work on different levels and different mutation types, so we do not exclude the

possibility that the same gene might show oncogenic and tumor suppressor features at the same time in different tumors, or even in the same cohort of patients (see Atypical tumor suppressor genes and oncogenes section).

Since the mutated genes reported in COSMIC are more than 18000, the known drivers in CGC accounts for less than 1% of all the mutated genes. These numbers indicate that the two classes are extremely unbalanced, and that a common "Receiving Operator Characteristic" analysis is not appropriate to address the goodness of our scores. We therefore calculated the Matthews correlation coefficient curves (MCC) for the two scores and maximize their values to obtain our thresholds (Figure 6).

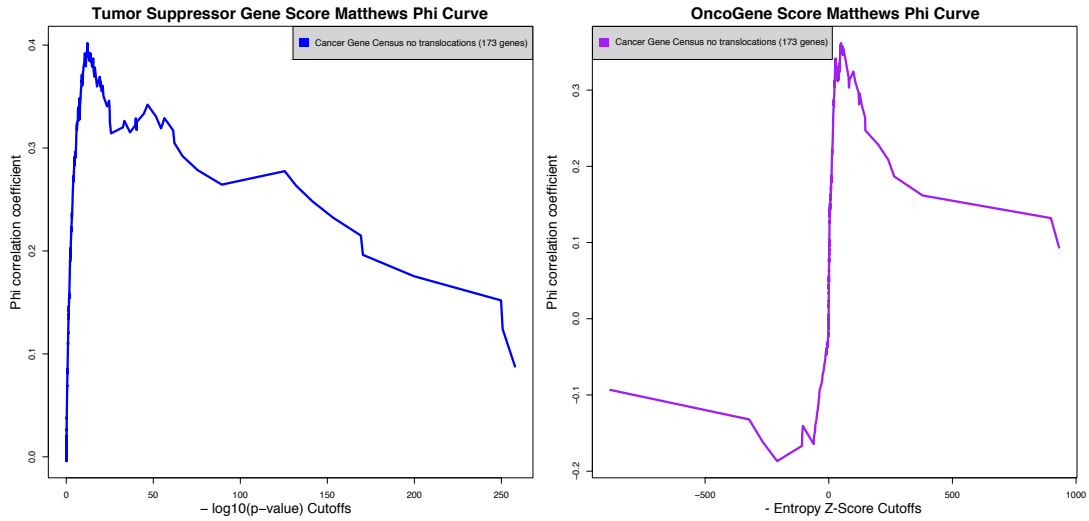

Figure 6 Matthews Phi correlation for the OG-Score and the TSG-Score. The plot shows the trend of the Matthews phi correlation for every possible cutoff of the classification of genes as oncogenes candidate or tumor suppressors. The OG-S and TSG-S are calculated on the COSMIC database v66 using as positive control the genes of Cancer Gene Census. The chosen cutoffs are the ones in which the two functions are maximized.

Compared to other common measures like accuracy, the MCC is much more informative for strongly unbalanced classes (Baldi et al., 2000). Our thresholds were also rescaled for every tumor type in order to take into account the setting-specific mutation rate and the number of samples at our disposal.

4.1.3.3 The Frequentist step: assessing the possible drivers

The genes that exceed at least one of the thresholds of the two scores, are classified as OGs or TSGs and four tests are then performed to assess if the mutational pattern in each gene shows a statistically defined "driver behavior". This analysis is complex, as it requires the proper estimation of the BMR, which is specific for each gene in each tumor type and patient. Indeed, we foresee at least seven sources of background mutation-rate heterogeneity: i) the specific mutation-rate of each tumor type; ii) the specific number of mutations in each patient; iii) the GC-content, as most of the mutations found in cancer are point mutations occurring in GC spots; iv) the gene size; v) the gene-specific SNP frequency; vi) the replication time; vii) the levels of gene expression. However, there might be other unknown parameters that could also influence the background mutations rate of a gene. Our method does not need to take into consideration either replication timing or gene expression levels, since they both require a great amount of new experimental data.

Briefly, the four tests used by DOTS-Finder are: 1) **Higher Frequency Test**. The rate of non-synonymous mutations per Mb in a gene is compared with the rate of mutations in the patients carrying mutations in that gene. 2) **Non-synonymous versus Synonymous Ratio Test**. Given the total number of mutations found in a specific gene, this test assesses whether the number of non-synonymous mutations is higher than the expected number of non-synonymous mutations. The expected value is calculated on the probabilistic ratio obtained by randomly placing the same number and type of mutations on the specific codon usage structure of the gene. 3) **Tumor-specificity Test.** This test prioritizes the driver genes in the different tumors, although it is not fundamental for the driver assessment. The frequency of non-synonymous mutations in the samples is compared with the frequency found in the COSMIC database across tumor types. The test verifies whether the frequency of non-synonymous mutations in a particular tumor

or situation is higher than the general frequency found in COSMIC. The idea is that some mutations are tissue-specific and might be driver only in certain kind of cancers. For example, *NPM1* is a clear driver gene specific for leukemias; similarly, *VHL* is specific for renal cancer. 4) **Functional Impact Test**. This test is used to verify whether the functional impact score of the gene mutations, calculated by Mutation Assessor, is higher than the average score in the patients affected by a mutation in that gene. The four p-values obtained from these tests are combined using the Stouffer's method with specific weights, in order to take into account both the dependencies between tests and their relative importance in the driver definition (see section **4.1.4.6**). The resulting p-value is then adjusted to correct for false discovery rate.

4.1.4 Material and Methods

4.1.4.1 Availability

DOTS-Finder can be downloaded at http://cgsb.genomics.iit.it/wiki/projects/DOTS-Finder under GNU GPLv3+. Full explanation on how to install DOTS-Finder, how to use it and how to interpret the results can be found at http://cgsb.genomics.iit.it/wiki/projects/DOTS-Finder/Documentation.

4.1.4.2 Input Format

DOTS-Finder accepts the following input formats:

1. MAF format version 2.3 (10 May 2012) and 2.4 (6 March 2013). The program is also a complete MAF format validator in case of submission to the TCGA. The MAF file specifications can be found at https://wiki.nci.nih.gov/display/TCGA/Mutation+Annotation+Format+(MAF)+Specification.
2. MARF format. The Mutation Annotation Reduced Format is a short version of the MAF format with just 13 columns instead of the canonical 34.

3. Annovar CSV (Wang et al., 2010). This is one of the most common annotator for exome/genome sequencing data; it is not directly supported, but we provide a simple step-by-step conversion method to MARF format.

4.1.4.3 Requirements

DOTS-Finder runs on MacOS and Unix based machines. The code is written in Python and contains embedded R codes. DOTS-Finder uses embedded version of bedtools and liftOver, thus it cannot be available for Windows users. In order to work properly, these freely available languages must be already installed with their libraries and packages:

- Python 2.7
- R >= 2.0.0
- CRAN package 'multicore'

4.1.4.4 Mutation data

We analyzed data from TCGA and COSMIC for a total of 8187 samples. The full database is the one used by TUSON Explorer (Davoli et al., 2013), available at http://elledgelab.med.harvard.edu/wp-content/uploads/2013/11/Mutation_Dataset.txt.zip. We removed from the Central_Nervous_System_NS dataset the patients not coming from the oligodendroglioma cancer type and integrated the original datasets used by TUSON Explorer with data from samples of diffuse large B-cell lymphoma (DLBCL) (Lohr et al., 2012) and chronic lymphocytic leukemia (CLL) (Wang et al., 2011). We also collect additional data from (Lawrence et al., 2014), available at www.tumorportal.org, including samples from other cancer types: multiple myeloma (MM) (Chapman et al., 2011), rhabdoid tumor (RHAB) (Lee et al., 2012) and carcinoid (CARC) (Francis et al., 2013).

4.1.4.5 Databases

The method is guided in all the different passages by sources of information on proteins and genes derived from several public databases. The exon length of the gene is calculated using the RefGene hg19 UCSC table (Pruitt et al., 2009) as the minimum number of exons (in base pairs) required to encompass all the possible annotated transcripts for that gene. In case a gene of interest is not annotated on RefGene, the length is set to the average value (3192 bp). The raw frequency of mutation per gene is derived from COSMIC v66 (Forbes et al., 2011) and calculated among all the samples stored in the database across any tumor types (947213 samples). The number of amino acids is derived from the UniProt database (Consortium, 2013) while the domains structure is taken from the "superfamilies" found on the NCBI Conserved Domain Database (Benson et al., 2013). The Functional Impact Score used for the **OG-S** (OncoGene Score) is taken from the Mutation Assessor database (Reva et al., 2011).

A Single Nucleotide Variation (SNV) can result into two different effects on the codon that will be transcribed: it can either change the amino acids (non-synonymous mutation) or mantain the same amino acid exploiting the redundancy of codons over amino acids (synonymous mutations). For every single base change (C>G, A>T etc.), we can derive how many changes lead to a non-synonymous variation or to a synonymous variation for every possible codon (Figure 7).

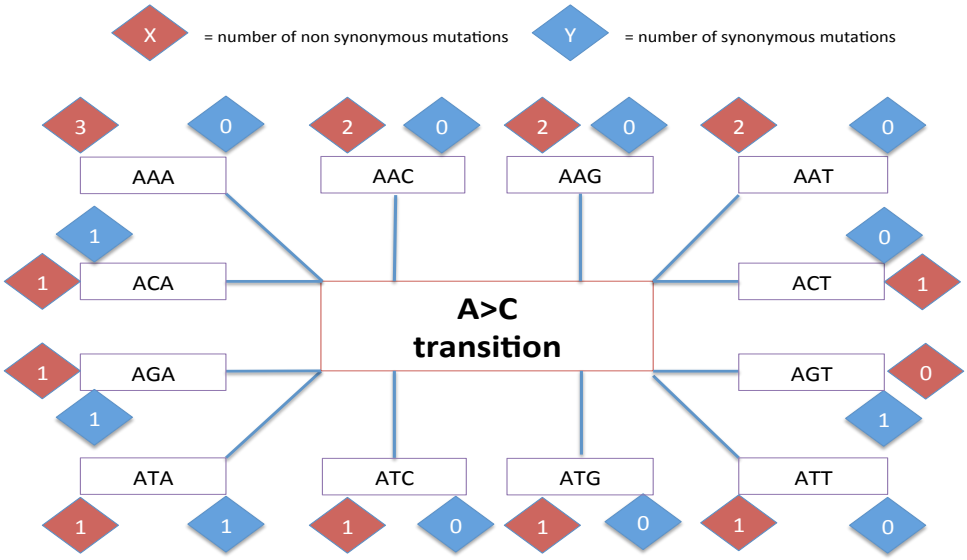

Figure 7 Effect of A>C transition on some codons. At top left corner, AAA codes for lysine and retains 3 spots of possible A>C mutations. All these A>C transitions lead to a change in the codified amino acids and are therefore non-synonymous mutations. ACA codes for threonine and is composed by 2 adenines. A change in the first A, lead to CCA, a proline (non-synonymous), while last A brings ACA to ACC that is still a threonine and therefore a synonymous SNV. According to the entire map of possible changes (A>T, T>A, C>G, etc.), weighted by gene specific codon usage, we can derive a comprehensive landscape of effects

We took the human codon usage from the NCBI GenBank (Benson et al., 2013) *via* the Kazusa website ftp://ftp.kazusa.or.jp/pub/codon/current/species/9606 to derive what we have called the *79 rule*: in a random mutation process on human genome, where all the types of transitions and transversions have the same probability to appear, given *n* random mutations on a human genome, 79% of them will be non-synonymous.

$$\lim_{n\to\infty} \frac{n_{non-synonymous}}{n} = 0.79$$

or, in other words:

$$\frac{n_{non-synonymous}}{n_{synonymous}} \sim 3.78$$

The number of non-synonymous mutations will be ~ 3.78 times higher than the number of synonymous mutations.

NON_SYN/Total	A	C	G	T	Average
A	0	0.816	0.759	0.838	0.810
C	0.759	0	0.807	0.587	0.745
G	0.679	0.848	0	0.848	0.815
T	0.790	0.690	0.824	0	0.780
Average	0.751	0.803	0.800	0.802	0.791

Table 1 Predicted non-synonymous mutations over total mutations divided by SNV type. Read by row, this table describes the effect of SNVs on the non-synonymous over total mutations ratio using the number of non-synonymous and synonymous changes per codon as described in **Figure 7**. This table refers to the effect of random mutations on an entire reference exome. It can be seen as a way to describe the dangerousness of a specific base change. For example, a C>T transition only leads to a non-synonymous SNV in 59% of the cases while a G>T transversion in 85% of the cases.

The *79 rule* derives from the average value of the weighted effects of all base substitutions (Table 1). For example, if we want to calculate the effect of the transversion A>C on the non-synonymous/total ratio ($\mathrm{NSY/total}_{A>C}$), we will have

$$\mathrm{NSY/total}_{A>C} = \frac{\overrightarrow{\mathrm{nsy}_{A>C}} \times \vec{W}}{(\overrightarrow{\mathrm{nsy}_{A>C}} + \overrightarrow{\mathrm{sy}_{A>C}}) \times \vec{W}}$$

where $\overrightarrow{\mathrm{nsy}_{A>C}}$ is the ordered non-synonymous variations vector (64x1) that a transversion A>C can cause, weighted for the human codon usage vector $\vec{W}$ divided by the total amount of A>C transversions that can be found on the 64 codons ($\overrightarrow{\mathrm{nsy}_{A>C}} + \overrightarrow{\mathrm{sy}_{A>C}} = \overrightarrow{\mathrm{total}_{A>C}}$) weighted for the same codon usage.

If we apply the same calculation to tumor sample datasets, like those provided by the TCGA, the results are surprisingly coherent with this simple probabilistic rule. The average ratio between non-synonymous and total mutations across patients for every tumor type spans between 0.74 and 0.81, suggesting that the mutational process is almost always random and therefore the large majority of mutations are passengers (Figure 8).

BLCA	0.7553825
BRCA	0.7951036
CESC	0.7354076
COADREAD	0.7473377
GBM	0.7537563
HNSC	0.7593664
KICH	0.7549399
KIRC	0.7653805
KIRP	0.7883589
LAML	0.8097577
LGG	0.7790602
LUAD	0.7824696
LUSC	0.7632863
OV	0.7926087
PAAD	0.8190069
PRAD	0.7780205
STAD	0.7658313
THCA	0.7987423
UCEC	0.7763582

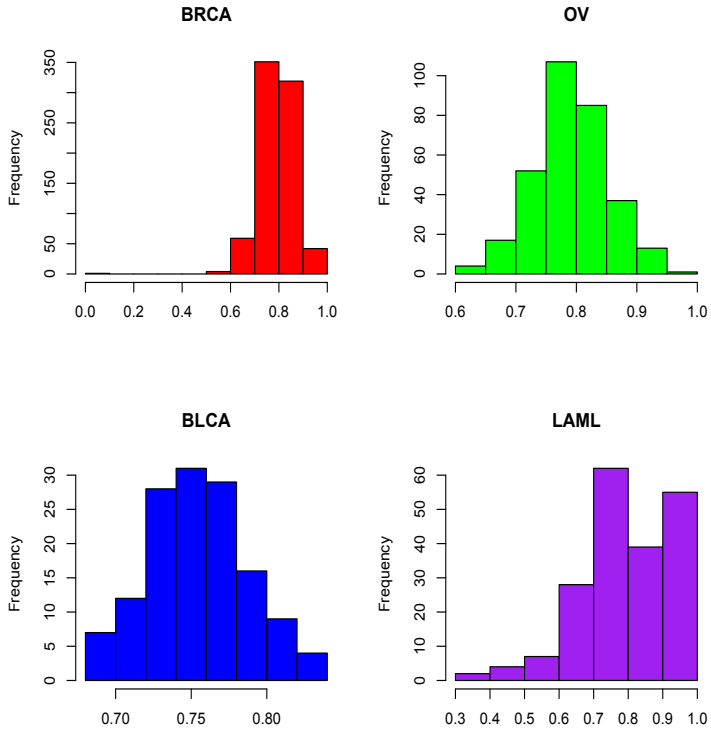

Figure 8 Non-synonymous mutations over total mutations distributions: the "*79 rule*". The 79 rule (see Additional File 1: Text S2.e) states that under the hypothesis of a random mutational process, the 79% of the SNVs lead to non–synonymous variations. If we look at the real data, this random process is still valid on average, giving another confirmation that the majority of mutations are passengers and are not under selective pressure

Each mRNA is composed by a distinctive percentage of codons that can vary significantly depending on the gene and can be completely different from the entire human codon usage. In addition, not all the types of SNVs have the same probability to be found. Transitions tends to happen more frequently and are generally less damaging compared to transversions (e.g. 2 out of 3 SNPs are transitions (Collins and Jukes, 1994)).

Moreover, the relative number of transitions and transversions in a sample are tumor dependent (Rubin and Green, 2009). For example, C>T transitions caused by misrepair of ultraviolet-induced covalent bonds between adjacent pyrimidines are frequent in melanoma, whereas C>A transversions caused by exposition to polycyclic aromatic hydrocarbons in tobacco smoke, characterize lung cancer (Lawrence et al., 2013). We therefore generalized the above formula for every gene-SNV couple:

$$\mathrm{NSY/total}^{g}_{i>j} = \frac{\overrightarrow{\mathrm{nsy}_{i>j}} \times \overrightarrow{\mathrm{w}_g}}{(\overrightarrow{\mathrm{nsy}_{i>j}} + \overrightarrow{\mathrm{sy}_{i>j}}) \times \overrightarrow{\mathrm{w}_g}}$$

where $i > j$ represents the SNV i to j with $i, j \in (A, C, G, T)$ and $i \neq j$, while $\overrightarrow{w_g}$ is the codon usage of the gene g.

4.1.4.6 DOTS-Finder step by step

Two main steps follow a preliminary analysis in DOTS-Finder: a functional assessment procedure and a statistical confirmation procedure. In the former, we identify a particular mutational pattern behavior that can be classified as "Oncogene", "Tumor Suppressor" or sometimes both. In the latter the two lists of possible oncogenes and tumor suppressors undergo 4 tests to assess their statistical probability of being true driver mutations.

1. Preliminary Step
    - Reannotation
    - Filtering
    - Descriptive Statistics
2. Functional Step
    - OG-Score
    - TSG-Score
3. Frequentist Step

3.1 Test 1: Higher Frequency Test

3.2 Test 2: Non-synonymous *versus* Synonymous Ratio Test

3.3 Test 3: Tumor-specificity Test

3.4 Test 4: Functional Impact Test

1. Preliminary Step

Before entering in the main DOTS-finder procedure, the MAF file is reannotated according to the refGene database and a few measures such as CG content, gene length,

number of amino acids and superfamily domains composition are added. This step is necessary to let every database coherently communicate to the others *via* the same annotation.

The tool automatically cuts the non-protein coding genes based on HUGO gene name database (19094 genes) (Gray et al., 2013) and discards all the mutations in non-coding regions like RNA mutations, intergenic mutations (IGR) and intron mutations (Intron). The user can change this setting *via* command options.

2. Functional Step

To calculate the **OG-S** we need the genomic coordinates of the missense mutations and the functional impact of the mutations according to Mutation Assessor. We associate the respective functional impact to every SNV and we assign to the Inframe InDels the average functional impact for that position (no score is provided for InDels in the database). The mutations are then mapped on the gene length and weighted by their impact. The discrete distribution of the mutations is smoothed with a Gaussian kernel estimation using a bandwidth that follows the Silverman's rule of thumb (Silverman, 1986). Thus, mutations that map close in the protein sequence increase the probability density function (PDF), creating a mutational hotspot with a higher density than the sum of the single-base discrete probabilities. The probability that the mutational profile has not arisen from non-selected passenger mutations is given by the comparison of the Shannon entropy index built on experimental data (X_m^g) compared with a bootstrapped uniform random distribution with the same numerosity (U_m^g) divided by the bootstrap interquartile range (bootIQR):

$$OGS_g = \frac{H(X_m^g) - \text{BootMedian}(H(U_m^g))}{\text{BootIQR}(H(U_m^g))}$$

where $H(X_m^g)$ is the sample entropy calculated on gene g with m missense mutations and $H(U_m^g)$ is the entropy of a uniform random sample of size m on gene g. The **OG-S** is therefore a modified Z-score, used to obtain robust bootstrap results even with small m.

The **TSG-S** reveals the characteristics of the driver genes that have diffuse truncating mutations in a non-specific pattern. To detect this particular pattern a large portion of all the mutations found on the gene must be truncating. The **TSG-S** is calculated as the $-\log_{10}(p-value)$ of a one-tail binomial test $(H_1: p > p_0)$ where the number of successes t_g is the number of truncating mutations on gene g and the number of trials n_g is the total number of mutations found on the gene. This ratio $(p = \frac{t_g}{n_g})$ is compared with a p_0 calculated as:

$$p_0 = \text{mean}(\frac{T_i^g}{N_i^g})$$

where T_i^g and N_i^g represent, respectively, the number of truncating mutations and the total number of mutations in patient i where gene g is mutated.

We can define the **TSG-S** for a gene g as:

$$TSGS_g = P(X \geq x | H_0) = \sum_{k=t_g}^{n_g} \binom{n_g}{k} {p_0}^k (1 - p_0)^{n_g - k}$$

3. Frequentist Step

The genes that pass at least one of the two thresholds (**OG-S** or **TSG-S**) are divided in the respective candidate categories (oncogene, tumor suppressor or both). Four statistical tests are run for these genes with specific modifications according to the categories they belong to. The four p-values obtained from the tests are pooled together using the Stouffer's method (Stouffer S et al., 1949) with a pattern of weights that take into account both the dependencies between tests and their relative importance in the driver definition. These suggested weights are set in order to take advantage of the full information provided by the four tests, but they can also be user-defined. The result is

finally adjusted using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995).

3.1 TEST 1: Higher Frequency Test

This test compares the rate of non-synonymous mutations per Mb in each gene with the rate of mutations in the patients carrying a mutation in that gene. The alternative hypothesis to reject the equality of these two proportions is

$$\frac{\mathrm{nsy}_g^t}{l_g * S_t} > \frac{\underline{\mathrm{NSY}_s^t}}{\text{exome length}}$$

Where nsy_g^t represents the number of non-synonymous mutations found on gene g and tumor t, l_g is the length of the gene in Mb, **S** is the total number of samples in tumor t and $\underline{\mathrm{NSY}_s^t}$ is the average number of non-synonymous mutations found in the patients with a mutation in gene **g**. This number is divided by the number of base pairs of an average exome sequencing (30Mb). Because of the low probability of mutation per Mb (from 0.1/Mb in AML to a maximum of 100/Mb in melanoma) a Poisson single tail test is run to assess if the rate of mutation of the gene is higher than the average mutation rate among the patients. This test is the same for both the TS and the OG groups. We apply a weight equal to 0.5 in the Stouffer's method because of the major relevance of this rate both in the literature (Dees et al., 2012; Lawrence et al., 2013; Wood et al., 2007) and for research/clinical purposes.

3.2 Test 2: Non-synonymous versus Synonymous Ratio Test

3.2.1 TEST 2 - OG : Non-synonymous versus Synonymous Ratio Test for Oncogenes

This test verifies if the rate between non-synonymous mutations and synonymous mutations is significantly high in the gene. To avoid zero division errors (some genes do not show synonymous mutations), the proposed test is based on the equivalent non-synonymous/total ratio. The rate of comparison is calculated on the expected ratio

obtained by randomly placing the same number and kind of mutations on the specific codon usage structure of the gene.

Since the effect of an InDel cannot be predicted in this way, we assume it will always produce a non-synonymous effect. So the total amount of mutations on gene g and tumor t is divided in

$$M_g^t = SNV_g^t + indel_g^t$$

the SNV_g^t are divided by their respective base substitution (A>C , G>T etc.) and put in the vector $\overrightarrow{bs_g^t}$ (12x1). We operate a vector product between $\overrightarrow{bs_g^t}$ and $\overrightarrow{NSY/total_g}$ calculated in our database in order to obtain the expected non-synonymous/total ratio in the SNVs. To obtain the final expected ratio we simply add the InDels we have subtracted before

$$expected\left(\frac{NSY^t}{total_g}\right) = \frac{SNV_g^t \cdot \left(\overrightarrow{bs_g^t} \times \overrightarrow{\frac{NSY}{total_g}}\right) + indel_g^t}{total_g^t}$$

Finally, we try to evaluate

$$\frac{nsy_g^t}{total_g^t} > expected\left(\frac{NSY^t}{total_g}\right)$$

using a one-tail binomial test.

3.2.2 TEST 2 - TSG : Non-synonymous versus Synonymous Ratio Test for Tumor Suppressor Genes

This test assesses if the rate between non-synonymous mutations and synonymous mutations in the gene is higher than the average rate in the patients who present the same mutation. We evaluate if

$$\frac{nsy_g^t}{total_g^t} > mean\left(\frac{NSY_s^t}{TOTAL_S^t}\right)$$

where nsy_g^t and $total_g^t$ represent, respectively, the number of non-synonymous mutations and the number of synonymous plus non-synonymous mutations found in

the gene g in tumor t while $\text{mean}(\frac{\text{NSY}_S^t}{\text{TOTAL}_S^t})$ is the average ratio calculated from all the samples with a mutation in the same gene. A one-tail binomial test is run in order to verify this inequality. This test is less precise than the previous one since the calculation of the null hypothesis ratio is made from a sample evaluation. Nevertheless this method of calculation of the null hypothesis is better for tumor suppressor candidates, since tumor suppressors are prone to have InDels and splice mutations that cannot be inserted in a probabilistic environment as we did for SNVs (the large majority of missense mutations are single spot mutations).

For TEST 2, we apply a weight of 0.2 in the Stouffer's method as this test has a lower statistical power and is linked to TEST 1; in fact, the total number of mutations depends on the sample size and the tumor specific mutation rate.

3.3 TEST 3: Tumor-specificity Test

This test verifies if the frequency of non-synonymous mutations in a particular tumor or situation is high compared with the general frequency found in COSMIC database.

Again, we evaluate if

$$\frac{\text{nsy}_g^t}{s^t} > F_g$$

where nsy_g^t represents the number of non-synonymous mutations found in the gene g in tumor t, s^t is the total number of patients/samples in tumor t, and F_g is the frequency of mutation across tumor types provided by the COSMIC database, by running a one-tail binomial test. However, we only apply a weight of 0.1 in the Stouffer's method, as this test is just used for ranking purposes in the chosen dataset. While we consider tumor specificity an important driver characteristic, we do not believe that not being tumor specific should be penalizing. For example, genes like *TP53* or *KRAS* should be considered important driver even in tumors where they are not frequently mutated.

3.4 Test 4: Functional Impact Test

3.4.1 TEST 4 - TSG: Functional Impact Test for TSGs

For every mutation in the gene we matched the respective patient it belongs to. We then compared the functional impact score of each mutation with the average score of all the other mutations in the patient. This test is used to assess if the distribution of the impact scores on the gene is stochastically higher than the average distribution.

We evaluate if

$$\text{mean}(\text{FI}_i^g) > mean(\text{FI}_i)$$

where FI_i^g represents the average functional impact on gene g in patient i while FI_i is the average functional impact in patient i without considering gene g. A Wilcoxon one-tail test for paired data was used to assess this inequality. Since no impact score is provided for truncating type mutations and silent mutations, in this work we applied the maximum score provided by Mutation Assessor to the first group (6) and the minimum to the silent mutations (0).

3.4.1 TEST 4 - OG: Functional Impact Test for Oncogenes

This test is like the above with the exception that since an oncogene is characterized by a majority of missense mutations, it is necessary to exclude all the truncating mutations from the calculation of the mean impact score, both at gene level and patient level.

As for the non-synonymous versus synonymous ratio test, an adequate sample size is fundamental for reaching a sufficient statistical power. The functional impact test is therefore weighted 0.2 in the Stouffer's method, the same as for TEST 2.

4.1.4.6.1 Setting the threshold for TSG-S and OG-S

The evaluation of our scores in classifying genes as driver or non-driver was set on the large database of COSMIC, using as positive control the genes of CGC (Futreal et al., 2004). We carry out the analysis by calculating and maximizing the Matthews phi curves

for the two scores against the list of true drivers (Figure 6). The **TSG-S** curve is maximized at $-\log_{10}(\mathrm{p}-\mathrm{value})$ of 12 (p-value=10^{-12}) reaching a Matthew's phi of 0.4 for the positive control of CGC genes that encompass somatic point mutations. The **OG-S** curve is instead maximized at $\mathrm{Entropy\ Z}-\mathrm{Score}$ of 49 reaching a Matthew's phi of 0.35. The calculation of the threshold cannot be directly applied to smaller tumor specific datasets because it is based on a huge amount of data provided by COSMIC (more than 7000 samples) and our scores are number-of-mutations dependent. In particular, the **OG-S** decreases if it is calculated on few mutations because the interquartile range of the uniform tends to increase by bootstrapping small samples. Similarly, the **TSG-S** enhances its statistical power with the increase in the number of trials (i.e. the number of mutations). We therefore derived a TSG coefficient and an OG coefficient that are calculated as

$$\mathrm{TSG}_{\mathrm{coefficient}} = \frac{\mathrm{COSMIC\ threshold\ for\ TSG}}{\mathrm{COSMIC\ mean\ number\ of\ truncating\ per\ gene}}$$

$$\mathrm{OG}_{\mathrm{coefficient}} = \frac{\mathrm{COSMIC\ threshold\ for\ OG}}{\mathrm{COSMIC\ mean\ number\ of\ missense\ per\ gene}}$$

These coefficients are multiplied for the mean number of truncating and missense mutations per gene in the single dataset during analysis in order to set specific tumor type thresholds. The mean number of missense and truncating mutations per gene is a way to aggregate both the information on the sample size (number of patients) and the mutation rate of the tumor type (number of mutations per patient).

We set a lower bound for these thresholds: 1 for **TSG-S** (p-value=10^{-1}) and 1 for **OG-S** (distance from the median uniform entropy of at least 1 interquartile range). For the **OG-S**, we also put an upper bound for this threshold at 3.5 as suggested in outlier analysis for the modified z-scores (Walfish, 2006).

4.1.5 Results

4.1.5.1 Application of DOTS-Finder to individual cancer types

We applied our methodology to 34 different cancer types (Colorectal Adenocarcinoma, Lung Squamous Cell Carcinoma, Uterin Carcinoma, Ovarian Adenocarcinoma, Lung Adenocarcinoma, Prostate Adenocarcinoma, Pancreatic Adenocarcinoma, Glioblastoma, Kidney Clear Cell Carcinoma, Kidney Papillary Cell Carcinoma, Kidney Chromophobe, Skin Melanoma, Low Grade Glioma, Esophageal Adenocarcinoma, Medulloblastoma, Stomach Adenocarcinoma, Head and neck squamous cell carcinoma, Oligodendroglioma, Acute Lymphoblastic Leukemia (ALL), Chronic Lymphocytic Leukemia (CLL), Soft Tissue Sarcoma, Lymphoma B-cell, Biliary Tract, Astrocytoma, Neuroblastoma, Liver Hepatocellular Carcinoma, Lung Small Cell Carcinoma, Rhabdoid_Tumor, Multiple Mieloma, Carcinoid, Breast Cancer, Thyroid Carcinoma, Acute Myeloid Leukemia (AML), Bladder Carcinoma). We analyzed the overall output in section **4.1.5.2.** In this section, we show the existence of a great variability among the different tumor types in terms of driver genes. In Table 2**,** we present the results of four cancer types: Breast carcinoma (BRCA) and Thyroid Carcinoma (THCA), described in sections **4.1.5.3** and **4.1.5.4**, and Acute Myeloid Leukemia (AML) and Bladder Carcinoma (BLCA), described in sections **4.1.5.5** and **4.1.5.6**. The rest of the DOTS-Finder results can be seen in Appendix Table 1. We also compared the DOTS-finder output with the output of the following methods: i) the main TCGA publications (when available); ii) TUSON Explorer (Davoli et al., 2013) (considering all the genes with a q-value <=0.1); iii) MuSiC (Tamborero et al., 2013b)(used for identifying significantly mutated genes in 12 cancer types); iv) MutSig (Lawrence et al., 2014) (used for identifying significantly mutated genes in 21 tumor types). Thus, we used the state-of-the-art results from official TCGA publications and from the latest release of the applications described above. We were not able to use exactly the same input data of all

the publications, since TUSON Explorer and MutSig (as used in (Lawrence et al., 2014)) are unavailable. Our results show that DOTS-Finder can identify known cancer genes involved in each tumor, confirm new discoveries reported by other groups, and detect novel driver gene candidates which are mutated at low frequency and not identified by other methods Appendix Table 2.

Acute Myeloid Leukemia		
S = 196		
MNSp = 11		
Gene name	**NS freq**	**q-value**
TSG		
CEBPA	0.066	0
NPM1	0.276	0
RUNX1	0.092	0
TET2	0.087	0
TP53	0.077	0
WT1	0.061	0
RAD21	0.026	3.27E-06
PHF6	0.031	3.40E-06
STAG2	0.031	1.38E-05
EZH2	0.015	0.0007
ASXL1	0.026	0.0014
HNRNPK	0.010	0.0083
CALR	0.010	0.0142
CBFB	0.010	0.0572
CBX7	0.005	0.0948
BCOR	0.010	0.1971
ONCOGENE		
CEBPA	0.066	0
DNMT3A	0.260	0
FLT3	0.270	0
IDH1	0.097	0
IDH2	0.102	0
NRAS	0.077	0
TP53	0.077	0
U2AF1	0.041	0

Thyroid Carcinoma		
S = 326		
MNSp = 19		
Gene name	**NS freq**	**q-value**
TSG		
TG	0.049	8.00E-10
EMG1	0.018	5.30E-08
RPTN	0.025	9.05E-06
PPM1D	0.015	0.0054
TMCO2	0.009	0.0056
IL32	0.009	0.0152
DNMT3A	0.015	0.2896
ONCOGENE		
BRAF	0.561	0
HRAS	0.037	0
NRAS	0.080	0
TG	0.049	3.47E-08
DNASE2	0.009	0.0694
PRDM9	0.018	0.0816
DICER1	0.009	0.1070
ZNF845	0.018	0.1070
PRG4	0.012	0.1085
PTTG1IP	0.012	0.1085

Breast Cancer		
S = 1046		
MNSp = 36		
Gene name	**NS freq**	**q-value**
TSG		
CBFB	0.021	0
CDH1	0.062	0
GATA3	0.095	0
MAP2K4	0.039	0
MAP3K1	0.070	0
PTEN	0.040	0
TP53	0.338	0
TBX3	0.022	1.11E-12
MLL3	0.065	5.93E-12
AOAH	0.019	3.98E-10
CTCF	0.021	7.90E-10
RUNX1	0.024	3.19E-06
NCOR1	0.038	3.97E-06
RB1	0.021	6.09E-06
NCOR2	0.032	0.0003
STXBP2	0.010	0.0004
AQP7	0.008	0.0017
ZFP36L1	0.012	0.0046
RBMX	0.012	0.0056
GPS2	0.007	0.0095
CASP8	0.015	0.0104
CDKN1B	0.008	0.0125
UBC	0.008	0.0155
MED23	0.013	0.0224
MYB	0.012	0.0407
CCDC144NL	0.008	0.1268
GNRH2	0.003	0.2062
HNF1A	0.009	0.7280
ONCOGENE		
AKT1	0.022	0
PIK3CA	0.285	0
TP53	0.338	0
TBX3	0.022	9.01E-10
SF3B1	0.017	3.36E-08

Bladder Carcinoma		
S = 145		
MNSp = 177		
Gene name	**NS freq**	**q-value**
TSG		
ARID1A	0.241	0
CDKN1A	0.145	0
KDM6A	0.214	0
TP53	0.262	0
ELF3	0.076	1.18E-10
MLL2	0.262	1.18E-10
EP300	0.152	3.03E-09
RB1	0.110	2.26E-08
SPTAN1	0.097	3.03E-06
MLL3	0.200	6.14E-06
CREBBP	0.131	1.16E-05
STAG2	0.090	7.55E-05
FOXQ1	0.048	0.0060
TXNIP	0.055	0.0079
FAT1	0.110	0.0370
FBXW7	0.069	0.0428
GCC2	0.069	0.0800
ZNF513	0.055	0.0911
KLF5	0.062	0.1184
GPS2	0.028	0.2599
NHLRC1	0.021	0.2635
ONCOGENE		
TP53	0.262	0
NFE2L2	0.076	6.08E-06
ERBB3	0.117	1.08E-05
RARG	0.069	1.53E-05
IRS4	0.014	0.6550
ELP5	0.014	0.6550
RPS6	0.021	0.6550

FOXA1	0.017	7.73E-05
HIST1H3B	0.008	0.0001
MEF2A	0.014	0.0002
PIK3R1	0.025	0.0008
ATN1	0.017	0.0425
AKD1	0.018	0.0431

Table 2 Significantly mutated genes identified by DOTS-Finder in four cancer types. Legend: S = Number of Samples , MNSp = Median number of Non-Silent mutations *per* patient , NS freq = Non-synonymous mutations frequency among samples , Underlined genes are near significance

4.1.5.2 Driver genes and tissue specificity

We used DOTS-Finder on samples from 34 tumor types and identified a total of 301 driver genes Table 2 and Appendix Table 1. Only 57 out of 301 genes were found in more than one tumor type, and most of the 25 genes present in at least three tumor types are well-known cancer driver genes (i.e. *TP53, PTEN, RB1, NRAS, IDH1, SF3B1, CTNNB1, BRAF, ARID1A, NFE2L2, MLL3, KRAS, KDM6A, CDKN2A, STAG2, SMARCA4, SMAD4, PIK3R1, PIK3CA, MLL2, IL32, CREBBP, CDKN1B, NPAP1, B2M*). Interestingly, genes found only in two different cancer types maintain tissue specificity, like, for example, *ATRX*, mutated only in low-grade glioma and in glioblastoma, probably being an important driver gene in tumors of the central nervous system. In addition, 244 genes displayed cancer specific patterns, being mutated in a single cancer type. Thus, the majority of tumor suppressor genes (TSGs) and oncogenes (OGs) are tissue-specific. For example, *NKX3-1* and *AR* are found only in prostate adenocarcinoma, *OGG1* is specific for renal cell carcinoma and *NOX4* is specific for glioblastoma. In addition, we also found that about 54% of the genes in our list (163 out of 301) were not present in the 300 TSGs and 250 OGs identified by TUSON Explorer. For example, Thyroglobulin (*TG)*, a well-studied gene in thyroid cancer (Rubio and Medeiros-Neto, 2009), is absent. We hypothesize that many new driver genes that are infrequently mutated might be tissue specific. Thus, it is very important to analyze the mutation signatures of individual tumor types, especially of those cancer types for which large sample size is unavailable and which will not reach saturation in the next future.

4.1.5.3 Breast carcinoma

We applied DOTS-Finder to the list of 1046 Breast carcinoma (BRCA) samples. We found a poor overlap between the TCGA official publication (Koboldt et al., 2012) and our results (Figure 9, Panel A), but all the known cancer genes for this tumor type are retained, while our results do not encompass any notorious "fishy gene" like *RYR2* or *OR6A2* (Lawrence et al., 2013), which are instead present in the TCGA publication. The TCGA publication also misses known breast cancer associated genes, like *FOXA1* (Robinson et al., 2013) and *CASP8* (Catucci et al., 2011). We identified 3 new driver candidates, not present in previous publications: *AQP7*, *MEF2A* and *UBC*. *AQP7* encodes the aquaporin 7, an integral-membrane protein that plays important roles in water and fluid transport and cell migration. Recent discoveries of *AQP* involvement in cell migration and proliferation suggest that *AQPs* play key roles in tumor biology (Verkman et al., 2008). *MEF2A* encodes a DNA-binding transcription factor that is involved in several cellular processes, including cell growth control and apoptosis. It was recently shown that NOTCH-MEF2 synergy may be significant for modulating human mammary oncogenesis (Pallavi et al., 2012). *UBC* is a member of the ubiquitin family and involved in cell cycle and DNA repair. The role of ubiquitination is well stablished in

cancer, especially in breast (Ohta and Fukuda, 2004).

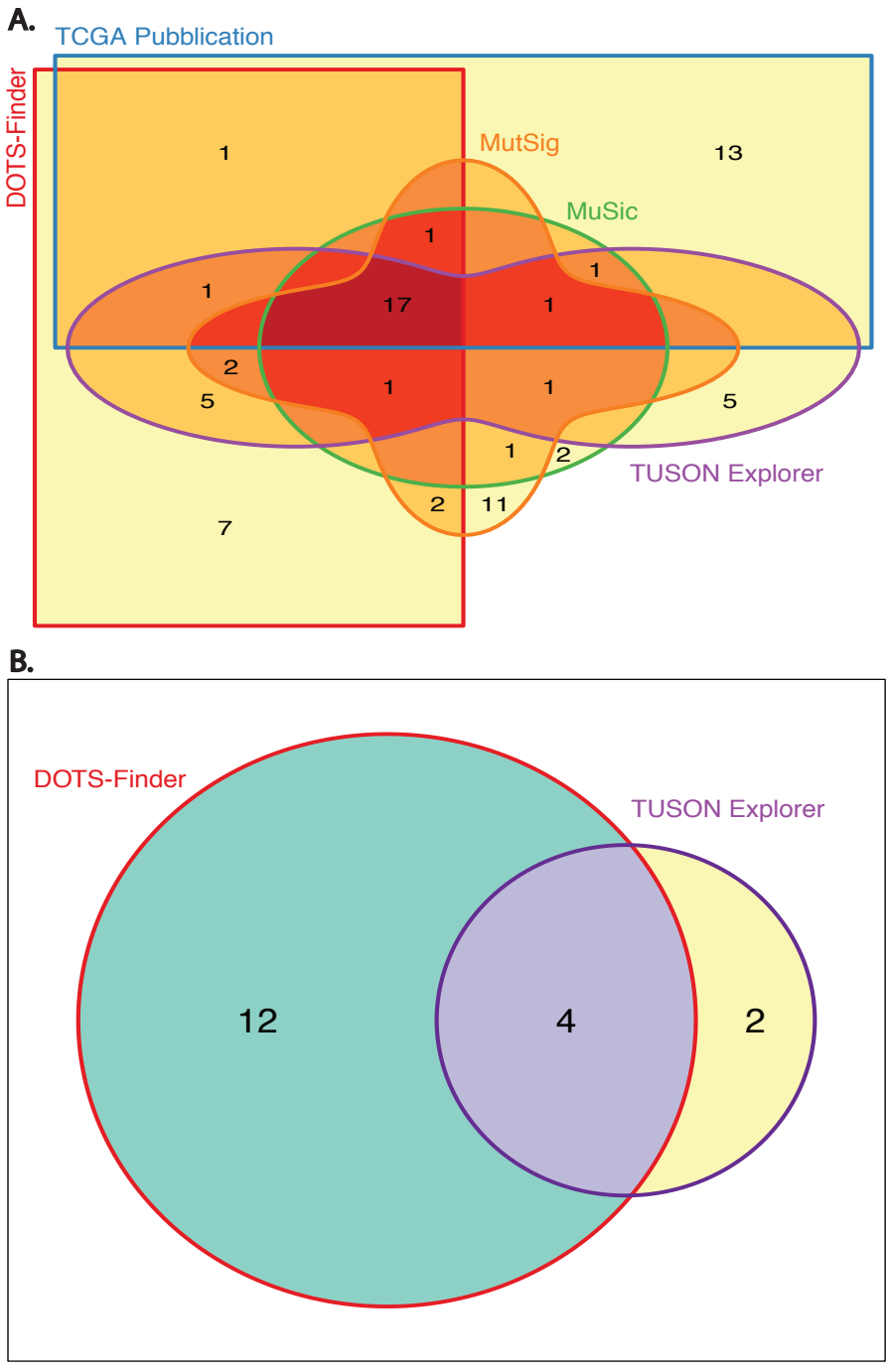

Figure 9 Comparative driver gene predictions in Breast Cancer and Thyroid Cancer. (A) The candidate driver genes predicted by DOTS-Finder in BRCA are compared against 4 previously reported predictions: MuSiC, MutSig, TUSON Explorer and the TCGA publication. The five-set Venn diagram shows the number of predicted genes that are in common between the different analyses and those uniquely predicted by each of them. The line delimiting each set and the name of the corresponding method are depicted in the same color. The diagram uses a graduated color ramp from light yellow to dark red to represent the overlap of an increasing number of tools that predict the same drivers. Although BRCA mutational landscape is highly heterogeneous among patients, all the methods agree on predicting the same 17 genes as drivers (darkest shade of red). In addition, DOTS-Finder is able to predict 7 genes that were never found by any method in BRCA. Also MutSig and TUSON Explorer retain unique predictions, respectively 11 and 5 possible driver candidates. This discrepancy is the reflection of the typical "mountains and hills" landscape of the BRCA genome, with few highly mutated genes (predicted by almost all the tools) and hundreds of low-frequency mutations (only identified by a specific tool). **(B)** Number of genes predicted by TUSON Explorer and DOTS-Finder in the THCA dataset. The former only

predicts a few driver genes (6); of these, two thirds are also identified by DOTS-Finder. Notably, our tool shows a much higher sensitivity than TUSON Explorer with 12 new predicted genes.

4.1.5.4 Thyroid Carcinoma

We applied DOTS-finder to the list of 326 Thyroid Carcinoma (THCA) samples from TCGA, identifying 12 driver genes. We could only compare the DOTS-Finder results with the results obtained by TUSON Explorer, since, to date, there are no published TCGA papers for Thyroid Carcinoma (Figure 9, Panel B). Three of our putative driver genes (*TG*, *BRAF* and *RPTN),* are also predicted by TUSON Explorer. *TG* and *BRAF* are known driver genes in THCA (Kimura et al., 2003; Rubio and Medeiros-Neto, 2009), while RPTN is a poorly characterized protein that has never been associated with THCA. We identified several putative driver genes that may have relevant functions in cancer development (Table 2): mutations in *EMG1* have been recently identified in a screen for mediators of *IGF-1* signaling in cancer (McMahon et al., 2010); germline mutations in *PRDM9* are thought to influence genomic instability, increasing the risk of acquiring genomic rearrangements associated with childhood leukemogenesis (Hussin et al., 2013); *PPM1D* is an important interactor of TP53, is amplified in different types of cancers and encodes wip1, a protein involved in oncogenesis (Bulavin et al., 2002). Recently, mutations and variants of this gene were associated with DNA damage response (Dudgeon et al., 2013). Although only slightly above our threshold, we also detected *PTTG1LP* and *DICER1* as putative OGs. Interestingly, pituitary tumor transforming gene (*PTTG*)-binding factor (*PTTG1IP*) encodes a poorly characterized proto-oncogene that has already been implicated in the etiology of thyroid tumors (Read et al., 2011; Stratford et al., 2005). Loss of *DICER1* is associated with the development of many cancers; somatic missense mutations affecting *DICER1* are common in non-epithelial ovarian tumors and these mutations show an oncogenic behavior (Heravi-Moussavi et al., 2012).

4.1.5.5 Acute Myeloid Leukemia

We applied DOTS-Finder to the 196 samples in TCGA Acute Myeloid Leukemia (AML) dataset and we were able to confirm the large majority of findings from previously reported analyses (Kandoth et al., 2013; Lawrence et al., 2014; Network, 2013) (Figure 10) and to discover three new driver candidates, as shown in Table 2. Unfortunately, we could not compare our results with TUSON Explorer, as AML samples were not analyzed. In particular, we identified as driver three genes with low mutations frequency (<=1%): *CBFC*, *CBX7* and *CALR*. *CBFC* and *CBX7* have been already implicated in AML pathogenesis. *CBFC* is the most common translocation target in AML, involved in a chromosomal rearrangement that results in the fusion of *CBFB* and *MYH11* genes, associated with the acute myeloid leukemia subtype M4Eo (Kundu and Liu, 2001). *CBFB* has a role in hematopoiesis (Kundu et al., 2002) and it is a direct target of *RUNX1* (Hart and Foroni, 2002), a well-known driver gene of AML. *CBX7* is a component of the Polycomb repressive complex 1 and it is causally linked to cancer development (Klauke et al., 2013). Interestingly, we classified this gene as a tumor suppressor, and this finding is consistent with the fact that loss of *CBX7* gene expression correlates with a highly malignant phenotype in thyroid cancer (Pallante et al., 2008) and reduces survival of colorectal cancer patients (Pallante et al., 2010) *CBX7* is specifically expressed in hematopoietic stem cells and its overexpression enhances self-renewal and can induce leukemia (Scott et al., 2007). CALR was recently found mutated in some forms of myeloproliferative neoplasms, a group of disorders related to AML (Klampfl et al., 2013). Although near significance, we also detected *BCOR*, a transcriptional corepressor. *BCOR* mutations are implicated in myelodysplastic syndromes and AML with normal karyotype (Grossmann et al., 2011). In addition, *BCOR* has been recently found in acute promyelocytic leukemia as a novel fusion partner of *RARA* (Yamamoto et al., 2010).

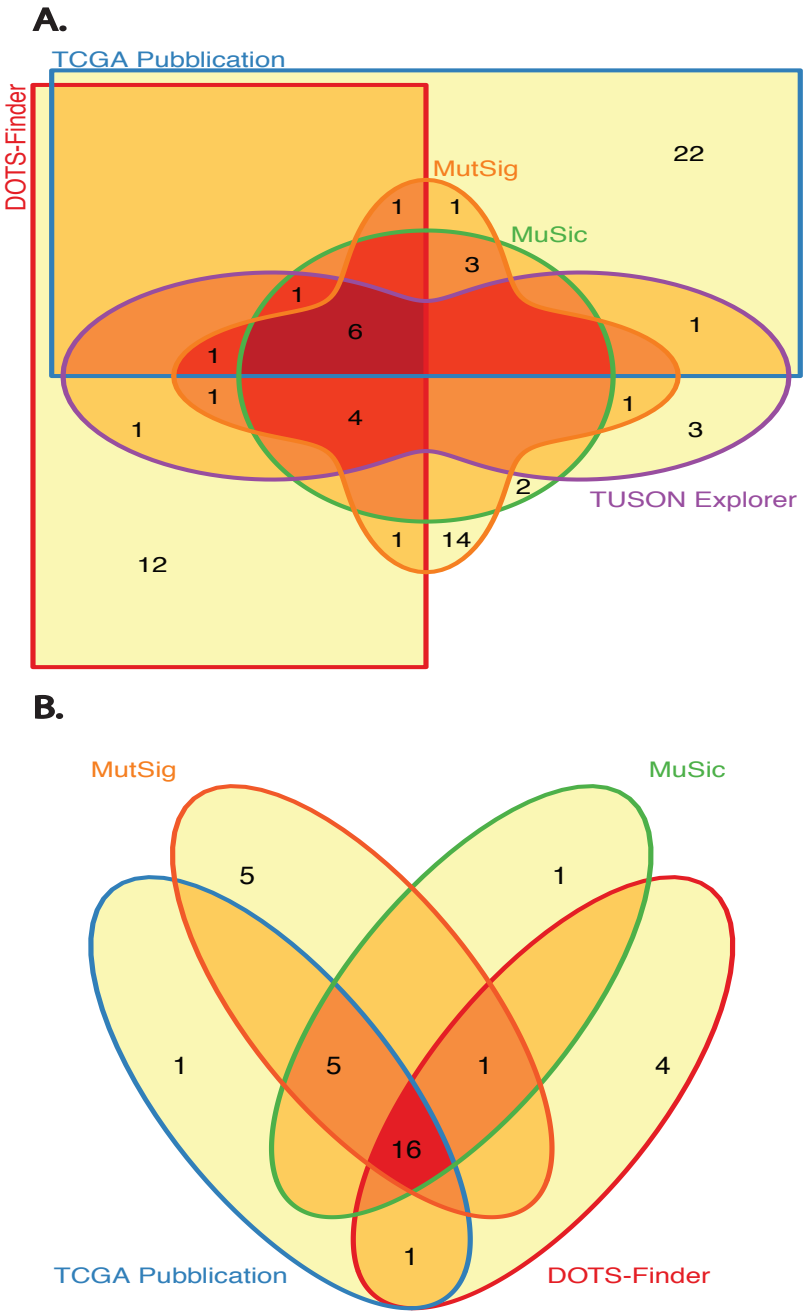

Figure 10 Venn diagrams of the set of candidate driver genes predicted by several tools in Bladder Carcinoma and Acute Myeloid Leukemia. (A) Comparison of driver genes predicted by five methods in Bladder Carcinoma. Only 8% of all the genes identified by at least one resource are identified by all methods. This percentage rises to 20% if we exclude the candidate driver genes coming from the TCGA publication. Nevertheless, there is a poor concordance among the methods as MutSig and DOTS-Finder identifying respectively 14 and 12 non-overlapping candidate drivers. **(B)** Comparison of driver genes predicted by four methods in Acute Myeloid Leukemia. The AML mutational spectrum has 50% of the genes shared by all the four resources analyzed. Nevertheless, DOTS-Finder was able to identify the following new driver candidates: *CBFC*, *CBX7*, *CALR and BCOR*.

4.1.5.6 Bladder Carcinoma

We applied DOTS-Finder to the list of 145 Bladder Carcinoma (BLCA) samples. We have identified 21 driver genes, of which 6 are also found in the official TCGA paper

(Guo et al., 2013) but their prediction is not properly comparable with our findings as it contains only 99 samples. Our results are instead consistent with MuSiC, MutSig and TUSON Explorer as shown in Appendix Table 2. Five driver genes were uniquely identified by DOTS-Finder and three of them (*SPTAN1*, *TXNIP*, *RARG*) have functions relevant to cancer development or have been previously associated with cancer. SPTAN1 encoded protein has been implicated in DNA repair and cell cycle regulation (Metral et al., 2009). *TXNIP* acts as a suppressor of tumor cell growth and loss of *TXNIP* expression facilitates BLCA. Notably *TXNIP* might be an important target for the prevention or treatment of bladder cancer (Nishizawa et al., 2011). Lastly, *RARG* encodes a retinoic acid receptor that acts as a ligand-dependent transcription factor that regulates cell growth and survival (Altucci et al., 2007). In addition, we also detected the following genes near significance: the known tumor suppressors *KLF5* and *GPS2* and the oncogenes *IRS4*, *RPS6* and *ELP5*. *KLF5* encodes a member of the Kruppel-like factor subfamily, which plays important roles in cell proliferation and cell cycle regulation (Chen et al., 2006) and it has been described as a tumor suppressor in several cancer types (Chen et al., 2003). Mutations in *GPS2* have been previously identified in medulloblastoma (Pugh et al., 2012). The insulin receptor substrate 4 (*IRS4*) and the Ribosomal Protein S (*RPS6*) may play a role in cancer development and progression *via* their effect on cell growth and proliferation. *ELP5* may play a role in cancer due to its involvement in histone acetyltransferase activity (Winkler et al., 2002).

4.1.5.7 Atypical tumor suppressor genes and oncogenes

The concept of TSG and OG has evolved over time. In conventional wisdom, TSGs are nonfunctional in tumors and require biallelic loss of function to manifest tumorigenicity (Payne and Kemp, 2005); OGs are typically characterized by acquired or enhanced function and a single mutated allele is sufficient (Xu et al., 2013). Thus, three levels of information are required to classify a cancer driver gene as an OG or a TSG: *functional,*

structural and *genetic*. The *functional* level is defined by a gain or loss of a biochemical function. It requires understanding of the actual role of the gene in tumorigenesis and of the pathways in which it is involved. Functional changes result from and can be predicted based on the *structural* information; this is what we ultimately do by dividing mutations into truncating (TSG related) or missense (OG related) ones and analyzing their pattern. The genetic effect defines the dominant or recessive characteristics of the driver gene. At the *genetic* level, a mutated gene can be dominant or recessive depending on how many dysfunctional copies are required to exert its effect (Table 3).

		FUNCTIONAL EFFECT	
		Gain	**Loss**
GENETIC EFFECT	**Dominant**	Typical OncoGene	Dominant Negative TSG
	Recessive	-	Typical TSG

Table 3 Genetic and Functional effect of mutations in oncogenes and tumor suppressors. A driver cancer gene is defined by a genetic effect (dominant, recessive) and a functional effect (gain or loss). These two components ultimately define the tumor suppressor and oncogene characteristics that we try to infer from the mutational landscape (structural effect)

Typically, the functional information is missing or poorly understood for new driver candidates and the genetic information (allelic-specific) is not directly available in cancer sequencing studies. Thus, the OG and TSG classification must be inferred from the structural level. It is not surprising that our tool can classify many genes as being both TSGs and OGs within the same cancer type, or even put them into different categories according to the tumor context. This apparent misclassification might cast a light on the particular behavior of some genes. There are four possible structural scenarios of mutations in a gene, as shown in Table 4. The first two are the same ones shown in Figure 2: a clustered missense mutation landscape with no truncating mutations, implying a typical gain-of-function OG like *KRAS*, and diffuse and predominant

truncating mutations with no missense pattern like *APC*, underlying a loss-of-function TSG.

		STRUCTURAL LANDSCAPE			
	Missense	clustered	no	clustered	Any
	Truncating	no	diffuse	diffuse	Clustered
BIOLOGICAL CLASSIFICATION	**Oncogene**	**Typical** (gain-of-function) e.g. KRAS	none found	none found	**Atypical** (gain of function through loss of inhibition). e.g. NPM1
	Tumor Suppressor	**Atypical** (dominant negative, gain-of-function) e.g. SMARCA4 in Lymphoma	**Typical** (loss-of-function) e.g. RB1	**Atypical** (possible dominant negative, gain-of-function*) e.g. TP53 in UCEC or DNMT3A in AML	none found

Table 4 Inference of biological classification by structural effect of mutational landscape. Inferring the biological role of oncogenes and TSGs in cancer *via* the mutational landscape can lead to borderline results in the classification. A careful confrontation with the literature can cast a light on the peculiar characteristics of driver genes in the different tumor types. *A mixed mutational landscape with diffuse truncating and clustered missense in the same tumor type must be carefully analyzed. We should understand whether truncating and missense mutations are mutually exclusive and what is the allelic status (heterozygosity or homozygosity) of the two different patterns.

In Figure 11**,** we present four genes with atypical patterns. *TP53* in endometrial carcinoma (Panel A) has a landscape of mutations that can be considered borderline for both the OG and TSG score definitions, with a consistent number of diffuse truncating mutations (around 20%) and a concentration of missense mutations on the DNA binding site. According to our tool, the duality of *TP53* is revealed in many tumor types and can mask a possible dominant negative effect, as summarized in (Oren and Rotter, 2010). Similarly, and strongly supported by the literature (Kim et al., 2013), *DNMT3A* in Panel B**,** presents diffuse truncating mutations and a visible missense cluster on the cytosine C5 DNA methylation domain. In both genes, a patient-specific mechanism, which can distinguish the two different patterns, is probably implicated. In Panel C, we analyze two different patterns of mutations in *SMARCA4* in different tumor types. Although considered a TSG (Medina et al., 2008), *SMARCA4* is classified as a true TSG only in lung adenocarcinoma, with 11 out of 18 truncating mutations diffuse all over the

gene body. In lymphoma, the situation is the opposite: none of the 6 mutations found is truncating, and 3 are clustered on amino acid 973 on the SNFN 2 domain of the protein. DOTS-Finder classifies this gene differently according to the tumor type, suggesting a dominant negative effect of *SMARCA4* which is able to regulate its own expression with just one mutated copy (Magnani and Cabot, 2009), as previously described for this cancer (Medina and Sanchez-Cespedes, 2008).

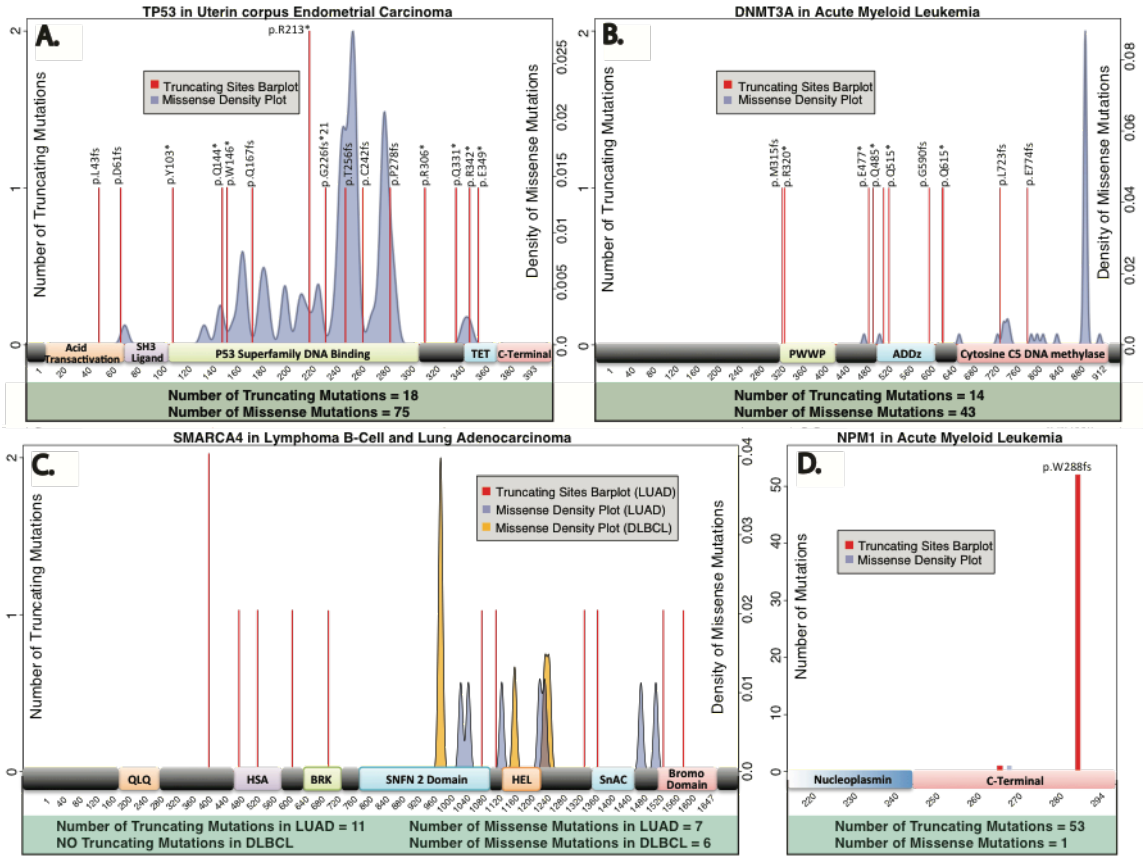

Figure 11 Mutational patterns of atypical tumor suppressor genes and oncogenes. (**A**) *TP53* mutational landscape in uterine corpus endometrial cancer. DOTS-Finder classifies this gene as a TSG as well as an oncogene. While this gene retains many truncating mutations, which are diffused all over the gene body, it also encompasses a high number of clusterized missense mutations affecting DNA binding. (**B**) *DNMT3A* mutational landscape in Acute Myeloid Leukemia. The pattern of mutations shows diffuse truncating mutations and an evident missense cluster on the cytosine C5 DNA methylation domain. The two types of mutations (truncating and missense) do not share the same domains. This pattern could reflect a double mechanism of action of this gene in different patients. (**C**) *SMARCA4* mutational landscape in lymphoma B-cell compared with lung adenocarcinoma. *SMARCA4* is reported in literature as a typical loss-of-function TSG and its mutational pattern in lung is consistent with this classification (diffuse truncating mutations). In lymphoma no truncating mutations are called, and half of the missense mutations affect amino acid 973. DOTS-Finder classifies *SMARCA4* as a TSG in lung but as an oncogene in lymphoma, following its clustered missense pattern. We suspect a possible dominant negative effect in this second example (see **Table 3**). (**D**) *NPM1* mutational landscape in acute myeloid leukemia. This gene is reported as a gain-of-function oncogene, however, it shows a peculiar mutational landscape: 99% of its mutations are truncating, but they are clustered on the c-terminal of amino acid 288. Mutation *p.W288fs* truncates the protein without deactivating it; NPM1 is instead delocalized from the nucleus to the cytoplasm. The total numbers of truncating sites and missense mutations are indicated in the panels. The mutations are

mapped on the corresponding canonical protein ideogram, therefore, not all the mutations can be represented (e.g. splice sites mutations are not included in the figure).

The last example in Panel D refers to *NPM1*, which is a shuttling protein involved in AML. Although *NPM1* is almost exclusively characterized by truncating mutations (53/54) and is classified as a TSG by DOTS-Finder, *NPM1* is instead a typical gain/switch-of-function gene (Mariano et al., 2006). The truncating mutations are, in fact, clustered as *p.W288fs*, a four base insertion that deactivates the c-terminal and delocalizes the protein (Grisendi et al., 2006).

4.1.5.8 The importance of considering subsets of samples

Analyzing the pattern of genetic alterations in tumor subsets classified by clinical or other biologic parameters can reveal important insight in individual pathogenic mechanisms and suggest possible therapeutic avenues. For instance, in lung adenocarcinoma (LUAD), about 25-30% of the cases are not attributable to tobacco smoking as they are found in people that have never smoked (never smokers - NS). Studies have revealed that LUAD in NS is a completely different disease from any type of lung cancer arising in smokers (LUAD included), as it differs in terms of clinical and pathological features, with diverse prognosis and strategy of care (Rudin et al., 2009). The difference in the mutational landscape (Govindan et al., 2012) supports the hypothesis that NS lung adenocarcinomas are driven by distinct genetic mechanisms. To identify additional driver genes with a role in the development of lung cancer in NS, we applied DOTS-Finder to the somatic mutations of the 50 NS patients present in the LUAD samples of the TCGA. These samples constitute approximately 10% of the population; our driver candidate predictions are reported in Appendix Table 3. At the top of the list of predicted OGs is *EGFR*, consistent with the fact that *EGFR* is a key oncogenic player in LUAD NS. Beside the identification of very well-known cancer genes such as *SMAD4*, *STK11*, *SETD2*, *MET*, *KEAP1*, *TP53* and *KRAS*, we also identified

several putative driver genes that might have relevant cancer development functions: somatic mutations in *GRM1* disrupt signaling with multiple downstream consequences (Esseltine et al., 2013); mutations in *RPL5* has been recently described as a potential oncogenic factor in T-cell acute lymphoblastic leukemia (De Keersmaecker et al., 2013); inactivating mutations in the *SHA* gene, which has a role as TSG, have been identified in familial paragangliomas (Bardella et al., 2011; Francis et al., 2013); *WRN* encodes a strand DNA breaks: defects in this gene are the cause of the aging-promoting Werner syndrome and copy number variations or epigenetic inactivation have been recently found in LUAD NS (Job et al., 2010) and non-small cell lung cancer (Agrelo et al., 2006) respectively.

Similarly, kidney cancer can be classified in different histological subtypes, the most common being Kidney Renal Clear Cell Carcinoma (KIRC), Papillary Cell Carcinoma (KIRP) and Kidney Chromophobe (KICH). Applying DOTS-Finder separately on each kidney dataset Appendix Table 1, we observed a subtype-specific pattern of genetic alterations. KIRC and KIRP share only *SETD2*, KIRC and KICH have only *TP53* in common, and there are no common driver genes between KIRP and KICH. By analyzing all the datasets together we can predict two new putative driver genes, *GFRAL* and *STAG2*, not appearing in the single analyses. Since the KIRC subset is predominant in terms of sample size, the aggregated analysis can recapitulate 69% of its genes, while it can only identify 50% of KICH and 27% of KIRP genes. In KIRP, we lose the following candidate driver genes, which then appear to be tumor specific: *KDM6A*, *SRCAP*, *SAV1*, *DARS*, *OGG1*, *MET*, *ATP10A;* similarly, in KICH we lose *CDKN1A*.

4.1.5.9 Small sample size analysis. The --lax option

DOTS-Finder sets the threshold for OG-S and TSG-S as a function of both the mutation rate of the analyzed tumor and the sample size of the input dataset (see section **4.1.4.6**). These thresholds have a default lower boundary. Nevertheless, for very small sample

sizes, these thresholds can still be too high to let genes pass the functional step. We decided to introduce an option called *--lax* that ignores the imposed lower boundary and allows more genes to pass the functional step in the presence of a small sample size. In Table 5 we show the analysis of two different tumors, the oligodendroglioma dataset (16 patients) and the carcinoid dataset (54 patients) obtained using the *--lax* option of DOTS-Finder.

	Oligodendroglioma			Carcinoid		
	Number of Patients=16 Median number of mutations per patient = 17.5			Number of Patients = 54 Median number of mutation per patient = 33		
	Gene name	NS freq	q-value	Gene name	NS freq	q-value
Default Option	FUBP1	0.125	0.05	CDKN1B	0.09259	0.00
	CIC	0.9375	0	PRDM9	0.05556	0.04
	IDH1	0.9375	0	CACNA1E	0.07407	0.05
Lax Option	CIC	0.9375	0	CDKN1B	0.09259	0.00
	FUBP1	0.125	0.17	PRDM9	0.05556	0.02
	CIC	0.9375	0	ATM	0.07407	0.03
	IDH1	0.9375	0	ERN2	0.03704	0.11
	NOTCH1	0.3125	0.00	MYBPC2	0.03704	0.18
	PIK3CA	0.25	0.00	MGAT2	0.03704	0.20
	PDCD6IP	0.125	0.02	TP53BP1	0.07407	0.10
	PKD1L2	0.125	0.02	NOP9	0.03704	0.22
	SLC26A3	0.125	0.02			
	FARP2	0.125	0.03			
	HIVEP2	0.125	0.04			
	KCNH6	0.125	0.04			
	RIN1	0.125	0.04			
	RNPEPL1	0.125	0.04			

LEGEND
TSG
oncogene

Table 5 Application of the *--lax* option to two small cancer datasets. Results of DOTS-Finder obtained analyzing Oligodendroglioma and Carcinoid datasets with default option and with the *--lax* option. The thresholds imposed by DOTS-Finder can be too high to let any driver candidate to pass the functional step. With small sample size or very low mutation rate tumors, an option called *--lax* can be used to make DOTS-Finder less stringent in the first step of the analysis. Legend: NS freq = Frequency of non-synonymous mutations among samples. Underlined genes are near significance

In the left column of Table 5 we present the result of the analysis of 16 exome sequencings from oligodendroglioma patients (Yip et al., 2012). Without the *--lax* option, DOTS-Finder recapitulates the knowledge regarding this rare brain tumor by identifying mutations in *CIC*, *IDH1* and *FUBP1* (Alentorn et al., 2012). The same dataset upon the *--lax* option reveals other possible driver candidates, like the known cancer genes *PIK3CA* and *NOTCH1*, the never reported *PDCD6IP*, a gene expressed in the nervous system and involved in cell death, *HIVEP2* and *KCNH6*, two genes previously reported in leukemia, and *RIN1*, an important Ras interactor.

In the right column instead are the results for the carcinoid tumor (Francis et al., 2013). *CDKN1B* has been already reported for this cancer type, but with the lax option on, other possible driver candidates have emerged. In particular, the known cancer-associated gene *ATM*, *TP53BP1*, an enhancer of *TP53* activation known to be involved in DNA damage response, *PRDM9*, described in the main text, in thyroid cancer and near significance, and *ERN2*, a pro-apoptotic gene involved in translational repression under endoplasmic reticulum stress.

4.1.5.10 Comparison of DOTS-Finder to existing tools using Pan-Cancer12 data

We compared the candidate driver genes predicted by DOTS-Finder against the predictions made by 5 methods: 1) MuSiC (Dees et al., 2012), 2) MutSig (Lawrence et al., 2013), 3) OncodriveFM (Gonzalez-Perez and Lopez-Bigas, 2012), 4) OncodriveCLUST (Tamborero et al., 2013a) and 5) ActiveDrive (Reimand et al., 2013), and described in a Pan Cancer comparative analysis of 12 different tumor types (Pan-Cancer12) (Tamborero et al., 2013b). All these methods, except MutSig, are publicly available and implemented as tools. Since the analysis described in Pan-Cancer12 contains the candidate driver genes derived from a cross-methodology that includes a pathway analysis and a series of sequential filters, we retrieved the output of each method from Synapse at the following accession numbers: syn1715784 for MutSig, syn1701498 for both OncodriveFM and OncodriveCLUST, and syn1713813 for MuSiC. As the original output of ActiveDriver was unavailable, we used the genes predicted by ActiveDriver that were present in the aggregated results. Then, we run DOTS-Finder on the Pan-Cancer12 dataset syn1729383. Furthermore, we compared our results with the predictions made by an additional available tool, MutSigCV version 1.4 (Lawrence et al., 2013), by using default parameters on the same input dataset (Figure 12). The predicted driver genes for all the above-cited tools can be found in Tables S1 in (Melloni et al., 2014). For statistical comparison, we evaluated precision and recall of all the methods

against 162 genes belonging to the Cancer Gene Census (version 68). We selected these 162 genes since they are the ones targeted by single nucleotide variants (SNVs) and small insertions and/or deletions (InDels) mutations. The other CGC genes are amplified, translocated or targeted by large insertions/deletions in cancer, thus being outside the scope of our study. To obtain a unique measure of accuracy of the predictions, we aggregated precision and recall through the F1-Score, a well-established balanced value of accuracy calculated as the harmonic mean of precision and recall. Since we have no *a priori* knowledge of the true negatives and we only know the true positives, measures that take into consideration only precision and recall are preferable in this context. In this sense, a method with a good balance between precision and recall ensures that the predicted genes that are not in CGC could be reliable driver candidates. For example, as shown in Figure 12**,** Panel D, a method like MuSiC shows a recall comparable to DOTS-Finder, but a lower precision.

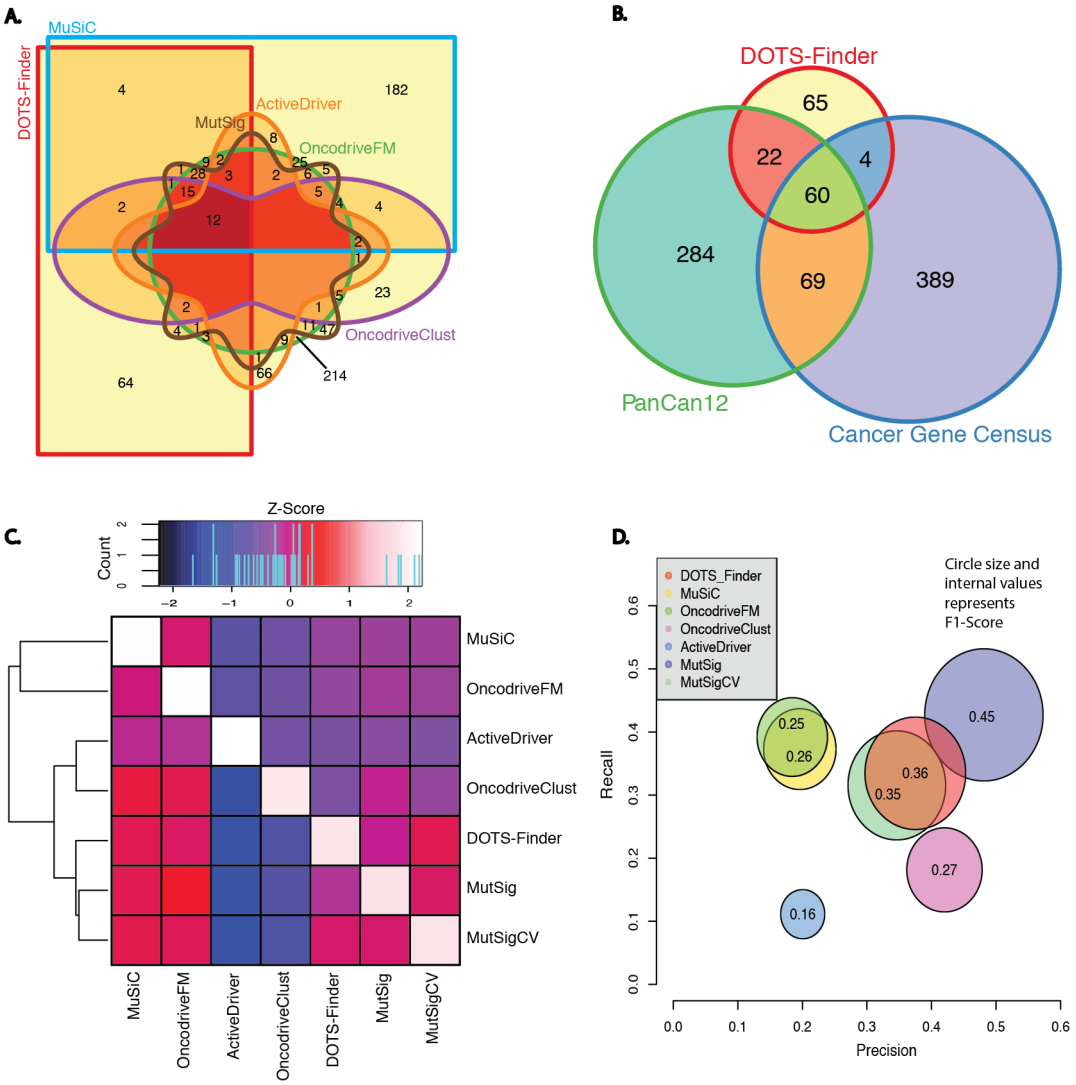

Figure 12 DOTS-Finder results compared to the Pan-Cancer12 analysis. (A) Six-way Venn diagram of DOTS-Finder and 5 other tools. This panel shows the number of putative driver genes that are predicted individually by each tool or in common by multiple tools. The diagram uses a graduated color ramp from light yellow to dark red to represent the overlap of an increasing number of tools that predict the same drivers. A full concordance between these methods can be obtained only for 12 genes among the 654 uniquely identified by at least one method (they are present in region with the darkest shade of red). This is due to the fact that each method is implemented for assessing different aspect of drivers' behaviors. Therefore, an approach that combines different complementary methods, as proposed by Tamborero *et al.*, is certainly preferable. **(B)** Pan-Cancer12 aggregated results compared to DOTS-Finder and CGC. This panel shows the existing overlap between the list of high confidence drivers and candidate drivers provided by both Pan-Cancer12 analysis and DOTS-Finder, crossed with the entire list of CGC genes (522). DOTS-Finder is able to retrieve 4 new CGC genes (*CALR, CREBBP, KDR, KIAA1549*) that none of the other methods were able to confirm. Interestingly, *CALR* has been recently added to the CGC. In addition, 65 new genes are predicted by DOTS-Finder as possible driver candidates, including *CBX7* and *UBC*, described in this paper. **(C)** Heatmap of the similarity between 7 methods. This heatmap is built on the number of overlapping genes between each pair of tools normalized by row. Therefore, the dendrogram on the left side of the plot indicates the similarity between pairs of methods compared to all the remaining ones. Results show that DOTS-Finder is close to MutSig and MutSigCV algorithms in terms of cross-predicted genes. It is instead very different from both MuSiC and OncodriveFM, which form an independent cluster disjointed from all the others. **(D)** Statistical comparison of all the methods against the 162 CGC genes targeted by SNVs and/or InDels mutations. In this plot we compared the precision (X-axis), recall (Y-axis) and F1-Score (harmonic mean between precision and recall; circles area) of 6 different available tools, including DOTS-Finder, against the 162 CGC genes used as a gold standard reference. In terms of F1-Score (harmonic mean between precision and recall), DOTS-Finder is the best performer. The aggregation of 3 different methods used by the latest MutSig strategy reaches an F1-Score of 0.45. However this strategy is not publicly available. DOTS-Finder and MutSig comprehensive approaches and the entire

Pan-Cancer12 analysis confirm that an approach that takes into consideration different sources of information is certainly preferable.

This indicates that we can provide the same number of true outputs with fewer attempts. According to this measure, DOTS-Finder is the best tool among the available ones with an F1-Score of 0.36 (precision=0.37, recall=0.35) (shown in Figure 12, Panel D).

4.1.5.11 Statistical power using a small number of cancer samples

One of the main strength of DOTS-Finder is its ability to retrieve reliable results even using a small number of cancer samples as input. Our double step procedure ensures a higher sensitivity to the deviation from the null hypothesis of being a passenger-mutated gene. In order to assess this characteristic, we collected the data from the latest bladder cancer TCGA dataset (238 patients) and run our pipeline against MutSigCV 1.4 using default parameters. We decided to use MutSigCV for this statistical comparison, as it is the available method with the best performance after DOTS-Finder. We retrieved 31 significant driver genes against the 26 of MutSigCV, with 16 common predictions. Then, we randomly down-sampled our dataset at several sampling fractions (5%, 10%, 15%, 20%, 30%, 40%, 50%, 70% and 90%) and selected 5 different subsamples for each fraction. We end up with 9x5 subsamples made up of a minimum of 12 to a maximum of 214 patients. We then run both DOTS-Finder and MutSigCV on all the 45 subsamples and collected the number of identified drivers. Our results show that DOTS-Finder is superior in terms of absolute output Figure 13, Panel A, especially for small sample size (from 12 to 48 patients).

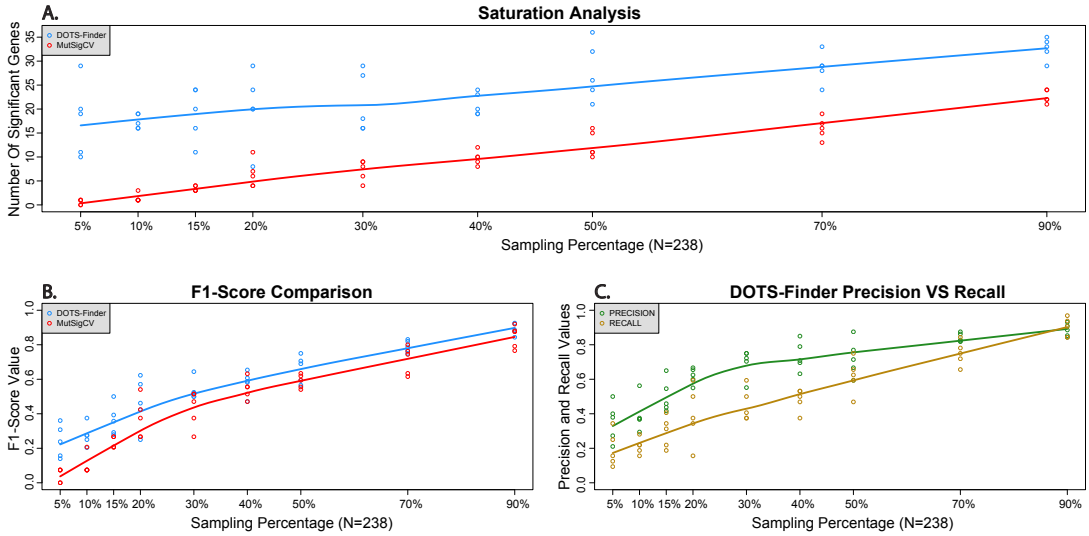

Figure 13 Comparative saturation analysis and performance analysis for a range of sample sizes. (A) Comparative saturation analysis. Here we show the absolute output in terms of number of significant genes found by DOTS-Finder (blue line) and MutSigCV (red line) for every subset from each down-sampling fraction. DOTS-Finder is able to provide a consistent output even with a very limited number of patients (a minimum of 10 genes identified with just 12 patients while MutSigCV retrieves 0 or 1 gene at best - always *TP53*). **(B)** Comparative F1-Score. In this panel we compared every prediction on the subsamples to the full output of each tool considering the whole dataset (N=238). Our predictions are not only consistent, but maintain an F1-Score distribution that is uniformly higher than MutSig at any downsampling level. This difference is much more evident for small samples. **(C)** Precision-Recall plot for DOTS-Finder. Here we present the precision-recall output of every subsample compared to the significant genes found on the entire dataset. With just the 5% of the entire dataset, DOTS-Finder is able to predict an average of 20% of the full output with a precision of almost 40%.

Our tool is also able to recapitulate its own results in terms of precision and recall better than MutSigCV, at any level of downsampling (Figure 13, Panel B). However, this difference is more evident for subsamples with very small fractions (from 5% to 30%). Finally, as shown in Figure 13, Panel C, we can observe that DOTS-Finder can recapitulate up to 40% of the results of the entire 238 patients-dataset, using just 5% of the dataset (12 patients), with a precision of almost 50%.

4.1.6 Discussion

DOTS-Finder is the first published software that can identify driver genes and can classify them as TSGs and/or OGs and it can also be used to identify driver genes with atypical patterns of mutations (Figure 11). In addition, it is the first software that can be used by a vast and diverse scientific community as it is easy to install and use, does not

need the availability of property software, and does not require the use of low-level and hard to access files (e.g. as bam files, coverage files).

We have applied DOTS-Finder on publicly available datasets containing the mutation profile of 34 cancer types. We have obtained plausible driver genes for many low mutation rate cancers like gliomas, acute myeloid leukemia and prostate cancer. Notably, we have obtained results that are consistent with the literature even with some high mutation rate tumor types, like Head and Neck Squamous Cell Carcinoma and Bladder Cancer, where the risk of falling into the "fishy genes" trap is higher.

Our tool outperforms other available methods in terms of precision-recall, considering CGC as a gold standard. Importantly, DOTS-Finder has confirmed the predictions made by other methods and discovered novel driver candidates never identified before.

Using DOTS-Finder, researchers can identify driver genes in large public databases and also in user-defined samples stratified for a given characteristic, as the software is specifically designed to identify driver genes even in small datasets (e.g. obese/normal weight, male/female etc.). The use of few samples in cancer is justified by the high molecular heterogeneity present in tumors. Indeed, we believe that the results produced by DOTS-Finder could be very useful for those researchers who want to identify driver genes in user-defined datasets, in order to investigate the significance or relevance of particular somatic mutations in relation to specific clinical questions.

4.2 LowMACA: exploiting protein family analysis for the identification of rare driver mutations in cancer

This section of the results is adapted from (Melloni et al., 2016). DOTS-Finder concept was to create a tool able to detect driver genes and divide them between tumor suppressors and oncogenes. Nevertheless, there are many genes whose frequency of mutation is so low that any statistics would fail to detect them because of the lack of statistical power. Therefore, in order to distinguish drivers from passengers, the only straightforward solution would be to sequence more cases that is of course impractical from many different points of view (time, cost, etc.). Nevertheless, mutations are entities with a large context and many properties that can be exploited by aggregating them in larger mutations clusters. For examples, mutations can have predictable effects on the final protein product, or again, they belong to genes, which in turn, belong to pathways or families. Aggregating mutations increase statistical power at the expense of losing the granularity of each single mutation. In particular, the tool presented in this section, LowMACA, exploits protein families introducing multiple sequence alignment as an approach to find connections between genes that show similarity in the secondary structure.

4.2.1 Abstract

The increasing availability of resequencing data has led to a better understanding of the most important genes in cancer development. Nevertheless, the mutational landscape of many tumor types is heterogeneous and encompasses a long tail of potential driver genes that are systematically excluded by currently available methods due to the low frequency of their mutations. We developed LowMACA (Low frequency Mutations Analysis via Consensus Alignment), a method that combines the mutations of various proteins

sharing the same functional domains to identify conserved residues that harbor clustered mutations in multiple sequence alignments. LowMACA is designed to visualize and statistically assess potential driver genes through the identification of their mutational hotspots. We analyzed the Ras superfamily exploiting the known driver mutations of the trio *K-N-HRAS,* identifying new putative driver mutations and genes belonging to less known members of the Rho, Rab and Rheb subfamilies. Furthermore, we applied the same concept to a list of known and candidate driver genes, and observed that low confidence genes show similar patterns of mutation compared to high confidence genes of the same protein family. LowMACA is a software for the identification of gain-of-function mutations in putative oncogenic families, increasing the amount of information on functional domains and their possible role in cancer. In this context LowMACA emphasizes the role of genes mutated at low frequency otherwise undetectable by classical single gene analysis. LowMACA is an R package available at http://www.bioconductor.org/packages/release/bioc/html/LowMACA.html. It is also available as a GUI standalone downloadable at: https://cgsb.genomics.iit.it/wiki/projects/LowMACA

4.2.2 Introduction

As previously described, the identification of driver mutations in cancer can be enhanced by considering the position of the mutations on the proteins rather than their simple frequency in cancer cohorts (Vogelstein et al., 2013). For this reason, tools that combine frequency of mutations and their position on the genome have been recently developed for the identification of potential drivers in small cohorts of patients to increase statistical power (Davoli et al., 2013; Melloni et al., 2014; Tamborero et al., 2013a). Furthermore, other methods based on network analysis were developed to aggregate mutational information at the level of interaction pathways (Mutation Consequences and Pathway Analysis working group of the International Cancer Genome Consortium,

2015). Nevertheless, as pointed out in a recent simulation based on saturation analysis on publicly available cancer data, we are still far from a true understanding of the genes mutated in less than 5% of the patients for almost any tumor type (Lawrence et al., 2014). Due to the lack of the required sample size, methods able to assess the role of rarely mutated genes are needed. LowMACA represents a solution to increase the information content of alteration patterns by summing up mutations on properly aligned amino acids in different proteins belonging to the same family. The accumulation of somatic mutations in specific Pfam domains has been already observed in cancer, introducing the concept of domain landscapes of somatic mutations in addition to the well-known genomic landscape (Nehrt et al., 2012; Peterson et al., 2010; Yang et al., 2015a).Nevertheless, these approaches only rely on the frequencies of mutated domains in cancer. We enhance this approach by adding the positional information of mutations within the domains, eventually increasing the statistical power of the domain level analysis. With LowMACA, we are able to assess various aspects of somatic mutations at the level of protein families, including clustering in hotspots, conservation of mutated residues, pattern similarity across proteins and co-occurrence or mutual exclusivity among positions resulting significant by LowMACA criteria. In fact, one of the significant improvements over existing methods is the ability of LowMACA to test single driver mutations and not only driver genes. All these unique aspects are illustrated here in the context of the Ras superfamily and in the analysis of a state-of-the-art set of high confidence and putative driver genes (Tamborero et al., 2013b).

4.2.3 Materials and Methods

4.2.3.1 Software Implementation and Overview

LowMACA is a computational tool for the analysis and visualization of somatic mutation data in cancer. It allows to properly assess the significance of hotspots of mutations shared across protein families and to show the interconnectivity among mutational patterns via different visualization methods. The software comes as an R package, fully integrated in the R/Bioconductor environment through the use of the AAMultipleAlignment class from the Biostrings library. The multiple alignment is performed with a wrapper around a clustal omega executable (Sievers et al., 2011) or the EBI soap webserver (McWilliam et al., 2013). At the present time, LowMACA is the only tool that allows using clustal omega within R storing results within a Biostrings class. Importantly, the LowMACA package implements a user-friendly GUI built with the shiny package, exploiting the interactive functionalities provided by D3 javascript and google charts plotting libraries. The tool comes with a pre-built annotation package named LowMACAAnnotation, that integrates the information of HGNC (Gray et al., 2014), UNIPROT un and Pfam (Finn et al., 2007) with the aim of guiding the user through the analysis of highly conserved classes of proteins belonging to common Pfam domains. The LowMACAAnnotation package creates a one-to-one match between UNIPROT canonical proteins and HGNC gene symbols and provides all the Pfam sequences of each protein entry.

LowMACA implements two conceptually different workflows: a Hypothesis Driven workflow and a Data Driven workflow.

The Hypothesis Driven workflow consists of:

1) Selecting proteins belonging to the same family (we suggest Pfam as a guideline).

2) Selecting one or more tumor types and classes of mutations that will be analyzed (see Methods section Input Data).

3) Retrieving mutations from specified cancer samples.

4) Aligning selected sequences along with their mutations

5) Calculating statistics and evaluating significant hotspots with different parameter settings

The Data Driven workflow consists of:

1) Providing a dataset of mutations from a cancer cohort in a format derived from TCGA standard maf files (see Input Data).

2) LowMACA collects all the genes that harbor at least one mutation and align their domains according to Pfam. Subsequently, the mutations are mapped on every consensus sequence created (one per Pfam analyzed).

3) LowMACA analyzes the mutational pattern of every protein by itself.

4) The hotspots found at point 2 and 3 are unified in one table and the list of putative driver mutations is presented (detailed information can be found in the package reference manual: http://bioconductor.org/packages/release/bioc/manuals/LowMACA/man/LowMACA.pdf).

4.2.3.2 Input Data

According to the choice of a Hypothesis Driven or Data Driven workflow, LowMACA requires different kinds of input. In the first case, LowMACA expects as input a Pfam ID of interest (e.g. "PF00001") and/or gene names, provided as Entrez Gene IDs (Maglott et al., 2005) or HUGO Gene Symbols (Gray et al., 2014). In case only a Pfam ID is provided, the LowMACAAnnotation package will look for all the genes that contain the specified domain, otherwise, only the chosen genes are retained. By selecting a Pfam ID of reference, only the portion of the proteins mapping to the Pfam domain will be

considered in the analysis. If a set of gene identifiers is selected without specifying any Pfam ID, the entire protein sequences are considered for the analysis. LowMACA admits also the use of non-ambiguous gene aliases. The LowMACAAnnotation package is designed to assign only canonical proteins to the relative gene creating a one-to-one unique match. LowMACA retrieves mutational data via the R/CRAN package "cgdsr" which queries the Cancer Genomics Data Server (CGDS) hosted by the Computational Biology Center at Memorial-Sloan-Kettering Cancer Center (MSKCC) (Cerami et al., 2012; Gao et al., 2013). Mutation data coming from personal databases can alternatively be used, following the instructions provided within the manual of our R-package. Since LowMACA looks for hotspots of mutations, the package keeps by default only the mutations that modify the protein without altering the reading frame or creating stop codons (collectively identified as missense type mutations). Other mutation types, such as frame shift InDels, nonsense mutations or splice-site mutations (collectively called truncating mutations), can be retrieved by modifying the parameters. By default, LowMACA will take into account all the tumors present within the cBioPortal repository, but mutations from specific cancer types can be selected. In case a data driven workflow is chosen, the user has to provide only mutation data. These data are a direct derivative of a common maf file as specified by TCGA and contains the mutations annotated by their gene, their amino acid change, sample of origin and type of mutation. A detailed description can be found in the package reference manual: http://bioconductor.org/packages/release/bioc/manuals/LowMACA/man/LowMACA.pdf.

4.2.3.3 Alignment and Mapping

Amino-acid sequences selected as described above are aligned using the multiple sequence alignment software Clustal Omega (Goujon et al., 2010; McWilliam et al., 2013; Sievers et al., 2011). Although the Pfam database is a comprehensive archive of cross-

species alignments, we only refer to human proteins and each Clustal Omega alignment represents a unique combination of conserved and not conserved residues. Using the original HMM model of the protein family is a limiting factor in this case, as we would lose portions of alignments specific to human proteins only. Moreover, Clustal Omega can handle alignments involving whole protein sequences, rather than only Pfam domains. From the output of the multiple alignment, a consensus sequence including the most represented amino acid found at every position is created that is representative of all the sequences under investigation. The mutations coming from aligned sequences are remapped directly on the consensus with the aim of obtaining a unique mutational profile. Considering that LowMACA specifically aims at highlighting mutations that fall on conserved residues, two measures of conservation are taken into account at this point. The first one concerns the specific positions of the alignment. LowMACA calculates the Trident conservation score for this purpose (Valdar, 2002), which is a mixed measure that encompasses three different aspects of a local alignment:

1) The entropy of the residues at the specific position. The more different amino acids are aligned the less conserved is the position.
2) The chemical similarity according to the substitution matrix BLOSUM62
3) The relative frequency of gaps

The second measure is global and involves the entire sequence. The alignment procedure of the LowMACA engine is delicate due to the fact that including dissimilar sequences in the analysis can invalidate the whole LowMACA workflow. For this reason, sequence similarity for every pair of amino acid sequences is calculated, based on the k-tuple measure (Wilbur and Lipman, 1983), and a warning is prompted whenever an amino-acid sequence differs too much from the others (threshold = 0.2). These measures are a safety net to avoid false positive results due to low quality alignments and become extremely useful if the user decides to perform analysis with

sequences not belonging to the same family. LowMACA provides the Pfam based framework as a guideline, but in theory every mutation profile can be compared.

4.2.3.4 Statistical Testing

4.2.3.4.1 Testing the randomness of the global mutational profile

Once the sequences are aligned and the mutations have been remapped on the consensus sequence, LowMACA measures the information contained in the mutational pattern using Shannon's definition of entropy (Melloni et al., 2014)

$$H(X) = -\sum_{i}^{K} P(x_i) ln P(x_i)$$

where $P(x_i) = \frac{n_i}{N}$ is the frequency of mutations mapping to the position i of the consensus alignment of length K and N is the total number of mutations. To statistically assess whether the pattern of mutations significantly differs from randomness, we compare $H(X)$ with the entropies of a bootstrap of one thousand random profiles. Each random profile is generated according to the following criteria: (i) the random profile has the same length of the consensus sequence generated from the analysis (i.e. K); (ii) the number of mutations that map on the random profile is equal to the total number of mutations that map on the consensus sequence (i.e. N); (iii) the probability of a mutation to fall onto a specific position of the random profile is proportional to the number of amino acids that map in the corresponding position of the multiple alignment. In this way, the more gaps are found in a position of the alignment, the lower is the probability that a mutation falls in that position in the random model. This last criterion is intended to correct the bias of finding more mutations in more conserved regions of the consensus. We fit the parameters of a Gamma distribution over the empirical distribution of the entropies calculated on the random profiles. This will be considered as the null distribution and used to assign a p-value to the global mutational profile.

4.2.3.4.2 Testing for the identification of hotspots of mutation

LowMACA is also able to identify significant positions along the consensus sequence, as opposed to the large majority of driver gene identification approaches (Tamborero et al., 2013b). The probability that the number of observed mutations n_i on position i of the consensus sequence derives from a random pattern of mutations is calculated estimating the per-position null distribution of the number of mutations that are expected to fall on that specific position. The null distribution is modeled using the Gamma distribution whose parameters are estimated from the bootstrapped random profiles generated for testing the randomness of the global mutational profile. A per-position p-value that the observed number of mutations originated from the null distribution is then calculated and p-values of residues that fall onto conserved positions (Trident score > 0.1) are corrected to obtain per-position q-values using the Benjamini-Hochberg procedure for multiple testing correction (Benjamini and Hochberg, 1995).

4.2.3.5 LowMACA Output

Using a Hypothesis Driven workflow, LowMACA outputs a detailed report of the mutational landscape of the consensus sequence. It specifies if the entire mutation profile can be considered random (global p-value), and it reports all the mutation hotspots that exceed the random distribution (per-position p-value and relative FDR corrected q-value); see Statistical Model section. Mutations that fall onto significant positions of the consensus sequence can be retrieved in their original position with a reverse mapping provided by LowMACA. The mutational profile can be visualized with many LowMACA methods. These plotting capabilities are considerably extended through the GUI. The interactivity that this implementation allows is particularly useful to observe the dynamic connections among mutational profiles of different proteins. The following plot types are offered by the package:

1) A stacked barplot that specifies the relative frequency of mutation per sequence in each position (in the GUI this plot has interactive features). This representation also includes a graphical view of the trident score and a logo plot of the most represented amino acids at every position.

2) A Protter style plot (Omasits et al., 2014) that represents the possible secondary structure of the consensus sequence with the significant positions found by LowMACA highlighted in red.

3) An interactive network plot in which the nodes represent the single sequences and the edges are drawn based on the number of shared mutated residues. The thicker are the edges, the more positions are in common. This representation provides an overview of the similarity among sequences in terms of mutational profile.

4) A heatmap of mutual exclusivity and co-occurrence of mutations at the entire sequence level and at single position level implemented with the R package co-occur (Griffith et al., 2016). For example, it can represent mutual exclusivity between mutations in *KRAS* and *NRAS* and between *KRAS* G12 and *NRAS* G12 positions (see Figure 1A).

The last two functionalities are only available through the LowMACA GUI. In a Data Driven workflow, the output is represented in a very similar way, but LowMACA takes care of analyzing all the Pfam domains through the mutations in the genes provided by the user in a single procedure. Every Pfam analysis can become a new LowMACA object and it can be viewed from a descriptive point of view as shown above in the Hypothesis Driven workflow.

4.2.4 Results

Our results are reported in three different sections. The first analysis is aimed at demonstrating the core concept of LowMACA using a known oncogenic family. Starting

from the cancer genes *KRAS*, *NRAS* and *HRAS (*that we will name *RAS trio)*, similar in structure and mutational profile, we seek to extend this conservation to all the Ras superfamily members (in total, 133 different proteins belonging to the PF00071). We demonstrate how LowMACA can be used to show the oncogenic potential of different positions of the family and to encompass new putative driver genes through the sharing of conserved mutations (see section **4.2.4.1**). We also evaluated mutual exclusivity of mutations that fall in specific positions of the consensus alignment (see section **4.2.4.2**). Moreover, by collecting all the observed mutations that fall in PF00071, we show that LowMACA hotspots fall in positions that are expected to be damaging by 8 different predictors of phenotypic effect. Although LowMACA predictions and mutation damage assessments are in agreement with the other predictors, our tool is more specific in assessing driver mutations against a gold standard of known cancer driver mutations and disease associated mutations (see section **4.2.4.3**). The second analysis is aimed at assessing the state-of-the-art in driver genes at a domain level. By taking a curated list of high confidence drivers (HCDs) and a list of candidate driver genes (CDGs) derived from 5 different bioinformatic tools (Tamborero et al., 2013b), we study the relationships in terms of common mutations among these genes (see section **4.2.4.4**). We show that 40% of all the HCDs share at least one domain with a CDG defining and expanding the same concept illustrated in the Ras example. Mutations that fall in known driver genes are shared both by other known drivers (like the tyrosine kinases *EGFR*, *BRAF*, *FLT3* and *JAK* family) but also by less frequently mutated genes with a similar structure (like the receptor L domain genes *ERBB2* with *ERBB4*). The third analysis shows, as a negative control, that silent mutations do not have the propensity to show significant pattern of mutations (see section **4.2.4.5**).

4.2.4.1 Ras superfamily analysis

We aligned and summarized the mutational landscape of the Ras superfamily, defined by PF00071 (Figure 14). This Pfam represents a large family of small GTPases that can be grouped in different subfamilies with specific biological characteristics (Wennerberg, 2005). We performed our analysis in two steps. First, we aligned all the mutations of the entire family encompassing 133 sequences. Second, we performed the same analysis dividing the mutations by the four main subfamilies: 1) Ras subfamily, involved in cell proliferation (Pylayeva-Gupta et al., 2011), 2) Rheb subfamily, involved in neural plasticity (Li et al., 2008), 3) Rho subfamily, involved in cytoskeletal morphology (Hall, 1998) and 4) Rab family, involved in cell trafficking (Stenmark, 2009). Analysis of the entire family found significant hotspots in the consensus alignment in positions 16, 17, 102, and 282, as highlighted in Figure 14, Panel A. In this analysis, we discuss genes that have at least two mutations in any of the identified hotspots. These mutations are well conserved in the superfamily but appear mainly represented in the Ras subfamily.

The main representative members of this proto-oncogenic subfamily are the known cancer genes that compose the *RAS trio.* Their mutations G12, G13, Q61 and A146, considered important drivers in many cancers (Janakiraman et al., 2010), map on the hotspots identified above. These three proteins share over 90% of sequence identity in the domain and are the most represented in terms of absolute number of mutations in these positions. Hotspots found in position 16 of the global alignment harbor mutations on residues G12 of *RAP1B,* on residue S13 of *RERGL* and G23 of *RALB*, which align with G12 of the *RAS trio*, while position 17 aligns with mutations on G85 in *GEM,* which aligns with G13 of the *trio.* Even if these proteins, (excluding the *trio*) are very rarely mutated, LowMACA identifies their alterations as putatively oncogenic (Figure 14, Panel B). All these proteins belong to the Ras subfamily, but a particular exception is represented by *RERGL* that harbors a recurrent S13F mutation: this protein is considered

part of the Ras subfamily but its sequence is very distant from the *RAS trio* (Figure 14, Panel C) and for this reason should be analyzed separately.

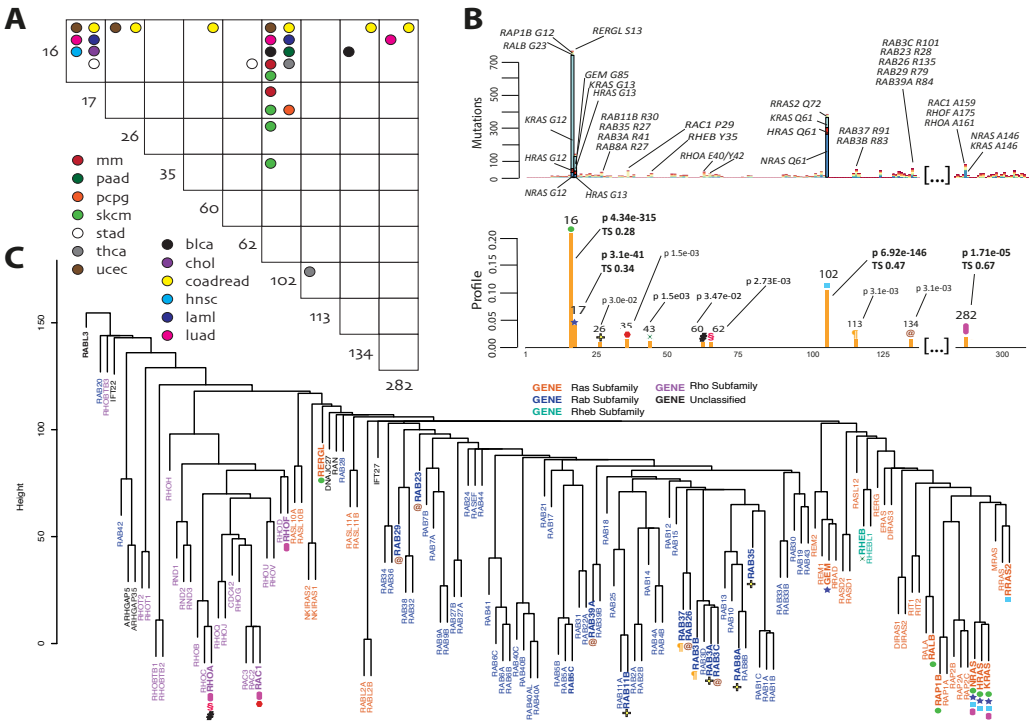

Figure 14 Ras Family A) LowMACA results based on the alignment of the Ras superfamily (PF00071). The first barplot reports the most mutated proteins under significant hotspots in their original position. These hotspots are also highlighted in the second barplot with colored symbols. Labels in the second barplot report the position of the consensus, the FDR corrected p-value and the trident score of conservation (TS). The TS is reported only for hotspots identified in the alignment of all the 133 family members. Both barplots are truncated on non-informative positions. **B)** The panel shows a plot representing the mutual exclusivity between mutations that fall in the same position of the global consensus alignment. Significant patterns are highlighted with the color corresponding to the tumor type where the mutual exclusivity was found. We consider mutually exclusive the pairs with a corrected p-value below 0.05 using the R package cooccur **C)** The dendrogram is built on hamming distances between all human sequences of the Ras superfamily aligned via clustal omega. Genes that belong to the same subfamily, as described in (Hall *et al.*, 1998), are represented with the same color. Significant hotspots (under gene names) are represented with the symbols used in Panel A.

Another highly conserved mutation is located in the aligned position 102 that corresponds to mutations in Q61 in *NRAS* prevalently and is one of the residues involved in the binding function of all the Ras family members to GTP (Prior et al., 2012). LowMACA analysis highlighted mutations aligned in position 102 also in other Ras members, in particular Q72 mutations in *RRAS2*. This gene has been extensively

analyzed at the transcriptional level but remains poorly investigated regarding the mutational context (Gutierrez-Erlandsson et al., 2013). *RRAS2* has a role in pathways activated by the *RAS trio*, however, while the *trio* exerts its pro-proliferative activity via the activation of the Raf-ERK pathway of MAP kinases, *RRAS2* activates this pathway poorly as it does not recruit Raf1 (Gutierrez-Erlandsson et al., 2013). Following the observation of several Q72 mutations in RRAS2, one might speculate on a possible activation of this gene in the same way as Q61 activates *NRAS*.

Position 282, corresponding to an alanine in 146 in the RAS trio, represents a completely different case. This hotspot is extremely well conserved in all the members of the superfamily and represents the only case of a significantly mutated residue shared by two different Ras subfamilies (Ras and Rho). This mutation does not impair the affinity with GTP (like G12/13 and Q61) but rather seems to have an effect on the GTP-Ras steady-state levels as reported by experimental assays (Janakiraman et al., 2010). *RAC1*, *RHOA* and *RHOF* emerge as putative oncogenes by this analysis, sharing mutations in this position. Among these, *RAC1* and *RHOA* are already present in the Cancer Gene Census (Futreal et al., 2004), adding confidence to the hypothesis that also *RHOF* might play a role in cancer. Moreover, relatively elevated levels of RHOF were observed in lymphomas derived from the germinal centre (Gouw et al., 2005).

Hotspots identified in the previous analysis correlate well with sequence similarity based on hamming distance (Figure 14, Panel C). For example, the aforementioned hotspots 16, 17 and 102 belong specifically to the Ras subfamily, identified in orange in the dendrogram. This subfamily harbors two glycines in position 16 and 17 that are not shared by the entire superfamily. In fact, the 16/17 glycines can be substituted by the couple serine/glycine (Rab subfamily) or the couple glycine/alanine (Rho subfamily) (Wennerberg, 2005). The Rheb subfamily instead, composed of just two genes *RHEB* and *RHEBL1,* does not conserve any of the two marker residues and carries a distinctive

leucine in position 16. By analyzing mutations that fall individually in each of the four subfamilies, we were able to identify new putative oncogenes and new hotspots of mutation. In order to keep the reference with positions identified with the global analysis, we maintained the full alignment of all the proteins of PF00071 and then subset the genes of interest according to the four subfamilies (this alignment parameter is called "datum" in the LowMACA package).

The analysis of the Rab subfamily (mostly represented in the central portion of the dendrogram in Figure 14, Panel C) highlights three new hotspots and 11 new putative oncogenes. Among these, *RAB29* harbors 4 mutations in position 134 of the alignment that are predicted to be damaging by most of the functional predictor tools used in section **4.2.4.3** and reported in Appendix Table 4 (R79W in Colorectal cancer and R79L in Lung adenocarcinoma). The involvement of members of this subfamily in cancer has been widely demonstrated (Chia and Tang, 2009).

The analysis of the Rho subfamily allowed the identification of new hotspots, which are mainly represented by *RAC1* and *RHOA*. *RAC1* marks a single hotspot found in position 35 corresponding to mutations of the residue proline 29 (*RAC1* P29). According to the most recent literature, P29 results altered in approximately 3.9% of TCGA skin cutaneous melanoma patients (Hodis et al., 2012) suggesting that *RAC1* is a melanoma oncogene. The biological significance of the *RAC1* P29 mutation remains unclear, although authors demonstrated that the mutation could destabilize the *RAC1* inactive GDP-bound state in favor of its active GTP-bound state, creating a gain-of-function oncogenic event (Watson et al., 2014). In fact, the expression of *RAC1* P29S in sensitive *BRAF*-mutant melanoma cell lines confers resistance to treatment with *RAF* inhibitors (Watson et al., 2014). Moreover, the P29S mutation has been reported in several cancers such as head and neck tumors (Stransky et al., 2011) and breast tumors (Forbes et al., 2011). Other Rho subfamily members also share the hotspot 35: *RAC2*, *RHOT1*, *RHOC*.

Even though one single mutation was found for each gene in our dataset, this position is extremely well conserved (a proline is present in all four genes) and all the mutations were found in melanoma patients without a *RAC1* P29 mutation (Appendix Table 4).

The mutational hotspots 60 and 62, respectively corresponding to glutamate 40 and tyrosine 42 in *RHOA*, were observed in seven tumors (six head and neck, one breast) and affect the effector domain of *RHOA*. *RHOA*, is considered a gene encoding a protein that is clearly involved in cell proliferation (Lawrence et al., 2014). As for the case of *RAC1*, also *RHOA* shares its hotspots with other Rho subfamily members (these results are not reported in Figure 14 since only one mutation was found in our dataset). These genes include *RHOH* E39K for hotspot 60 and *RHOC* Y42C and *RAC1* Y40S for hotspot 62. Both positions are still well conserved in the subfamily (Appendix Table 4).

The analysis of the Rheb subfamily shows a significant number of mutations that fall in the hotspot 43. These mutations are mostly represented by Y35N hosted by *RHEB* and found present in Kidney Renal Clear Cell and Uterine Corpus Endometrioid Carcinomas in TCGA patients. Moreover, authors observed that mutations of *RHEB* (Y35N/C/H) increase phosphorylation of endogenous substrate S6 kinase (S6K1) of the mTOR signaling pathway (Grabiner et al., 2014), a protein kinase that plays key roles in cellular regulation (Wang and Proud, 2011). For the presence of the Y35N mutation, *RHEB* was recently highlighted as a novel cancer gene involved in cell proliferation (Lawrence et al., 2014), and cancer associated mutations in *RHEB* inducing mTORC1 activity have been reported (Grabiner et al., 2014). The only other member of the subfamily (*RHEBL1)* shares a Y35H mutation in the same hotspot in one melanoma case in our dataset.

4.2.4.2 Mutual exclusivity analysis

In order to corroborate LowMACA results reported above, we performed mutual exclusivity analysis on significant mutations and hotspots. Mutual exclusivity between

mutations on genes of the same pathway is a critical measure to assess if the pathway is relevant for cancer. The reason is that after the first mutation occurs, there is no selective pressure for a second mutation in another gene of the same pathway (Vandin et al., 2012). While generally performed gene-wise (Ciriello et al., 2012), the particular characteristics of LowMACA allow us to extend this concept to mutations that map on conserved residues within Pfam domains. If a putative driver mutation is found to be mutually exclusive with a known driver, its significance is enhanced as it possibly exerts the same function in cancer. We implemented mutual exclusivity analysis using the R package *cooccur* for a genomic analysis (Griffith et al., 2016), stratifying mutation data by tumor type. Our results revealed that hotspots in positions 16, 17, and 102 cover the large majority of mutually exclusive patterns (Figure 14, Panel A). This is a confirmation of the known exclusivity pattern of the mutations in *KRAS* and *NRAS* even among different positions within the genes themselves (Figure 15, right panel). In general, mutations in position 16 and 102 can be seen as a signature of two types of cancer: colorectal, characterized by *KRAS* G12, and melanoma, characterized by *NRAS* Q61 (Figure 15, left panel) (Janakiraman et al., 2010).

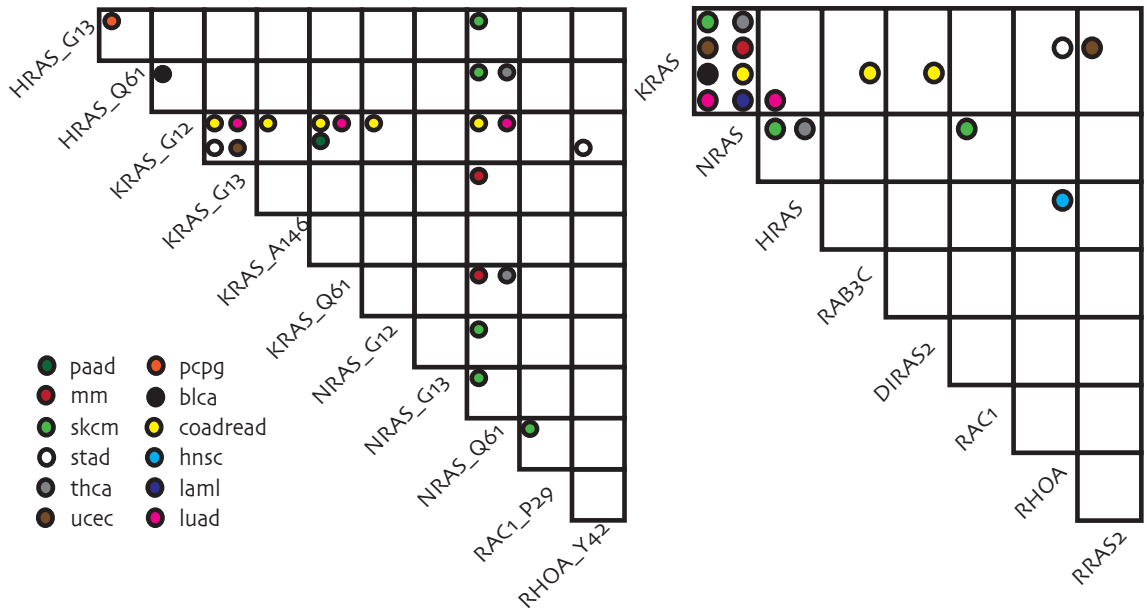

Figure 15 Mutual Exclusivity in the Ras Family. In these panels we show a plot representing the mutual exclusivity between mutations that fall in the same residue of individual proteins (**left panel**) and genes (**right panel**). Significant patterns are highlighted with the color corresponding to the tumor type where the exclusivity was found. We consider mutual exclusive the pairs with a corrected p-value below 0.05

using the R package *cooccur*. In the left panel we narrow down the patterns described in figure 1B, highlighting the major role of *KRAS* G12 and *NRAS* Q61. Notably, *RAC1* P29 and *RHOA* Y42 (described in the main text as potential new driver mutations) retain a pattern of mutual exclusivity with *NRAS* Q61 and *KRAS* G12, respectively. In the right panel it is possible to appreciate that mutual exclusivity between minor genes always occurs with the RAS trio. These genes cover many of the RAS subfamilies, in particular *RAB3C* (Rab subfamily), *DIRAS2* and *RRAS2* (Ras subfamily) and *RAC1* and *RHOA* (Rho subfamily).

These two highly frequent mutations allowed us to infer a possible driver role for less frequent mutations. For example, mutations in positions 26, 60 and 134 in colorectal cancer are mutually exclusive with position 16. Both hotspots are supported by this analysis in the Rab and Rho subfamilies. Similarly, position 102 is mutually exclusive with 26 and 35 in melanoma and 113 in thyroid cancer, further supporting the role of the aforementioned subfamilies.

4.2.4.3 Comparison with functional impact tools

We retrieved every amino acid substitution occurring in the RAS superfamily from the cBioPortal database (more than 10'000 different samples) and we annotated our predictions (if a mutation falls under a significant hotspot of LowMACA, as presented in the Ras superfamily analysis in section **4.2.4.1**). The 2294 unique substitutions found in PF00071 cover 2264 positions and 130 proteins of the 133 RAS superfamily genes. LowMACA predicts as significant 150 mutated residues under 11 hotspots (Figure 14), which correspond to 215 different substitutions. We also annotated this dataset including the predictions of functional impact from 8 different tools using ANNOVAR (Wang et al., 2010). These tools include PolyPhen2 (Adzhubei et al., 2010), Mutation Assessor (Reva et al., 2011), Mutation Taster (Schwarz et al., 2014), SIFT (Sim et al., 2012), MetaLR (Dong et al., 2015), LRT (Chun and Fay, 2009), FATHMM (Shihab et al., 2013a) and RadialSVM (http://genomics.usc.edu/software/11-icages) and are aggregated in the dbNSFP database (Liu et al., 2013). The functional impact of an amino acid variation can be assessed in many different ways (across species conservation, stoichiometric similarity between original and substitute amino acid, change of protein

conformation etc.), but all the algorithms share a similar output, assessing if a mutation could be considered "tolerated" or "damaging". We summarize this information as a damaging comprehensive score: the proportion of tools that predict the variation as damaging, ranging from 0 (all prediction as "tolerated") to 1 (8 out 8 prediction of damaging mutation). In Figure 16, Panel A, we show how the 2264 positions are classified in terms of this damaging score. This score is calculated on the actual amino acid substitution and not on the position, so in case there are more possible variations, the median damaging score is considered (like in the case of *KRAS* G12 that can be substitute by V, A, K and other amino acids and a unique damaging score exists for every change).

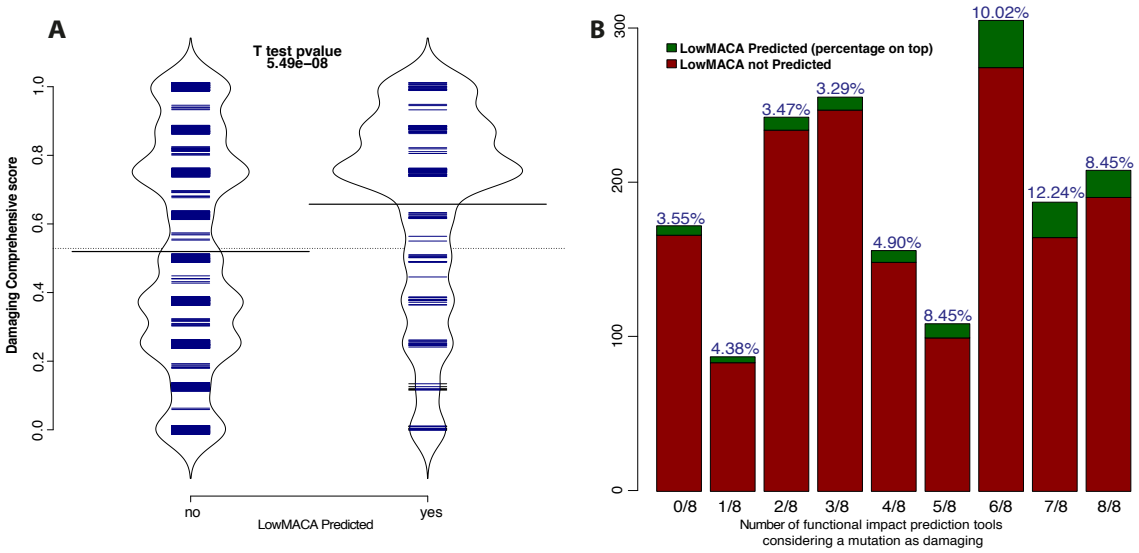

Figure 16 Comparison of LowMACA Ras mutations with functional impact tools. A) Distribution of damaging scores with respect to LowMACA classification in significant and non significant mutations **B)** Fraction of significant mutations in LowMACA for each class of damaging score. Damaging score is calculated as the fraction of dbNSFP tools (8 tools) that classify a specific mutation as damaging.

There is a significant difference (p-value of the two tails t-test = 5.49e-08) between the score calculated on the positions not considered by LowMACA and those that fall under LowMACA hotspots. This can be interpreted as a positive concordance between the LowMACA predictions (based on actual data) on the RAS family and their impact on the protein they belong to as calculated by the dbNSFP tools. This simply means that in cancer, the most frequent mutations are also the most damaging. In Figure 16, Panel B,

we show how the predictions of LowMACA are distributed. The majority of our predictions fall beyond the majority voting of the tools (5/8 and higher). To better understand the difference between dbNSFP tools and LowMACA, we further annotated our dataset with four databases of manually curated variations in the human genome predicted as disease-associated. These four databases include two databases for disease-associated variations, Humsavar (UNIPROT database of human polymorphism at protein level, www.uniprot.org/docs/humsavar), clinvar version 20140929 (Landrum et al., 2014) and two cancer-specific databases created by the Washington University, CiViC (Clinical Interpretation of Variants in Cancer, https://civic.genome.wustl.edu/#/home) and DoCM (Database of Curated Mutations http://docm.genome.wustl.edu/) . CiViC and DoCM are not published yet. We then test the predictions of LowMACA against the union of Humsavar and clinvar and against the union of CiViC and DoCM. In both cases, there is a significant positive association (p-value << 0.01 and OR >>1) (Appendix Table 5), meaning that predictions made by LowMACA are strongly in accordance with known results. Furthermore, we can appreciate a good overall recall and accuracy against these databases: 74% (32/43) and 95% (21/22) of recall for disease-associated and cancer-associated variants respectively with an accuracy of 15% and 10%. We performed the same analysis against those variants evaluated as damaging by more than 50% of the dbNSFP tools. Although still positively associated to pathogenic variations, the results are less striking (p-value 3.6e-05 and 1.89e-03 for disease-associated and cancer-associated respectively). While still maintaining a good recall, the accuracy is extremely poor. This is not surprising since functional prediction tools are not intended to find mutations that actually occur in cancer or other diseases, but simply to assess if a possible variation could be harmful to the affected protein. LowMACA has the ability to discern those mutations that actually occur in patients, enhancing the accuracy of a true functional prediction.

4.2.4.4 Analysis of driver genes: comparison with available tools

In this section, we analyzed the state-of-the-art driver genes identified with different bioinformatics tools under the lens of the protein families they belong to. In particular, we focused our attention on the 435 genes identified by a unifying approach as presented in (Tamborero et al., 2013b). In this study, driver genes are divided in two categories, 291 High Confidence Driver (HCDs) and 144 Candidate Driver Genes (CDGs), according to several criteria, which include: 1) the number of bioinformatic tools that identify the gene as potential driver (5 tools were taken into consideration), 2) if the gene belongs to a list of manually curated cancer genes as provided by the Cancer Gene Census (CGC) (Futreal et al., 2004), 3) if the gene belongs to the same pathway in the KEGG database (Kanehisa et al., 2014). With this analysis we want to add address two questions: what Pfam domains are contained in driver genes and what are the candidate driver mutations shared between HCDs and CDGs according to LowMACA criteria.

Since we are considering missense mutations, most of the tumor suppressors contained in the driver gene list will not be covered by LowMACA. In fact, tumor suppressors tend to lose their function during tumorigenesis and mutational landscapes are typically represented by sparse truncating mutations all over the gene body (Vogelstein et al., 2013). In this case, no clear clusters can be seen at single amino acid level because for a gene to lose its protein function there are generally no preferential positions. Furthermore, many tumor suppressors are singletons in the Pfam database, in the sense that their main domain can only be found in the genes themselves or in few other members (e.g. P53 Pfam, PF00870, is only shared by three genes *TP53*, *TP63*, *TP73*, Suppressor *APC*, PF11414, belongs to *APC* and *APC2* only). Nevertheless, highly mutated tumor suppressors like *TP53*, *VHL*, *RB1 ARID1*, *PTEN* and *APC* form actual hotspots that resulted significant in the LowMACA analysis (Zhao et al., 2013). To appreciate the full results, see Table S2 in (Melloni et al., 2016). Other known tumor

suppressors such as *WT1*, *CEBPA* or *CDKN1A* are instead missed by our analysis. The case of *TP53* is particularly interesting as it tends to form clusters of missense mutations specifically on its *P53* domain that probably exert oncogenic or dominant negative functions (Melloni et al., 2014). The fact that some tumor suppressors are identified and some are not depends in large part from the frequency of mutations. As the frequency increases, the sensitivity is enhanced and preferential positions of disruption emerge. Preferential mutation spots, even in tumor suppressors, are generally explained by possible dominant negative or oncogenic signature of certain tumor suppressors (Mahmoud et al., 1997; Papa et al., 2014) but also by a higher susceptibility to carcinogens of certain codons in these genes compared to other codons (Rivlin et al., 2011). 577 different Pfam domains are covered by the driver gene list, approximately one tenth of the entire Pfam-A database: 440 in the HCD list, 223 in the CDG list and 86 in common (Figure 17, Panel A). To assess whether the overlap between the Pfam domains contained in the lists of CDG and HCD is greater than expected, we randomly sampled the same amount of genes that are contained in the two lists and measured the overlap of the contained Pfam domains. On average, we found a smaller overlap (57 ± 7), but also a smaller number of Pfam domains in the CDG-sized samples (194 ± 11) and in the HCD-sized samples (355 ± 15).

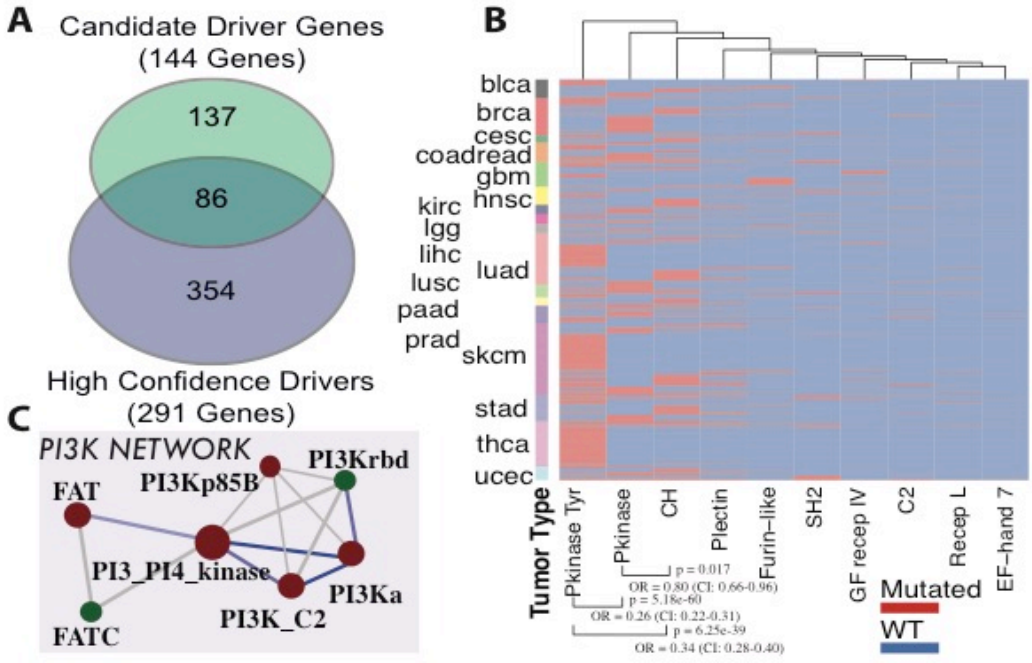

Figure 17 Other oncogenic Pfam families A) Venn diagram of the represented Pfam domains in the list of 291 high confidence drivers and 144 candidate drivers. A total of 577 different Pfam domains are covered by these genes with 86 Pfam domains shared between the two lists. **B)** Heatmap representation of significant Pfam domains in the "Kinase" network. Every row represents a patient of 17 different tumor types. A strong mutual exclusivity between tyrosine kinases, kinases and CH domain is shown. **C)** PI3K networks in driver genes. Every circle represents a distinct Pfam domain and the size represents the number of genes that contain the specified Pfam domain. Color indicates if significant hotspots were found in the LowMACA analysis (red is significant, green is not significant). Two domains are connected if they are found together on the same gene/protein. Edge thickness represents the number of genes that harbor both Pfam domains at the vertices (minimum 2). Blue color indicates mutual exclusivity and orange depicts significant co-occurrence.

We conclude that driver genes contain more domains than the rest of the other human genes (p=7e-9 and p=4e-3 for HCD and CDG, respectively, via z-test) but their overlap is not significant (p=0.38 via chi-squared test). The first two significant p-values can be interpreted as an expected enrichment in functional portions for the driver gene list compared to the rest of human genes. The not significant overlap instead could be interpreted as an enrichment of singletons caused by the great amount of tumor suppressors but also as a lack of connections between the two lists from the domain point of view. We performed LowMACA analysis in order to find significant hotspots of mutations at two different levels: 1) all the domains were analyzed by aligning the specific sequences of each HCD and CDG that harbors them and 2) the entire protein

was scanned for hotspots considering just its sequence, without any alignment. The second analysis was performed to look for protein-specific hotspots that could be found outside of the Pfam domains and to prevent the exclusion of genes that are not considered by the Pfam-A database (e.g. *WT1*). Obviously, conservation plays no role in this case. Our results identified hotspots of mutation in 11 out of the 137 Pfam domains that were found only in CDG (8%), 32 out of the 86 Pfam domains that were shared both by CDG and HCD (37%) and 188 Pfam domains that were found only in HCD (53%). The higher number of domains that were found significant in HCD compared to CDG reflects the increased number of mutations in each category. Overall, 52 out of 144 candidates (36%) and 177 out of 291 drivers (60%) are supported by LowMACA analysis, either by single sequence analysis or Pfam analysis. Hotspots that are supported with single-sequence analysis (found in 140 genes for HCDs and in 35 genes for CDGs) highlight genes that do not need further support from Pfam companion genes for their identification. Pfam analysis added support to further 37 driver genes and 17 candidates. Compared to the number of genes identified on single sequences, the analysis of the Pfam domains increased the number of identified genes by 26% in HCD and by 50% in the CDG categories, reflecting the fact that LowMACA is particularly useful in identifying genes that mutate at low frequency. In fact, the major gain is found in the CDG category whose genes are typically less frequently mutated. To better characterize recurrence of Pfam domains within the CDG and HCD genes, we built a group of networks where vertices are Pfam domains and edges connect domains that are included together in at least two protein sequences (Figure 18).

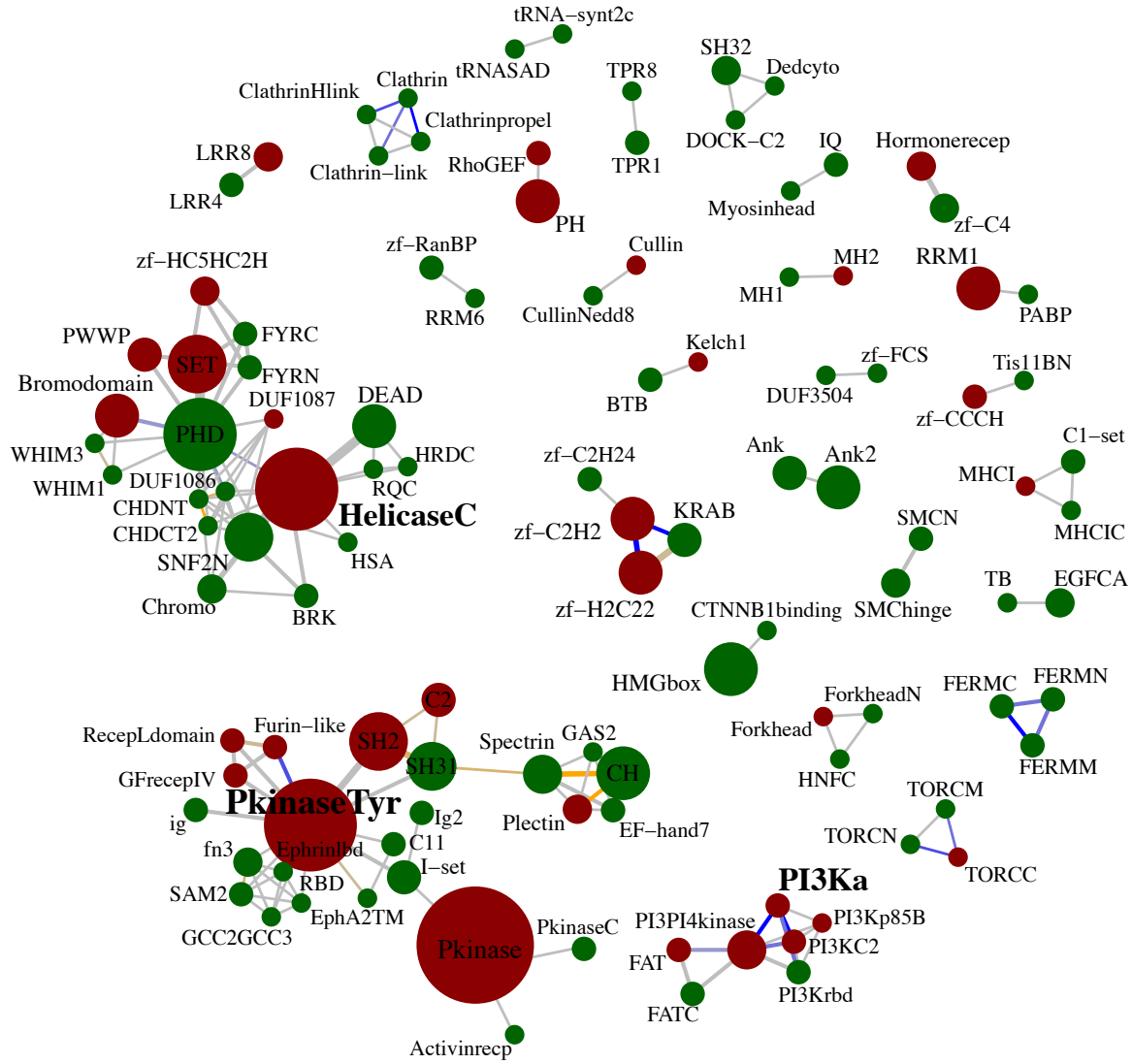

Figure 18 Complete Network of Pfam domains harbored by driver genes. 577 different domains are included in the list of 453 driver genes from Tamborero *et al.* that are represented by each circle in the plot. In red, we highlight those Pfams that harbor at least one significant hotspot, in green those that resulted not significant. An edge connects two domains if at least two genes harbor both the Pfam domains at the vertices. Blue edges are drawn if the domains are mutually exclusive (Fisher test < 0.05 is light blue, < 0.01 is deep blue and OR < 1), yellow if co-occurrening (Fisher test < 0.05 is light brown, < 0.01 is orange and OR > 1), grey if not significant. Excluding domains connected by only one gene, 110 out of the 577 are represented in this figure.

The three main connected graphs are represented by the "PkinaseTyr" network, the "PI3K" network (Figure 17, Panel C) and the "HelicaseC" network, which were named after their main hub. The "PkinaseTyr" network encompasses major oncogenes like *BRAF*, *EGFR*, *FLT3* and *ERBB2* for PF07714 (Pkinase_tyr, Additional file 2: Table S3 highlighted in green) and *STK11*, *CHEK2*, MAPKinases (*MAP3K1/3/4*) and activin receptors (*ACVR1B*) for PF00069 (PKinase). We specifically analyzed the 10 domains which resulted significant with LowMACA and represent them as a heatmap (Figure 17, Panel B): mutated subjects in at least one of the Pfam sequences are depicted in red,

while subjects with a wild type domain are depicted in blue. For many tumor types, in particular bladder (BLCA, in black), breast (BRCA, in red) and colorectal (COADREAD, in orange), a clear mutually exclusive pattern is visible, where subjects with mutations in Pkinase have a wild type tyrosin kinase and vice versa (p=5.18e-60, Odds Ratio 0.26 under Fisher exact test). In glioblastoma (GBM, in green), the majority of patients have a mutation on the Furin-like domain (PF00757), mutually exclusive with tyrosine kinases. The most studied missense mutation in this tumor type is in fact *EGFR* A289V/D/T, known for being resistant to anti-EGFR inhibitor used in lung cancer (Vivanco et al., 2012). This alanine residue is perfectly conserved within the Furin-like domain among other epidermal growth factor genes and appears mutated also in *ERBB2* and *ERBB4*, although not in glioblastoma.

The "HelicaseC" network encompasses genes of various families, which are not strictly connected to each other at the functional level. The Helicase_C domain (PF000271) is the largest significant member of this module and encompasses HCDs as *CHD4*, *SMARCA4* and *ATRX* with two highly conserved arginine residues mutated at low frequency in various tumor types. These mutations affect the corresponding arginine of *CDH7*, *SMARCAD1* and *DDX3X*, which are considered as candidate drivers by the analysis of Tamborero and colleagues (Tamborero et al., 2013b).

The "PI3K" network is instead a strictly interconnected module with a strong degree of mutual exclusivity between the domains that compose it (blue edges in (Figure 17, Panel C). The mutations in these Pfam domains belong for the large majority to three main HCDs (*PIK3CA*, *PIK3CB* and *PIK3CG*). In particular, *PIK3CA* is one of the most mutated genes in many types of cancers. The most relevant mutations appear to be in position 24, 27, and 28 of the multiple alignment of PF00613 (PI3Ka domain) that correspond to E542, E545 and Q546 in *PIK3CA* (Appendix Table 6). These mutations can be found conserved also in the other two HCDs at low frequency and a similar role

has been already assessed for *PIK3CB* (Pang et al., 2009). As we have shown, the overlap between Pfam domains in HCDs and CDGs is not significantly higher than expected from random sampling. This suggests that the current concept of driver genes could be biased due to inappropriate consideration of infrequently mutated genes within the same family. For this reason, we decided to extend our analysis to other possible candidates not present in the list of Tamborero *et al.* (Tamborero et al., 2013b) in the same way as we did for the Ras family. We thus analyzed all the proteins within the following Pfam domains: PF00794 (PI3K_rbd) PF00792 (PI3K_C2) PF00454 (PI3_PI4_kinase), PF02192 (PI3K_p85B) and PF00613 (PI3Ka). These domains are all shared by the 3 aforementioned HCDs and encompass the majority of their mutations. We found low frequency mutations in *PIK3C2A*, *PIK3C2G* and *PIK3CD,* other members of this kinase family, which were never considered as potential driver candidates before. The first two genes belong to the class II of PI3Ks and their role in human diseases is still unclear (Vanhaesebroeck et al., 2010). *PIK3CD*, instead, belongs to the same class I of *PIK3CA/B/G* and has been found amplified or overexpressed in cancer (Kok et al., 2009).

4.2.4.5 Analysis of silent mutations

We run as a negative control a LowMACA analysis using a database of silent mutations on the Pfam domains which were involved with a major role in the previous sections: Ras supefamily (PF00071), Pkinase_tyr (PF07714), Helicase_C (PF00271) and PI3Ka (PF00613). This analysis is aimed at assessing whether non-random pattern emerge from silent mutations. We downloaded TCGA data from TCGA original repositories and performed the analysis on this subset since the cBioportal database exclude silent mutations. The analysis of 676, 1144, 216 and 37 silent mutations that fall on the Ras, Pkinase_tyr, Helicase_C and PI3Ka, respectively, do not show any significant hotspot. On the contrary, 5 hotspots are identified in Ras domain, 10 in Pkinase_tyr, 2 in

Helicase_C and 3 in PI3Ka when analyzed with non-silent mutations (Figure 19).

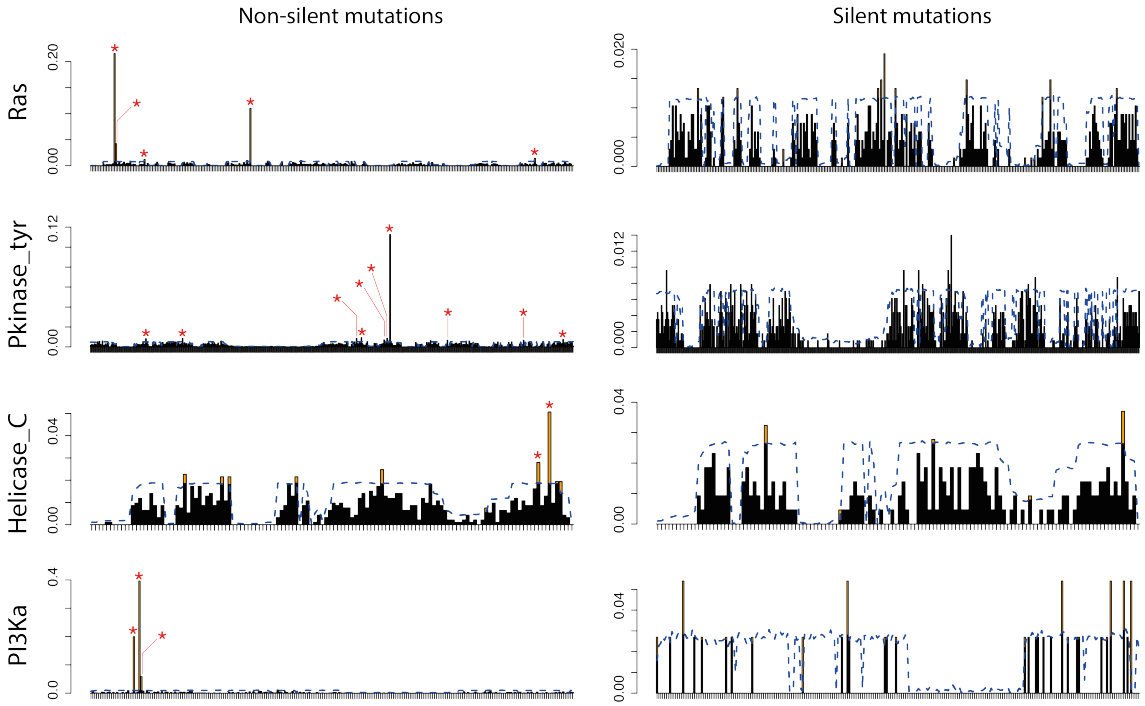

Figure 19 Barplots showing the stacking of silent and non-silent mutations within the 4 main Pfams discussed within the text (PF00071 - Ras superfamily, PF07714 - Pkinase_tyr, PF00271 - Helicase_C and PF00613 - PI3Ka). On the x-axis it is depicted the position in the global alignment, while on the y-axis the mutation frequency of ach position. The blue dashed line represent the threshold of significant mutation frequency, which is different for each position of the global alignment. Bars above the dashed blue line are significant in terms of their p-vlaue. Red asterisks highlight the residues that are significant after Benjamini-Hochberg procedure for multiple testing correction of p-value, which is performed only on conserved positions (see Methods). For silent-mutations analysis, a database was collected from TCGA original repositories and supplied as external repository for LowMACA analysis (see software Vignette for further information). The analysis shows that no hotspots are identified in any of the Pfams checked with the use of the silent mutations database, while at least two hotspots per Pfam are identified when the repository of non-silent mutations is used (canonical analysis). In particular, 5 hotspots are identified in Ras domain, 10 in Pkinase_tyr, 2 in Helicase_C and 3 in PI3Ka.

4.2.5 Discussion

We developed LowMACA, a software aimed at characterizing low frequency mutations involving specific residues within the consensus sequence of protein families. LowMACA maps the mutations observed in different members of a protein family to the multiple alignment of the family members. The resulting consensus protein is suitable to summarize the mutation patterns of different proteins and increases the amount of information on functional domains and their possible role in cancer. All the mutations selected by LowMACA frequently fall upon specific positions of the consensus protein and these can be considered as “highly conserved” in cancer.

Moreover, we have identified patterns of statistically significant mutual exclusivity (mutex) among the identified mutations. The presence of these patterns helps to clarify the meaning of all the mutations belonging to specific pathways indicating exclusive roles of the involved genes in cancer. For example, the mutex analysis between *RAC1* and *NRAS* in skin melanomas (Figure 15) confirms the relevance of the role of *RAC1*, which is co-mutated with *NRAS*, in gain-of-function oncogenic GTP mediated events. The *RAC1* P29L mutation has been experimentally expressed in C. Elegans neurons displaying defects in axon guidance and branching errors that were not seen in equivalent transgenic lines expressing wild-type *Rac1*. Loss of function of the *Rac1* gene did not show any pattern of alteration of axon guidance, demonstrating that *Rac1* P29L is a gain of function mutation (Alan and Lundquist, 2013). These results suggest that a sort of "code switch" between mutations in *NRAS* and in *RAC1* occurs, probably generating different patterns of cell migration. Translating the experimental observations concerning *RAC1* from a neuronal system to cancer is not straightforward. However, it is tempting to speculate that cancer can orchestrate a complex mechanism of choices depending on the environmental context where it develops. The mutex analysis between Rho members and the *RAS trio* in cancer represents an example of how one out of the many mechanisms underlying cell growth and metastatic processes can provide a selective advantage to cancer cells.

The identification of mutex patterns concerning other proteins belonging to the Ras family suggests that beyond *KRAS*, *HRAS* and *NRAS* other minor genes, such as *RRAS2*, could play a "Ras-like" role in promoting pro-proliferative activity via the activation of the Raf-ERK pathway of MAP kinases (Gutierrez-Erlandsson et al., 2013) in uterine and cervical cancers (Figure 15). This finding supports the hypothesis that *RRAS2* has a vicariant role in wild type *KRAS* cancers. Other mutual exclusivities have been observed between *HRAS* and *RHOA*, in head and neck squamous cell carcinoma (HNSC) and

between *DIRAS2* and *KRAS* in colorectal cancer. The phenomenon by which minor proteins in a family domain can harbor the "same" mutations harbored by known drivers is observable also in other Pfam domains encompassed in the PI3K family. These findings highlight a possible role of minor members of this kinase family in cancer (e.g. *PIK3C2A*, *PIK3C2G* and *PIK3CD)*. LowMACA allows focusing on this phenomenon and helps formulating a possible explanation: cancers cells that gain a selective advantage from major driver mutations in one type of cancer may gain a similar selective advantage from corresponding mutations in closely related proteins in other types of cancer where the related protein plays a prominent role due to tissue specific differences in gene expression or environmental constraints such as exposure to therapeutic agents. In extending LowMACA analyses to other Pfam domains we also demonstrated the existence of liaisons among genes considered high confidence drivers with other genes that are considered candidate drivers. The presence of low-frequency mutations in *ERBB2* and *ERBB4* that correspond to known driver mutations in tyrosine kinases such as *EGFR*, *BRAF*, *FLT3* and *JAK* further strengthens this concept.

Nevertheless, Ras subfamilies also show specific hotspots that reflect the subtle differences played by genes of each subfamily in cellular homeostasis. The Rho subfamily genes have roles in regulating cytoskeletal dynamics and deregulation of Rho proteins contributes to tumorigenesis and metastasis, while Ras subfamily proteins mainly function in regulating cell proliferation (Wennerberg, 2005).

LowMACA is intended as an algorithm that emphasizes low-frequency mutations in genes containing a Pfam domain. Nevertheless, we cannot generalize this concept to all driver genes. For example, genes such as *TP53*, *VHL*, *RB1* or *APC*, show distinct patterns of somatic driver mutations that are not shared by other members of their family (like *TP63* and *TP73* or *APC2*). These tumor suppressors should be considered as singletons and this characteristic underlines the difference between tumor suppressors and

oncogenes. Thus, LowMACA is particularly useful for the identification of gain-of-function mutations in putative oncogenic families.

LowMACA emphasizes the role of genes mutated at minor frequency in cancer, which are often neglected by current analyses. The possibility to classify patients associated to signatures of low-frequency mutations identified by our software represents a promising route for future work. At the same time, a more accurate classification of driver genes may shed light on molecular mechanisms underlying cancer that until now were not yet considered.

4.3 A knowledge-based framework for the discovery of cancer predisposing variants using large-scale sequencing breast cancer data.

This last section of the results represents an attempt to apply a genomic-based approach, like the ones seen in the first two sections, to a typical case-control genetic study on exome sequencing data. What is generally performed in a genetic case-control framework is to seek for SNPs with a strong imbalance in terms of allele count between cases and controls, taking into account how rare or common is the SNP in the population. While this seems to be the most unbiased approach, it is hard to reach the required statistical power when using exome sequencing data because the number of SNPs to test is extremely high (possibly millions of tests) and the cost to reach an adequate sample size is impractical. What we know is that cancer has a double "mutation" mechanism. It is not only a familial disease but also a somatic disease, with changes in DNA at the somatic level that, as pointed out in *Rhaman 2014*, are substantially overlap with each other. At least half of the known cancer predisposing genes are also known somatic driver genes and that suggests that what we know about

the somatic changes can be used to get more insights into predisposition. The search of somatic mutations that correspond to germline variants was the starting point of this section that has been further expanded to design a complete framework for predisposing variants discovery.

4.3.1 Abstract

The landscape of cancer predisposing genes has been extensively investigated in the last 30 years with various methodologies ranging from candidate gene to genome-wide association studies. However, sequencing data are still poorly exploited in cancer predisposition studies due to the lack of statistical power when comparing millions of variants at once.

Here, to overcome these power limitations, we propose a knowledge based framework trained on the characteristics of known cancer predisposing variants and genes. Under our framework, we take advantage of a combination of previously generated datasets of sequencing experiments to identify novel breast cancer predisposing variants comparing the normal genomes of 673 breast cancer patients of European origin against 27,173 controls matched by ethnicity.

We detect several expected variants on known breast cancer predisposing genes like *BRCA1* and *BRCA2* and 19 variants on genes associated with other cancer types, like *RET* and *AKT1*. Furthermore, we detect 185 variants that overlap with somatic mutations in cancer and 50 variants associated with 41 possible loss-of-function-genes, including *PIK3CB* and *KMT2C*. Finally, we find a set of 19 variants as potentially pathogenic and negatively associated with age at onset that have never been associated to breast cancer.

In this study we demonstrate the usefulness of a genomic-driven approach nested in a classic case-control study to prioritize cancer predisposing variants. In addition, we provide a resource containing variants that may affect susceptibility to breast cancer.

4.3.2 Introduction

Breast cancer is one of the most common cancers with greater than 1,300,000 cases and 450,000 deaths per year worldwide (Koboldt et al., 2012) and it is estimated that ~5-10% of women have germline mutations that lead to hereditary predisposition to breast cancer (Ripperger et al., 2008). Specific mutations in *BRCA1* and *BRCA2* are known to be responsible for inherited susceptibility to breast cancer in families with early-onset disease(Campeau et al., 2008). In particular, it has been demonstrated that the risk for first-degree relatives of an affected person is increased by two-fold and BRCA1/2 mutation carriers account for just 20% of this enhanced risk (Ripperger et al., 2008). Mutations in other genes, such as *PALB2*, *PTEN* and *TP53*, have been associated with increased risk of breast cancer. Unfortunately, many familiar breast cancers (~50%) are still unexplained at the genetic level and many predisposing variants are yet to be found (Fachal and Dunning, 2015).

Historically, beside the use of linkage analysis, which requires families with a penetrant phenotype, the analysis of candidate genes has allowed the discovery of the majority of well-known cancer predisposing genes (Rahman, 2014). Conversely, genome wide association studies (GWAS), which have been extensively used, have the ability to discover cancer predisposing genes on a genome wide scale with a pure data driven approach but suffer from lack of precision (Ward and Kellis, 2012). In fact, the results of GWAS can only indicate regions where the real pathogenic variants actually reside. The use of whole-exome sequencing (WES) and whole-genome sequencing (WGS) data, although able to overcome the aforementioned limitations of GWAS, has been poorly exploited due to power limits imposed by testing millions of variants simultaneously. Furthermore, Next Generation Sequencing (NGS) data are more expensive and less reliable in terms of accuracy of the genotype call, so that a statistical power comparable to the largest GWAS studies is technically unreachable (Zheng et al., 2013). Nevertheless,

WES or WGS have the advantage that they can identify rare variants that may influence cancer risk while the concept of linkage disequilibrium (LD) used by GWAS mainly relies on common single nucleotide polymorphisms (SNPs) with minor allele frequency (MAF) generally greater than 5%. Actually, we can hypothesize that only a fraction of the heritability can be ascribed to common genetic variants while rare variants can convey the remaining heritability (Kiezun et al., 2012).

A straightforward case-control comparison on allele frequencies would be both underpowered and incomplete, since most rare potentially pathogenic variants would be excluded because they lack the required statistical power. Thus, we propose a computational framework that is trained on our knowledge of the characteristics of known cancer predisposing genes and variants. Under our framework, we studied the normal genomes coming from 673 breast cancer patients of European origin from The Cancer Genome Atlas (TCGA) against over 27'000 control genotypes unselected for cancer phenotype from the Exome Aggregation Consortium (ExAC) database with matched ethnicity (Figure 20).

We filtered and integrated allele-counting comparisons with custom annotations coming from the state-of-the-art databases, somatic mutations data and GWAS studies to assess the probability of facing a true pathogenic variant. In particular, we take advantage of the characteristics of somatic driver genes (like their gain or loss-of-function) to emulate a candidate gene analysis. Cancer is in fact a unique case where disease causes and disease predisposition are strongly tightened, with a clear definition of gain-of-function/oncogenes and loss-of-function/tumor suppressor genes (Rahman, 2014). Giving cancer unique characteristics, we were able to prioritize a set of genes and variants with a top-down approach in contrast to GWAS analysis: first, we isolate the best candidates through fine-mapping and second, we rank them using statistical analysis.

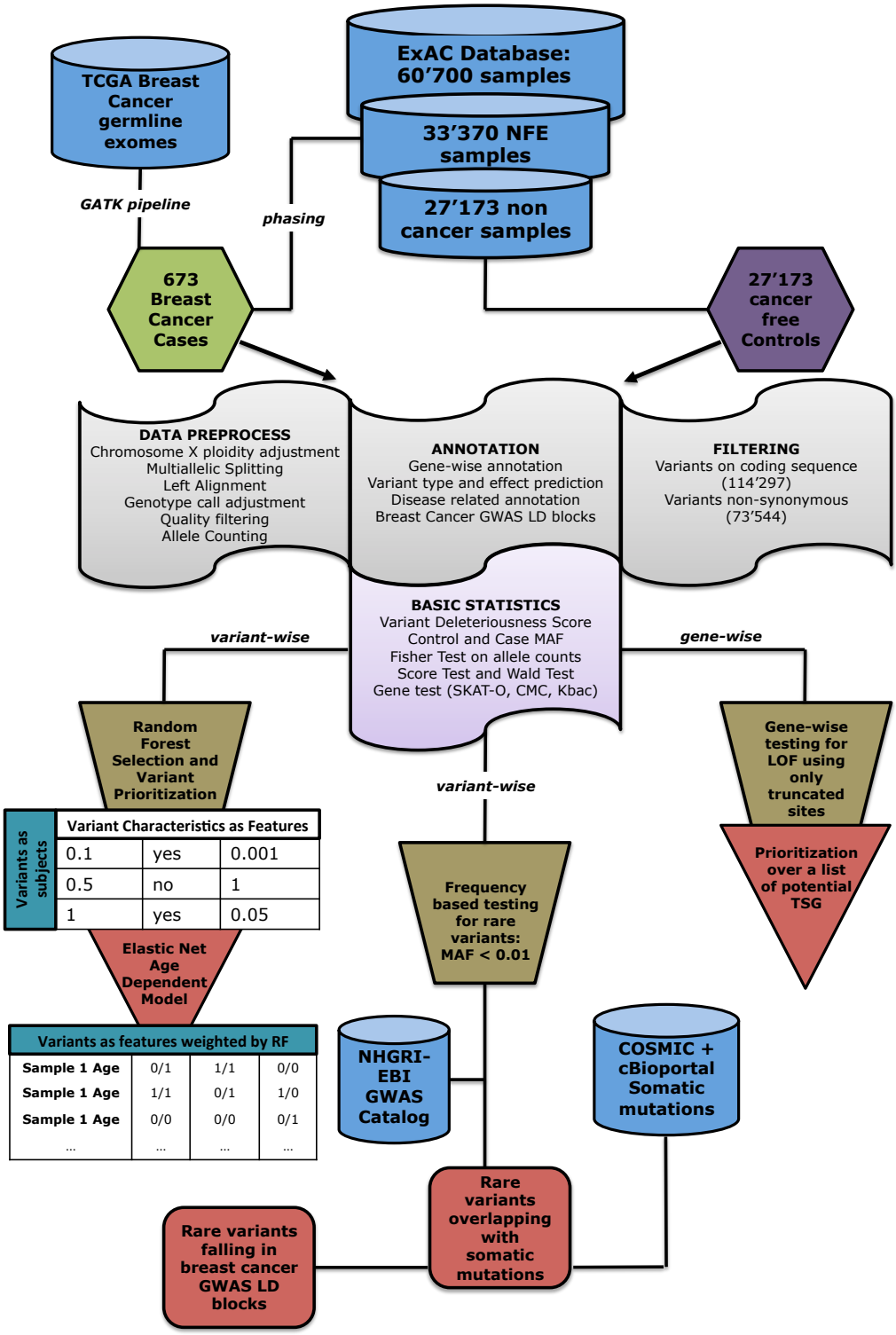

Figure 20 Workflow scheme for the whole analysis. Blue cylinders represent data (both obtained from available databases or processed during the analysis), hexagons are the analyzed datasets of cases and controls, while red squares and triangles represent analysis and output. Flag shapes represent post process annotation and statistical testing. Brown trapezoids represent the three main analysis branches presented in this paper.

4.3.3 Materials and Methods

4.3.3.1 Control Data

We used the aggregated results from the ExAC database (http://exac.broadinstitute.org/) as control population (Lek et al., 2016). This resource aggregates data from more than 60'000 samples with germline genotype data, of which 33'370 were classified as of European origin. The original data source is both from population studies (1000 Genome Project, HapMap, Exome Sequencing Consortium) and from disease related studies (including part of the TCGA). To overcome the overlap with tumor samples, we used the data cleaned from any cancer sample, for a total of 53'105 samples of which 27'173 are of Caucasian origin.

4.3.3.2 Case Data

We downloaded from the TCGA (http://cancergenome.nih.gov/) the original BAM files of the normal sample for all the 695 women and men of Caucasian origin diagnosed with breast cancer. We used 673 (7 men and 666 women) out of the 695 samples discarding whole genomes and samples not derived from blood specimens. We analyzed the BAM files, following the exact same GATK pipeline and the same level of sensitivity used by ExAC (see section **4.3.3.4**). We retrieved from the TCGA open access database, the available clinical information for these patients, including age and sex.

4.3.3.3 Annotation Data

To better characterize our variants, we took advantage of several external databases and in-house datasets for annotation. In particular, we used ANNOVAR to obtain information on gene, protein change, and type of variant (missense mutations, truncating mutations, InDels etc.) (Wang et al., 2010). Only variants found within the coding sequence and classified as non-synonymous were retained. ANNOVAR also annotated the variants with the 9 different tools for prediction of phenotypic effect

(SIFT, Polyphen2_HDIV, Polyphen2_HVAR, LRT, MutationTaster, MutationAssessor, FATHMM, RadialSVM, LR) that are included in dbNSFP (Liu et al., 2013). We summarized this information as a comprehensive deleteriousness score calculated as the proportion of tools that calls a particular variant as damaging or probably damaging. Finally, ANNOVAR provided information about the presence of the variant as a somatic mutation in the COSMiC database (Forbes et al., 2011) and if the mutation is present in the ClinVar database (Landrum et al., 2014). This information was integrated with custom annotations from other public resources. We integrated our annotation by checking for the presence of the variant in the cBioPortal database (Cerami et al., 2012) for the same amino acidic change and in other database of known disease causing and cancer related mutations: CIViC (https://CIViC.genome.wustl.edu), DoCM (http://docm.genome.wustl.edu/) and Humsavar (http://www.uniprot.org/docs/humsavar). These resources provided the evidence of overlap between our case variants and somatic mutation in cancer. The annotated genes were flagged if they belong to three categories according to biological and cancer related characteristics: genes known to be predisposing for cancer, genes known to be driver in cancer at somatic level and genes involved in DNA repair. The first list was created based on the most recent literature, including (Futreal et al., 2004; Rahman, 2014; Vogelstein et al., 2013; Walsh et al., 2010) and represents the state–of-the-art of the knowledge on cancer predisposing genes (323 genes). The list of known somatic drivers was created using an in-house tool for detecting driver genes (Melloni et al., 2014) and adding the most recent literature and state-of-the-art tools (Davoli et al., 2013; Dees et al., 2012; Futreal et al., 2004; Lawrence et al., 2014; Vogelstein et al., 2013) for a total of 413 genes. Finally a comprehensive list of genes involved in DNA repair was retrieved from (Lange et al., 2011) including 166 genes. In total, we considered as our target gene list a total of 758 unique genes. In addition, we also tried to classify these genes as potential oncogenes

or tumor suppressors based on the joint results of (Davoli et al., 2013; Melloni et al., 2014; Rahman, 2014; Schroeder et al., 2014; Vogelstein et al., 2013). The tools and literature used in this classification are the ones able to distinguish between tumor suppressor and oncogenes based on their mutational pattern or experimental results. In case of discordant results, the gene is considered both as tumor suppressor and oncogene. In total, we were able to classify 119 genes as oncogenes and 235 genes as tumor suppressors based on the results of both literature and bioinformatics algorithms. It is noteworthy that, when the same gene is found both as predisposing gene and somatic driver, known predisposing loss-of-function genes corresponds to somatic tumor suppressors and known gain-of-function genes to driver oncogenes. The extent of this overlap between cancer predisposing genes and somatic driver genes has been estimated to be over the 50% of the cancer predisposing genes list (Rahman, 2014).

We retrieved a dataset of known breast cancer associated SNPs from GWAS studies included in the Human Genome Research Institute's Catalog of Published Genome-wide Association Studies (version 2016-05-08) (Welter et al., 2014). A p-value of $5X10^{-8}$ was used as threshold. We manually selected the publications included in the catalog under the ontology "breast cancer" with a study in a fully European origin cohort during discovery phase and presenting SNPs associated with the disease and not with some of its characteristics. For example, we removed studies about drug resistance, chemotherapy adverse events or levels of proteins in breast tissue. These variants are not directly associated with the disease but represent a flag for a probable region where the disease-causing variant could be found. In total we collected a list of 130 SNPs from 23 studies. Using HapMap recombination data (Frazer et al., 2007), we created the boundaries of such regions as all the DNA regions surrounding the GWAS SNPs below a recombination rate of 20 cM/Mb (Machiela et al., 2015). If one of our variants fall into one these regions, its distance from the flag GWAS SNP is annotated. In fact, there is no

direct relationship between physical distance and genetic distance but since we are inside low recombination regions, we can consider bp distance as a proxy for cM distance.

4.3.3.4 Data Preprocess

Case data preprocess was based on the whole GATK pipeline used by the ExAC consortium (Lek et al., 2016). This included Picard MarkDuplicates, local realignment around InDels, base quality recalibration, haplotype call, joint genotyping and variant quality score recalibration (http://picard.sourceforge.net/) (McKenna et al., 2010). Our pipeline included also splitting multiallelic sites and left aligning them both for case and controls in order to obtain a perfect match of Chromosome, Position, Reference and Alternative alleles. Working with biallelic sites is generally preferred, especially during annotation. A multiallelic site in fact, would have complete different variant effect according to the alternative alleles. The genotype call was retained if the genotype quality was higher than 20 and the depth of sequencing was higher than 10. Such filter was used to obtain robust genotype calls. Even if breast cancer is way more common in women compared to men, our case dataset is composed by 7 men and 666 women. We therefore fixed the ploidity for men on chromosome X in non-pseudo autosomal regions. For every heterozygous call, only the most probable allele was retained and one single allele was counted for every homozygous call. We used bcftools/vcftools, variant tools (vt) and in-house scripts (Danecek et al., 2011; Tan et al., 2015) to process the post-call data.

We created a custom allele counting procedure so that for every biallelic variant, the reference count was constant for every possible alternative allele. This procedure is not standard but allowed us to test every alternative allele against the exact number of reference alleles called at the site and it is useful in case of filtering for any criteria because the number of reference alleles is never lost and we could easily re-aggregate the data to create a multiallelic test as explained in the *Statistical Analysis* section.

4.3.3.5 Statistical Analysis

As mentioned in the introduction, statistical power is a critical issue in genome wide case-control studies. In particular, exome data are even more underpowered compared to GWAS since potentially millions of variants can be tested at a time. The initial call from all the 673 samples included millions of variants that were filtered to keep only coding and non-synonymous events. Since we did not perform any imputation and we applied a strict quality filter after the raw calls, retaining only exonic variants was the best way to maximize coverage in a dataset composed for the large majority by exome sequencing data (including all cases). At this point, we divide the testing procedure into different branches (Figure 20):

1. Frequency and annotation based analysis
2. Loss-of-function gene-wise testing
3. Age-dependent polygenic modeling

4.3.3.5.1 Frequency and annotation based analysis

The frequency and annotation based analysis is made up of simple annotation and filtering step-wise procedure (as summarized in Figure 22). Rare (control MAF below 1%) non-synonymous variants were retained and only the ones with a case MAF greater than the control are kept. Subsequently, only damaging mutations with a deleterious score of at least 0.5 (majority of tools for prediction of phenotypic effect considering the mutation as damaging) were selected. The pipeline is then divided in two branches. On one side (left arm of Figure 22) we sought for those variants that classified as somatic in any type of tumors using COSMIC and cBioPortal databases. On the other side, we took into consideration only variants that fall into LD blocks of previously annotated breast cancer associated SNP in GWAS studies.

4.3.3.5.2 Loss-of-function gene-wise testing

The LOF testing is a gene-wise test that seeks for imbalance in allele count in truncation events between cases and controls. In this testing procedure, we looked for truncations, frame shift InDels or nonsense mutations that retain by definition a higher probability of creating a loss-of-function event. In this context, we wanted to emulate the way driver somatic tumor suppressors genes are generally discovered, using the frequency of any rare truncation controlled by the same frequency in control cohort (Davoli et al., 2013; Melloni et al., 2014; Vogelstein et al., 2013). We first filtered out common events (over 5% in the control cohort) and then we performed, for every variant, a simple one-tail fisher count test between minor/major allele count between cases and controls. For every gene, we aggregated all the p-values obtained from the tests using Stouffer method (Stouffer S et al., 1949) to obtain a single value per gene. In this procedure a weight that is proportional to the inverse of control frequency, was applied so that the more a variant is rare, the higher is the weight applied in the aggregation step. Finally, we retained only genes belonging to our target gene list with an FDR corrected p-value below 0.05.

4.3.3.5.3 Age-dependent polygenic model

The Age-dependent polygenic model branch is instead a stepwise procedure. Like for the LOF procedure, we calculated a minor/major fisher count test variant-wise between cases and controls, including all variants, without applying any filter to the MAF of the control cohort. For multiallelic sites, reference and alternatives composed a matrix of 2*(n+1) where 1 represents the reference counts and n the number of different alternative alleles. A bootstrap version of fisher test was used in this case. The calculated p-values were added as an explanatory variable in the step-wise procedure. We also run other commonly used human genetics statistical tests to be added as explanatory variables, both gene-wise and variant-wise. Using RVtest, we were able to run SKAT-O, CMC, Kbac tests for the gene-wise level and Wald and SingleScore tests for the variant-

wise level (Zhan et al., 2016). All the aforementioned tests used the 1000 genome original genotype calls as control cohort, since we need the full genotypes in order to run them and they were not available for the ExAC database.

The workflow is composed by:

1) Creating a set of variants that accounts for every variation in our case samples reported as pathogenic in at least one of the following databases: Humsavar, DoCM, ClinVar or CIViC. Variants were further subset for a manually curated list of cancer related keywords. This list included both direct cancer or neoplastic events predisposition as well as cancer related syndromes (like Li-Fraumeni or Von Hippel Lindau syndromes) for a total of 38 variants in 24 different genes (Appendix Table 7)
2) Creating a set of negative controls from the list of ClinVar annotated variants that have been tested as non-pathogenic.
3) Implementing a random forest classifier using a dichotomous response variable (pathogenic, non-pathogenic) with a training set that included all the variants in point 1) and 2) (Breiman, 2001). The features used for classification are reported in Figure 23 and included all the tests described above, MAF in cases and controls, number of homozygous and heterozygous calls, deleteriousness score and a dummy variable describing if the variant was a truncation event or a simple missense variant. A tree based algorithm like the random forest is particularly powerful in problems where the interaction of various features is a critical point for the model. The output of this analysis was the relative number of trees classifying a variant as pathogenic. We call this score Pathogenicity Score.
5) The variants that did not belong to the training set were selected and we filtered for the ones that show a Pathogenicity Score of at least 0.5 (majority voting in the random forest procedure)

6) The final step of the model was to correlate our variants with the age at initial pathological diagnosis. The variants found at point 5) switched from being subjects to become explanatory variables with a value of 0, 1 and 2 according to the state of double major allele, heterozygous or homozygous minor allele. With such a dataset, we build a robust elastic net linear model by running 100 models in parallel under various subsets of the dataset (Zou and Hastie, 2005). This procedure guarantees that the average beta values and the number of times a feature is in all the elastic net models remain stable. A penalized linear model like the elastic net is preferable for its ability of assessing the direction and the magnitude of the contribution of each variant. In the case of the age for example, we are interested in understanding what are the variant with a negative beta or, in other terms, the ones that contributes to the decrease of age at onset. Considering that a male patient has a risk of getting cancer one hundred times less than a female of the same age, a male patient age is rescaled with a logit function in order to correspond to the same risk of a younger woman. To build the risk function at age classes, we used data from the Cancer Research UK report 1996-2011, available at http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer.

4.3.4 Results

4.3.4.1 Pathogenic and Breast Cancer Related variants

We first asked whether known breast cancer predisposing variants were present in our dataset. We collected from the literature a list of 17 known breast cancer susceptibility genes (Table 6) (Rahman, 2014; Vogelstein et al., 2013; Walsh et al., 2010). We expected to find some pathogenic variants in these genes, as they comprise a part of the known genetic cause of breast onset.

Gene	Somatic Driver Gene	Total Number of variants	Number of Pathogenic Variants	Number of Truncating Variants	Number of highly damaging mutations
ATM	**X**	21			5
BRCA1	x	18	2		3
BRCA2	**X**	21	5		2
BRIP1		5	1		
CDH1	**X**	3	1		
CHEK2	x	6	2	1	
MRE11A		4			1
NBN		5	1	2	
PALB2		1			
PRKAR1A					
PTEN	**X**				
RAD50		5			
RAD51C		3		1	
STK11	x	3			
TP53	**X**	4			1

Table 6 List of the most important breast cancer predisposing genes and variants found in our case dataset. The second column reports if the gene is considered also a somatic driver gene (breast cancer somatic driver genes are reported with a bold capital X). Other columns report the number of non-synonymous variants found in total, the number of variants considered pathogenic and the number of rare truncating variants (control minor allele frequency below 1%) that are not already included in the list of pathogenic variants. The last column shows instead all the missense variants that are not considered pathogenic with a very high deleteriousness score (8/9 tools to predict functional damage report the variant as damaging). Our pathogenic reference is ClinVar and Humsavar databases.

Considering both known pathogenic and truncating variants on these 17 genes (Table 6), we obtained a total of 16 different mutations that cover 36 out of 673 of our cases (~5%). We decided to take into account also rare truncating variants because they are generally considered *de facto* pathogenic when the gene exerts its oncogenic function via loss-of-function. This is the case for all the known predisposing genes in breast cancer and, in general, for the large majority of Cancer Predisposing Genes (CPGs) (Rahman, 2014). The frequency of the identified variants in the breast cancer dataset is compatible with a sample of sporadic cases, especially given the fact that many potential pathogenic variations are still not reported in databases like ClinVar (Landrum et al., 2014). Furthermore, the complete lack of any variation on *PTEN* and *PRKAR1A* can be explained by the rarity of finding mutations on these genes. The cancer syndromes connected to these genes (Cowden Syndrome and Carney Complex) are in fact extremely infrequent in the population: the first has an incidence of 1 in 200'000

individuals (Hobert and Eng, 2009), the latter a total prevalence of few hundreds reported cases (Stratakis et al., 2001). It is noteworthy that 8 out of 15 of the genes reported in Table 6 are also known somatic driver genes and 5 of these 8 genes are specifically considered driver in breast cancer. All of them are predicted or possess tumor suppressor functions. The second question we asked was whether other cancer pathogenic variants could be found in our case dataset. It is in fact known that many cancer predisposing genes can lead to complex tumor syndromes in which more than one tumor type can arise (Rahman, 2014). Known examples are the aforementioned *BRCA1* and *BRCA2* that are linked to both breast and ovarian cancer (Petrucelli et al., 2010) or the more recent discovery of *PALB2*, connected to breast and pancreatic tumors (Jones et al., 2009; Rahman et al., 2007). We therefore seek for all those variants connected to additional cancer or cancer syndrome genes and we found 28 different variants on 24 genes. Among them, the 19 variants with a control MAF below 1% are reported in Table 7.

Variant	Control MAF	Case MAF	log2 MAF Ratio	Summary of ClinVar and Humsavar Annotation
COL7A1 - R1538C - (3,48619779,G,A)	0.002%	0.07%	5.35	Malignant Melanoma
RET - V804M - (10,43614996,G,A)	0.017%	0.54%	4.96	MEN2A Syndrome|Thyroid Carcinoma
AKT1 - E17K - (14,105246551,C,T)	0%	0.08%	4.47	Colon Ovary and Breast Cancer
FANCC - R185 - (9,97912338,G,A)*	0.006%	0.07%	3.76	Fanconi Anemia
FLCN - H429fs - (17,17119708,-,G)	0.054%	0.70%	3.68	Renal Cell Carcinoma
MSH6 - T955fs - (2,48030639,-,C)	0.213%	2.61%	3.62	Lynch Syndrome
ELAC2 - R741H - (17,12896274,C,T)	0.072%	0.23%	1.66	Prostate Cancer
RET - Y791F - (10,43613908,A,T)	0.244%	0.69%	1.50	MEN2A Syndrome|Thyroid Carcinoma
FLCN - R239C - (17,17125879,G,A)	0.033%	0.08%	1.20	Renal Cell Carcinoma
PKHD1 - T36M - (6,51947999,G,A)	0.075%	0.15%	0.98	Colorectal Cancer
GALNT12 - D303N - (9,101594229,G,A)	0.185%	0.30%	0.72	Colorectal Cancer
PRF1 - N252S - (10,72358722,T,C)	0.501%	0.82%	0.72	Non-Hodgkin Lymphoma
SDHD - G12S - (11,111957665,G,A)	0.992%	1.04%	0.07	Cowden Disease 3
TSC1 - H681Y - (9,135779052,G,A)	0.561%	0.52%	-0.11	Neoplastic Syndrome
AR - R727L - (X,66937326,G,T)	0.083%	0.07%	-0.16	Prostate Cancer
SDHD - H50R - (11,111958677,A,G)	0.975%	0.82%	-0.25	Cowden Disease 3
MUTYH - Y165C - (1,45798475,T,C)	0.256%	0.15%	-0.78	MYH-associated polyposis|Endometrial Carcinoma
APC - R396C - (5,112154969,C,T)	0.155%	0.08%	-1.00	Gardner syndrome
ASCC1 - N290S - (10,73892817,T,C)	0.173%	0.07%	-1.22	Esophageal Carcinoma

Table 7 List of rare cancer-related pathogenic variants (control MAF below 1%). This list includes all those genes that are not considered breast cancer predisposing but are connected to other types of cancer or cancer syndromes.

Although only 13 variants out of 19 have a minor allele frequency in the cases higher than controls, the results from this simple annotation are quite unexpected. For example, we found *COL7A1*, a collagen gene linked to epydermolysis bullosa that is a severe skin syndrome with elevated life-time risk of melanoma (Martins et al., 2009). MAF frequency in our dataset is at least 1 order of magnitude higher than in controls. We also detected two variants on *RET*, a gene connected to MEN2A syndrome with an extremely high penetrant risk of thyroid cancer (Eng, 1999) that to our knowledge has been connected to breast cancer at the level of expression and thus as a possible therapeutic target (Morandi et al., 2011). Evidences of a connection to another thyroid cancer related syndrome (MEN1) have been recently demonstrated in breast cancer (Dreijerink et al., 2014), but a suggestion to a link to MEN2A is completely novel and it would represent an unusual case of an gain-of-function mutation linked to breast cancer risk. Interestingly, we identified 3 truncating or frameshift alterations on *FANCC, FLCN* and *MSH6*, three loss-of-function genes respectively associated to Fanconi Anemia (as *PALB2, BRCA1 and RAD51C* reported in Table 6) (D'Andrea, 2010), renal cell carcinoma (Stamatakis et al., 2013) and Lynch Syndrome (Baglietto et al., 2010), with no previous direct connections to breast cancer. Lastly, we discovered *AKT1* E17K, a variant linked to many cancer types, including breast cancer, at the somatic level. It is reported in databases such as ClinVar or OMIM (that are generally focused on hereditary genetic traits) because it is considered a high frequency somatic driver mutation (Bleeker et al., 2008). This gene has also been connected to a minority of Cowden Syndrome cases along with *PIK3CA* because it belongs to the same pathway as *PTEN,* whose mutations are causative of 85% of the cases (Hobert and Eng, 2009). This variant is particularly relevant because it represents both a case of gain-of-function mutation in a breast cancer

oncogene that is frequently seen somatically mutated in tumors and also a risk associated germline variant in our dataset.

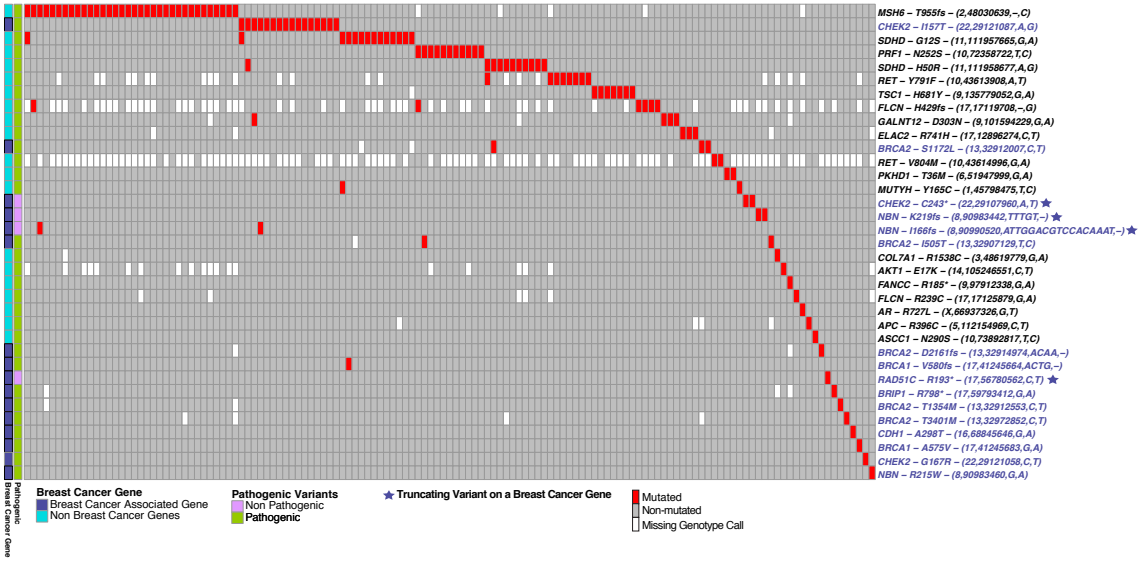

Figure 21 Distribution of pathogenic and truncating variants on breast cancer genes in our case dataset of 673 breast cancer patients. This oncoprint plot reports three classes of high confidence breast cancer predisposing genes (rows) and each column represents one of the samples with at least one of these mutations. In blue, we report variants on known breast cancer predisposing genes (complete list in Table 1). A star is reported if the variant is a truncation but is not reported as pathogenic in databases ClinVar or Humsavar. Otherwise the variant is present in these databases. Pathogenic variants that affect genes related to cancer or cancer syndromes but are not strictly listed as breast cancer pathogenic are reported in black and include genes like *RET* (thyroid cancer) or *APC* (colon cancer)

To summarize our findings, we draw a heatmap of all the aforementioned variants in our dataset (Figure 21). If we sum up all the cases with at least 1 of these mutations, we approximately cover the 20% of our dataset with 19 non breast related pathogenic variations, 12 pathogenic breast related and 4 truncating variants on breast CPGs. It is noteworthy, that *MSH6* alone covers the 5% of patients, although the MAF is much lower because this calculation considers also missing genotype calls and heterozygous and homozygous calls. We can also notice that co-occurent mutations are quite rare: only 13 out of 135 samples have more than one variant, while the remaining 122 are hit by a single variation. Furthermore, variant frequency in the dataset is extremely unbalanced: the top 7 variants in Figure 21 cover the 15% of the patients while the remaining 28 the missing 5%.

4.3.4.2 Analysis of rare variants in target cancer genes

The most simple and straightforward way of prioritizing predisposition candidates is to look at rare variants, which can be defined as variants with MAF < 1% in the controls. We concentrated our efforts on non-synonymous variants (~70'000) and we filter for rare variants where the prevalence in the cases is higher than controls, retaining only ~50'000 variants (see Figure 22).

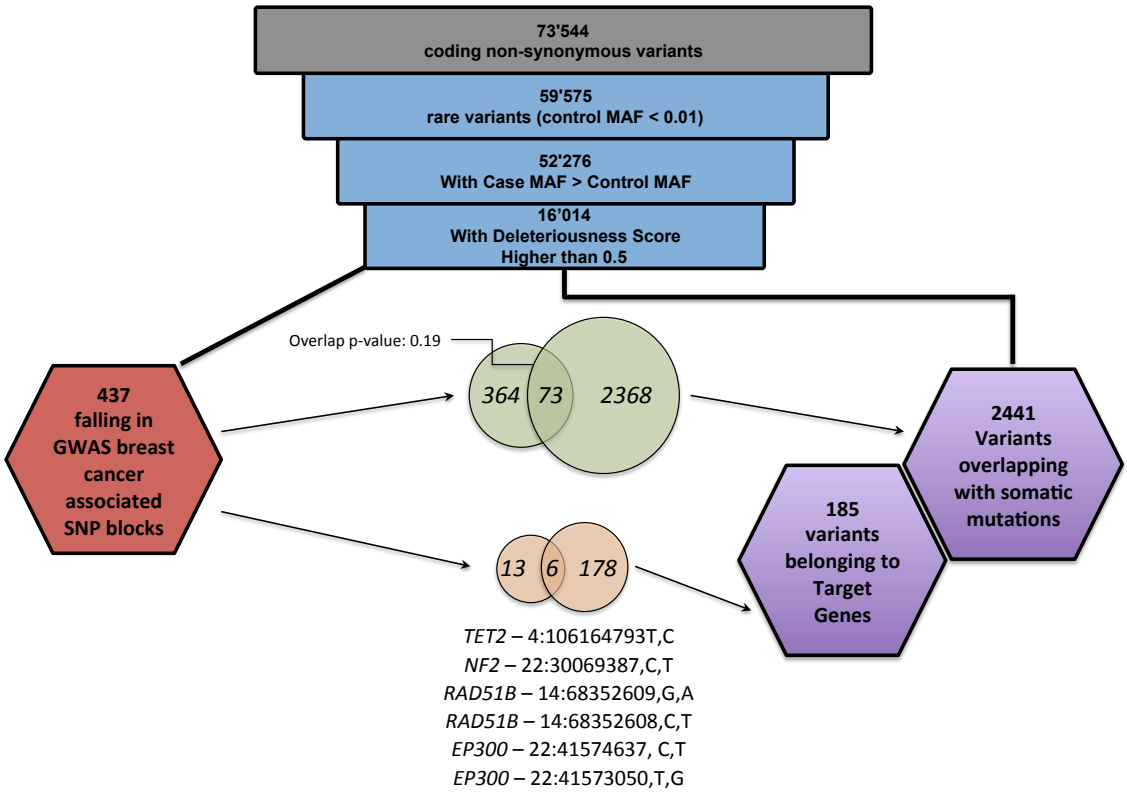

Figure 22 Analysis of rare variants. This flowchart represents the step-wise procedure in the central arm of Figure 20 and is performed by filtering from 73544 coding non-synonymous variants to 16014 rare variants (MAF < 1%) with a deleteriousness score over 0.5 and where MAF in the cases is higher than controls. Rare variants are prioritized into two branches: on the left variants falling in GWAS breast cancer LD blocks are retained, on the right, variants overlapping with cancer somatic mutations from COSMIC or cBioPortal are reported. For both datasets, overlaps are reported at the initial level and after filtering for variants belonging to our list of 758 target genes (known cancer predisposing genes, known somatic driver genes and DNA repair genes). Final common list of 6 variants on our target gene list that was found both as overlapping with somatic mutations and falling into a GWAS LD block is reported.

Then, we filtered out variants with a deleteriousness score lower than 0.5 or, in other words, where 4 or less of the 9 methods included in dbNSFP evaluated the variant as possibly damaging (see section **4.3.3.5.1**) (Liu et al., 2013). The dataset at this is point is composed by over 16'000 variants to prioritize and we look for two specific

characteristics: we explored what are the rare variants that overlap with cancer somatic mutations and we also check if some of them fall into regions of low recombination (thus, possible LD) with breast cancer associated SNPs from GWAS studies to further confirm our results. We used COSMIC and cBioportal databases to create the largest set of somatic variants from WGS and WES studies, including over 50 different tumor subtypes, and we look for a perfect match between our variants and these somatic mutations (Forbes et al., 2011; Gao et al., 2013). To create GWAS blocks we designed regions around 130 manually selected SNPs from the NHGRI-GWAS catalog (see section **4.3.3.3**) (Welter et al., 2014). 2441 variants on 16'000 are also found somatically mutated from cBioPortal database or COSMIC database. Among those matched variants, only 73 falls into GWAS regions over a total of 437 (Figure 22 and Appendix Table 8). The overlap of these two groups is apparently random as this is not significantly different from a bootstrap of random overlaps (p-value of permutation Z test = 0.19). This result suggests two important aspects of this section: first, the missing enrichment of somatic mutations in GWAS associated regions confirmed the results of Machiela *et al.* (Machiela et al., 2015) and secondly while GWAS are designed to work on common variants, somatic mutations are usually rare. Thus these two types of analysis represent two different layers of hereditability.

Since somatic mutations in cancer are mainly passenger, a simple overlap with a somatic mutation cannot suggest a real involvement in cancer predisposition (Vogelstein et al., 2013). Therefore, we decided to further subset our 2441 SNPs to include only variants on a list of manually curated target genes (see section **4.3.3.3**) with a higher probability of being real drivers. Only 6 variants in 4 genes ended up having all the characteristics included in this analysis. These variants form a list of highly valuable candidates (Figure 22) as theorized by one of the ICOGS flagship paper (Michailidou et al., 2013). In particular, *RAD51B* is a known breast cancer associated gene (Golmard et al., 2013)

TET2 variant discovered in our dataset is only ~80kb away from an ICOGs SNP rs9790517. In addition, *TET2* has already been associated with breast cancer at the RNA level (Yang et al., 2015b) and it is considered a known somatic driver in leukemia and melanoma (Ficz and Gribben, 2014). Another ICOGs variant (rs132390 on *EMID1*) is in a low recombination region along with *NF2* R335C variation. *NF2* has been associated to the hereditary neurofibromatosis syndrome 2 and mutates both at germinal and somatic level in breast cancer (Schroeder et al., 2013). The same ICOGs SNP has been found in LD with *CHEK2,* a known breast cancer associated gene. Although our HapMap data do not support this linkage disequilibrium, we found a variant on this gene (rs201206424) at approximately the same distance as the *NF2* variant described above (~400kb) (Michailidou et al., 2013). This *CHEK2* variant has also been found as somatically mutated in breast cancer. Two different alterations were found on *EP300* in LD with the ICOGs SNP rs6001930. EP300 has a well-established role as a tumor suppressor but it is poorly investigated as a breast cancer predisposing gene (Gayther et al., 2000).

Excluding the aforementioned 6 SNPs, 37 variants are monomorphic in the ExAC database that represents our control (Appendix Table 9). The first positions sorted by MAF ratio are occupied by truncating variants on *PIK3CB, KMTC2* and *NBN*. The first two genes are known somatic drivers, and in particular the second one has also been classified as a tumor suppressor (as the truncating mutation suggests). The same genes will be considered, as a whole, as significant loss-of-function genes in the next section of the results (see section **4.3.4.3**). *NBN* instead, has been already associated with increased risk of breast cancer via the Nijmegen syndrome being part of an important DNA repair pathway (Varon et al., 1998). Nevertheless, the specific frameshift mutation found (I166fs), has never been associated with this syndrome before but has been found somatically mutated in breast cancer using our annotation. The most relevant result of this analysis branch is probably the variant E17K on *AKT1* (rs121434592) that has been

already described in the previous section. This gene is a known somatic driver kinase and this mutation has been found in 46 different samples in the cBioPortal database in many different tumor types, including breast. E17K is also reported by CIViC and DoCM databases list of curated somatic driver mutations (Bleeker et al., 2008). This variant, along with *ATM* R337C (rs138398778) is reported in the list of cancer hotspots curated by Chang et al. (Chang et al., 2016) and they both represents a case of known somatic driver mutation that can be considered a cancer predisposing variant. In addition, we found other germline variants present in more than 2 samples in COSMIC or cBioPortal on the following genes: *HNF1A, FGFR3*, *ASXL1*. Interestingly, all these genes are included in our list of cancer predisposing genes or somatic driver genes and none of them has been connected to breast cancer predisposition before.

4.3.4.3 Analysis of loss-of-function genes

We decided to focus on possible loss-of-function genes involved in predisposition to breast cancer because the large majority of cancer predisposing genes are in fact recessive loss-of-function variants (Rahman, 2014). In particular, we wanted to explore the existing overlap between somatic driver tumor suppressors and loss-of-function predisposition to breast cancer following a somatic driver gene discovery pipeline as discussed in the previous section. It is known that truncating mutations plays a major role in targeting potential tumor suppressors (Vogelstein et al., 2013), so we selected from our dataset only the truncating variants under a softer filter of frequency of 5% in the control population, for a total of 2522 different truncating events on 1931 different genes. On this reduced dataset, we looked for imbalance between control and case frequency in any of the truncation spots with a gene-wise testing procedure (see section **4.3.3.5.2**). After testing and correcting for false discovery rate, we filtered for candidate genes in our 758 target gene list to seek for really potential overlapping driver/predisposing genes and to emulate a candidate gene analysis. Only 94 genes have

at least one truncating variant with a frequency in control cohort below 0.05, of which 41 passed the p-value threshold (Appendix Table 10). As a proof of concept, known breast cancer predisposing genes like *BRCA1*, *BRCA2* and *CHEK2* are in fact selected by our procedure. Other known breast cancer predisposing genes, such as *TP53* or *PALB2*, are instead not found truncated in our dataset because they are too rare for our detection power in a non-familiar selected dataset (Table 6) (Antoniou et al., 2014). Nevertheless, *TP53* has one missense variant included in the list of the 176 overlapping with somatic mutations and this particular variant has never been reported as pathogenic before (rs138729528), being completely novel in our control dataset. Among the 41 significant LOF candidates, *FGFR3, PIK3CB, HNF1A* and *KMTC2* were also highlighted as somatically mutated by the previous analysis but in this case we were able to add a defined loss-of-function role. In addition, another member of the homeobox family (*HNF1B*) has one of the lowest p-values. This gene has been connected to predisposition to ovarian cancer, but no association with breast cancer has been previously described (Shen et al., 2013). With similar characteristics, we found the anaplastic lymphoma kinase (*ALK)*, a known driver gene and predisposing gene in lung cancer and neuroblastoma, with few evidence of association with breast cancer (Siraj et al., 2015).

The majority of the genes in this list harbor 1 to 2 different truncation points. *CRIPAK* appears to be a particular exception with 27 different truncations in various point of the gene body. This abundance of frameshifts and non-sense variants at various points of the protein can be partially explained by the fact that *CRIPAK* is intronless and like other genes with this feature (like *CDR1* or *AD7C-NTP*) it tends to accumulate these variations for evolutionary reasons (Okamura et al., 2006) and is probably a false positive result.

4.3.4.4 Polygenic age-dependent model

In the last section of the results, we moved from a pure case-control study to a more association-like study. In all the previous analysis, we always put a filter on the

frequency, selecting rare (control MAF < 1%) or low frequency variants (control MAF < 5%). In this analysis, exploiting a trait that can be considered complex as age at pathological diagnosis, we used every non-synonymous variant in our dataset (Figure 20). As explained in section **4.3.3.5.3**, we implemented a double step machine learning approach composed by 1) a tree-based classification with variants as subjects (dimensionality reduction step) 2) a penalized linear model regressing age to the genotypes of the cases, so that the variants become now covariates (feature selection step). In the first step, the final goal is to assign a "Pathogenicity Score", or in other words, a probability value that represents how similar to the prototypes in the training set and far from the negative set our variants are. These two sets are represented by known pathogenic variations and a series of variants tested as simple polymorphisms. An interesting side effect of the procedure is also that we could assign a score of importance to the features that are responsible for the classification machinery. We started with a dataset of prototypes (our training set) composed by 38 pathogenic and 706 non-pathogenic variants (see section **4.3.3.5.3**). The overall model on the training set reports a very low out-of-bag error of 3.5% in the classification process with an area under the ROC curve of 0.84 (see Figure 23, Panel B).

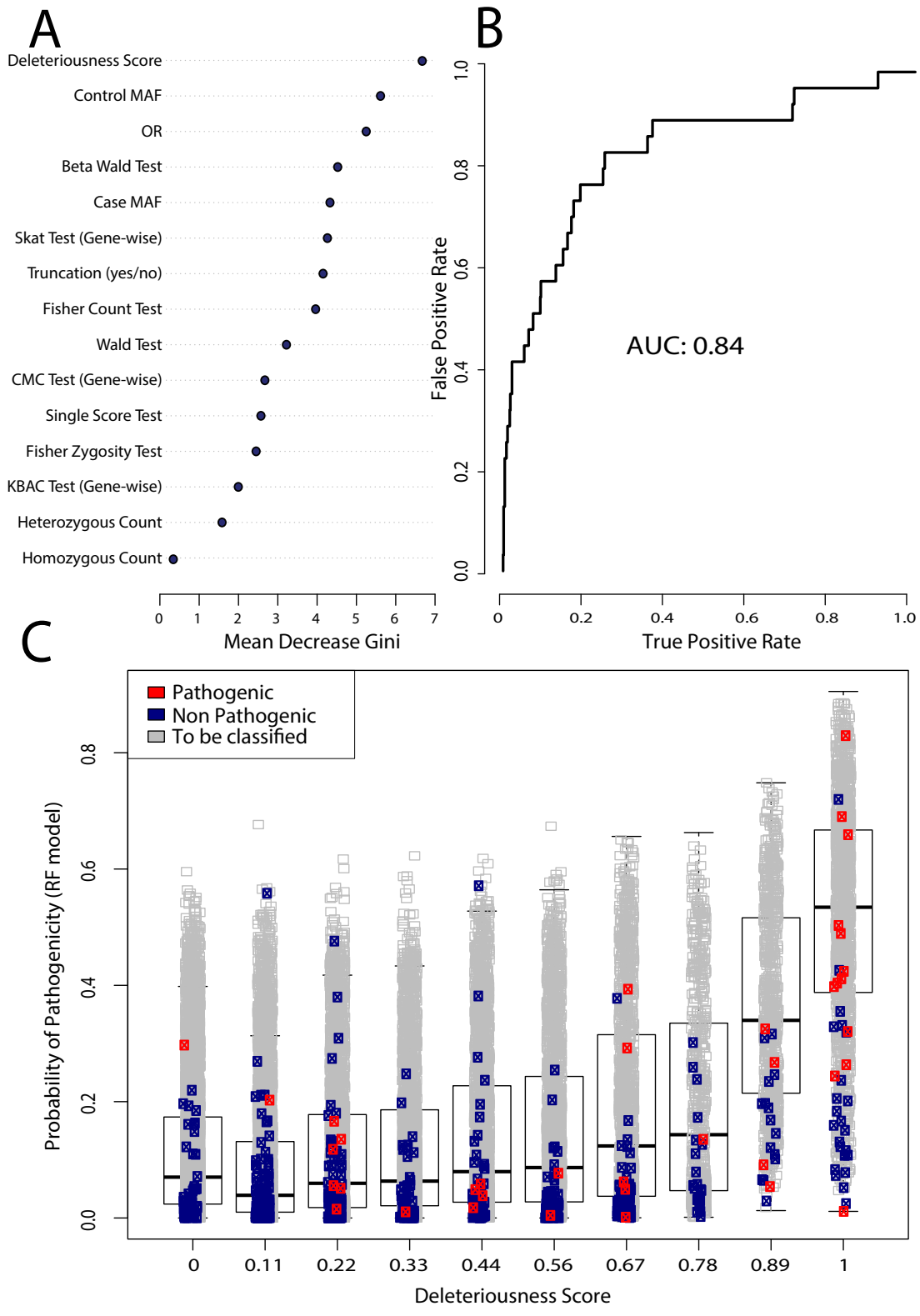

Figure 23 Polygenic age-dependent model breakdown. A) The feature rank of the Random Forest model according to the mean decrease of Gini index is reported. At the top, the most important variable is deleteriousness score (see Panel C). **B)** ROC curve on random forest training model. An AUC of 0.84 is reached under the supervision of the training dataset formed by reported pathogenic and non-pathogenic variants according to a dataset of curated cancer predisposing variants. **C)** The top predictor in our random forest model is reported without the influence of the other variants. Although it cannot represent the real tree scheme of the model, there is a clear positive trend between increased deleteriousness score (X-axis) and the number of trees classifying a variable as pathogenic (Y-axes). Furthermore, the pathogenic score starts increasing after the 0.56 threshold (5 over 9 predictors of phenotypic effect classifying the variant as damaging).

Using this algorithm, we could learn the main characteristics of pathogenic and non-pathogenic variants, as shown in Figure 23, Panel A, and we used these characteristics to classify a test set of unknown variants. The mean decrease in the Gini index represents the ability of each feature to separate the class pathogenic from the non-pathogenic as the amount of homogeneity gained after each node split that contains the feature under exam. Using our random forest model, we then classified the unknown variants in the test set. We tried to represent the behavior of each top feature in the classification problem by comparing the feature value with the corresponding probability of pathogenicity assigned by the random forest model. This probability is calculated as the proportion of trees that classify the variant as pathogenic in a total of 100'001 trees built in the training procedure. For example, the random forest model has the tendency of assigning high probability to more deleterious variants, as a clear linear trend is visible between deleteriousness and RF score assignment (see in Figure 23, Panel A). In fact, the majority of the known pathogenic variants (the red dots) fall into the top two deleteriousness score category compared to the non-pathogenic variations (the blue dots) that appear to fall in every category without a specific pattern. Interestingly, the results of widely used test of associations like SKAT or SingleScore do not seem to provide sufficient adherence to pathogenic variations, according to the Gini Index ranking (Figure 23, Panel A) and justify our use of a more genomic and knowledge-based approach rather than a pure statistical method. However, to develop the random forest model, the prototypes we used for pathogenic and non-pathogenic variants are not exclusively breast-related and have no direct connection to our dataset. Thus, we decided to define the variants as features that could be associated with patient characteristics.

If we used the whole dataset, we would end up with a matrix in form of 673 samples (one for each patient) and more than 70'000 features (one for each variant). This model would

suffer from a heavy curse of dimensionality. In order to reduce the number of features entering in this second clinical model, we set a standard threshold of 0.5 in the pathogenicity probability coming from the random forest that allows a high level of specificity (almost 100%) while still retaining a good sensitivity (~60%). This threshold allows retaining a variant with some evidence of being pathogenic and discarding the majority of non-pathogenic noisy variants. A variant is retained in the clinical model if:

1. The Pathogenicity Score is greater than 0.5 (majority voting in random forest procedure)
2. It is not part of the training set

With this filter we ended up with 4045 variants entering in the second step and therefore reducing the risk of an inflated dimensionality. The model we used for the polygenic analysis is a robust elastic net. More details about the procedure can be found in section **4.3.3.5.3**. We run a regression model of age at initial pathological diagnosis to the genotype of our subjects. The controls are therefore not included in this procedure. While in the first procedure the output was represented by the Pathogenicity Score, in this case we ranked the features (now the variants) as being negatively associated with age. The result of the elastic net model can be influenced in case the number of subjects is lower than the number of features (4045 variants as features and 673 age values as subjects). That is why we used a permutation based multi model that allows a robust ranking. Furthermore, like any other shrinkage method, not all the features are retained, in order to reduce the degrees of freedom of the model. The variants are ordered by the number of time a feature is retained in the model with a negative beta since the higher the number of times, the lower the age when the variant is present. Final results include a list of 19 variants retained in at least 10% of the 100 models, of which 15 with a negative beta in more than 50% of the 100 models (Table 8).

Variant	Approved Name	Control MAF	Case MAF	Protein Change	Mean Beta ElasticNet	Negative Beta Percentage
MRPL24 - 1,156708335,C,T	mitochondrial ribosomal protein L24	0.0000%	0.074%	W54*	-2.78	1.00
CST4 - 20,23667825,-,C	cystatin S	0.0129%	0.300%	V81fs	-5.09	1.00
PARD6A - 16,67696278,C,T	par-6 family cell polarity regulator alpha	0.0018%	0.078%	R256*	-1.86	1.00
TRIOBP - 22,38121788,-,C	TRIO and F-actin binding protein	0.0059%	0.471%	S1075fs	-3.64	1.00
ZNF85 - 19,21132125,C,T	zinc finger protein 85	0.0000%	0.085%	R205*	-4.36	1.00
FOXP4 - 6,41553185,A,G	forkhead box P4	0.0018%	0.091%	K147R	-8.04	1.00
PKHD1 - 6,51890490,A,C	polycystic kidney and hepatic disease 1 (autosomal recessive)	0.0000%	0.075%	M1373R	-5.33	1.00
SURF1 - 9,136218808,A,T	surfeit 1	0.0000%	0.081%	L179Q	-6.49	1.00
HIST2H2AB - 1,149859084,TT…GT,-	histone cluster 2, H2ab	0.0000%	0.074%	T121fs	-3.59	0.97
STIM2 - 4,27004586,G,A	stromal interaction molecule 2	0.0000%	0.081%	V281I	-1.65	0.97
CPA3 - 3,148597632,C,T	carboxypeptidase A3 (mast cell)	0.0000%	0.074%	R178*	-5.47	0.94
TMCO3 - 13,114188422,-,G	transmembrane and coiled-coil domains 3	0.0326%	0.742%	A469fs	-1.93	0.93
SERPINF2 - 17,1649022,CCTG,-	serpin peptidase inhibitor, clade F	0.0000%	0.080%	A62fs	-1.74	0.84
PYGL - 14,51383751,G,A	phosphorylase, glycogen, liver	0.0037%	0.149%	R276C	-0.08	0.71
FNIP2 - 4,159790466,C,A	folliculin interacting protein 2	0.0016%	0.101%	S893*	-0.86	0.58
CPPED1 - 16,12758817,G,A	calcineurin-like phosphoesterase domain containing 1	0.0000%	0.074%	R149*	-0.14	0.44
OR52B4 - 11,4388943,G,A	olfactory receptor, family 52, subfamily B, member 4 (gene/pseudogene)	0.0018%	0.076%	R195*	4.81	0.09
SCN10A - 3,38755496,G,A	sodium channel, voltage gated, type X alpha subunit	0.0037%	0.074%	R1155C	1.62	0.08
ZNF683 - 1,26694960,G,A	zinc finger protein 683	0.0000%	0.089%	R35*	1.18	0.03

Table 8 Results from the polygenic age-dependent model. A double-step machine learning algorithm selects variant based on a series of pathogenic prototypes and then further selects them using a permutation based multi-model regression over age at onset. Variants in this set are negatively associated with age and are divided in three layers: on top, variants negatively associated in at least 80% of the models and with an average beta less than -1.5, in the middle, variants retained in at least 40% of the model with poor average beta, at the bottom, variants found negatively associated only in a few model.

We noticed several desired features of the final set of variants. First, without imposing a filter on the control MAF, we selected for rare variants in the population, so that all our 19 variants have a MAF in the control set way below the 1% threshold and a MAF in the cases above the corresponding one in the controls. Furthermore, more than a half of these variants are completely novel in the ExAC dataset. Second, 13 of the 19 variants are classified as truncation events and all the other 6 missense events have a deleteriousness score higher than 0.8, thus evaluated as almost certainly damaging. Lastly, another confirmation of the importance of evaluating somatic events overlapping with germline

mutations is the fact that we noticed a double enrichment in variants found also as somatic.

Among the initial dataset of 73'354 variants, only the ~13% of them are found as somatic events in COSMIC or cBioPortal. After the random forest procedure, this frequency increases up to ~17% among the 4045 retained variants (p-value of binomial test: 8.78e-11) and after the elastic net selection to the ~26% (although not significant, 5 of the 19 variants are also found as somatic).

None of the genes found using this procedure belong to the list of target genes nor are found within breast cancer associated SNPs low recombination regions and there are very few literature reports of a known involvement in cancer, making their selection a completely novel finding (Table 8). Excluding variants on *TMCO3*, *TRIOBP, PYGL* and *CST4*, all the remaining 15 involved one single sample in our dataset, therefore so rare that any simple statistical approach would probably not detect them. Like briefly mentioned before, this set of variants and genes are mostly not involved in cancer, except for *PKHD1*, a gene involved in polycystic kidney disease and a high risk of renal cancer that has also been mentioned in the first section of the results for another known pathogenic variant (Sharp, 2005). Other genes reported in Table 8 with some evidence of cancer involvement includes *STIM2*, which have been associated to allelic loss in 4p in several tumor types, including breast (Shivapurkar et al., 1999) and *FOXP4*, an important member of the forkhead box transcription factor that are widely known to be involved in tumorigenesis and cell-growth (Myatt and Lam, 2007). Although not directly implicated in tumorigenesis, other genes appears to be promising candidates being part of families involved in cancer, including *SERPINF2*, a member of serpin family that has a clear role in cancer cell survival (Valiente et al., 2014) *PARD6A,* member of the PAR family, involved in cell cycle gatekeeping and interactor of other major cancer pathways like MAPK and PI3K (Marques and Klefström, 2015) and finally *HIST2H2AB*, part of

the cluster 2 of histones whose parallel family in cluster 1 is highly mutated in many cancer types (Lawrence et al., 2014; Timp and Feinberg, 2013)

4.3.5 Discussion

The use of NGS technologies has revolutionized the study of human cancers by allowing the simultaneous identification of multiple somatic mutations but it can also offer the possibility to look for the presence of cancer susceptibility variants and genes. Interestingly, taking advantage of sequenced normal genomes of cancer patients, recent studies have suggested that the susceptibility due to rare variants in sporadic cancers can be much more common that previously anticipated (Schrader KA et al., 2016). However, it remains challenging to determine the pathogenicity and the clinical significance of these germline variants since many of them are rare and not well characterized. Our study represents one of the first attempts to prioritize germline variants that may predispose to breast cancer using sequencing data.

We developed a computational framework based on the characteristics of somatic mutations to identify putative cancer predisposing variants. In particular, we provided an analysis of rare variants and we detected 185 variants that overlap with somatic mutations in cancer. Furthermore, we carried out an analysis of truncating mutations on suspected tumor suppressors, revealing known and new possible loss-of-function candidates. We detected 50 variants associated with 41 possible loss-of-function-genes, including *PIK3CB* and *KMT2C*. Lastly, we built a robust age-dependent polygenic model that involves a mixture of supervised and regression based algorithm to uncover variants at any frequency level. With this model, we identified a set of 19 variants potentially pathogenic and negatively associated with age at onset belonging to genes that have never been associated to breast cancer. Furthermore, we checked if any of the identified candidate variants falls into GWAS known breast cancer susceptibility regions.

In our study we detected several expected variants on known breast cancer predisposing genes like *BRCA1* and *BRCA2,* which are a confirmation of the validity of this study. We also identified 19 variants on genes that are known to be predisposing for other cancer types or cancer syndromes, like *RET* and *AKT1,* that have never been previously associated with breast cancer predisposition.

To our knowledge, there are few examples in the literature attempting an analysis on predisposing genetic makeup in cancer that exploit sequencing data (Kanchi et al., 2014; Lu et al., 2015). While these works design an in depth analysis of known predisposing genes, they lack of a sufficiently extended control dataset, using respectively ~400 normal controls against a dataset of ovarian cancer cases of approximately the same size (Kanchi et al., 2014) and ~1000 samples against ~4000 cases of various cancer types (Lu et al., 2015). The use of the ExAC database, that comprises over 27'000 control samples, allowed a higher resolution that we emphasize at the level of the single variants within a candidate predisposing gene, discerning variants of scarce significance from true candidate pathogenic variations. Furthermore, we introduce more variables in our knowledge-based approach, including also over 20 years of breast cancer GWAS data and patients' characteristics like age of onset. In particular, the latter information is used as a new explanatory variable to further enlarge our set of candidates beyond the limits of already known cancer-related genes and not only as a confirmation of association between early onset and known predisposing genes.

We know that our analysis have several limitations. First, to improve our understanding on the association of rare variants to breast cancer hereditability, we should sequence a larger number of individuals and possibly extend our analysis to other ethnicities. For example, we should use an independent longitudinal cohort to clarify the prevalence of the identified variants or a smaller cohort of suspected familial cases. Secondly, genomic data could be associated to patients' family history, since in the TCGA clinical data this

information is missing. Lastly, we have provided a valuable resource of potential new cancer-related variants that could be characterized from a functional point of view.

In this study we have developed a genomic-driven approach able to prioritize cancer predisposing variants using a case-control genetic scheme. We demonstrate the use of public available sequencing data to better characterize known susceptibility genes and to identify novel cancer predisposing variants. The opportunity to classify individuals according to their risk of developing hereditary-based cancer, will improve clinical management of breast cancer patients in terms of genome-tailored prevention strategies, programs for early diagnosis and possible treatments.

5 Discussion

Although each section is independent from each other, this work has a clear common background and objective. As anticipated in the introduction, one of the most important targets of NGS genomics was to create a sort of catalogue of what is driving cancer in human. The three sections can be therefore summarized in 3 sentences:

1. Distinguish driver genes from passengers and divide them in tumor suppressors and oncogenes (section **4.1**)
2. Focusing on oncogenes, expand the reservoir of cancer genes by finding connections in secondary structure between proteins. Even very low mutation frequency could have a functional meaning by transferring knowledge between proteins (section **4.2**)
3. Apply an approach similar to the first two sections to a dataset of germline variants (section **4.3**)

The common thread in fact, despite the type of data, either somatic or germline, remains the hunt for cancer genes and to distinguish driver mutations or pathogenic variants from passenger mutants. Since the first articles about genome-wide mutation profiles in

cancer (2006-2007), the necessity of a clear picture for each tumor type cancer gene has become a major challenge in cancer genomics, in particular under the grand design of the personalized medicine field. As anticipated in section **2.5**, from the bioinformatics point of view, the real game changer was the possibility to obtain sufficient data to reach statistically meaningful conclusion. Around 2013, when a sufficient amount of the TCGA data was made available, it was soon realized that even hundreds of samples were not sufficient to disentangle the extraordinary heterogeneity in the mutational spectrum of the various cancer types. In particular for highly unstable and fast-mutating tumor types like melanoma or lung cancer, where variability is even more accentuated. In the course of a few years, the subject of distinguish drivers from passengers will probably drying over because we are probably reaching a point of saturation of what can be discovered through this data and techniques. It is true of course that we lack of a sufficiently large sample size to overcome the variability issues but it also true that the first genomic and personalized based clinical trials have shown poor results and this should be the fundamental reason behind this research field (Tourneau et al., 2015). What is probably still missing is the knowledge around other level of the same kind of data. For example, the lessons learnt from the TCGA experience (a project that is now at its conclusion) could be applied to data from metastases, where mutational spectrum is even more elusive than the primary and only small size studies have been published. Another very close field that could benefit from driver and passengers analyses is the clonal evolution field and how mutations evolve and spread in that layer of heterogeneity that is called intra-tumor. The "discovery phase" on primary tumors has probably reached its peak, but the entire experience and tools could be transferred to new areas of interest.

5.1 Driver gene discovery

At the time DOTS-Finder was conceived, there was already an explosion of interest in creating tools for driver discovery. Our vision of mutational process, in fact, is the result of a long evolution of DNA from normal cell up to point of sequencing. The order of the events is unknown and particularly hard to reconstruct, since what we know from exome or genome sequencing is a snapshot of a heterogeneous tumor at a precise moment in time. The need of statistical methods to distinguish noisy passenger mutations from important drivers was therefore a necessity to understand how a tumor evolved, what pathways were mainly involved and ultimately, how to restore the normal phenotype. Many approaches were already present that tackled this problem from various angles (frequency, position of mutations, severity of the mutations, etc.) but an overall view was evident. Driver genes can be elusive if mutated at very low frequency, so that a positional approach that tries to check for mutational pattern, rather than mutational frequency, have proven to be a very effective compared to simple frequency based methods (see section **4.1.2**). In fact, the research of patterns of mutations sounds more "ratiometric", although statistically more disputable, since it relies on a clear side effect (or better the original cause) of oncogene and tumor suppressor behavior. Nevertheless, there is a plethora of other factors to take into account, like length of the gene, position of the gene along the genome, expression levels, replication time. The positional approach, in this sense, is less greedy in terms of required information because it closes the gene in its own environment and do not need the estimation of a global background mutation rate. Expression or replication time data are not necessary and gene length is taken automatically into account. Nevertheless, it is not sufficient to distinguish all oncogenes and tumor suppressor, so that's why we implemented the two-step approach.

In section **4.1.5.11**, we show instead one of the main feature of DOTS-Finder. Compared to our best competitor (MutSigCV), DOTS-Finder is superior in terms of accuracy and

recall when subsets of various sizes are used instead of all the available data (Lawrence et al., 2013). This feature allows calling driver genes even in a situation of scarce evidence of not being passenger. Large sequencing studies like TCGA encompass thousands of patients of the most common tumor types, but rare tumors do not have the same amount of free data, both for a difficulty in finding patients diagnosed with that particular disease but also because of smaller investments compared to major tumor types like breast or lung cancer. Driver discovery tools are not particularly useful in everyday bioinformatics work, because they represent the very final step of an analysis on mutation data that is hardly performed more than once. Furthermore, a tool that needs hundreds of samples to reach a sufficient statistical power becomes useful only for very large and expensive studies. Pilot studies with sample size around 20-30 patients are way more common and so DOTS-Finder can become handy in situations like these.

Another important point that has been emphasized during DOTS-Finder development is both the small amount of data preprocessing in order to run the tool and very little system requirements. DOTS-Finder was originally developed in python but soon we realized that R could overcome certain mathematical passages more easily (in particular related to oncogene and tumor suppressor scores) and python was used mostly as a shell for the whole architecture. MutSigCV, for example, is written in Matlab, which is not freely available. Moreover, tools like MuSiC or OncodriveFM require demanding input file or a certain amount of data preprocessing that we specifically avoid (Dees et al., 2012; Gonzalez-Perez and Lopez-Bigas, 2012). MuSiC, being in fact only a part of a larger toolset, can only work with the original BAM files that are both hard to obtain in many cases and more difficult to manage compared to VCF or MAF. OncodriveFM instead requires calculations of SIFT and Polyphen2 scores and therefore requires a non-standard input format. DOTS-Finder tries to overcome such difficulties requiring a

standard MAF input file (see section **4.1.4.2**) and uses python under virtualenv and R implementation with very few dependencies.

The creation of this tool was therefore powered by i) the need of a more comprehensive approach to discover cancer driver that could be superior to the sum of its parts ii) simplicity, in the sense that the data required to run the tool is minimal iii) to create an instrument that could be really used for a pilot study on a few cases.

5.2 Oncogenes and driver mutations discovery

LowMACA was born under different premises compared to DOTS-Finder. The first attempt at the realization of this tool was having the possibility to characterize families of proteins rather than unique genes. Driver discovery tools are generally aimed at a global analysis on a set of specific samples but they lack the possibility to interrogate a specific set of genes given all available knowledge across tumor types. LowMACA set up uses a list of genes and Pfam IDs as input and mutation data are collected for the requested input only. Compared to DOTS-Finder and other similar tools, this tool is indeed way more usable on a regular basis. In the same way a web resource can be useful to check the mutations of a specific gene, LowMACA was created under a similar concept. What we did was to rely on a web resource (the cBioPortal) that updates constantly its database, so that mutations could be downloaded on the fly to check for potential driver mutations (Gao et al., 2013). What is new in LowMACA is the possibility to aggregate genes under the same Pfam or on other criteria and align them to form a new consensus protein where all the mutations of the original genes are remapped. Again, usability is a crucial point like it is for DOTS-Finder, but switching the input data from mutation to genes makes it even more appealing from the standing point of a usage on the long run. For LowMACA, the whole implementation was packed in an R library in order to maintain compatibility with the database package cgdsr from cBioPortal. Furthermore, an accompanying data package was implemented too, that greatly simplify the task of

searching for families of genes. The entire Pfam along with Uniprot is automatically available to the user and a perfect one-to-one match between gene symbols and canonical proteins was created to uniform sequence databases and dictionaries (Finn et al., 2007; Gray et al., 2014; The UniProt Consortium, 2014). Moreover, a Shiny implementation of the package was created outside of the Bioconductor repository that further simplifies the possibility of an on-the-fly analysis. A completely web-based implementation was in fact discarded as a viable possibility because the calculations performed by R are sometimes too complex and time consuming to fit usual web timing.

LowMACA also introduces a few new ideas into the driver discovery universe. In particular, being an aggregating method, it tries to overcome the limits of rare mutation boundaries. A very simple but effective analysis of what is in fact the required sample size to be confident enough to distinguish any driver from passenger can be found in (Lawrence et al., 2014). The statistical power of MutSig (and by inference of any other driver discovery tool) is inversely proportional to the mutation rate of the specific tumor type, so that a tumor like melanoma, with a very high number of mutation per patient would require thousands of sequenced cases to be completely saturated in term of driver genes discovered. LowMACA, by aggregating mutational profiles, represents a sort of shortcut that has been proposed in various forms in recent years, in particular via pathway or network analysis (Ciriello et al., 2012; Vandin et al., 2012). The advantage of LowMACA over pathway analysis is that the position of the mutation is retained and doesn't lose its meaning. This allowed us to be able not only to collect information over driver genes but also on driver mutations, finding connections between genes that are not clearly visible unless the sequences and mutational profiles are blatantly similar (see the case of *KRAS, NRAS* and *HRAS*). In fact, the special case of the RAS trio served as a proof of concept that a similar mechanism do exists and that mutations are also cancer specific in terms of gene and position (see section **4.2.4.1**) within the same gene (a

particularly interesting case is represented by *EGFR* in lung and brain tumors, as shown in section **4.2.4.4**). Nevertheless, there are certain limitations. LowMACA, for example, uses an amino acidic dictionary based on secondary structure connections. This choice creates a series of inherent biases, that can be summarized as i) the Pfam database, based on predicted similarities given by an HMM model ii) the alignment algorithm that could get stuck in local minima, in particular in the case of hundreds of sequences aligned at the same time iii) the difficulty of judging the alignment goodness of each single base. Under this view, pathway analysis does not suffer from these biases, since it works on a more biological level that is constituted mostly by literature findings and partially by predicted connections. For example, in the case of HotNet2, an efficient network-based driver discovery tools, the pathway architecture is not even superimposed (Leiserson et al., 2015). It is the algorithm itself that creates the network reducing any source of external bias caused by erroneous database entries. These drawbacks are therefore insurmountable in case we want to investigate single mutations using secondary structure similarities and a couple of safety nets have been implemented. First of all, the Valdar score can be fine tuned according to user specification in order to accept a minimum level of similarity for each base aligned (Valdar, 2002). Secondly, the possibility to aggregate mutations using a Gaussian density was borrowed from the oncogene score of DOTS-Finder framework without any imposition on the choice of bandwidth (in DOTS-Finder, the bandwidth was set with Silverman's rule of thumb (Silverman, 1986). Both Valdar score and bandwidth can be easily changed in the Shiny based application following user specifications.

As explained above, LowMACA is dedicated to exploratory analysis rather than deterministic results like normal driver discovery tools, as many parameters can be used at the same time to fine-tune the results. In particular the choice of which are the genes to align is of paramount importance. In the section dedicated to RAS family we show a

way to overcome the limits of Pfam by using literature to separate a large family (around 130 different proteins) into more homogeneous subfamilies, mixing *de facto* secondary structure similarities with defined biological functions. Three possible follow-ups of this work come from this possibility:

1. Deconstruct Pfam families following other criteria or even implement a more stringent HMM to create user-defined families. For example, subdividing large families into more biologically meaningful clusters, like we did for the RAS superfamily.
2. Implement new dictionaries, for example using amino acid motifs. In this case the sequences to align become shorter and as a result, families increase in size.
3. Changing alphabet. Amino acids are convenient to work with because there exists an intrinsic stechiometric similarity that facilitates alignment evaluation. Furthermore, databases such as GenBank or Pfam are a useful precompiled resource. Using directly genomic regions based on nucleobases alphabet, other kind of structures can be analyzed, in particular outside of the coding region, like for example, binding site motifs.

Even though LowMACA was able to show connections between proteins via their mutational patterns, it does not answer to the question of why a gene mutates way more frequently than others in the same family. This is the case of the RAS family too, where *KRAS* is the leading gene and all the others follow. If the structure and the function are the same or at least very similar, why *KRAS* mutates and *RRAS2* does not? We think there are at least three possible scenarios to disentangle this question that involve the probability of mutation and the function of the proteins:

1. If the probability of mutation is constant in the family and the function of all the proteins is truly interchangeable, it could be a matter of expression. A mutation on *KRAS* or *RRAS2* could be seen with same probability but the selective growth

advantage is much stronger in the first case because *KRAS* is more expressed and its oncogenic potential is greater. This explains the fact that *KRAS* (ubiquitously expressed) is mutated in many cancer types, including pancreas, lung and colon, while *NRAS* is typically found in melanoma and leukemia (Downward, 2003).

2. The probability of mutation is constant but the function is slightly different. Although the RAS trio shares ~85% of their protein sequence, if the function is different so it is their oncogenic potential. This could be true both at wild type level and in the mutated form. It has demonstrated, for example, that *NRAS* and *HRAS* are not essential for mouse embryonic development while *KRAS* knock-down showed embryonic lethality (Johnson et al., 1997). Furthermore, drugs developed on *HRAS* models showed no effect on *KRAS* mutations, confirming different behaviors also in the mutated form (Baines et al., 2011).
3. The probability of mutation is different from gene to gene. While this is certainly true, because mutation rate depends on replication time, expression and upstream epigenetic factors (Lawrence et al., 2013), these differences were seen on large portion of the genome and it is probably hard to demonstrate for genes within the same family.

5.3 Bridging the gap between genetics and genomics

As briefly mentioned in the introduction of this final discussion, the scope of section **4.3** was an attempt at bridging the gap between genetic and genomic analysis. Broadly speaking, the two main categories of genetic studies can be summarized as case-control (our case) and trait association studies. Case-control genetic studies follow a scheme starting from statistical evaluation of variants enriched in cases compared to controls and subsequently fine-maps variants with a biological meaning. This is certainly a rigorous approach from the statistical point of view but it has its drawbacks. Sample size, for example, is a crucial point because the more positions are tested the higher is the

probability to find false positive results. In the same way, rare pathogenic variants are more prone to be called as false negative because of their scarce odds to find them. Rare variants play an important role in disease predisposition because they are generally associated with penetrant phenotypes. Genomic studies, on the other hand, do not always follow such scheme. Often, a knowledge-based approach that exploits ontologies, pathways or simply literature enters in the decision process of highlighting relevant results from the very beginning of the study. These information sources could be used as simple filters to retain the best candidates for the analysis or in other cases through what we can call an "approach by prototypes". This methodology is based on previous knowledge on how a predisposing gene or variant should present (the prototypes) and searches for all the variations that are as closed as possible to known pathogenic events. At the same time, variants that are proved not to be pathogenic can be used as a negative control. This is not any different from any machine learning approach with a training and test set.

As mentioned in section **4.3.2**, our approach wants to use a typical genomic (knowledge-based) approach applied to genetic data. Using regular genetic techniques is not particularly indicated with exome sequencing data and up to the date, very few examples have proven to work on cancer data (see section **4.3.5**). Both the examples reported (Kanchi et al., 2014; Lu et al., 2015) used a sort of pre filtering technique, in particular to show the ability to highlight genes like *BRCA1* or *BRCA2* that are known CPGs and are usually tested via small target sequencing panels. Most of their results, however, are aimed at pinpointing genes rather than variants. On the contrary, we wished to maintain a variant-wise approach (that is way more informative as it highlights potential pathogenic variations), but that comes at the cost, again, of losing statistical power. We therefore tried to move a step forward and add new layers of complexity to the problem.

Under our framework, only coding non-synonymous variants are retained and various genomic-style methodologies are used:

1. Use variants that overlap with somatic mutations, in known driver genes, with a high deleteriousness score (see section **4.3.4.2** and **4.3.3.3**). This approach is based on the assumption that what is seen in cancer as somatic could probably be harmful in the germline too. Passenger mutations are possibly filter out using the knowledge from what are the driver gene candidates (Rahman, 2014) and a prediction of damage based on the estimation of phenotypic effect (Liu et al., 2013)
2. Select truncating variants on tumor suppressor candidates, which is a technique borrowed from DOTS-Finder, adapted to take into account a case-control scheme that is not present when dealing with somatic mutations only
3. Develop a hybrid approach based first on selecting those variants that have characteristics similar to known pathogenic variants (prototype approach). Use the genotype of the selected variants like an association trait study, by running a regression over age at disease onset (GWAS-like approach)

In the same way as LowMACA represented a shortcut for a lack of sufficient sample size when dealing with rare somatic mutations, this approach can be seen as a shortcut to overcome the same lack on different data. The 673 cases we used in this study are orders of magnitude smaller than GWAS studies on cancer like the COGs consortium that comprises over 200'000 genotyped samples (Sakoda et al., 2013). In fact, an approach based on alignment that emphasizes oncogenes (gain-of-function) rather than tumor suppressors (loss-of-function) was also hypothesized in the early phase of this analysis. Unfortunately, there are potential pitfalls in following this approach:

- While oncogenes are common in cancer, they are very rare in the germline. A part from a few cases (like for example *RET*), at least 90% of known cancer

predisposing genes are in fact loss-of-function. This is not surprising since oncogenes are dominant and with a strong neoplastic transformation potential. This is in most cases, deleterious for embryonal development.

- LowMACA doesn't take into account the minor allele frequency of the variant in both cases and controls because the individual itself represents the control in cancer and somatic mutations are generally rare by definition. An extra effort should be considered to implement an approach that take into account cases and controls that is a potentially fruitful follow-up of this work.

5.4 Conclusion

In this work, by the results of three distinct studies, we built a comprehensive computational framework to study cancer mutations. Cancer is indeed a unique disease where the edge between predisposition and disease development factors does not create a defined distinction that can be exploited as a source of mutual biological information. In the NGS era, we envisage a unified approach that clearly defines cancer etiology from DNA somatic mutations based on a deep understanding of the genetic risk components.

6 References

Adzhubei, I. a, Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.

Agrelo, R., Cheng, W.-H., Setien, F., Ropero, S., Espada, J., Fraga, M.F., Herranz, M., Paz, M.F., Sanchez-Cespedes, M., Artiga, M.J., et al. (2006). Epigenetic inactivation of the premature aging Werner syndrome gene in human cancer. Proc. Natl. Acad. Sci. U. S. A. *103*, 8822–8827.

Alan, J.K., and Lundquist, E. a (2013). Mutationally activated Rho GTPases in cancer. Small GTPases *4*, 159–163.

Alentorn, A., Sanson, M., and Idbaih, A. (2012). Oligodendrogliomas: new insights from the genetics and perspectives. Curr. Opin. Oncol. *24*, 687–693.

Alioto, T.S., Buchhalter, I., Derdak, S., Hutter, B., Eldridge, M.D., Hovig, E., Heisler, L.E., Beck, T.A., Simpson, J.T., Tonon, L., et al. (2015). A comprehensive assessment

of somatic mutation detection in cancer using whole-genome sequencing. Nat. Commun. *6*.

Altucci, L., Leibowitz, M.D., Ogilvie, K.M., de Lera, A.R., and Gronemeyer, H. (2007). RAR and RXR modulation in cancer and metabolic disease. Nat. Rev. Drug Discov. *6*, 793–810.

Antoniou, A.C., Casadei, S., Heikkinen, T., Barrowdale, D., Pylkäs, K., Roberts, J., Lee, A., Subramanian, D., De Leeneer, K., Fostira, F., et al. (2014). Breast-Cancer Risk in Families with Mutations in PALB2. N. Engl. J. Med. *371*, 497–506.

Armitage, P., and Doll, R. (1954). The Age Distribution of Cancer and a Multi-stage Theory of Carcinogenesis. Br. J. Cancer *8*, 1–12.

Baglietto, L., Lindor, N.M., Dowty, J.G., White, D.M., Wagner, A., Gomez Garcia, E.B., Vriends, A.H.J.T., Cartwright, N.R., Barnetson, R.A., Farrington, S.M., et al. (2010). Risks of Lynch Syndrome Cancers for MSH6 Mutation Carriers. JNCI J. Natl. Cancer Inst. *102*, 193–201.

Baines, A.T., Xu, D., and Der, C.J. (2011). Inhibition of Ras for cancer treatment: the search continues. Future Med. Chem. *3*, 1787–1808.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A.F., and Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics *16*, 412–424.

Bardella, C., Pollard, P.J., and Tomlinson, I. (2011). SDH mutations in cancer. Biochim. Biophys. Acta BBA - Bioenerg. *1807*, 1432–1443.

Bashashati, A., Haffari, G., Ding, J., Ha, G., Lui, K., Rosner, J., Huntsman, D.G., Caldas, C., Aparicio, S.A., and Shah, S.P. (2012). DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. Genome Biol. *13*, R124.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Ser. B Methodol. *57*, 289–300.

Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Sayers, E.W. (2013). GenBank. Nucleic Acids Res. *41*, D36–D42.

Berger, M.F., Lawrence, M.S., Demichelis, F., Drier, Y., Cibulskis, K., Sivachenko, A.Y., Sboner, A., Esgueva, R., Pflueger, D., Sougnez, C., et al. (2011). The genomic complexity of primary human prostate cancer. Nature *470*, 214–220.

Bleeker, F.E., Felicioni, L., Buttitta, F., Lamba, S., Cardone, L., Rodolfo, M., Scarpa, A., Leenstra, S., Frattini, M., Barbareschi, M., et al. (2008). AKT1E17K in human solid tumours. Oncogene *27*, 5648–5650.

Bodini, M., Ronchini, C., Giacò, L., Russo, A., Melloni, G.E.M., Luzi, L., Sardella, D., Volorio, S., Hasan, S.K., Ottone, T., et al. (2014). The hidden genomic landscape of acute myeloid leukemia: subclonal structure revealed by undetected mutations. Blood blood-2014-05-576157.

Boveri, T. (2008). Concerning the Origin of Malignant Tumours by Theodor Boveri. Translated and annotated by Henry Harris. J. Cell Sci. *121*, 1–84.

Breiman, L. (2001). Random forests. Mach. Learn. *45*, 5–32.

Bulavin, D.V., Demidov, O.N., Saito, S. 'ichi, Kauraniemi, P., Phillips, C., Amundson, S.A., Ambrosino, C., Sauter, G., Nebreda, A.R., Anderson, C.W., et al. (2002). Amplification of PPM1D in human tumors abrogates p53 tumor-suppressor activity. Nat. Genet. *31*, 210–215.

Campeau, P.M., Foulkes, W.D., and Tischkowitz, M.D. (2008). Hereditary breast cancer: new genetic developments, new therapeutic avenues. Hum. Genet. *124*, 31–42.

Catucci, I., Verderio, P., Pizzamiglio, S., Manoukian, S., Peissel, B., Zaffaroni, D., Roversi, G., Ripamonti, C.B., Pasini, B., Barile, M., et al. (2011). The CASP8 rs3834129 polymorphism and breast cancer risk in BRCA1 mutation carriers. Breast Cancer Res. Treat. *125*, 855–860.

Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E., et al. (2012). The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. Cancer Discov. *2*, 401–404.

Chang, M.T., Asthana, S., Gao, S.P., Lee, B.H., Chapman, J.S., Kandoth, C., Gao, J., Socci, N.D., Solit, D.B., Olshen, A.B., et al. (2016). Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. Nat. Biotechnol. *34*, 155–163.

Chapman, M.A., Lawrence, M.S., Keats, J.J., Cibulskis, K., Sougnez, C., Schinzel, A.C., Harview, C.L., Brunet, J.-P., Ahmann, G.J., Adli, M., et al. (2011). Initial genome sequencing and analysis of multiple myeloma. Nature *471*, 467–472.

Check Hayden, E. (2013). Huge cancer study uncovers 74 genetic risk factors. Nature.

Chen, S., and Parmigiani, G. (2007). Meta-analysis of BRCA1 and BRCA2 penetrance. J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol. *25*, 1329–1333.

Chen, C., Bhalala, H.V., Vessella, R.L., and Dong, J.-T. (2003). KLF5 is frequently deleted and down-regulated but rarely mutated in prostate cancer. The Prostate *55*, 81–88.

Chen, C., Benjamin, M.S., Sun, X., Otto, K.B., Guo, P., Dong, X.-Y., Bao, Y., Zhou, Z., Cheng, X., Simons, J.W., et al. (2006). KLF5 promotes cell proliferation and tumorigenesis through gene regulation and the TSU-Pr1 human bladder cancer cell line. Int. J. Cancer *118*, 1346–1355.

Cheng, W.-C., Chung, I.-F., Chen, C.-Y., Sun, H.-J., Fen, J.-J., Tang, W.-C., Chang, T.-Y., Wong, T.-T., and Wang, H.-W. (2014). DriverDB: an exome sequencing database for cancer driver gene identification. Nucleic Acids Res. *42*, D1048–D1054.

Chia, W.J., and Tang, B.L. (2009). Emerging roles for Rab family GTPases in human cancer. Biochim. Biophys. Acta - Rev. Cancer *1795*, 110–116.

Chun, S., and Fay, J.C. (2009). Identification of deleterious mutations within three human genomes. Genome Res. *19*, 1553–1561.

Ciriello, G., Cerami, E., Sander, C., and Schultz, N. (2012). Mutual exclusivity analysis identifies oncogenic network modules. Genome Res. *22*, 398–406.

Collins, D.W., and Jukes, T.H. (1994). Rates of Transition and Transversion in Coding Sequences since the Human-Rodent Divergence. Genomics *20*, 386–396.

Consortium, T.U. (2013). Update on activities at the Universal Protein Resource (UniProt) in 2013. Nucleic Acids Res. *41*, D43–D47.

D'Andrea, A.D. (2010). Susceptibility Pathways in Fanconi's Anemia and Breast Cancer. N. Engl. J. Med. *362*, 1909–1919.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158.

Davoli, T., Xu, A.W., Mengwasser, K.E., Sack, L.M., Yoon, J.C., Park, P.J., and Elledge, S.J. (2013). Cumulative Haploinsufficiency and Triplosensitivity Drive Aneuploidy Patterns and Shape the Cancer Genome. Cell *155*, 948–962.

De Keersmaecker, K., Atak, Z.K., Li, N., Vicente, C., Patchett, S., Girardi, T., Gianfelici, V., Geerdens, E., Clappier, E., Porcu, M., et al. (2013). Exome sequencing identifies mutation in CNOT3 and ribosomal genes RPL5 and RPL10 in T-cell acute lymphoblastic leukemia. Nat. Genet. *45*, 186–190.

Dees, N.D., Zhang, Q., Kandoth, C., Wendl, M.C., Schierding, W., Koboldt, D.C., Mooney, T.B., Callaway, M.B., Dooling, D., Mardis, E.R., et al. (2012). MuSiC: Identifying mutational significance in cancer genomes. Genome Res. *22*, 1589–1598.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498.

Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K., and Liu, X. (2015). Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum. Mol. Genet. *24*, 2125–2137.

Downward, J. (2003). Targeting RAS signalling pathways in cancer therapy. Nat. Rev. Cancer *3*, 11–22.

Dreijerink, K.M.A., Goudet, P., Burgess, J.R., and Valk, G.D. (2014). Breast-Cancer Predisposition in Multiple Endocrine Neoplasia Type 1. N. Engl. J. Med. *371*, 583–584.

Dudgeon, C., Shreeram, S., Tanoue, K., Mazur, S.J., Sayadi, A., Robinson, R.C., Appella, E., and Bulavin, D.V. (2013). Genetic variants and mutations of PPM1D control the response to DNA damage. Cell Cycle *12*, 2656–2664.

Dulbecco, R. (1986). A turning point in cancer research: sequencing the human genome. Science *231*, 1055–1056.

Eng, C. (1999). RET Proto-Oncogene in the Development of Human Cancer. J. Clin. Oncol. *17*, 380–380.

Esseltine, J.L., Willard, M.D., Wulur, I.H., Lajiness, M.E., Barber, T.D., and Ferguson, S.S.G. (2013). Somatic mutations in GRM1 in cancer alter metabotropic glutamate receptor 1 intracellular localization and signaling. Mol. Pharmacol. *83*, 770–780.

Fachal, L., and Dunning, A.M. (2015). From candidate gene studies to GWAS and post-GWAS analyses in breast cancer. Curr. Opin. Genet. Dev. *30*, 32–41.

Fasano, O., Aldrich, T., Tamanoi, F., Taparowsky, E., Furth, M., and Wigler, M. (1984). Analysis of the transforming potential of the human H-ras gene by random mutagenesis. Proc. Natl. Acad. Sci. U. S. A. *81*, 4008–4012.

Ficz, G., and Gribben, J.G. (2014). Loss of 5-hydroxymethylcytosine in cancer: Cause or consequence? Genomics *104*, 352–357.

Finn, R.D., Tate, J., Mistry, J., Coggill, P.C., Sammut, S.J., Hotz, H.-R., Ceric, G., Forslund, K., Eddy, S.R., Sonnhammer, E.L.L., et al. (2007). The Pfam protein families database. Nucleic Acids Res. *36*, D281–D288.

Forbes, S.A., Bhamra, G., Bamford, S., Dawson, E., Kok, C., Clements, J., Menzies, A., Teague, J.W., Futreal, P.A., and Stratton, M.R. (2008). The Catalogue of Somatic Mutations in Cancer (COSMIC). Curr. Protoc. Hum. Genet. Editor. Board Jonathan Haines Al *CHAPTER*, Unit-10.11.

Forbes, S.A., Bindal, N., Bamford, S., Cole, C., Kok, C.Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A., et al. (2011). COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res. *39*, D945–D950.

Francis, J.M., Kiezun, A., Ramos, A.H., Serra, S., Pedamallu, C.S., Qian, Z.R., Banck, M.S., Kanwar, R., Kulkarni, A.A., Karpathakis, A., et al. (2013). Somatic mutation of CDKN1B in small intestine neuroendocrine tumors. Nat. Genet. *45*, 1483–1486.

Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M., et al. (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851–861.

Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M.R. (2004). A CENSUS OF HUMAN CANCER GENES. Nat. Rev. Cancer *4*, 177–183.

Gao, J., Aksoy, B.A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S.O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., et al. (2013). Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. Sci. Signal. *6*, pl1–pl1.

Gayther, S.A., Batley, S.J., Linger, L., Bannister, A., Thorpe, K., Chin, S.-F., Daigo, Y., Russell, P., Wilson, A., Sowter, H.M., et al. (2000). Mutations truncating the EP300 acetylase in human cancers. Nat. Genet. *24*, 300–303.

Gillespie, J.H. (2000). Genetic Drift in an Infinite Population: The Pseudohitchhiking Model. Genetics *155*, 909–919.

Goldman, N., and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. *11*, 725–736.

Golmard, L., Caux-Moncoutier, V., Davy, G., Al Ageeli, E., Poirot, B., Tirapo, C., Michaux, D., Barbaroux, C., d'Enghien, C.D., Nicolas, A., et al. (2013). Germline mutation in the RAD51B gene confers predisposition to breast cancer. BMC Cancer *13*, 484.

Gonzalez-Perez, A., and Lopez-Bigas, N. (2012). Functional impact bias reveals cancer drivers. Nucleic Acids Res. *40*, e169.

Goujon, M., McWilliam, H., Li, W., Valentin, F., Squizzato, S., Paern, J., and Lopez, R. (2010). A new bioinformatics analysis tools framework at EMBL-EBI. Nucleic Acids Res. *38*, W695–9.

Gouw, L.G., Reading, N.S., Jenson, S.D., Lim, M.S., and Elenitoba-Johnson, K.S.J. (2005). Expression of the Rho-family GTPase gene RHOF in lymphocyte subsets and malignant lymphomas. Br. J. Haematol. *129*, 531–3.

Govindan, R., Ding, L., Griffith, M., Subramanian, J., Dees, N.D., Kanchi, K.L., Maher, C.A., Fulton, R., Fulton, L., Wallis, J., et al. (2012). Genomic Landscape of Non-Small Cell Lung Cancer in Smokers and Never-Smokers. Cell *150*, 1121–1134.

Grabiner, B.C., Nardi, V., Birsoy, K., Possemato, R., Shen, K., Sinha, S., Jordan, A., Beck, A.H., and Sabatini, D.M. (2014). A diverse array of cancer-associated MTOR mutations are hyperactivating and can predict rapamycin sensitivity. Cancer Discov. *4*, 554–563.

Gray, K. a, Yates, B., Seal, R.L., Wright, M.W., and Bruford, E. a (2014). Genenames.org: the HGNC resources in 2015. Nucleic Acids Res. *1*, 1–7.

Gray, K.A., Daugherty, L.C., Gordon, S.M., Seal, R.L., Wright, M.W., and Bruford, E.A. (2013). Genenames.org: the HGNC resources in 2013. Nucleic Acids Res. *41*, D545–D552.

Greenman, C. (2006). Statistical Analysis of Pathogenicity of Somatic Mutations in Cancer. Genetics *173*, 2187–2198.

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G.L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C., et al. (2007). Patterns of somatic mutation in human cancer genomes. Nature *446*, 153–158.

Griffith, D.M., Veech, J.A., and Marsh, C.J. (2016). **cooccur** : Probabilistic Species Co-Occurrence Analysis in *R*. J. Stat. Softw. *69*.

Grisendi, S., Mecucci, C., Falini, B., and Pandolfi, P.P. (2006). Nucleophosmin and cancer. Nat. Rev. Cancer *6*, 493–505.

Grossmann, V., Tiacci, E., Holmes, A.B., Kohlmann, A., Martelli, M.P., Kern, W., Spanhol-Rosseto, A., Klein, H.-U., Dugas, M., Schindela, S., et al. (2011). Whole-exome sequencing identifies somatic mutations of BCOR in acute myeloid leukemia with normal karyotype. Blood *118*, 6153–6163.

Guo, G., Sun, X., Chen, C., Wu, S., Huang, P., Li, Z., Dean, M., Huang, Y., Jia, W., Zhou, Q., et al. (2013). Whole-genome and whole-exome sequencing of bladder cancer identifies frequent alterations in genes involved in sister chromatid cohesion and segregation. Nat. Genet. *45*, 1459–1463.

Gutierrez-Erlandsson, S., Herrero-Vidal, P., Fernandez-Alfara, M., Hernandez-Garcia, S., Gonzalo-Flores, S., Mudarra-Rubio, A., Fresno, M., and Cubelos, B. (2013). R-RAS2 overexpression in tumors of the human central nervous system. Mol. Cancer *12*, 127.

Haber, D.A., and Settleman, J. (2007). Cancer: Drivers and passengers. Nature *446*, 145–146.

Hall, A. (1998). Rho GTPases and the Actin Cytoskeleton. Science *279*, 509–514.

Hall, J.M., Lee, M.K., Newman, B., Morrow, J.E., Anderson, L.A., Huey, B., and King, M.C. (1990). Linkage of early-onset familial breast cancer to chromosome 17q21. Science *250*, 1684–1689.

Hart, S.M., and Foroni, L. (2002). Core binding factor genes and human leukemia. Haematologica *87*, 1307–1323.

Heravi-Moussavi, A., Anglesio, M.S., Cheng, S.-W.G., Senz, J., Yang, W., Prentice, L., Fejes, A.P., Chow, C., Tone, A., Kalloger, S.E., et al. (2012). Recurrent Somatic DICER1 Mutations in Nonepithelial Ovarian Cancers. N. Engl. J. Med. *366*, 234–242.

Hobert, J.A., and Eng, C. (2009). PTEN hamartoma tumor syndrome: An overview. Genet. Med. *11*, 687–694.

Hodis, E., Watson, I.R., Kryukov, G.V., Arold, S.T., Imielinski, M., Theurillat, J.P., Nickerson, E., Auclair, D., Li, L., Place, C., et al. (2012). A landscape of driver mutations in melanoma. Cell *150*, 251–263.

Hussin, J., Sinnett, D., Casals, F., Idaghdour, Y., Bruat, V., Saillour, V., Healy, J., Grenier, J.-C., de Malliard, T., Busche, S., et al. (2013). Rare allelic forms of PRDM9 associated with childhood leukemogenesis. Genome Res. *23*, 419–430.

Hwang, S., Kim, E., Lee, I., and Marcotte, E.M. (2015). Systematic comparison of variant calling pipelines using gold standard personal exome variants. Sci. Rep. *5*, 17875.

Janakiraman, M., Vakiani, E., Zeng, Z., Pratilas, C.A., Taylor, B.S., Chitale, D., Halilovic, E., Wilson, M., Huberman, K., Ricarte Filho, J.C., et al. (2010). Genomic and biological characterization of exon 4 KRAS mutations in human cancer. Cancer Res. *70*, 5901–5911.

Job, B., Bernheim, A., Beau-Faller, M., Camilleri-Broët, S., Girard, P., Hofman, P., Mazières, J., Toujani, S., Lacroix, L., Laffaire, J., et al. (2010). Genomic Aberrations in Lung Adenocarcinoma in Never Smokers. PLoS ONE *5*.

Johnson, L., Greenbaum, D., Cichowski, K., Mercer, K., Murphy, E., Schmitt, E., Bronson, R.T., Umanoff, H., Edelmann, W., Kucherlapati, R., et al. (1997). K-ras is an essential gene in the mouse with partial functional overlap with N-ras. Genes Dev. *11*, 2468–2481.

Jones, S., Hruban, R.H., Kamiyama, M., Borges, M., Zhang, X., Parsons, D.W., Lin, J.C.-H., Palmisano, E., Brune, K., Jaffee, E.M., et al. (2009). Exomic Sequencing Identifies PALB2 as a Pancreatic Cancer Susceptibility Gene. Science *324*, 217–217.

Kanchi, K.L., Johnson, K.J., Lu, C., McLellan, M.D., Leiserson, M.D.M., Wendl, M.C., Zhang, Q., Koboldt, D.C., Xie, M., Kandoth, C., et al. (2014). Integrated analysis of germline and somatic variants in ovarian cancer. Nat. Commun. *5*.

Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A., et al. (2013). Mutational landscape and significance across 12 major cancer types. Nature *502*, 333–339.

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res. *42*, D199-205.

Kiezun, A., Garimella, K., Do, R., Stitziel, N.O., Neale, B.M., McLaren, P.J., Gupta, N., Sklar, P., Sullivan, P.F., Moran, J.L., et al. (2012). Exome sequencing and the genetic basis of complex traits. Nat. Genet. *44*, 623–630.

Kim, S.J., Zhao, H., Hardikar, S., Singh, A.K., Goodell, M.A., and Chen, T. (2013). A DNMT3A mutation common in AML exhibits dominant-negative effects in murine ES cells. Blood *122*, 4086–4089.

Kimura, E.T., Nikiforova, M.N., Zhu, Z., Knauf, J.A., Nikiforov, Y.E., and Fagin, J.A. (2003). High Prevalence of BRAF Mutations in Thyroid Cancer. Cancer Res. *63*, 1454–1457.

Klampfl, T., Gisslinger, H., Harutyunyan, A.S., Nivarthi, H., Rumi, E., Milosevic, J.D., Them, N.C.C., Berg, T., Gisslinger, B., Pietra, D., et al. (2013). Somatic Mutations of Calreticulin in Myeloproliferative Neoplasms. N. Engl. J. Med. *369*, 2379–2390.

Klauke, K., Radulović, V., Broekhuis, M., Weersing, E., Zwart, E., Olthof, S., Ritsema, M., Bruggeman, S., Wu, X., Helin, K., et al. (2013). Polycomb Cbx family members mediate the balance between haematopoietic stem cell self-renewal and differentiation. Nat. Cell Biol. *15*, 353–362.

Knudson, A.G. (1971). Mutation and cancer: statistical study of retinoblastoma. Proc. Natl. Acad. Sci. U. S. A. *68*, 820–823.

Koboldt, D.C., Fulton, R.S., McLellan, M.D., Schmidt, H., Kalicki-Veizer, J., McMichael, J.F., Fulton, L.L., Dooling, D.J., Ding, L., Mardis, E.R., et al. (2012). Comprehensive molecular portraits of human breast tumours. Nature *490*, 61–70.

Kok, K., Geering, B., and Vanhaesebroeck, B. (2009). Regulation of phosphoinositide 3-kinase expression in health and disease. Trends Biochem. Sci. *34*, 115–127.

Kundu, M., and Liu, P.P. (2001). Function of the inv(16) fusion gene CBFB-MYH11. Curr. Opin. Hematol. *8*, 201–205.

Kundu, M., Chen, A., Anderson, S., Kirby, M., Xu, L., Castilla, L.H., Bodine, D., and Liu, P.P. (2002). Role of Cbfb in hematopoiesis and perturbations resulting from expression of the leukemogenic fusion gene Cbfb-MYH11. Blood *100*, 2449–2456.

Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M., and Maglott, D.R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. *42*, D980–D985.

Lange, S.S., Takata, K., and Wood, R.D. (2011). DNA polymerases and cancer. Nat. Rev. Cancer *11*, 96–110.

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.

Lawrence, M.S., Stojanov, P., Mermel, C.H., Robinson, J.T., Garraway, L.A., Golub, T.R., Meyerson, M., Gabriel, S.B., Lander, E.S., and Getz, G. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. Nature *505*, 495–501.

Lee, R.S., Stewart, C., Carter, S.L., Ambrogio, L., Cibulskis, K., Sougnez, C., Lawrence, M.S., Auclair, D., Mora, J., Golub, T.R., et al. (2012). A remarkably simple genome underlies highly malignant pediatric rhabdoid cancers. J. Clin. Invest. *122*, 2983–2988.

Leiserson, M.D.M., Blokh, D., Sharan, R., and Raphael, B.J. (2013). Simultaneous Identification of Multiple Driver Pathways in Cancer. PLOS Comput Biol *9*, e1003054.

Leiserson, M.D.M., Vandin, F., Wu, H.-T., Dobson, J.R., Eldridge, J.V., Thomas, J.L., Papoutsaki, A., Kim, Y., Niu, B., McLellan, M., et al. (2015). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nat. Genet. *47*, 106–114.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

Ley, T.J., Mardis, E.R., Ding, L., Fulton, B., McLellan, M.D., Chen, K., Dooling, D., Dunford-Shore, B.H., McGrath, S., Hickenbotham, M., et al. (2008). DNA sequencing of a cytogenetically normal acute myeloid leukemia genome. Nature *456*, 66–72.

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., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Li, Y.H., Werner, H., and Püschel, A.W. (2008). Rheb and mTOR regulate neuronal polarity through Rap1B. J. Biol. Chem. *283*, 33784–33792.

Liu, X., Jian, X., and Boerwinkle, E. (2013). dbNSFP v2.0: A database of human non-synonymous SNVs and their functional predictions and annotations. Hum. Mutat. *34*, 2393–2402.

Lohr, J.G., Stojanov, P., Lawrence, M.S., Auclair, D., Chapuy, B., Sougnez, C., Cruz-Gordillo, P., Knoechel, B., Asmann, Y.W., Slager, S.L., et al. (2012). Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. Proc. Natl. Acad. Sci. *109*, 3879–3884.

Lu, C., Xie, M., Wendl, M.C., Wang, J., McLellan, M.D., Leiserson, M.D.M., Huang, K., Wyczalkowski, M.A., Jayasinghe, R., Banerjee, T., et al. (2015). Patterns and functional implications of rare germline variants across 12 cancer types. Nat. Commun. *6*, 10086.

Machiela, M.J., Ho, B.M., Fisher, V.A., Hua, X., and Chanock, S.J. (2015). Limited evidence that cancer susceptibility regions are preferential targets for somatic mutation. Genome Biol. *16*, 193.

Maglott, D., Ostell, J., Pruitt, K.D., and Tatusova, T. (2005). Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res. *33*, D54–8.

Magnani, L., and Cabot, R.A. (2009). Manipulation of SMARCA2 and SMARCA4 transcript levels in porcine embryos differentially alters development and expression of SMARCA1, SOX2, NANOG, and EIF1. Reproduction *137*, 23–33.

Mahmoud, N.N., Boolbol, S.K., Bilinski, R.T., Martucci, C., Chadburn, A., and Bertagnolli, M.M. (1997). Apc gene mutation is associated with a dominant-negative effect upon intestinal cell migration. Cancer Res. *57*, 5045–5050.

Mariano, A.R., Colombo, E., Luzi, L., Martinelli, P., Volorio, S., Bernard, L., Meani, N., Bergomas, R., Alcalay, M., and Pelicci, P.G. (2006). Cytoplasmic localization of NPM in myeloid leukemias is dictated by gain-of-function mutations that create a functional nuclear export signal. Oncogene *25*, 4376–4380.

Marques, E., and Klefström, J. (2015). Par6 family proteins in cancer. Oncoscience *2*, 894–895.

Martins, V.L., Vyas, J.J., Chen, M., Purdie, K., Mein, C.A., South, A.P., Storey, A., McGrath, J.A., and O'Toole, E.A. (2009). Increased invasive behaviour in cutaneous squamous cell carcinoma with loss of basement-membrane type VII collagen. J. Cell Sci. *122*, 1788–1799.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, 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.

McMahon, M., Ayllón, V., Panov, K.I., and O'Connor, R. (2010). Ribosomal 18 S RNA Processing by the IGF-I-responsive WDR3 Protein Is Integrated with p53 Function in Cancer Cell Proliferation. J. Biol. Chem. *285*, 18309–18318.

McWilliam, H., Li, W., Uludag, M., Squizzato, S., Park, Y.M., Buso, N., Cowley, A.P., and Lopez, R. (2013). Analysis Tool Web Services from the EMBL-EBI. Nucleic Acids Res. *41*, W597–600.

Medina, P.P., and Sanchez-Cespedes, M. (2008). Involvement of the chromatin-remodeling factor BRG1/SMARCA4 in human cancer. Epigenetics *3*, 64–68.

Medina, P.P., Romero, O.A., Kohno, T., Montuenga, L.M., Pio, R., Yokota, J., and Sanchez-Cespedes, M. (2008). Frequent BRG1/SMARCA4–inactivating mutations in human lung cancer cell lines. Hum. Mutat. *29*, 617–622.

Melloni, G.E., Ogier, A.G., de Pretis, S., Mazzarella, L., Pelizzola, M., Pelicci, P.G., and Riva, L. (2014). DOTS-Finder: a comprehensive tool for assessing driver genes in cancer genomes. Genome Med. *6*, 44.

Melloni, G.E.M., de Pretis, S., Riva, L., Pelizzola, M., Céol, A., Costanza, J., Müller, H., and Zammataro, L. (2016). LowMACA: exploiting protein family analysis for the identification of rare driver mutations in cancer. BMC Bioinformatics *17*, 80.

Metral, S., Machnicka, B., Bigot, S., Colin, Y., Dhermy, D., and Lecomte, M.-C. (2009). AlphaII-spectrin is critical for cell adhesion and cell cycle. J. Biol. Chem. *284*, 2409–2418.

Michailidou, K., Hall, P., Gonzalez-Neira, A., Ghoussaini, M., Dennis, J., Milne, R.L., Schmidt, M.K., Chang-Claude, J., Bojesen, S.E., Bolla, M.K., et al. (2013). Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nat. Genet. *45*, 353–361.

Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P.A., Harshman, K., Tavtigian, S., Liu, Q., Cochran, C., Bennett, L.M., and Ding, W. (1994). A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science *266*, 66–71.

Morandi, A., Plaza-Menacho, I., and Isacke, C.M. (2011). RET in breast cancer: functional and therapeutic implications. Trends Mol. Med. *17*, 149–157.

Murphree, A.L., and Benedict, W.F. (1984). Retinoblastoma: clues to human oncogenesis. Science *223*, 1028–1033.

Mutation Consequences and Pathway Analysis working group of the International Cancer Genome Consortium (2015). Pathway and network analysis of cancer genomes. Nat. Methods *12*, 615–621.

Myatt, S.S., and Lam, E.W.-F. (2007). The emerging roles of forkhead box (Fox) proteins in cancer. Nat. Rev. Cancer *7*, 847–859.

Nehrt, N.L., Peterson, T. a, Park, D., and Kann, M.G. (2012). Domain landscapes of somatic mutations in cancer. BMC Genomics *13 Suppl 4*, S9.

Network, T.C.G.A.R. (2013). Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. N. Engl. J. Med. *368*, 2059–2074.

Nishizawa, K., Nishiyama, H., Matsui, Y., Kobayashi, T., Saito, R., Kotani, H., Masutani, H., Oishi, S., Toda, Y., Fujii, N., et al. (2011). Thioredoxin-interacting protein suppresses bladder carcinogenesis. Carcinogenesis *32*, 1459–1466.

Nordling, C.O. (1953). A New Theory on the Cancer-inducing Mechanism. Br. J. Cancer *7*, 68–72.

Nowell, P.C. (1976). The clonal evolution of tumor cell populations. Science *194*, 23–28.

Ohta, T., and Fukuda, M. (2004). Ubiquitin and breast cancer. Oncogene *23*, 2079–2088.

Okamura, K., Feuk, L., Marquès-Bonet, T., Navarro, A., and Scherer, S.W. (2006). Frequent appearance of novel protein-coding sequences by frameshift translation. Genomics *88*, 690–697.

Omasits, U., Ahrens, C.H., Müller, S., and Wollscheid, B. (2014). Protter: interactive protein feature visualization and integration with experimental proteomic data. Bioinforma. Oxf. Engl. *30*, 884–6.

Oren, M., and Rotter, V. (2010). Mutant p53 Gain-of-Function in Cancer. Cold Spring Harb. Perspect. Biol. *2*.

Pallante, P., Federico, A., Berlingieri, M.T., Bianco, M., Ferraro, A., Forzati, F., Iaccarino, A., Russo, M., Pierantoni, G.M., Leone, V., et al. (2008). Loss of the CBX7 gene expression correlates with a highly malignant phenotype in thyroid cancer. Cancer Res. *68*, 6770–6778.

Pallante, P., Terracciano, L., Carafa, V., Schneider, S., Zlobec, I., Lugli, A., Bianco, M., Ferraro, A., Sacchetti, S., Troncone, G., et al. (2010). The loss of the CBX7 gene expression represents an adverse prognostic marker for survival of colon carcinoma patients. Eur. J. Cancer Oxf. Engl. 1990 *46*, 2304–2313.

Pallavi, S.K., Ho, D.M., Hicks, C., Miele, L., and Artavanis-Tsakonas, S. (2012). Notch and Mef2 synergize to promote proliferation and metastasis through JNK signal activation in Drosophila. EMBO J. *31*, 2895–2907.

Pang, H., Flinn, R., Patsialou, A., Wyckoff, J., Roussos, E.T., Wu, H., Pozzuto, M., Goswami, S., Condeelis, J.S., Bresnick, A.R., et al. (2009). Differential enhancement of breast cancer cell motility and metastasis by helical and kinase domain mutations of class IA phosphoinositide 3-kinase. Cancer Res. *69*, 8868–8876.

Papa, A., Wan, L., Bonora, M., Salmena, L., Song, M.S., Hobbs, R.M., Lunardi, A., Webster, K., Ng, C., Newton, R.H., et al. (2014). Cancer-associated PTEN mutants act in a dominant-negative manner to suppress PTEN protein function. Cell *157*, 595–610.

Payne, S.R., and Kemp, C.J. (2005). Tumor suppressor genetics. Carcinogenesis *26*, 2031–2045.

Perou, C.M., Sørlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A., et al. (2000). Molecular portraits of human breast tumours. Nature *406*, 747–752.

Peterson, T. a, Adadey, A., Santana-Cruz, I., Sun, Y., Winder, A., and Kann, M.G. (2010). DMDM: domain mapping of disease mutations. Bioinforma. Oxf. Engl. *26*, 2458–9.

Petrucelli, N., Daly, M.B., and Feldman, G.L. (2010). Hereditary breast and ovarian cancer due to mutations in BRCA1 and BRCA2. Genet. Med. *12*, 245–259.

Pleasance, E.D., Stephens, P.J., O'Meara, S., McBride, D.J., Meynert, A., Jones, D., Lin, M.-L., Beare, D., Lau, K.W., Greenman, C., et al. (2010). A small cell lung cancer genome reports complex tobacco exposure signatures. Nature *463*, 184–190.

Prior, I. a, Lewis, P.D., and Mattos, C. (2012). A comprehensive survey of ras mutations in cancer.

Pruitt, K.D., Tatusova, T., Klimke, W., and Maglott, D.R. (2009). NCBI Reference Sequences: current status, policy and new initiatives. Nucleic Acids Res. *37*, D32–D36.

Pugh, T.J., Weeraratne, S.D., Archer, T.C., Pomeranz Krummel, D.A., Auclair, D., Bochicchio, J., Carneiro, M.O., Carter, S.L., Cibulskis, K., Erlich, R.L., et al. (2012). Medulloblastoma exome sequencing uncovers subtype-specific somatic mutations. Nature *488*, 106–110.

Pylayeva-Gupta, Y., Grabocka, E., and Bar-Sagi, D. (2011). RAS oncogenes: weaving a tumorigenic web. Nat. Rev. Cancer *11*, 761–74.

Rahman, N. (2014). Realizing the promise of cancer predisposition genes. Nature *505*, 302–308.

Rahman, N., Seal, S., Thompson, D., Kelly, P., Renwick, A., Elliott, A., Reid, S., Spanova, K., Barfoot, R., Chagtai, T., et al. (2007). PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. Nat. Genet. *39*, 165–167.

Read, M.L., Lewy, G.D., Fong, J.C.W., Sharma, N., Seed, R.I., Smith, V.E., Gentilin, E., Warfield, A., Eggo, M.C., Knauf, J.A., et al. (2011). Proto-oncogene PBF/PTTG1IP regulates thyroid cell growth and represses radioiodide treatment. Cancer Res. *71*, 6153–6164.

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.

Reimand, J., Wagih, O., and Bader, G.D. (2013). The mutational landscape of phosphorylation signaling in cancer. Sci. Rep. *3*.

Reva, B., Antipin, Y., and Sander, C. (2011). Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. *39*, e118–e118.

Ripperger, T., Gadzicki, D., Meindl, A., and Schlegelberger, B. (2008). Breast cancer susceptibility: current knowledge and implications for genetic counselling. Eur. J. Hum. Genet. *17*, 722–731.

Rivlin, N., Brosh, R., Oren, M., and Rotter, V. (2011). Mutations in the p53 Tumor Suppressor Gene. Genes Cancer *2*, 466–474.

Robinson, J.L.L., Holmes, K.A., and Carroll, J.S. (2013). FOXA1 mutations in hormone-dependent cancers. Front. Oncol. *3*.

Rubin, A.F., and Green, P. (2009). Mutation patterns in cancer genomes. Proc. Natl. Acad. Sci. *106*, 21766–21770.

Rubio, I.G.S., and Medeiros-Neto, G. (2009). Mutations of the thyroglobulin gene and its relevance to thyroid disorders. Curr. Opin. Endocrinol. Diabetes Obes. *16*, 373–378.

Rudin, C.M., Avila-Tang, E., Harris, C.C., Herman, J.G., Hirsch, F.R., Pao, W., Schwartz, A.G., Vahakangas, K.H., and Samet, J.M. (2009). LUNG CANCER IN NEVER SMOKERS: MOLECULAR PROFILES AND THERAPEUTIC IMPLICATIONS. Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. *15*, 5646–5661.

Sakoda, L.C., Jorgenson, E., and Witte, J.S. (2013). Turning of COGS moves forward findings for hormonally mediated cancers. Nat. Genet. *45*, 345–348.

Schrader KA, Cheng DT, Joseph V, and et al (2016). GErmline variants in targeted tumor sequencing using matched normal dna. JAMA Oncol. *2*, 104–111.

Schroeder, M.P., Rubio-Perez, C., Tamborero, D., Gonzalez-Perez, A., and Lopez-Bigas, N. (2014). OncodriveROLE classifies cancer driver genes in loss of function and activating mode of action. Bioinformatics *30*, i549–i555.

Schroeder, R.D., Angelo, L.S., and Kurzrock, R. (2013). NF2/Merlin in hereditary neurofibromatosis 2 versus cancer: biologic mechanisms and clinical associations. Oncotarget *5*, 67–77.

Schwarz, J.M., Cooper, D.N., Schuelke, M., and Seelow, D. (2014). MutationTaster2: mutation prediction for the deep-sequencing age. Nat. Methods *11*, 361–2.

Scott, C.L., Gil, J., Hernando, E., Teruya-Feldstein, J., Narita, M., Martínez, D., Visakorpi, T., Mu, D., Cordon-Cardo, C., Peters, G., et al. (2007). Role of the chromobox protein CBX7 in lymphomagenesis. Proc. Natl. Acad. Sci. U. S. A. *104*, 5389–5394.

Shah, S.P., Morin, R.D., Khattra, J., Prentice, L., Pugh, T., Burleigh, A., Delaney, A., Gelmon, K., Guliany, R., Senz, J., et al. (2009). Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. Nature *461*, 809–813.

Sharp, A.M. (2005). Comprehensive genomic analysis of PKHD1 mutations in ARPKD cohorts. J. Med. Genet. *42*, 336–349.

Shen, H., Fridley, B.L., Song, H., Lawrenson, K., Cunningham, J.M., Ramus, S.J., Cicek, M.S., Tyrer, J., Stram, D., Larson, M.C., et al. (2013). Epigenetic analysis leads to identification of HNF1B as a subtype-specific susceptibility gene for ovarian cancer. Nat. Commun. *4*, 1628.

Shihab, H. a., Gough, J., Cooper, D.N., Stenson, P.D., Barker, G.L. a, Edwards, K.J., Day, I.N.M., and Gaunt, T.R. (2013a). Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models. Hum. Mutat. *34*, 57–65.

Shihab, H.A., Gough, J., Cooper, D.N., Day, I.N.M., and Gaunt, T.R. (2013b). Predicting the functional consequences of cancer-associated amino acid substitutions. Bioinformatics *29*, 1504–1510.

Shivapurkar, N., Sood, S., Wistuba, I.I., Virmani, A.K., Maitra, A., Milchgrub, S., Minna, J.D., and Gazdar, A.F. (1999). Multiple Regions of Chromosome 4 Demonstrating Allelic Losses in Breast Carcinomas. Cancer Res. *59*, 3576–3580.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. *7*, 539.

Silverman, B.W. (1986). Density estimation for statistics and data analysis (CRC press).

Sim, N.-L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., and Ng, P.C. (2012). SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res. *40*, W452–W457.

Siraj, A.K., Beg, S., Jehan, Z., Prabhakaran, S., Ahmed, M., R.Hussain, A., Al-Dayel, F., Tulbah, A., Ajarim, D., and Al-Kuraya, K.S. (2015). ALK alteration is a frequent event in aggressive breast cancers. Breast Cancer Res. *17*, 127.

Sjöblom, T., Jones, S., Wood, L.D., Parsons, D.W., Lin, J., Barber, T.D., Mandelker, D., Leary, R.J., Ptak, J., Silliman, N., et al. (2006). The Consensus Coding Sequences of Human Breast and Colorectal Cancers. Science *314*, 268–274.

Smith, J.M., and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. Genet. Res. *23*, 23–35.

Stamatakis, L., Metwalli, A.R., Middelton, L.A., and Linehan, W.M. (2013). Diagnosis and Management of BHD-Associated Kidney Cancer. Fam. Cancer *12*, 397–402.

Stehelin, D., Varmus, H.E., Bishop, J.M., and Vogt, P.K. (1976). DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. Nature *260*, 170–173.

Stenmark, H. (2009). Rab GTPases as coordinators of vesicle traffic. Nat. Rev. Mol. Cell Biol. *10*, 513–525.

Stouffer S, DeVinney L, and Suchmen E (1949). The American soldier: Adjustment during army life. (Princeton US: Princeton University Press).

Stransky, N., Egloff, A.M., Tward, A.D., Kostic, A.D., Cibulskis, K., Sivachenko, A., Kryukov, G.V., Lawrence, M.S., Sougnez, C., McKenna, A., et al. (2011). The Mutational Landscape of Head and Neck Squamous Cell Carcinoma. Science *333*, 1157–1160.

Stratakis, C.A., Kirschner, L.S., and Carney, J.A. (2001). Clinical and Molecular Features of the Carney Complex: Diagnostic Criteria and Recommendations for Patient Evaluation. J. Clin. Endocrinol. Metab. *86*, 4041–4046.

Stratford, A.L., Boelaert, K., Tannahill, L.A., Kim, D.S., Warfield, A., Eggo, M.C., Gittoes, N.J.L., Young, L.S., Franklyn, J.A., and McCabe, C.J. (2005). Pituitary Tumor Transforming Gene Binding Factor: A Novel Transforming Gene in Thyroid Tumorigenesis. J. Clin. Endocrinol. Metab. *90*, 4341–4349.

Tamborero, D., Gonzalez-Perez, A., and Lopez-Bigas, N. (2013a). OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. Bioinformatics *29*, 2238–2244.

Tamborero, D., Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Kandoth, C., Reimand, J., Lawrence, M.S., Getz, G., Bader, G.D., Ding, L., et al. (2013b). Comprehensive identification of mutational cancer driver genes across 12 tumor types. Sci. Rep. *3*.

Tan, A., Abecasis, G.R., and Kang, H.M. (2015). Unified representation of genetic variants. Bioinformatics *31*, 2202–2204.

The UniProt Consortium (2014). Activities at the Universal Protein Resource (UniProt). Nucleic Acids Res. *42*, D191–8.

Tian, D. (2011). Remarkable difference of somatic mutation patterns between oncogenes and tumor suppressor genes. Oncol. Rep.

Timp, W., and Feinberg, A.P. (2013). Cancer as a dysregulated epigenome allowing cellular growth advantage at the expense of the host. Nat. Rev. Cancer *13*, 497–510.

Tourneau, C.L., Delord, J.-P., Gonçalves, A., Gavoille, C., Dubot, C., Isambert, N., Campone, M., Trédan, O., Massiani, M.-A., Mauborgne, C., et al. (2015). Molecularly targeted therapy based on tumour molecular profiling versus conventional therapy for advanced cancer (SHIVA): a multicentre, open-label, proof-of-concept, randomised, controlled phase 2 trial. Lancet Oncol. *16*, 1324–1334.

Tyzzer, E.E. (1916). Tumor Immunity. J. Cancer Res. *1*, 125–156.

Valdar, W.S.J. (2002). Scoring residue conservation. Proteins *48*, 227–41.

Valiente, M., Obenauf, A.C., Jin, X., Chen, Q., Zhang, X.H.-F., Lee, D.J., Chaft, J.E., Kris, M.G., Huse, J.T., Brogi, E., et al. (2014). Serpins Promote Cancer Cell Survival and Vascular Cooption in Brain Metastasis. Cell *156*, 1002–1016.

Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr. Protoc. Bioinforma. Ed. Board Andreas Baxevanis Al *43*, 11.10.1-33.

Vandin, F., Upfal, E., and Raphael, B.J. (2012). De novo discovery of mutated driver pathways in cancer. Genome Res. *22*, 375–385.

Vanhaesebroeck, B., Guillermet-Guibert, J., Graupera, M., and Bilanges, B. (2010). The emerging mechanisms of isoform-specific PI3K signalling. Nat. Rev. Mol. Cell Biol. *11*, 329–341.

Varon, R., Vissinga, C., Platzer, M., Cerosaletti, K.M., Chrzanowska, K.H., Saar, K., Beckmann, G., Seemanová, E., Cooper, P.R., Nowak, N.J., et al. (1998). Nibrin, a Novel DNA Double-Strand Break Repair Protein, Is Mutated in Nijmegen Breakage Syndrome. Cell *93*, 467–476.

Verkman, A.S., Hara-Chikuma, M., and Papadopoulos, M.C. (2008). Aquaporins—new players in cancer biology. J. Mol. Med. Berl. Ger. *86*, 523–529.

Visscher, P.M., Brown, M.A., McCarthy, M.I., and Yang, J. (2012). Five Years of GWAS Discovery. Am. J. Hum. Genet. *90*, 7–24.

Vivanco, I., Robins, H.I., Rohle, D., Campos, C., Grommes, C., Nghiemphu, P.L., Kubek, S., Oldrini, B., Chheda, M.G., Yannuzzi, N., et al. (2012). Differential Sensitivity of Glioma- versus Lung Cancer-Specific EGFR Mutations to EGFR Kinase Inhibitors. Cancer Discov. *2*, 458–471.

Vogelstein, B., and Kinzler, K.W. (2004). Cancer genes and the pathways they control. Nat. Med. *10*, 789–799.

Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., and Kinzler, K.W. (2013). Cancer Genome Landscapes. Science *339*, 1546–1558.

Walfish, S. (2006). A review of statistical outlier methods. Pharm. Technol. *30*, 82.

Walsh, T., Lee, M.K., Casadei, S., Thornton, A.M., Stray, S.M., Pennil, C., Nord, A.S., Mandell, J.B., Swisher, E.M., and King, M.-C. (2010). Detection of inherited mutations for breast and ovarian cancer using genomic capture and massively parallel sequencing. Proc. Natl. Acad. Sci. *107*, 12629–12633.

Wang, X., and Proud, C.G. (2011). mTORC1 signaling: what we still don't know. J. Mol. Cell Biol. *3*, 206–220.

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. *38*, e164.

Wang, L., Lawrence, M.S., Wan, Y., Stojanov, P., Sougnez, C., Stevenson, K., Werner, L., Sivachenko, A., DeLuca, D.S., Zhang, L., et al. (2011). SF3B1 and Other Novel Cancer Genes in Chronic Lymphocytic Leukemia. N. Engl. J. Med. *365*, 2497–2506.

Ward, L.D., and Kellis, M. (2012). Interpreting noncoding genetic variation in complex traits and human disease. Nat. Biotechnol. *30*, 1095–1106.

Watson, I.R., Li, L., Cabeceiras, P.K., Mahdavi, M., Gutschner, T., Genovese, G., Wang, G., Fang, Z., Tepper, J.M., Stemke-Hale, K., et al. (2014). The RAC1 P29S Hotspot Mutation in Melanoma Confers Resistance to Pharmacological Inhibition of RAF. Cancer Res. *74*, 4845–4852.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., et al. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. *42*, D1001–D1006.

Wennerberg, K. (2005). The Ras superfamily at a glance. J. Cell Sci. *118*, 843–846.

Whitman, R.C. (1919). Somatic Mutations as a Factor in the Production of Cancer: A Critical Review of v. Hansemann's Theory of Anaplasia in the Light of Modern Knowledge of Genetics. J. Cancer Res. *4*, 181–202.

Wilbur, W.J., and Lipman, D.J. (1983). Rapid similarity searches of nucleic acid and protein data banks. Proc. Natl. Acad. Sci. U. S. A. *80*, 726–730.

Winkler, G.S., Kristjuhan, A., Erdjument-Bromage, H., Tempst, P., and Svejstrup, J.Q. (2002). Elongator is a histone H3 and H4 acetyltransferase important for normal histone acetylation levels in vivo. Proc. Natl. Acad. Sci. U. S. A. *99*, 3517–3522.

Wood, L.D., Parsons, D.W., Jones, S., Lin, J., Sjöblom, T., Leary, R.J., Shen, D., Boca, S.M., Barber, T., Ptak, J., et al. (2007). The Genomic Landscapes of Human Breast and Colorectal Cancers. Science *318*, 1108–1113.

Xu, J., Haigis, K.M., Firestone, A.J., McNerney, M.E., Li, Q., Davis, E., Chen, S.-C., Nakitandwe, J., Downing, J., Jacks, T., et al. (2013). Dominant Role of Oncogene Dosage and Absence of Tumor Suppressor Activity in Nras-Driven Hematopoietic Transformation. Cancer Discov. *3*, 993–1001.

Yamamoto, Y., Tsuzuki, S., Tsuzuki, M., Handa, K., Inaguma, Y., and Emi, N. (2010). BCOR as a novel fusion partner of retinoic acid receptor alpha in a t(X;17)(p11;q12) variant of acute promyelocytic leukemia. Blood *116*, 4274–4283.

Yang, F., Petsalaki, E., Rolland, T., Hill, D.E., Vidal, M., and Roth, F.P. (2015a). Protein domain-level landscape of cancer-type-specific somatic mutations. PLoS Comput. Biol. *11*, e1004147.

Yang, L., Yu, S.-J., Hong, Q., Yang, Y., and Shao, Z.-M. (2015b). Reduced Expression of TET1, TET2, TET3 and TDG mRNAs Are Associated with Poor Prognosis of Patients with Early Breast Cancer. PLOS ONE *10*, e0133896.

Yang, Z., Ro, S., and Rannala, B. (2003). Likelihood models of somatic mutation and codon substitution in cancer genes. Genetics *165*, 695–705.

Yip, S., Butterfield, Y.S., Morozova, O., Chittaranjan, S., Blough, M.D., An, J., Birol, I., Chesnelong, C., Chiu, R., Chuah, E., et al. (2012). Concurrent CIC mutations, IDH mutations, and 1p/19q loss distinguish oligodendrogliomas from other cancers. J. Pathol. *226*, 7–16.

Zhan, X., Hu, Y., Li, B., Abecasis, G.R., and Liu, D.J. (2016). RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. Bioinformatics *32*, 1423–1426.

Zhao, M., Sun, J., and Zhao, Z. (2013). TSGene: a web resource for tumor suppressor genes. Nucleic Acids Res. *41*, D970-6.

Zheng, W., Zhang, B., Cai, Q., Sung, H., Michailidou, K., Shi, J., Choi, J.-Y., Long, J., Dennis, J., Humphreys, M.K., et al. (2013). Common genetic determinants of breast-cancer risk in East Asian women: a collaborative study of 23 637 breast cancer cases and 25 579 controls. Hum. Mol. Genet. *22*, 2539–2550.

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B Stat. Methodol. *67*, 301–320.

7 List of Abbreviations

AML: Acute Myeloid Leukemia

BAM: Binary Annotation Format

BLCA: Bladder Carcinoma

BMR: Background Mutation Rate

BRCA: Breast Carcinoma

Bp: base pair

CDG: Candidate Driver Gene

CGC: Cancer Gene Census

CIViC: Clinic Interpretation of Variants in Cancer

CMC: Combined Multivariate and Collapsing Method

COSMIC: Catalogue of Somatic Mutations in Cancer

DoCM: Database of Curated Mutations

DOTS-Finder: Driver Oncogene and Tumor Suppressor Finder

ExAC: Exome Aggregation Consortium

FDR: False Discovery Rate

GATK: Genome Analysis Tool Kit

GUI: Graphic User Interface

GWAS: Genome Wide Association Study

HCD: High Confidence Driver Gene

InDel: Small Insertion-Delition

Kbac: Kernel-based Adaptive Cluster Test

LOF: Loss-of-function

LowMACA: Low frequency Mutation Analysis using Consensus Alignment

MAF: Minor Allele Frequency and Mutation Annotation Format

MCC: Matthews Correlation Coefficient

MNSp: Median number of Non-Silent mutations *per* patient

NCDG: New Candidate Driver Gene

NGS: Next Generation Sequencing

NS: Never Smoker

OG: Oncogene

OG-S: Oncogene Score

Ras Trio: the genes *KRAS, NRAS* and *HRAS*

SKAT-O: Sequence Kernel Association Test Optimal

SNP: Single Nucleotide Polymorphism

SNV: Single Nucleotide Variant

TCGA: The Cancer Genome Atlas

THCA: Thyroid Cancer

TSG: Tumor Suppressor Gene

TSG-S: Tumor Suppressor Gene Score

WES: Whole Exome Sequencing

WGS: Whole Genome Sequencing

8 Appendix

Multiple Mieloma		
205		
53		
Gene name	**NS freq**	**q-value**
TP53	0.0732	0.0000
SP140	0.0293	0.1088
FGFR3	0.0244	0.2249
PLXDC2	0.0195	0.2249
EZR	0.0146	0.3487
NRAS	0.1805	0.0000
CCND1	0.0439	0.0000
TP53	0.0732	0.0000
ACTG1	0.0341	0.0000
BRAF	0.0634	0.0000
EGR1	0.0341	0.0003
KRTDAP	0.0098	0.0003
DTX1	0.0293	0.0010
MOGAT3	0.0195	0.0056
TRAF2	0.0049	0.0150
IDH1	0.0146	0.0817

Carcinoid		
54		
33		
Gene name	**NS freq**	**q-value**
CDKN1B	0.093	0.000
PRDM9	0.056	0.035
CACNA1E	0.074	0.054
LAX		
Gene name	**NS freq**	**q-value**
CDKN1B	0.0926	0.0000
PRDM9	0.0556	0.0182
ATM	0.0741	0.0291
ERN2	0.0370	0.1107
TP53BP1	0.0741	0.1005

	tumor suppressors
	oncogenes
	additional genes below the threshold of significance
NS freq	frequency of non silent mutations in the patients

Lung Small Cell Carcinoma		
73		
178		
Gene name	**NS freq**	**q-value**
RB1	0.4384	0.0000
TP53	0.7808	0.0000
MYH2	0.1644	0.0009
MNDA	0.0822	0.0045
HCN1	0.2192	0.0045
KIF21A	0.1507	0.0045
NLRP8	0.0685	0.0518
ZIM3	0.0685	0.0531
IL1RAPL2	0.0959	0.0557
IL26	0.0411	0.0572
ROBO4	0.0959	0.0572
SLIT2	0.0959	0.0572
ELAVL2	0.1096	0.0733
TP53	0.7808	0.0000
COL22A1	0.2055	0.0000
OR4K17	0.0548	0.0142

Rhabdoid_Tumor		
33		
5		
Gene name	**NS freq**	**q-value**
SMARCB1	0.2121	0.0000
LAX		
Gene name	**NS freq**	**q-value**
SMARCB1	0.2121	0.0000
ZNF433	0.0606	0.0165
SMARCB1	0.2121	0.0000
GABRB2	0.0606	0.0167

MCF2	0.0822	0.0159
OR5F1	0.0411	0.0159
IL1RAPL2	0.0959	0.0249
UBE2NL	0.0411	0.0249
NPAP1	0.1096	0.0479
OR5B2	0.0411	0.0479
DIP2C	0.0959	0.0771

Neuroblastoma

283

14

Gene name	NS freq	q-value
ZNF717	0.021201413	0.000151487
ALK	0.084805654	0
IL18RAP	0.014134276	0.094614427
SIGLEC1	0.014134276	0.094614427

LAX

Gene name	NS freq	q-value
ZNF717	0.021201413	0.000540186
ALK	0.084805654	0
MYCN	0.017667845	0.002450297

Liver Hepatocellular Carcinoma

152

44

Gene name	NS freq	q-value
ALB	0.078947368	0
ARID1A	0.098684211	1.03E-09
AXIN1	0.059210526	1.30E-09
ERRFI1	0.026315789	0.000314282
ARID2	0.059210526	0.000542467
RPS6KA3	0.039473684	0.001846214
SIGLEC12	0.026315789	0.019178482
ZNF226	0.026315789	0.019580451
ACVR2A	0.032894737	0.051847305
BRD7	0.019736842	0.133207108
CTNNB1	0.118421053	0
TP53	0.263157895	0
WWP1	0.052631579	2.25E-06
NFE2L2	0.046052632	2.62E-05
UBR3	0.052631579	0.001837446
USP25	0.026315789	0.00675368
IGSF10	0.046052632	0.01772374
VIM	0.026315789	0.020421845
ZNF804B	0.032894737	0.036041033

Biliary Tract

9

22

Gene name	NS freq	q-value
TP53	0.666666667	7.87E-09
ROBO2	0.222222222	0.014403669
MLL3	0.222222222	0.033543191

LAX

Astrocytoma

52

1

Gene name	NS freq	q-value
FGFR1	0.076923077	6.84E-06

Gene name	NS freq	q-value
TP53	0.666666667	5.11E-08
SMAD4	0.444444444	1.68E-05
RNF43	0.333333333	0.001132554
NDC80	0.222222222	0.007054856
ROBO2	0.222222222	0.037449541
MLL3	0.222222222	0.109015372
ARID1A	0.111111111	0.945938479
TP53	0.666666667	6.71E-10
SMAD4	0.444444444	7.31E-06
GNAS	0.222222222	0.008241586
RNF43	0.333333333	0.057650829

Hematological NS			Lymphoma B-cell		
473			83		
10			70		
Gene name	**NS freq**	**q-value**	**Gene name**	**NS freq**	**q-value**
MLL2	0.063424947	0	SGK1	0.120481928	0
FAM46C	0.019027484	4.94E-08	TNFRSF14	0.108433735	2.02E-13
NPM1	0.014799154	1.08E-07	FBXO11	0.096385542	1.58E-10
WT1	0.019027484	1.08E-07	CREBBP	0.240963855	1.30E-08
TMEM30A	0.014799154	0.000904284	B2M	0.072289157	1.33E-08
RUNX1	0.016913319	0.001985932	DUSP2	0.048192771	0.000766125
FAM5B	0.010570825	0.087584538	HNRNPU	0.048192771	0.0430406
ZRSR2	0.00845666	0.087584538	PFN1	0.024096386	0.091884294
ASXL1	0.012684989	0.226676493	NFKBIA	0.036144578	0.105566134
BCOR	0.00845666	0.634568733	C10orf12	0.048192771	0.222787593
BCL2	0.090909091	0	ETS1	0.048192771	0.409245387
CCND3	0.029598309	0	MLL2	0.168674699	0.937228598
CREBBP	0.038054968	0	BCL2	0.240963855	0
EZH2	0.073995772	0	MYC	0.204819277	0
ID3	0.031712474	0	PIM1	0.180722892	0
MYC	0.044397463	0	SGK1	0.120481928	0
MYD88	0.025369979	0	TP53	0.216867470	0
SF3B1	0.035940803	0	ACTB	0.084337349	4.96E-10
SGK1	0.025369979	0	MYD88	0.084337349	2.98E-09

	9			9	
	0.09090909			0.24096385	
TP53	1	0	CREBBP	5	3.94E-09
	0.03171247			0.07228915	
CARD11	4	2.90E-13	SMARCA4	7	1.18E-07
	0.02748414			0.07228915	
NRAS	4	6.62E-12	EZH2	7	0.000143132
	0.02325581				
MEF2B	4	1.08E-11	CARD11	0.13253012	0.00068373
	0.02325581			0.04819277	
PIM1	4	1.47E-11	STAT6	1	0.00246889
	0.01479915			0.08433734	
IRF4	4	2.92E-09	BCR	9	0.003671572
	0.02325581			0.06024096	
FLT3	4	5.74E-07	ZNF608	4	0.010139367

Chronic Lymphocytic Leukemia (CLL)		
223		
15		
Gene name	**NS freq**	**q-value**
	0.03587443	
ATM	9	1.03E-06
CHD2	0.02690583	0.001993193
	0.02242152	
KDM6A	5	0.00913055
ARID1A	0.01793722	0.968290143
SPEN	0.01793722	0.968290143
	0.05381165	
MYD88	9	0
	0.11210762	
SF3B1	3	0
	0.09417040	
TP53	4	0
RPS15	0.01793722	1.77E-08
MED12	0.01793722	0.018970565
	0.01345291	
LYN	5	0.020293661

Soft Tissue Sarcoma		
15		
5		
Gene name	**NS freq**	**q-value**
	0.06666666	
ZIC3	7	0.00425854

Oligodendroglioma		
16		
17.5		
Gene name	**NS freq**	**q-value**
FUBP1	0.125	0.04722897
	0.41666666	
CIC	7	0
	0.41666666	
IDH1	7	0

LAX		
Gene name	**NS freq**	**q-value**
	0.41666666	
CIC	7	0
FUBP1	0.125	0.173117289
CIC	0.9375	0
IDH1	0.9375	0
NOTCH1	0.3125	5.12E-05
PIK3CA	0.25	0.00027747
PDCD6IP	0.125	0.019122225
PKD1L2	0.125	0.022348156

Acute Lymphoblastic Leukemia (ALL)		
125		
7		
LAX		
Gene name	**NS freq**	**q-value**
PHF6	0.056	3.99E-06
TP53	0.048	6.92E-05
PHF6	0.056	2.30E-05

SLC26A3	0.125	0.022348156
FARP2	0.125	0.034964368
HIVEP2	0.125	0.039569068
KCNH6	0.125	0.039569068
RIN1	0.125	0.039569068
RNPEPL1	0.125	0.039569068

Stomach Adenocarcinoma		
300		
127.5		
Gene name	**NS freq**	**q-value**
CBWD1	0.0967	0.0000
SEC31A	0.0300	0.0000
TP53	0.3867	0.0000
ARID1A	0.1667	0.0166
TP53	0.3867	0.0000
PGM5	0.0833	0.2196

Head and neck squamous cell carcinoma		
416		
116		
Gene name	**NS freq**	**q-value**
CASP8	0.0841	0.0000
CDKN2A	0.1731	0.0000
FAT1	0.1947	0.0000
NOTCH1	0.1707	0.0000
TP53	0.6779	0.0000
MLL2	0.1538	0.0000
AJUBA	0.0433	0.0000
EPHA2	0.0409	0.0000
NSD1	0.0962	0.0001
ZNF750	0.0361	0.0003
RASA1	0.0409	0.0009
B2M	0.0168	0.0035
BAGE5	0.0264	0.0150
PRB3	0.0216	0.0252
ITGA8	0.0361	0.0422
HRAS	0.0409	0.0000
PIK3CA	0.1755	0.0000
TP53	0.6779	0.0000
NFE2L2	0.0505	0.0000
FBXW7	0.0505	0.0004
EP300	0.0673	0.0022
HIST1H2BF	0.0096	0.0048
RHOA	0.0120	0.0477

Esophageal Adenocarcinoma		
159		
117		
Gene name	**NS freq**	**q-value**
TP53	0.3208	0.0000
CDKN2A	0.0566	0.0000
ARID1A	0.0566	0.1499
CNTNAP4	0.0566	0.1499
SMARCA4	0.0629	0.1499
TP53	0.3208	0.0000
DOCK2	0.1069	0.0000
ZNF208	0.0755	0.0000
EPHA6	0.0440	0.0204
CNBD1	0.0314	0.0318

Medulloblastoma		
336		
9		
Gene name	**NS freq**	**q-value**
TCH1	0.0506	0.0000
MLL2	0.0565	0.0000
CTDNEP1	0.0238	0.0000
CREBBP	0.0268	0.0005
GPS2	0.0089	0.0151
LDB1	0.0119	0.0510
FBXW7	0.0149	0.1464
TCF4	0.0179	0.1464
BCOR	0.0149	0.1512
ARID2	0.0089	0.2960

ELMO1	0.0629	0.0425	CTNNB1	0.0744	0.0000
RYR3	0.1069	0.0484	DDX3X	0.1101	0.0000
NYAP2	0.0377	0.0662	SMARCA4	0.0565	0.0000
GABRA6	0.0503	0.0897	TP53	0.0298	0.0000
			SMO	0.0208	0.0000
			CREBBP	0.0268	0.0001
			SF3B1	0.0089	0.0090
			TBR1	0.0089	0.0090
			CLEC12B	0.0119	0.0387

Skin Melanoma			**Low Grade Glioma**		
390			228		
237.5			42		
Gene name	**NS freq**	**q-value**	**Gene name**	**NS freq**	**q-value**
CDKN2A	0.1385	0.0000	ATRX	0.4211	0.0000
DNAH7	0.2487	0.0000	CIC	0.1886	0.0000
TP53	0.1641	0.0000	FUBP1	0.0965	0.0000
PTEN	0.0897	0.0000	TCF12	0.0395	0.0000
B2M	0.0231	0.0282	IL32	0.0263	0.0000
BRAF	0.5154	0.0000	EMG1	0.0175	0.0009
C10orf71	0.0718	0.0000	IDH1	0.7675	0.0000
CCDC141	0.0949	0.0000	TP53	0.5088	0.0000
CDKN2A	0.1385	0.0000			
MUC16	0.4487	0.0000			
NPAP1	0.1564	0.0000			
NRAS	0.2564	0.0000			
PRB3	0.0615	0.0000			
RAC1	0.0795	0.0000			
RGPD4	0.1154	0.0000			
STK19	0.0385	0.0000			
TCEB3CL	0.0590	0.0000			
TRIOBP	0.1231	0.0000			
HNRNPCL1	0.0641	0.0001			
IDH1	0.0385	0.0057			

Kidney Papillary Cell Carcinoma			**Kidney Chromophobe**		
111			65		
66			11		
Gene name	**NS freq**	**q-value**	**Gene name**	**NS freq**	**q-value**
IL32	0.0360	0.0006	PTEN	0.0923	0.0000
NF2	0.0631	0.0006	CDKN1A	0.0308	0.0187
SETD2	0.0631	0.0086	KIAA0947	0.0462	0.0232
SCAF11	0.0450	0.0098	TP53	0.2000	0.0000
KDM6A	0.0450	0.0595			
SMARCB1	0.0270	0.0595			
SRCAP	0.0721	0.0643			
SAV1	0.0270	0.0690			
DARS	0.0270	0.0917			
CDC27	0.0360	0.0993			
OGG1	0.0270	0.0993			

MET	0.0811	0.0002
ATP10A	0.0450	0.0688
NFE2L2	0.0270	0.0688
PCF11	0.0631	0.0688

Kidney All		
534		
64		
Gene name	**NS freq**	**q-value**
BAP1	0.0712	0.0000
PBRM1	0.2416	0.0000
SETD2	0.0899	0.0000
VHL	0.3052	0.0000
KDM5C	0.0412	0.0000
PTEN	0.0356	0.0000
NF2	0.0243	0.0049
GFRAL	0.0131	0.0337
STAG2	0.0281	0.0685
SMARCB1	0.0131	0.0852
MTOR	0.0693	0.0000
MUC4	0.0843	0.0000
TP53	0.0449	0.0000
MUC2	0.0300	0.0145
SMARCA4	0.0243	0.0145
SPAM1	0.0150	0.0145
NFE2L2	0.0150	0.0255

Kidney Clear Cell Carcinoma		
355		
71		
Gene name	**NS freq**	**q-value**
BAP1	0.0958	0.0000
KDM5C	0.0620	0.0000
PBRM1	0.3465	0.0000
SETD2	0.1127	0.0000
VHL	0.4479	0.0000
SCAF4	0.0254	0.0355
BAP1	0.0958	0.0000
MTOR	0.0901	0.0000
MUC4	0.1183	0.0000
SETD2	0.1127	0.0000
TP53	0.0254	0.0038
MUC2	0.0338	0.0108
TRIM51	0.0113	0.0376
SPAM1	0.0169	0.0405
OR5H1	0.0085	0.0501
SMARCA4	0.0225	0.0603

Pancreatic Adenocarcinoma		
401		
16		
Gene name	**NS freq**	**q-value**
RNF43	0.0449	0.0000
SMAD4	0.1471	0.0000
TP53	0.2195	0.0000
ARID1A	0.0349	0.0000
MLL3	0.0574	0.0002
MEN1	0.0175	0.0497
KRAS	0.3541	0.0000
TP53	0.2195	0.0000
CTNNB1	0.0224	0.0000
GNAS	0.0224	0.0000
SF3B1	0.0200	0.0001
CDH10	0.0200	0.0019
ANKRD20A4	0.0050	0.0025
GABRQ	0.0100	0.0091
PRAMEF11	0.0075	0.0142
CHGB	0.0150	0.0223
MYH6	0.0150	0.0406
TEX2	0.0175	0.0566

Glioblastoma		
361		
65		
Gene name	**NS freq**	**q-value**
NF1	0.1191	0.0000
PIK3R1	0.1136	0.0000
PTEN	0.2825	0.0000
RB1	0.0776	0.0000
TP53	0.2521	0.0000
NOX4	0.0249	0.0000
RPL5	0.0194	0.0000
ZNF431	0.0111	0.0000
STAG2	0.0332	0.0001
TPTE2	0.0222	0.0052
ATRX	0.0443	0.0054
CHD8	0.0277	0.0378
NBPF9	0.0166	0.0669
EGFR	0.2188	0.0000
PIK3CA	0.1025	0.0000
PIK3R1	0.1136	0.0000
TP53	0.2521	0.0000
IDH1	0.0388	0.0000

GPR133	0.0150	0.0910	ABCC9	0.0305	0.0117
KANSL1	0.0175	0.0910	KCNB2	0.0194	0.0273
			PIK3R5	0.0305	0.0283
			TRABD2A	0.0111	0.0283
			AMER3	0.0194	0.0658
			KHDC3L	0.0083	0.0658
Colorectal Adenocarcinoma			**Lung Squamous Cell Carcinoma**		
328			179		
90			290		
Gene name	**NS freq**	**q-value**	**Gene name**	**NS freq**	**q-value**
APC	0.6463	0.0000	CDKN2A	0.1453	0.0000
TP53	0.5152	0.0000	CSMD3	0.4525	0.0000
TGIF1	0.0183	0.0813	TP53	0.7877	0.0000
KRAS	0.4329	0.0000	COL11A1	0.1955	0.0000
SMAD4	0.1372	0.0000	MROH2B	0.1453	0.0000
TP53	0.5152	0.0000	EPB41L3	0.0950	0.0006
NRAS	0.0762	0.0000	CLSTN2	0.0950	0.0013
KRTAP1-3	0.0122	0.0157	PTEN	0.0782	0.0032
			REG3G	0.0447	0.0035
			MYH8	0.1173	0.0054
			DPPA4	0.0670	0.0061
			MLL2	0.1955	0.0070
			BAI3	0.1229	0.0086
			PRIM2	0.0726	0.0086
			RB1	0.0670	0.0086
			ADAM2	0.0670	0.0137
			DNAH5	0.1788	0.0174
			ELTD1	0.1006	0.0284
			ACSM2B	0.0726	0.0371
			NFE2L2	0.1508	0.0000
			TP53	0.7877	0.0000
			ZNF208	0.1341	0.0004
Uterin Carcinoma			**Ovarian Adenocarcinoma**		
248			503		
66			51		
Gene name	**NS freq**	**q-value**	**Gene name**	**NS freq**	**q-value**
ARID1A	0.3347	0.0000	TP53	0.6759	0.0000
ARID5B	0.1169	0.0000	BRCA1	0.0398	0.0000
CTCF	0.1815	0.0000	NF1	0.0497	0.0000
IK	0.0323	0.0000	CDK12	0.0298	0.0005
PIK3R1	0.3347	0.0000	RB1	0.0278	0.0007
PTEN	0.6492	0.0000	IL21R	0.0159	0.0212
RPL22	0.1250	0.0000	SNTG1	0.0159	0.0418
ZFHX3	0.1774	0.0000	TP53	0.6759	0.0000
CTNNB1	0.2984	0.0000	PPP2R1A	0.0139	0.4807
KRAS	0.2137	0.0000			
PIK3R1	0.3347	0.0000			
PTEN	0.6492	0.0000			

TP53	0.2782	0.0000
PPP2R1A	0.1089	0.0016

Lung Adenocarcinoma			Prostate Adenocarcinoma		
244			403		
231			39		
Gene name	**NS freq**	**q-value**	**Gene name**	**NS freq**	**q-value**
COL11A1	0.2049	0.0000	PTEN	0.0496	0.0000
STK11	0.0820	0.0000	KDM6A	0.0273	0.0001
TP53	0.5205	0.0000	OR2T35	0.0099	0.0001
TPTE	0.1025	0.0000	CDKN1B	0.0174	0.0009
CDKN2A	0.0574	0.0070	GPATCH4	0.0124	0.0954
CHDC2	0.0451	0.0073	APC	0.0323	0.1001
RBM10	0.0492	0.0435	SPOP	0.0844	0.0000
TAAR5	0.0328	0.0486	TP53	0.1315	0.0000
PNLIP	0.0287	0.2889	FOXA1	0.0298	0.0000
SMARCA4	0.0697	0.2889	NKX3-1	0.0174	0.0059
KRAS	0.2746	0.0000	AR	0.0223	0.0253
TP53	0.5205	0.0000	CTNNB1	0.0199	0.0337
EGFR	0.1189	0.0287	LPAR1	0.0149	0.0974
NPAP1	0.1393	0.0568			

Appendix Table 1 Application of DOTS-Finder to 30 tumor types. The frequency of non-silent mutation is reported for every gene (NS Freq) with a q-value < 0.1. Genes around the threshold of significance are reported in yellow. In white, detected oncogenes are reported, in green, detected tumor suppressors. Under each tumor type name, number of samples and median number of non-silent mutations per sample are reported in the order.

Breast Cancer						
1046						
36						
DOTS-Finder			TCGA - Publication	Music	TUSON Explorer	Mutsig
Gene name	NS freq	q-value				
CBFB	0.0210	0.0000	TP53	MAP2K4	TP53	PIK3CA
CDH1	0.0621	0.0000	PIK3CA	PIK3CA	GATA3	TP53
GATA3	0.0946	0.0000	GATA3	KRAS	MAP3K1	GATA3
MAP2K4	0.0392	0.0000	MAP3K1	TP53	CDH1	MAP3K1
MAP3K1	0.0698	0.0000	MLL3	TBL1XR1	MLL3	PTEN
PTEN	0.0402	0.0000	CDH1	PIK3R1	PTEN	AKT1
TP53	0.3375	0.0000	MAP2K4	CBFB	MAP2K4	CTCF
TBX3	0.0220	0.0000	RUNX1	GATA3	RB1	CBFB
MLL3	0.0650	0.0000	PTEN	MAP3K1	NCOR1	MLL3
AOAH	0.0191	0.0000	TBX3	CDH1	TBX3	MAP2K4
CTCF	0.0210	0.0000	PIK3R1	NCOR1	AOAH	RUNX1
RUNX1	0.0239	0.0000	AKT1	RB1	RUNX1	CDH1
NCOR1	0.0382	0.0000	CBFB	MALAT1	MED23	SF3B1
RB1	0.0210	0.0000	TBL1XR1	TBX3	ARID1A	PIK3R1
NCOR2	0.0315	0.0003	NCOR1	PTEN	RBMX	ARID1A
STXBP2	0.0096	0.0004	CTCF	ARID1A	NF1	NCOR1
AQP7	0.0076	0.0017	ZFP36L1	CTCF	CDKN1B	KRAS
ZFP36L1	0.0115	0.0046	GPS2	AKT1	HNF1A	SPEN
RBMX	0.0124	0.0056	SF3B1	RUNX1	CCDC144NL	RB1
GPS2	0.0067	0.0095	CDKN1B	MLL3	MYB	MLL
CASP8	0.0153	0.0104	USH2A	SF3B1	KDM6A	ERBB2
CDKN1B	0.0076	0.0125	RPGR	NF1	ZFP36L1	TBL1XR1
UBC	0.0076	0.0155	RB1	FOXA1	SETD2	CDKN1B
MED23	0.0134	0.0224	AFF2	VEZF1	NCOR2	HIST1H3B
MYB	0.0115	0.0407	NF1	CDKN1B	TBL1XR1	FOXA1
CCDC144NL	0.0076	0.1268	PTPN22		ARID2	CASP8
GNRH2	0.0029	0.2062	RYR2		FOXA1	MED23
HNF1A	0.0086	0.7280	PTPRD		DUSP16	TBX3
AKT1	0.0220	0.0000	OR6A2		BRCA2	CUL4B
PIK3CA	0.2849	0.0000	HIST1H2BC		CBFB	STAG2
TP53	0.3375	0.0000	GPR32		CTCF	MYB
TBX3	0.0220	0.0000	CLEC19A		PIK3CA	RAB40A
SF3B1	0.0172	0.0000	CCND3		AKT1	EP300
FOXA1	0.0172	0.0001	SEPT13		SF3B1	FGFR2
HIST1H3B	0.0076	0.0001	DCAF4L2			GNPTAB
MEF2A	0.0143	0.0002				ERBB3
PIK3R1	0.0249	0.0008				ACVR1B
ATN1	0.0172	0.0425				
AKD1	0.0182	0.0431				
Thyroid Carcinoma						

326			
19			
DOTS-Finder			
Gene name	**NS freq**	**q-value**	**TUSON Exlorer**
TG	0.0491	0.0000	TG
EMG1	0.0184	0.0000	RPTN
RPTN	0.0245	0.0000	MLL3
PPM1D	0.0153	0.0054	DNMT3A
TMCO2	0.0092	0.0056	CHD2
IL32	0.0092	0.0152	BRAF
DNMT3A	0.0153	0.2896	
BRAF	0.5613	0.0000	
HRAS	0.0368	0.0000	
NRAS	0.0798	0.0000	
TG	0.0491	0.0000	
DNASE2	0.0092	0.0694	
PRDM9	0.0184	0.0816	
DICER1	0.0092	0.1070	
ZNF845	0.0184	0.1070	
PRG4	0.0123	0.1085	
PTTG1IP	0.0123	0.1085	

AML					
196					
11					
DOTS-Finder					
Gene name	**NS freq**	**q-value**	**TCGA - Publication**	**Music**	**Mutsig**
CEBPA	0.0663	0.0000	CEBPA	NPM1	FLT3
NPM1	0.2755	0.0000	DNMT3A	FLT3	DNMT3A
RUNX1	0.0918	0.0000	FLT3	DNMT3A	NPM1
TET2	0.0867	0.0000	IDH1	IDH2	IDH2
TP53	0.0765	0.0000	IDH2	IDH1	IDH1
WT1	0.0612	0.0000	NPM1	RUNX1	TET2
RAD21	0.0255	0.0000	NRAS	TET2	NRAS
PHF6	0.0306	0.0000	RUNX1	NRAS	RUNX1
STAG2	0.0306	0.0000	TET2	TP53	WT1
EZH2	0.0153	0.0007	TP53	CEBPA	U2AF1
ASXL1	0.0255	0.0014	WT1	WT1	TP53
HNRNPK	0.0102	0.0083	KRAS	KRAS	KRAS
CALR	0.0102	0.0142	U2AF1	KIT	PTPN11
CBFB	0.0102	0.0572	KIT	U2AF1	KIT
CBX7	0.0051	0.0948	PTPN11	PTPN11	SMC3
BCOR	0.0102	0.1971	PHF6	MIR142	STAG2
CEBPA	0.0663	0.0000	SMC3	PHF6	PHF6
DNMT3A	0.2602	0.0000	FAM5C	SMC3	RAD21
FLT3	0.2704	0.0000	SMC1A	SMC1A	CEBPA
IDH1	0.0969	0.0000	RAD21	STAG2	ASXL1
IDH2	0.1020	0.0000	STAG2	RAD21	SFRS2
NRAS	0.0765	0.0000	HNRNPK	ASXL1	SMC1A

TP53		0.0765	0.0000	EZH2	EZH2	PAPD5
U2AF1		0.0408	0.0000			EZH2
						PDSS2
						MXRA5
						KDM6A

Bladder Carcinoma
145
177

DOTS-Finder			TCGA - Publication	Music	TUSON Explorer	Mutsig
Gene name	NS freq	q-value				
ARID1A	0.2414	0.0000	UTX	TP53	ARID1A	TP53
CDKN1A	0.1448	0.0000	TP53	ARID1A	KDM6A	KDM6A
KDM6A	0.2138	0.0000	ARID1A	KDM6A	CDKN1A	RB1
TP53	0.2621	0.0000	CREBBP	MALAT1	MLL2	PIK3CA
ELF3	0.0759	0.0000	EP300	CDKN1A	TP53	ARID1A
MLL2	0.2621	0.0000	HRAS	MLL2	MLL	MLL2
EP300	0.1517	0.0000	RB1	RB1	FAT1	CDKN1A
RB1	0.1103	0.0000	PIK3CA	ELF3	MLL3	ERCC2
SPTAN1	0.0966	0.0000	FGFR3	PIK3CA	ELF3	STAG2
MLL3	0.2000	0.0000	STAG2	FBXW7	RB1	RXRA
CREBBP	0.1310	0.0000	SYNE1	PRX	STAG2	TBC1D12
STAG2	0.0897	0.0001	ERCC2	ERCC2	FBXW7	NFE2L2
FOXQ1	0.0483	0.0060	KRAS	EP300	EP300	C3orf70
TXNIP	0.0552	0.0079	MLL	MLL3	CREBBP	ERBB3
FAT1	0.1103	0.0370	NF1	FGFR3	ARHGAP35	ELF3
FBXW7	0.0690	0.0428	SYNE2	STAG2	ASXL2	FBXW7
GCC2	0.0690	0.0800	ANK3		TSC1	FGFR3
ZNF513	0.0552	0.0911	CSMD3		FOXQ1	FOXQ1
KLF5	0.0621	0.1184	ELF3		PIK3C2B	CREBBP
			ESPL1		No Oncogenes	
GPS2	0.0276	0.2599				HRAS
NHLRC1	0.0207	0.2635	LRP2			SNX25
FOXA1	0.0414	0.2872	ANK2			TSC1
TP53	0.2621	0.0000	ATM			MGA
NFE2L2	0.0759	0.0000	CHD6			EZR
ERBB3	0.1172	0.0000	ERBB2			CDKN2A
RARG	0.0690	0.0000	ERBB3			DDX5
IRS4	0.0138	0.6550	FAT4			RHOA
ELP5	0.0138	0.6550	KALRN			PHF6
RPS6	0.0207	0.6550	LAMA4			MLL3
			MLL3			BCLAF1
			NCOR1			TGFBR2
			NFE2L3			EPHA2
			PDZD2			SETD2
			PIK3R4			
			TRAK1			
			TRRAP			

Appendix Table 2 Results of DOTS-Finder on 4 tumor types and comparison with existing tools. The frequency of non-silent mutation is reported for every gene (NS Freq) with a q-value < 0.1. Genes around

the threshold of significance are reported in yellow. In white, detected oncogenes are reported, in green, detected tumor suppressors. Under each tumor type name, number of samples and median number of non-silent mutations per sample are reported in this order.

Lung adenocarcinoma non smoker			Lung Adenocarcinoma		
62			244		
83.5			231		
Gene name	**NS freq**	**q-value**	**Gene name**	**NS freq**	**q-value**
TP53	0.3065	0.0000	COL11A1	0.2049	0.0000
NBPF1	0.1290	0.0000	STK11	0.0820	0.0000
IL32	0.0645	0.0005	TP53	0.5205	0.0000
KEAP1	0.0806	0.0005	TPTE	0.1025	0.0000
STK11	0.0806	0.0005	CDKN2A	0.0574	0.0070
RPL5	0.0484	0.0008	CHDC2	0.0451	0.0073
SDHA	0.0484	0.0008	RBM10	0.0492	0.0435
OR5B3	0.0484	0.0012	TAAR5	0.0328	0.0486
PSME2	0.0323	0.0019	PNLIP	0.0287	0.2889
OR4C16	0.0484	0.0098	SMARCA4	0.0697	0.2889
SETD2	0.0645	0.0100	KRAS	0.2746	0.0000
ECI1	0.0323	0.0200	TP53	0.5205	0.0000
FKBP2	0.0323	0.0495	EGFR	0.1189	0.0287
CEBPZ	0.0484	0.0495	NPAP1	0.1393	0.0568
SMAD4	0.0968	0.0517			
RBM10	0.0484	0.0665			
UXS1	0.0323	0.0764			
MET	0.0645	0.0904			
EGFR	0.3226	0.0000			
KRAS	0.1290	0.0000			
CASP8	0.0484	0.0000			
PLP2	0.0323	0.0024			
AQP10	0.0323	0.0029			
PAPPA2	0.1129	0.0029			
REG1B	0.0484	0.0029			
BROX	0.0323	0.0030			
MC5R	0.0323	0.0030			
OR52I1	0.0323	0.0030			
SPTA1	0.1290	0.0030			
KCNMB1	0.0484	0.0047			
NDUFAF3	0.0323	0.0050			
PAQR9	0.0323	0.0074			
EFCAB12	0.0323	0.0084			
PRSS45	0.0161	0.0134			
OR2T4	0.0323	0.0162			
C11orf63	0.0484	0.0195			
HSD17B6	0.0323	0.0228			
MCF2	0.0484	0.0228			
TCRB	0.0323	0.0228			
PSG3	0.0323	0.0272			
ITGA2B	0.0323	0.0296			
COL25A1	0.0645	0.0379			
OR1M1	0.0484	0.0410			

	tumor suppressors
	oncogenes
	additional genes below the threshold of significance
NS freq	frequency of non silent mutations in the patients

OR2W3	0.0484	0.0410
PROKR2	0.0484	0.0410
WRN	0.0484	0.0410
NCKIPSD	0.0323	0.0488
PLD2	0.0323	0.0603
CR2	0.1129	0.0619
DPY19L2	0.0484	0.0622
HSPA5	0.0323	0.0622
OR14A16	0.0323	0.0695
GRM1	0.0484	0.0869

Appendix Table 3 Results of DOTS-Finder obtained from the complete LUAD dataset and the non-smoker LUAD subset. The frequency of non-silent mutation is reported for every gene (NS Freq) with a q-value < 0.1. Genes around the threshold of significance are reported in yellow. In white, detected oncogenes are reported, in green, detected tumor suppressors. Under each tumor type name, number of samples and median number of non-silent mutations per sample are reported in this order.

Gene Symbol	Protein change	SIFT	Polyphen2	LRT	Mutation Taster	Mutation Assessor	FATHMM	Radial SVM	Meta LR	% Damaging
RAB29	A16T	T	P	N	D	N	T	T	T	0.25
RAB29	E116K	T	B	N	D	L	T	T	T	0.125
RAB29	E68Q	D	D	D	D	M	D	D	D	0.875
RAB29	P117L	T	P	N	D	L	T	T	T	0.25
RAB29	Q60E	D	P	D	D	M	T	D	D	0.75
RAB29	R69H	D	D	D	D	M	T	D	D	0.75
RAB29	R79L	D	D	D	D	M	T	D	D	0.75
RAB29	R79W	D	D	D	D	H	T	D	D	0.875
RAB29	V13E	D	D	D	D	M	T	D	D	0.75
RAB29	W151*	T	.	D	D	.	.	.	.	0.25
RAB29	W62L	D	D	D	D	H	D	D	D	1
RAC1	A159V	D	D	D	D	H	D	D	D	1
RAC1	A59T	D	D	D	D	M	D	D	D	0.875
RAC1	A88E	D	P	D	D	N	T	T	T	0.5
RAC1	C18F	D	D	D	D	H	T	D	D	0.875
RAC1	C18Y	D	D	D	D	H	T	D	D	0.875
RAC1	D63H	D	D	D	D	H	T	D	D	0.875
RAC1	D63N	D	D	D	D	M	T	D	D	0.75
RAC1	D65N	D	P	D	D	M	T	D	D	0.75
RAC1	E31D	D	B	D	D	N	T	T	T	0.375
RAC1	G142S	D	D	D	D	H	T	D	D	0.875
RAC1	G15S	D	D	D	D	H	D	D	D	1
RAC1	I21M	D	D	D	D	L	T	D	D	0.75
RAC1	K116N	D	D	D	D	H	D	D	D	1
RAC1	K116R	D	D	D	D	H	D	D	D	1
RAC1	K116T	D	D	D	D	H	D	D	D	1
RAC1	L177V	D	B	D	D	L	T	T	T	0.375
RAC1	L53V	D	P	D	D	M	D	D	D	0.875
RAC1	N39S	D	P	D	D	L	T	D	D	0.75
RAC1	N92I	D	D	D	D	H	T	D	D	0.875
RAC1	N92K	D	P	D	D	M	T	D	D	0.75
RAC1	P140L	D	B	D	D	N	T	T	T	0.375
RAC1	P29L	D	D	D	D	M	T	D	D	0.75
RAC1	P29S	D	P	D	D	L	T	T	D	0.625
RAC1	P29T	D	P	D	D	M	T	D	D	0.75
RAC1	P34H	D	D	D	D	H	T	D	D	0.875
RAC1	P34S	D	D	D	D	M	T	D	D	0.75
RAC1	P87L	D	B	D	D	M	T	D	D	0.625
RAC1	Q162R	D	D	D	D	N	T	T	T	0.5
RAC1	Q61R	D	D	D	D	M	D	D	D	0.875
RAC1	R102L	T	B	D	D	L	T	T	T	0.25
RAC1	R68H	D	P	D	D	H	T	D	D	0.875
RAC1	S71F	D	D	D	D	M	T	D	D	0.75
RAC1	S86I	D	P	D	D	M	T	D	D	0.75
RAC1	V14E	D	D	D	D	H	D	D	D	1
RAC1	V46G	D	D	D	D	H	T	D	D	0.875

RAC1	V85M	D	P	D	D	M	T	D	D	0.75
RAC1	Y32C	D	D	D	D	H	T	D	D	0.875
RAC1	Y40S	D	P	D	D	M	T	D	D	0.75
RAC2	A27V	T	B	D	D	L	T	T	T	0.25
RAC2	C18R	D	D	D	D	H	T	D	D	0.875
RAC2	D124E	T	B	D	D	N	T	T	T	0.25
RAC2	E62K	D	P	D	D	H	D	D	D	1
RAC2	F82L	D	B	D	D	H	D	D	D	0.875
RAC2	G15D	D	D	D	D	H	D	D	D	1
RAC2	G30R	D	P	D	D	L	T	T	T	0.5
RAC2	I110F	T	B	D	D	L	T	T	T	0.25
RAC2	I21M	D	D	D	D	L	T	D	D	0.75
RAC2	K130R	T	B	D	D	N	T	T	T	0.25
RAC2	P136H	D	D	D	D	M	T	D	D	0.75
RAC2	P29L	T	T	T	T	T	T	T	T	0
RAC2	Q162H	D	P	D	D	M	T	D	T	0.625
RAC2	R102Q	T	B	D	D	L	T	T	T	0.25
RAC2	R102W	D	D	D	D	H	T	D	D	0.875
RAC2	R174W	D	D	D	D	M	T	D	D	0.75
RAC2	T35I	D	D	D	D	H	D	D	D	1
RAC2	V168M	D	D	D	D	M	T	D	D	0.75
RAC2	V36A	.	B	D	D	M	T	D	D	0.5
RAC2	V93I	D	B	D	D	N	T	T	T	0.375
RAC2	W97*	T	.	D	D	.	.	.	.	0.25
RHOC	D120N	D	D	U	D	H	D	D	D	0.875
RHOC	D124fs	D	D	D	D	D	D	D	D	1
RHOC	D59E	D	P	D	D	M	D	D	D	0.875
RHOC	E125Q	T	B	.	D	N	T	T	T	0.125
RHOC	E142K	D	B	.	D	M	T	T	T	0.25
RHOC	E64K	D	D	D	D	H	D	D	D	1
RHOC	G178D	D	B	.	D	N	T	T	T	0.25
RHOC	K162N	T	B	.	D	H	T	T	T	0.25
RHOC	P31S	D	D	D	D	L	T	T	T	0.5
RHOC	R145W	D	D	.	D	L	T	T	T	0.375
RHOC	R150W	D	P	.	D	M	T	T	T	0.375
RHOC	R168L	D	P	.	D	L	T	T	T	0.375
RHOC	R68Q	D	B	D	D	L	T	T	T	0.375
RHOC	S73A	D	B	D	D	L	T	T	T	0.375
RHOC	S73fs	D	D	D	D	D	D	D	D	1
RHOC	V24I	D	B	D	D	M	T	T	T	0.375
RHOC	Y42C	D	P	D	D	M	T	D	T	0.625
RHOT1	A83V	T	D	D	D	M	T	D	D	0.625
RHOT1	D106H	T	B	D	D	N	T	T	T	0.25
RHOT1	D91N	D	B	D	D	L	T	T	T	0.375
RHOT1	E12K	D	D	D	D	M	T	D	D	0.75
RHOT1	E39Q	T	D	D	D	M	T	D	D	0.625
RHOT1	P30L	T	D	D	D	M	T	D	D	0.625
RHOT1	P43S	D	D	D	D	M	T	D	D	0.75
RHOT1	P48fs	D	D	D	D	D	D	D	D	1
RHOT1	R104K	T	B	D	D	N	T	T	T	0.25

RHOT1	S156L	D	D	D	D	M	T	D	D	0.75
RHOT1	V84I	T	D	D	D	L	T	T	T	0.375
RHOT1	Y82H	D	D	N	D	M	D	D	D	0.75

Appendix Table 4 Breakdown of the analysis of some Rab and Rho subfamily members. This table represents all the mutations found in more than 10000 cancer patient (cBioportal Database) on the RAS superfamily Pfam (PF00071). An extended version comes as a supplementary of (Melloni et al., 2016).
SIFT - D: Deleterious (sift<=0.05); T: tolerated (sift>0.05)
Polyphen2 - D: Probably damaging (>=0.957), P: possibly damaging (0.453<=pp2_hdiv<=0.956); B: benign (pp2_hdiv<=0.452)
LRT - D: Deleterious; N: Neutral; U: Unknown
Mutation Taster - A" ("disease_causing_automatic"); "D" ("disease_causing"); "N" ("polymorphism"); "P" ("polymorphism_automatic")
Mutation Assessor - H: high; M: medium; L: low; N: neutral. H/M means functional and L/N means non-functional
FATHMM - D: Deleterious; T: Tolerated
Radial SVM - D: Deleterious; T: Tolerated
LR - D: Deleterious; T: Tolerated

LowMACA VS Disease Associated snps	Fisher Test Result			**Disease Associated Mutations**	
p-value	5.30E-27				
alternative hypothesis:	OR different from 1			no	yes
Confidence Interval	18.30 - 84.98	**LowMACA**	Not Predicted	2257	11
Odds Ratio	38.17336152		Predicted	172	32
Accuracy	0.156862745				
Recall	0.744186047				
F1-Score	0.259109312				

dbNSFP VS Disease Associated snps	Fisher Test Result			**Disease Associated Mutations**	
p-value	3.56E-05				
alternative hypothesis:	OR different from 1			no	yes
Confidence Interval	1.98 - 10.11	**dbNSFP**	Tolerated	1286	9
Odds Ratio	4.250413143		Damaging	1143	34
Accuracy	0.028887001				
Recall	0.790697674				
F1-Score	0.055737705				

LowMACA VS Cancer Associated snps	Fisher Test Result			**Cancer Associated Mutations**	
p-value	1.36E-22				
alternative hypothesis:	OR different from 1			no	yes
Confidence Interval	41.14 - 9985.34	**LowMACA**	notPredicted	2267	1
Odds Ratio	260.147541		Predicted	183	21
Accuracy	0.102941176				
Recall	0.954545455				
F1-Score	0.185840708				

dbNSFP VS Cancer Associated snps	Fisher Test Result			**Cancer Associated Mutations**	
p-value	1.89E-03				
alternative hypothesis:	OR different from 1			no	yes
Confidence Interval	1.64 - 20.40	**dbNSFP**	Tolerated	1291	4
Odds Ratio	5.012510785		Damaging	1159	18
Accuracy	0.015293118				
Recall	0.818181818				
F1-Score	0.030025021				

Appendix Table 5 Statistical Results from the comparison of dbNSFP results and LowMACA results. A set of known pathogenic variants taken from Humsavar and Clinvar and a set of known cancer related mutations taken from DoCM and CiviC was compared to the results detected by LowMACA and the aggregated score from 8 different predictors of phenotypic effect.

Pfam ID	Pfam Name	Multiple Alignment Position	Consensus Amino Acid	Trident Conservation Score	Genes Mutated in Multiple Align Position	Position Qvalue
PF00454	PI3_PI4_kinase	57	C	0.4922	MTOR	1.31E-02
PF00454	PI3_PI4_kinase	365	L	0.3596	ATM|ATR|PI4KAP2|MTOR|PIK3C2A|PIK3CA|PRKDC	1.35E-03
PF00613	PI3Ka	24	E	0.1122	PIK3C2A|PIK3CA	5.68E-61
PF00613	PI3Ka	27	E	0.2313	PIK3CA|PIK3CB|PIK3CG	1.31E-136
PF00613	PI3Ka	28	E	0.2391	PIK3CA|PIK3CB	1.90E-16
PF00792	PI3K_C2	1	E	0.1445	PIK3C2A|PIK3CA	4.21E-02
PF00792	PI3K_C2	16	E	0.3616	PIK3CA|PIK3C2A|PIK3CB	4.21E-02
PF00792	PI3K_C2	108	C	0.1332	PIK3CA|PIK3C2A	1.15E-08
PF00792	PI3K_C2	140	E	0.3173	PIK3C2G|PIK3CA|PIK3CG|PIK3CD	4.21E-02
PF00792	PI3K_C2	142	E	0.1362	PIK3CA|PIK3CB	5.44E-08
PF00792	PI3K_C2	173	A	0.2408	PIK3CA|PIK3CD|PIK3CG	4.87E-02
PF02192	PI3K_p85B	8	R	1.0000	PIK3CA|PIK3CB|PIK3CD	1.57E-07
PF02192	PI3K_p85B	51	E	1.0000	PIK3CA|PIK3CD	1.10E-03
PF02192	PI3K_p85B	58	R	1.0000	PIK3CA|PIK3CD	1.90E-12
PF02192	PI3K_p85B	63	R	0.3856	PIK3CA|PIK3CB	2.03E-04
PF02192	PI3K_p85B	76	G	0.3591	PIK3CA	5.27E-03
PF02192	PI3K_p85B	78	R	0.3661	PIK3CA	1.67E-04

Appendix Table 6 Results of LowMACA on the main PI3K families. In purple, the main Pfam PF00613 encompasses the family I of PIK3, composed by known cancer genes *PIK3CA* and *PIK3CB*.

Variant	Humsavar Disease	ClinVar Disease
SDHB - 1,17354297,A,G		Neoplastic_Syndromes\x2c_Hereditary
MUTYH - 1,45798475,T,C		Endometrial_carcinoma|Neoplastic_Syndromes\x2c_Hereditary
RNASEL - 1,182554557,C,T		Prostate_cancer\x2c_susceptibility_to
RET - 10,43613908,A,T	Multiple_neoplasia_2A_(MEN2A)_[MIM:171400]	Familial_medullary_thyroid_carcinoma
RET - 10,43614996,G,A	Medullary_thyroid_carcinoma_(MTC)_[MIM:155240]	MEN2A_and_FMTC,MEN2_phenotype
PRF1 - 10,72358722,T,C	Familial_hemophagocytic_lymphohistiocytosis_2_(FHL2)_[MIM:603553]	Malignant_lymphoma\x2c_non-Hodgkin
ASCC1 - 10,73892817,T,C		Barrett_esophagus|Esophageal_adenocarcinoma
TYR - 11,89017961,G,A		Cutaneous_malignant_melanoma_8
SDHD - 11,111957665,G,A		Cowden_disease_3|Paragangliomas_1|Carcinoid_tumor_of_intestine|Pheochromocytoma
SDHD - 11,111958677,A,G		Carcinoid_tumor_of_intestine|Neoplastic_Syndromes\x2c_Hereditary
BRCA2 - 13,32907129,T,C	Breast_cancer_(BC)_[MIM:114480]	BRCA1_and_BRCA2_Hereditary_Breast_and_Ovarian_Cancer|Neoplastic_Syndromes\x2c_Hereditary
BRCA2 - 13,32912007,C,T		BRCA1_and_BRCA2_Hereditary_Breast_and_Ovarian_Cancer|Neoplastic_Syndromes\x2c_Hereditary
BRCA2 - 13,32912553,C,T		BRCA1_and_BRCA2_Hereditary_Breast_and_Ovarian_Cancer|Neoplastic_Syndromes\x2c_Hereditary
BRCA2 - 13,32914974,ACAA,-		BRCA1_and_BRCA2_Hereditary_Breast_and_Ovarian_Cancer|Neoplastic_Syndromes\x2c_Hereditary
BRCA2 - 13,32972852,C,T		BRCA1_and_BRCA2_Hereditary_Breast_and_Ovarian_Cancer|Neoplastic_Syndromes\x2c_Hereditary
TSHR - 14,81609723,A,C		Malignant_melanoma
AKT1 - 14,105246551,C,T	Breast_cancer_(BC)_[MIM:114480]|Proteus_syndrome_(PROTEUSS)_[MIM:176920]	Breast_adenocarcinoma|Carcinoma_of_colon|Neoplasm_of_ovary|Proteus_syndrome
CDH1 - 16,68845646,G,A		not_provided|Neoplastic_Syndromes\x2c_Hereditary|Hereditary_diffuse_gastric_cancer
ELAC2 - 17,12896274,C,T		Prostate_cancer\x2c_hereditary\x2c_2
ELAC2 - 17,12915009,G,A	Prostate_cancer,_hereditary,_2_(HPC2)_[MIM:614731]	Prostate_cancer\x2c_hereditary\x2c_2
FLCN - 17,17119708,-,G		Multiple_fibrofolliculomas|Pneumothorax\x2c_primary_spontaneous|not_provided|Neoplastic_Syndromes\x2c_Hereditary
FLCN - 17,17125879,G,A	Renal_cell_carcinoma_(RCC)_[MIM:144700]	
BRCA1 - 17,41245664,ACTG,-		BRCA1_and_BRCA2_Hereditary_Breast_and_Ovarian_Cancer|Breast-ovarian_cancer\x2c_familial_1
BRCA1 - 17,41245683,G,A		Familial_cancer_of_breast|Breast-ovarian_cancer\x2c_familial_1
BRIP1 - 17,59793412,G,A		Fanconi_anemia\x2c_complementation_group_J|Neoplastic_Syndromes\x2c_Hereditary
MSH6 - 2,48030639,-,C		Lynch_syndrome|not_provided|Neoplastic_Syndromes\x2c_Hereditary
CHEK2 - 22,29121058,C,T	Prostate_cancer_(PC)_[MIM:176807]	
CHEK2 - 22,29121087,A,G		Li-Fraumeni_syndrome_2|Colorectal_cancer\x2c_susceptibility_to
VHL - 3,10183605,C,T	Pheochromocytoma_(PCC)_[MIM:171300]	not_specified|Von_Hippel-Lindau_syndrome|not_provided
COL7A1 - 3,48619779,G,A		Malignant_melanoma
APC - 5,112154969,C,T		Gardner_syndrome|not_provided|Neoplastic_Syndromes\x2c_Hereditary|not_specified|Adenomatous_polyposis_coli
PKHD1 - 6,51947999,G,A	Polycystic_kidney_disease,_autosomal_recessive_(ARPKD)_[MIM:263200]	Polycystic_kidney_disease\x2c_infantile_type|COLORECTAL_CANCER\x2c_PROTECTION_AGAINST|not_provided
MSR1 - 8,16012594,G,A		Malignant_tumor_of_prostate|BARRETT_ESOPHAGUS/ESOPHAGEAL_ADENOCARCINOMA
NBN - 8,90983460,G,A		Microcephaly\x2c_normal_intelligence_and_immunodeficiency|Neoplastic_Syndromes\x2c_Hereditary|not_specified
FANCC - 9,97912338,G,A		Fanconi_anemia\x2c_complementation_group_C|Neoplastic_Syndromes\x2c_Hereditary
GALNT12 - 9,101594229,G,A	Colorectal_cancer_1_(CRCS1)_[MIM:608812]	
TSC1 - 9,135779052,G,A		Neoplastic_Syndromes\x2c_Hereditary
AR - X,66937326,G,T		Prostate_cancer_susceptibility

Appendix Table 7 Positive set of cancer associated variants. This set of 38 variants is composed by mutations associated with any kind of cancer, including breast, and is used as a reference set of positive controls in our age-dependent polygenic model (see sections **4.3.3.5.3** and **4.3.4.4**).

Variant	log2FoldFreq	Control MAF	Case MAF	avsnp142	Protein Change	Deleteriousness Score
MYH7B,20,33585277,33585277,C,A	25.7930	0.0000	0.0058	rs540290584	A1236D	0.78
NAALADL1,11,64812885,64812885,G,A	23.6046	0.0000	0.0013		P694L	0.89
FKBP8,19,18650369,18650369,C,T	23.4982	0.0000	0.0012		V152I	0.89
SLC12A7,5,1075570,1075570,G,A	23.3124	0.0000	0.0010	rs372681736	A628V	0.67
EHBP1L1,11,65348582,65348582,C,T	23.1859	0.0000	0.0010		R230*	1.00
PELI3,11,66241362,66241362,G,A	23.1318	0.0000	0.0009		R162Q	0.78
TET2,4,106164793,106164793,T,C	22.9502	0.0000	0.0008		C1221R	0.56
NF2,22,30069387,30069387,C,T	22.9362	0.0000	0.0008		R335C	0.78
COL1A1,17,48263191,48263191,C,T	22.8817	0.0000	0.0008	rs146035171	R1399H	0.89
AMPD1,1,115218614,115218614,G,A	22.8794	0.0000	0.0008	rs587779370	R496C	1.00
MYH7B,20,33577664,33577664,G,A	22.8639	0.0000	0.0008		R612Q	0.56
DFFA,1,10523562,10523562,C,T	22.8464	0.0000	0.0008	rs574523820	R186H	0.67
SGCA,17,48246600,48246600,G,A	22.8356	0.0000	0.0007		W244*	1.00
ANO8,19,17441686,17441686,G,A	22.8291	0.0000	0.0007		T315M	0.56
DHTKD1,10,12126673,12126673,C,T	22.8248	0.0000	0.0007	rs141125831	R149W	0.67
NAALADL1,11,64824854,64824854,G,A	22.8248	0.0000	0.0007		R198C	0.67
COLGALT1,19,17690300,17690300,C,T	22.8248	0.0000	0.0007		R426W	0.78
NDUFA6,22,42482261,42482261,G,A	22.8248	0.0000	0.0007		R131W	0.89
ARHGEF5,7,144075893,144075893,C,T	22.8248	0.0000	0.0007		R1524*	1.00
PIK3R2,19,18266970,18266970,G,A	7.1063	0.0003	0.0357		R94H	0.56
CORO1B,11,67209552,67209552,C,T	6.3551	0.0000	0.0015		R70H	0.78
NEFH,22,29881809,29881809,C,T	5.6188	0.0000	0.0008		A394V	0.89
BBS1,11,66297334,66297334,C,T	5.4727	0.0000	0.0008	rs577426256	R462C	0.78
SYVN1,11,64896178,64896178,-,G	5.4211	0.0001	0.0025		R534fs	1.00
DDX49,19,19035507,19035507,C,T	5.3482	0.0000	0.0007		R310W	0.89
PIK3C2B,1,204438071,204438071,-,G	5.3437	0.0000	0.0008		R287fs	1.00
SF3A1,22,30733026,30733026,G,A	5.3352	0.0000	0.0007		R699C	0.56
DUSP18,22,31059662,31059662,C,T	5.3349	0.0000	0.0007	rs202138261	R110H	1.00
RAD51B,14,68352609,68352609,G,A	5.3349	0.0000	0.0007	rs548280411	R159H	0.56
MRPS30,5,44811233,44811233,C,T	5.3344	0.0000	0.0007	rs201364888	R242*	1.00
SLC12A7,5,1064287,1064287,G,A	4.9346	0.0000	0.0011		R840C	1.00
OR2A5,7,143747859,143747859,G,A	4.3353	0.0000	0.0007	rs372476887	R122Q	0.78
PDE4DIP,1,144879312,144879312,G,A	4.3352	0.0000	0.0007	rs371331495	R1380W	0.56
APITD1-CORT,1,10511574,10511574,-,C	4.0169	0.0003	0.0056		A80fs	1.00
CEP250,20,34067191,34067191,C,T	3.8338	0.0001	0.0008	rs199810583	R744W	0.56
TBX10,11,67400532,67400532,C,T	3.7520	0.0001	0.0007	rs535008516	V198M	0.89
GDF5,20,34025551,34025551,-,G	3.7513	0.0001	0.0009		L53fs	1.00
RIN1,11,66100043,66100043,G,A	3.5774	0.0002	0.0022	rs2282532	P686S	0.56
RAD51B,14,68352608,68352608,C,T	3.3348	0.0001	0.0007	rs61755649	R159C	0.67
SLC6A19,5,1221267,1221267,T,G	3.1453	0.0002	0.0016	rs483352699	F514V	0.89
TP53INP2,20,33296585,33296585,-,C	3.0891	0.0001	0.0008		S14fs	1.00
EP300,22,41574637,41574637,C,T	3.0133	0.0001	0.0007	rs145312648	R2308C	0.78
MYO9B,19,17311582,17311582,C,T	2.7561	0.0001	0.0007		R1503C	1.00
NNT,5,43655960,43655960,G,A	2.7455	0.0002	0.0015	rs139987446	R693H	1.00
CNBD2,20,34560629,34560629,C,T	2.6349	0.0002	0.0015	rs150690141	R44W	0.67
ARHGEF5,7,144077001,144077001,A,G	2.5476	0.0001	0.0008		E1549G	0.67
GHR,5,42718765,42718765,C,T	2.5409	0.0001	0.0007	rs34853905	R364C	0.78
PC,11,66616566,66616566,G,A	2.3293	0.0003	0.0016	rs148492494	A1114V	1.00

EP300,22,41573050,41573050,T,G	2.1793	0.0011	0.0050		C1779G	1.00
PLEKHH1,14,68045912,68045912,G,A	2.0428	0.0006	0.0023	rs201225859	G971R	0.89
PDE4DIP,1,144866636,144866636,C,T	2.0135	0.0002	0.0007	rs367741522	R1869Q	0.56
SPAG4,20,34207652,34207652,G,T	1.9482	0.0003	0.0010	rs369926602	R354L	0.67
SGSM3,22,40800319,40800319,C,T	1.9127	0.0002	0.0009	rs138251871	R13W	0.56
FBN1,15,48704816,48704816,G,A	1.8760	0.0008	0.0030	rs61746008	R2726W	0.56
PDE4DIP,1,144866687,144866687,C,T	1.7908	0.0006	0.0022	rs139494606	R1852Q	0.56
PPA2,4,106320294,106320294,G,A	1.7468	0.0002	0.0007	rs138215926	P62L	0.89
NOBOX,7,144096940,144096940,C,T	1.7361	0.0002	0.0008	rs201947677	R355H	1.00
CLSTN1,1,9804590,9804590,T,C	1.4929	0.0003	0.0008	rs375488055	N356S	0.89
PDE4DIP,1,145015874,145015874,G,A	1.4282	0.0003	0.0007		R72*	1.00
TRIP13,5,916035,916035,A,G	1.0875	0.0003	0.0007	rs143798038	S384G	0.89
MVB12A,19,17535470,17535470,C,T	0.9719	0.0004	0.0008	rs143800574	A248V	0.56
TMEM134,11,67235051,67235051,G,A	0.9680	0.0066	0.0129	rs143199541	R84*	1.00
FAM83C,20,33876601,33876601,T,G	0.9654	0.0024	0.0047	rs200589769	H225P	0.56
CYP2D6,22,42524814,42524814,A,G	0.7631	0.0045	0.0076	rs199535154	L162P	0.78
NIM1K,5,43246067,43246067,G,C	0.6935	0.0005	0.0007	rs55663207	E64Q	0.78
MYH7B,20,33582133,33582133,C,T	0.6846	0.0029	0.0046	rs200371401	R919C	0.89
MYH7B,20,33575964,33575965,AT,-	0.6768	0.0009	0.0015	rs571047145	M538fs	1.00
DPP3,11,66249736,66249736,C,T	0.2753	0.0006	0.0008	rs142478050	A22V	0.67
HNF4G,8,76470800,76470800,C,T	0.2467	0.0006	0.0007	rs201625743	R251C	1.00
FCHO1,19,17886852,17886852,G,A	0.2403	0.0007	0.0008	rs199761608	R305H	0.56
CARNS1,11,67191572,67191572,C,T	0.2290	0.0020	0.0024	rs200939791	R662C	0.78
DNAJB7,22,41257815,41257815,G,A	0.1175	0.0027	0.0030	rs149771105	R62W	1.00
PDE4DIP,1,144857705,144857705,G,A	0.0125	0.0022	0.0022	rs146619065	R2117W	0.56

Appendix Table 8 List of variants falling in GWAS LD blocks of breast cancer associated SNPs. A list of 437 variants falls in this category, of wich 73 are also found somatically mutated in cancer and are hereby reported. For the complete list of filters used to obtain this table, refer to **Figure 22**.

Variant	log2FoldFreq	Case MAF	avsnp142	Protein Change	Deleteriousness Score
PIK3CB,3,138413709,138413709,-,G	25.2329	0.0039		R116fs	1.00
KMT2C,7,151945349,151945349,T,A	23.8270	0.0015	rs201039690	K724*	1.00
NBN,8,90990520,90990536,ATTGGACGTCCACAAAT,-	23.8270	0.0015		I166fs	1.00
HNF1A,12,121416792,121416792,C,T	23.4119	0.0011		T74M	0.78
ASXL1,20,31022331,31022331,C,T	23.2945	0.0010		R606W	0.56
POLE,12,133225574,133225574,G,A	23.0025	0.0008		R1364C	0.89
PTCH1,9,98211539,98211539,G,A	22.9761	0.0008		R1206C	0.89
AKT1,14,105246551,105246551,C,T	22.9666	0.0008	rs121434592	E17K	0.67
TET2,4,106164793,106164793,T,C	22.9502	0.0008		C1221R	0.56
NF2,22,30069387,30069387,C,T	22.9362	0.0008		R335C	0.78
MAP2K2,19,4117473,4117473,C,T	22.9201	0.0008		G83S	0.89
MSH6,2,48027887,48027887,G,A	22.8884	0.0008		R792Q	0.67
APC,5,112177788,112177788,G,A	22.8573	0.0008		R2148Q	0.78
FGFR1,8,38275843,38275843,G,A	22.8421	0.0008		R356W	1.00
DDB2,11,47256422,47256422,C,T	22.8399	0.0008		R273C	0.89
FMR1,X,147011711,147011711,G,A	22.8323	0.0007		R193H	0.67
ASPM,1,197073484,197073484,G,A	22.8313	0.0007	rs200202166	R1633C	0.56
CEP57,11,95546134,95546134,C,T	22.8291	0.0007	rs387906977	R81*	1.00
FANCD2,3,10133905,10133905,G,A	22.8291	0.0007		R1273Q	0.56
ASPM,1,197073381,197073381,C,T	22.8270	0.0007		R1667H	1.00
NOX4,11,89088203,89088203,G,A	22.8270	0.0007	rs374112961	R357*	1.00
SETBP1,18,42532994,42532994,C,T	22.8270	0.0007		T1230I	0.56
FAT1,4,187549401,187549401,C,G	22.8270	0.0007	rs138797966	E1573Q	0.78
JAK2,9,5069154,5069154,C,T	22.8270	0.0007		R487C	0.89
DHX9,1,182841496,182841496,C,T	22.8248	0.0007		R528C	0.67
DHX9,1,182841497,182841497,G,A	22.8248	0.0007		R528H	0.67
ATM,11,108235818,108235818,T,C	22.8248	0.0007	rs371619067	Y2954H	0.56
KCNJ5,11,128781800,128781800,G,A	22.8248	0.0007		R211Q	1.00
ABCC11,16,48234267,48234267,C,T	22.8248	0.0007	rs200200325	V668M	0.89
TP53,17,7578407,7578407,G,A	22.8248	0.0007	rs138729528	R43C	1.00
TRIM37,17,57165733,57165733,C,T	22.8248	0.0007	rs201317687	R67H	0.56
EPB41L3,18,5396207,5396207,G,A	22.8248	0.0007	rs138017302	S767L	0.78
POLQ,3,121195392,121195392,C,T	22.8248	0.0007		A2134T	0.78
FAT1,4,187584680,187584680,G,C	22.8248	0.0007		P1118R	0.56
FBN2,5,127641568,127641568,C,T	22.8248	0.0007	rs140276399	R1832H	0.67
MYB,6,135539105,135539105,C,T	22.8248	0.0007		T552M	0.67
KMT2C,7,151945051,151945051,A,G	22.8248	0.0007	rs2838171	I823T	0.78

Appendix Table 9 List of variants that overlap with cancer somatic mutations. From a total of 185 overlapping variants on candidate driver genes, we reported all the monomorphic sites in the ExAC database (MAF in the controls = 0).

Gene Symbol	REF count	ALT count	Exac REF count	Exac ALT count	Number of truncating events	Stouffer Compoud Qvalue
ALK	1093	189	55873	809	1	1.13E-115
HNF1B	1204	142	64362	1052	1	9.89E-64
CNOT3	1216	66	63843	184	1	1.09E-52
MSH6	1270	34	64712	138	1	1.22E-22
HNF1A	688	22	32536	88	1	2.94E-14
ASXL1	341	13	43003	85	1	2.48E-11
SMOX	1174	8	61840	47	1	1.09E-04
CRIPAK	28672	521	1490842	3057	27	1.16E-04
FLCN	854	6	62572	34	1	1.61E-04
ANAPC1	2476	22	105814	269	2	5.67E-04
FGFR3	1244	4	64820	13	1	2.07E-03
AHNAK2	3865	5	193999	10	3	2.50E-03
PIK3CB	1263	5	65030	29	1	3.15E-03
KMT2C	6174	134	296450	5956	5	7.93E-03
JAG1	1314	2	64504	2	1	1.43E-02
CHEK2	1342	2	65439	3	1	2.23E-02
POLR1A	1162	24	58860	664	1	2.79E-02
FANCM	2627	3	130509	99	2	4.59E-02
SPRY4	351	1	61286	1	1	4.66E-02
HLA-B	1050	6	87400	2338	2	4.66E-02
COL18A1	278	105	44002	33065	2	4.66E-02
IL32	141	1	10067	0	1	4.66E-02
CDC27	1193	1	63698	0	1	4.66E-02
BBS10	1249	1	64820	0	1	4.66E-02
PARP1	1273	1	65248	0	1	4.66E-02
IKBKB	1279	1	65106	0	1	4.66E-02
FBN2	1295	1	65443	0	1	4.66E-02
BRCA2	1315	1	64690	0	1	4.66E-02
ERCC6	1331	1	65474	0	1	4.66E-02
UROD	1339	1	65431	0	1	4.66E-02
CEP57	1341	1	65470	0	1	4.66E-02
RNASEL	1343	1	65468	0	1	4.66E-02
BRCA1	1343	1	65444	0	1	4.66E-02
POLN	1339	1	65234	0	1	4.66E-02
ATR	1345	1	65482	0	1	4.66E-02
SIN3A	1345	1	65412	0	1	4.66E-02
PTPRB	1345	1	65360	0	1	4.66E-02
TYR	1343	1	65176	0	1	4.66E-02
MSH4	1343	1	65122	0	1	4.66E-02
NLRP3	1345	1	65138	0	1	4.66E-02
NOX4	1343	1	59130	0	1	4.99E-02

Appendix Table 10 List of candidate loss-of-function genes. We report all those genes with an excess of truncating events in cases compared to controls with a compound q-value < 0.05.