



UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Università degli Studi di Padova

Department of Biology

Ph.D. COURSE IN: BIOSCIENCES

CURRICULUM: GENETICS, GENOMICS AND BIOINFORMATICS

SERIES XXXI

**THE GENOMIC LANDSCAPE OF SOLID AND HEMATOLOGIC MALIGNANCIES CHARACTERIZED BY
NEW BIOINFORMATICS TOOLS**

Coordinator: Ch.mo Prof. Ildikò Szabo

Supervisor: Ch.mo Prof. Stefania Bortoluzzi

Co-Supervisor: Dr. Alessandro Coppe

Ph.D. student: Andrea Binatti

Summary

1. RIASSUNTO	3
2. ABSTRACT	5
3. INTRODUCTION	7
3.1. GENETIC AND GENOMIC ALTERATIONS IN CANCER	7
3.2. PERSONALIZED MEDICINE FOR CANCER TREATMENT.....	11
3.3. DNA SEQUENCING	12
3.4. NEXT-GENERATION SEQUENCING	14
<i>Illumina</i>	14
<i>IonTorrent</i>	16
3.5. WHOLE EXOME SEQUENCING	19
3.6. SOMATIC VARIANTS DETECTION BY WES	20
<i>WES raw data</i>	20
<i>Short reads to the reference genome alignment</i>	22
<i>Post-alignment read processing</i>	23
<i>Somatic variant calling</i>	24
3.7. VARIANT ANNOTATION AND PRIORITIZATION	28
<i>Annotation of somatic variants and prediction of functional impact</i>	29
<i>Somatic variant filters based on population allele frequencies</i>	29
<i>Additional criteria for variant prioritization</i>	30
<i>Recurrence as indicator of importance?</i>	31
<i>Pathways and interaction networks topological analysis</i>	32
4. AIMS OF THE STUDY	37
5. MATERIALS AND METHODS	38
5.1. PROGRAMMING LANGUAGES.....	38
5.2. DATABASES AND WEB TOOLS FOR SOMATIC VARIANT PRIORITIZATION	39
5.3. THIRD-PARTY SOFTWARE TOOLS.....	41
5.4. BIOINFORMATICS TOOLS FOR PATHWAY-DERIVED NETWORK CONSTRUCTION AND ANALYSIS.....	46
5.5. LARGE GRANULAR LYMPHOCYTE LEUKEMIA.....	47
<i>WES data</i>	47
<i>Identification of somatic variants from LGL-L tumors</i>	49
<i>Pathway-derived gene meta-networks construction</i>	50
<i>Validations of selected somatic variants</i>	50
5.6. FOLLICULAR LYMPHOMA OF THE PEDIATRIC AGE.....	50
<i>WES data</i>	50
<i>Somatic variants identification in PTNFL and PFLT patients</i>	53

<i>Mutated gene network analysis</i>	54
<i>Validations of somatic variants and TNFRSF14 exon1 investigation</i>	54
5.7. HIGH-RISK NEUROBLASTOMA.....	55
<i>WES data</i>	55
<i>Identification of somatic variants from HR-NB patients</i>	57
<i>Pathway analysis</i>	58
<i>Gene network analysis</i>	58
<i>Ultra-deep sequencing for variants validation</i>	59
<i>Independent cohort of HR-NB</i>	59
6. RESULTS	60
6.1. iWHALE: A COMPUTATIONAL PIPELINE FOR CANCER EXOME SEQUENCING ANALYSES.....	60
6.2. PATHWAY-DERIVED META-NETWORK OF MUTATED GENES.....	61
6.3. GENOMIC CHARACTERIZATION OF LGL-L PATIENTS.....	65
<i>Detected somatic variants</i>	65
<i>Beyond the mutated STAT genes</i>	67
<i>Novel mutated functional modules in STAT-negative patients</i>	68
<i>Confirmed somatic variants</i>	70
<i>High mutational burden in CD4-positive T-LGLL patients</i>	72
<i>Mutations in the network for each LGL-L subgroup</i>	73
<i>Discussion</i>	75
6.4. NEW GENES AND PATHWAYS MUTATED IN FOLLICULAR LYMPHOMA OF THE PEDIATRIC AGE PATIENTS.....	77
<i>Recurrent somatically mutated genes in pediatric FLs</i>	77
<i>Exon 1 of TNFRSF14 is a mutational hotspot in PTNFL</i>	80
<i>Pathways recurrently altered in FL of the pediatric age</i>	80
<i>Highly prioritized and confirmed somatic variants</i>	82
<i>Discussion</i>	85
6.5. GENES AND PATHWAYS LINKED TO HR-NB AGGRESSIVENESS.....	88
<i>Selected group-specific somatic variants</i>	88
<i>SS and LS patients have similar mutational landscapes</i>	92
<i>Mutations exclusive of SS or LS</i>	93
<i>Different pathways are altered in HR-NB subgroups</i>	95
<i>Only mutations of SS patients tend to cluster into specific subnetworks</i>	98
<i>Discussion</i>	100
7. CONCLUSIONS AND FUTURE PERSPECTIVES	105
8. REFERENCES	107

1. Riassunto

Il sequenziamento dell'esoma (WES) rileva efficacemente varianti in cellule tumorali, identificando le caratteristiche molecolari coinvolte nella patogenesi e nella progressione della malattia, con importanti risvolti per la diagnosi e per lo sviluppo e la scelta di terapie personalizzate. L'analisi di dati WES di tumori presenta tuttavia varie complicazioni dovute all'eterogeneità tumorale, ad alterazioni della ploidia, a contaminazioni dei campioni o ad artefatti tecnici.

La pipeline iWhale, basata su Docker e SCons, è stata sviluppata per analizzare dati WES di tumori con l'obiettivo di rilevare ed annotare mutazioni somatiche tramite l'uso di quattro diversi software (MuTect, MuTect2, Strelka2 e VarScan2) e l'integrazione di informazioni provenienti da vari database. Inoltre, ho collaborato allo sviluppo di un metodo per la costruzione di meta-reti di geni mutati che sono annotati in database di pathway e ho costruito una struttura di dati customizzata per rilevare statisticamente pathway ricorrentemente mutati in cellule tumorali.

In collaborazione con diversi gruppi di ricerca, ho utilizzato ed adattato di volta in volta versioni progressivamente più rifinite della mia pipeline in studi riguardanti la leucemia linfocitica granulosa a grandi cellule T (LGL-L), due tipi di linfomi follicolari pediatrici (PTNFL e PFLT) e Neuroblastoma ad alto rischio (HR-NB).

LGL-L è una leucemia cronica rara caratterizzata da una persistente crescita clonale di cellule citotossiche T o natural killer (NK) dovuta all'attivazione del pathway JAK/STAT. Mediante analisi WES sono state identificate nuove mutazioni somatiche in geni ricorrentemente mutati in 19 pazienti con LGL-L, comprendenti casi senza mutazioni nei geni STAT. Sono state selezionate per validazione con sequenziamento Sanger 16 varianti in diversi geni, tra le quali l'oncosoppressore *FAT4* e il regolatore epigenetico *KMT2D*. Nuove varianti Q706L e S715F in *STAT5B* sono state anche caratterizzate funzionalmente. Grazie ad analisi di reti derivate da pathway, sono state identificate delle componenti funzionali composte da geni mutati, funzionalmente o direttamente interagenti con i geni STAT, in pazienti STAT negativi. Altre componenti funzionali con una possibile rilevanza nella patogenesi di LGL-L in assenza di mutazioni nei geni STAT sono emerse dalle analisi.

Una coorte di pazienti affetti da linfomi follicolari pediatrici è stata analizzata tramite WES. Sono state confermate mutazioni presenti in *TNFRSF14*, *IRF8* e *MAP2K1*, geni precedentemente associati a PTNFL, e sono stati caratterizzati nuove mutazioni e geni con possibile coinvolgimento nello sviluppo di PTNFL. Undici varianti presenti in *ARHGEF1*,

MAP2K1, TNFRSF14, ATG7, GNA13, RSF1, UBAP2 e ZNF608 sono state validate e selezionate come possibili eventi driver in PTNFL e PFLT. I nostri risultati hanno per la prima volta permesso di associare il pathway GPCR ed enzimi modificatori della cromatina ai linfomi follicolari pediatrici.

NB è un tumore solido che origina dalle cellule della cresta neurale primitiva ed è caratterizzato da un'alta eterogeneità clinica e da pochi geni ricorrentemente mutati (*MYCN, ALK, ATRX*). Per investigare sulle basi biologiche coinvolte nell'aggressività di NB, è stato effettuato WES di pazienti affetti da HR-NB con metastasi e divisi in base alla sopravvivenza (pazienti SS e LS, rispettivamente con sopravvivenza inferiore o uguale e superiore a 5 anni). Solo i geni *SMARCA4, SMO, ZNF44* e *CHD2* sono stati trovati mutati ricorrentemente in modo specifico in pazienti SS. HotNet2 ha rivelato che le mutazioni rilevate nei due gruppi ricadevano in pathway diversi. Le mutazioni dei pazienti SS si sono raggruppate in sei sotto-reti significativamente mutate, coinvolte nell'organizzazione della matrice extracellulare tramite MAPK pathway, nella motilità cellulare tramite PTK2, nell'attività delle metalloproteinasi della matrice, nella maturazione del centrosoma e nel rimodellamento dei cromosomi. Grazie all'esistenza di farmaci già approvati dalla FDA che hanno come bersaglio alcune delle proteine mutate o delle pathway identificate, i risultati ottenuti possono facilitare lo sviluppo di terapie mirate ai pazienti con le forme più aggressive di HR-NB.

2. Abstract

Whole Exome Sequencing (WES) has high power to discover variants in cancer cells, allowing the identification of molecular features underlying diseases development and progression, with important outcomes for cancer diagnosis/prognostication as well as for development and selection of molecularly targeted therapies in personalized medicine. WES projects pose as well different challenges due to biological factors, such as tumour heterogeneity, altered ploidy, low tumor purity, and technical artifacts, that make not obvious the identification of relevant variants.

IWhale, an easy-to-use and customizable pipeline based on Docker and SCons, was developed to analyze cancer WES data, to detect and annotate somatic mutations by a combination of four different callers and integration of information deriving from different databases. Moreover, a systems genetics approach and custom data structures were built up to construct pathway-derived meta-networks of mutated genes depicting their direct interactions and functional relations, to ultimately identify key functions and pathways recurrently hit in cancer cells.

In collaboration with different groups, increasingly refined and customized versions of the pipeline were applied in three WES studies regarding Large granular lymphocyte leukemia (LGL-L), pediatric follicular lymphomas (PTNFL and PFLT) and High-Risk Neuroblastoma (HR-NB).

LGL-L is a rare chronic leukemia with persistent clonal increase of cytotoxic T cells or natural killer (NK) cells often associated to JAK/STAT pathway activation. By analysis of WES data in 19 patients, including cases without STAT mutations (STAT- patients), novel somatic mutations in recurrently mutated genes were identified. 16 selected variants, including those in the tumor suppressor gene *FAT4* and in the epigenetic regulator *KMT2D*, were validated. The new Q706L and S715F *STAT5B* variants has been also functionally characterized. With pathway-derived network analysis, functional modules composed by several STAT-interacting or STAT-functional connected genes mutated in STAT-negative patients were discovered. Additional modules with putative pathogenic relevance in LGL-L and mutated in the absence of STAT mutations were identified.

In PTNFL, recently recognized as a defined clinicopathological entity, WES analysis of the largest cohort collected so far uncovered mutations in the few genes, *TNFRSF14*, *IRF8* and *MAP2K1* previously associated to PTNFL, identifying as well novel mutations and genes. Eleven validated variants prioritized as possible drivers hit the recurrently mutated *ARHGEF1*, *MAP2K1* and *TNFRSF14* genes, as well as *ATG7*, *GNAI3*, *RSF1*, *UBAP2*, and *ZNF608*. G-

protein coupled receptor signaling and chromatin modifying enzyme alterations was linked for the first time to PTNFL and PFLT according to obtained findings.

NB, a solid cancer arising from primitive neural crest cells and accounting for 9% of pediatric tumors, is characterized by high clinical heterogeneity and low mutation recurrence even in known driver (*MYCN*, *ALK*, *ATRX*). To clarify the biological basis of disease aggressiveness, WES was used to examine the genomic landscape of HR-NB patients at metastatic stage with short survival (SS) and long survival (LS). A few genes, including *SMARCA4*, *SMO*, *ZNF44* and *CHD2*, were recurrently mutated only in the SS group and HotNet2 analysis revealed that in the two patient groups, mutations occurred in different pathways. Notably mutations of SS patients clustered into a six significantly mutated subnetworks, involved into MAPK pathway associated with the organization of the extracellular matrix, to cell motility through PTK2 signaling, to matrix metalloproteinase activity, to centrosome maturation and chromosome remodeling, to metabolism of nucleotides and lipoproteins, and to transport of small molecules. Since FDA-approved compounds targeting the deregulated pathways are available these findings may help to improve the treatment of HR-NB patients with most aggressive disease.

3. Introduction

3.1. Genetic and genomic alterations in cancer

Cancer is a complex and heterogeneous genetic disease which develops, broadly speaking, when a group of cells start to uncontrollably proliferate invading an organ or tissue. Cancer cells can also spread from original tumor site and colonize other parts of the body in a process called metastasis¹.

After cardiovascular diseases, cancer is the second leading cause of death in the world. The incidence is rising in developed countries because of the aging of the population and to an increasing exposition of environmental risk factors². About 90.5 million people had cancer in 2015 and cancer burden worldwide is projected to redouble within the next two decades^{3,4}. It is established that the 90-95% of cancer cases are caused by DNA mutations directly or indirectly caused by environmental factors such as tobacco, diet, obesity, infections, radiation, stress or lack of physical activity whereas only the 5-10% of cases are induced by hereditary genetic factors^{5,6}.

The shift from normal to cancer phenotype is mainly due to an accumulation of somatic mutations over time that modify essential cellular functions, with contributions from epigenetic and transcriptional alterations. Only a minority of somatic mutations harbored by malignant cells are “driver” conferring selective advantages and leading to the development of typical features of malignancy such as sustaining proliferative signaling, evading growth suppressors, avoiding immune destruction, enabling replicative immortality, promoting inflammation, activating metastasis, inducing angiogenesis, generating genome instability and mutation, resisting cell death and deregulating cellular energetics^{7,8}. Most of the mutations carried by cancer cells are “passenger” having a little or no impact on cancer expansion.

The accumulation of somatic variants over time can lead to formation of different cancer clones that are subject to selective pressure induced by tissue ecosystems (**Figure 1**) and resulting in, both intra-tumor and longitudinal, cancer heterogeneity.

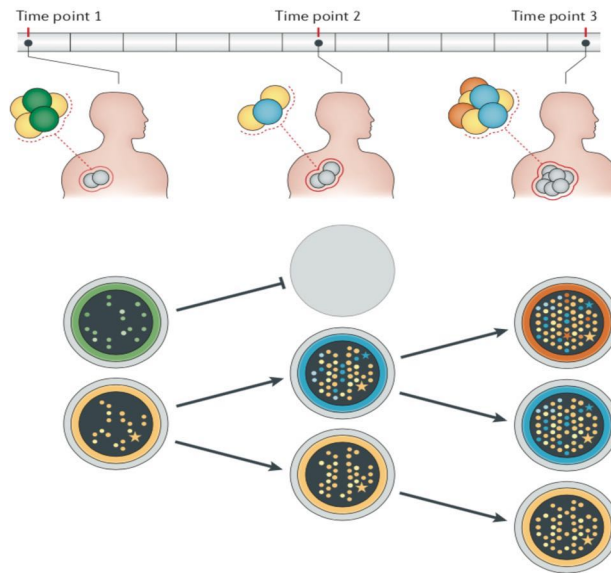


Figure 1. A conceptual example of clonal evolution model in cancer. *The founding clone indicated in yellow persists during disease progression, whereas the green clone extinguishes before time point 2. The blue clone emerges from yellow clone thanks to accumulation of somatic variants at time point 2. In the same way, another cancer clone emerges (orange) from blue clone and the founding clone survives also at time point 3. The tumor mass exhibits a heterogeneous profile made of small number of driver mutations and many passenger mutations⁹.*

Clones in cancer compete for space as well as resources, such as oxygen and glucose. Cells with somatic mutations that increase their fitness will proliferate more than non-mutated cells favoring their expansion and tissue invasion. For instance, in colorectal cancer the first mutations often affect *APC* gene, resulting in a slowly growing adenoma, but a second mutation in *KRAS* induces a second round of clonal growth allowing an expansion of the cellular mass. This process of mutation followed by clonal expansion continues with mutation in *PIK3CA* and *TP53* generating a malignant tumor that can invade through the underlying basement membrane and metastasize to lymph node and distant organs¹⁰.

In truth, the accumulation of somatic variants is thought to occur in a context of individual-specific germline variants that can predispose to cancer. The most known and extreme example is represented by the germline variants in *BRCA1* and *BRCA2* in hereditary breast ovarian cancer syndrome^{11,12}. Germline variants can behave as pre-existing driver events, that together with later acquired somatic collectively lead to cancer development. Alfred Knudson hypothesized that one germline and one somatic variant are required to affect both alleles of a tumor suppressor gene and resulting in cancer development¹³. In addition to gene level, germline variants can predispose to cancer on the basis of their co-occurrence in different genes participating the same pathway. Although germline variants may have a functional impact on typical cancer hallmarks, acquired somatic variants are fundamental for malignant transformation. Germline variants may remain silent because of compensatory pathways in

normal tissues and lack of alteration in other cancer-associated pathways until somatic mutations impair the compensatory or activate oncogenic pathways¹⁴.

Somatic and germline mutations that affect cancer cells can be of different types:

- single-nucleotide variants (SNVs),
- insertions or deletions of nucleotides (indels),
- copy-number variations (CNVs),
- large structural variants (SVs).

When one of these mutations hits coding-regions, oncogenes can be activated promoting cell proliferation, or inhibit tumor suppressor allowing cancer cells to avoid the cellular survival and division control systems. In general, somatic SNV or indels in oncogenes tend to target a specific region of the gene, whereas, tumor suppressor genes are mutated throughout their entire length (**Figure 2**)^{10,15}.

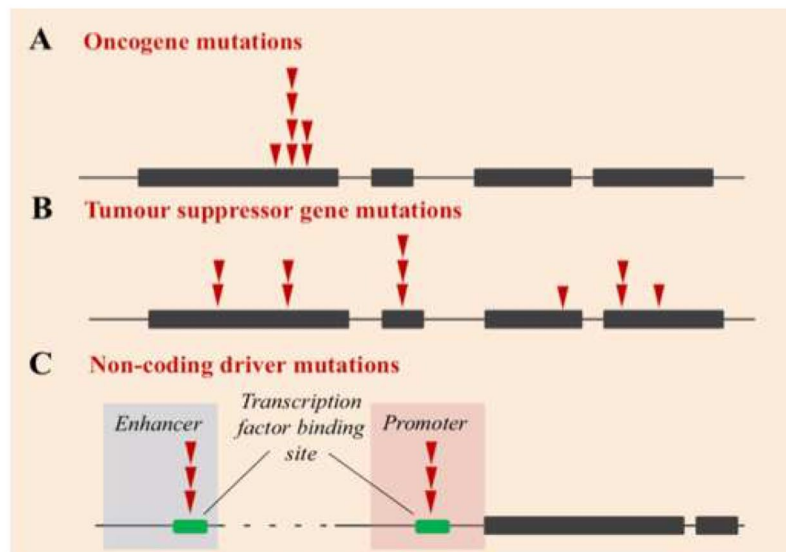


Figure 2. Diagram depicting distribution of different types of driver variants. Bars represent exons, and red triangles depict an example pattern of somatic mutations across a cancer cohort. A) Activating mutations hitting oncogenes are confined to a specific region of amino acids; B) Inactivating mutations usually affect tumor suppressor genes arising throughout the gene; C) Non-coding driver mutations may target cis-regulatory elements, for example, removing a transcription factor binding site, or create a de novo binding motif. Figure adapted from Poulos RC, et al., 2018¹⁵.

Vogelstein et. al¹⁰ proposed the “20/20 rule” to determine which genes are oncogenic or tumor suppressors evaluating mutations recorded in The Catalogue of Somatic Mutations in Cancer (COSMIC). An oncogene must have more than 20% of somatic missense mutations in recurrent positions, whereas tumor suppressor genes require that more than 20% of hitting mutations must be inactivating (truncating or frameshift variants). Thanks to this method, *IDH1* gene, which was considered a tumor suppressor gene, has been classified as oncogene since almost

the mutations hit a substrate binding site at codon 132. The oncogenic role of *IDHI* was also supported by biochemical experiments^{16,17}.

Moreover, mutations in functional non-coding regions of the genome, can be associated to cancer development. Variants in introns can alter splicing or induce the loss of regulatory elements. For instance, a rare germline mutation in an intron of *BRC A2* induces aberrant splicing and is associated to Fanconi anemia, a recessive disease linked to high cancer risk¹⁸. Mutations can also impact on transcription factor binding, as exemplified by the *TERT* gene promoter, mutated in more 50 cancer types¹⁹. Single-nucleotide mutations in *TERT* promoter creating new transcription factor binding sites and upregulating *TERT* expression were first reported in melanoma^{20,21}. Mutations in non-coding RNA (ncRNA) may have a driver role, impacting on ncRNA fielding, interactions and functions. *MALAT1*, a long non-coding RNA that regulates expression of genes associated to metastasis, is mutated in bladder cancer²².

Copy number alterations (CNAs) can drive cancer by duplicating oncogenes or deleting onco-suppressors. The CNA-associated oncogenes and tumor suppressors comprise the focal amplification of 8q24.21 (*MYC*), 11q13.3 (*CCND1*), 7p11.2 (*EGFR*), 17q12 (*ERBB2/HER2*) and 7q31.2 (*MET*), whereas focal deletion involved 13q14.2 (*RBI*), 9p21.3 (*CDKN2A*) and 10q23.31 (*PTEN*)^{23,24}.

Similarly, large structural mutations may lead to aberrant gene fusion events that create a new oncogenic protein or truncate a tumor suppressor gene²⁵. Specific structural variants in leukemia and sarcoma drive to activation of oncogenes or generation of specific gene fusion, some of which are used in diagnosis, such as *STY-SSX1* fusion in synovial sarcoma and *EWS-FLII* fusion in Ewing sarcoma²⁶. In prostate cancer, an aberrant fusion between 5'UTR of *TMPRSS2* and ETS family genes (*ERG* and *ETVI*) is frequently detected²⁷. This event causes an *ERG* overexpression, which disrupts androgen receptor (AR) signaling by inhibiting the expression of AR and its target genes²⁸.

The number of somatic mutations needed for cancer development is variable, depending on several factors. For instance, lung tumors and melanomas usually harbor high number of nonsynonymous variants per tumor (~200) because of protracted exposition to mutagens, such as tobacco smoke or ultraviolet radiation, and/or DNA repair defects¹⁰. Genetic alterations of the proofreading domain of DNA polymerases *POLE* or *POLD1* has been also associated with high number of somatic mutations in cancer²⁹⁻³¹. The number of somatic mutations in tumors of self-renewing tissues tends also to increase with age³². It has been showed that more than half of somatic mutations in these tumors emerge during the self-renewal of normal cells, for instance the gastrointestinal epithelium, and do not directly contribute to malignancy

development. This observation partially explains why tumors of non-self-renewing tissues and pediatric cancers harbor few mutations. Indeed, pediatric cancers usually develop in non-self-renewing tissues and when they arise in renewing tissues (leukemias) emerge from precursors that have not renovated themselves as often as in adults¹⁰.

3.2. Personalized medicine for cancer treatment

Hundreds of driver genes, such as *TP53*, *BRAF*, *EGFR*, *PIK3CA* to cite the commonest, have been identified and gathered in databases (Cancer Gene Census³³ and IntOGen³⁴). Pharmacological therapies have been developed which target some of these genes inhibiting cancer growth and invasion. The tyrosine-kinase inhibitor Imatinib has been successfully used to target cells carrying the *BCR-ABL* fusion gene in chronic myeloid leukemia³⁵. The protein coded by *EGFR* hit by activating mutations can be inhibited by gefitinib in lung adenocarcinoma³⁶. However, highly recurrently mutations targetable by specific drug therapy are unknown for the most cancer types. In addition, the clonal expansion of cancer considerably complicates the scenario for therapeutic treatments. Indeed, therapies induce an additional selective pressure where sensitive cancer cells will die, but resistant cells might survive, possibly acquiring new driver mutations not present in primary tumors as demonstrated in breast cancer^{37,38}, and lead to tumor relapse.

Other targeted therapeutic strategies are immune-based therapies where the aim is augmenting the patient immune response against cancer cells, with a general boosting of immune defenses (Interferon or Interleukins), or using endogenous immune cells expanded (Tumor-infiltrating lymphocytes) or genetically engineered. Checkpoint inhibition and cellular therapy with autologous chimeric antigen receptor T cells (CAR T cells) have shown efficacy as molecularly targeted salvage therapy in several cancer types (small cell lung cancer³⁹, advanced melanoma⁴⁰, renal cancer⁴¹, and acute lymphocytic leukemia⁴²) and are promising approaches, especially as they will be refined, to reduce serious side effects, and will be used in combination with other conventional therapies⁴³. However, it has been reported that tumor immunogenicity differs greatly between different types of cancers and cancers of the same type in different patients⁴⁴. Consequently, it will be crucial to develop therapeutic strategies where the reactivity of T-cells is selectively enhanced against tumor-specific clonal neoantigens⁴⁵.

From above explained approaches is evident that is necessary to design therapeutic strategies considering the heterogeneity observed in cancer. For this reason, it is increasingly taking place the approach of personalized medicine where individual characteristics, including the

mutations profile, environment, and lifestyle, are exploited to develop new therapeutic strategies maximizing efficacy and minimizing toxicity (**Figure 3**).

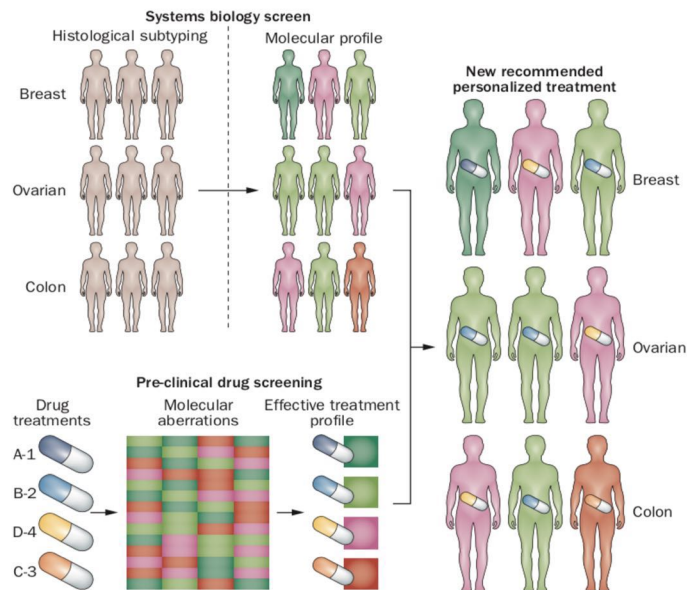


Figure 3. Tumor mutation profile for personalized therapy. *The analysis of mutation profile for cancer patients integrating tumor driver and pharmacogenomics data is promising to effectively predict personalized therapeutic options⁴⁶.*

The identification of specific genomic subsets of particular tumors already allows the association to prognostic characteristics that can be translated into therapy reduction or intensification. For example, medulloblastoma patients affected by WNT pathway alterations have a better prognosis and less intensive therapy can be applied^{47,48}.

The search of driver mutations in cancer genomes and variants influencing pharmacogenomics⁴⁹ are crucial to move toward a personalized medicine in cancer. Thus, an increase of knowledge about cancer biology and genome is needed to develop new effective therapeutic strategies applicable in the most types of cancer.

3.3. DNA sequencing

Sequencing consists in reading the exact sequence of nucleotides of a nucleic acid. Already in 1977, two fundamental works concerning DNA sequencing were published. Allan Maxam and Walter Gilbert described a method in which DNA fragments marked in terminal position were subjected to basis-specific chemical cleavage and the products of the reaction were separated by gel electrophoresis⁵⁰. Sanger and Coulson described an enzymatic approach (Sanger sequencing) which uses DNA polymerase to extend a short synthetic oligonucleotide primer hybridized to a single-stranded DNA template^{51,52}. Sanger sequencing has been progressively

improved with various refinements such as: 1) nucleotide-specific fluorescent dyes, so that the chain-termination reaction for the four dideoxynucleotides to be carried out in a single DNA synthesis reaction⁵³; 2) polyacrylamide gels in capillary electrophoresis, to shorten run times⁵⁴; 3) automatic laser fluorescence detection⁵⁵. Thanks to these gradual improvements, the Sanger sequencing can sequence fragments up to 1000 bases with a per-base accuracy as high as 99.999%⁵⁶. The refinement and improvements made to the Sanger sequencing allowed its utilization on large scale and the realization of Human Project Genome in 2001⁵⁷. However, this method is affected by high costs and time required, as well as low throughput. One more limitation of Sanger sequencing is the impossibility to detect variants with low frequency, such as in cancer samples, due to high background levels.

The advent of Next-Generation Sequencing (NGS) technologies on the marketplace in 2004 was revolutionary for basic, applied and clinical research because the amount of data produced by each sequencing run was considerably increased with a substantial reduction in costs and time of execution (from hundreds to billions of pair of bases for sequencing run). The term NGS refers to a series of technologies that have in common the parallel and massive sequencing of clonally amplified and spatially separated DNA molecules on a solid support. The sequencing phase consists of repeated cycles of nucleotide extensions by a DNA polymerase or alternatively by oligonucleotide ligation cycles. A first advantage of NGS is the use of clonal amplification (PCR) of the fragments to sequence avoiding the cloning step into plasmids. In addition, different samples can be processed in a unique sequencing run through a method for marking the templates (barcoding), requiring informatic analysis to identify the fragments belonging to each sample.

NGS technologies can accurately detect alleles at low frequency, but they have some limitations as well: the first is short length of produced fragments leading to major issues to address during alignment phase to a reference genome, and even more for de novo assembly of genome and transcriptome, requiring complex data analysis. In addition, error frequency in base calling is higher than Sanger sequencing, even if this limitation is partially solved through large number of sequences and utilization of highly-efficient DNA polymerase. “Paired-end sequencing” strategy, where both ends of the fragments are sequenced yielding two paired reads at a known distance, is used to even map reads over repetitive regions of genome improving detection of small indels. Another issue of NGS is the high computational request due to huge amount of data to storage and process⁵⁸. Complex bioinformatics tools are essential to improve base calling and to perform every phase of subsequent data analysis. Each sequencing technology uses platform-specific parameters and scores making complex to

compare results obtained from different sequencers. This complexity in data interpretation and management of a computer system requires the presence of high-specialized personnel.

3.4. Next-generation sequencing

The various technologies differ from each other based on the combinations of biochemical processes and protocols followed to perform four basic common steps: library preparation, sequencing, imaging or signal processing, and data analysis. Library preparation is accomplished by random fragmentation through sonication or nebulization (“Shotgun method”) into a target size (from 150 to many hundred base pairs) depending on the platform read length and chemistry. Specific nucleotide sequences (“adaptors”) are, then, ligated at both ends of each fragment to hybridize them to solid surfaces covered by adapter-complementary oligonucleotide anchors. Adaptors can contain sample-specific sequence (“barcode”) allowing simultaneous sequencing of more samples per run dividing platform throughput between the total number of samples. The anchored templates are amplified by “emulsion PCR” or “bridge PCR” generating clusters of identical fragments to obtain a detectable signal. Signal detection is performed on all DNA fragments cyclically and in parallel by an optical system composed by a microscope with a Charge-Coupled-Device (CCD) camera plus a computer and storage system (*Illumina*), or, in the case of semiconductor sequencing, by a semiconductor chip (*IonTorrent*)⁵⁹.

Illumina

Illumina sequencing technology was first put on the market in 2006 through the production of Genome Analyzer by the Solexa company, which was purchased a year later by Illumina. Library preparation is accomplished by random fragmentation of DNA sample into segments of some hundreds of bases which are modified to generate 5’ phosphorylated ends. An adenine is added to 3’ ends of fragments for improving ligation process with adaptor sequences having a thymine at 3’ ends (**Figure 2**). The adaptor sequences are perfectly complementary to the anchoring oligonucleotides fixed on the flow cell, a planar optically transparent surface similar to a microscope slide (**Figure 4**).

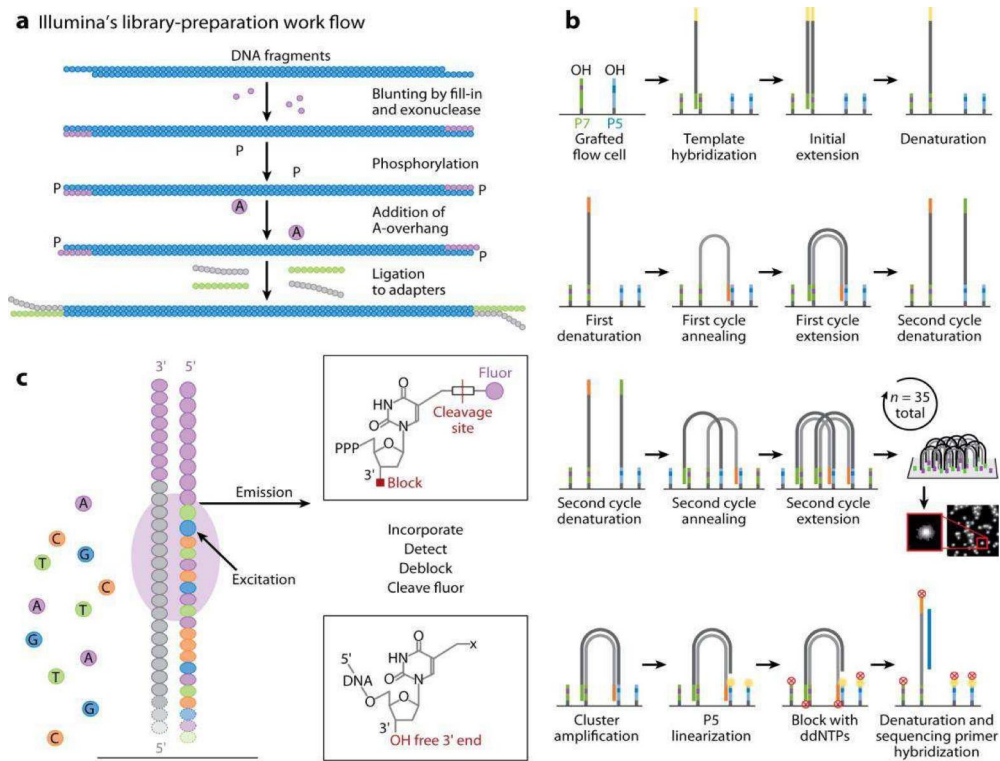


Figure 4. Illumina sequencing workflow. a) Library preparation. b) Clusters of templates are generated by bridge amplification. c) Sequencing by synthesis with reversible dye terminators⁶⁰.

Oligonucleotides of the flow cell are linked to the surface at its 5' end, leaving the 3' end free for the polymerase action. DNA templates are amplified by “bridge PCR” which is based on the folding of the arch-shaped DNA strands to hybridize to an adjacent anchoring-oligonucleotide and complementary to the adapter present at the free end of the filament itself. The amplification occurs through DNA polymerase that synthesizes the complementary filaments to those present. The double-stranded structures are then denatured (chemically or thermally) giving single-stranded filaments ready for another cycle of amplification. This process is repeated many times to obtain, for each initial DNA template, a cluster composed of approximately one million of clonal amplicons (**Figure 4**). Bridge-PCR can produce 100-200 million spatially separated template clusters, providing free ends to which a universal sequencing primer can be hybridized to initiate the sequencing reaction⁵⁸. Illumina uses a chemistry called “Cyclic Reversible Termination” (CRT), which sequences the template strand one nucleotide at a time through progressive rounds of base incorporation, washing, imaging, and cleavage⁶¹. Each cycle involves a DNA polymerase and 4 deoxynucleotides modified with the incorporation of a fluorescent marker and the addition of a reversible terminator. The chemical terminator group (“3'-O-azidomethyl group”) plays a key role in CRT sequencing,

blocking the hydroxyl present in the 3' position of the nucleotide and preventing the incorporation of the next nucleotide⁶² (**Figure 4**).

In this way, only a single nucleotide is added for each filament in each cycle. The process is defined reversible because, after the removal of the terminator group and the restoration of the oxidant in the 3' position, the polymerization continues. The 3'-O-azidomethyl nucleotides are incorporated based on the complementarity of the sequence of each strand, within each clonal cluster. The presence of blocking group allows a synchronized process and the remaining unincorporated bases are washed away. The Illumina technology can yield paired-end data sequencing both ends of fragments for each DNA clusters. After the first round of sequencing, the single stranded flow-cell bound DNA undergo again bridge amplification, but this time forward strand is washed away, leaving clusters of the reverse strand, which can be sequenced as before. Imaging is then performed by two lasers that interrogate the fluorescent labels of the attached base to get an image in which each cluster will have a different color representing the inserted nucleotide. The images are then processed in order to extract numerical signals for every base at every synthesis event from all the parallel reactions allowing the base calling. These raw image files represent terabytes of data and require substantial storage resources. After the removal of the blocking group and the fluorescent label, and the restoration of the OH group in 3' position, the sequencing reaction continues with the next cycle⁶². The number of cycles, corresponding to the read lengths, is limited by multiple factors that cause signal decay and dephasing. The precision of base calling is affected by the increase in interference signals as the length of sequences increases. This is mainly due to excess or inadequate incorporation of nucleotides or failure in removing the terminator assembly. With subsequent cycles, the errors can be accumulated by producing heterogeneous populations of filaments with various lengths within a cluster generating an inaccuracy in base calling especially at the 3' end. The most common errors in Illumina are substitutions with a large percentage of them occurring when the previously incorporated nucleotide is a Guanine⁶³. Furthermore, genomic analyzes of Illumina data have found that sequences having regions rich in AT or GC are underrepresented, probably due to amplification bias during library preparation^{63,64}. Despite these problems, Illumina is a precise and reliable technology with an error rate lower than 1%^{65,66}.

IonTorrent

IonTorrent technology was first described in July 2011 by Rothberg, who devised an innovative sequencing method called ionic semiconductor sequencing⁶⁷. In a short time, this technology

has become very successful despite being the last created, both for the speed with which it performs sequencing run and for costs that are much lower than other NGS platforms. Semiconductor sequencing is based on the detection of hydrogen ions released after nucleotide insertion by semiconductor chip allowing faster sequencing times through omission of time-consuming imaging steps. The IonTorrent protocol also follows the basic steps of all other NGS platforms: library preparation, clonal amplification, and sequencing reaction. The library preparation consists in the random fragmentation of the samples by sonication, nebulization or time-dependent enzymatic reactions, in order to obtain fragments of about 200 base pairs. Following is the addition of the adapter oligonucleotides containing the barcode sequences at the ends of the doubled-strand DNA fragments through ligation.

The clonal amplification is carried out by *emulsion PCR*, which consists in the preparation of an emulsion of water and oil where micro-bubbles of water act as microreactors for PCR reactions. All reagents together with non-paramagnetic beads are added in aqueous solution to carry out the amplification reaction. Fragments to amplify are bound to the beads through complementarity between universal adapter sequence P1 and the anchoring oligonucleotides exposed on the beads surface. The aqueous solution is subsequently mixed with an oil solution which, being hydrophobic, forms microbubbles of water called *micelles*. Each ideal micelle must contain a single fragment of library DNA, PCR primers, one bead, and the PCR mix. Amplification starts with denaturation of DNA fragments and only reverse strand will hybridize with the anchoring oligonucleotides through the adapter sequence P1. Subsequently, the DNA polymerase begins the amplification of the forward strand starting from the end bounded to the bead. Once the amplification is completed, the reverse template is denatured by the bead and bind another free anchoring oligonucleotide, while the newly produced forward strand remains fixed on the surface of the bead (**Figure 5**).

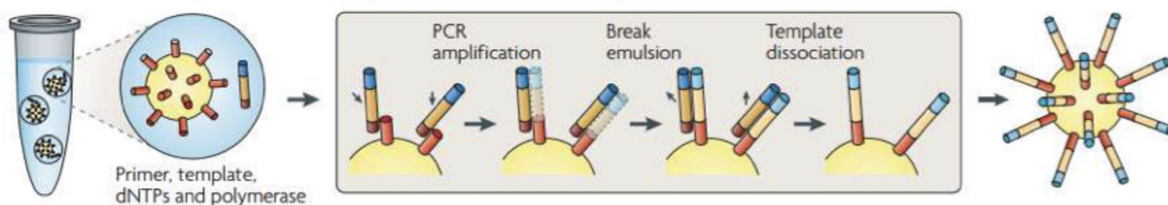


Figure 5. Emulsion PCR. Inside each micelle there is a bead and a fragment of DNA. Through the adapter sequences the reverse filament is fixed on the surface and acts as template for polymerization of the forward filament. At the end of the PCR cycle, the filaments are denatured, and the forward strand remains fixed on the surface of the bead while the reverse hybridizes with another anchoring oligonucleotide to act as template for another PCR cycle. At the end of amplification phase, the surface of the bead will be covered with clonal amplifiers. Figure adapted from Metzker ML, 2010⁶⁸.

There will be a further amplification cycle that will exploit the forward and the reverse fragments connected to the bead as templates. After 16-18 cycles, the beads will be covered with clonal amplifiers. The product of the reaction is enriched removing beads that have no amplified product attached. Since emulsion PCR is a passage used in other NGS platforms, the real peculiarity of IonTorrent is the sequencing process that has two specific characteristics:

- The use of semiconductors as a structural element of the supports in which the sample is dispensed for sequencing (*IonChip*);
- The use of a detection system not based on luminescent reactions but on potential variations.

The main nucleus of sequencing chemistry is the IonChip (**Figure 6**). It is constituted by an upper surface layer, in which there are microwells designed to accommodate only one bead and the sequencing reagents. Below the surface layer, there is a second layer consisting of semiconductors, which allows transmission of signals to the underlying layer. The latter is structured as a sensor plate, one for each well, which has the ability to record small pH changes that occur in each well during sequencing reaction, converting them to potential differences and subsequently into digital data.

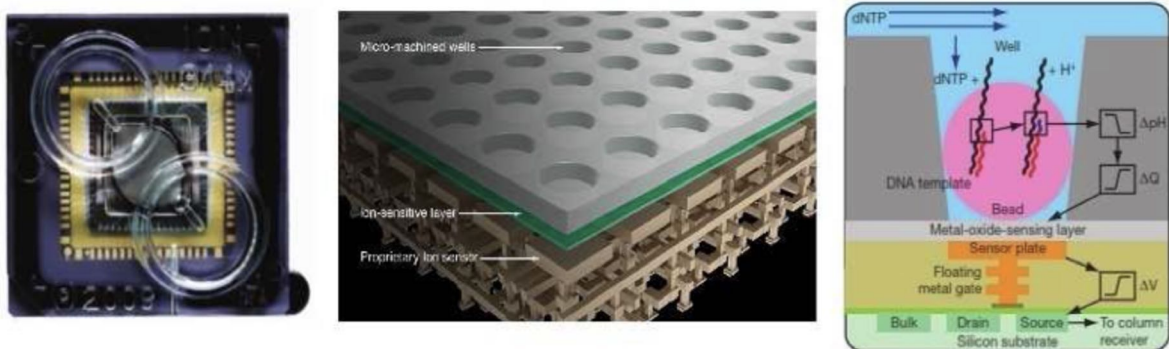


Figure 6. Features of the IonChip. *IonChip* is shown on the left of the image. The figure in the center represents the three-dimensional structure of the *IonChip*, in which the 3 layers are visible. On the right there is a schematic representation of the sequencing chemistry. Figure adapted from Rothberg JM, et al., 2010⁶⁷ and *Ion AmpliSeq™ Library Preparation User Guide*⁶⁸.

IonTorrent sequencers exploit a natural feature of the polymerization reaction to detect which nucleotide has been added, that is the release of hydrogen ions which occurs when a nucleotide is added to a DNA sequence. Sequencing process is characterized by subsequent cycles consisting in the release on the chip of one of the 4 unmodified deoxynucleotides (A, T, G, or C) and a subsequent washing step to prepare the conditions for the release of the next nucleotide. The incorporation of a nucleotide complementary to the first free base of the

template by a polymerase determines the hydrolysis of the triphosphate group of the nucleotide with the release of a proton. The release of protons leads to a pH change inside the microwell that is detected by the underlying sensors and the variation is converted into digital information (**Figure 6**).

During each cycle, the recorded pH variation is directly proportional to the number of bases incorporated in the nascent sequence. The data are represented through a graph called Ionogram, which shows the number of bases incorporated for each flow performed by the sequencer.

The main advantages of this sequencing technology are the absence of imaging system which is usually expensive and for which data acquisition is time-consuming, and in the ability to use unmodified deoxynucleotides⁵⁹. A problem of IonTorrent technology is the errors (insertions or deletions) during base calling in homopolymeric zones. Sequences that have more than 4 repeated nucleotides have a high percentage of error that is to be considered during data analysis. Because of this problem, the percentage of error compared to Illumina is slightly higher and around 1,7%⁶⁵.

3.5. Whole exome sequencing

Whole exome sequencing (WES) consists in sequencing virtually all known protein-coding regions (“exons”) in a genome. WES was used for the first time in 2009 to rediscover a previously known mutation that causes Freeman-Sheldon syndrome in a four-member family⁶⁹. Over the last 10 years, WES has been used with success in thousands of cancer patients leading to new discoveries, including mutations in different forms of cancers including leukemia^{70–72}. WES covers about 30 Mb scattered across 180,000 exons in the genome⁶⁹. The sequencing of exons is both timesaving and less expensive compared to whole genome sequencing. Moreover, the reduction in sequencing costs led to a substantial increase of the target region coverage making easier and more likely to detect relevant mutations. WES can be used to detect mainly different point mutations (SNPs), small insertions and deletions (indels), and copy number variations (CNVs).

It has been estimated that up to 85% of disease-causing mutations can be found in the exome regions⁷³. To perform WES, it is essential to efficiently isolate the exons that are spread out in the genome. Currently, there are three companies that offer commercial kits for capturing exons: Agilent, Roche/Nimblegen and Febit⁷⁴. They all apply the same principle to extract exons: randomly fragmented DNA is hybridized to oligonucleotide baits complementary to the

exome targets. The difference between the commercial kits is in the way hybridization is performed: the most commonly used methods are solid-phase and liquid-phase hybridization. In solid-phase hybridization, probes complementary to the sequences of interest are fixed to a solid support, microarray or filter, and allowed to hybridize with total DNA⁷⁵. The non-selected genomic fragments are washed away while the target fragments remain attached via probes to the solid support, and are later eluted, PCR amplified and sequenced. In liquid-phase hybridization, the used probes are biotinylated. The probes hybridize with their target DNA and are then captured with magnetic streptavidin beads⁷⁶. The beads are then removed, and the selected fragments can be amplified and sequenced. In 2009, the commercial kits have been setup to target the human consensus coding sequence region, covering about 83% of the RefSeq coding exon bases (29 Mb of the genome)⁷⁷. Manufacturers are continuously improving their kits by adding more baits to increase the percentage of exonic regions captured. For example, the newest version of Agilent Sure Select Kit (v7) covers approximately 50 Mb. Coming to WES limitations, biases and inefficiencies introduced during the targeting capture process needs to be considered in the interpretation of the sequencing results. For example, the capture efficiency across different exons is not uniform leading to uneven read coverage of sequenced exons and loss of coverage regions⁷⁸. In addition, since WES is a technology relying on previous knowledge, unknown exons are skipped by capture and potential important variant can be miss⁷⁹. WES is often used to study cancer genomes because it overcomes difficulties encountered with other technologies. Cancer samples are problematic to sequence because are almost always a mixture of cancerous and healthy cells. Investigating the cancer genome with Sanger sequencing will miss most variants at low frequency⁸⁰.

3.6. Somatic variants detection by WES

WES raw data

WES results in a huge amount of data, which need to be analyzed with bioinformatics tools. This thesis focuses on methods used for WES analysis and the issues to address for detection, prioritization and interpretation of somatic variants from WES data of cancers. The big volume of information generated by NGS experiments is the first problem for the management, storage and data analysis.

The information captured is encoded in images (Illumina) or in digital data (IonTorrent) that must be recorded, managed and processed. Next, acquired data are converted into sequences in a process called “Base calling” which requires intensive computation. Each platform uses

specific computer algorithms that evaluate several parameters, such as the detected fluorescence or the potential difference, the background noise, and the presence of any non-specific signals, to generate nucleotide sequences. A quality score, related to the probability of error, is assigned to each called sequence bases. Base quality scores are very important for the subsequent steps of analysis. Although quality assessment varies according to the type of platform, the calculations are based on a *Phred score index* introduced in 1998 for sequence data obtained with Sanger sequencing^{81,82}. A *Q* Phred quality score is an integer mapping of the *p* probability that the corresponding base call is incorrect:

$$Q = -10 \log_{10} p$$

The higher the quality of the base, the higher the Phred score: for example, the chances that base with a *Q*=30 is incorrectly called are 1 in 1000 (99,9% accuracy). Base quality score during following analysis is considered to exclude fragments or bases having low quality and also to optimize the precision of alignment to the reference sequences⁸³.

Once finished a sequencing, millions of short sequences called *reads* (usually long between 25-250 bp long) are obtained and stored in text files encoded in *FASTQ* format. Each read consists of four rows. The first row, beginning with the '@' character, is a header uniquely identifying the corresponding read and an optional description; the second row reports the read sequence in *FASTA* format indicating with 'A', 'T', 'G', 'C' characters the corresponding bases, or 'N' when the base calling failed; the third row is delimited by '+' character, optionally followed by the sequence identifier and description, indicating the end of the sequence and the beginning of the quality scores line; the last line contains the quality scores for each called base, encoded in a way such that each single character corresponds to the quality of the base in that position.

An example of FASTQ read is given below:

```
@K00171:147:H57LJBBXX:5:1101:32289:2123 1:N:0:AAGGACAC
GCCTCTCTGGAGAGAATGAGCTGGTGTTCGGGGTGCAGGTGACCTGTCAGGTGAGGCCATCCCC
+
AAAFAFAFJFJJ7AJJFJJFJJAJJJJJJJJJFAFJAJJJJJFFFJJJFJJJJF<FJJJJJJFJJFJJJJJJJAF
```

Phred quality scores from 0 to 93 are encoded using ASCII 33 ('!') to 126 ('~'). Raw reads undergo to quality control and pre-processing which includes the removal of adapter sequences and sequences with low quality.

SNPs and indels can be detected from FASTQ files by an analysis comprising three main steps:

Alignment: the reads must be aligned to a reference genome.

Post-alignment processing: the alignment of reads must be improved to limit the calling of false positive variants.

Variant calling: Detection of sites in sequenced regions that are different respect to the reference genome.

Next, usually the detected variants are annotated with information deriving from databases and filtered to remove low confidence variants.

Short reads to the reference genome alignment

Obtaining good alignments is a critical task for most NGS projects and is crucial to identify true somatic variants. The short-length of NGS reads, the uneven coverage of the sequenced genome following the Poisson distribution, the repetitive regions of the genome, indels, and sequencing errors making the alignment step computationally intensive and time-consuming. Several alignment algorithms have been developed to address these challenges. Since alignment can be time consuming, a good aligner must be a good compromise between accuracy and speed of computation. *Bowtie*^{84,85} and *Burrows-Wheeler Aligner (BWA)*⁸⁶ exploit *Burrows-Wheeler Transform (BWT)* to efficiently store a compressed prefix tree (Trie) of the reference genome in memory. A Trie is a data structure which stores every prefix of a string such that every exactly repeated substring is only recorded once. The time required to test if a query string is an exact match of a reference string represented by a Trie is linear with the length of the query string. BWA uses the standard search algorithm of prefix trees allowing also mismatches and gaps in the NGS reads. For paired-end reads, BWA first aligns both reads of a pair separately and then joins them. If paired-end reads map to different positions in the reference genome, the region where the reads align close to each other is preferred. The alignment data are stored into a text file named *SAM* (“*Sequence Alignment/Map*”) format, that informs for each sequence about chromosome, alignment position, strand, and quality estimate of alignment. The SAM format consists of a header and an alignment section⁸⁷. The header contains various information about sequenced sample (e.g. sequencer, reference sequence dictionary, program used for alignment or optional comments). Each header line begins with ‘@’ character to distinguish them from alignment section. Alignment data are organized in 11 mandatory fields, as well as a variable number of optional fields⁸⁷.

SAM files are usually very large, posing significant issues for storage and for manipulation in subsequent analysis steps. SAM files can be compressed into an indexed binary format named *BAM*, which is more efficient in term of space and speed since the data are more accessible by software to perform downstream analysis.

Column	Name	Type	Description
1	QNAME	String	Query name of the read or the read pair
2	FLAG	Integer	Bitwise FLAG
3	RNAME	String	Reference sequence name
4	POS	Integer	1-based leftmost position of clipped alignment
5	MAPQ	Integer	Mapping quality (Phred score)
6	CIGAR	String	Extended CIGAR string
7	MRNM	String	Mate reference name
8	MPOS	Integer	1-based leftmost mate position
9	ISIZE	Integer	Inferred insert size
10	SEQ	String	Query sequence on the same strand as the reference
11	QUAL	String	ASCII of Phred-scaled base quality+33

Table 1. Description of the fields included in SAM/BAM file format.

Post-alignment read processing

Before variant calling, a post-alignment processing is important to enhance the quality of the alignment used to detect variants. Since PCR amplification introduces duplicates of reads in the data, influencing coverage and downstream analysis, the removal of PCR duplicates is fundamental to accurately represent the sequencing depth and reduce false positive variants. Many tools have been developed for PCR duplicates removal, such as Picard (<http://broadinstitute.github.io/picard>) or SAMtools⁸⁷. Recently, random small oligonucleotide barcodes called Unique Molecular Identifiers (UMIs) are increasingly used in sequencing experiments to easily recognize PCR duplicates and correct artifacts inserted during sequencing or amplification step⁸⁸⁻⁹⁰. UMIs are attached to the original fragments through ligation or primer extension prior to PCR amplification, and then are retrieved from sequencing reads allowing the molecular-tracking of each reads. Sequenced fragments with the same UMI sequence derive from the same molecule and all except one are marked as duplicates. In addition, reads with identical UMI sequence can be used to reconstruct a consensus read by majority voting or weighted scoring at each base, in order to correct sequencing or PCR errors through base-call consensus and UMI counting⁹⁰. This method can introduce artifacts arisen during the first-cycle PCR that propagate through amplification or errors hitting UMI sequences. Moreover, the correction of errors could lead to rare variants missing.

Another important parameter to consider in variant calling is the base quality score. Studies have reported that base quality scores are not accurately assigned by sequencer-software and not reflecting the real error probability⁹¹. In particular, base quality scores are largely affected by the position within a read and the sequence context of a considered base. In addition, certain dinucleotides are more prone to errors on each sequencing platform, and specific platforms

present systematic errors (errors in rich-GC regions in Illumina or errors in repeated regions in IonTorrent). Since base quality score is fundamental for accurate variant calling, bioinformatics tools have been developed to perform base quality score recalibration. A tool of Genome Analysis Toolkit (GATK) recomputes base quality scores after alignment, considering position, dinucleotide context, and a baseline expected error rate calculated from loci without SNPs expected⁹². Post-alignment processing comprises also local realignment of reads around indels. The mapping of short reads including indels is difficult for aligners, resulting in many mismatches near the misalignment respect to the reference genome, which ultimately results into false SNPs. In addition, reads having an indel near their start or end are often incorrectly aligned with respect to the true indel. Local realignment converts sequences with indels into reads containing a consensus indel suitable for correct variant detection (**Figure 7**).

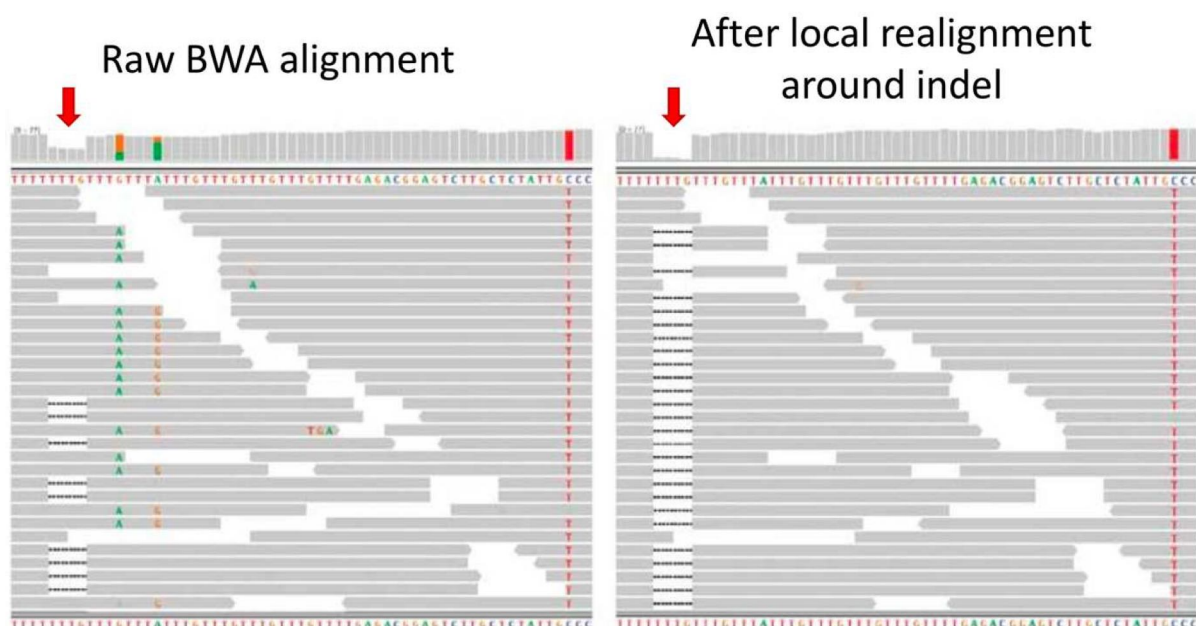


Figure 7. Alignment visualization before (left) and after (right) local realignment around indels⁹². Reads are indicated by grey arrows; highlighted bases are mismatches with the reference whereas dashed lines represent a deletion of four bases. Local realignment around indel enhances reads alignment indicating that the two variants G/A and A/G (highlighted letters) are misalignment artifacts. Figure adapted from DePristo MA, et al., 2011.

Somatic variant calling

After alignment enhancement, the next step of analysis is variant calling. Four different types of variants can be detected by WES: SNPs, indels, copy number variations and structural variants (duplications, translocations, etc.). The calling of each type of variant is based on specific assumptions requiring different algorithms and few variant callers are enough versatile to detect more than one type of variant. The general strategy to detect SNPs and short indels is to search for differences between the reads that cover a considered region and the sequence of the reference genome. Considering that NGS reads are short, structural variants and long indels

are detected locating the breakpoints based on the sudden change of read depth or misalignment patterns by exploiting split-reads and de-novo assembly methods⁹⁰. Germline and somatic variant (SNPs and indels) calling also requires different algorithms. Germline variants are expected at allele frequencies of 50 or 100% and sequencing artifacts are easily identifiable because are usually present at low frequency. Conversely, some real somatic variants have low frequencies in case of contaminated samples, circulating DNA, or tumor rare subclones, requiring complex statistical modeling and advanced error correction in order to disambiguate them from artifacts. In this thesis, only the detection and analysis of somatic SNPs and indels will be considered.

As previously discussed, one of the main aims of cancer sequencing projects is to discover specific somatic mutations leading to tumorigenesis, whose identification may give information on disease mechanisms being clinically relevant, for diagnostic and prognostic aims, and to guide the optimization of the type and intensity of the therapy approach.

To distinguish somatic variants from germline and loss of heterozygosity (LOH) variants, the genome of tumor sample is compared with the genome of normal tissue collected from the same patient in the so-called tumor-control matched sequencing approach (**Figure 8**).

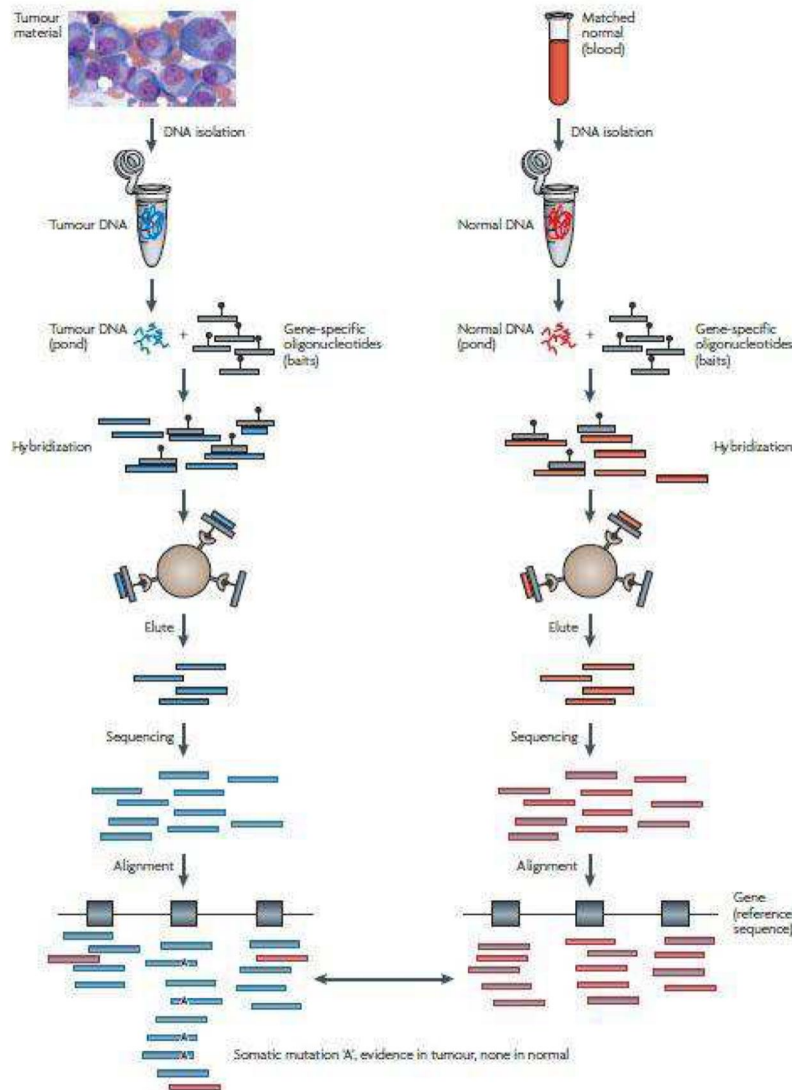


Figure 8. Schematic diagram of tumor-control matched sequencing workflow. Library preparation, sequencing, and alignment of tumor (blue) and control (red) samples are performed separately. Somatic mutations in the tumor DNA can be detected from comparison with control DNA⁸⁰.

Despite the design of the analysis seems very simple, actually a mere subtraction of variants called in the control sample from the tumor sample is absolutely not adequate to detect somatic variants. This step of analysis is challenging because of complexity of cancer samples characterized by altered ploidy⁹³, intra-tumor heterogeneity⁹⁴, low tumor purity⁹⁵, and insertion of false positives or artifacts during tumor tissue conservation⁹⁶, library preparation, sequencing or alignment⁹⁷. For an optimized somatic variant detection, the tumor DNA should be extracted from a biopsy of pure population of tumor cells and without significant necrosis or inflammation⁹⁸. In this condition, most of the somatic mutations are expected to be heterozygous. Actually, the tumor samples are collection of cells, comprising both normal cells and different populations of malignant cells. In addition, tumor biopsies are often single snapshots of the whole tumor affected by a selection bias of cell populations influencing the

frequencies of the detected somatic variants. Therefore, the expected allele frequency of somatic SNPs in tumors may be substantially less than 50%, making difficult to distinguish true variants from errors or artifacts even with high sequence coverage (**Figure 9**).

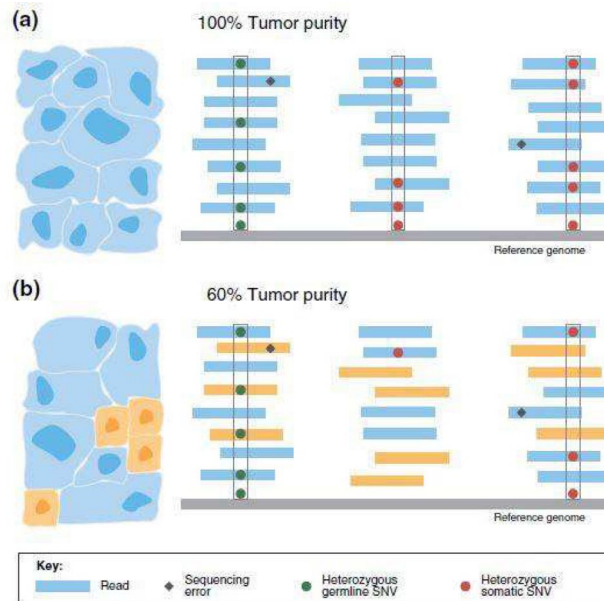


Figure 9. Variant detection in tumor samples. Reads are aligned to the reference genome (gray bar). Germline variants are indicated by green circles, somatic variants by red circles and sequencing errors by black diamonds. **(a)** Representation of variant detection in a pure tumor sample. Assuming that the locus is not affected by copy number variations, a heterozygous germline or somatic SNP is present in about half of the reads covering the region. **(b)** Representation of variant detection when tumor purity decreases. The reads deriving from tumor and control cells are indicated with blue and orange bars, respectively. The number of reads supporting somatic variants proportionally decreases with tumor purity, thus reducing the signal to detect somatic variants. In this example the somatic variant in the middle is not detected because not distinguishable from a sequencing error⁹³.

Several variant callers for somatic mutations have been developed to try and overcome above-cited issues with different approaches. VarScan2⁹⁹ and VarDict¹⁰⁰ exploit a heuristic approach to identify potential variants present in reads respecting algorithm-specific thresholds and then apply statistical tests or rules to call somatic variants. SAMtools¹⁰¹ and SomaticSniper¹⁰² make genotype analysis assuming diploidy in both tumor and control. They consider the probability of the joint genotypes with Bayes' rule (SomaticSniper) or log-likelihood ratio (SAMtools) to calculate a "somatic score". The higher is the score, the higher is the probability that tumor has different genotype respect to control and the site is marked as potential somatic variant subject to post-filters. Another caller, MuTect¹⁰³, considers allele frequencies instead of genotypes not assuming the diploidy with the aim to also detect somatic variants at low frequencies. Independently for each locus, MuTect detects variants in tumor sample by a Bayesian classifier to evaluate if each observed non-reference base can be due to a sequencing error. The allele frequency for each site is estimated as the fraction of tumor reads that carry the variant and not

assuming it is heterozygous. Then, site-based filters are applied to eliminate false positives considering six type of known sequencing errors and artifacts (proximal gap, poor mapping, triallelic site, strand bias, clustered position, observed in control). Finally, the remaining variant sites undergo to another statistical evaluation by a different Bayesian classifier to sort out somatic and germline variants. Recently, some variant callers, such as MuTect2¹⁰⁴ and Strelka2¹⁰⁵, were developed exploiting the haplotype-based strategy because the detection of indels and structural variants is notably better respect to position-based strategy. These tools locally assemble reads in a region and generate candidate haplotypes that may be represented by de Bruijn-like graphs⁹⁰. The likelihood of each haplotype is estimated by aligning each read to the haplotype and counting the read support⁹⁰. This approach is more reliable in regions dense of variants because it is not based on local alignment which is prone to errors particularly in difficult regions⁹⁰.

Since each variant caller has its own merits and defects, the choice of which to use strongly depends on what type of variants are of interest, the desired VAF and the coverage of the sequenced samples. Variant callers based on genotype approach are designed for low-coverage data and not sensitive enough to detect low frequency variants because the diploidy assumption in tumor implies that real variants' allele frequency is around 0.5 or 1.0⁹⁰. MuTect and callers that estimate allele frequencies directly are the most efficient to detect somatic variants at low frequency, especially with high depth of coverage. Heuristic analysis-based callers are efficient for low-frequency variants, but the caller-specific thresholds must be chosen very carefully^{106,107}. If indels and structural variants are of interest, haplotype-based variant callers are the most convenient.

Variant caller	Type of variant	Type of core algorithm
VarScan2	SNV, indel	Heuristic threshold
VarDict	SNV, indel, SV	
SAMtools	SNV, indel	Genotype analysis
SomaticSniper	SNV	
MuTect	SNV	Allele frequency analysis
MuTect2	SNV, indel	Haplotype-based analysis
Strelka2	SNV, indel	

Table 2. List of some of tumor-normal somatic variant callers grouped for type of core algorithm.

3.7. Variant annotation and prioritization

Somatic variants detected in each considered sample by callers are very numerous and the majority of them are passenger variants. Several approaches can be used to prioritize somatic mutations including variant effect prediction, evaluation of allele frequency in populations,

observation of variant or mutated gene recurrence in disease samples and functional prioritization of gene products by pathway or protein-protein interaction (PPI) network.

Annotation of somatic variants and prediction of functional impact

Many genome annotation databases have been developed and are continuously updated, such as Ensembl¹⁰⁸ or UCSC Genome Browser¹⁰⁹. There are bioinformatics tools that exploit information gathered in these databases in order to annotate variants in coding and non-coding regions genome wide. For instance, ANNOVAR¹¹⁰ and SnpEff¹¹¹ allow functional impact prediction by annotating transcripts and giving information about amino acidic changes. Missense, nonsense, stop-loss, frameshift and splice site variants have all potential to affect protein function and they are prioritized over synonymous variants by these tools. In particular, truncating and splicing variants are of interest because of their high cellular and systemic impact. In addition, several algorithms and methods have been developed to predict deleterious and tolerated variants basing on sequence identity, sequence conservation, evolutionary relationship, protein primary and secondary structure, entropy-based protein stability. The most used are SIFT¹¹², PolyPhen2¹¹³ and MutationAssessor¹¹⁴, but they lack specificity and sensitivity to sufficiently reduce the large number of candidate mutations from exome sequencing^{115,116}. They are based on *a priori* assumptions that may not be enough exhaustive to predict an oncogenic impact of a mutation. Some epitopes of phospho-kinases and signal transduction proteins can be quite complex, and these approaches could not capture this complexity missing important oncogenic variants¹¹⁷. For example, MutationAssessor predicts as tolerated a well-known activating variant (H1047R) in *PIK3CA*¹¹⁸. In addition, predictions of different tools are very often in contrast each other because based on different assumptions. More complex methods have been developed to attempt increasing predictive precision combining more predictors. MetaSVM and MetaLR are based on Support Vector Machine (SVM) and Logistic Regression (LR), respectively, to combine 10 different predictor scores and the maximum frequency indicated in 1000 genomes populations in order to predict for deleterious variants¹¹⁹. These tools are not sufficient alone to exhaustively predict driver variants, and other approaches are needed to reduce false negatives.

Somatic variant filters based on population allele frequencies

Various databases could be used to delete common polymorphisms considering allele frequencies both as an average across all the populations studied and for individual population groups, including African, Admixed American, East Asian, Finnish, Non-Finnish European,

South Asian, and other. NCBI dbSNP database¹²⁰, established in 2001, is continuously updated to gather both well-known and rare variants from many organisms giving additional information about disease association, genotype origin, and somatic and germline variant information. One of the most popular databases for population allele frequencies is 1000 Genomes Project¹²¹ which uses data from the sequencing of more than 1000 healthy people of five ethnicity groups. More recently, the Exome Aggregation Consortium (ExAC)¹²² has assembled and reanalyzed WES data of 60,706 unrelated individuals who are part of various disease-specific and population genetic studies¹²³. In 2017, the authors of ExAC published the Genome Aggregation Database (gnomAD) providing allele frequencies of variants detected from 123,136 exomes and 15,496 genomes of unrelated individuals giving a more comprehensive data¹²². Currently, ExAC and gnomAD are the most used databases considering that population allele frequencies are calculated from much more individuals than 1000 Genomes Project. Nevertheless, ExAC and gnomAD have some limitations: individuals are not randomly sampled from the population, some populations are not represented, some exons have confounded or no coverage owing to exome capture technologies, absence of individuals' clinical data, some control cancer exomes obtained from blood may be contaminated by circulating cancer cells. In addition, the variants causing cancer are expected to show an incomplete penetrance and/or variable expressivity of the clinical phenotype implicating that there are carriers considered as healthy in these databases¹²⁴. A study reported that 24% of 5,700 asymptomatic individuals were carriers of confirmed causal alleles for at least 1 severe condition¹²⁵. Finally, the most of cancers arises at adult age not impacting on population fitness and variants associated to cancer development could have population allele frequencies larger than usually used thresholds (1%). Thus, interpretation based on population allele frequencies should be done with care requiring other approaches to further prioritize variants.

Additional criteria for variant prioritization

The identification of candidate variants for a particular cancer or clinical features is not always obvious, despite the application of variant filtering protocols. The remaining variants with likely functional impact are much more than those can be experimentally validated. For a practical point of view, individuation of candidate genes carrying functional variants in the context of existing biomedical knowledge can be useful to obtain a set of variants for further functional validations or experimentations. Bioinformatics tools, such as above-cited ANNOVAR and SnpEff, can annotate variants with information from published cancer and disease databases. ClinVar¹²⁶ reports association between genomic variants and diseases

providing clinical data. Specific for cancer studies are Cancer Gene Census³³, COSMIC¹²⁷, and The Cancer Genome Atlas (<https://cancergenome.nih.gov/>). Cancer Gene Census is a continuously updated catalogue of genes carrying mutations that have been causally implicated in cancer. While, COSMIC and The Cancer Genome Atlas are large databases that report variants detected in different types of tumor allowing for identification of relevant cancer-related variants.

Recurrence as indicator of importance?

Another approach to identify possibly driver variants is to look for recurrent somatic variants in cohorts composed by patients with the same or similar phenotype. Recurrence analysis of variants is based on the concept that high number of cancer samples harboring the same mutation supports its involvement in unregulated cellular growth or other cancer hallmarks. For example, the mutation V600E in BRAF was found in more than 50% of melanoma patients¹²⁸ or the mutation V617F in JAK2 has a frequency of about 90% polycythemia vera patients¹²⁹. Nevertheless, the recurrence of variants in cancer studies is often very low and little informative leading to search recurrences at higher biological levels, such as genes. Different patients can be affected by different variants hitting the same gene inducing similar functional protein changes and, thus, phenotype. Through NGS, *KRAS*, *TP53*, and *APC* was found mutated by different variants in 112 colorectal cancer patients with frequency of 36%, 39%, and 30%, respectively¹³⁰. Other cancer types failed to show highly recurrent genes, such as medulloblastoma where cBioPortal database reported *CTNBB1* as the most recurrently mutated genes in 9% of patients.

However, many recurrently-mutated genes in cancer studies do not seem to be associated to cancer development considering their biological function and/or genomic features. In some cases, recurrences are due to false positive variants created by technical or alignment errors. In other cases, the variants are artifacts inserted during sample conservation. Indeed, most cancer biopsies are in formalin-fixed and paraffin-embedded, but nucleic acids of these samples are subject to have cross-linking and degradation¹³¹.

In addition to artifacts, it is known that the mutation rate is not constant across the human genome but depends on the genomic context of a nucleotide and the type of mutation. In addition, some genes are more prone to mutate due to their length, expression rate, and/or replication timing. The background mutation rate (BMR) is not constant across patients with same cancer, and hypermutated cases are often found. Lawrence et. al reported that pediatric tumors have a mutation rate much lower (0.1/Mb) than other cancers often induced by

environmental factors, such as lung cancer or melanoma (100/Mb)¹³². Finally, certain genomic regions display localized somatic hypermutation, called kataegis⁹³.

Several tools have been developed to identify driver recurrently mutated genes to solve the problem of extensive false positive recurrently mutated genes in cancer cohorts. Examples are MutSigCV¹³² and MuSiC¹³³ which find recurrently-mutated genes evaluating whether the observed number of mutations in a considered gene is significantly higher than the number expected according to a BMR. The BMR is the probability of observing a mutation in a specific location of the genome¹³². The main differences between methods are in how BMR is estimated and how many complications are considered. These tools require very large cohorts to identify high-confident recurrently-mutated genes.

Pathways and interaction networks topological analysis

Initially, network and pathway-based methods have been largely used with expression data to evaluate biological functions affected by aberrant expression of genes⁹. Cancer development is well known to be characterized by alteration of key functions affected by somatic mutations^{8,10}. In fact, protein-coding genes do not act in isolation, but rather interact participating to complex cellular reactions linked to specific biological functions. Moreover, many cancer studies reported that patients with similar phenotypic characteristics carried a few recurrently mutated genes and a long tail of rarely mutated genes^{10,134}. These evidences lead to focus the research of driver events at pathway level where the frequently and rarely mutated genes can alter same biological functions or can concur to alter them. Pathways are small-scale systems of well-studied processes where interactions comprise biochemical reactions and events of regulation and signaling between genes, proteins and other biomolecules that carry out biological functions¹³⁵. The aim of this approach is to identify groups or combinations of genes associated to specific pathways recurrently mutated prioritizing also rare driver variants that are crucial for precision oncology (**Figure 10**).

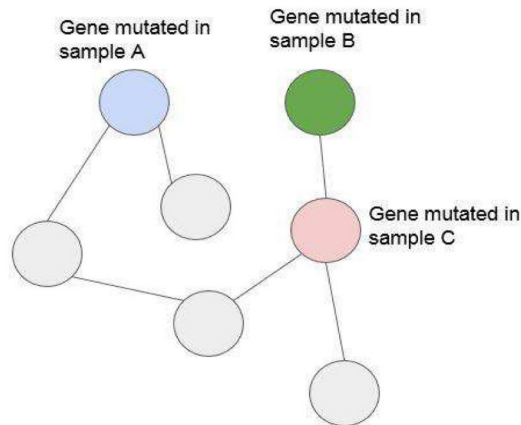


Figure 10. Example of recurrently mutated gene network. *Nodes represent genes and edges represent gene interactions. The three colored nodes indicate mutated genes detected in three different patients. This reconstruction shows a group of connected genes which is recurrently mutated suggesting that biological functions associated to it are altered.*

A first approach to determine if groups of genes are recurrently mutated in a cohort is to compare them to known pathways gathered in public databases without considering information about interactions. This approach consists in a gene set enrichment analysis applied to pre-defined lists where the overlap between mutated genes and sets of genes with known functional annotations (pathways) is statistically evaluated through hypergeometric test or Fisher's exact test followed by a correction for multiple testing. Various public databases give functional information such as KEGG¹³⁶, Reactome¹³⁷ and Gene Ontology (GO)¹³⁸. More complex methods, such as PathScan¹³⁹, were developed to perform per-patient enrichment analyses and, then, combine the results of tests across all of the considered patients. These approaches result in a list of processes and pathways significantly enriched of mutated genes giving a first view on affected biological systems, but they have four main limitations. First, the annotated gene sets used for enrichment analyses are often large and these methods may not capture significant enrichment of mutations in a smaller subset belonging to the considered gene set. Second, these methods are based on already well-known pathways making not possible to detect driver mutations in less-characterized pathways and to predict novel cancer pathways. Third, the pathways are considered as separated gene groups not considering their crosstalk which is, instead, fundamental for cancer development¹⁴⁰. Finally, the topology of interactions between genes is ignored and consider the contribute of each gene to a pathway equally.

A second approach that can be used to infer novel cancer genes and pathways is based on gene or protein networks, overcoming the limitations of pathway analyses. The altered genes and

their “interacting genes” are extracted from public databases to reconstruct an interaction network. This depicts interactions between mutated genes, but also relations with non-mutated genes that could participate in cancer development due to their interactions. Several public databases can be used to reconstruct a gene or protein network. The interactions of network can be experimentally verified, being more reliable, or only computationally predicted. STRING¹⁴¹ gathers both types of interactions whereas HPRD¹⁴² contains only those experimentally validated. The above-cited KEGG¹³⁶ and Reactome¹³⁷ are considered the most reliable sources of experimentally verified reactions.

Networks are often large and complex making their interpretation not obvious. In order to obtain the significantly mutated subnetworks making the interpretation of the result more comprehensible, methods performing quantitative analyses on biological networks integrating somatic mutations data have been developed. The topological features of interactions network can provide insights on the biological significance of participating proteins or genes. Usually, topological analysis considers two main aspects: centrality analysis and clustering analysis. Centrality analysis estimates the importance of single nodes in the connectivity of the whole network. Centrality analysis considers two aspects: node degree and degree of global centrality. Node degree is the number of edges of the considered node, but this estimation is only local because it estimates the centrality of the node with its surrounding proteins/genes. A global centrality analysis is required to estimate position of the node considering the whole network through the measures of closeness centrality¹⁴³ and betweenness centrality¹⁴⁴.

In contrast, clustering analysis is aimed at identifying groups of nodes (functional modules) where nodes are more connected to each other than to the rest of the network reducing considerably the complexity of the considered network. The detection of functional modules from large networks is still challenging, and many algorithms have been developed to investigate about different biological scenarios.

One of the most interesting clustering methods for somatic variant analysis is HotNet2¹⁴⁵, which is based on an insulated heat diffusion model to detect statistically significant mutated subnetworks evaluating both recurrence of gene mutations and local topology of protein’s interactions (**Figure 11**).

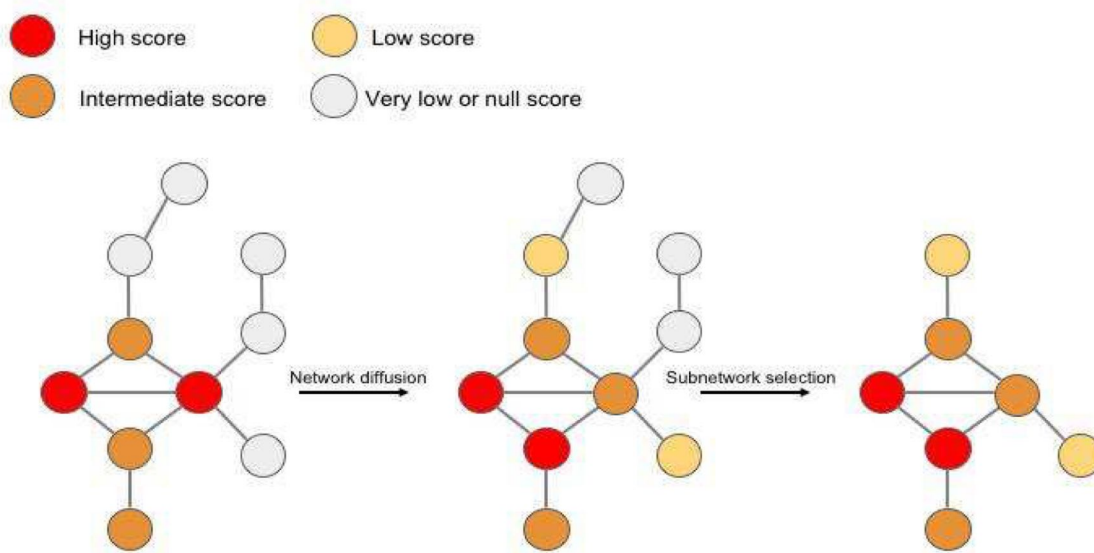


Figure 11. Hot diffusion model for analyses of networks (HotNet2¹⁴⁵). Nodes indicated genes and edges gene interactions. Color of nodes indicate the amount of “heat” concentrated into nodes. Hot diffusion across the network is used to estimate how strongly the aberrant genes are connected in the network. HotNet2 compute and assign to each node a score proportional to the frequency of mutations in the gene. Heat score starts to spread to neighboring nodes exploiting an insulated diffusion model until an equilibrium state is reached. Finally, subnetworks are identified in the network according to the amount of heat exchange between pairs of nodes and “cold” nodes are deleted.

HotNet2 algorithm is divided in four steps. Initially, HotNet2 estimates the influence between all gene pairs of the network by an insulated heat diffusion model where the heat is transferred to neighbors and a fraction of heat is retained by each node, until an equilibrium is reached. The estimated amount of heat that is transferred from a node to another represents the influence estimation which is strongly dependent by topology of interactions. Second, the “heat” scores for each gene are calculated proportionally to frequencies of gene mutations and they are placed onto the network in according to influence estimation obtained in previous step. Third, the mutated subnetworks are identified by removing gene-gene interactions where the amount of “heat” exchange is below a threshold δ , which is the median of minimum δ such that all subnetworks detected in 100 permuted networks have a maximum size L_{\max} . For all sizes k lower than L_{\max} , HotNet2 computes the number of subnetworks, having a minimum size k , obtained by the application of δ on the considered network. Finally, HotNet2 evaluates the statistical significance of the number of detected subnetworks for each minimum size k (X_k) using a two-stage statistical test. In the first stage, a p-value is computed by a permutation test where the empirical distribution of X_k is estimated on permuted networks where the heat scores are randomly changed on genes. The second stage consists of False Discovery Rate computation by Benjamini-Hochberg method for each p-value calculated previously. Finally,

the subnetworks with minimum size k having the lower corrected p-value are selected.

Although network analyses have high power of discovery in cancer studies and overcome limitations of pathway-based approach, they are limited by the quality and coverage of the interaction network. At the moment, high-quality interaction networks are relatively scarce. Most interaction networks are based on high-throughput screens to predict protein interactions, such as yeast-two-hybrid, mass spectrometry and computational methods, which are subject to high false positive rates¹⁴⁶. In addition, the databases of interaction networks report relations between protein that occur in different tissues, different cellular locations, or at different developmental time points or cell-cycle phase⁹³. Finally, in some cases the interpretation of the results is not obvious because many predicted interactions are not functionally annotated. The continuous improvement of network-based methods and an always more complete knowledge about human interactome will help understanding how somatic mutations can act to drive cancer progression.

4. Aims of the study

This PhD project aimed at improving available bioinformatics software and developing new tools for WES data analysis obtained from solid or blood cancers, to identify somatic driver events and to associate variants to different clinical conditions or features. The project has been characterized by both methodological development and new experimental data analysis:

- The development of a computational pipeline to detect and prioritize reliable somatic variants from WES data that could perform customizable, more complete and easier analyses addressing most of the challenges presented by WES studies in cancer. Moreover, the development of a systems genetics approaches to reconstruct pathway-derived meta-networks that can be easier to interpret than the most used protein-protein interaction networks. The use of bioinformatics tools to perform quantitative analyses on pathway-derived networks with the aim at detecting significantly mutated subnetworks easily associated to altered biological functions.
- During my PhD, different, progressively improved and updated versions of the pipeline and of the systems genetics methods whose final version is presented in this thesis have been applied in three cancer studies conducted in collaboration with different groups of researchers, that contributed data and biology expertise. These studies characterized by experimental work, including variant validation and functional study relying on the results from bioinformatics analysis presented in this thesis are:
 1. Discovery of putative driver variants and functional modules of mutated genes in Large Granular lymphocyte Leukemia (LGL-L).
 2. Genomic characterization of pediatric High-Risk Neuroblastoma (HR-NB) prognostic subgroups having different survival time.
 3. Detection of molecular mechanisms driving pediatric-type nodal follicular lymphoma (PTNFL) and primary follicular lymphoma of the testis (PFLT) development.

5. Materials and methods

5.1. Programming languages

Python

Python (www.python.org) is a general-purpose, high-level programming language. Its design philosophy emphasizes programmer productivity and code readability. It is widely used in the scientific community and a library specifically developed for computational molecular biology is freely available (Biopython).

R

R (www.r-project.org/) is a high-level and an interpreted programming language and a software environment mainly used for statistical computing, statistical software development and data analysis. In addition, Bioconductor (<https://www.bioconductor.org/>), an open source and development software project for the analysis and comprehension of high-throughput genomic data, is based on R programming language.

RStudio

RStudio (<https://www.rstudio.com/>) is an integrated development environment (IDE) which is designed to facilitate programming in R language. The RStudio project provides various advantages including a four-panel layout with a console for interactive R session, a source-code editor to organize a project's files, and panels with notebooks to organize fewer central components. Moreover, the console and source code-editor are tightly linked to internal help system of R and It is possible to set up different projects and switching between them is easy. Finally, it provides many easy-to-use tools for managing packages, the workspace, files, and more.

SCons

SCons (<http://www.scons.org>) is a software tool written in Python designed to facilitate software development by managing the building and compilation of large software projects. SCons' actions are generic and can be implemented in Python scripts that will be custom for the specific application.

Docker

Docker (<https://www.docker.com/>) is the most used framework that implements operating system-level virtualization. It allows to develop, test and distribute informatics code in standardized conditions inside units isolated from the physical computer, called “containers”. A container can be considered as a very light virtual machine which contains everything necessary to make software work, such as libraries, dependencies, code, and system tools. Moreover, Docker facilitates code distribution making the software work with any operating system.

5.2. Databases and web tools for somatic variant prioritization

The Single Nucleotide Polymorphism Database (dbSNP)

dbSNP¹²⁰ is a free public database of genetic variations of different species, curated and periodically updated by National Center for Biotechnology Information (NCBI) in collaboration with the National Human Genome Research Institute (NHGRI). dbSNP collects SNPs, indels, short tandem repeats (STRs), multinucleotide polymorphisms (MNPs), heterozygous sequences and names variants.

Catalogue of Somatic Mutations in Cancer (COSMIC)

COSMIC¹²⁷ is a public archive of somatic mutations detected in human cancer studies, which is curated from the Cancer Genome Project of the Sanger Institute. The data are collected from the scientific literature and large-scale resequencing studies of cancer samples undertaken by the Cancer Genome Project. In addition, provides a list of cancer genes already confirmed by experimental studies, named Cancer Gene Census³³.

ClinVar

ClinVar¹²⁶ is a free database curated by NCBI reporting variants associated to clinical phenotypes. It is based on dbSNP, to maintain information about the location of variation on human assemblies, and MedGen (<http://www.ncbi.nlm.nih.gov/medgen>), to give phenotypic descriptions.

dbNSFP

The dbNSFP¹⁴⁷ is an integrated database of functional predictions from multiple algorithms and annotations of SNPs of the human genome. It comprises prediction scores from 20

prediction algorithms, 6 conservation scores, and information about population allele frequencies, gene expression, gene interactions and etc.

Exome Aggregation Consortium (ExAC)

ExAC¹²² is a public collection of human variants detected by WES in 60,706 unrelated individuals providing information about allele frequencies in various populations and functional data. Individuals affected by severe pediatric disease were removed from data set. The WES raw data of all individuals were analyzed with the same workflow and jointly variant-called to increase consistency across projects.

Genome Aggregation Consortium (gnomAD)

The gnomAD¹²² is an improvement of ExAC in term of dataset size to compute population allele frequencies and type of variants reported. Indeed, it aggregates both exome data of 123,136 and genome data of 15,496 unrelated individuals sequenced in different population genetic and disease-specific studies.

Kyoto Encyclopedia of Genes and Genomes (KEGG)

KEGG¹³⁶ includes various databases gathering information about biological pathways, genomes, diseases, drugs, and chemical substances. This project was started in 1995 at the Kyoto University with the aim to improve interpretation of genome sequencing data developing KEGG PATHWAY database. KEGG PATHWAY is a collection of manually drawn pathway maps building a network of molecular interactions and reactions representing experimental-verified knowledge on metabolism and various cellular and organismal functions.

Reactome

Reactome¹³⁷ is a free catalogue of biological pathways representing molecular details of signal transduction, transport, DNA replication, metabolism, and other cellular processes. Each pathway is made up of biological reactions between entities (nucleic acids, proteins, complexes and small molecules) forming a network. Reactome reports physical interactions where substrates interact, possibly facilitated by enzymes or other molecular catalysts, to result into products. The reported interactions are various including classical chemical interconversions of intermediary metabolism, binding events, complex formation, transport events that direct molecules between cellular compartments, and events such as the activation of a protein by

cleavage of one or more of its peptide bonds. Each interaction is experimentally verified and supported by literature citations.

MutationMapper

MutationMapper (http://www.cbioportal.org/mutation_mapper.jsp) is a web tool part of cBioPortal¹⁴⁷ which builds lollipop plots by mapping inserted mutations on protein sequences of the isoforms functional annotated in Pfam database¹⁴⁸.

5.3. Third-party software tools

Burrows-Wheeler Aligner (BWA)

BWA⁸⁶ is a free fast and accurate aligner of relatively short nucleotide reads against long reference sequence such human genome. It exploits Burrows-Wheeler transform (BWT) to index reference genome and accelerating alignment process. BWA aligns both single-end and paired-end reads.

Picard

Picard (<https://broadinstitute.github.io/picard/>) is a set of tools for manipulating NGS data and file formats such as BAM, SAM and VCF.

SAMtools

SAMtools⁸⁷ is a set of utilities for post-processing alignments in the SAM and BAM formats. It is common used to index SAM/BAM files, view alignments, and call variants.

Genome Analysis Toolkit (GATK)

GATK⁹² is a large open source of software package that are useful to make both germline and somatic genome analysis.

MuTect

MuTect¹⁰³ is a SNPs caller for reliable and accurate detection somatic variants at low frequencies in cancer samples.

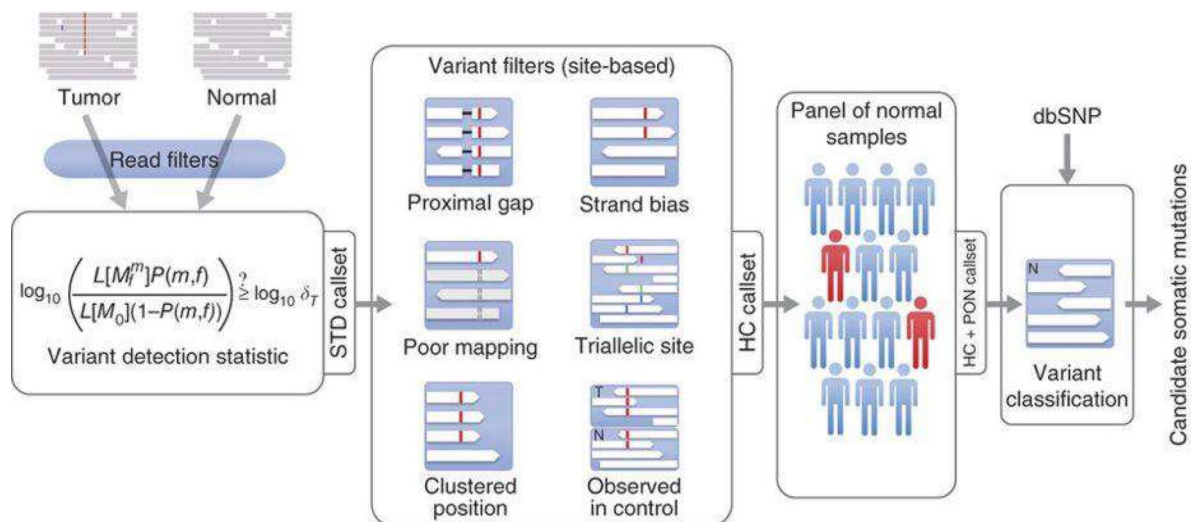


Figure 12. Diagram of the detection of somatic variants using MuTect. *MuTect filters tumor and normal reads and evaluates if each variant is more detected than expected random sequencing errors, then, candidate variants have to pass six filters considering variant site features; optionally, variants are filtered by using a panel of control samples to remove both germline events and artefacts; finally, determination if each variant is somatic or germline is achieved¹⁰³.*

This method consists of four main steps (**Figure 12**). First, the aligned reads with too many mismatches or low-quality scores are removed in the tumor and control samples. Second, a statistical analysis to detect variants in tumor sample is performed by using a Bayesian classifier. Third, candidate variants must pass six filters where variants near gap, triallelic sites, variants observed in a single direction of reads, clustered variants, variants observed in matched control sample beyond a fixed number of times, and variants caused by sequence similarity in the genome. Optionally, artifacts and germline variants are filtered by using a panel of normal samples sequenced and analyzed with the same workflow of considered cohort. Finally, a second Bayesian classifier is used to evaluate the somatic or germline status for each candidate variant.

MuTect2

MuTect2¹⁰⁴ is part of GATK4 and exploits original algorithm of MuTect with the assembly-based algorithm of HaplotypeCaller, a GATK tool for germline variant calling, to detect somatic SNPs and indels.

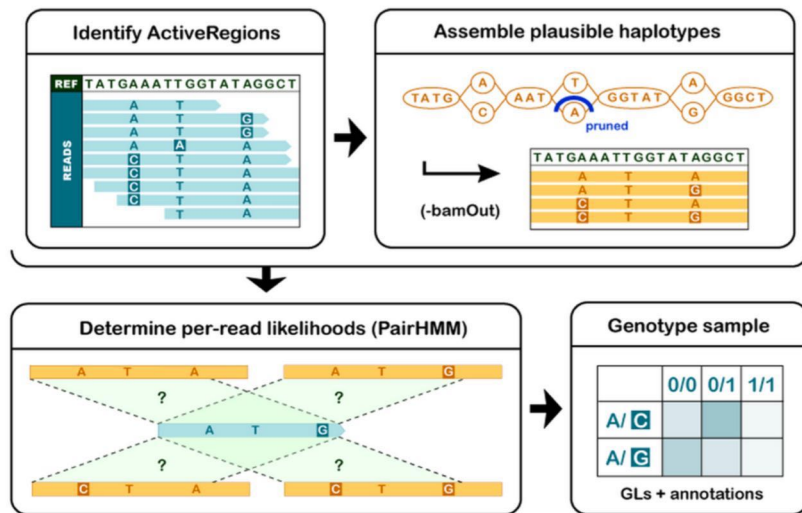


Figure 13. Diagram of the detection of variants through assembly-based algorithm. Initially, active regions are defined, and their re-assembly is performed to determine haplotypes. For each active region is calculated the likelihood of haplotypes given the read data. Finally, it assigns the most likely genotypes to the sample. Figure from <https://software.broadinstitute.org/gatk/documentation/presentations>.

Variant calling is divided in four steps (**Figure 13**). First, the program defines which regions of the genome have significant evidence of variation (ActiveRegions), using original MuTect somatic statistics methods. Second, for each ActiveRegion, the program constructs a De Bruijn-like graph to reassemble the ActiveRegion and generate a list of a possible haplotypes. Then, MuTect2 realigns each haplotype against the reference¹⁴⁹ in order to detect potentially variants. Third, for each ActiveRegion, MuTect2 performs a pairwise alignment of each individual read against each haplotype in turn including the reference haplotype using a pairHMM algorithm¹⁵⁰, which consider information about base quality scores and indel quality scores. At the end of this process, a score for each read-haplotype pair is provided, representing the probability of observing that read given that haplotype. These scores are then marginalized to calculate how much evidence there is for individual alleles at the candidate sites that were identified in the previous step. A table is produced containing per-read allele likelihoods for each candidate variant site under consideration. Finally, MuTect2 evaluates the likelihoods in aggregate to determine what is the most likely genotype of the sample at each site, by applying the Bayes' theorem. In this selection, it also considers somatic status of variants in a similar way used by MuTect. In addition, MuTect2 allows for varying allelic fraction for each variant taking in consideration the low purity of tumor samples, the cancer heterogeneity and possible presence of copy number variations.

Strelka2

Strelka2¹⁰⁵ is a software optimized for germline and somatic variants detection. The algorithm to call somatic variants is based on prior version Strelka¹⁵¹. Somatic calling model is based on a Bayesian approach accounting for any level of allele frequency variation in the tumor sample without requiring an estimate of tumor purity. Candidate regions with indels and read realignment is performed to improve accuracy for indel detection (**Figure 14**).

Strelka2 is improved respect to Strelka in the somatic-variant probability model that considers for tumor-cell contamination in the matched-normal sample, improving somatic recall for liquid tumor analysis and not requiring information about normal sample contamination. The probability model is extended by a final empirical variant scoring (EVS) step, inspired in part by machine-learning-based variant classification approaches^{152,153}. It uses a random forest model trained on call-quality features to improve precision by considering for errors that are not adequately represented in the generative variant probability model¹⁰⁵. The EVS model for somatic variants is pretrained on data of curated tumor-cell lines considered as somatic truth sets. Finally, for each variants an aggregate score is assigned, which will be used for following prioritization steps.

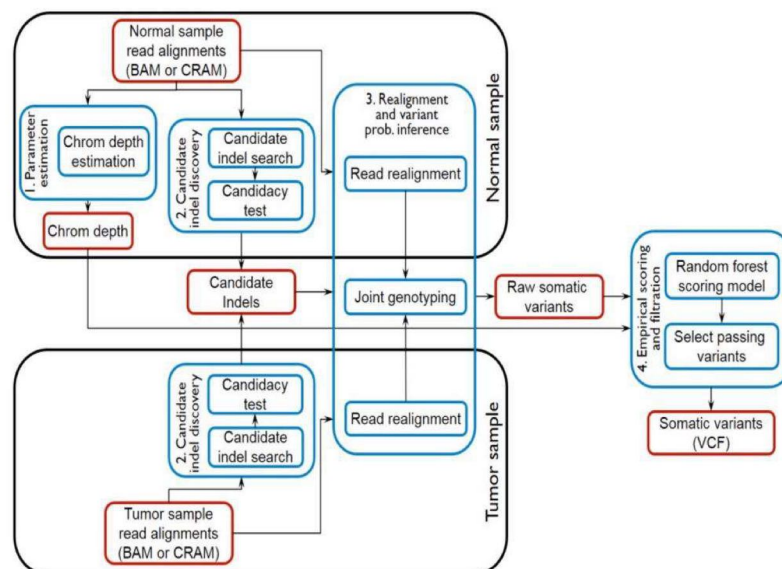


Figure 14. Strelka2 variant calling workflows. Somatic variants from matched tumor-normal sample pairs are detected acting several steps, including: (1) parameter estimation, (2) candidate variant discovery, (3) realignment and variant probability inference, and (4) empirical scoring and filtration. Figure adapted from Kim S., et al., 2018¹⁰⁵.

VarScan2

VarScan2⁹⁹ algorithm applies heuristic methods and a statistical test to detect somatic and germline SNPs, indels, and CNVs by exome sequence data from matched tumor and control

samples. In order to detect SNPs and indels, it simultaneously compares reads data from tumor and control samples for each position. VarScan2 evaluates for each position if both samples have a minimum coverage and determines a genotype for each sample individually based on reads observed. If genotype do not match between matched tumor and control samples, their reads count is evaluated by one-tailed Fisher's exact test comparing the number of reads supporting reference and variant between the two matched samples. If the resulting p-value is statistically significant, then the variant is assigned somatic status. In contrast, if p-value is not significant or genotypes of matched samples match, then the variant is considered as germline. When genotypes match, a one-tailed Fisher's exact test is used to compare the total number of reads supporting variant and reference (tumor and control are combined) to the expected distribution for a variant position due to sequencing error. Finally, VarScan2 categorizes detected variants as high confident (HC) or low confident (LC) based on allele frequencies in tumor and control samples, and p-value previously calculated. Positions that are homozygous in normal but heterozygous in tumor or where the variant allele is not the same are presumed to be sequencing errors or artifacts and are removed.

Torrent Variant Caller (TVC)

Torrent Variant Caller (<https://github.com/domibel/IonTorrent-VariantCaller>) is a plugin of Torrent Suite to call variants in optimized way for IonTorrent data attempting to address the high rate of false positive indels introduced by IonTorrent technology. It has an option to call somatic variants where, prior to any statistical calculation, a check of the variants is carried out. If in normal sample the same position of tumor allele is not sequenced, has a low coverage or has an allele frequency ranging from 0% to 10%, the variant is labelled as non-confident. For each residual variant is estimated the likelihood that mutated allele in tumor sample is not present in the control sample by building a binomial model (Poisson approximation). In case tumor allele is seen above error rate in control sample, the variant is marked as germline.

SnpEff and SnpSift

SnpEff¹¹ is a bioinformatics tool for rapidly annotating genomic variants (SNPs, indels and MNPs) and predicting variant impact considering information about genomic location and changes induced on coded protein. The impact of variant can be defined, in order of deleterious effect, as HIGH, MODERATE, LOW, or MODIFIER (**Table 3**).

Impact	Description	Type of variant
HIGH	The variant is assumed to have disruptive effect in the protein	Large structural variants, deletion/gain/duplication of exons, frameshift variants, gene fusion or rearrangement, protein-protein interaction variants, protein structural interaction variants, splice acceptor/donor site variants, stop lost/gained
MODERATE	A non-disruptive variant that might change protein functions	Missense variants, in-frame insertions/deletions, gene duplications, variants affecting splicing branch point from U12 splicing machinery, 5' or 3' UTR deletions
LOW	Assumed to be mostly harmless or unlikely to change protein functions	Synonymous variants, synonymous starts/stops, premature start codon variants in 5' UTR, splice region variants
MODIFIER	Usually non-coding variants where predictions are difficult or there is no evidence of impact	Downstream/upstream gene variants, intergenic variants, intron variants, miRNA variants, regulatory region variants, 5' or 3' UTR variants

Table 3. Type of variants associated to predicted impact by SnpEff.

SnpSift¹⁵⁴ is a toolbox to annotate variants with external databases of interest, to filter and manipulate annotated files.

5.4. Bioinformatics tools for pathway-derived network construction and analysis

Graphite

Graphite¹⁵⁵ is an R package of Bioconductor that allows for pathways analysis originally considering gene expression data from both microarray and RNA-sequencing. In addition, Graphite converts the complex pathway structures, including different types of direct and indirect gene and gene products relations – as regulatory relations, molecular complexes and biosynthetic pathways with compound intermediates, to cite only a few - into path-derived gene networks, solving complexity with appropriate biology-driven rules. There is also a web tool of Graphite but less updated (<https://graphiteweb.bio.unipd.it/>)¹⁵⁶.

HotNet2

HotNet2¹⁴⁵ is a tool written in Python which exploits an insulated heat diffusion model and applies a permutation test in order to detect significant mutated subnetworks in large and complex protein-protein interaction networks.

Cytoscape

Cytoscape¹⁵⁷ is an open source software for visualizing molecular interaction networks and integrating with high-throughput data. The software is extensible with plugins to perform additional analysis on networks, and to provide new layouts, additional file format support, connection with databases, and other features. The Cytoscape plugin community development is constantly evolving providing always more features.

5.5. Large Granular Lymphocyte Leukemia

WES data

The Hematology Research Unit of the Department of Clinical Chemistry and Hematology of University of Helsinki directed by Prof. Satu Mustjoki provided data, expertise, and performed variant validation and functional studies based on results obtained by bioinformatics analysis presented in this thesis.

LGL leukemia patients (n=19) (**Table 4**) were diagnosed based on the WHO 2008 guidelines. The samples were collected from Helsinki University Central Hospital (n=15), Cleveland Clinic (n=3) and Cologne University Hospital (n=1). All studies were conducted in accordance of the principles of the Helsinki declaration and were approved by the Helsinki University Central Hospital, Cleveland Clinic and Cologne University Hospital Ethics Committees. Written informed consents were obtained from all patients prior to sample collection.

Patient number	T/N K LGL	Sex	Age at diagnosis	Leukocytes (x10E9/L)	Lymphocytes (x10E9/L)	Vbeta expansion	Associated disorders	Treatment
246	CD8+	M	70	10.9	9.5	Vb.16: 94%	Anemia, neutropenia	None
274	CD8+	F	69	7.2	5.8	Vb.7.1: 28%	Neutropenia	Methotrexate, cyclosporine A, prednison
1352	CD8+	M	40	12.6	6.7	Vb 13.2: 69%	None	None
1149	CD8+	M	77	10.8	9.4	Vb 20: 73%	Neutropenia, anemia, MGUS	None
1147	CD8+	F	72	11.8	7	Vb 22: 91%	Collagenosis	None
1148	CD8+	F	50	16.1	12.9	Vb 17: 94%	None	None
1403	CD8+	M	59	12.7	3.7	Vb.3: 89%	None	None
2803	CD8+	M	66	5.4	4.5	Vb.5.1: 65%	Anemia, neutropenia, thrombocytopenia	None
1527	CD8+	M	64	9.3	3.6	Vb 14: 31%	None	None
1255	CD8+	M	59	12.9	9.5	Vb. 20: 78%	MGUS, anemia	Cyclosporine A
1235	CD8+	M	60	9.5	7.9	Vb.17: 96%	Anemia, neutropenia, hypergamma-globulinemia	Methotrexate, cyclosporine A
1256	CD8+	F	74	5.4	4.5	Na	Anemia, neutropenia	Cytosan, cyclosporine A
1265	CD8+	M	69			Na	None	None
1254	CD4+CD8+	M	61	15.9	8.5	Vb.13.1: 98%	None	None
1526	CD4+	F	70	10.2	8.4	Vb.8: 86%	None	None
1525	CD4+	F	80	8.7	5.0	Vb 13.1:77%	None	None
1272	NK	F	46	7.7	6.9	Na	Anemia, neutropenia, gamma heavy chain	Methotrexate
1253	NK	F	58	5.0	4.3	Na	Linear IgA disease, hypergamma-globulinemia	None
1260	NK	F	47	11.4	8.9	Na	Neutropenia	None

Table 4. Clinical features and data of the studied LGL leukemia patients.

Exome sequencing was performed from sorted T or NK cell fractions as tumor cells by using the Illumina GAII instrument as 82 base pairs paired end reads. In CD8+ T-LGL leukemias, polyclonal CD4+ cells were used as a germline control and in CD4+ T-LGL leukemias CD8+ cells were used as a germline control. For NK LGL cases germline DNA was extracted from CD3+ T-cells.

Identification of somatic variants from LGL-L tumors

Variants from whole-exome sequencing were called using a custom bioinformatics pipeline (Figure 15) developed based on the already available and in-house made scripts.

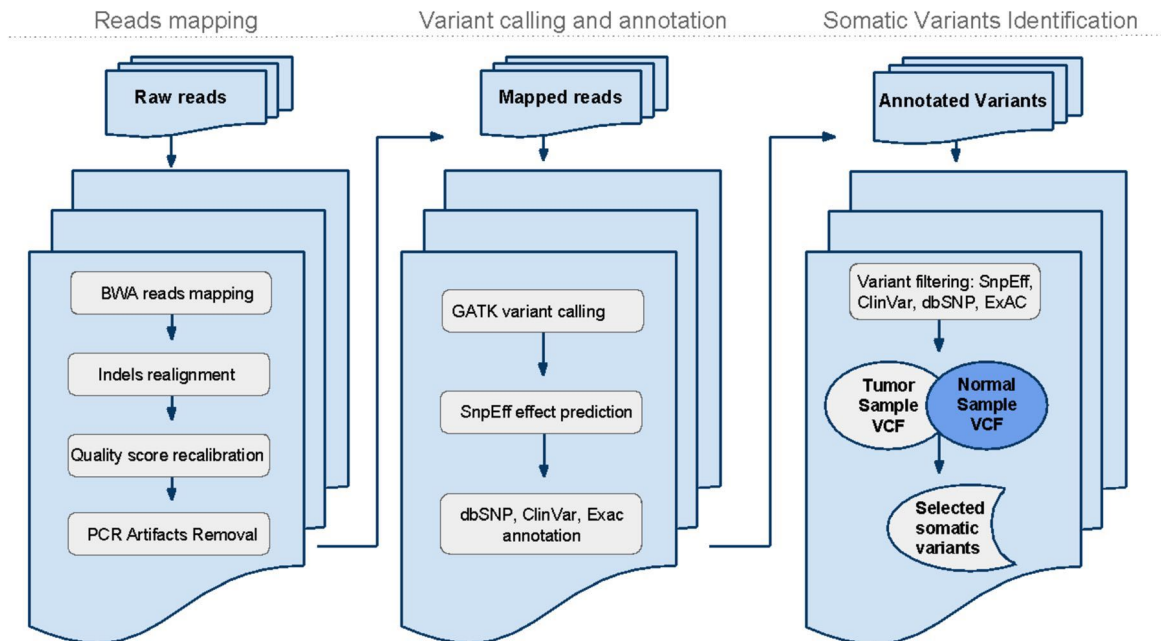


Figure 15. Bioinformatic pipeline used for somatic variants identification in LGLL study. All steps were connected together with SCons software building tool.

Paired-end reads were aligned to human genome version hg19 using BWA⁸⁶ with default parameters. Samtools (<http://www.htslib.org>) and GATK Genome Analysis Tool Kit⁹² were used to process and improve the BAM files generated during alignment to the reference genome phase. To reduce the risk of false positive variant calling, reads mapping to multiple genomic positions were removed from the alignment. Putative PCR duplicates, produced by over-amplification of the same DNA fragment, identifiable by multiple read pairs aligned exactly to the same genomic position, were removed using the *rmDup* command from Samtools package. As BWA aligns each read independently, once the primary mapping is produced, a local realignment around indels was carried out to improve the alignment considering the information from all reads mapping to the region. This step was carried out by the *RealignerTargetCreator* and *IndelRealigner* commands from GATK suite. To improve reads base quality scores accuracy using all reads aligned the considered region, we performed base quality score recalibration using the *BaseRecalibrator* command from GATK. Variant calling and filter tagging were performed by *UnifiedGenotyper* and *VariantFiltration* from the GATK suite⁹². In particular variants with coverage under 30 were tagged as “low coverage”, variants

with quality less than 30 were tagged as “low quality”. Moreover, variants with a FS score under 60 were tagged as “Strand Bias”. Finally, if more than 3 variants were found in a genomic window of 10 nucleotides width, these variants were tagged as “SNP cluster”. All tagged variants were subsequently removed leaving only high confidence variants for further analyses. Variants annotation was performed by SnpSift¹⁵⁴ and SnpEff¹¹¹, in particular *annotate* command of SnpSift was used to associate dbSNP¹¹⁸ identifier while SnpEff *eff* command was used to add functional and putative impact annotation to variants. Then, only variants with MODERATE or HIGH SnpEff predicted impact were retained. Variants with a dbSNP id having benign or likely benign clinical significance indicated by Clinvar¹²⁶ were discarded. Finally, variants with ExAC¹²² allele frequencies more than 5% were removed. Putative somatic variants were identified by subtracting variants found in normal sample from the set of variants found in the matched tumor sample.

Pathway-derived gene meta-networks construction

Mutated genes were mapped, in parallel, to networks derived from KEGG¹³⁶ and from Reactome¹³⁷ pathways using Graphite webtool (<https://graphiteweb.bio.unipd.it/>)¹⁵⁶. From path-derived networks obtained by Graphite including at least one mutated gene each, gene-gene relations were extracted and merged to build-up a non-redundant KEGG-Reactome union network. We considered two types of interactions with different strength: direct interactions according to pathway-topology and participation to the same pathway. Network visualization, optimization, annotation and analysis were performed using Cytoscape v3.1¹⁵⁷.

Validations of selected somatic variants

Selected somatic variants were validated by Sanger sequencing of both control and tumor samples. Based on exome sequencing results, specific primers were designed for candidate mutations using the Primer Blast search (<http://blast.ncbi.nlm.nih.gov/>, National Center for Biotechnology Information, Bethesda, MD, USA).

5.6. Follicular lymphoma of the pediatric age

WES data

The Pediatric Onco-Hematology Research Unit of the Department Women’s and Children’s Health of University of Padova, directed by Prof. Lara Mussolin, provided WES data and

expertise, and conducted validation of variants detected by bioinformatics analysis presented in this thesis.

The study population included 21 patients with lymphoproliferative disorders belonging to the spectrum of follicular lymphoid neoplasms of the pediatric age (2008 WHO Classification of Tumours of the Haematopoietic and Lymphoid Tissues and its 2016 update). In compliance with the Helsinki Declaration, informed written consent was obtained from parents or legal guardians on behalf of the children enrolled in the study. The main clinical characteristics of the patients are listed in **Table 5**. Median age at diagnosis was 14.2 years (range: 4.0-17.7 years), with a male to female ratio of 10:1. All cases were originally diagnosed as primary nodal pediatric-type follicular lymphoma (18 cases) or primary testicular follicular lymphoma (3 cases) and all of them were centrally reviewed. The immunohistochemical profile disclosed positivity for pan-B cell markers, Bcl6 and CD10, with negative (19/21 cases) or weak positive (2/21 cases) Bcl2 in all the cases. The mean proliferation index (Ki-67 immunostaining) was greater than 40%. MUM1 immunostaining was negative in all tested cases (18/18). B-cell monoclonality was confirmed by polymerase chain reaction of the immunoglobulin genes, according to BIOMED-2 guidelines¹⁵⁸. As for the clinical management, 15/21 cases were treated with surgical excision alone, while 6/21 cases were also treated with 2 to 4 cycles of chemotherapy according to the AIEOP-LNH97 protocol. A single case experienced a second malignancy after 4 courses of chemotherapy. At present all patients are alive and in good clinical condition (mean follow-up: 41.7 months, range 8.3-124.8 months).

Patient ID	Gender	Age	Stage	Anatomic site	CD 10	Bcl 6	Ki 67	MUMI	Bcl 2	FISH BC L2	FISH BC L6	Tumor cells %	TNFRS F14 mutations	Treatment Protocol	Outcome
W1	M	10,9	I	peripheral lymph nodes	pos	pos	≥ 40 %	neg	neg	neg	neg	50	pos	complete resection, w&w	alive, CR
W2	M	15,8	II	peripheral lymph nodes	pos	pos	≥ 40 %	na	neg	na	na	30-40	pos	LNH974 courses	alive, CR
W3	M	17,4	I	peripheral lymph nodes	pos	pos	≥ 40 %	neg	neg	neg	neg	60-70	neg	complete resection, w&w	alive, CR
W6	M	5,0	I	testis	pos	pos	≥ 40 %	neg	neg	neg	neg	30	neg	complete resection, w&w	alive, CR
W7	M	16,7	I	peripheral lymph nodes	pos	pos	≥ 40 %	neg	neg	neg	neg	70	neg	LNH972 courses	alive, CR
W8	M	13,4	I	Ear-nose-throat	pos	pos	≥ 40 %	neg	neg	neg	neg	30-40	neg	LNH974 courses	alive, CR2
W10	M	6,7	I	testis	pos	pos	≥ 40 %	neg	neg	neg	neg	20-30	neg	complete resection, w&w	alive, CR
W13	M	14,2	I	peripheral lymph nodes	pos	pos	≥ 40 %	neg	neg	neg	neg	80	neg	complete resection, w&w	alive, CR
W14	F	14,5	I	peripheral lymph nodes	pos	pos	≥ 40 %	neg	weak	neg	neg	30-40	neg	complete resection, w&w	alive, CR
S1	M	7,8	I	peripheral lymph nodes	pos	pos	≥ 40 %	neg	neg	neg	neg	40	neg	complete resection, w&w	alive, CR
S2	M	15,0	I	peripheral lymph nodes	pos	pos	≥ 40 %	neg	neg	neg	neg	70	pos	complete resection, w&w	alive, CR
S3	M	14,8	II	peripheral lymph nodes	pos	pos	≥ 40 %	neg	neg	neg	neg	30-40	neg	LNH974 courses	alive, CR
S4	M	16,8	I	Ear-nose-throat	pos	pos	≥ 40 %	neg	neg	neg	neg	20	pos	complete resection, w&w	alive, CR
S5	M	5,0	I	testis	pos	pos	≥ 40 %	neg	neg	neg	neg	10-15	pos	complete resection, w&w	alive, CR
S6	M	15,4	I	Ear-nose-throat	pos	pos	≥ 40 %	neg	neg	neg	neg	40-50	neg	complete resection, w&w	alive, CR
S7	F	17,7	I	na	pos	pos	≥ 40 %	neg	neg	neg	neg	70	neg	complete resection, w&w	alive, CR
S8	M	17,1	I	peripheral lymph nodes	pos	pos	≥ 40 %	neg	neg	neg	neg	60	pos	complete resection, w&w	alive, CR

S9	M	5,9	II	peripheral lymph nodes	pos	pos	≥ 40 %	na	neg	na	na	80	neg	LNH97 4 courses	alive, CR
S10	M	10,1	I	peripheral lymph nodes	pos	pos	≥ 40 %	neg	weak	na	na	na	neg	complete resection, w&w	alive, CR
S11	M	4,0	I	peripheral lymph nodes	pos	pos	≥ 40 %	neg	neg	neg	neg	30	pos	complete resection, w&w	alive, CR
S12	M	4,5	I	Ear-nose-throat	pos	pos	≥ 40 %	na	neg	na	na	80	pos	LNH97 2 courses	alive, CR

Table 5. Clinical features of the 21 patients with follicular lymphoma of the pediatric age spectrum considered in the study. Abbreviations: M= male; F= female; na= not available; w&w= watch and wait; CR=first complete remission; CR2=second complete remission.

Purified genomic DNA from nine paired tumor samples and peripheral blood were outsourced for Whole Exome Sequencing (WES) analysis (Biodiversa srl, Rovereto (TN), Italy). Samples were enriched in protein coding sequences using the SureSelect Human All Exon V5 (Agilent Technologies, Santa Clara, CA), following the manufacturer's instructions. The resulting libraries were subjected to paired-end sequencing (2 x 150 bp) on an Illumina HiSeq4000 system, with a theoretical coverage of 150X for tumor samples and 100X for paired germline samples.

Somatic variants identification in PTNFL and PFLT patients

Reads were aligned to human genome version hg19 using BWA⁸⁶. Picard and GATK Genome Analysis Tool Kit⁹² were used to process the files generated during alignment to the reference genome and to remove reads mapping to multiple genomic positions and putative PCR duplicates in order to reduce the risk of false positive variant calling. *RealignerTargetCreator* and *IndelRealigner* commands from GATK suite were used to perform a local realignment around indels considering the information from all reads mapping to the region. To improve reads base quality scores accuracy using all reads aligned the considered region, base quality score recalibration was obtained using the *BaseRecalibrator* command from GATK. Somatic SNPs were detected by MuTect¹⁰³ and then annotated by SnpEff¹¹¹ and SnpSift¹⁵⁴. SnpEff predicted functional and putative impact of detected variants. SnpSift's *annotate* command provided the association of known variants to dbSNP (v. 150)¹²⁰ and COSMIC (v. 82)¹²⁷ identifiers, clinical significance from Clinvar (updated on 05/09/2017)¹²⁶. Known variants annotated in Clinvar as benign or likely benign were discarded and only variants with SnpEff predicted impact HIGH or MODERATE were further considered. Finally, variants with non-Finnish European population allele frequency >5% according to ExAC (Exome Aggregation Consortium; <http://exac.broadinstitute.org>)¹²² data were discarded (**Figure 16**).

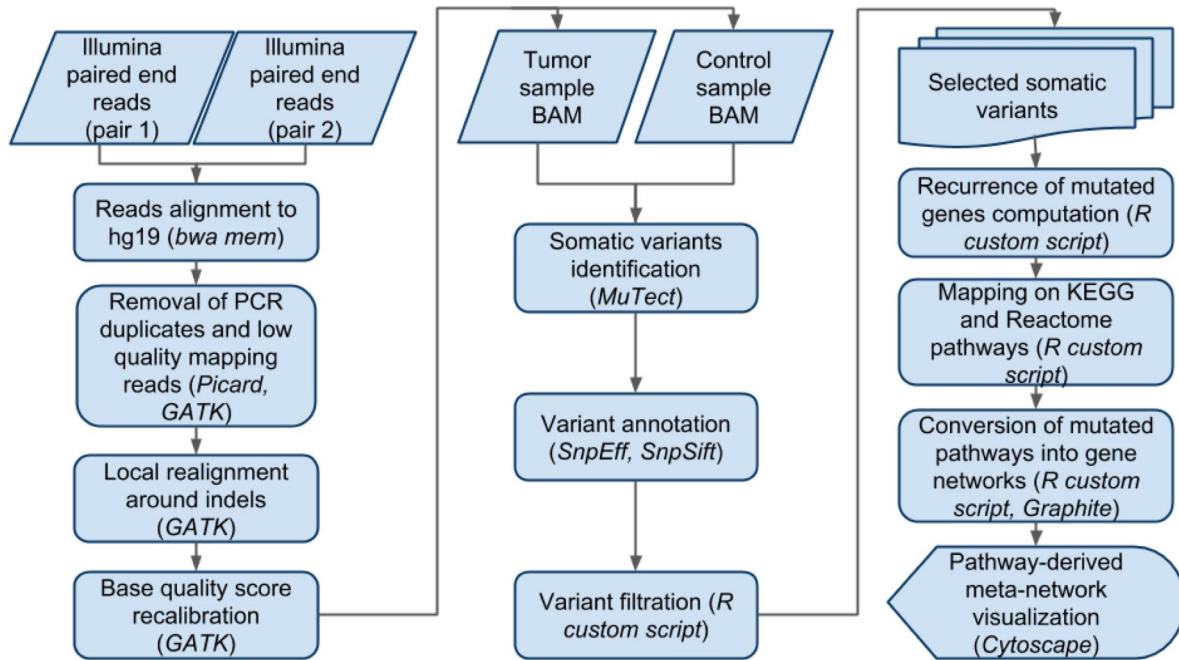


Figure 16. Bioinformatic workflow for detection of high confident somatic SNPs and reconstruction of a mutated pathway-derived network. Reads mapping, reads processing and variant calling followed by annotation were automated by using SCons software building tool.

Mutated gene network analysis

R Graphite Bioconductor package (v. 1.20.1)¹⁵⁵ was used to map and convert pathway topologies annotated in KEGG¹³⁶ and Reactome¹³⁷ into pathway-derived gene networks. Only pathways including at least one mutated gene were converted into networks and merged in order to build up a unique non-redundant KEGG-Reactome meta-network of 299,387 pairwise interactions between 10,672 genes. Two types of interactions were considered with different relevance: direct connections according to pathway-topology and participation to same pathway. Visualization, optimization and annotation of the network were performed using Cytoscape v3.5.1¹⁵⁷(**Figure 16**).

Validations of somatic variants and TNFRSF14 exon1 investigation

Selected somatic variants with variant allele frequency (VAF) >15% were validated by direct Sanger sequencing of both control and tumor samples. Based on exome sequencing results, specific primers were designed for candidate mutations by using Primer Express 3.0 (Applied Biosystems, Foster City, CA). PCR products of variants with VAF < 15% were cloned into TOPO TA cloning vector (Invitrogen, Carlsbad, CA) and at least 20 colonies were sequenced. Mutations on exon 1 of TNFRSF14 were evaluated in an extended series of other 12 patients

(11 PTNFL and one PFLT), either by direct sequencing or PCR cloning and sequencing of at least 20 colonies.

5.7. High-Risk Neuroblastoma

WES data

This study was conducted in collaboration with Neuroblastoma Laboratory of Fondazione Istituto di Ricerca Pediatrica Città della Speranza di Padova conducted by Prof. Gian Paolo Tonini, that provided data, expertise, and performed validation of results obtained by bioinformatics analysis presented in this thesis.

A cohort of stage M NB patients from the Italian Neuroblastoma Registry with complete clinical data and follow-up over 10 years was considered. Frozen tissue from the primary tumor at onset was available for each patient. Patients were stratified into two groups according to their overall survival: the SS group (n = 14), including patients with rapid disease progression and rapid fatal outcome, all with a survival time < 60 months, and the LS group (n = 15), including patients who are responsive to therapy and survived at least 60 months from diagnosis (**Table 6**). Five SS patients (ID2475; ID2368; ID2181; ID1995; ID2100) were also included in the previous NB report¹⁵⁹. Out of the 15 LS patients, 11 were still in complete remission, two were alive with disease and two died of disease at last follow-up. Informed consent was received for the use of biological material from legal tutors, and the study was approved by the Institutional Board of the participating Institutions. All tumor samples were classified as NB Schwannian stroma-poor according to criteria established by the International Neuroblastoma Pathology Committee¹⁶⁰. The presence of at least 60% of neuroblasts in tumor samples was verified.

PATIENT ID	SEX	AGE AT DIAGNOSIS (months)	MYCN STATUS	DNA INDEX	SURVIVAL (months)	OUTCOME	GROUP
2475	M	208	Gain	No data	33	DOD	SS
1965	M	83	Not amplified	1.51	34	DOD	
1955	F	77	Not amplified	No data	6	DOT	
2368	M	75	Not amplified	No data	48	DOD	
3060	F	118	Unknown	No data	33	DOD	
1900	M	37	Not amplified	1.14	24	DOD	
2181	M	47	Amplified	No data	28	DOD	
2384	M	58	Gain	No data	45	DOD	
1995	M	22	Amplified	2.37	23	DOD	
1920	M	14	Not amplified	No data	9	DOT	
2100	M	27	Not amplified	No data	12	DOD	
2513	M	52	Not amplified	No data	20	DOD	
2852	M	50	Gain	No data	43	DOD	
2578	F	23	Gain	1.07	42	DOD	
1409	F	34	Not amplified	1.00	159	CR	
1641	M	33	Not amplified	1.96	105	CR	
2121	M	61	Not amplified	1.88	144	CR	
2393	F	73	Gain	No data	62	CR	
2140	M	55	Not amplified	1.00	75	CR	
1905	M	15	Not amplified	1.96	80	CR	
2528	M	61	Gain	No data	86	CR	
2035	M	17	Not amplified	1.52	144	CR	
2488	M	47	Not amplified	No data	65	AWD	
2951	M	68	Gain	1.00	64	DOD	
2251	F	12	Amplified	No data	110	CR	
2576	F	32	Not amplified	No data	71	AWD	

2426	F	7	Not amplified	No data	53	CR
2613	F	12	Not amplified	No data	39	CR
2828	M	8	Not amplified	1.92	71	CR

Table 6. NB patient cohort description. Abbreviations: M, male; F, Female; Unknown, Physician did not have data; No data, data was not made available; DOD, Dead of disease; DOT, Dead of toxicity of the treatment; AWD, Alive with disease; CR, Complete remission

Exome sequencing for each matched tumor-control sample was performed by Ion Proton Sequencing.

Identification of somatic variants from HR-NB patients

Read mapping and variant calling were performed with Torrent Suite and Ion Reporter™ software, provided by the Ion Proton™ System. The Proton Run Browser was used for quality control metrics (percent bead loading, usable sequences, read length, alignment metrics to hg19 reference genome and mean raw accuracy). The samples were processed using the workflow: “Somatic – Proton – High Stringency Configuration”. Bam files of the tumor and blood samples of each patient were uploaded to Ion Reporter™ (IR) software using the available plug-in, IonReporterUploader_V1_2. Variant calling was done using Torrent Variant Caller (v. 5.0–9). Next, the files were processed using a workflow AmpliSeq Exome paired sample (tumor/normal) to subtract variants [SNV, multiple nucleotide variant (MNV), indel and copy number variant (CNV)] discovered in the peripheral blood DNA against the tumor DNA. The annotation of somatic variants was performed by SnpSift¹⁵⁴ and SnpEff¹¹¹. SnpSift’s *annotate* command provided the association of known variants to dbSNP (v. 147)¹²⁰ and COSMIC (v. 77)¹²⁷ identifiers, clinical significance from Clinvar (updated on May 02, 2016)¹²⁶, as well as functional prediction indicated by MetaSVM and MetaLR¹¹⁹ from dbNSFP database (v. 2.9.1)¹⁴⁷. The two algorithms predict whether the variant is tolerated or deleterious, considering nine scores present in dbNSFP (SIFT, Polyphen – 2, GERP ++, MutationTaster, MutationAssessor, FATHMM, LRT, SiPhy and PhyloP) and MMAF observed in different populations of 1,000 genomes. SnpEff predicts the functional and putative impact of detected variants. Known variants annotated in Clinvar as benign or likely benign were discarded, and only variants with HIGH or MODERATE SnpEff predicted impact were further considered. After integration of exome sequencing and Exome Aggregation Consortium (ExAC, <http://exac.broadinstitute.org>)¹²² data, variants with a non-Finnish European population allele frequency > 5% were discarded (**Figure 17**). The interpretation of variants’ impact was also

obtained by mapping selected variants to protein sequences and their domain annotation using MutationMapper (http://www.cbioportal.org/mutation_mapper.jsp)¹⁶¹.

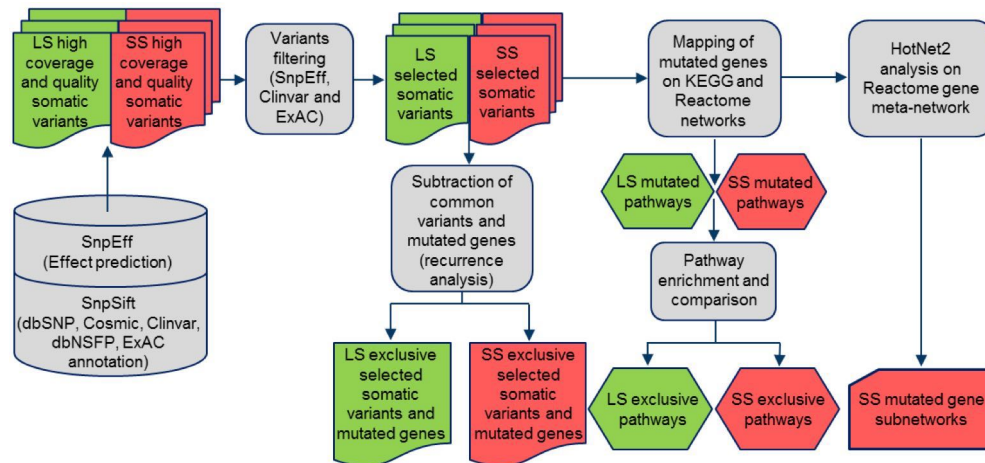


Figure 17. Bioinformatic pipeline and study design used for somatic variants identification in HR-NB patients. The flowchart summarizes the workflow followed for the identification of group-exclusive somatic variants and mutated pathways; from these lists we chose the most promising variants for validation with Sanger sequencing and Roche 454.

Pathway analysis

Genes mutated in SS and in LS were separately mapped to the KEGG¹³⁶ and Reactome¹³⁷ pathways. The pathways with at least three genes mutated in a group and none in the other group were defined as “group-specific pathways.” Significant pathway enrichment was calculated considering the separately mutated genes in the SS and LS patient groups. Significantly enriched pathways only in one group and with a number of genes mutated in the group by at least 1.5x the number of genes mutated in the other group were considered to be “group-specifically enriched pathways.” The detected pathways were organized after the architecture of the KEGG and Reactome databases to have a less redundant description of altered molecular signaling and biological functions, gathering pathways into functional classes.

Gene network analysis

The R Graphite Bioconductor package (v. 1.20.1)¹⁵⁵ was used to convert complex pathway topologies into Reactome pathway-derived gene networks. Reactome networks were merged into a pathway-derived gene network of 186,808 pairwise interactions between 8,678 genes. The HotNet2 algorithm was applied to Reactome-derived gene networks to statistically identify group-specific subnetworks of mutated protein-coding genes, defining groups of functionally

related genes in which mutations significantly converge (**Figure 17**). Hotnet2 analysis was first conducted using all the somatically mutated genes in the whole cohort of patients, and then the two patient groups were considered separately. Due to the cardinality of our cohort and the considerable dimension of the considered network, multiple testing correction applied in this analysis, considerably increases the p-value.

Ultra-deep sequencing for variants validation

Validation of tumor variants was performed by ultra-deep sequencing on Amplicon libraries (from 200 to 400 bp) using the 454 Junior Titanium sequencer (Roche). Target regions of the genome reference sequences corresponding to the amplicons were obtained from the human reference genome (GRCh37/hg19) using the *getfasta* command of BEDTools¹⁶². Reads were mapped to these sequences through the *bwa-sw* command of Burrows-Wheeler Aligner¹⁶³ and variants were called using GATK – Genome Analysis Tool Kit⁹². Variants were also confirmed using IGV – Integrated Genome Viewer¹⁶⁴.

Independent cohort of HR-NB

To confirm results, we analyzed the largest available group of stage M NB with survival data profiled by WES (Pugh cohort)¹⁶⁵. The 4,120 genes with non-silent somatic mutations reported in the Pugh cohort were analyzed, separating the 240 patients into SS (221; with overall survival ≤ 5 years) and LS (19; with overall survival >5 years) according to survival classification.

6. Results

6.1. iWhale: a computational pipeline for cancer exome sequencing analyses

A computational pipeline for the automatic analyses of cancer whole-exome sequencing data has been developed (**Figure 18**). It starts from the pairs of cancer and control samples sequencing data and produces VCFs with annotated cancer mutations of high impact. The pipeline is built using Docker and SCons that glue together all bioinformatic tools used by the analyses. Docker allows to easily download and install the pipeline in any computer (Linux, Windows, macOS) and run the analysis independently from the used computer. SCons allows to split the analyses by steps that once they have run, any stop, like killing by error the process, or even shutting down the computer will automatically recover from the last run point.

At the moment four variant calling software are used by the pipeline: Mutect¹⁰³, Mutect2¹⁰⁴, VarScan2⁹⁹, and Strelka2¹⁰⁵, but in the future more ones will be added. The user can choose which programs to use and change the default settings.

From the biologically point of view, the pipeline uses publicly available databases to filter out mutations with no effects or mark as important other ones in the final VCFs produced. dbSNP database¹²⁰, gnomeAD genome aggregation database¹²², Cosmic database¹²⁷, Clinvar¹²⁶ and Cancer Gene Census (CGC)³³ databases.

Once published, the computational pipeline will be made publicly available in Github (github.com) for free download. The data needed by the pipeline to run, will be automatically downloaded and installed by using Ansible; a radically simple IT automation platform that makes your systems easier to deploy and use.

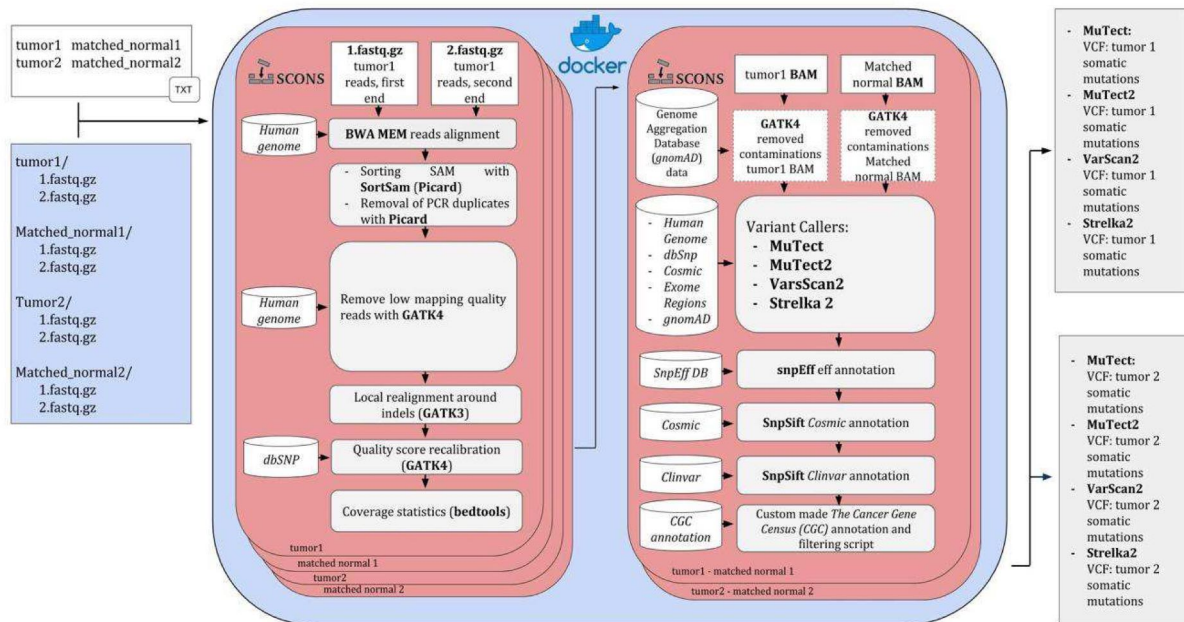


Figure 18. Diagram of iWhale. The pipeline automatically recognizes paired-end FASTQ files for each sample to start analysis with alignment phase. Once alignment for each sample is completed, iWhale automatically recognizes BAM files of tumor-control matched samples to perform variant calling by using four different tools. Finally, called variants are annotated by exploiting information of different databases.

Other pipelines have been developed for WES analysis, such as SeqMule¹⁶⁶, Fastq2vcf¹⁶⁷, or IMPACT¹⁶⁸. At the moment, all the pipelines for WES data are developed with programming languages that require solid computer skills making the analysis non-user-friendly. Other pipelines are more user-friendly but are based on web platforms, such as Galaxy¹⁶⁹, where there are space and time limitations. Thanks to Docker, iWhale is more user-friendly and can be used in any operating system not constraining the use of Linux ambient. In addition, iWhale requires only one text file organized in two columns and containing tumor-control matched samples. Through a one more optional text file including all parameters to change, iWhale allows for custom WES analysis, considering that cancer studies often require specific precautions to obtain optimal results. Overall, iWhale makes easy a complex and articulated analysis of WES data from cancer samples.

6.2. Pathway-derived meta-network of mutated genes

A systems genetics approach has been developed to construct pathway-derived meta-network depicting direct and functional interactions between mutated genes (**Figure 19**). The tool exploits Graphite R package¹⁵⁵ to map in parallel a list of mutated genes on KEGG¹³⁶ and Reactome¹³⁷ pathways, and only pathways carrying at least one mutated gene are converted into gene networks. Finally, all mutated pathway-derived networks are merged into a unique

non-redundant meta-network. The tool generates a tab-separated text file containing for each line the symbols of genes participating to interaction and functional annotation about interaction, such as direction, type of interaction and pathways where is involved. The text file is loaded into Cytoscape to obtain a visualization of the generated meta-network.

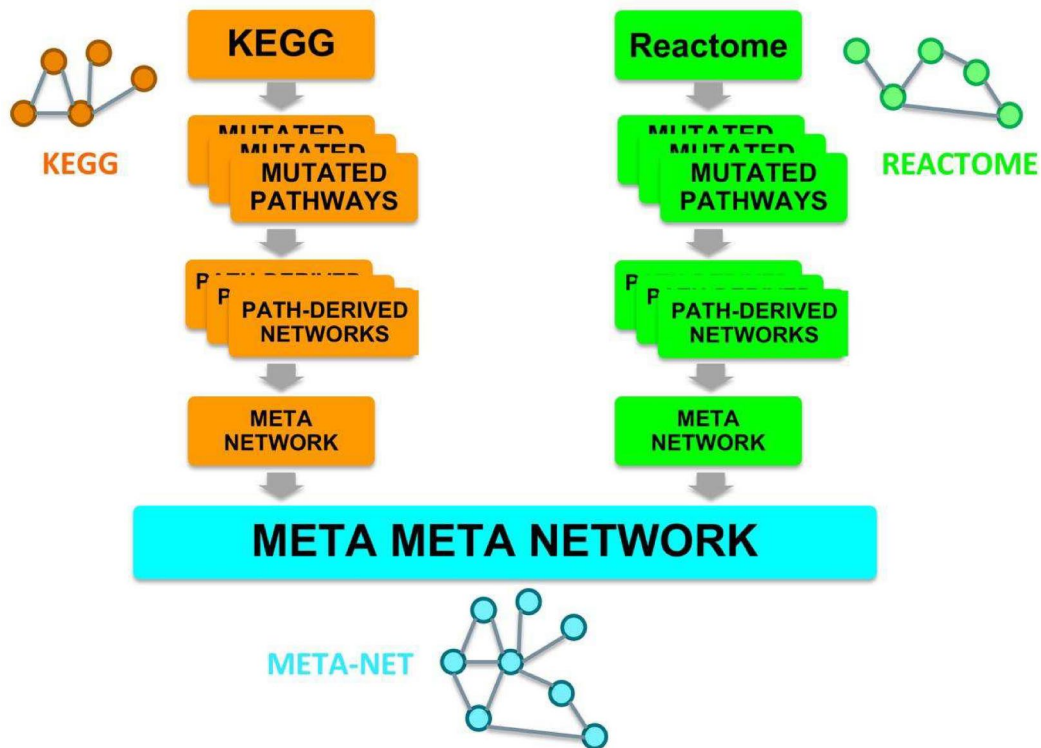


Figure 19. Diagram of construction of pathway-derived meta-network. *The analysis starts in parallel for KEGG (orange) and Reactome (green) pathways. Mutated pathways are transformed into gene networks that, finally, are overlapped to obtain a meta-network composed by unique gene-gene interactions.*

This method is a gene network-based approach which allows to obtain topological information about mutated gene verified interactions. Qualitative evaluation can be performed on the network to identify altered biological functions and infer how they can interact leading to malignant transformation. Moreover, the identification of mutated pathways and aberrant biological functions are simplified respect to protein-protein interactions network where interactions are not functional annotated. The last concept is an advantage of this approach but also a limitation. Indeed, pathway-derived network analysis are strongly based on prior knowledge and suffer of well-studied pathways bias. Many mutated genes detected by WES are not gathered in pathway databases leading to a limited power discovery.

Another complication is that often the resulting networks from cancer studies are very large and complex making challenging or impossible to infer any evaluation. For this reason, quantitative analysis is fundamental to interpret a large gene network.

HotNet2¹⁴⁵ is a bioinformatics tools that detects statistically mutated subnetworks analyzing large networks. HotNet2 has been originally developed for analysis of protein-protein (PPI) interaction networks for which pre-computed ready-to-use data are provided. Since the versions of PPI networks of HotNet2 are outdated and they are challenging to functionally annotate, I setup a custom updated data structure to analyze more biologically informative Reactome-derived network.

HotNet2 algorithm is based on insulated heat diffusion model where each node of a network diffuses an amount of own heat to its immediate neighbors and to the rest of the network. The amount of heat that spreads from each node is strongly dependent by the initial heat, the local topology of the network, and the fraction β of heat retained by each node. Therefore, β parameter balances the heat diffusion across the network and is particularly network-specific. The best β value is when a source node concentrates the most of its heat into its neighbors without spreading across the rest of the network. To determine best β value, I generated 20 different diffusion matrices of Reactome-derived network using β values from 1 to 0.05 (**Figure 20**), contained in Hierarchical Data Format (HDF5, which is a file format for storing huge amounts of numerical data hierarchically organized). A heat diffusion matrix contains the values of heat that diffused along an edge linking two nodes at equilibrium. This is an estimate of the influence that a source node has on a destination node, given a determined β value.

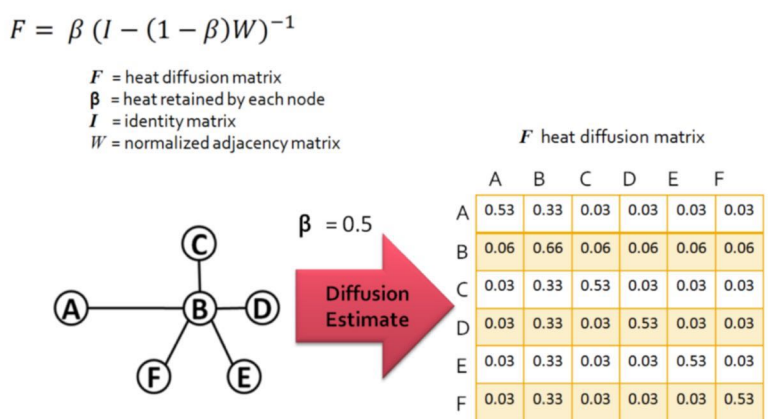


Figure 20. Example of heat diffusion matrix of HotNet2. The value of heat on each node at equilibrium is calculated by a matrix-based equation. A network of 6 nodes is modeled as a normalized adjacency matrix where information about edges are stored. If edge exist between two nodes a value of 1 is assigned, otherwise the value is 0. An identity matrix is used to keep track of nodes within a network, assigning 1 to considered node and 0 to the rest of the matrix. Performing the steps in the equation using the selected β parameter, an insulated heat diffusion matrix is generated. Figure from Yan, Melissa, "Implementation of the HotNet2 Network Diffusion-based Analysis Method in Java" (2016). Scholar Archive. 3860. <https://digitalcommons.ohsu.edu/etd/3860>.

As β decreases from 1, the amount of heat on neighbors increases as less and less heat is retained in the source node¹⁴⁵. Then, I observed how the heat of highly connected genes, such

as TP53, spreads to the direct neighbors and to the rest of the network (**Figure 21**). Diminishing β , there was an inflection point at which the amount of heat on neighbors decreased, as more and more heat diffused towards the rest of the network. The smallest β before this inflection point was chosen as the best value.

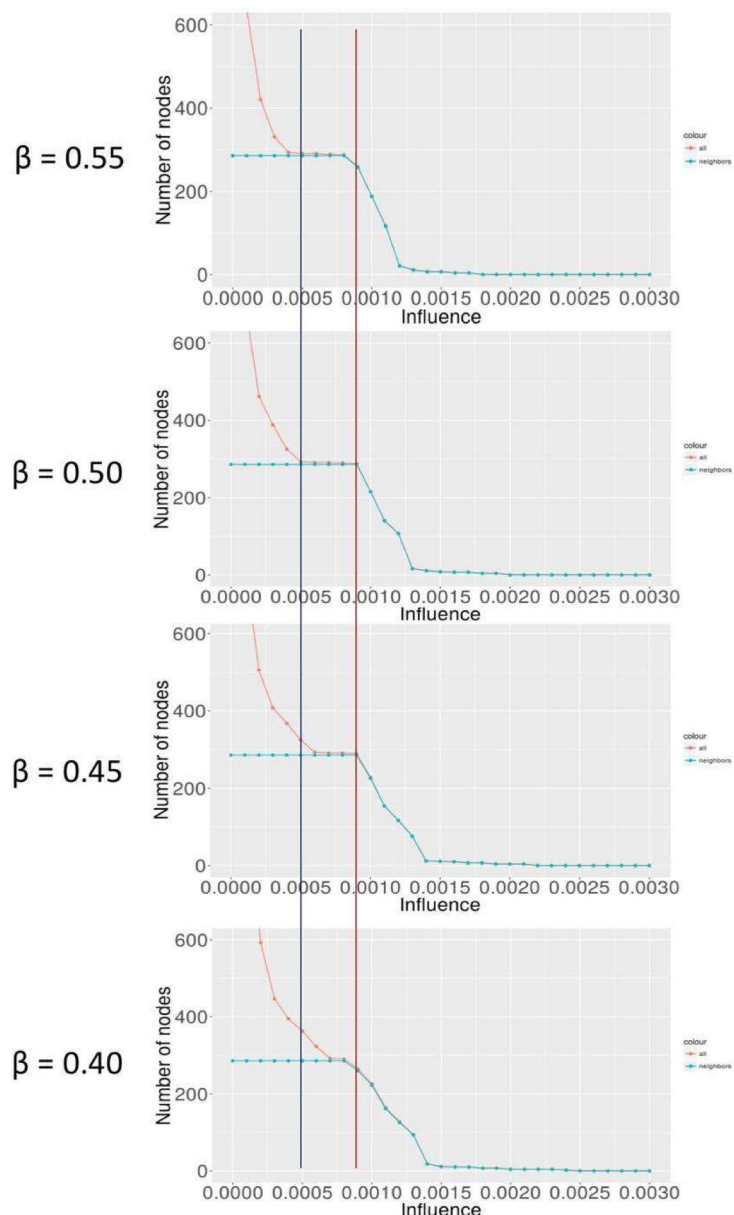


Figure 21. Examples of the distributions used to set β value for Reactome-derived network. Figures in different rows represent distributions of the influence across Reactome network with different β values for TP53 gene. The x-axis indicates a cut off of influence. The y-axis represents the number of nodes in the network with influence larger than the cut off of influence. Red and blue dotted circles represent the number of all nodes and neighbor nodes, respectively. The vertical red line indicates the location of inflection point in neighbors while the blue line indicates the point where the heat starts to spread across all network from neighbors.

The best β for Reactome-derived network is equal to 0.5 for two reasons. First, the smallest β before inflection point is equal to 0.45, but the heat spreads across network more than 0.5. Second, the amount of heat concentrated on neighbors are equal between 0.5 and 0.45.

Concluding, I setup a data structure for optimized HotNet2 analysis on Reactome-derived network including a heat diffusion matrix with β equal to 0.5, two text files containing the index of nodes, and all the edges of the network.

6.3. Genomic characterization of LGL-L patients

Detected somatic variants

LGL-L is a rare clonal disease with persistent increase of CD 8+ cytotoxic T cells or, in a small percentage of cases, CD4+ phenotype or CD16/56+ natural killer (NK) cells. LGL-L affects especially elderly patients with a median age of 60 years. Despite, the progression is indolent and asymptomatic for a long time after diagnosis, about 60% of the patients become symptomatic during the course of the disease. The most common complications are cytopenia, recurrent infections, splenomegaly, and autoimmune disorders such as rheumatoid arthritis. A fundamental pathogenic role for LGL-L is played by JAK/STAT pathway activation due to gain-of-function mutations in *STAT3*¹⁷⁰⁻¹⁷³ or *STAT5B*^{174,175} genes. JAK/STAT activation was found activated in LGL-L patients also through non-mutational mechanisms, such as an increased interleukin-6 secretion by normal mononuclear cells supported by the frequent epigenetic inactivation of *SOCS3*, which is an inhibitor of JAK/STAT pathway¹⁷⁶. Nevertheless, in some cases LGL-L patients develop disease not carrying JAK/STAT alterations. In order to investigate about other mutational mechanisms that could be involved in LGL-L or inducing a persistent JAK/STAT activation, I analyzed matched tumor-control WES data of 19 LGL-L patients including 11 STAT-mutation-negative patients.

The average sequencing coverage in the tumor samples was 32x (**Figure 22**).

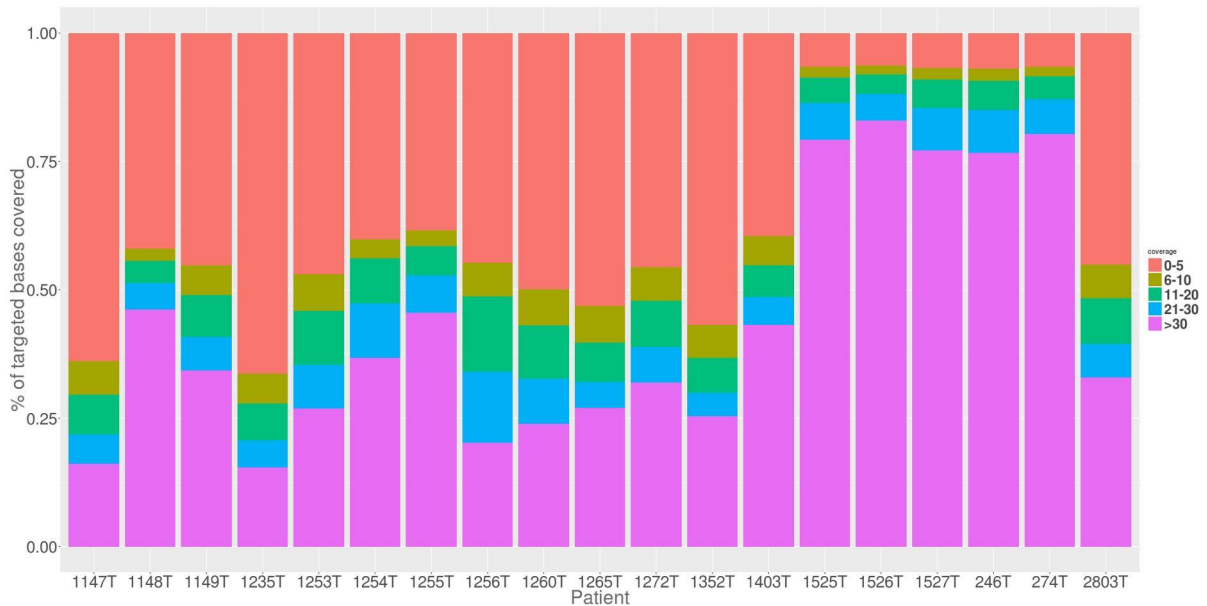


Figure 22. Coverage profile per patient in LGLL tumor samples. Each bar on the x-axis represents a patient; on the y-axis the percentage of targeted bases coverage is shown. Different colors represent different coverage ranges.

After selecting high confidence variants and filtering out variants already described in human populations single nucleotide polymorphism database and/or with allele frequency higher than 5% in ExAC, 28,508 somatic variants in 16,518 genes were identified in the whole cohort. Next, among high confidence and rare variants, 370 variants in 347 genes with a strong predicted functional impact were selected. The observed differences in numbers of somatic mutations (range 5–40, average 20) and genes involved (range 4–41, 19) per patient were not because of coverage differences (**Figure 23**).

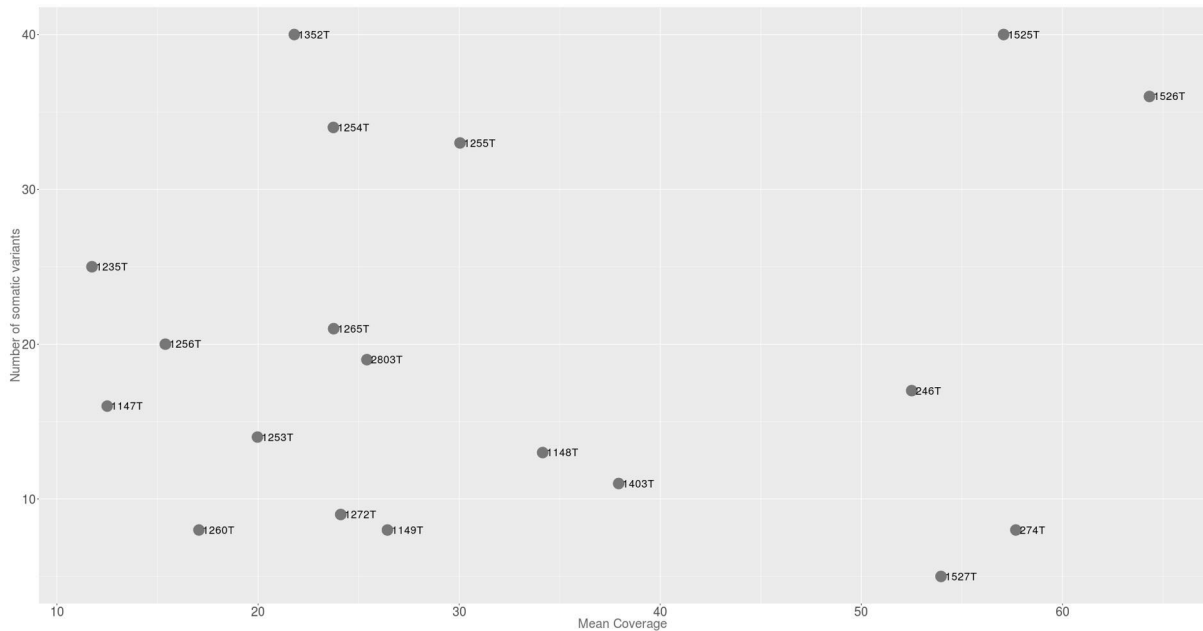


Figure 23. Somatic variants detected versus tumor mean coverage. *Number of high confident, rare and high impact somatic variants identified in each sample in relation with the mean coverage in the sample (no correlation is observed, rho: -0.17, p-value: 0.49).*

A slight tendency toward more mutated genes per patient in STAT-mutation-positive (22.9 in average) versus negative patients (18.4 in average) was noticed.

Beyond the mutated STAT genes

STAT3 (all in CD8+ T-LGL) and *STAT5B* (CD4+ and CD8+ cases) were the most recurrently mutated genes in the cohort (in 8/19 patients, 42%).

In addition to *STAT3* (all in CD8+ T-LGL) and *STAT5B* (CD4+ and CD8+ cases) mutations (in 8/19 patients, 42%), 14 other genes had recurrent mutations including transcriptional/epigenetic regulator, tumor suppressor and cell proliferation genes (**Figure 24A-B**).

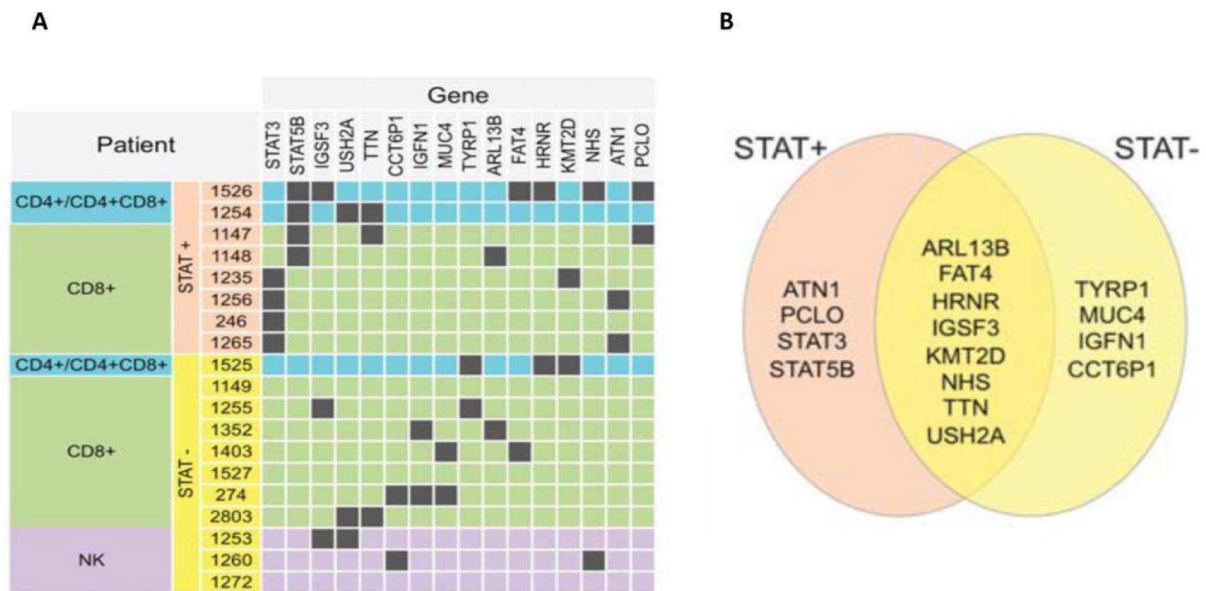


Figure 24. Recurrent somatic mutations in LGL leukemia patients. (A) The table indicates the genes that carry somatic variants in more than one patient, with a color code showing STAT3 and STAT5B status and classification of patients. (B) Recurrently mutated gene sets found only in STAT-mutation-negative patients (STAT⁻), only in STAT-mutation-positive patients (STAT⁺) or in both groups.

Novel mutated functional modules in STAT-negative patients

The custom knowledge-based systems genetics approach (**Figure 19**) provided the functional prioritization of mutated genes. We identified 119 KEGG and 426 Reactome pathway-derived networks, each including at least one of the 347 previously prioritized mutated genes associated to high confidence, rare and high-impact variants. The union of all path- derived networks generated a meta-network with 118 (34%) mutated genes, giving a non-redundant representation of functional relations, based on direct interactions between somatically mutated genes. Remarkably, 47 mutated genes were directly connected to at least another mutated gene in 18 multigene components (groups of genes whose products directly interact, that is, encode proteins taking part in the same molecular complex or regulating each other). Considering co-participation of mutated genes in pathways including STAT genes as additional functional link, seven multigene components connected by direct relations and three isolated genes converged into a component of 26 genes. In this reconstructed LGL leukemia network (**Figure 25**) 61 somatically mutated genes (occurring in many cases only in one sample) preferentially fall into a limited number of highly connected pathways, and in this manner collectively form a functional module hit by somatic mutations in LGL leukemia. The largest network component

included 24 mutated genes either directly linked to STAT genes, to their neighbors and/or participating in pathways including STAT genes (**Figure 25**).

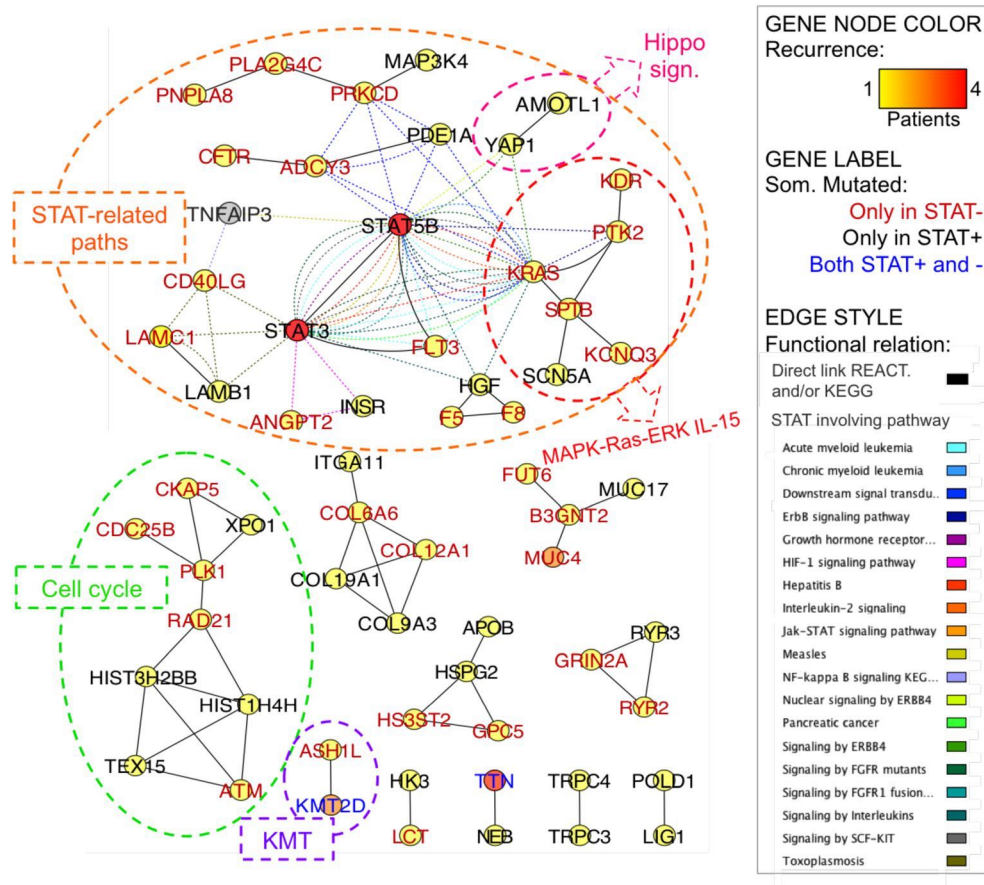


Figure 25. LGL leukemia mutation network. The network shows the functional relations of genes somatically mutated in LGL leukemia patients, according to the integration of KEGG and Reactome pathways topology (See the text and Supplementary Methods for details on the pathway-derived network reconstruction procedure); network nodes represent somatically mutated genes, node color indicates recurrence (according to the legend heat color scale) in the cohort, node label indicates the gene Symbol (different label colors indicates genes that are mutated only in STAT mutation positive (STAT+), only in STAT mutation negative (STAT-) or in both patient groups, as shown in the legend); genes are connected with black solid lines if they are directly connected in KEGG- and/or Reactome-derived networks or with colored dashed lines if they participate to pathways including STAT3 and/or with STAT5B (the legend indicates the color code for different pathways).

In 16 out of 19 patients, at least one gene of the largest component was mutated with some patients showing more than one hit in the gene group. For instance, one STAT-mutation-negative CD4+ patient presented with mutated alleles in three genes of the component (CD40LG, F8 and PLA2G4C). The similar variant allele frequency values of the variants support their co-presence in the dominant LGL leukemic clone. Altogether, 8 of 11 STAT-mutation-negative patients carried validated somatic mutations in at least one of the ‘STAT-related component’ genes, such as in FLT3, KRAS, ADCY3, ANGPT2 and PTK2. These mutated genes also connect the STAT component to the MAPK-Ras-ERK (**Figure 25**) pathway

and to the IL-15, all known to be deregulated in LGL leukemia¹⁷⁷. Other relevant variants confirmed in STAT-mutation-negative patients and connected to the STAT pathway were *KRAS* and the kinase KDR/VEGFR2.

Other components (and pathways) not directly linked to the main lesions were also of interest. Nine genes were linked to cell cycle regulation, and include the *CDC25B* gene and *ATM*, which is involved in apoptosis and P53 signaling (**Figure 25**). Furthermore, the epigenetic nodule included the recurrently mutated *KMT2D*, which is connected to *ASH1L*. Both are histone methyltransferases involved in epigenetic regulation of gene expression programs and are part of the ASCOM complex, involved in transcriptional co-activation.

Confirmed somatic variants

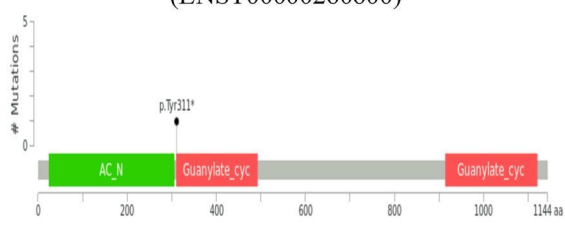
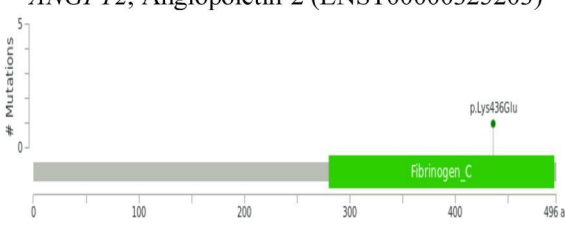
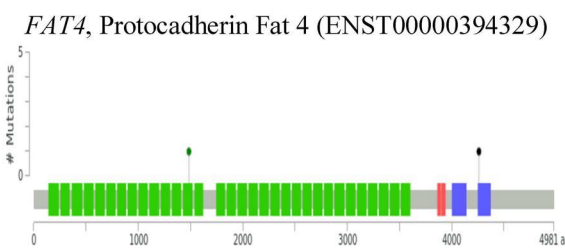
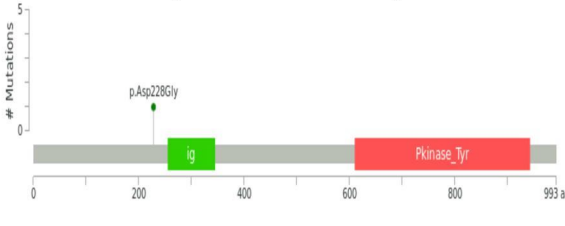
Sanger sequencing validations of somatic variants were obtained in 14 genes (**Table 7** and **Figure 26**) being recurrent or prioritized according to functional criteria and/ or connections emerged by integrated pathway-derived networks. Allele frequencies of variants detected in a same patient were very similar suggesting their co-occurrence in the same tumor clone.

Gene	Description	Variant	Variant Allele Frequency	Patient	Class	STAT3 STAT5B mutation state	STAT3 STAT5B Variant Allele Frequency
<i>PCLO</i>	Piccolo Presynaptic Cytomatrix Protein	M1182I	0.36	1147	CD8+	STAT5B	0.42
<i>ANGPT2</i>	Angiopoietin 2	K436E	0.33	1352	CD8+	NEG	
<i>FAT4</i>	FAT Atypical Cadherin 4	-2501	0.38	1403	CD8+	NEG	
<i>NRP1</i>	Neuropilin 1	V391M	0.36				
<i>FAT4</i>	FAT Atypical Cadherin 4	D1485N	0.45	1526	CD4+	STAT5B	0.48
<i>PCLO</i>	Piccolo Presynaptic Cytomatrix Protein	-1205	0.48				
<i>FLT3</i>	Fms Related Tyrosine Kinase 3	D228G	0.51	2803	CD8+	NEG	
<i>KDR</i>	Kinase Insert Domain Receptor	R176G	0.41				
<i>CD40LG</i>	CD40 Ligand	V216G	0.29	1525	CD4+	NEG	
<i>KMT2D/MLL2</i>	Lysine Methyltransferase 2D	E226G	0.25				
<i>PLA2G4C</i>	Phospholipase A2 Group IVC	-281	0.28				
<i>CDC25B</i>	Cell Division Cycle 25B	R526H	0.33	1253	NK	NEG	
<i>KRAS</i>	Kirsten Rat Sarcoma Viral Oncogene Homolog	A59G	0.21				

<i>PTK2</i>	Protein Tyrosine Kinase 2	R688Q	0.31			
<i>ADCY3</i>	Adenylate Cyclase 3	Y261*	0.47	1260	NK	NEG
<i>RAB12</i>	RAB12, Member RAS Oncogene Family	P235F	0.51	1272	NK	NEG

Table 7. Probably pathogenetic validated somatic variants.

The positions of the mutations in protein domains of selected genes are shown in **Figure 26**.

GENE SYMBOL, PROTEIN (TRANSCRIPT ENSEMBL ID)	AA CHANGE	SAMPLE S	COMMENT
<p><i>ADCY3</i>, Adenylate Cyclase type 3 (ENST00000260600)</p> 	T311*	1260	Stop gained variant cutting Guanylate cyclase, ATP, Mg ²⁺ domains and phosphorylation and glycosylation sites (UniprotKB, NextProt)
<p><i>ANGPT2</i>, Angiopoietin-2 (ENST00000325203)</p> 	K436E	1352	Missense variant lying in Fibrinogen C-terminal domain implicated in protein-protein interactions
<p><i>FAT4</i>, Protocadherin Fat 4 (ENST00000394329)</p> 	D1485N	1526	Missense variant lying in Cadherin 14 domain
	H4261fs	1403	Frameshift variant inducing a premature stop codon before the last Laminin G-like domain
<p><i>FLT3</i>, Receptor-type tyrosine-kinase FLT3 (ENST00000241453)</p> 	D228G	2803	Missense variant

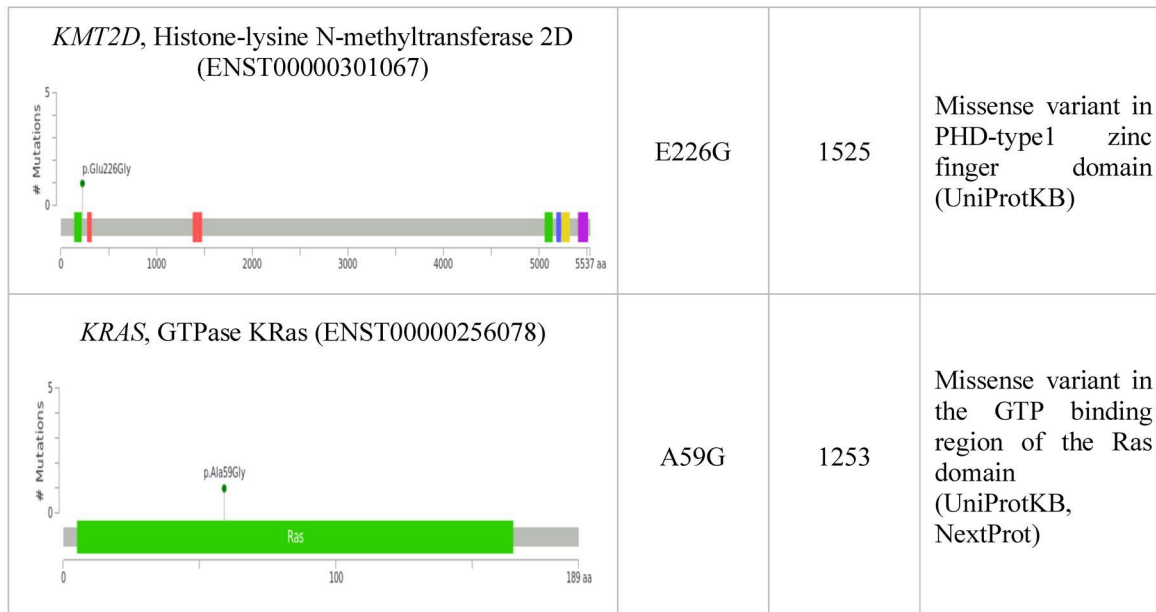


Figure 26. Impact of selected somatic variants to protein products. For each mutated gene and considering the protein isoform encoded by main transcript of the gene, Lollipop plots show the type and the position of selected somatic variants in relation to the protein sequence and domain structure and the table on the right provides details on one or more variants identified, as variant type, protein change, and possible functional relevance.

All validated variants impacted on protein functional domain or they induced truncating protein leading to loss of function.

High mutational burden in CD4-positive T-LGLL patients

When comparing the mutation profile between three different phenotypic LGL subgroups, qualitative and quantitative differences were observed, although the clinical characteristics of patients did not markedly differ. Interestingly, the average number of mutated genes per patient in the CD4+ class (36.0) was higher than in the other two classes (**Figure 27A**). Notably, in the entire cohort the number of somatic variants per patient did not correlate with the mean sequencing coverage of the patient sample (the correlation was close to zero, **Figure 23**), thus, the differences in mutation load likely reflect a different natural history of the LGL phenotypes. The number of mutated genes per NK patient was slightly, but not significantly, lower than in the CD8+ patients.

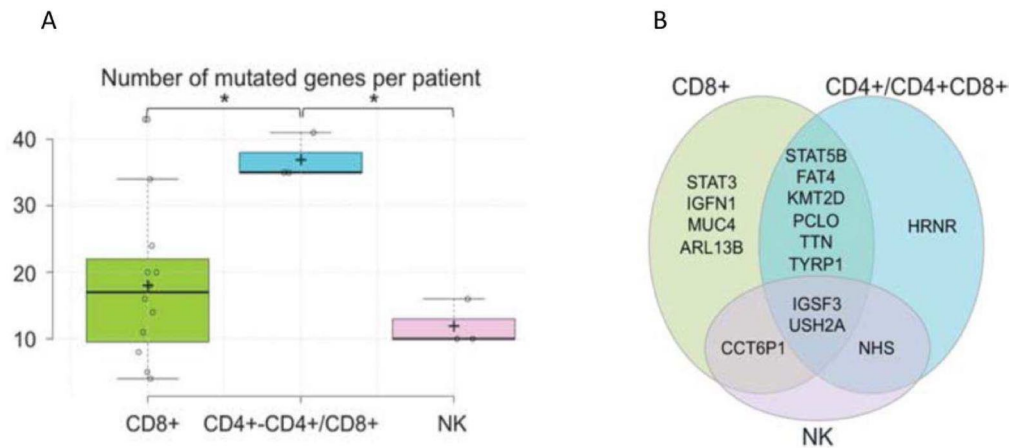


Figure 27. Mutational comparison between T-LGLL subgroups. (A) Number of mutations per patient in each class. Normal distribution of values was confirmed with the Shapiro–Wilk test ($P = 0.099$). Both analysis of variance ($P = 0.009$) and pairwise Tukey *s.d.* post hoc tests (P -values 0.010 and 0.019 in the comparisons of CD4+/CD4+CD8+ with NK and CD8+, respectively) confirmed the statistical significance of the observed difference. (B) Recurrently mutated genes that are found only in one or are shared among patient classes (CD8+, CD4+/CD4+CD8+ and NK+).

In CD8+ T-LGL leukemia, 227 genes presented somatic mutations (17.8 per patient on average). Genes recurrently mutated in the CD8+ class were largely overlapping with those recurrent in the whole cohort since more than 2/3 of the cohort was comprised of CD8+ cases (**Figure 27B**). Four patients were *STAT3* and 2 were *STAT5B* mutated. *IGFN1*, *MUC4*, *TTN*, *AKIRIN2*, *ARL13B*, *SVEP1*, and *ATN1* were mutated each in 2 CD8+ patients. Of these, *IGFN1* and *MUC4* were mutated only in *STAT* mutation negative patients, only in the CD8+ class and not in other classes.

In CD4+ and CD4+CD8+ LGL patients 108 genes showed somatic mutations. In addition to *STAT5B*, also *HRNR* (hornerin) was recurrently mutated in CD4+ patients (**Figure 27B**).

In NK LGL leukemia cases (all *STAT* mutation negative), 31 genes with somatic mutations with 10.3 hits per patient on average were identified. In addition to *KRAS*, *PTK2*, *NOTCH2*, and *CDC25B*, other genes known to be recurrently mutated in cancer were mutated only in NK patients, i.e. *HRASLS* (*HRAS*-like suppressor), *RAB12* (*RAB12*, member *RAS* oncogene family), *PTPRT* (Protein tyrosine phosphatase, receptor type, T), and *LRBA* (LPS-Responsive Vesicle Trafficking, Beach and Anchor Containing).

Mutations in the network for each LGL-L subgroup

The networks of genes mutated in individual CD8+ and CD4+ or NK LGL leukemia patients and in each subgroup are presented in **Figure 28**.

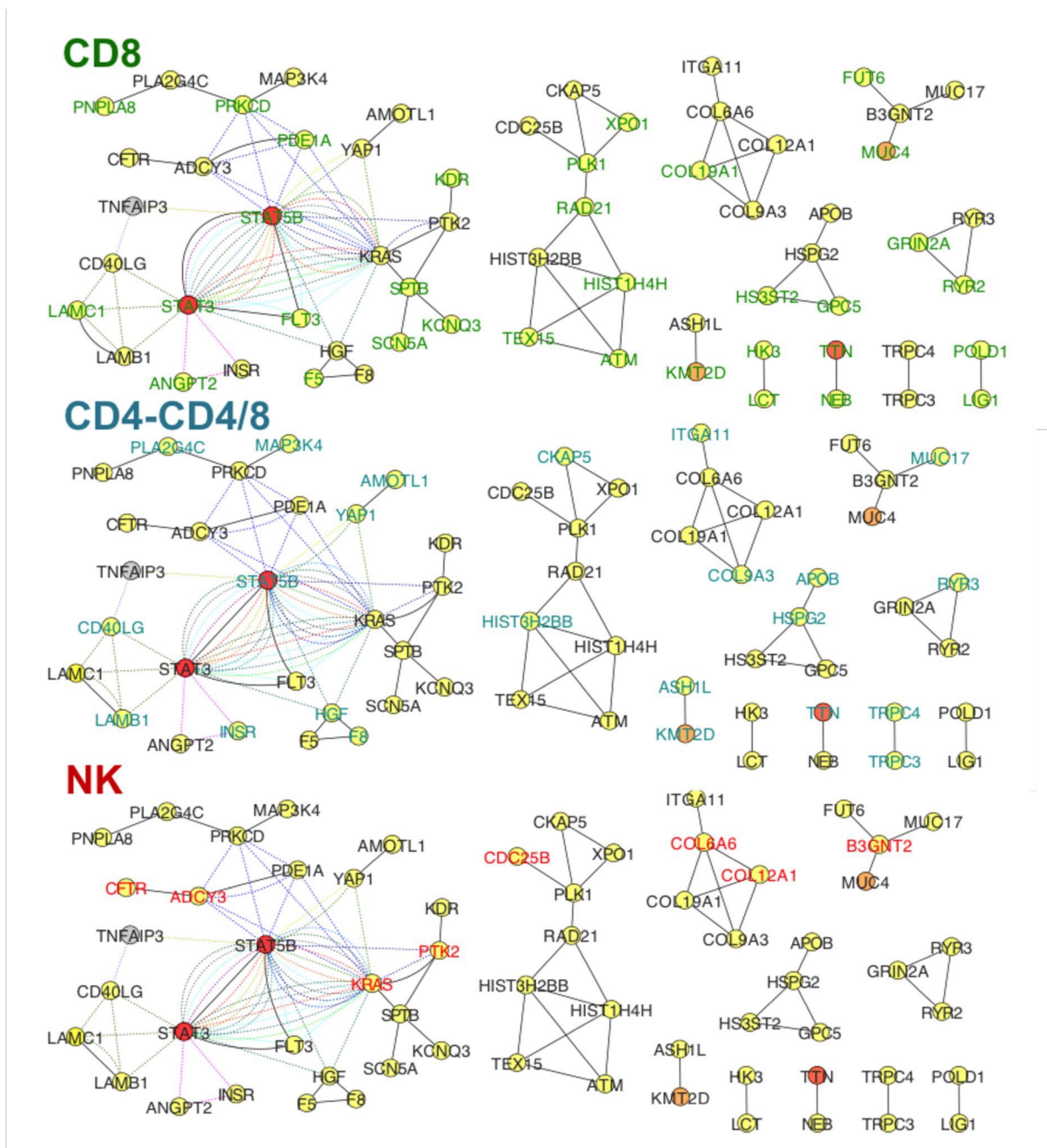


Figure 28. Genes in the network that are mutated in each patient subgroup. For each class, the genes mutated in at least one patient of the class are highlighted (gene symbol in a class-specific color) whereas the genes mutated only in the other classes are shown in black; the node color indicates the recurrence in the whole cohort as in *Figure 25*).

The CD8⁺ patients give the major contribute to network reporting mutations in many genes involved in all identified altered pathways.

The CD4⁺ patients carry a mutation in different genes of the STAT pathway, including *CD40LG* and *PLA2G4C* and in both genes (*KMT2D* and *ASH1L*) of the “epigenetic regulation” component. In addition, non-recurrent somatic mutations of *YAP1* and of its inhibitor *AMOTL1* point toward an involvement of Hippo signaling deregulation in CD4-positive T-LGLL.

NK LGL leukemia patients carry a mutation in the STAT-related genes *ADCY3* and *CFTR*. They also carry somatic mutations in the connected *KRAS*, *PTK2*, which are involved in MAPK-Ras-ERK signaling, and *CDC25B* (involved in cell cycle pathway).

Discussion

JAK-STAT pathway activation is the hallmark of LGL leukemia. It can be triggered by activating mutations in the *STAT3* and *STAT5B* genes and also by non-mutational mechanisms¹⁷⁸, such as by increased interleukin-6 (IL-6) secretion and epigenetic inactivation of JAK-STAT pathway inhibitors¹⁷⁶.

These WES data obtained from the largest cohort of LGL leukemia patients thus far revealed novel mutated genes connected to STAT signaling, further supporting the central role of the JAK-STAT pathway in LGL leukemia. Of note, JAK-mutations were not seen in this LGL-leukemia cohort, although prevalent in another mature T-cell malignancies¹⁷⁹.

Somatic variants in 347 genes in LGL leukemia patients were detected and the variant number per patient was significantly higher (more than doubled) in CD4+ LGL leukemia patients than in the other two phenotypic classes (CD8+, NK). The clinical characteristics of these immunophenotypically segregated patient groups did not differ markedly. As reads coverage across patients was of low variability, and the number of somatic variants did not correlate with the coverage per patient, the observed differences are likely to reflect a different natural history of the LGL phenotypes. Cytomegalovirus-derived stimulation and restricted usage of TCR V β has been associated with CD4+ T-LGL cases¹⁸⁰, and this could relate to the higher variant rate. Focusing on recurrently mutated genes, 16 genes were found to be recurrently mutated (**Figure 24**), including *STAT3* and *STAT5B* genes. *STAT5B* was found recurrently mutated in two out of three CD4+ patients. The two mutations Q706L and S715F detected in CD4+ patients hit a serine phosphorylation site in the transactivation domain of *STAT5B*. Functional validations were conducted on these two new detected variants showing that the S175F variant has an activating effect on *STAT5B* significantly increasing its phosphorylation respect to wild-type *STAT5B*. In contrast, the Q706L showed no increased phosphorylation. In addition, mutations in transactivation domain of *STAT5B* were reported in 55% of CD4+ LGLL patients analyzed by deep amplicon sequencing or WES¹⁸¹.

One of the other recurrently mutated genes with confirmed somatic variants was *FAT4*, which is a member of the protocadherin family that regulates cell polarity and is an upstream regulator of stem cell genes both during development and cancer. Furthermore, *FAT4* acts as a tumor growth suppressor via activation of Hippo signaling. It has been found recurrently mutated in

several types of human cancer, and also previously described as an age-related hit in chronic myelomonocytic leukemia¹⁸², acute lymphoblastic leukemia¹⁸³, other myeloproliferative neoplasms¹⁸⁴, and in extranodal natural killer/T cell lymphoma¹⁸⁵. Not only *FAT4*, but also the recurrently mutated *ARL13B* (ADP-ribosylation factor-like 13B) is linked to Hippo signaling. *ARL13B* encodes a small GTPase found in the ciliary membrane. Primary cilia are both chemo- and mechano-sensors whose role in cell cycle control was recently recognized and whose importance in cancer cells is gradually understood as they crosstalk with several signaling pathways including Hippo. The Hippo signaling-linked *ARL13B* and *FAT4* genes were mutated in a mutually exclusive way and 2 out of 4 patients with *ARL13B* or *FAT4* mutations were *STAT*-mutation negative. Notably, both Hippo and JAK/STAT are among the pathways implicated in cell competition processes, in which cells with different relative fitness compete locally for tissue development, which are active also in tumorigenesis¹⁸⁶. Hippo signaling activation can be due to loss-of-function mutations of its inhibitors, as *FAT4*, but also constitutive activity of the Ras pathway confers proliferative advantage through inhibition of the Hippo pathway¹⁸⁷. In addition to *FAT4* and *ARL13B*, *KMT2D* was recurrently mutated. It is a transcriptional and epigenetic regulator which is frequently mutated in a variety of cancers¹⁸⁸ and whose disruption has been recently linked to lymphomagenesis suggesting that it acts as a tumor suppressor controlling the epigenetic landscape of cancer precursor cells¹⁸⁹. Interestingly, *HRNR* was found recurrently mutated specifically in CD4+ patients. *HRNR* is a calcium-binding protein involved in hematopoietic progenitor cell differentiation that has been reported as mutated, amplified, or overexpressed in cancer and was previously connected with acute myeloid leukemia transformed from myelodysplastic syndrome with t(1;2)(q21;q37)¹⁹⁰.

The second part of the study focused on functional prioritization of mutated genes by a custom knowledge-based “systems genetic” approach. The reconstructed LGL leukemia network (**Figure 25**) includes 61 somatically mutated and functionally related genes that were found to be affected in the cohort. Apparently, mutated genes preferentially fall into a limited number of highly connected pathways. Among the discovered modules, the *STAT*-related network component is especially significant. It should be noted that in several cases *STAT*-mutation negative patients carried a somatic mutation in a gene functionally connected to *STAT3* and/or to *STAT5B* (such as *FLT3*, *KRAS*, *ADCY3*, *ANGPT2*, and *PTK2*). These genes also connect the *STAT* component to the MAPK-Ras-ERK (**Figure 25**) pathway and to the IL-15, all deregulated in LGL leukemia¹⁷⁷. For example, *PTK2* (focal adhesion kinase 1, FAK1) is a non-receptor protein-tyrosine kinase which is highly expressed in T-cells and it regulates several

processes, including cell cycle progression, cell proliferation, and apoptosis, activating numerous pathways as PI3K/AKT signaling MAPK/ERK, and MAP kinase signaling cascades. Also, the mutated *ANGPT2* (Angiopoietin 2) gene is linked to PI3K-AKT and RAS signaling pathways that it antagonizes. *ANGPT2* is expressed in lymphocytes and controls T-cell proliferation^{191,192} and several studies reported the involvement of *ANGPT2* and other angiogenic factors in chronic lymphocytic leukemia where they exert pro-survival effects^{193,194}. Some of the STAT-connected genes are receptors. *CD40LG* is expressed on the T-cell surface that regulates B-cell function by engaging CD40, regulating immune systems and participating in STAT3 as well as in IL and NFAT signaling pathways. Interestingly, *CD40LG* was annotated in the same KEGG pathways as *TNFAIP3* (**Figure 25**), which is a negative regulator of NF- κ B signaling and known tumor suppressor gene and was recently found to be mutated in 8% of T-LGL leukemia patients¹⁹⁵. *FLT3* (fms-related tyrosine kinase 3) is a class III receptor tyrosine kinase that, promoting the phosphorylation of various proteins and kinases in the PI3K/AKT/mTOR, RAS, and JAK/STAT signaling pathways, regulates differentiation, proliferation, and survival of hematopoietic cells and is causally implicated in acute myeloid leukemia.

In addition to the STAT-related and the previously discussed “epigenetic” components of the LGL leukemia network, the cell cycle regulation module includes the validated *CDC25B* gene, and other genes mutated in STAT negative patients such as *ATM*, that connects the P53 and apoptosis pathways with the cell cycle pathway genes. The functional and pathogenic relevance of genes in these modules remains to be determined by future studies.

With the systems genetic approach, individual mutations found in LGL leukemia patients were mapped in novel functional modules. The central role of JAK-STAT network was further highlighted, and these data provide important new insights of the activation of this pathway in those LGL leukemias that do not carry STAT mutations. In addition, new insights about molecular features characterizing the three different LGL subtypes were provided.

6.4. New genes and pathways mutated in follicular lymphoma of the pediatric age patients

Recurrent somatically mutated genes in pediatric FLs

Peripheral B-cell lymphomas of the pediatric age are a well-defined group of B-cell neoplasms, encompassing both aggressive and indolent tumors. Biologically, indolent B-cell lymphomas of the pediatric age derive from either germinal center (GC) (follicular lymphoma [FL]) or

extra-GC (marginal B-cell lymphoma) B-cells¹⁹⁶. Classic FL is very rare in children and adolescents, whereas specific FL variants are more frequent than in adults and include the “pediatric-type nodal follicular lymphoma” (PTNFL) and the primary FL of the testis (PFLT)¹⁹⁶. PTNFL has long been considered a localized variant of follicular lymphoma (FL) with high grade morphology and a benign clinical course¹⁹⁷, distinguished from typical adult FL by the absence of the t(14;18)(q32;q21) translocation and the lack of Bcl2 expression¹⁹⁸. In the revised World Health Organization 2016 classification of lymphoid neoplasms, PTNFL has been recognized as a definite clinico-pathological entity, not restricted to the pediatric age but presenting, albeit rarely, also in adults¹⁹⁶. Although the mutational landscape of adult FL has been extensively investigated^{199,200}, only few genetic alterations involved in the pathogenesis of PTNFL has been reported so far by two independent studies^{201,202}. Among genes recurrently mutated in adult FL, only *TNFRSF14* alterations were detected in PTNFL. Both mutations in *IRF8* DNA binding domain²⁰³ and mutations activating *MAP2KI*^{201,204} have been recently proposed as potential drivers of PTNFL pathogenesis. Even if these studies provided the first insights on mutations subtending PTNFL development, the biological mechanisms and signaling pathways involved in this malignancy remain to be fully elucidated and very little is known about the molecular features of FL histological variants in children.

In order to investigate about biological mechanisms and signaling pathways involved in FL of the pediatric age development, I analyzed matched tumor-control WES data of 7 PTNFL and 2 PFLT patients.

After WES sequence read quality selection and alignment to the reference genome, average sequence coverage of 216x for tumor samples and 87x for paired peripheral blood was obtained. Sequence coverage, considering both tumor and control samples, was homogeneous (**Figure 29**).

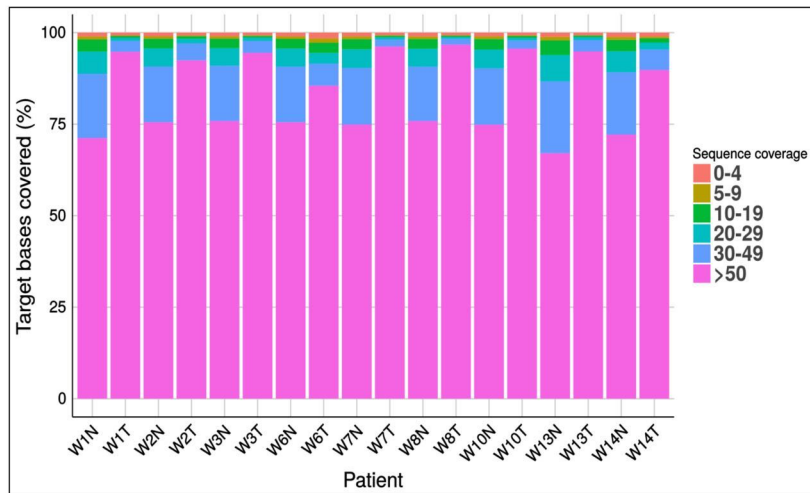


Figure 29. Sequence coverage profile per patient. Each bar on the x-axis represents a patient; on the y-axis the percentage of targeted bases coverage is shown; different colors represent different coverage ranges; T, tumor sample, N, paired peripheral blood.

Variant annotation, filtering and analysis identified 184 high confidence somatic variants in 169 genes, with “Moderate” or “High” SnpEff predicted impact and a normalized allele frequency >0.1.

Recurrent mutations were observed in 10 genes (**Figure 30**), including *MAP2K1*, *IRF8* and *TNFRSF14*, previously linked to PTNFL^{201–204}. In addition to already known mutated genes, *ARHGEF1*, *NSD1*, *PABPC1*, and *RHPN2* have important cellular functions.

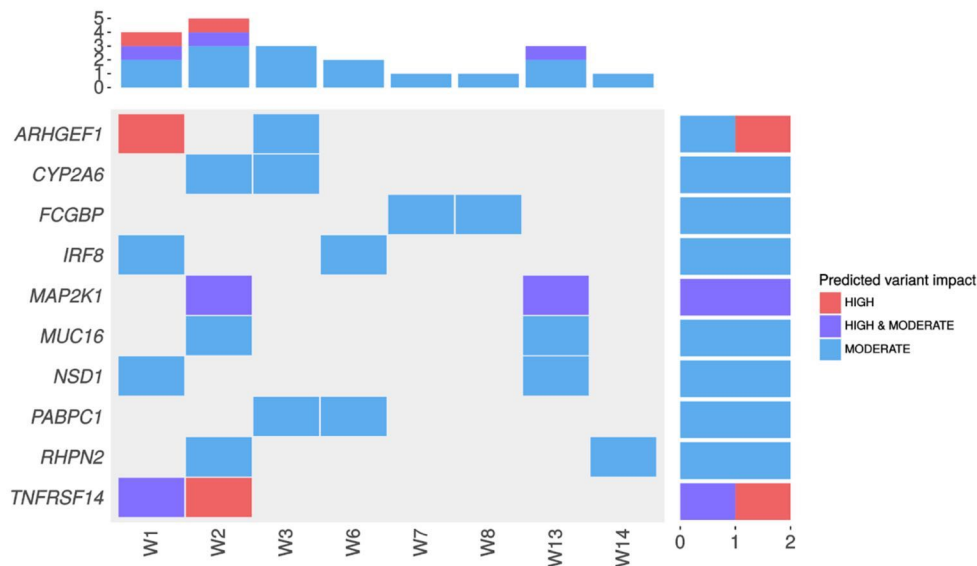


Figure 30. Recurrent somatic mutations in PTNFL and PFLT patients. The figure indicates the genes carrying somatic variants in more than one patient.

Exon 1 of TNFRSF14 is a mutational hotspot in PTNFL

Screening by Sanger sequencing of *TNFRSF14* exon 1 in 12 additional PTNFL and PFLT cases disclosed mutations in six of them, thus overall 8/21 (38%) of the cases had *TNFRSF14* exon 1 mutations (**Figure 31**).

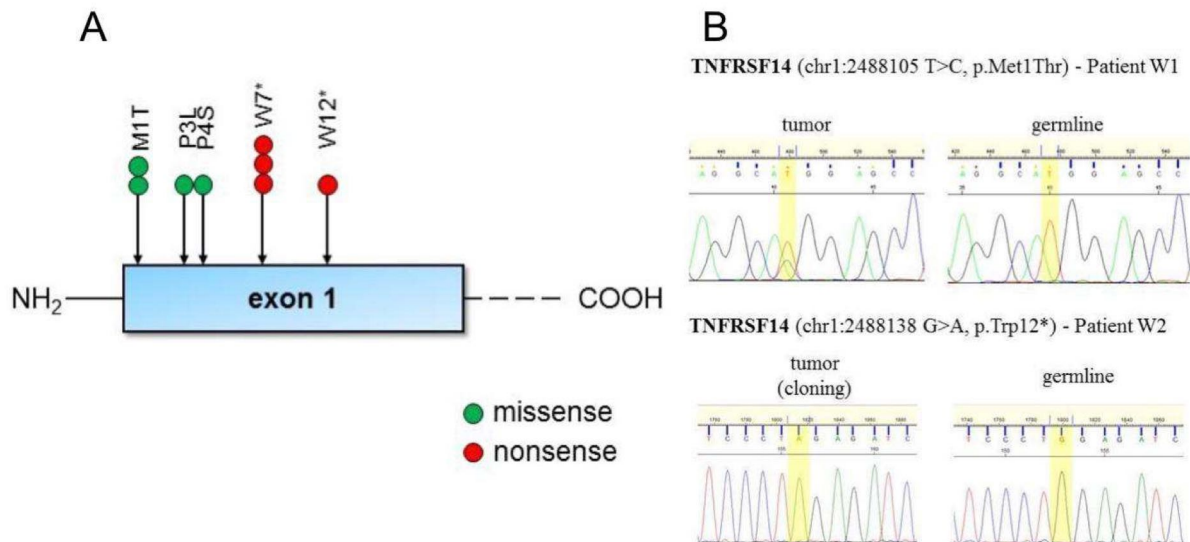


Figure 31. Mutations detected in exon 1 of TNFRSF14. (A) Schematic diagram of *TNFRSF14* 8 mutations detected in exon 1 positioning detected in 21 cases of PTNFL and PFLT. (B) Validation by Sanger sequencing of the two variants, both hitting *TNFRSF14* exon 1, detected by WES.

The p.Met1Thr variant in W1 patient is a start loss variant while p.Trp12* is a stop gain variant and both are predicted to be inactivating mutations.

Pathways recurrently altered in FL of the pediatric age

A meta-network (**Figure 32A**) was reconstructed from KEGG and Reactome pathways depicting direct interactions and functional relationships between 66 genes somatically mutated in pediatric FLs, mostly carrying previously undescribed variants.

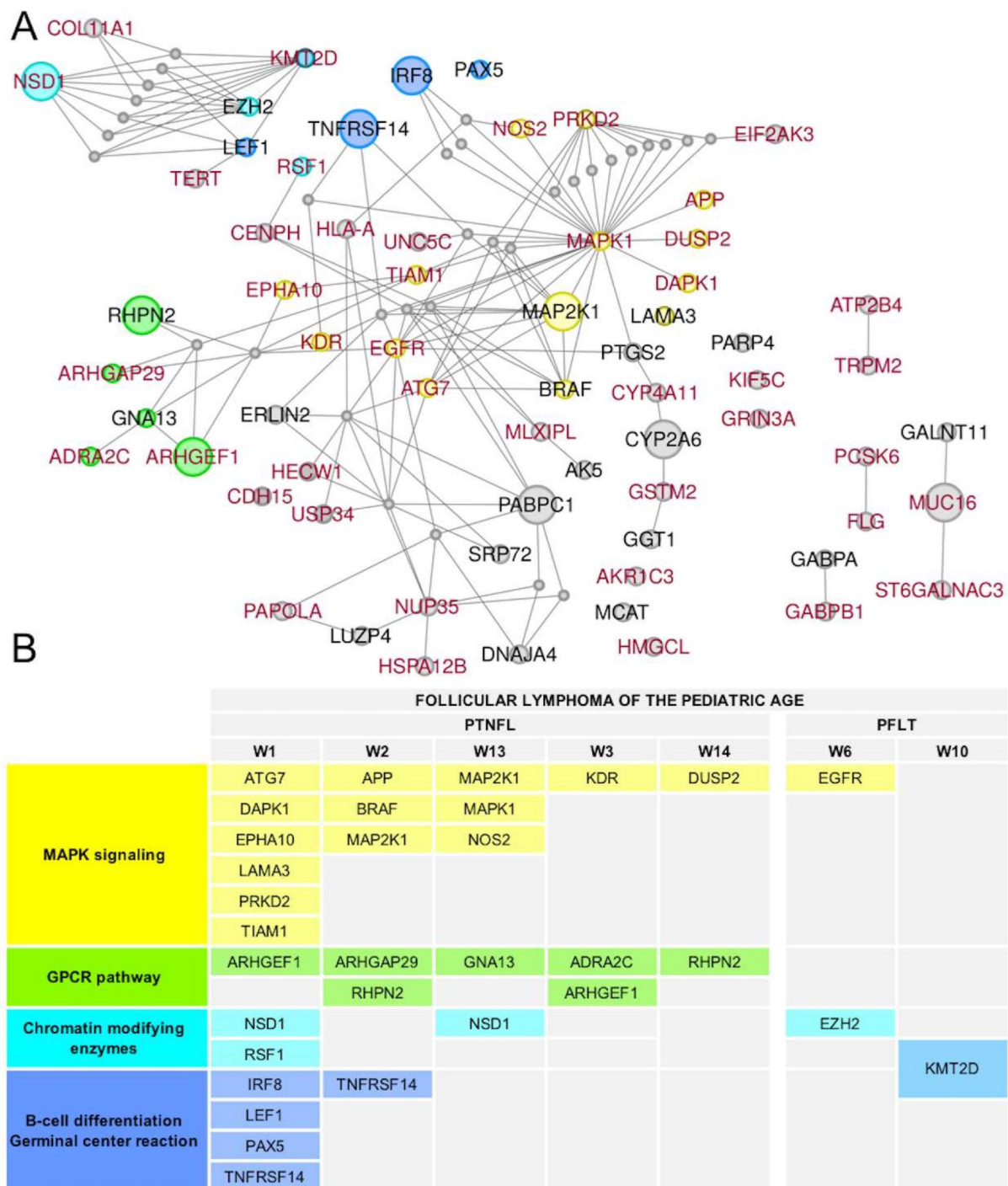


Figure 32. KEGG and Reactome pathways derived meta-network of genes somatically mutated in FL of the pediatric age. **A)** Net nodes indicate genes somatically mutated in the cohort, with node size proportional to recurrence in the cohort analyzed by WES; genes involved in the negative regulation of MAPK, G-protein coupled receptor pathways and chromatin modifying enzymes are indicated in different colors, using panel B as color legend; direct functional links between mutated genes are depicted, and small grey nodes indicate pairs of mutated genes connected through a single non mutated gene in the network; isolated nodes located close to the connected components are associated to them according to annotated function and pathway participation; red node labels indicate the genes for which new mutations were identified. **B)** Mutation table showing which patient carries mutations in the genes of the main identified pathways. Patients are grouped according to the primary site of the tumor. Patients W7 and W8, negative for mutations connected with those represented in the figure, are not shown.

Six connected components of multiple mutated genes each were identified in the net: a large group of 51 genes (“main component”), a group of six genes, and four smaller groups. Notably, the net included nine genes recurrently mutated in the cohort, most of them in the main component, comprising *ARHGEF1*, *RHPN2*, *CYP2A6*, and *PAPBC1* in addition to the already known in PTNFL *TNFRSF14*, *IRF8* and *MAP2K1*. Almost half of the somatic mutations observed in the main component targeted genes belonging to two highly interlaced signaling pathways: “*negative regulation of MAPK*” and “*G-protein coupled receptor*” (**Figure 32A**). MAPK pathway component was made up of 14 mutated genes including the recurrently mutated *MAP2K1*. New variants in 12 additional genes of the MAPK pathway (*APP*, *ATG7*, *DUSP2*, *DAPK1*, *EGFR*, *EPHA10*, *KDR*, *MAPK1*, *NOS2*, *PRKD2* and *TIAMI*) were uncovered. Overall, 6/9 patients carried one or more mutations in genes directly involved in MAPK signaling (**Figure 32B**).

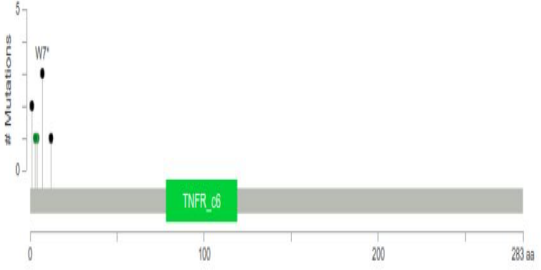
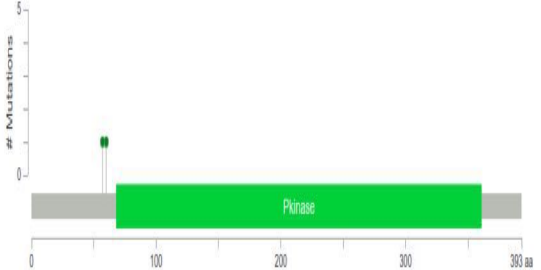
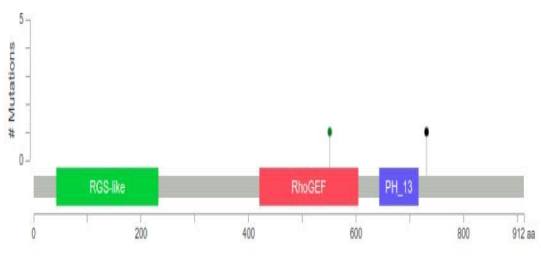
The “*G-protein coupled receptor*” signaling pathway was mutated in 5 cases of PTNFL (5/7, 71,4%) and was preserved in the examined PFLT cases, comprising the recurrently mutated *ARHGEF1* and *RHPN2* and other 3 mutated genes (*GNA13*, *ARHGAP29* and *ADRA2C*) (**Figure 32B**). This signaling is comprised in the large component and it is linked to MAPK pathway.

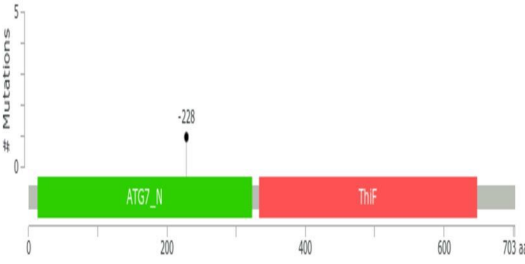
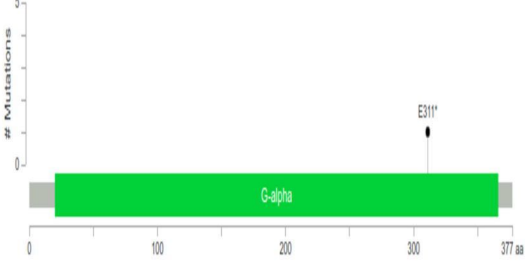
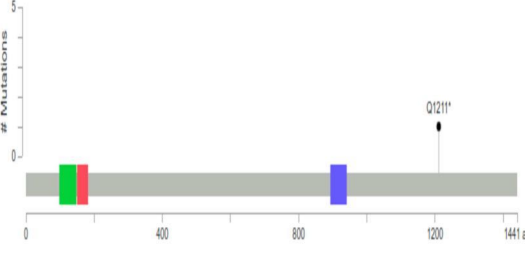
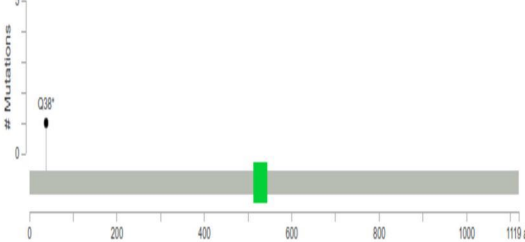
Moreover, the “*chromatin modifying enzymes*” pathway was hit in both PTNFL and PFLT: *RSF1* and *NSD1* genes were mutated in PTNFL cases, while mutations on *EZH2* and *KMT2D* genes were found in PFLT (**Figure 32B**).

Five mutated genes are involved in germinal center B-cell development (**Figure 32A**) including the two recurrently mutated *TNFRSF14* and *IRF8*, *LEF1*, *PAX5* and *KMT2D*, which also participates to chromatin modifying enzymes pathway. The mutations in this functional component are almost all detected in W1 patient, except two mutations in *TNFRSF14* and *KMT2D* detected in W2 and W10 patients, respectively.

Highly prioritized and confirmed somatic variants

Eleven “High” impact variants in eight genes, prioritized according to recurrence, functional criteria and/or previous reports in PTNFL or other lymphoid malignancies, were validated by Sanger sequencing. Six variants in the recurrently mutated *ARHGEF1*, *MAP2K1* and *TNFRSF14* genes, and 5 variants in *ATG7*, *GNA13*, *RSF1*, *UBAP2*, and *ZNF608* (**Figure 33**). All validated variants are loss of function variants or missense variants impacting protein domains.

GENE SYMBOL, PROTEIN (TRANSCRIPT ENSEMBL ID)	VARIANT (VAF)	AA CHANGE	SAMPLES	COMMENT
<p><i>TNFRSF14</i>, Tumor necrosis factor receptor superfamily member 14 (ENST00000355716)</p> 	chr1:2488105 T>C (0.36)	M1?	W1, S4	Start lost variant inducing loss of function
	chr1:2488111 C>T (0)	P3L	S13	Missense variant in N-terminal signal peptide
	chr1:2488113 C>T (0)	P4S	S5	Missense variant in N-terminal signal peptide
	chr1:2488123 – 2488124 G>A	W7*	S2, S9, S12	Stop gain variant inducing loss of function
	chr1:2488138 G>A (0.10)	W12*	W2	Stop gain variant inducing loss of function
<p>MAP2K1, Dual specificity mitogen-activated protein kinase 1 (ENT00000307102)</p> 	chr15:66727453 A>G (0.10)	K57E	W2	Missense variant lying closely to protein kinase domain, already found in hairy-cell leukemia (COSM1315807)
	chr15:66727463 T>G (0.22)	V60G	W13	Missense variant lying closely to protein kinase domain
<p><i>ARHGEF1</i>, Rho guanine nucleotide exchange factor 1 (ENST00000354532)</p> 	chr19:42406962 G>A (0.26)	R551H	W3	Missense variant hitting RhoGEF domain (COSM6206921)
	chr19:42409131 G>A (0.16)	W731*	W1	Stop gain variant cutting the phosphotyrosine site in 738 position and a phosphoserine in 863 position

<p><i>ATG7</i>, Autophagy-related protein 7 (ENST00000354449)</p> 	<p>chr3:11356968 G>A (0.22)</p>	<p>-228</p>	<p>W1</p>	<p>Stop gain variant inducing loss of function</p>
<p><i>GNAI3</i>, Guanine nucleotide-binding protein subunit alpha 13 (ENST00000439174)</p> 	<p>chr17:63010578 C>A (0.47)</p>	<p>E311*</p>	<p>W13</p>	<p>Stop gain variant already detected in colorectal cancer (COSM1680008)</p>
<p><i>RSF1</i>, Remodeling and spacing factor 1 (ENST00000308488)</p> 	<p>chr11:77383207 G>A (0.13)</p>	<p>Q1211*</p>	<p>W1</p>	<p>Stop gain variant inducing loss of function cutting 13 phosphoserine, 2 phosphothreonine and 2 N6-acetyllysine sites</p>
<p><i>UBAP2</i>, Ubiquitin-associated protein 2 (ENST00000360802)</p> 	<p>chr9:33998850 G>A (0.11)</p>	<p>Q38*</p>	<p>W2</p>	<p>Stop gain variant inducing loss of function cutting the most of protein including Ubiquitin associated domain</p>

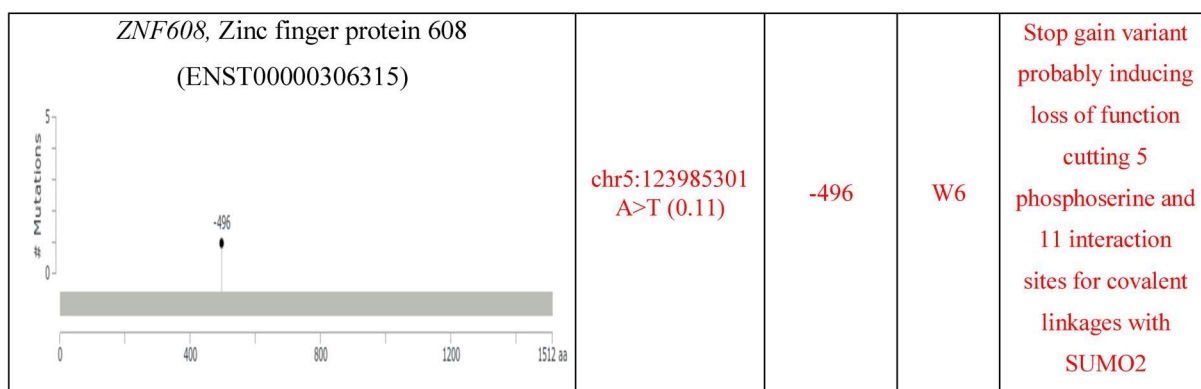


Figure 33. Impact of validated somatic variants to protein products. For each mutated gene and considering the protein isoform encoded by the main transcript of the gene, Lollipop plots show the type and the position of selected somatic variants in relation to the protein sequence and domain structure. The columns on the right provide details on identified variants, as variant type, protein change, and possible functional relevance. Previously unknown variants are indicated in red.

Discussion

Peripheral B-cell lymphomas of the pediatric age comprise both aggressive and indolent tumors. Specific variants of the classic FL are more frequent in children and adolescents than in adults, such as PTNFL and PFLT. Even though recent studies indicated some mutations as implicated in pediatric FLs, especially in PTNFL, the molecular mechanisms and the signaling pathways involved in this disease are still unknown.

These WES data obtained from a series of FL of the pediatric age revealed new variants and genes mutated in PTNFL and PFLT, along with a few previously known mutations. In addition, pathway-derived networks of functionally connected genes recurrently hit in these lymphomas were outlined leveraging systems genetics approaches.

Somatic variants with deleterious predicted impact hit 169 genes, including 10 recurrently mutated genes (each mutated in 2/9 patients) (**Figure 30**). These data confirm the recurrence of mutations affecting *TNFRSF14*, *MAP2K1*, and *IRF8* in PTNFL. *TNFRSF14* is a member of TNF receptor superfamily, which is involved in signal transduction pathways that activate inflammatory and limit T-cell immune response²⁰⁵. Both the two *TNFRSF14* variants p.Met1Thr and p.Trp12* hit the exon 1 and they are predicted to be inactivating mutations. Validation of these two variants and investigation about *TNFRSF14* exon 1 in additional 12 pediatric FL patients are in accordance with previous studies that reported *TNFRSF14* exon 1 as a mutational hotspot in PTNFL^{201,202}.

Two missense mutations affecting interaction residues of *MAP2K1* were detected. The variant p.Lys57Gly has been reported in natural killer/T-cell lymphoma²⁰⁶. *MAP2K1*, whose mutations

results in the activation of ERK signaling pathway²⁰⁷, could play a very central role in the pediatric FLs mutation network.

IRF8 (Interferon Regulatory Factor 8) is a transcription factor of the Interferon Regulatory Factor (IRF) family, mainly expressed in myeloid cells. It regulates expression of genes involved in various complex networks such as apoptosis, cell cycle, differentiation and maturation. Two *IRF8* missense mutations, p.Lys66Arg in W1 and p.Tyr23His in W6, were detected and both predicted to be very deleterious by MetaSVM and MetaLR. The variant *IRF8* p.Lys66Arg was already described in co-occurrence with mutations of *TNFRSF14* in 50%²⁰³ and 10%²⁰² of PTNFL patients, supporting the hypothesis of immune system deregulation as one of the main drivers of PTNFL development.

Other recurrently mutated genes showed to have important biological functions, such as *NSDI*, *PABPC1*, *ARHGEF1*, and *RHPN2*.

NSDI is a histone methyltransferase which acts as transcriptional factor capable of both negatively or positively influencing transcription, depending on the cellular context. Mutations in *NSDI* were associated to Beckwith-Wiedemann Syndrome, an overgrowth usually present at birth that predisposes to childhood cancer²⁰⁸.

PABPC1 regulates expression of mRNAs binding the 3' poly(A) tail and promoting ribosome recruitment and translation initiation²⁰⁹. It is even involved in regulation of nonsense-mediated decay of mRNA with premature codon stops²¹⁰. High expression of *PABPC1* has been found as involved into gastric cancer and poor survival by regulating miR34-c²¹¹.

ARHGEF1 is a Rho GTPase which is involved in numerous cellular processes activated by extracellular stimuli. Mutations in *ARHGEF1* have been associated to germinal center-type diffuse large B-cell lymphoma²¹².

RHPN2 is a member of rhophilin family of (Rho)-GTPase binding proteins. *RHPN2* amplification leads to mesenchymal transformation in aggressive glioblastoma by activation of RhoA which is associated with a dramatically decreasing in the survival of patients²¹³.

A custom knowledge-based “systems genetic” approach reconstructed pediatric FL network (**Figure 32**) includes 65 somatically mutated and functionally related genes that were found to be mutated in the cohort. The mutated genes seem to group in a limited number of highly connected pathways such as “negative regulation of MAPK”, “G protein-coupled receptor”, and “Chromatin modifying enzymes”.

The “negative regulation of MAPK” signaling pathway comprises *MAP2K1* and *BRAF* hit by the p.Lys601Asn mutation (W2) previously reported as oncogenic since affecting the tyrosine kinase domain²¹⁴. Other specific mutations are particularly worthy of note: MAPK1

p.Tyr131His falls in the catalytic domain, ATG7 truncation due to a stop gain at amino acid 228 eliminates the C-terminal region of the protein, essential for dimerization and key molecular interactions; DUSP2 p.Ser249Asn falls in the tyrosine-protein phosphatase domain involved in the negative regulation of MAPK members associated with cellular proliferation and differentiation; the detected *PRKD2* mutation is predicted to produce both p.Phe81Ile substitution and also to impair the main transcript splicing, probably resulting in a loss of function. “High” impact variants in other genes of the pathway are those causing stop loss in *EPHA10* and a possible intron retention in *NOS2*. The 66.7% of the cohort carried mutations in MAPK pathway.

Our data strengthen the relevance of MAPK pathway in the pathogenesis of PTNFL and uncovered specific molecular alterations not described previously, indicating as well that the entire signaling cascade, rather than a single gene, is altered in this lymphoma.

In the GPCR pathway, a previously unreported ARHGEF1 p.Trp746* nonsense mutation was disclosed in patient W1 and the known ARHGEF1 pathogenic variant p.Arg566His was found in W3. Notably, *GNA13*, encoding Gα13, a direct interactor of ARHGEF1, carried a truncating mutation (p.Glu311*) eliminating the C-terminal part on the G protein alpha domain. Two *GNA13* missense mutations, both affecting Gly60 (p.Gly60Ser and Gly60Asp)²⁰² in the nucleotide phosphate-binding region of the protein were previously reported by Schmidt *et al.* in two cases of PTNFL. Disruption of Gα13-dependent pathway by loss of function mutations has already been reported in more aggressive lymphomas, such as diffuse large B-cell lymphoma and Burkitt’s lymphoma²¹⁵.

RSF1 and three genes of the second largest component (*EZH2*, *KMT2D* and the recurrently mutated *NSD1*) belong to the “chromatin modifying enzymes” pathway. The detected *EZH2* p.Tyr646Phe mutation has been frequently found in both FL¹⁹⁹ and in PTNFL²⁰¹ of adult patients, while the variants uncovered in *NSD1* and *RSF1* were previously unreported. Mutations in epigenetic modifiers have been formerly considered nearly exclusive of FL of adults²⁰¹, although one study reported *KMT2D* as the most frequently mutated histone-modifying enzyme²⁰² and its disruption has been previously proven to alter germinal center B-cell development and promote lymphomagenesis¹⁸⁹. Taken together, these data confirm that epigenetic mechanisms, hit in 2/7 of the cases analyzed, can contribute to PTNFL pathogenesis. Interestingly, both the PFLT cases profiled by WES (W6 and W10) carry mutations of epigenetic modifiers. Further investigation on larger cohorts is needed to clarify the role of alterations of epigenetic modifiers in PFLT.

Besides *KMT2D*, the above-described *TNFRSF14* and *IRF8* genes are also connected with germinal center B-cell development (**Figure 32A**). Moreover, LEF1 (Lymphoid enhancer binding factor 1), a transcription factor involved in regulating B-cell development, carried a p.Ala150Val variant, reported in COSMIC as pathogenic, in the Proline-rich region implicated in the activation of the protein. *PAX5* is also part of the transcriptional network, under *IRF8* control, that orchestrates B-cell lineage specification, commitment, and differentiation. The p.Gly183Ala mutation uncovered in this cohort is very close to the p.Gly183Ser *PAX5* variant, previously shown to impact on B-lymphoid development reducing the transcription factor activity²¹⁶. Mutations of the genes directly linked to B-cell differentiation and/or germinal center reaction occurred in three cases of our cohort, two of which carried also hit in the other three identified pathways, namely the “*negative regulation of MAPK*”, “*G-protein coupled receptor*”, and “*chromatin modifying enzymes*”, which are mutated and putatively altered in most cases. Of note, most of the mutations on genes directly linked to B-cell development occurred in patient W1, which might molecularly resemble a more aggressive lymphoma. Since the subject has been recently cured with complete resection only, and 10 years of follow-up have been planned with annual monitoring.

In conclusion, beyond confirming a few mutations previously reported in PTNFL, we identified several novel variants in genes involved in negative regulation of MAPK and provided evidence on the involvement of GPCR downstream signaling in PTNFL pathogenesis. We also detected mutations in genes encoding chromatin modifying enzymes in 2/7 PTNFL and in 2/2 PFLT patients, some of them hitting genes previously not associated to FL of the pediatric age. Our analysis at network level considerably extended previous data on the mutational landscape of FL of the pediatric age, further indicating the signaling pathways of possible pathogenic relevance in these malignancies.

6.5. Genes and pathways linked to HR-NB aggressiveness

Selected group-specific somatic variants

Neuroblastoma is an embryonic tumor arising from primitive neural crest cells and accounting for 9% of pediatric cancers. It is characterized by a remarkable clinical heterogeneity with low recurrence of driver genes (*MYCN*^{165,217}, *ALK*^{165,218}, *ATRX*^{159,165,218}). HR-NB patients are characterized by metastatic disease (stage M) e show an overall survival lower than 40% at 5 years from diagnosis²¹⁹. Even if HR-NB patients respond well to the first line therapy, most of them relapse. Stigliani et. al²¹⁷ divided a cohort into short-survival (SS; patients with disease

progression and survival at most 5 years from diagnosis) and long-survival groups (LS; responsive to the therapy and survival over 5 years from diagnosis) and showed that SS group is characterized by high number of structural CNAs and high chromosomal instability. To further clarify the biological basis of disease aggressiveness focusing on SNVs and indels, I analyzed WES data of a sizeable cohort composed by 29 matched tumor-control samples (14 SS and 15 LS) affected by HR-NB at stage M.

Alignment of 2,189,622,787 reads to the reference exome (37.1 million reads per sample on average) yielded a 97x average coverage and a 76.29% of the target exome with at least 30x coverage, ranging from 53% to 84% in different patients. Sequence coverage in the SS and LS groups, considering both the tumor and peripheral blood cell samples, was homogeneous (**Figure 34**).

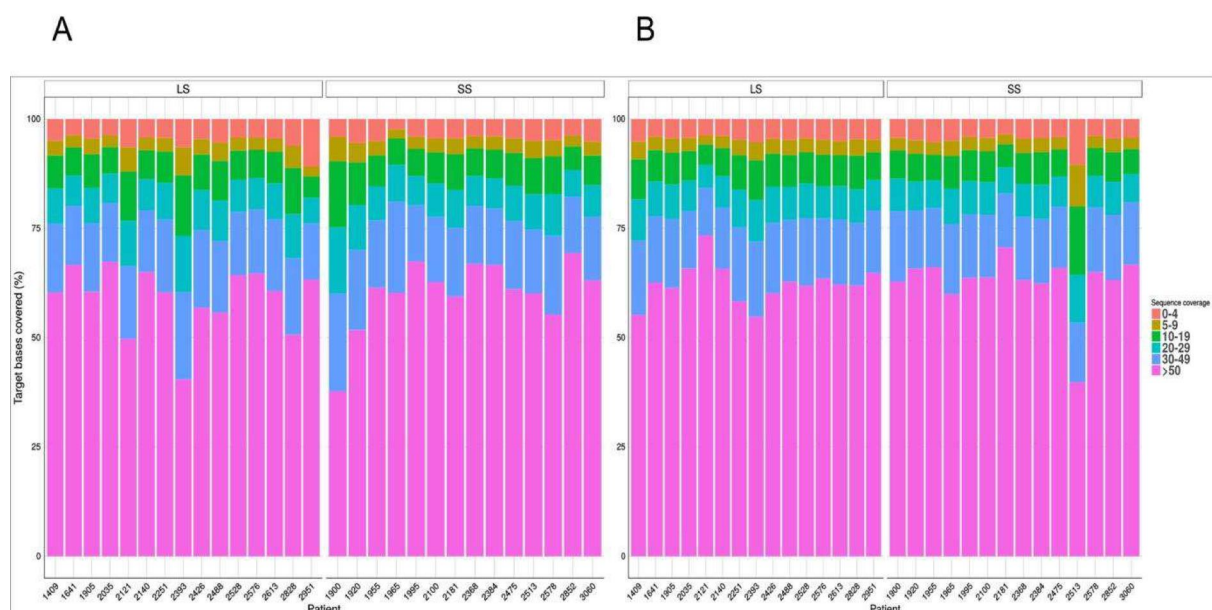


Figure 34. Sequencing coverage profile of the samples included in the LS and SS patient groups. The barplot displays the percentage of targeted bases coverage per patient, with different colors representing different coverage ranges (see legend) in the tumour (A) and control (B) samples.

A total of 2,301 and 1,805 high quality and coverage somatic variants for the SS and LS groups, respectively, were detected after read mapping, variant calling and identification of somatic variants comparing the tumor and control data. In accordance with the mutations types observed before^{159,165,220}, somatic variants resulted in enrichment in C > A (LS = 32.3%, SS = 25.2%) transversions at TCT sites and in C > T transitions (LS = 20.3%, SS = 22.6%) at GCG trinucleotide substitution types, normally due to deamination of 5-methylcytosine (**Figure 35**).

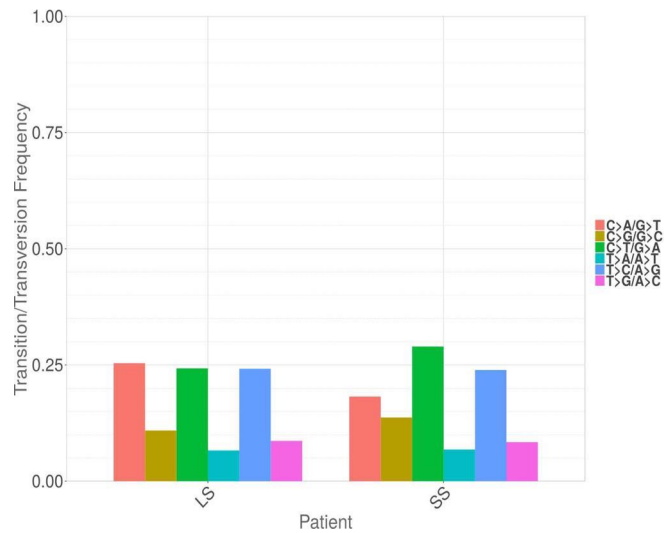


Figure 35. Mutation spectrum of high coverage and quality somatic variants. For each patient group, the plot shows the proportions of the six transition and transversion categories observed among the SNVs.

Next, 1,288 high-quality, detrimental and rare somatic variants in 1,043 genes, 580 variants detected in LS patients and 708 in SS passed the variant effect- and frequency-based filtering steps. Variants were later examined considering recurrence, hit gene and pathways, and the possible impacts of mutations on disease progression. Variant effects, group-exclusivity, intra-group recurrence, gene products biological function and relationships among the mutated genes were used to prioritize group-specific variants for validation. The selected variants confirmed are 50 in 49 genes (**Table 8**).

GENE	ID	PROTEIN	METHOD OF VALIDATION	GROUP	
LARP4	1409	p.Lys650Glu	SANGER	LS	
HMCN1	1905	p.Tyr2004Ter			
PTPRE	1905	p.Glu428Asp			
FBN1	1905	p.Glu1297Asp			
PIK3C2B	2121	p.His626Asn			
ARFGEF2	2121	p.Glu1558Ter			
CREBBP	2393	p.Ala407Ser			
PCCB	2393	p.Asn339Ile			
C10orf53	2488	p.Ala121Asp			
SCN11A	2488	p.Trp666Ter			
YTHDC1	2488	p.Val15Phe			
PKMYT1	2528	p.Gln111His			
ABCA8	2528	p.Val171Leu			
THAP9	2613	p.Leu444Pro			
PCDH12	2613	p.Leu726fs			
RAP2B	2951	p.Ser83Thr p.Leu84fs			
OR5T1	2951	p.Ile294Phe			454 Roche
MYCN	2393	p.Pro44Leu			
TBL1X	1905	p.Gly568Cys	SANGER	SS	
SMO	2578	p.Arg451Gly			
INTS2	1965	p.Pro754His			
PHGDH	2100	p.Gln416Lys			
CXXC1	2368	p.Cys395Tyr			
NTNG2	2384	p.Asp156Glu			
GREB1 rs	1965	p.(=)			
NME4	1965	p.Gly178Trp			
CACNA1G	2852	p.Val174Leu			
GPR45	2181	p.Tyr60Ter			
FGFR1	2100	p.Asn544Lys			
FGF4	2852	p.Glu159Asp			
SLC9A9	2368	p.Leu94Met			
KMT2C	2852	p.Gln3534Leu			
VCL	2475	p.Asn531Thr			
BBS10	2513	p.Asp70Tyr			
NALCN	2368	p.Ser278Leu			
NID2	3060	p.Asp1036fs			
SPTLC2	2513	p.Gly33Val			
AK7	2852	p.Ala722Leu			
SLC12A1	2475	p.Pro412Leu			
CHD2	2852	p.Trp151Cys			
SFI1	2100	p.Arg261Trp			
ARHGEF11	2384	p.Arg812Gln			
ANK3	1955	p.Tyr923Ter			454 Roche
LAMA2	1955	p.Glu352Ter			
CCR3	2513	p.Ala98Asp			
PTH2R	2513	p.Ser82Ter			
PTK2	2181	p.Arg569Leu			
PTPRA	2578	p.Asp368Tyr			
SMARCA4	2513//3060	p.Arg468Cys//p.Arg906His			

Table 8. Technically validated variants in genes somatically mutated only in one of the considered NB patient groups.

SS and LS patients have similar mutational landscapes

Comparing the number, type and effect of somatic mutations observed in the two patient groups, no significant differences were observed.

in the numbers of somatic variants per patient in the SS (median 37) and LS (36) groups were observed (Wilcoxon test p -value = 0.98; **Figure 36A**).

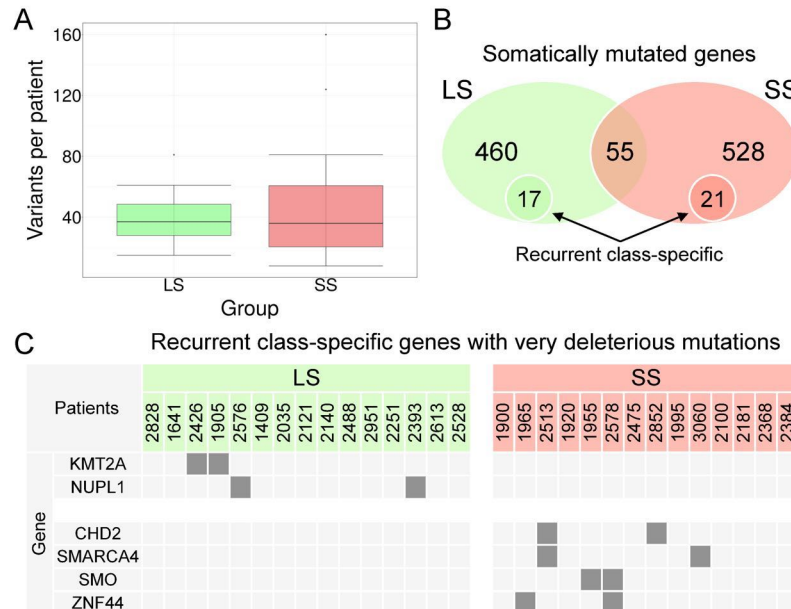


Figure 36. Comparison of mutation landscapes in SS and LS NB patients. (A) The number of variants in LS and SS groups were not significantly different, as shown by the boxplot of the distribution of selected somatic variants per patient ($p = 0.98$ of Wilcoxon test of median equality, conducted after Shapiro–Wilk test of normal distribution p -value = 3.79×10^{-5}). (B) Venn chart of number of somatically mutated genes in LS and SS groups, showing class-specifically mutated genes and their subset of genes being both class-specific and recurrent intra-class. (C) Mutation matrix indicating in which class and patients are mutated class-specific and recurrent genes, hit by particularly deleterious mutations.

Very close numbers of selected somatic variants per Mb were observed in the two groups (median values of 0.62 and 0.64, respectively). The number of somatic variants per patient in this cohort was higher than previously reported^{159,165,218,220}, but a direct comparison between WES studies was hampered by several factors, on top the different sequencing depth or technology and the different analysis methods and settings used. Moreover, similar patterns in the SS and LS patients were present considering the variant type (**Figure 37A**; G test $p = 0.11$) and predicted variant effect on the protein sequences (**Figure 37B**, G test $p = 0.06$) without evident of differences in relation to different therapy responses and outcomes.

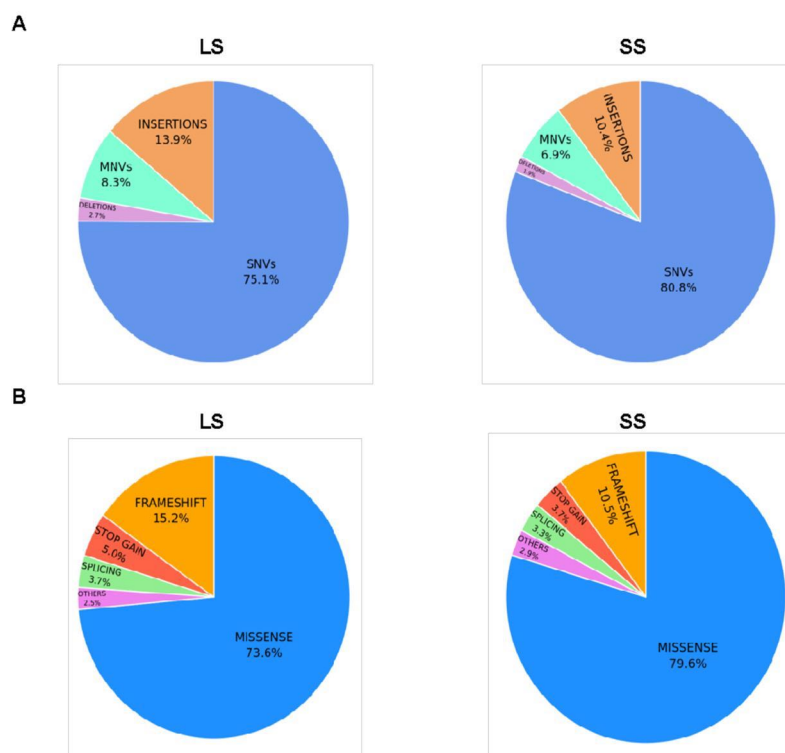


Figure 37. Comparison of type of selected variants detected in the LS and SS patient groups. Proportions of variants of different type and annotation observed in the two patient groups. Cake plots show A) the proportions of variants being Single Nucleotide Variants (SNV), Multiple Nucleotide Variants (MNP) or indels in the LS and SS patient groups, and B) the proportions of variants by annotations class (missense, frameshift, stop gain and splicing) in the same groups.

Mutations exclusive of SS or LS

The above-cited 708 and 580 high-confidence damaging rare variants detected in SS and LS patients, fell into 583 and 515 individual genes, respectively. Of these, 690 and 562 variants are group-specific. Only 102 (9.8%) out of 1,043 genes mutated in the whole NB cohort were recurrently mutated in two or more patients. Fifty-five genes were recurrently mutated in patients of both classes (**Figure 36B**), including 17 genes detected in more than two patients, with *DPCR1* (mutated in nine patients), *AHNAK2* (6), and *CBX4* and *ZNF717* (both mutated in four patients) being the most recurrently observed.

Notably, 528 genes were specifically mutated in SS and 460 in LS patients (**Figure 36B**), including 21 and 17 recurrent and group-specific genes, respectively. Six of these genes carried particularly damaging variants (**Figure 36C**). Only *KMT2A* (Lysine Methyltransferase 2A; E2926Q in patient ID2426; S3291C in patient ID1905) and *NUPL1* (Nucleoporin 58; N153fs in patient ID2393 and ID2576) resulted in recurrently mutated and group-specific pathways LS patients. In SS patients, four genes (*SMO*, *SMARCA4*, *ZNF44* and *CHD2*), all known to be expressed in neural tissues, were recurrently mutated, group-specific and carried variants not

described before and particularly deleterious. *SMO* (Frizzled Class Receptor Smoothened) encodes nonclassical G-protein-coupled receptors that are highly expressed in neural tissues and involved in Hedgehog signaling. Two *SMO* variants (**Figure 38**) were detected, R451G in ID2578 in the Frizzled/Smoothened family membrane region, and T640fs (ID1955) that induces a premature stop codon ending the protein 132 amino acids before the C-terminus. The other recurrent SS-specific genes, *CHD2* and *SMARCA4*, are transcriptional regulators. *SMARCA4* somatic variants (R468C in patient ID2513; R906H in ID3060; **Figure 38**) in two SS patients were validated. The deleterious *SMARCA4* R906H mutation was annotated in the COSMIC database (COSM5576007) as previously being observed in gastric cancer²²¹, whereas the putatively damaging R468C variant was not described previously. Both *SMARCA4* variants were localized in the transcription activator chain of the protein falling, respectively, in the Helicase Sant-associated domain (HSA) and in the helicase ATP-binding domain of the protein (**Figure 38**), responsible for DNA and ATP binding and ATP hydrolysis. SS-specific *FGFR1* N577L (detected in ID2100) and *PTK2* R569L (detected in ID2181) variants were validated and described in a previous NB study¹⁵⁹. These two genes were already associated with NB tumorigenesis^{159,220} although not specifically in relation to patient survival. The *PTK2/FAK1* (focal adhesion kinase) variant is located close to the Tyr576 phosphorylation site of the kinase domain “catalytic loop” (**Figure 38**) required for *PTK2* activation²²² and for mediating NB progression and aggressiveness²²³.

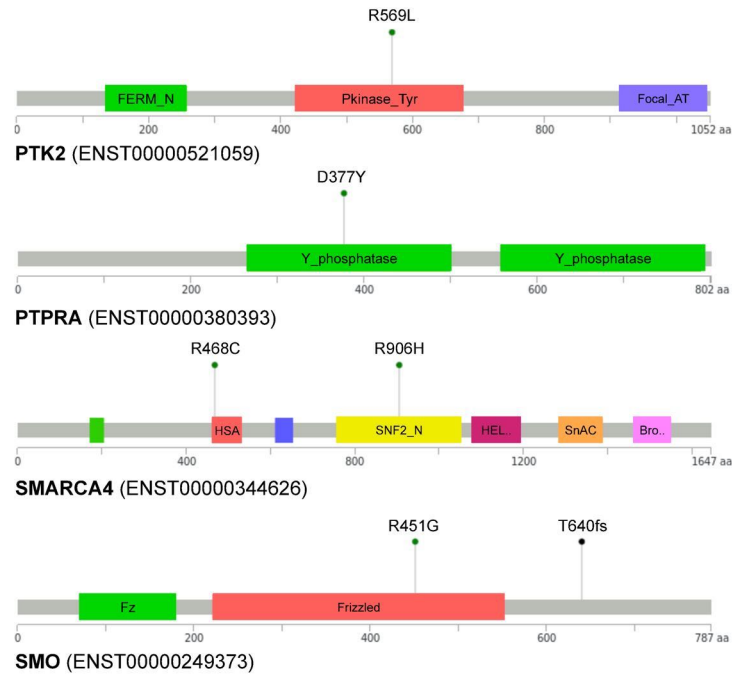


Figure 38. Impact on proteins of somatic variants detected in NB patients in PTK2, PTPRA, SMARCA4 and SMO genes. For each gene, considering the protein encoded by the reference transcript, lollipop plots show the type and the position of somatic variants in relation to the protein sequence and domains (colored portions) according to Pfam annotation (<http://pfam.xfam.org/>); different lollipop colors indicate variant annotation types.

Different pathways are altered in HR-NB subgroups

Group-specific pathways were investigated considering mutated genes annotated in Reactome (48%) and KEGG (24%). Thirty-Four Reactome and 12 KEGG pathways were specifically enriched in the LS group, and 12 Reactome and 17 KEGG pathways were recurrently mutated in LS patients and never mutated in SS (LS-specific). In SS patients, 17 Reactome and one KEGG pathways were specifically enriched, and 25 Reactome and 4 KEGG pathways were SS-specific. The genes somatically mutated in each of the two patient groups tend to participate to different pathways and pathway classes indicating a link between disease aggressiveness and specific functions hit by mutations (**Figure 39**). Mutations in 15 cell cycle genes were present in LS patients, with several deleterious mutations in genes linked to mitosis, including *CDC27*, *CDCA5*, *CENPC* and *AURKB*. Notch-related genes are mutated in both groups, but several genes (*TBLIX*, *CREBBP*, *NOTCH4*, *NOTCH3*, *TNRC6B* and *TLE2*) specifically belonging to NOTCH1 signaling are mutated only in LS patients.

In SS patients, the axon guidance pathway was hampered by mutations in 20 genes, involved particularly in NCAM signaling for neurite outgrowth by MAPK2 and MAPK activation

(including *ARHGEF11*, *CACNA1G*, *FGF4*, *PTPRA*, *PTK2*, *ANK3*, *SMO* and *NTNG2*) processes, which is important for neurodevelopment and oncogenesis. The MAPK pathway is linked through PTK2 signaling to ERBB4 (including *FGF4*, *PTPRA* and *PTK2*), and MET (*LAMA2*, *PTK2* and *LAMA4*). ERBB4 signaling was specifically enriched, and both MET signaling and the Cilium assembly pathway (*BBS10*, *SMO*, *INPP5E*) were exclusively mutated in SS patients.

Class	Pathway		Genes		
DISEASE	Diseases of signal transduction	Oncogenic MAPK signaling	Paradoxical activation of RAF signaling by kinase inactive BRAF Signaling by high-kinase activity BRAF mutants Signaling by moderate kinase activity BRAF mutants Signaling by RAS mutants	FGG, VCL, KSR2	
DEVELOPMENTAL BIOLOGY	Axon guidance	NCAM signaling for neurite out-growth	MAP2K and MAPK activation	ARHGEF11, CACNA1G, KCNQ2, RPS6KA4, CACNA1H, FGG, FGF4, VCL, PTPRA, SCN4A, PLXNA3, CLTCL1, PTK2, ANK3, SPRED3, KSR2, SEMA6D, ROBO1, PTPRC, RGMA	
		RET signaling	MAP2K and MAPK activation	FGG, VCL, KSR2	
EXTRACELLULAR MATRIX ORGANIZATION	Keratinization	Formation of the cornified envelope		PPL, CELA2A, EVPL, KLK5, PKP3	
	Collagen formation	Assembly of collagen fibrils and other multimeric structures		PLOD2, TLL1, COL1A1, COL5A2, COL11A1, COL11A2, COL18A1, P3H3, COLGALT1, COL12A1	
		Collagen biosynthesis and modifying enzymes		TLL1, PRSS1, COL18A1, MMP14, CAPN9, CAPN15, COL12A1	
Degradation of the extracellular matrix	Activation of Matrix Metalloproteinases		HSPG2, ITGA7, COL13A1, ITGB8, FBN1		
SIGNAL TRANSDUCTION	Integrin cell surface interactions			GH2, FGG, FGF4, VCL, PTPRA, PTK2, SPRED3, KSR2, TNRC6C, GFAP, RPS6KB2	
	Signaling by ERBB4			LAMA2, PTK2, LAMA4	
	Signaling by MET	MET promotes cell motility	MET activates PTK2 signaling	FGG, VCL, KSR2	
	MAPK1/MAPK3 signaling	RAF/MAP kinase cascade	MAP2K and MAPK activation	TBL1X, CREBBP, NOTCH4, NOTCH3, TNRC6B, TLE2	
	Signaling by NOTCH	Signaling by NOTCH1	NOTCH1 Intracellular Domain Regulates Transcription	ZWINT, CENPC, CENPP, KNL1, AJRKB	
TRANSPORT OF SMALL MOLECULES	SLC-mediated transmembrane transport	Transport of inorganic cations/anions and amino acids/oligopeptides		SLC2A4, SLC12A1, SLC5A8, SLC38A4, POM121, SLC4A9, NUP85, SLC40A1, SLC01B3, SLC9A9, SLC17A6	
	Ion channel transport	Stimuli-sensing channels		RYR1, TRDN, NALCN, TPCN2, UNC79	
GENE EXPRESSION	RNA Polymerase I Transcription			ERCC6, TAF1B, PTRF	
ORGANELLE BIOGENESIS AND MAINTENANCE	Cilium assembly	Cargo trafficking to the periciliary membrane		BBS10, SMO, INPP5E	
	Metabolism of Lipids	Phospholipid metabolism		INPP1, ACHE, PNPLA7, TNFAIP8L1, PCYT2, OSBPL5, INPP5E	
METABOLISM	Nucleotide metabolism	Synthesis and interconversion of nucleotide di- and triphosphates		AK9, AK7, NME4, DPYS	
	Carbohydrate metabolism	Glycosaminoglycan metabolism	Heparan sulfate/heparin (HS-GAG) metabolism	MGAM, GLCE, ENO3, SLC9A1, NAGLU, PFKP, GLYCTK, NUP93, UST, MAN2B2, HSPG2, KERA, CSPG4	
	Metabolism of vitamins and cofactors			APOB, PCCB, TCN1, HSPG2, LRP1, PTGIS	
METABOLISM OF PROTEINS	Translation			EIF3G, RPL12, RPS7	
	Post-translational protein modification	Deubiquitination	UCH proteinases	HCFC1, NFRKB, INO80	
		Asparagine N-linked glycosylation	Transport to the Golgi and subsequent modification		MIA2, SPTB, F5, GRIA1, DCTN5, COG6
		SUMOylation	SUMO E3 ligases SUMOylate target proteins	ER to Golgi Anterograde Transport Cargo concentration in the ER	CBX4, MDC1, NUP93, AJRKB, SP100, MITF
	Peptide hormone metabolism			MME, CPA3, CPN1, ISL1	
IMMUNE SYSTEM	Adaptive Immune System	Costimulation by the CD28 family		EREG, PHLPP1, HLA-DQA2, PRKCB, LYN, CDC27, CD80, CD79B, AP1S1, ZBTB16, ITK, TRIP12, HECTD2, RNF25, DCTN5, TNRC6B, LILRA1	
HEMOSTASIS	Platelet activation, signaling and aggregation			F13A1, APOB, PRKCB, LYN, F5, JAK3, CREBBP, GNAS, PRKCZ, A2M, DGKZ, MAG, DOCK4	
REPRODUCTION	Fertilization			ADAM20, IZUMO2, CATSPER1	
NEURONAL SYSTEM	Transmission across Chemical Synapses	Neurotransmitter Receptor Binding And Downstream Transmission In The Postsynaptic Cell		ADCY3, PRKCB, GRIA1, ARHGEF9, CHRNG	
CELL CYCLE	Cell Cycle, Mitotic	M Phase	Mitotic Metaphase and Anaphase Separation of Sister Chromatids Mitotic Prometaphase Resolution of Sister Chromatid Cohesion	CEP290, ZWINT, PCNT, PRKCB, CDC27, CENPC, CENPP, NCAPG2, NUP93, KNL1, CDCA5, AJRKB, LIN37, PKMYT1, CDC14A	

Figure 39. Summary of Reactome pathways exclusively mutated or exclusively enriched in LS or in SS NB patients. The figure depicts the hierarchy of Reactome pathways that resulted or contained pathways exclusively mutated in LS (green fill) or in SS (light red fill) patients, or that were enriched in a class-specific way (bold text); the gray fill indicates more general classes at high hierarchical level being not class-specific; for the most high-level specific or specifically enriched class of each group, the corresponding mutated genes are indicated in the right part of the figure.

Only mutations of SS patients tend to cluster into specific subnetworks

A further analysis of the topological structure of mutation gene networks derived from the Reactome pathway annotation detected significant functionally-connected gene networks in HR-NB SS patients, which were in accordance with previous observations at the pathway level. Neither the 463 genes mutated in the whole cohort (adjusted p-value of global mutation clustering 0.43) nor the 213 genes mutated in LS patients (adjusted p-value 0.93) showed significant clustering according to Hotnet2 analysis. Conversely, a more pronounced clustering was observed of the 268 genes mutated in SS patients, 79 of which converged into 18 subnetworks of at least three genes (adjusted p-value of global mutation clustering 0.24). The six most relevant network components, comprising 31 functionally connected genes that are somatically mutated specifically in SS patients, were selected (**Figure 40**). The largest component, which was recurrently identified in almost two thirds of SS patients (9 of 14), included nine genes (*NID2*, *LAMA4*, *LAMA2*, *PTK2*, *PTPRA*, *FGG*, *VCL*, *MMP14* and *KSR2*) of the RAF/MAPK signaling pathway and extracellular matrix organization. In addition to the previously observed PTK2 variant¹⁵⁹, the D377Y variant was validated, which fell into the Y phosphatase domain of PTPRA (**Figure 38**), closely connected with PTK2 in the RAS/MAPK pathway, and the stop gaining variant (E352*) of LAMA2.

A second component of six genes (*NALCN*, *UNC79*, *SLC9A9*, *SLC12A1*, *SLC5A8* and *SLC4A9*) that was linked to the transmembrane transport of small molecules was mutated in four SS patients. Two patients carried mutations in two genes of the component (*NALCN* and *SLC9A9* co-mutated in ID2368; *SLC5A8* and *SLC4A9* in ID1955). The third component, which was linked to centrosome maturation, included five genes (*CDK5RAP2*, *CDK11A*, *CEP89*, *TUBGCP6* and *SF11*) mutated in three different patients (ID1965, ID2100, ID2384) (*CEP89*, *TUBGCP6*, and *CDK11A* co-mutated in the patient ID1965).

Four genes were involved in lipid and lipoprotein (*SPTLC2* and *ACSL6*) or nucleotide (*AK7*, *AK9* and *ACSL6*) metabolism, which were mutated in three SS patients (ID2368, ID2513, ID2852), with *SPTLC2* and *ACSL6* in the same patient. Two additional SS-specific components were defined by *SMO*, recurrently mutated in two patients and functionally connected *BBS10* and *GAS8* genes co-mutated in a third, and by *KMT2C*, *HOXB3*, and *HOXC4* mutated in ID1965 and ID2852 patients.

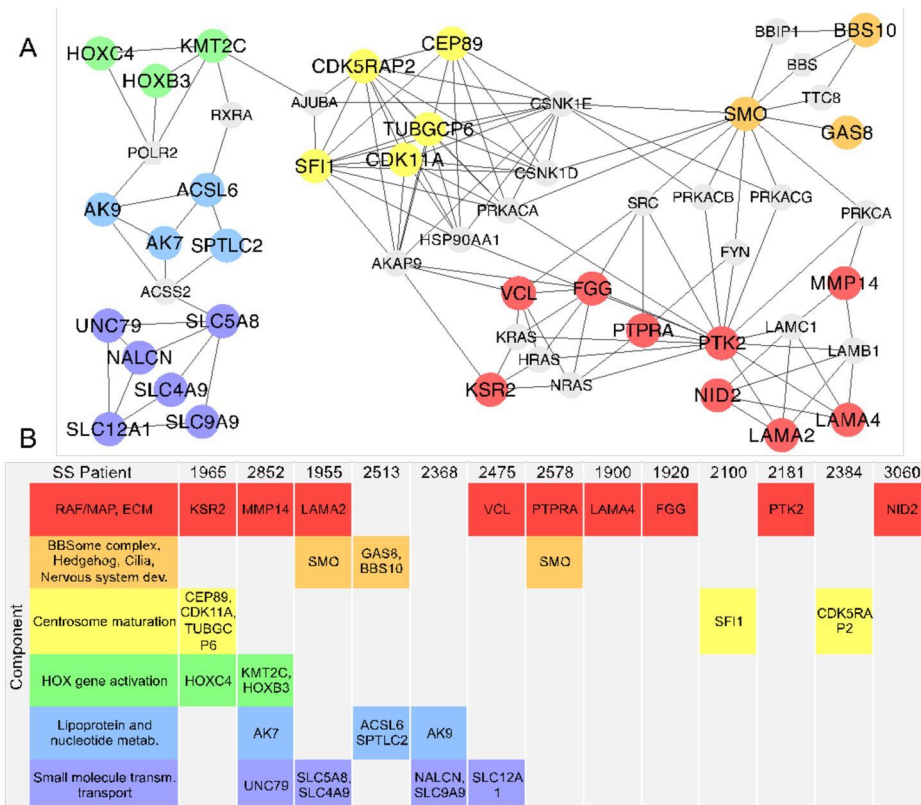


Figure 40. Reactome-derived network of genes somatically mutated formed by six SS-specific components. (A) Colored nodes in the net indicate genes mutated in SS patients, with six different components (groups of functionally connected genes somatically mutated in NB patients with rapid disease progression identified by Hotnet2 analysis) in different colors; gray nodes represent genes directly connecting the components according to pathway topology and non-mutated in the analyzed cohort (edges between gray nodes are omitted). (B) Each component was recurrently mutated in different patients, and specific tumors carried mutations in multiple genes and components of the network.

Large independent cohort

SS-specific genes, pathways and component identified in the study were compared to the findings in Pugh cohort (**Figure 40**). Of the genes with SS-specific recurrence in study cohort, *NFATC1* and *OR14J1* were recurrent with SS-specificity also in the Pugh cohort. Furthermore, five genes (*CHD2*, *DIDO1*, *KRTAP4-8*, *ZNF44* and *ZNF91*) with SS-specific recurrence in study cohort were mutated with SS-specificity also in Pugh patients (**Figure 40A**). The SMARCA4 mutation was reported in a patient with a survival of 61 months according to survival classification. Nineteen of the 78 genes prioritized because involved in SS-specific pathways identified (**Figure 38**) were mutated with SS-specificity in the Pugh cohort, including five (*ANK3*, *COL11A1*, *COL12A1*, *COL1A1*, *PNPLA7*) that were also recurrent. Five genes (*AK7*, *NALCN*, *PTK2*, *SLC5A8*, *TUBGCP6*) included both in SS-specific pathways and in significant subnetworks (**Figure 39**) identified in the study were mutated only in SS patients of the Pugh cohort. Furthermore, analysis with HotNet2 was performed considering the 1,810

genes somatically mutated in SS patients of the Pugh cohort and mapped in the Reactome-derived network, detecting 14 significant (adjusted p value 0.05) subnetworks involving 143 genes. Extracellular matrix organization, carbohydrate and lipid metabolism emerged from both studies' data. *PTK2*, which was shown to be mutated with SS-specificity both in study data and in the Pugh cases, in the network of mutations detected in Pugh patients was directly linked to two gene groups involved in extracellular matrix organization (**Figure 40B**).

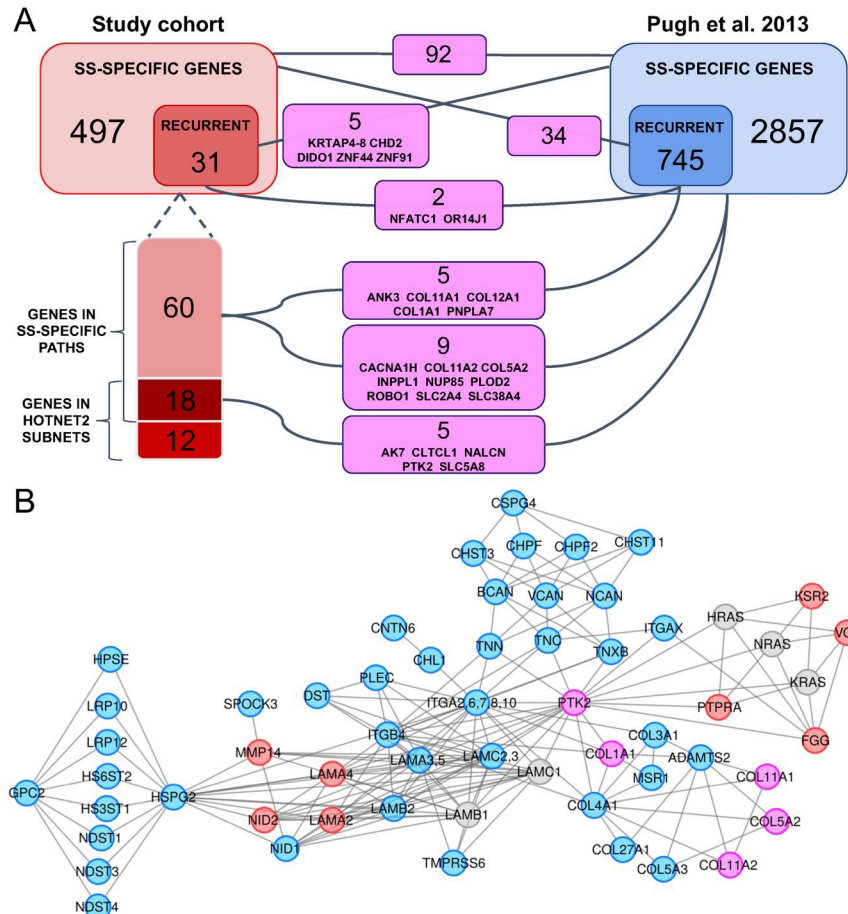


Figure 38. Overlap and functional connections of SS-specific genes prioritized in this study and genes mutated in SS patients in Pugh cohort. A) The flowchart shows the overlap of genes mutated with SS-specificity in the study cohort, and of subsets prioritized according to recurrence, or by pathway and/or HotNet2 analysis, with genes mutated in SS patients of the 240 stage M NB Pugh et al. (2013) cohort. B) Meta-network obtained merging components related to ECM identified by parallel HotNet2 analysis of genes mutated in study cohort (red nodes) and in Pugh cohort (blue nodes); Violet nodes show common genes, and network topology, such as edges and first neighbour genes (not mutated, grey nodes), outlines functional connections between genes identified in one or in both cohorts.

Discussion

One of the major challenges for oncologists treating HR-NB is the high percentage of patients showing rapid disease progression despite multimodal treatment. Of these, approximately 60% of HR-NBs have a fatal course within 5 years of diagnosis. To identify genetic abnormalities

associated with disease aggressiveness, somatic mutation profiles of HR-NB patients with SS and LS were compared. General tumor genomic landscapes of HR-NB patients with SS and LS were similar, exhibiting close frequencies of variants and numbers of somatically mutated genes per patient. Nevertheless, few genes were recurrently mutated specifically in the SS group, including *SMARCA4*, *SMO*, *ZNF44* and *CHD2*. Chromodomain Helicase DNA Binding Protein 2 (*CHD2*) is important for neurogenesis and de novo mutations in this gene were found in neurodevelopmental disorders²²⁴. *CHD2* is a tumor suppressor chromatin remodeler, previously observed to be mutated and proposed as a cancer driver in chronic lymphocytic leukemia²²⁵. Notably, the transcription co-activator and tumor suppressor *SMARCA4* (SWI/SNF Related, Matrix Associated, Actin Dependent Regulator of Chromatin, Subfamily A, Member 4) were recurrently mutated in the SS group. *SMARCA4* encodes a member of the SWI/SNF nucleosome-remodeling complex whose mutations impact growth control, differentiation, development and cell adhesion²²⁶. *SMARCA4* also known as BRG1, is a tumor suppressor gene^{227–229} that shows inactivating mutations or overexpression in several adult cancers^{230,231}. Jubierre et al. showed that the *SMARCA4* gene has a role in the proliferation of NB cells both in vitro and in vivo²³². A correlation between *SMARCA4* mutations and loss of function in lung cancer cell lines was observed, indicating an association with aggressive tumor behavior and worse patient survival²³³. As it has been demonstrated that *SMARCA4* and *TERT* are functionally linked²³⁴, SWI/SNF damage could alter *TERT* function²³⁵, which one of the most important genes rearranged in NB²³⁶.

Remarkably, somatic mutations occurring in SS or LS patients hit different pathways. In addition, functional gene networks, corresponding to sub pathways, hit only SS patients. Numerous gene variants observed in the tumor of SS patients affected the RAF/MAP kinase cascade, as well as MET and ERBB4 pathways linked to PTK2 signaling. MAP2K and MAPK activation, specific of SS tumors, are of interest because they can be involved in cell motility by triggering PTK2 signaling and Matrix Metalloproteinases activation. These results agree with previous data on the enrichment of somatic mutations in FAK signaling and cell adhesion signaling¹⁵⁹.

Furthermore, several genes connected to RAF/MAPK signaling were mutated in SS (*NID2*, *LAMA4*, *LAMA2*, *PTK2*, *PTPRA*, *FGG*, *VCL*, *MMP14* and *KSR2*), impacting extracellular matrix organization, regulation of cell adhesion and migration. A previous observation of *PTK2* mutation in a HR-NB patient, further strengthens the importance of PTK2 signaling in aggressive tumors. The validation of *PTK2* and *PTPRA* mutations, network data and previous

observation of mutated PTK2¹⁵⁹ further strengthen the importance of PTK2 signaling in aggressive tumors.

Particularly relevant groups of clustered genes mutated in SS were involved in centrosome maturation, in the regulation of the cell cycle, in ciliary basal body docking (*CDK5RAP2*, *CDK11A*, *CEP89*, *TUBGCP6* and *SFII*) and in cilium assembly (the recurrently and SS specifically mutated *SMO*, and *GAS8*, *BBS10*). The observation of mutations linked to the chromosome remodeling pathway in SS tumors support the role of chromosome instability in NB²³⁷, providing further explanation for the observed CNA in patients with fatal outcomes^{217,238}.

The obtained results from the analysis of a sizeable independent cohort of 240 stage M NB patients¹⁶⁵ gave additional strength to findings. Mutations in Pugh SS patients targeting genes prioritized in the cohort (*ANK3*, *COL11A1*, *COL12A1*, *COL1A1*, *PNPLA7*, *AK7*, *NALCN*, *PTK2*, *SLC5A8* and *TUBGCP6*) as SS-specific based on recurrence, pathway enrichment and/or pathway- derived network topology analysis, were particularly noteworthy and supported the results. The reconstruction and analysis of pathway-derived mutation networks reported in Pugh SS patients further backed the observations done in study cohort about the deregulation of lipid metabolism and RAF/MAP signaling in relation to extracellular matrix organization mutated genes.

Potential targets for pharmacological therapies of more aggressive HR-NB

Recent comparison of matched primary and relapsed NB tumors revealed that disease progression is accompanied by an increased mutational load, exhibiting new mutations in the MAPK pathway that were not present at the onset of disease, and accumulated in tumors of relapsing patients^{239,240}. These findings of specific MAPK signaling pathway damages may be relevant for more efficacious therapeutic management of patients at diagnosis. Specific genes mutated at diagnosis exclusively in pathways belonging to the SS group could be candidates for pharmacological targeting. *SMO*, *PTK2*, *MMP14* and *SDHB* are quite interesting as they are targeted by FDA-approved drugs according to the Drug Gene Interaction Database (DGID, <http://dgidb.genome.wustl.edu/>) (**Table 8**).

Gene ID	Gene Description	Variant	FDA approved drugs	Drug class	References for drugs use
SDHB	Succinate dehydrogenase complex, subunit B, iron sulfur (Ip)	L7FS	Succinic acid	Small molecule	He et al., 2004, Citric acid cycle intermediates as ligands for orphan G-protein-coupled receptors., Nature; Southern et al., 2013, Screening β -arrestin recruitment for the identification of natural ligands for orphan G-protein-coupled receptors., J Biomol Screen
SMO	Smoothed, frizzled family receptor	R451G; T640FS	Vismodegib	Small molecule inhibitor	Yauch et al., 2009, Smoothed mutation confers resistance to a Hedgehog pathway inhibitor in medulloblastoma., Science; Wang et al., 2012, Identification of a novel Smoothed antagonist that potently suppresses Hedgehog signaling., Bioorg. Med. Chem.
			Fluocinonide	Small molecule	Wang et al., 2010, Identification of select glucocorticoids as Smoothed agonists: potential utility for regenerative medicine., Proc. Natl. Acad. Sci. U.S.A.
			Halcinonide	Small molecule	Wang et al., 2011, Glucocorticoid hedgehog agonists in neurogenesis., Vitam. Horm.; Wojnar et al., 1986, Androstene-17-thioketals. 1st communication: glucocorticoid receptor binding, antiproliferative and antiinflammatory activities of some novel 20-thiasteroids (androstene-17-thioketals),. Arzneimittelforschung
PTK2	PTK2 protein tyrosine kinase 2	R569L	Masitinib	Kit inhibitor	Dubreuil et al., 2009, Masitinib (AB1010), a potent and selective tyrosine kinase inhibitor targeting KIT., PLoS ONE
MMP14	Matrix metalloproteinase 14 (membrane-inserted)	P8FS	Prinomastat	Mmp inhibitor	Abbenante et al., 2005, Protease inhibitors in the clinic., Med Chem

Table 8. Information on drugs available in relation to genes carrying deleterious mutations in NB patients.

Recently, Padovan-Merhar et al.²⁴¹ reported an increased SMO mutation frequency in tumors of HR-NB patients at relapse, showing that most of these new mutations are targetable and give an additional tool to treat relapsing patients. Functional investigation is mandatory to assess the potential significance of mutated genes as therapeutic targets, and further study is needed to evaluate drugs, such as Masitinib and Vismodegib, for NB therapy.

Two groups of HR-NB patients with different outcome were characterized, providing new data on mutations recurrently affecting specific pathways and functions in patients with SS, informing the molecular features, beyond well-defined CNA patterns, that are associated with high tumor aggressiveness. These data may help to address an early treatment of HR-NB patients using FDA-approved compounds targeting the deregulated pathways and mutated genes present at onset of disease.

7. Conclusions and future perspectives

This thesis presents an automated, modular and easy-to-use computational pipeline (iWhale) for the detection and annotation of somatic variants from WES data by selecting and combining the different software tools. Moreover, a systems genetics approach to reconstruct and to statistically analyze pathway-derived networks composed by functionally and directly connected genes was developed.

The application of different versions of these methods provided additional molecular information about the pathogenesis and differential disease progress of three different cancer studies resulting in new data that could become useful for diagnosis and eventually in personalized therapies.

The developed systems genetics approach has contributed to add knowledge about JAK/STAT activation, especially in LGL-L patients not affected by STAT mutations. The qualitative analysis of pathway-derived network of a series of pediatric FL confirmed the central role played by alterations of MAPK pathway and identified mutations in genes participating to GPCR and chromatin modification pathways. These findings suggested that mutations in chromatin modifiers are not exclusive for adult-form of follicular lymphoma but may also have a role in PTNFL pathogenesis.

A custom data structure for statistical analysis on pathway-derived networks with HotNet2 was setup in order to analyze large and complex gene networks, that otherwise they would be impossible to interpret. HotNet2 analysis on Reactome-derived network was performed in Neuroblastoma study to identify molecular features underlying the heterogeneity of HR-NB patients in term of survival time. This analysis resulted in the detection of 6 components associated to specific functions, such as extracellular matrix organization via MAPK pathway, primary cilium assembly, centrosome maturation, HOX gene activation, nucleotide and lipoprotein metabolism, and small molecule transmembrane transport. Some of the gene variants present at onset of the disease and included in the detected components, especially in MAPK signaling, can be targeted by FDA-approved drugs to improve early treatment for patients with more aggressive form of neuroblastoma.

WES data of the three studies were analyzed with different, progressively improved and updated versions of the pipeline now implemented in iWhale, following the continuous updating over the years of tools and methods used for analyzing cancer exome data, and also customizing the analysis according to each specific study design and aims. In all the studies the human reference genome hg19 was used due to the incompatibility with hg38 of ExAC and

gnomAD databases used for variant prioritization and since the presence of ALT contigs in hg38 would have affected the alignment step reducing variant calling sensitivity.

In LGL-L study, the somatic variants were called with an in-house method based on a not optimized software for somatic variants detection and on a subtraction of the variants called in the control samples from the ones called in matched-tumor samples resulting in a loss of low frequency variants. The subsequent integration of MuTect in the pipeline and its use for pediatric FL study produced more robust results allowing for the detection of somatic SNPs even at low frequencies. For the study of neuroblastoma cases, Torrent Variant Caller, a commercial software optimized for IonTorrent data, was preferred to detect somatic variants in order to address the difficulties of IonTorrent technology in sequencing repetitive regions giving numerous false indels.

The version of iWhale presented in this thesis is the version of the pipeline up and running at the end on 2018. In the future, iWhale will be implemented with variant prioritization steps where the variants detected by the four different callers will be automatically filtered by using information about population allele frequency, predicted functional impact, and clinical significance. The prioritization will be further improved with the implementation of software to identify statistically-significant mutated genes considering background mutation rate (MutSigCV and MuSiC) and for germline and CNV variant detection.

Although HotNet2 analysis is already rather effective, it may be improved by using, in addition to those based on mutation recurrence of mutations, more informative heat scores, considering the variant impact or scores assigned to genes that predict their driver status. Finally, I would like to automate HotNet2 analysis and data structure construction to perform custom analysis, since they are complex steps that require robust computational knowledge.

8. References

1. Sudhakar, A. History of Cancer, Ancient and Modern Treatment Methods. *J. Cancer Sci. Ther.* **1**, 1–4 (2009).
2. Torre, L. A. *et al.* Global cancer statistics, 2012. *CA. Cancer J. Clin.* **65**, 87–108 (2015).
3. Stewart, B. W. *et al.* Cancer prevention as part of precision medicine: ‘plenty to be done’. *Carcinogenesis* **37**, 2–9 (2016).
4. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet Lond. Engl.* **388**, 1545–1602 (2016).
5. Anand, P. *et al.* Cancer is a Preventable Disease that Requires Major Lifestyle Changes. *Pharm. Res.* **25**, 2097–2116 (2008).
6. Islami, F. *et al.* Proportion and number of cancer cases and deaths attributable to potentially modifiable risk factors in the United States. *CA. Cancer J. Clin.* **68**, 31–54 (2018).
7. Tomasetti, C., Marchionni, L., Nowak, M. A., Parmigiani, G. & Bert Vogelstein. Only three driver gene mutations are required for the development of lung and colorectal cancers. *Proc. Natl. Acad. Sci.* **112**, 118–123 (2015).
8. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
9. Ding, L., Wendl, M. C., McMichael, J. F. & Raphael, B. J. Expanding the computational toolbox for mining cancer genomes. *Nat. Rev. Genet.* **15**, 556–570 (2014).
10. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).
11. Wooster, R. *et al.* Identification of the breast cancer susceptibility gene BRCA2. *Nature* **378**, 789–792 (1995).
12. Miki, Y. *et al.* A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* **266**, 66–71 (1994).
13. Knudson, A. G. Mutation and cancer: statistical study of retinoblastoma. *Proc. Natl. Acad. Sci. U. S. A.* **68**, 820–823 (1971).
14. Agarwal, D., Nowak, C., Zhang, N. R., Pusztai, L. & Hatzis, C. Functional germline variants as potential co-oncogenes. *NPJ Breast Cancer* **3**, 46 (2017).
15. Poulos, R. C. & Wong, J. W. H. Finding cancer driver mutations in the era of big data research. *Biophys. Rev.* (2018). doi:10.1007/s12551-018-0415-6
16. Dang, L. *et al.* Cancer-associated IDH1 mutations produce 2-hydroxyglutarate. *Nature* **465**, 966 (2010).

17. Ward, P. S. *et al.* The common feature of leukemia-associated IDH1 and IDH2 mutations is a neomorphic enzyme activity converting alpha-ketoglutarate to 2-hydroxyglutarate. *Cancer Cell* **17**, 225–234 (2010).
18. Bakker, J. L. *et al.* A novel splice site mutation in the noncoding region of BRCA2: implications for Fanconi anemia and familial breast cancer diagnostics. *Hum. Mutat.* **35**, 442–446 (2014).
19. Bell, R. J. A. *et al.* Understanding TERT Promoter Mutations: A Common Path to Immortality. *Mol. Cancer Res. MCR* **14**, 315–323 (2016).
20. Horn, S. *et al.* TERT promoter mutations in familial and sporadic melanoma. *Science* **339**, 959–961 (2013).
21. Huang, F. W. *et al.* Highly recurrent TERT promoter mutations in human melanoma. *Science* **339**, 957–959 (2013).
22. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333–339 (2013).
23. Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**, 1134–1140 (2013).
24. Beroukhi, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905 (2010).
25. Mertens, F., Johansson, B., Fioretos, T. & Mitelman, F. The emerging complexity of gene fusions in cancer. *Nat. Rev. Cancer* **15**, 371–381 (2015).
26. Oda, Y. & Tsuneyoshi, M. Recent advances in the molecular pathology of soft tissue sarcoma: implications for diagnosis, patient prognosis, and molecular target therapy in the future. *Cancer Sci.* **100**, 200–208 (2009).
27. Tomlins, S. A. *et al.* Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310**, 644–648 (2005).
28. Yu, J. *et al.* An integrated network of androgen receptor, polycomb, and TMPRSS2-ERG gene fusions in prostate cancer progression. *Cancer Cell* **17**, 443–454 (2010).
29. Govindan, R. *et al.* Genomic landscape of non-small cell lung cancer in smokers and never-smokers. *Cell* **150**, 1121–1134 (2012).
30. Palles, C. *et al.* Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nat. Genet.* **45**, 136–144 (2013).
31. Gryfe, R. & Gallinger, S. Microsatellite instability, mismatch repair deficiency, and colorectal cancer. *Surgery* **130**, 17–20 (2001).
32. Tomasetti, C., Vogelstein, B. & Parmigiani, G. Half or more of the somatic mutations

in cancers of self-renewing tissues originate prior to tumor initiation. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 1999–2004 (2013).

33. Futreal, P. A. *et al.* A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183 (2004).
34. Gonzalez-Perez, A. *et al.* IntOGen-mutations identifies cancer drivers across tumor types. *Nat. Methods* **10**, 1081–1082 (2013).
35. Goldman, J. M. & Melo, J. V. Chronic myeloid leukemia--advances in biology and new approaches to treatment. *N. Engl. J. Med.* **349**, 1451–1464 (2003).
36. Paez, J. G. *et al.* EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science* **304**, 1497–1500 (2004).
37. Bertucci, F. *et al.* Comparative genomic analysis of primary tumors and metastases in breast cancer. *Oncotarget* **7**, 27208–27219 (2016).
38. Meric-Bernstam, F. *et al.* Concordance of genomic alterations between primary and recurrent breast cancer. *Mol. Cancer Ther.* **13**, 1382–1389 (2014).
39. Johnson, D. B., Rioth, M. J. & Horn, L. Immune checkpoint inhibitors in NSCLC. *Curr. Treat. Options Oncol.* **15**, 658–669 (2014).
40. Karlsson, A. K. & Saleh, S. N. Checkpoint inhibitors for malignant melanoma: a systematic review and meta-analysis. *Clin. Cosmet. Investig. Dermatol.* **10**, 325–339 (2017).
41. Atkins, M. B., Clark, J. I. & Quinn, D. I. Immune checkpoint inhibitors in advanced renal cell carcinoma: experience to date and future directions. *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.* **28**, 1484–1494 (2017).
42. Maude, S. L. *et al.* Chimeric antigen receptor T cells for sustained remissions in leukemia. *N. Engl. J. Med.* **371**, 1507–1517 (2014).
43. Trapani, J. A. & Darcy, P. K. Immunotherapy of cancer. *Aust. Fam. Physician* **46**, 194–199 (2017).
44. Blankenstein, T., Coulie, P. G., Gilboa, E. & Jaffee, E. M. The determinants of tumour immunogenicity. *Nat. Rev. Cancer* **12**, 307–313 (2012).
45. Lu, Y.-C. & Robbins, P. F. Targeting neoantigens for cancer immunotherapy. *Int. Immunol.* **28**, 365–370 (2016).
46. Werner, H. M. J., Mills, G. B. & Ram, P. T. Cancer Systems Biology: a peek into the future of patient care? *Nat. Rev. Clin. Oncol.* **11**, 167–176 (2014).
47. von Bueren, A. O. *et al.* Treatment of Children and Adolescents With Metastatic Medulloblastoma and Prognostic Relevance of Clinical and Biologic Parameters. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **34**, 4151–4160 (2016).

48. Ramaswamy, V. *et al.* Risk stratification of childhood medulloblastoma in the molecular era: the current consensus. *Acta Neuropathol. (Berl.)* **131**, 821–831 (2016).
49. Relling, M. V. & Evans, W. E. Pharmacogenomics in the clinic. *Nature* **526**, 343–350 (2015).
50. Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 560–564 (1977).
51. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. 1977. *Biotechnol. Read. Mass* **24**, 104–108 (1992).
52. Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* **94**, 441–448 (1975).
53. Smith, L. M., Fung, S., Hunkapiller, M. W., Hunkapiller, T. J. & Hood, L. E. The synthesis of oligonucleotides containing an aliphatic amino group at the 5' terminus: synthesis of fluorescent DNA primers for use in DNA sequence analysis. *Nucleic Acids Res.* **13**, 2399–2412 (1985).
54. Swerdlow, H. & Gesteland, R. Capillary gel electrophoresis for rapid, high resolution DNA sequencing. *Nucleic Acids Res.* **18**, 1415–1419 (1990).
55. Smith, L. M. *et al.* Fluorescence detection in automated DNA sequence analysis. *Nature* **321**, 674–679 (1986).
56. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **26**, 1135–1145 (2008).
57. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
58. Metzker, M. L. Sequencing technologies - the next generation. *Nat. Rev. Genet.* **11**, 31–46 (2010).
59. Morey, M. *et al.* A glimpse into past, present, and future DNA sequencing. *Mol. Genet. Metab.* **110**, 3–24 (2013).
60. Mardis, E. R. Next-generation sequencing platforms. *Annu. Rev. Anal. Chem. Palo Alto Calif* **6**, 287–303 (2013).
61. Reuter, J. A., Spacek, D. V. & Snyder, M. P. High-throughput sequencing technologies. *Mol. Cell* **58**, 586–597 (2015).
62. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
63. Dohm, J. C., Lottaz, C., Borodina, T. & Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* **36**, e105 (2008).

64. Hillier, L. W. *et al.* Whole-genome sequencing and variant discovery in *C. elegans*. *Nat. Methods* **5**, 183–188 (2008).
65. Quail, M. A. *et al.* A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **13**, 341 (2012).
66. Salipante, S. J. *et al.* Performance comparison of Illumina and ion torrent next-generation sequencing platforms for 16S rRNA-based bacterial community profiling. *Appl. Environ. Microbiol.* **80**, 7583–7591 (2014).
67. Rothberg, J. M. *et al.* An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**, 348–352 (2011).
68. *Ion AmpliSeq™ Library Preparation User Guide.*
69. Ng, S. B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272–276 (2009).
70. Wilkerson, M. D. *et al.* Integrated RNA and DNA sequencing improves mutation detection in low purity tumors. *Nucleic Acids Res.* **42**, e107 (2014).
71. Grossmann, V. *et al.* Whole-exome sequencing identifies somatic mutations of BCOR in acute myeloid leukemia with normal karyotype. *Blood* **118**, 6153–6163 (2011).
72. Yan, X.-J. *et al.* Exome sequencing identifies somatic mutations of DNA methyltransferase gene DNMT3A in acute monocytic leukemia. *Nat. Genet.* **43**, 309–315 (2011).
73. Choi, M. *et al.* Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 19096–19101 (2009).
74. Clark, M. J. *et al.* Performance comparison of exome DNA sequencing technologies. *Nat. Biotechnol.* **29**, 908–914 (2011).
75. Albert, T. J. *et al.* Direct selection of human genomic loci by microarray hybridization. *Nat. Methods* **4**, 903–905 (2007).
76. Teer, J. K. & Mullikin, J. C. Exome sequencing: the sweet spot before whole genomes. *Hum. Mol. Genet.* **19**, R145–151 (2010).
77. Pruitt, K. D. *et al.* The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.* **19**, 1316–1323 (2009).
78. Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.* **12**, 745–755 (2011).
79. Chaitankar, V. *et al.* Next generation sequencing technology and genomewide data

- analysis: Perspectives for retinal research. *Prog. Retin. Eye Res.* **55**, 1–31 (2016).
80. Meyerson, M., Gabriel, S. & Getz, G. Advances in understanding cancer genomes through second-generation sequencing. *Nat. Rev. Genet.* **11**, 685–696 (2010).
 81. Ewing, B., Hillier, L., Wendl, M. C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8**, 175–185 (1998).
 82. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**, 186–194 (1998).
 83. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851–1858 (2008).
 84. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
 85. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
 86. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinforma. Oxf. Engl.* **25**, 1754–1760 (2009).
 87. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.* **25**, 2078–2079 (2009).
 88. Fu, Y., Wu, P.-H., Beane, T., Zamore, P. D. & Weng, Z. Elimination of PCR duplicates in RNA-seq and small RNA-seq using unique molecular identifiers. *BMC Genomics* **19**, 531 (2018).
 89. Clement, K., Farouni, R., Bauer, D. E. & Pinello, L. AmpUMI: design and analysis of unique molecular identifiers for deep amplicon sequencing. *Bioinformatics* **34**, i202–i210 (2018).
 90. Xu, C. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Comput. Struct. Biotechnol. J.* **16**, 15–24 (2018).
 91. Bravo, H. C. & Irizarry, R. A. Model-based quality assessment and base-calling for second-generation sequencing data. *Biometrics* **66**, 665–674 (2010).
 92. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
 93. Raphael, B. J., Dobson, J. R., Oesper, L. & Vandin, F. Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. *Genome Med.* **6**, 5 (2014).
 94. Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23–28 (1976).

95. Ding, L., Raphael, B. J., Chen, F. & Wendl, M. C. Advances for studying clonal evolution in cancer. *Cancer Lett.* **340**, 212–219 (2013).
96. Do, H. & Dobrovic, A. Sequence artifacts in DNA from formalin-fixed tissues: causes and strategies for minimization. *Clin. Chem.* **61**, 64–71 (2015).
97. Pfeiffer, F. *et al.* Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Sci. Rep.* **8**, 10950 (2018).
98. Yu, B., O’Toole, S. A. & Trent, R. J. Somatic DNA mutation analysis in targeted therapy of solid tumours. *Transl. Pediatr.* **4**, 125–138 (2015).
99. Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
100. Lai, Z. *et al.* VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res.* **44**, e108 (2016).
101. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinforma. Oxf. Engl.* **27**, 2987–2993 (2011).
102. Larson, D. E. *et al.* SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinforma. Oxf. Engl.* **28**, 311–317 (2012).
103. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
104. Auwera, G. A. V. der. Abstract 3590: Somatic variation discovery with GATK4. *Cancer Res.* **77**, 3590–3590 (2017).
105. Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* **15**, 591–594 (2018).
106. Xu, C., Nezami Ranjbar, M. R., Wu, Z., DiCarlo, J. & Wang, Y. Detecting very low allele fraction variants using targeted DNA sequencing and a novel molecular barcode-aware variant caller. *BMC Genomics* **18**, 5 (2017).
107. Spencer, D. H. *et al.* Performance of common analysis methods for detecting low-frequency single nucleotide variants in targeted next-generation sequence data. *J. Mol. Diagn. JMD* **16**, 75–88 (2014).
108. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2018).
109. Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
110. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).

111. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin)* **6**, 80–92 (2012).
112. Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M. & Ng, P. C. SIFT missense predictions for genomes. *Nat. Protoc.* **11**, 1–9 (2016).
113. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. *Curr. Protoc. Hum. Genet. Editor. Board Jonathan Haines Al* **0 7**, Unit7.20 (2013).
114. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* **39**, e118–e118 (2011).
115. Gilissen, C., Hoischen, A., Brunner, H. G. & Veltman, J. A. Disease gene identification strategies for exome sequencing. *Eur. J. Hum. Genet. EJHG* **20**, 490–497 (2012).
116. Flanagan, S. E., Patch, A.-M. & Ellard, S. Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations. *Genet. Test. Mol. Biomark.* **14**, 533–537 (2010).
117. Ubersax, J. A. & Ferrell, J. E. Mechanisms of specificity in protein phosphorylation. *Nat. Rev. Mol. Cell Biol.* **8**, 530–541 (2007).
118. Bader, A. G., Kang, S., Zhao, L. & Vogt, P. K. Oncogenic PI3K deregulates transcription and translation. *Nat. Rev. Cancer* **5**, 921–929 (2005).
119. Dong, C. *et al.* Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum. Mol. Genet.* **24**, 2125–2137 (2015).
120. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
121. Consortium, T. 1000 G. P. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
122. Karczewski, K. J. *et al.* The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Res.* **45**, D840–D845 (2017).
123. Hintzsche, J. D., Robinson, W. A. & Tan, A. C. A Survey of Computational Tools to Analyze and Interpret Whole Exome Sequencing Data. *Int. J. Genomics* **2016**, (2016).
124. Cooper, D. N., Krawczak, M., Polychronakos, C., Tyler-Smith, C. & Kehrer-Sawatzki, H. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Hum. Genet.* **132**, 1077–1130 (2013).
125. Lizarin, G. A. *et al.* An empirical estimate of carrier frequencies for 400+ causal Mendelian variants: results from an ethnically diverse clinical sample of 23,453 individuals. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **15**, 178–186 (2013).

126. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).
127. Forbes, S. A. *et al.* COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45**, D777–D783 (2017).
128. Davies, H. *et al.* Mutations of the BRAF gene in human cancer. *Nature* **417**, 949–954 (2002).
129. Yoo, J.-H. *et al.* JAK2 V617F/C618R mutation in a patient with polycythemia vera: a case study and review of the literature. *Cancer Genet. Cytogenet.* **189**, 43–47 (2009).
130. Jauhri, M. *et al.* Prevalence and coexistence of KRAS, BRAF, PIK3CA, NRAS, TP53, and APC mutations in Indian colorectal cancer patients: Next-generation sequencing-based cohort study. *Tumour Biol. J. Int. Soc. Oncodevelopmental Biol. Med.* **39**, 1010428317692265 (2017).
131. Gilbert, M. T. P. *et al.* The isolation of nucleic acids from fixed, paraffin-embedded tissues-which methods are useful when? *PloS One* **2**, e537 (2007).
132. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
133. Dees, N. D. *et al.* MuSiC: Identifying mutational significance in cancer genomes. *Genome Res.* **22**, 1589–1598 (2012).
134. Garraway, L. A. & Lander, E. S. Lessons from the cancer genome. *Cell* **153**, 17–37 (2013).
135. Creixell, P. *et al.* Pathway and network analysis of cancer genomes. *Nat. Methods* **12**, 615–621 (2015).
136. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361 (2017).
137. Fabregat, A. *et al.* The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* **46**, D649–D655 (2018).
138. The Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.* **45**, D331–D338 (2017).
139. Wendl, M. C. *et al.* PathScan: a tool for discerning mutational significance in groups of putative cancer genes. *Bioinforma. Oxf. Engl.* **27**, 1595–1602 (2011).
140. McCormick, F. Signalling networks that cause cancer. *Trends Cell Biol.* **9**, M53-56 (1999).
141. Jensen, L. J. *et al.* STRING 8--a global view on proteins and their functional

- interactions in 630 organisms. *Nucleic Acids Res.* **37**, D412-416 (2009).
142. Keshava Prasad, T. S. *et al.* Human Protein Reference Database--2009 update. *Nucleic Acids Res.* **37**, D767-772 (2009).
143. Koschützki, D. & Schreiber, F. Centrality Analysis Methods for Biological Networks and Their Application to Gene Regulatory Networks. *Gene Regul. Syst. Biol.* **2**, 193–201 (2008).
144. Yu, H., Kim, P. M., Sprecher, E., Trifonov, V. & Gerstein, M. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput. Biol.* **3**, e59 (2007).
145. Leiserson, M. D. M. *et al.* Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* **47**, 106–114 (2015).
146. Deane, C. M., Salwiński, Ł., Xenarios, I. & Eisenberg, D. Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol. Cell. Proteomics MCP* **1**, 349–356 (2002).
147. Liu, X., Wu, C., Li, C. & Boerwinkle, E. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Non-synonymous and Splice Site SNVs. *Hum. Mutat.* **37**, 235–241 (2016).
148. Finn, R. D. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **36**, D281–D288 (2008).
149. Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197 (1981).
150. Yoon, B.-J. Hidden Markov Models and their Applications in Biological Sequence Analysis. *Curr. Genomics* **10**, 402–415 (2009).
151. Saunders, C. T. *et al.* Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinforma. Oxf. Engl.* **28**, 1811–1817 (2012).
152. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* **43**, 11.10.1-33 (2013).
153. Ding, J. *et al.* Feature-based classifiers for somatic mutation detection in tumour-normal paired sequencing data. *Bioinforma. Oxf. Engl.* **28**, 167–175 (2012).
154. Cingolani, P. *et al.* Using *Drosophila melanogaster* as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. *Front. Genet.* **3**, 35 (2012).
155. Sales, G., Calura, E., Cavalieri, D. & Romualdi, C. graphite - a Bioconductor package to convert pathway topology to gene network. *BMC Bioinformatics* **13**, 20 (2012).

156. Sales, G., Calura, E., Martini, P. & Romualdi, C. Graphite Web: web tool for gene set analysis exploiting pathway topology. *Nucleic Acids Res.* **41**, W89–W97 (2013).
157. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
158. van Dongen, J. J. M. *et al.* Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 Concerted Action BMH4-CT98-3936. *Leukemia* **17**, 2257–2317 (2003).
159. Lasorsa, V. A. *et al.* Exome and deep sequencing of clinically aggressive neuroblastoma reveal somatic mutations that affect key pathways involved in cancer progression. *Oncotarget* **7**, 21840–21852 (2016).
160. Shimada, H. *et al.* The International Neuroblastoma Pathology Classification (the Shimada system). *Cancer* **86**, 364–372 (1999).
161. Gao, J. *et al.* Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. *Sci. Signal.* **6**, p11 (2013).
162. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
163. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinforma. Oxf. Engl.* **26**, 589–595 (2010).
164. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).
165. Pugh, T. J. *et al.* The genetic landscape of high-risk neuroblastoma. *Nat. Genet.* **45**, 279–284 (2013).
166. Guo, Y., Ding, X., Shen, Y., Lyon, G. J. & Wang, K. SeqMule: automated pipeline for analysis of human exome/genome sequencing data. *Sci. Rep.* **5**, 14283 (2015).
167. Gao, X., Xu, J. & Starmer, J. Fastq2vcf: a concise and transparent pipeline for whole-exome sequencing data analyses. *BMC Res. Notes* **8**, 72 (2015).
168. Hintzsche, J. *et al.* IMPACT: a whole-exome sequencing analysis pipeline for integrating molecular profiles with actionable therapeutics in clinical samples. *J. Am. Med. Assoc. JAMIA* **23**, 721–730 (2016).
169. Afgan, E. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* **46**, W537–W544 (2018).
170. Koskela, H. L. M. *et al.* Somatic STAT3 Mutations in Large Granular Lymphocytic

- Leukemia. *N. Engl. J. Med.* **366**, 1905–1913 (2012).
171. Fasan, A. *et al.* STAT3 mutations are highly specific for large granular lymphocytic leukemia. *Leukemia* **27**, 1598–1600 (2013).
172. Jerez, A. *et al.* STAT3 mutations unify the pathogenesis of chronic lymphoproliferative disorders of NK cells and T-cell large granular lymphocyte leukemia. *Blood* **120**, 3048–3057 (2012).
173. Andersson, E. *et al.* Activating somatic mutations outside the SH2-domain of STAT3 in LGL leukemia. *Leukemia* **30**, 1204–1208 (2016).
174. Rajala, H. L. M. *et al.* Discovery of somatic STAT5b mutations in large granular lymphocytic leukemia. *Blood* **121**, 4541–4550 (2013).
175. Rajala, H. L. M., Porkka, K., Maciejewski, J. P., Jr, T. P. L. & Mustjoki, S. Uncovering the pathogenesis of large granular lymphocytic leukemia—novel STAT3 and STAT5b mutations. *Ann. Med.* **46**, 114–122 (2014).
176. Teramo, A. *et al.* Intrinsic and extrinsic mechanisms contribute to maintain the JAK/STAT pathway aberrantly activated in T-type large granular lymphocyte leukemia. *Blood* **121**, 3843–3854 (2013).
177. Leblanc, F., Zhang, D., Liu, X. & Loughran, T. P. Large granular lymphocyte leukemia: from dysregulated pathways to therapeutic targets. *Future Oncol. Lond. Engl.* **8**, 787–801 (2012).
178. Bailey, N. G. & Elenitoba-Johnson, K. S. J. Mature T-cell leukemias: Molecular and Clinical Aspects. *Curr. Hematol. Malig. Rep.* **10**, 421–428 (2015).
179. Kiel, M. J. *et al.* Integrated genomic sequencing reveals mutational landscape of T-cell prolymphocytic leukemia. *Blood* **124**, 1460–1472 (2014).
180. Rodríguez-Caballero, A. *et al.* Expanded cells in monoclonal TCR-alpha-beta+/CD4+/NKa+/CD8-/dim T-LGL lymphocytosis recognize hCMV antigens. *Blood* **112**, 4609–4616 (2008).
181. Andersson, E. I. *et al.* High incidence of activating STAT5B mutations in CD4-positive T-cell large granular lymphocyte leukemia. *Blood* **128**, 2465–2468 (2016).
182. Mason, C. C. *et al.* Age-related mutations and chronic myelomonocytic leukemia. *Leukemia* **30**, 906–913 (2016).
183. Garg, M. *et al.* Profiling of somatic mutations in acute myeloid leukemia with FLT3-ITD at diagnosis and relapse. *Blood* **126**, 2491–2501 (2015).
184. Tenedini, E. *et al.* Targeted cancer exome sequencing reveals recurrent mutations in myeloproliferative neoplasms. *Leukemia* **28**, 1052–1059 (2014).

185. Dobashi, A. *et al.* Frequent BCOR aberrations in extranodal NK/T-Cell lymphoma, nasal type. *Genes. Chromosomes Cancer* **55**, 460–471 (2016).
186. Penzo-Méndez, A. I. & Stanger, B. Z. Cell competition in vertebrate organ size regulation. *Wiley Interdiscip. Rev. Dev. Biol.* **3**, 419–427 (2014).
187. Harvey, K. F., Zhang, X. & Thomas, D. M. The Hippo pathway and human cancer. *Nat. Rev. Cancer* **13**, 246–257 (2013).
188. Guo, C. *et al.* KMT2D maintains neoplastic cell proliferation and global histone H3 lysine 4 monomethylation. *Oncotarget* **4**, 2144–2153 (2013).
189. Zhang, J. *et al.* Disruption of KMT2D perturbs germinal center B cell development and promotes lymphomagenesis. *Nat. Med.* **21**, 1190–1198 (2015).
190. Wang, L., Wang, Y.-Y., Cao, Q., Chen, Z. & Chen, S.-J. Hornerin gene was involved in a case of acute myeloid leukemia transformed from myelodysplastic syndrome with t(1;2)(q21;q37). *Leukemia* **20**, 2184–2187 (2006).
191. Kawaguchi, M. *et al.* Serum levels of angiopoietin-2, but not angiopoietin-1, are elevated in patients with erythrodermic cutaneous T-cell lymphoma. *Acta Derm. Venereol.* **94**, 9–13 (2014).
192. Kelly, S. *et al.* Angiogenic gene expression and vascular density are reflected in ultrasonographic features of synovitis in early Rheumatoid Arthritis: an observational study. *Arthritis Res. Ther.* **17**, 58 (2015).
193. Xia, Y., Lu, R.-N. & Li, J. Angiogenic factors in chronic lymphocytic leukemia. *Leuk. Res.* **36**, 1211–1217 (2012).
194. Kopparapu, P. K. *et al.* MCPH1 maintains long-term epigenetic silencing of ANGPT2 in chronic lymphocytic leukemia. *FEBS J.* **282**, 1939–1952 (2015).
195. Johansson, P. *et al.* Recurrent alterations of TNFAIP3 (A20) in T-cell large granular lymphocytic leukemia. *Int. J. Cancer* **138**, 121–124 (2016).
196. Swerdlow, S. H. *et al.* The 2016 revision of the World Health Organization classification of lymphoid neoplasms. *Blood* **127**, 2375–2390 (2016).
197. Sandlund, J. T. & Perkins, S. L. Uncommon non-Hodgkin lymphomas of childhood: pathological diagnosis, clinical features and treatment approaches. *Br. J. Haematol.* **169**, 631–646 (2015).
198. Campo, E. *et al.* The 2008 WHO classification of lymphoid neoplasms and beyond: evolving concepts and practical applications. *Blood* **117**, 5019–5032 (2011).
199. Morin, R. D. *et al.* Somatic mutations altering EZH2 (Tyr641) in follicular and diffuse large B-cell lymphomas of germinal-center origin. *Nat. Genet.* **42**, 181–185 (2010).

200. Okosun, J. *et al.* Integrated genomic analysis identifies recurrent mutations and evolution patterns driving the initiation and progression of follicular lymphoma. *Nat. Genet.* **46**, 176–181 (2014).
201. Louissaint, A. *et al.* Pediatric-type nodal follicular lymphoma: a biologically distinct lymphoma with frequent MAPK pathway mutations. *Blood* **128**, 1093–1100 (2016).
202. Schmidt, J. *et al.* Genome-wide analysis of pediatric-type follicular lymphoma reveals low genetic complexity and recurrent alterations of TNFRSF14 gene. *Blood* **128**, 1101–1111 (2016).
203. Ozawa, M. G. *et al.* A study of the mutational landscape of pediatric-type follicular lymphoma and pediatric nodal marginal zone lymphoma. *Mod. Pathol. Off. J. U. S. Can. Acad. Pathol. Inc* **29**, 1212–1220 (2016).
204. Schmidt, J. *et al.* Mutations of MAP2K1 are frequent in pediatric-type follicular lymphoma and result in ERK pathway activation. *Blood* **130**, 323–327 (2017).
205. Kotsiou, E. *et al.* TNFRSF14 aberrations in follicular lymphoma increase clinically significant allogeneic T-cell responses. *Blood* **128**, 72–81 (2016).
206. Jiang, L. *et al.* Exome sequencing identifies somatic mutations of DDX3X in natural killer/T-cell lymphoma. *Nat. Genet.* **47**, 1061–1066 (2015).
207. Bromberg-White, J. L., Andersen, N. J. & Duesbery, N. S. MEK genomics in development and disease. *Brief. Funct. Genomics* **11**, 300–310 (2012).
208. Baujat, G. *et al.* Paradoxical NSD1 mutations in Beckwith-Wiedemann syndrome and 11p15 anomalies in Sotos syndrome. *Am. J. Hum. Genet.* **74**, 715–720 (2004).
209. Kozlov, G., Ménade, M., Rosenauer, A., Nguyen, L. & Gehring, K. Molecular determinants of PAM2 recognition by the MLLE domain of poly(A)-binding protein. *J. Mol. Biol.* **397**, 397–407 (2010).
210. Singh, G., Rebbapragada, I. & Lykke-Andersen, J. A competition between stimulators and antagonists of Upf complex recruitment governs human nonsense-mediated mRNA decay. *PLoS Biol.* **6**, e111 (2008).
211. Zhu, J., Ding, H., Wang, X. & Lu, Q. PABPC1 exerts carcinogenesis in gastric carcinoma by targeting miR-34c. *Int. J. Clin. Exp. Pathol.* **8**, 3794–3802 (2015).
212. Leeksa, O. C., de Miranda, N. F. & Veelken, H. Germline mutations predisposing to diffuse large B-cell lymphoma. *Blood Cancer J.* **7**, e532 (2017).
213. Danussi, C. *et al.* RHPN2 drives mesenchymal transformation in malignant glioma by triggering RhoA activation. *Cancer Res.* **73**, 5140–5150 (2013).
214. Luk, P. P. *et al.* BRAF mutations in non-small cell lung cancer. *Transl. Lung Cancer*

Res. **4**, 142–148 (2015).

215. Muppidi, J. R. *et al.* Loss of signalling via $G\alpha 13$ in germinal centre B-cell-derived lymphoma. *Nature* **516**, 254–258 (2014).
216. Shah, S. *et al.* A recurrent germline PAX5 mutation confers susceptibility to pre-B cell acute lymphoblastic leukemia. *Nat. Genet.* **45**, 1226–1231 (2013).
217. Stigliani, S. *et al.* High genomic instability predicts survival in metastatic high-risk neuroblastoma. *Neoplasia N. Y. N* **14**, 823–832 (2012).
218. Molenaar, J. J. *et al.* Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes. *Nature* **483**, 589–593 (2012).
219. Luksch, R. *et al.* Neuroblastoma (Peripheral neuroblastic tumours). *Crit. Rev. Oncol. Hematol.* **107**, 163–181 (2016).
220. Sausen, M. *et al.* Integrated genomic analyses identify ARID1A and ARID1B alterations in the childhood cancer neuroblastoma. *Nat. Genet.* **45**, 12–17 (2013).
221. Kuboki, Y. *et al.* Comprehensive analyses using next-generation sequencing and immunohistochemistry enable precise treatment in advanced gastric cancer. *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.* **27**, 127–133 (2016).
222. Owen, J. D., Ruest, P. J., Fry, D. W. & Hanks, S. K. Induced focal adhesion kinase (FAK) expression in FAK-null cells enhances cell spreading and migration requiring both auto- and activation loop phosphorylation sites and inhibits adhesion-dependent tyrosine phosphorylation of Pyk2. *Mol. Cell. Biol.* **19**, 4806–4818 (1999).
223. Lee, S. *et al.* FAK is a critical regulator of neuroblastoma liver metastasis. *Oncotarget* **3**, 1576–1587 (2012).
224. Carvill, G. L. *et al.* Targeted resequencing in epileptic encephalopathies identifies de novo mutations in CHD2 and SYNGAP1. *Nat. Genet.* **45**, 825–830 (2013).
225. Rodríguez, D. *et al.* Mutations in CHD2 cause defective association with active chromatin in chronic lymphocytic leukemia. *Blood* **126**, 195–202 (2015).
226. Reisman, D., Glaros, S. & Thompson, E. A. The SWI/SNF complex and cancer. *Oncogene* **28**, 1653–1668 (2009).
227. Decristofaro, M. F. *et al.* Characterization of SWI/SNF protein expression in human breast cancer cell lines and other malignancies. *J. Cell. Physiol.* **186**, 136–145 (2001).
228. Schneppenheim, R. *et al.* Germline nonsense mutation and somatic inactivation of SMARCA4/BRG1 in a family with rhabdoid tumor predisposition syndrome. *Am. J. Hum. Genet.* **86**, 279–284 (2010).
229. Wilson, B. G. & Roberts, C. W. M. SWI/SNF nucleosome remodellers and cancer. *Nat.*

Rev. Cancer **11**, 481–492 (2011).

230. Sentani, K. *et al.* Increased expression but not genetic alteration of BRG1, a component of the SWI/SNF complex, is associated with the advanced stage of human gastric carcinomas. *Pathobiol. J. Immunopathol. Mol. Cell. Biol.* **69**, 315–320 (2001).

231. Bai, J. *et al.* BRG1 expression is increased in human glioma and controls glioma cell proliferation, migration and invasion in vitro. *J. Cancer Res. Clin. Oncol.* **138**, 991–998 (2012).

232. Jubierre, L. *et al.* BRG1/SMARCA4 is essential for neuroblastoma cell viability through modulation of cell death and survival pathways. *Oncogene* **35**, 5179–5190 (2016).

233. Matsubara, D. *et al.* Lung cancer with loss of BRG1/BRM, shows epithelial mesenchymal transition phenotype and distinct histologic and genetic features. *Cancer Sci.* **104**, 266–273 (2013).

234. Wu, S. *et al.* BRG1, the ATPase subunit of SWI/SNF chromatin remodeling complex, interacts with HDAC2 to modulate telomerase expression in human cancer cells. *Cell Cycle Georget. Tex* **13**, 2869–2878 (2014).

235. Maida, Y. *et al.* Involvement of telomerase reverse transcriptase in heterochromatin maintenance. *Mol. Cell. Biol.* **34**, 1576–1593 (2014).

236. Peifer, M. *et al.* Telomerase activation by genomic rearrangements in high-risk neuroblastoma. *Nature* **526**, 700–704 (2015).

237. Tonini, G. P. Growth, progression and chromosome instability of Neuroblastoma: a new scenario of tumorigenesis? *BMC Cancer* **17**, 20 (2017).

238. Defferrari, R. *et al.* Influence of segmental chromosome abnormalities on survival in children over the age of 12 months with unresectable localised peripheral neuroblastic tumours without MYCN amplification. *Br. J. Cancer* **112**, 290–295 (2015).

239. Eleveld, T. F. *et al.* Relapsed neuroblastomas show frequent RAS-MAPK pathway mutations. *Nat. Genet.* **47**, 864–871 (2015).

240. Schramm, A. *et al.* Mutational dynamics between primary and relapse neuroblastomas. *Nat. Genet.* **47**, 872–877 (2015).

241. Padovan-Merhar, O. M. *et al.* Enrichment of Targetable Mutations in the Relapsed Neuroblastoma Genome. *PLoS Genet.* **12**, e1006501 (2016).

Somatic mutations in specific and connected subpathways are associated with short neuroblastoma patients' survival and indicate proteins targetable at onset of disease

Maria Rosaria Esposito¹, Andrea Binatti², Marcella Pantile¹, Alessandro Coppe³, Katia Mazzocco⁴, Luca Longo⁵, Mario Capasso^{6,7,8}, Vito Alessandro Lasorsa⁷, Roberto Luksch⁹, Stefania Bortoluzzi² and Gian Paolo Tonini¹

¹Neuroblastoma Laboratory, Fondazione Istituto di Ricerca Pediatrica Città della Speranza, Padua, Italy

²Department of Molecular Medicine, University of Padua, Padua, Italy

³Department of Women's and Children's Health, University of Padova, Padua, Italy

⁴Translational Research Department, Laboratory Medicine, Diagnostics and Services U.O.C. Pathological Anatomy, IRCCS Giannina Gaslini Institute, Genoa, Italy

⁵U.O.C. Bioterapie, Ospedale Policlinico San Martino, Genoa, Italy

⁶Department of Molecular Medicine and Medical Biotechnology, University of Naples Federico II, Naples, Italy

⁷CEINGE Biotechnologie Avanzate, Naples, Italy

⁸IRCCS SDN, Istituto di Ricerca Diagnostica e Nucleare, Naples, Italy

⁹Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy

Neuroblastoma (NB) is an embryonic malignancy of the sympathetic nervous system with heterogeneous biological, morphological, genetic and clinical characteristics. Although genomic studies revealed the specific biological features of NB pathogenesis useful for new therapeutic approaches, the improvement of high-risk (HR)-NB patients overall survival remains unsatisfactory. To further clarify the biological basis of disease aggressiveness, we used whole-exome sequencing to examine the genomic landscape of HR-NB patients at stage M with short survival (SS) and long survival (LS). Only a few genes, including *SMARCA4*, *SMO*, *ZNF44* and *CHD2*, were recurrently and specifically mutated in the SS group, confirming the low recurrence of common mutations in this tumor. A systems biology approach revealed that in the two patient groups, mutations occurred in different pathways. Mutated genes (*ARHGEF11*, *CACNA1G*, *FGF4*, *PTPRA*, *PTK2*, *ANK3*, *SMO*, *NTNG2*, *VCL* and *NID2*) regulate the MAPK pathway associated with the organization of the extracellular matrix, cell motility through PTK2 signaling and matrix metalloproteinase activity. Moreover, we detected mutations in *LAMA2*, *PTK2*, *LAMA4*, and *MMP14* genes, impairing MET signaling, in *SFI1* and *CHD2* involved in centrosome maturation and chromosome remodeling, in *AK7* and *SPTLC2*, which regulate the metabolism of nucleotides and lipoproteins, and in *NALCN*, *SLC12A1*, *SLC9A9*, which are involved in the transport of small molecules. Notably, connected networks of somatically mutated genes specific for SS patients were identified. The detection of mutated genes present at the onset of disease may help to address an early treatment of HR-NB patients using FDA-approved compounds targeting the deregulated pathways.

Key words: neuroblastoma, whole-exome sequencing, somatic mutations, pathways, gene networks, target protein

Additional Supporting Information may be found in the online version of this article.

†M.R.E. and A.B. contributed equally to this work

Maria Rosaria Esposito's current address is: Fondazione Istituto di Ricerca Pediatrica Città della Speranza; Department of Industrial Engineering (DII), University of Padua Corso Stati Uniti, 4 35127, Padua, Italy

Conflict of interest: The authors declare that they have no conflict of interest.

Grant sponsor: Fondazione Italiana per la Lotta al Neuroblastoma; **Grant sponsor:** Fondazione Cassa di Risparmio di Padova e Rovigo;

Grant sponsor: Ministero dell'Istruzione, dell'Università e della Ricerca; **Grant numbers:** 2010NYKNS7_002; **Grant sponsor:** University of Padua

DOI: 10.1002/ijc.31748

History: Received 27 Oct 2017; Accepted 21 Jun 2018; Online 11 Jul 2018

Correspondence to: Maria Rosaria Esposito, Fondazione Istituto di Ricerca Pediatrica Città della Speranza, Neuroblastoma Laboratory, Corso Stati Uniti, 4 35127, Padua, Italy, E-mail: mr.esposito@irpcds.org; or Stefania Bortoluzzi, Department of Molecular Medicine, University of Padua, Padua, Italy, E-mail: stefania.bortoluzzi@unipd.it

What's new?

Most patients with metastatic neuroblastoma don't survive 5 years from diagnosis, despite responding well to first-line treatments. Previous work comparing short-survival and long-survival patients identified some key chromosomal differences. These authors take the search deeper, conducting whole-exome sequencing to compare somatic mutations between patients who survived at least 5 years and those who did not. They determined that mutations among the short-survival group affected different pathways than those afflicting the long-survival patients. In some cases, drugs already exist that target these proteins, suggesting that testing for these mutations at the time of diagnosis could indicate specific treatments.

Introduction

Neuroblastoma (NB) is a pediatric cancer of the sympathetic nervous system. Metastatic disease (stage M) usually involves liver, skin, the bone marrow and/or skeleton.¹ Stage M patients are classified as high-risk NB (HR-NB) and show an overall survival lower than 40% at 5 years from diagnosis.² The majority of HR-NB stage M patients responds well to the first-line therapy, but relapse occurs in the majority of patients. Several studies demonstrated that both copy number aberrations (CNAs)³⁻⁵ and genomic variants⁶⁻¹¹ contribute to the tumor aggressiveness. In 2012, Stigliani et al.⁴ investigated the genomic features of HR-NB, dividing patients into a short-survival (SS; with disease progression and survival at most 5 years from diagnosis) and a long-survival group (LS; responsive to the treatments and a survival over 5 years from diagnosis). The study showed that tumor cells from the SS group are characterized by a high number of structural CNAs and high chromosomal instability. A previous study showed that, beyond CNAs, also mutations hitting specific pathways could be implicated in HR-NB progression.⁹ After this line of evidence, to elucidate the biology underlying differences between the SS and LS groups in term of different outcome, our study characterized by whole-exome sequencing (WES) the genomic landscape of primary tumors, with focus to single nucleotide variants (SNVs) and indels. Aberrant somatic mutations exclusive for the SS or LS patients were found, as well as pathways and subpathways that are specifically targeted in SS tumors, which were confirmed by the analysis of a large independent cohort.⁸

Materials and Methods**Patients and tumor samples**

A cohort of stage M NB patients from the Italian Neuroblastoma Registry with complete clinical data and follow-up over 10 years was considered. Frozen tissue from the primary tumor at onset was available for each patient. Patients were stratified into two groups according to their overall survival: the SS group ($n = 14$), including patients with rapid disease progression and rapid fatal outcome, all with a survival time < 60 months, and the LS group ($n = 15$), including patients who are responsive to therapy and survived at least 60 months from diagnosis. Five SS patients (ID2475; ID2368; ID2181; ID1995; ID2100) were also

included in the previous NB report.⁹ Informed consent was received for the use of biological material from legal tutors, and the study was approved by the Institutional Board of the participating Institutions. Total genomic DNAs (gDNA) from 29 tumors and matched constitutional DNA of patients was purified according to the standard protocol with Invisorb[®] Spin Tissue Mini Kit (SPA-Stratec molecular). The amount and quality of gDNA were assessed by Nanodrop and Qubit Instruments (Invitrogen), respectively, and only high-quality samples (DNA/protein ratio, A260/A280: 1.8–2.0) were processed. All tumor samples were classified as NB Schwannian stroma-poor according to criteria established by the International Neuroblastoma Pathology Committee.¹ The presence of at least 60% of neuroblasts in tumor samples was verified.

Exome library preparation and WES

For each sample, 100 ng of DNA (determined by Qubit[®] 2.0 Fluorometer) was used for exome library preparation by the AmpliSeq[™] exome kit (Life Technologies) targeting approximately 35 Mb of human exons. Briefly, the gDNA was amplified by oligo pools/primers to perform ultra-high multiplex PCR enrichment of the exonic regions of the genome. Next, the amplicons were ligated to adaptors with Ion Xpress Barcode Adapters Kit and purified with Agencourt AMPure XP kit (Beckman Coulter Genomics). The library was quantified with Quantitation RT-PCR with the Ion Library Quantitation Kit (Life Technologies), diluted to 100 pM and loaded on a P1 chip for Ion Proton Sequencing according to the manufacturer's protocol.

WES variant calling

Read mapping and variant calling were performed with Torrent Suite and Ion Reporter[™] software, provided by the Ion Proton[™] System. The Proton Run Browser was used for quality control metrics (percent bead loading, usable sequences, read length, alignment metrics to hg19 reference genome and mean raw accuracy). The samples were processed using the workflow: "Somatic – Proton – High Stringency Configuration". Bam files of the tumor and blood samples of each patient were uploaded to Ion Reporter[™] (IR) software using the available plug-in, IonReporterUploader_V1_2. Variant calling was done using Torrent Variant Caller (v. 5.0–9). Next, the files were processed using a workflow AmpliSeq

Exome paired sample (tumor/normal) to subtract variants [SNV, multiple nucleotide variant (MNV), indel and copy number variant (CNV)] discovered in the peripheral blood DNA against the tumor DNA.

Variant annotation and prioritization

The annotation of somatic variants was performed by SnpSift¹² and SnpEff.¹³ SnpSift's *annotate* command provided the association of known variants to dbSNP (v. 147) and COSMIC (v. 77)¹⁴ identifiers, clinical significance from Clinvar (updated on May 02, 2016),¹⁵ as well as functional prediction indicated by MetaSVM and MetaLR¹⁶ from dbNSFP database (v. 2.9.1).¹⁷ The two algorithms predict whether the variant is tolerated or deleterious, considering nine scores present in dbNSFP (SIFT, Polyphen – 2, GERP ++, MutationTaster, MutationAssessor, FATHMM, LRT, SiPhy and PhyloP) and MMAF observed in different populations of 1,000 genomes. SnpEff predicts the functional and putative impact of detected variants. Known variants annotated in Clinvar as benign or likely benign were discarded, and only variants with *HIGH* or *MODERATE* SnpEff predicted impact were further considered. After integration of exome sequencing and Exome Aggregation Consortium (ExAC, Cambridge, MA, <http://exac.broadinstitute.org>)¹⁸ data, variants with a non-Finnish European population allele frequency > 5% were discarded (Supporting Information Fig.S1). The interpretation of variants' impact was also obtained by mapping selected variants to protein sequences and their domain annotation using MutationMapper (http://www.cbioportal.org/mutation_mapper.jsp). Additional structural predictions and analysis on the mutated protein sequences were obtained using Phyre2.¹⁹

Pathway analyses

Genes mutated in SS and in LS were separately mapped to the KEGG²⁰ and Reactome²¹ pathways. The pathways with at least three genes mutated in a group and none in the other group were defined as “group-specific pathways.” Significant pathway enrichment was calculated considering the separately mutated genes in the SS and LS patient groups. Significantly enriched pathways only in one group and with a number of genes mutated in the group by at least 1.5x the number of genes mutated in the other group were considered to be “group-specifically enriched pathways.” The detected pathways were organized after the architecture of the KEGG and Reactome databases to have a less redundant description of altered molecular signaling and biological functions, gathering pathways into functional classes.

Gene network analysis

The R Graphite Bioconductor package (v. 1.20.1)²² was used to convert complex pathway topologies into Reactome pathway-derived gene networks using appropriate biology-

driven rules to transform different types of direct and indirect relations between genes and gene products annotated in pathways (i.e., regulatory relations, participation to molecular complexes and biosynthetic pathways, also with compound intermediates) into pairwise gene connections. Reactome networks were merged into a pathway-derived gene network of 186,808 pairwise interactions between 8,678 genes. We applied the HotNet2 algorithm²³ to Reactome-derived gene networks to statistically identify group-specific subnetworks of mutated protein-coding genes, defining groups of functionally related genes in which mutations significantly converge (Supporting Information Fig. S1). HotNet2 consists of an insulated heat diffusion model to detect significantly mutated gene subnetworks, evaluating both the heat score of nodes and the local network topology. The heat score for each node was calculated from the number of samples carrying prioritized somatic variants in the corresponding gene. Hotnet2 analysis was first conducted using all the somatically mutated genes in the whole cohort of patients, and then the two patient groups were considered separately. Due to the cardinality of our cohort and the considerable dimension of the considered network, multiple testing correction applied in this analysis, considerably increases the p-value. Gene groups emerging from Hotnet2 analysis were also linked to Gene Ontology biological processes (The Gene Ontology Consortium, 2015).

Variants validation by ultra-deep sequencing

Validation of tumor variants was performed by ultra-deep sequencing on Amplicon libraries using the 454 Junior Titanium sequencer (Roche) according to the protocol for Amplicon amplification, Lib-A (Roche). Amplicons were obtained by one-step PCR using the FastStartTM High Fidelity PCR System, dNTPack (Roche) and specific adaptors were ligated for each patient. Amplicon lengths ranged from 200 to 400 bp, including forward and reverse Phusion primers, intermediate patient-specific sequence MID and the target template. The initial PCR was performed with 10 ng of gDNA input according to the manufacturer's protocol, the FastStartTM High Fidelity PCR System, and the dNTPack (Roche). After PCR amplification, the Library with Agencourt AMPure beads (Beckman Coulter) was purified and the libraries were quantified according to the Quant-iT PicoGreen dsDNA Assay Kit (Thermo Fisher Scientific). Finally, the amount of library to be used in the emPCR was determined according to the Method Manual Lib-A (Roche). Target regions of the genome reference sequences corresponding to the amplicons were obtained from the human reference genome (GRCh37/hg19) using the *getfasta* command of Bedtools.²⁴ Reads were mapped to these sequences through the *bwa-sw* command of Burrows-Wheeler Aligner²⁵ and variants were called using GATK – Genome Analysis Tool Kit.²⁶ Variants were also confirmed using IGV – Integrated Genome Viewer.²⁷

Results

Somatic mutations in HR-NB patients

To clarify the genomic features associated with rapid disease progression in NB patients, we analyzed an Italian cohort of 29 tumors of HR-NB stage M patients classified in SS and LS by WES (Table 1). The survival ranged from 6 to 48 months in SS and from 62 to 159 months in LS. Two SS patients out of 14 died as a result of therapy toxicity. Out of the 15 LS patients, 11 were still in complete remission, two were alive with disease and two died of disease at last follow-up. Alignment of 2,189,622,787 reads to the reference exome (37.1 million reads per sample on average) yielded a 97x average coverage and a 76.29% of the target exome with at least 30x coverage, ranging from 53% to 84% in different patients. Sequence coverage in the SS and LS groups, considering both the tumor and peripheral blood cell samples, was homogeneous (Supporting Information Fig. S2). A total of 2,301 and 1,805 high quality and

coverage somatic variants for the SS and LS groups, respectively, were detected after read mapping, variant calling and identification of somatic variants comparing the tumor and control data. In accordance with the mutations types observed before,⁷⁻⁹ somatic variants resulted in enrichment in C > A (LS = 32.3%, SS = 25.2%) transversions at TCT sites and in C > T transitions (LS = 20.3%, SS = 22.6%) at GCG trinucleotide substitution types, normally due to deamination of 5-methylcytosine (Supporting Information Fig. S3). Next, 1,288 high-quality, detrimental and rare somatic variants in 1,043 genes, 580 variants detected in LS patients (Supporting Information Table S1A), and 708 in SS (Supporting Information Table S1B) passed the variant effect- and frequency-based filtering steps (Supporting Information Fig. S1). Variants were later examined considering recurrence, hit gene and pathways, and the possible impacts of mutations on disease progression. Variant effects, group-exclusivity, intra-group recurrence, gene

Table 1. NB patient cohort description.

Patient ID	Sex	Age at diagnosis (months)	MYCN status	DNA index	Survival (months)	Outcome	Group
2,475	M	208	Gain	No data	33	DOD	SS
1,965	M	83	Not amplified	1.51	34	DOD	
1,955	F	77	Not amplified	No data	6	DOT	
2,368	M	75	Not amplified	No data	48	DOD	
3,060	F	118	Unknown	No data	33	DOD	
1,900	M	37	Not amplified	1.14	24	DOD	
2,181	M	47	Amplified	No data	28	DOD	
2,384	M	58	Gain	No data	45	DOD	
1,995	M	22	Amplified	2.37	23	DOD	
1,920	M	14	Not amplified	No data	9	DOT	
2,100	M	27	Not amplified	No data	12	DOD	
2,513	M	52	Not amplified	No data	20	DOD	
2,852	M	50	Gain	No data	43	DOD	
2,578	F	23	Gain	1.07	42	DOD	
1,409	F	34	Not amplified	1.00	159	CR	LS
1,641	M	33	Not amplified	1.96	105	CR	
2,121	M	61	Not amplified	1.88	144	CR	
2,393	F	73	Gain	No data	62	CR	
2,140	M	55	Not amplified	1.00	75	CR	
1,905	M	15	Not amplified	1.96	80	CR	
2,528	M	61	Gain	No data	86	CR	
2,035	M	17	Not amplified	1.52	144	CR	
2,488	M	47	Not amplified	No data	65	AWD	
2,951	M	68	Gain	1.00	64	DOD	
2,251	F	12	Amplified	No data	110	CR	
2,576	F	32	Not amplified	No data	71	AWD	
2,426	F	7	Not amplified	No data	53	CR	
2,613	F	12	Not amplified	No data	39	CR	
2,828	M	8	Not amplified	1.92	71	CR	

Abbreviations: M, male; F, Female; Unknown, Physician did not have data; No data, data was not made available; DOD, Dead of disease; DOT, Dead of toxicity of the treatment; AWD, Alive with disease; CR, Complete remission

products biological function and relationships among the mutated genes were used to prioritize group-specific variants for validation. We confirmed 50 selected variants in 49 genes (Supporting Information Table S2).

Mutation landscapes in SS and LS patients

We compared the number, type and effect of somatic mutations observed in the two patient groups. No significant difference in the numbers of somatic variants per patient in the SS (median 37) and LS (36) groups were observed (Wilcoxon test p value = 1; Fig. 1a). Very close numbers of selected somatic variants per Mb were observed in the two groups (median values of 0.62 and 0.64, respectively). The number of somatic variants per patient in our cohort was higher than previously reported,^{6–9} but a direct comparison between WES studies is hampered by several factors, on top the different sequencing depth or technology and the different analysis methods and settings used.

Moreover, similar patterns in the SS and LS patients were present considering the variant type (Supporting Information Fig. S4A; G test $p = 0.11$) and predicted variant effect on the protein sequences (Supporting Information Fig. S4B, G test $p = 0.06$) without evident of differences in relation to different therapy responses and outcomes.

Somatic variants and mutated genes exclusive of SS or LS

The above-cited 708 and 580 high-confidence damaging, and rare variants detected in SS and LS patients, fell into 583 and 515 individual genes, respectively. Eighteen variants in 18 different genes occurred in both patient groups, resulting in recurrence in NB-HR patients considered as a whole, whereas there were 690 and 562 group-specific variants. Only 102 (9.8%) out of 1,043 genes mutated in the whole NB cohort were recurrently mutated in two or more patients. Fifty-five genes were recurrently mutated in patients of both classes (Fig. 1b), including 17 genes detected in more than two patients, with *DPCR1* (mutated in nine patients), *AHNAK2* (6), and *CBX4* and *ZNF717* (both mutated in four patients) being the most recurrently observed. Notably, 528 genes were specifically mutated in SS and 460 in LS patients (Fig. 1b), including 21 and 17 recurrent and group-specific genes, respectively, and six that carry particularly damaging variants (Fig. 1c). Of these genes, only *KMT2A* (Lysine Methyltransferase 2A; E2926Q in patient ID2426; S3291C in patient ID1905) and *NUPL1* (Nucleoporin 58; N153 fs in patient ID2393 and ID2576) resulted in recurrently mutated and group-specific pathways LS patients.

In SS patients, four genes (*SMO*, *SMARCA4*, *ZNF44* and *CHD2*), all known to be expressed in neural tissues, were recurrently mutated and group-specific and carried particularly

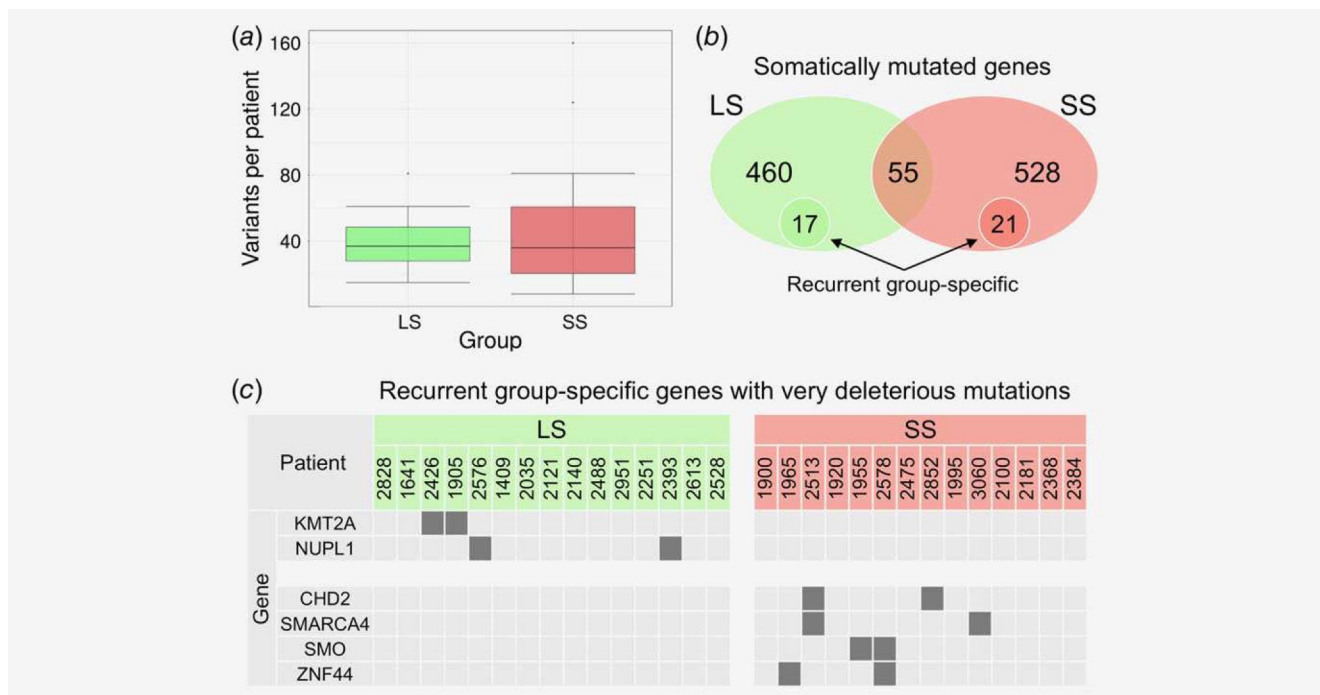


Figure 1. Comparison of mutation landscapes in SS and LS NB patients. (a) The number of variants in LS and SS groups were not significantly different, as shown by the boxplot of the distribution of selected somatic variants per patient ($p = 0.98$ of Wilcoxon test of median equality, conducted after Shapiro–Wilk test of normal distribution p value = 3.79×10^{-5}). (b) Venn chart of number of somatically mutated genes in LS and SS groups, showing class-specifically mutated genes and their subset of genes being both class-specific and recurrent intra-class. (c) Mutation matrix indicating in which class and patients are mutated class-specific and recurrent genes, hit by particularly deleterious mutations.

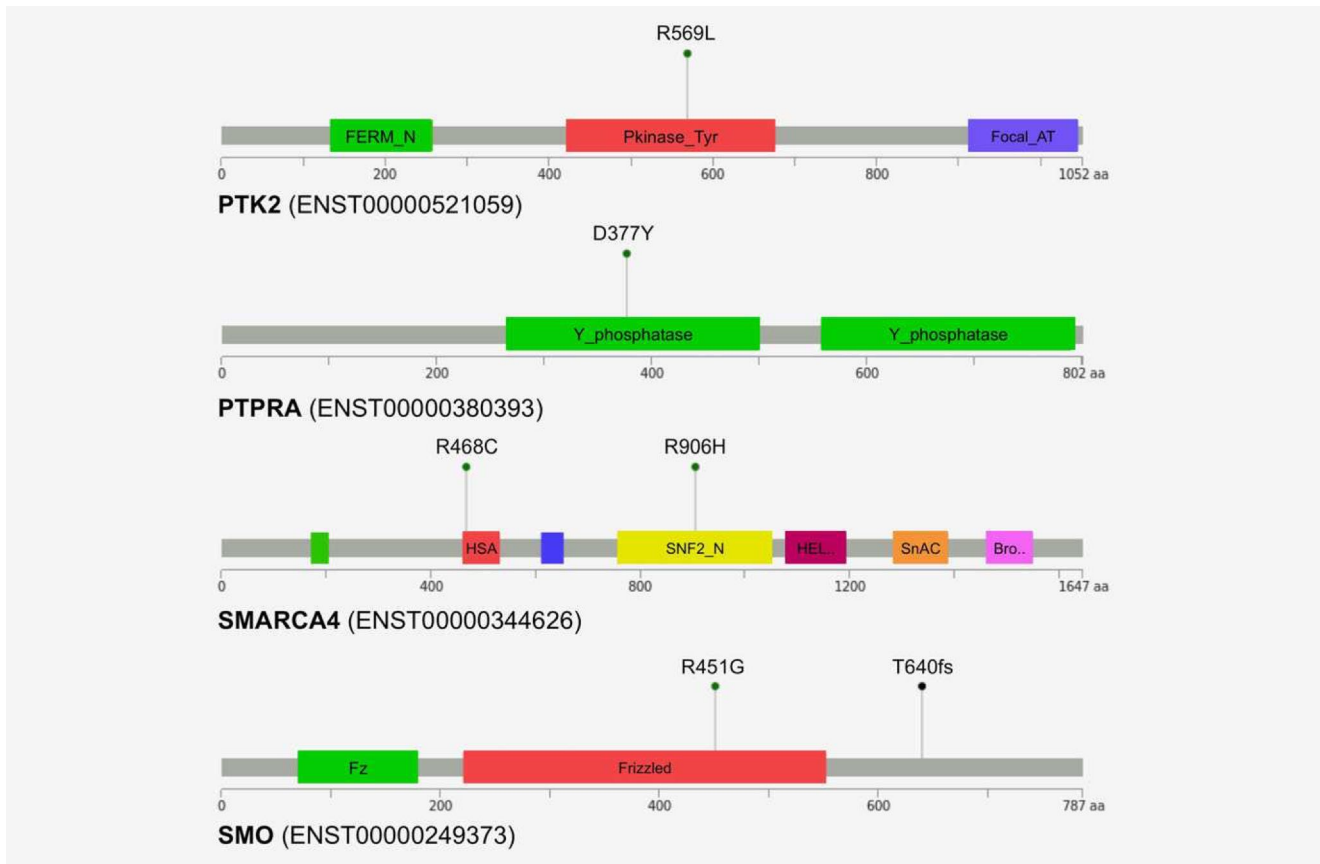


Figure 2. Impact on proteins of somatic variants detected in NB patients in *PTK2*, *PTPRA*, *SMARCA4* and *SMO* genes. For each gene, considering the protein encoded by the reference transcript, lollipop plots show the type and the position of somatic variants in relation to the protein sequence and domains (colored portions) according to Pfam annotation (<http://pfam.xfam.org/>); different lollipop colors indicate variant annotation types (See Supporting Information Fig. S5 for additional protein plots).

deleterious variants not described before, except *SMARCA4* R906H. *SMO* (Frizzled Class Receptor Smoothened) encodes nonclassical G-protein-coupled receptors that are highly expressed in neural tissues and involved in Hedgehog signaling. Two *SMO* variants (Fig. 2; Supporting Information Fig. S5) were detected, R451G in ID2578 (Supporting Information Table S2) in the Frizzled/Smoothened family membrane region, and T640 fs (ID1955) that induces a premature stop codon ending the protein 132 amino acids before the C-terminus. The other recurrent SS-specific genes, *CHD2* and *SMARCA4*, were transcriptional regulators. Chromodomain Helicase DNA Binding Protein 2 (*CHD2*) is important for neurogenesis and de novo mutations in this gene were found in neurodevelopmental disorders.²⁸ *CHD2* is a tumor suppressor chromatin remodeler, previously observed to be mutated and proposed as a cancer driver in chronic lymphocytic leukemia.²⁹ Notably, the transcription co-activator and tumor suppressor *SMARCA4* (SWI/SNF Related, Matrix Associated, Actin Dependent Regulator of Chromatin, Subfamily A, Member 4) were recurrently mutated in the SS group. *SMARCA4* encodes a member of the SWI/SNF nucleosome-remodeling complex whose mutations impact

growth control, differentiation, development and cell adhesion.³⁰ *SMARCA4* somatic variants (R468C in patient ID2513; R906H in ID3060; Fig. 2; Supporting Information Fig. S5) in the two SS patients were validated (Supporting Information Table S2). The deleterious *SMARCA4* R906H mutation was annotated in the COSMIC database (COSM5576007), as previously being observed in gastric cancer,³¹ whereas the putatively damaging R468C variant was not described previously. Both *SMARCA4* variants were localized in the transcription activator chain of the protein, falling, respectively, in the Helicase Sant-associated domain (HSA) and in the helicase ATP-binding domain of the protein (Fig. 2). Phyre2 structural analysis, in particular, suggested a possible strong impact of R906H on ATP binding that is essential for transcriptional activation. SS-specific FGFR1 N577 L (detected in ID2100) and *PTK2* R569L (detected in ID2181) variants were validated and described in a previous NB study (Supporting Information Fig. S5 and Table S2).⁹ These two genes were already associated with NB tumorigenesis^{7,9} although not specifically in relation to patient survival.

The *PTK2*/FAK1 (focal adhesion kinase) variant is located close to the Tyr576 phosphorylation site of the kinase domain

“catalytic loop” (Fig. 2) required for PTK2 activation³² and for mediating NB progression and aggressiveness.³³

Different pathways and functions are hit by mutations in SS and LS patients

Beyond the gene level analysis, somatic mutations falling into different genes that co-participate in the same pathways and/or are linked together in specific functional or interaction networks were investigated, considering mutated genes annotated in Reactome (48%) and KEGG (24%). Thirty-Four Reactome and 12 KEGG pathways were specifically enriched in the LS group, and 12 Reactome and 17 KEGG pathways were recurrently mutated in LS patients and never mutated in SS (LS-specific; Supporting Information Tables S3 and S4). In SS patients, 17 Reactome and one KEGG pathways were specifically enriched, and 25 Reactome and 4 KEGG pathways were SS-specific (Supporting Information Tables S3 and S4). Considering the hierarchical structure of pathway annotation and redundancy and most deleterious mutations, the somatically mutated genes in each of the two patient groups tend to participate to different pathways and pathway classes (Fig. 3; Supporting Information Figure S6) indicating a link between disease aggressiveness and specific processes and functions hit by mutations. Mutations in 15 cell cycle genes were present in LS patients, with several deleterious mutations in genes linked to mitosis, including *CDC27*, *CDCA5*, *CENPC* and *AURKB*. Notch-related genes are mutated in both groups, but several genes (*TBL1X*, *CREBBP*, *NOTCH4*, *NOTCH3*, *TNRC6B* and *TLE2*) specifically belonging to Notch1 signaling are mutated only in LS patients.

In SS patients, the axon guidance pathway was possibly hampered by mutations in 20 genes, involved particularly in NCAM signaling for neurite outgrowth by MAPK2 and MAPK activation (including *ARHGEF11*, *CACNA1G*, *FGF4*, *PTPRA*, *PTK2*, *ANK3*, *SMO* and *NTNG2*) processes, which is important for neurodevelopment and oncogenesis. The MAPK pathway is linked through PTK2 signaling to ERBB4 (including *FGF4*, *PTPRA* and *PTK2*), and *MET* (*LAMA2*, *PTK2* and *LAMA4*). ERBB4 signaling was specifically enriched, and both *MET* signaling and the Cilium assembly pathway (*BBS10*, *SMO*, *INPP5E*) were exclusively mutated in SS patients.

Genes mutated in SS patients cluster into specific pathway-derived subnetworks

A further analysis of the topological structure of mutation gene networks derived from the Reactome pathway annotation, encoding direct relations among genes and their products, detected significant associations of somatically mutated genes in NB SS patients belonging to functionally connected gene networks specific to the worst survival group, which were in accordance with previous observations at the pathway level. Neither the 463 genes mutated in the whole cohort (adjusted *p* value of global mutation clustering 0.43) nor the 213 genes mutated in LS patients (adjusted *p* value 0.93) showed

significant clustering according to Hotnet2 analysis. Conversely, a more pronounced clustering was observed of the 268 genes mutated in SS patients, 79 of which converged into 18 subnetworks of at least three genes (adjusted *p* value of global mutation clustering 0.24) (Supporting Information Table S5). Figure 4 shows the six most relevant network components, comprising 31 functionally connected genes that are somatically mutated specifically in SS patients. The largest component, which was recurrently identified in almost two thirds of SS patients (9 of 14), included nine genes (*NID2*, *LAMA4*, *LAMA2*, *PTK2*, *PTPRA*, *FGG*, *VCL*, *MMP14* and *KSR2*) of the RAF/MAPK signaling pathway and extracellular matrix organization. In addition to the previously observed *PTK2* variant, we validated the D377Y variant, which fell into the Y phosphatase domain of *PTPRA* (Fig. 2), closely connected with *PTK2* in the RAS/MAPK pathway, and the stop gaining variant (E352*) of *LAMA2* (Supporting Information Table S2 and Fig. S5). A second component of six genes (*NALCN*, *UNC79*, *SLC9A9*, *SLC12A1*, *SLC5A8* and *SLC4A9*) that was linked to the transmembrane transport of small molecules was mutated in four SS patients. Two patients carried mutations in two genes of the component (*NALCN* and *SLC9A9* co-mutated in ID2368; *SCL5A8* and *SLC4A9* in ID1955). The third component, which was linked to centrosome maturation, included five genes (*CDK5RAP2*, *CDK11A*, *CEP89*, *TUBGCP6* and *SF11*) mutated in three different patients (ID1965, ID2100, ID2384) (*CEP89*, *TUBGCP6*, and *CDK11A* co-mutated in the patient ID1965). Four genes were involved in lipid and lipoprotein (*SPTLC2* and *ACSL6*) or nucleotide (*AK7*, *AK9* and *ACSL6*) metabolism, which were mutated in three SS patients (ID2368, ID2513, ID2852), with *SPTLC2* and *ACSL6* in the same patient. Two additional SS-specific components were defined by *SMO*, recurrently mutated in two patients and functionally connected *BBS10* and *GAS8* genes co-mutated in a third, and by *KMT2C*, *HOXB3*, and *HOXC4* mutated in ID1965 and ID2852 patients.

Specific mutated genes and deregulated pathways of SS patients are confirmed by analysis of a large independent cohort

To confirm our findings, we analyzed the largest available group of stage M NB with survival data profiled by WES (Pugh cohort).⁸ The 4,120 genes with nonsilent somatic mutations reported in the Pugh cohort were analyzed, separating the 240 patients into SS (221; with overall survival ≤ 5 years) and LS (19; with overall survival >5 years) according to our classification.

SS-specific genes, pathways and component identified in our study were compared to the findings in Pugh cohort (Supporting Information Fig. S7). Of the genes with SS-specific recurrence in our cohort, NFATC1 and OR14J1 were recurrent with SS-specificity also in the Pugh cohort. Furthermore, five genes (*CHD2*, *DIDO1*, *KRTAP4-8*, *ZNF44* and

Class		Pathway		Genes	
DISEASE	Diseases of signal transduction	Oncogenic MAPK signaling	Paradoxical activation of RAF signaling by kinase inactive BRAF	FGG, VCL, KSR2	
			Signaling by high-kinase activity BRAF mutants		
			Signaling by moderate kinase activity BRAF mutants		
			Signaling by RAS mutants		
DEVELOPMENTAL BIOLOGY	Axon guidance	NCAM signaling for neurite out-growth	MAP2K and MAPK activation	ARHGEF11, CACNA1G, KCNQ2, RPS38A4, CACNA1H, FGG, FGF4, VCL, PTPRA, SCN4A, PLXNA3, CLTCL1, PTK2, ANK3, SPRED3, KSR2, SEMA6D, ROBO1, PTPRC, RGMA	
			RET signaling		MAP2K and MAPK activation
		Keratinization	Formation of the cornified envelope	PPL, CELA2A, EVPL, KLK5, PKP3	
EXTRACELLULAR MATRIX ORGANIZATION	Collagen formation	Assembly of collagen fibrils and other multimeric structures	Collagen biosynthesis and modifying enzymes	PLOD2, TLL1, COL1A1, COL5A2, COL11A1, COL11A2, COL18A1, P3H3, COLGALT1, COL12A1	
		Degradation of the extracellular matrix			Activation of Matrix Metalloproteinases
	Integrin cell surface interactions		HSPG2, ITGA7, COL13A1, ITGB8, FBN1		
	Signaling by ERBB4		GH2, FGG, FGF4, VCL, PTPRA, PTK2, SPRED3, KSR2, TNRC6C, GFAP, RPS6KB2		
SIGNAL TRANSDUCTION	Signaling by MET	MET promotes cell motility	MET activates PTK2 signaling	LAMA2, PTK2, LAMA4	
	MAPK1/MAPK3 signaling	RAF/MAP kinase cascade	MAP2K and MAPK activation	FGG, VCL, KSR2	
	Signaling by NOTCH	Signaling by NOTCH1	NOTCH1 Intracellular Domain Regulates Transcription	TBL1X, CREBBP, NOTCH4, NOTCH3, TNRC6B, TLE2	
	Signaling by RHO GTPases	RHO GTPase effectors	RHO GTPases Activate Formins	ZWINT, CENPC, CENPP, KNL1, AURKB	
TRANSPORT OF SMALL MOLECULES	SLC-mediated transmembrane transport	Transport of inorganic cations/anions and amino acids/oligopeptides	SLC2A4, SLC12A1, SLC5A8, SLC38A4, POM121, SLC4A9, NUP85, SLC40A1, SLC01B3, SLC9A9, SLC17A6		
	Ion channel transport	Stimuli-sensing channels	RYR1, TRDN, NALCN, TPCN2, UNC79		
GENE EXPRESSION	RNA Polymerase I Transcription			ERCC6, TAF1B, PTRF	
ORGANELLE BIOGENESIS AND MAINTENANCE	Cillium assembly	Cargo trafficking to the periciliary membrane		BBS10, SMO, INPP5E	
METABOLISM	Metabolism of Lipids	Phospholipid metabolism		INPPL1, ACHE, PNPLA7, TNFAIP8L1, PCYT2, OSBP5, INPP5E	
	Nucleotide metabolism	Synthesis and interconversion of nucleotide di- and triphosphates		AK9, AK7, NME4, DPYS	
	Carbohydrate metabolism	Glycosaminoglycan metabolism	Heparan sulfate/heparin (HS-GAG) metabolism	MGAM, GLCE, ENO3, SLC9A1, NAGLU, PFKP, GLYCTK, NUP93, UST, MAN2B2, HSPG2, KERA, CSPG4	
	Metabolism of vitamins and cofactors				APOB, PCCB, TCN1, HSPG2, LRP1, PTGIS
METABOLISM OF PROTEINS	Translation			EIF3G, RPL12, RPS7	
	Post-translational protein modification	Deubiquitination	UCH proteinases	HCFC1, NFRKB, INO80	
		Asparagine N-linked glycosylation	Transport to the Golgi and subsequent modification		MIA2, SPTB, F5, GRIA1, DCTN5, COG6
		SUMOylation	SUMO E3 ligases SUMOylate target proteins	ER to Golgi Anterograde Transport	Cargo concentration in the ER
	Peptide hormone metabolism			MME, CPA3, CPN1, ISL1	
IMMUNE SYSTEM	Adaptive Immune System	Costimulation by the CD28 family		EREG, PHLPP1, HLA-DQA2, PRKCB, LYN, CDC27, CD80, CD79B, AP1S1, ZBTB16, ITK, TRIP12, HECTD2, RNF25, DCTN5, TNRC6B, LILRA1	
HEMOSTASIS	Platelet activation, signaling and aggregation			F13A1, APOB, PRKCB, LYN, F5, JAK3, CREBBP, GNAS, PRKCZ, A2M, DGKZ, MAG, DOCK4	
REPRODUCTION	Fertilization			ADAM20, IZUMO2, CATSPER1	
NEURONAL SYSTEM	Transmission across Chemical Synapses		Neurotransmitter Receptor Binding And Downstream Transmission In The Postsynaptic Cell	ADCY3, PRKCB, GRIA1, ARHGEF9, CHRNG	
CELL CYCLE	Cell Cycle, Mitotic	M Phase	Mitotic Metaphase and Anaphase	Separation of Sister Chromatids	
			Mitotic Prometaphase	Resolution of Sister Chromatid Cohesion	

Figure 3. Summary of Reactome pathways exclusively mutated or exclusively enriched in LS or in SS NB patients. The figure depicts the hierarchy of Reactome pathways that resulted or contained pathways exclusively mutated in LS (green fill) or in SS (light red fill) patients, or that were enriched in a class-specific way (bold text); the gray fill indicates more general classes at high hierarchical level being not class-specific; for the most high-level specific or specifically enriched class of each group, the corresponding mutated genes are indicated in the right part of the figure (See Supporting Information Tables S3,S4 and Figs. S4,S5 for additional information).

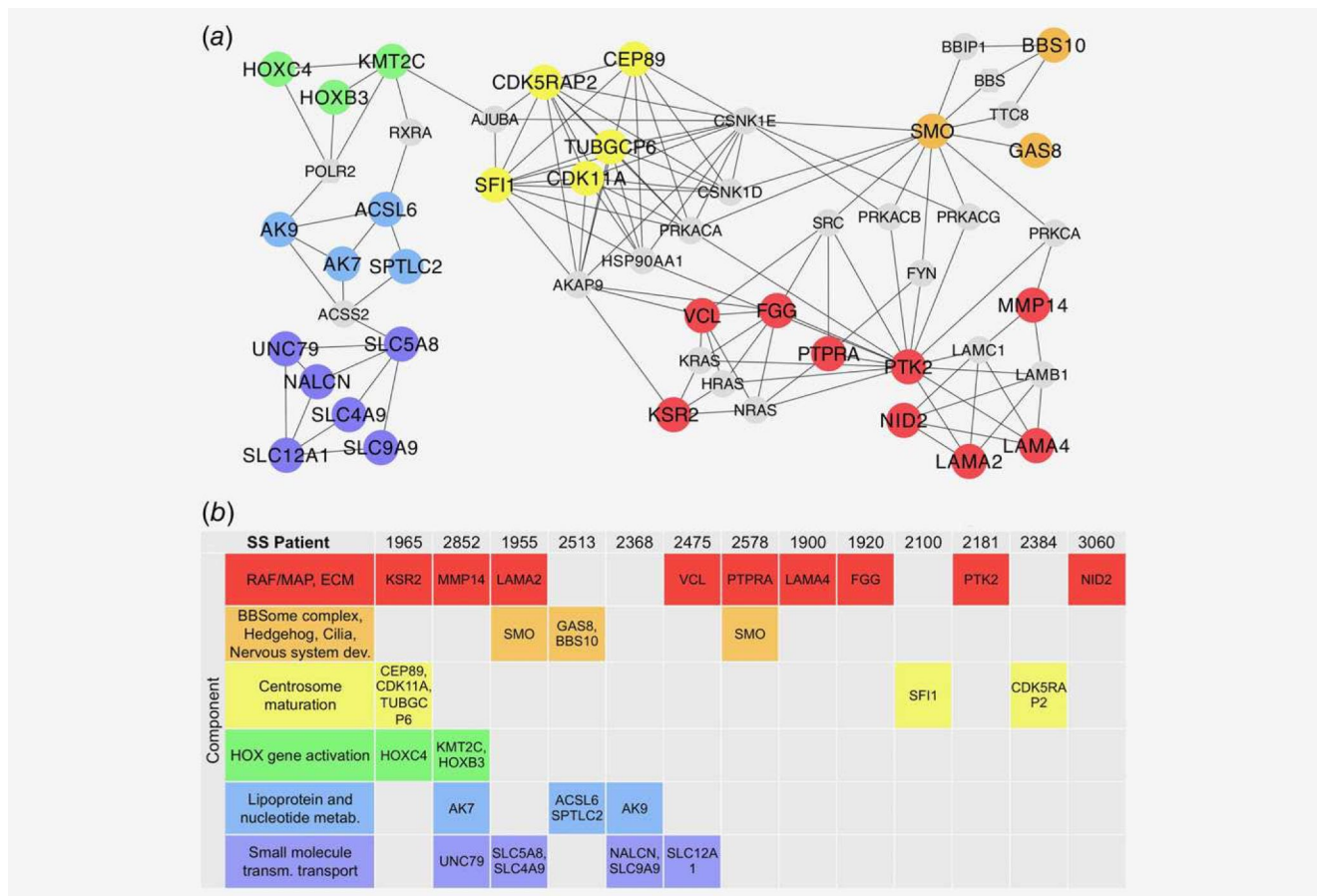


Figure 4. Reactome-derived network of genes somatically mutated in NB patients formed by six SS-specific components. (a) Colored nodes in the net indicate genes mutated in SS patients, with six different components (groups of functionally connected genes somatically mutated in NB patients with rapid disease progression identified by Hotnet2 analysis) in different colors; gray nodes represent genes directly connecting the components according to pathway topology and non mutated in the analyzed cohort (edges between gray nodes are omitted). (b) Each component was recurrently mutated in different patients, and specific tumors carried mutations in multiple genes and components of the network.

ZNF91) with SS-specific recurrence in our cohort were mutated with SS-specificity also in Pugh patients (Supporting Information Fig. S7A). The *SMARCA4* mutation was reported in a patient with a survival of 61 months according to our classification. Our analysis indicated 34 genes mutated in one SS patient of our cohort and recurrently mutated with SS-specificity in the Pugh cohort, including *ABCA13* which was mutated in seven patients.

Nineteen of the 78 genes prioritized because involved in SS-specific pathways identified (Fig. 3) were mutated with SS-specificity in the Pugh cohort, including five (*ANK3*, *COL11A1*, *COL12A1*, *COL1A1*, *PNPLA7*) that were also recurrent (Supplementary Figure 7A). Five genes (*AK7*, *NALCN*, *PTK2*, *SLC5A8*, *TUBGCP6*) included both in SS-specific pathways and in significant subnetworks (Fig. 4) identified in our study were mutated only in SS patients of the Pugh cohort.

Furthermore, analysis with HotNet2 was performed considering the 1,810 genes somatically mutated in SS patients of the Pugh cohort and mapped in the Reactome-derived

network, detecting 14 significant (adjusted *p* value 0.05) pathway-derived subnetworks involving 143 genes (Supporting Information Table S6). Extracellular matrix organization, carbohydrate and lipid metabolism emerged both from our study and (Figs. 3 and 4) Pugh data. *PTK2*, which was shown to be mutated with SS-specificity both in our data and in the Pugh cases, in the network of mutations detected in Pugh patients (Supporting Information Table S6), was directly linked to two gene groups involved in ECM (Supporting Information Fig. S7B).

Discussion

One of the major challenges for oncologists treating HR-NB is the high percentage of patients showing rapid disease progression despite multimodal treatment. Of these, approximately 60% of HR-NBs have a fatal course within 5 years of diagnosis. To identify genetic abnormalities associated with disease aggressiveness, we compared a somatic mutation profile of HR-NB patients with SS and LS.

Table 2. Information on drugs available in relation to genes carrying deleterious mutations in NB patients.

Gene ID	Gene Description	Variant discovered	FDA approved drugs	Drug class	References for drugs use
SDHB	Succinate dehydrogenase complex, subunit B, iron sulfur (lp)	L7FS	Succinic acid	Small molecule	He et al., 2004, Citric acid cycle intermediates as ligands for orphan G-protein-coupled receptors., <i>Nature</i> ; Southern et al., 2013, Screening β -arrestin recruitment for the identification of natural ligands for orphan G-protein-coupled receptors., <i>J Biomol Screen</i>
SMO	Smoothed, frizzled family receptor	R451G; T640FS	Vismodegib	Small molecule inhibitor	Yauch et al., 2009, Smoothed mutation confers resistance to a Hedgehog pathway inhibitor in medulloblastoma., <i>Science</i> ; Wang et al., 2012, Identification of a novel Smoothed antagonist that potently suppresses Hedgehog signaling., <i>Bioorg. Med. Chem.</i>
			Fluocinonide	Small molecule	Wang et al., 2010, Identification of select glucocorticoids as Smoothed agonists: potential utility for regenerative medicine., <i>Proc. Natl. Acad. Sci. U.S.A.</i>
			Halcinonide	Small molecule	Wang et al., 2011, Glucocorticoid hedgehog agonists in neurogenesis., <i>Vitam. Horm.</i> ; Wojnar et al., 1986, Androstene-17-thioketals. 1st communication: glucocorticoid receptor binding, antiproliferative and antiinflammatory activities of some novel 20-thiasteroids (androstene-17-thioketals)., <i>Arzneimittelforschung</i>
PTK2	PTK2 protein tyrosine kinase 2	R569L	Masitinib	Kit inhibitor	Dubreuil et al., 2009, Masitinib (AB1010), a potent and selective tyrosine kinase inhibitor targeting KIT., <i>PLoS ONE</i>
MMP14	Matrix metalloproteinase 14 (membrane-inserted)	P8FS	Prinomastat	Mmp inhibitor	Abbenante et al., 2005, Protease inhibitors in the clinic., <i>Med Chem</i>

General tumor genomic landscapes of HR-NB patients with SS and LS were similar, exhibiting close frequencies of variants and numbers of somatically mutated genes per patient.

Nevertheless, few genes were recurrently mutated specifically in the SS group, including *SMARCA4*, *SMO*, *ZNF44* and *CHD2*. *SMARCA4* also known as *BRG1*, is a tumor suppressor gene of the SWI/SNF complex^{34–37} that shows inactivating mutations or overexpression in several adult cancers.^{38,39} We found two missense mutations, R468C and R906H, in the Transcription Activator chain region responsible for DNA and ATP binding and ATP hydrolysis, which are predicted to be very dangerous. The loss of function of the *SMARCA4* protein likely impaired its activity with damage of the SWI/SNF that is involved in chromatin remodeling. Jubierre et al. showed that the *SMARCA4* gene has a role in the proliferation of NB cells both *in vitro* and *in vivo*.⁴⁰ Matsubara et al. observed a correlation between *SMARCA4* mutations and loss of function in lung cancer cell lines, indicating an association with aggressive tumor behavior and worse patient survival.⁴¹ Thus, *SMARCA4* mutations may determine the loss of function associated with tumor aggressiveness and poor NB patient survival. As it has been demonstrated that *SMARCA4* and *TERT* are functionally linked,⁴² SWI/SNF damage could alter *TERT* function,⁴³ which one of the most important genes rearranged in NB.⁴⁴

Among the nonrecurrent gene mutations, we validated the FGFR1 N546 K variant, as well as novel deleterious variants of *CREBBP* and *OR5T1* (Supporting Information Fig.S5 and Table S2), whose mutations were found to be associated with

NB aggressiveness.^{7,8} Remarkably, somatic mutations occurring in SS or LS patients hit different pathways. In addition, functional gene networks, corresponding to sub pathways, hit only SS patients. Numerous gene variants observed in the tumor of SS patients affected the RAF/MAP kinase cascade, as well as MET and ERBB4 pathways linked to PTK2 signaling. MAP2K and MAPK activation, specific of SS tumors, are of interest because they can be involved in cell motility by triggering PTK2 signaling and Matrix Metalloproteinases activation. These results agree with previous data on the enrichment of somatic mutations in FAK signaling and cell adhesion signaling.⁹

Furthermore, several genes connected to RAF/MAPK signaling were mutated in SS (*NID2*, *LAMA4*, *LAMA2*, *PTK2*, *PTPRA*, *FGG*, *VCL*, *MMP14* and *KSR2*), impacting extracellular matrix organization, regulation of cell adhesion and migration. A previous observation of PTK2 mutation in a HR-NB patient by Lasorsa et al., further strengthens the importance of PTK2 signaling in aggressive tumors. Mutations of *NELL1*, *UNC79* and *COL5A2* genes in one SS patient of our cohort (Supporting Information Table S1B), were previously reported in NB patients with SS, albeit with different variants.⁹ Particularly relevant groups of clustered genes mutated in SS were involved in centrosome maturation, in the regulation of the cell cycle, in ciliary basal body docking (*CDK5RAP2*, *CDK11A*, *CEP89*, *TUBGCP6* and *SFI1*) and in cilium assembly (the recurrently and SS specifically mutated *SMO*, and *GAS8*, *BBS10* and *AK7*). Our observations on the mutations linked to the chromosome remodeling pathway in SS tumors support the role of chromosome instability in NB,⁴⁵ providing further

explanation for the observed CNA in patients with fatal outcomes.^{4,46}

The clustering of somatic mutations observed in SS patients reflected two phenomena: specific functions targeted in several SS patients, as observed for the RAF/MAPK signaling component, and co-occurrence in the same patient of mutations in two or more functionally connected genes.

The analysis of a sizeable independent cohort of 240 stage M NB patients⁸ gave additional strength to our findings. Mutations in Pugh SS patients targeting genes prioritized in our cohort (*ANK3*, *COL11A1*, *COL12A1*, *COL1A1*, *PNPLA7*, *AK7*, *NALCN*, *PTK2*, *SLC5A8* and *TUBGCP6*) as SS-specific based on recurrence, pathway enrichment and/or pathway-derived network topology analysis, were particularly noteworthy and supported our results. The reconstruction and analysis of pathway-derived mutation networks reported in Pugh SS patients further backed the observations done in our cohort about the deregulation of lipid metabolism and RAF/MAP signaling in relation to ECM mutated genes.

Recent comparison of matched primary and relapsed NB tumors revealed that disease progression is accompanied by an increased mutational load in MAPK pathway genes, exhibiting new mutations in the MAPK pathway that were not present at the onset of disease, and accumulated in tumors of relapsing patients.^{47,48} Our findings of specific MAPK signaling pathway damages (also observed by Eleveld *et al.*⁴⁷ and Schramm *et al.*⁴⁸) may be relevant for more efficacious therapeutic management of patients at diagnosis. Specific genes mutated at diagnosis exclusively in pathways belonging to the SS group could be candidates for pharmacological targeting. *SMO*, *PTK2*, *MMP14* and *SDHB* are quite interesting as they are targeted by FDA approved drugs according to the Drug Gene Interaction Database (DGID: <http://dgidb.genome.wustl.edu/>) (Table 2). Recently, Padovan-Merhar *et al.*⁴⁹ reported an increased *SMO* mutation frequency in tumors of HR-NB patients at relapse, showing that most of these new mutations

are targetable and give an additional tool to treat relapsing patients. Functional investigation is mandatory to assess the potential significance of mutated genes as therapeutic targets, and further study is needed to evaluate drugs, such as Masitinib and Vismodegib, for NB therapy.

In our study, two groups of HR-NB patients with different outcome were characterized, providing new data on mutations recurrently affecting specific pathways and functions in patients with SS, informing the molecular features, beyond well-defined CNA patterns, that are associated with high tumor aggressiveness.

Author Contributions

GPT, MRE have designed, planned and performed the study. MP realized DNA extraction from tumor biopsy, exome library and sequencing on Ion proton sequencer. AB, AC and SB performed bioinformatics analyses and contributed new systems biology methods. KM and LL have contributed for collection of matched tumor biopsies and peripheral blood samples, DNA extraction from peripheral blood. MC, AVL and RL revised the manuscript. MRE, AB, SB and GPT wrote the manuscript that has been revised and approved by all Authors.

Acknowledgements

The authors would like to thank Drs. Alberto Garaventa, Angela Rita Sementa, Riccardo Haupt for providing patient data. The authors would like also to thank Dr. Francesca Dal Pero of Roche Company for support on 454 sequencing, Dr. Silvia Bresolin for technical support on Roche 454-GS Junior DNA sequencing platform and Prof. Giuseppe Basso for fruitful result discussion. The present work was mainly supported by Fondazione Italiana per la Lotta al Neuroblastoma. The authors thank for financial support to SB Fondazione Cassa di Risparmio di Padova e Rovigo Progetti di Eccellenza 2011/2012 by Ministero dell'Istruzione, dell'Università e della Ricerca PRIN 2010/11 (2010NYKNS7_002) and University of Padova. AB is recipient of a PhD fellowship from the University of Padova (PhD in Biosciences)

References

- Shimada H, Ambros IM, Dehner LP, *et al.* The International Neuroblastoma Pathology Classification (the Shimada system). *Cancer* 1999;86:364–72.
- Luksch R, Castellani MR, Collini P, *et al.* Neuroblastoma (peripheral neuroblastic tumours). *Crit Rev Oncol Hematol* 2016;107:163–81.
- Coco S, Theissen J, Scaruffi P, *et al.* Age-dependent accumulation of genomic aberrations and deregulation of cell cycle and telomerase genes in metastatic neuroblastoma. *Int J Cancer* 2012;131:1591–600.
- Stigliani S, Coco S, Moretti S, *et al.* High genomic instability predicts survival in metastatic high-risk neuroblastoma. *Neoplasia* 2012;14:823–32.
- Schleiermacher G, Mosseri V, London WB, *et al.* Segmental chromosomal alterations have prognostic impact in neuroblastoma: a report from the INRG project. *Br J Cancer* 2012;107:1418–22.
- Molenaar JJ, Koster J, Zwijnenburg DA, *et al.* Sequencing of neuroblastoma identifies chromothripsis and defects in neurogenesis genes. *Nature* 2012;483:589–93.
- Sausen M, Leary RJ, Jones S, *et al.* Integrated genomic analyses identify ARID1A and ARID1B alterations in the childhood cancer neuroblastoma. *Nat Genet* 2013;45:12–7.
- Pugh TJ, Morozova O, Attiyeh EF, *et al.* The genetic landscape of high-risk neuroblastoma. *Nat Genet* 2013;45:279–84.
- Lasorsa VA, Formicola D, Pignataro P, *et al.* Exome and deep sequencing of clinically aggressive neuroblastoma reveal somatic mutations that affect key pathways involved in cancer progression. *Oncotarget* 2016;7:21840–52.
- Calabrese FM, Clima R, Pignataro P, *et al.* A comprehensive characterization of rare mitochondrial DNA variants in neuroblastoma. *Oncotarget* 2016;7:49246–58.
- Capasso M, Diskin SJ. Genetics and genomics of neuroblastoma. *Cancer Treat Res* 2010;155:65–84.
- Cingolani P, Patel VM, Coon M, *et al.* Using *Drosophila melanogaster* as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Front Genet* 2012;3:35.
- Cingolani P, Platts A, Wang le L, *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* 2012;6:80–92.
- Forbes SA, Beare D, Gunasekaran P, *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res* 2015;43:D805–11.
- Landrum MJ, Lee JM, Benson M, *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res* 2016;44:D862–8.

16. Dong C, Wei P, Jian X, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum Mol Genet* 2015;24:2125–37.
17. Liu X, Jian X, Boerwinkle E. dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Hum Mutat* 2013;34:E2393–402.
18. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 2016;536:285–91.
19. Kelley LA, Mezulis S, Yates CM, et al. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc* 2015;10:845–58.
20. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;28:27–30.
21. Fabregat A, Sidiropoulos K, Garapati P, et al. The Reactome pathway Knowledgebase. *Nucleic Acids Res* 2016;44:D481–7.
22. Sales G, Calura E, Cavalieri D, et al. graphite - a Bioconductor package to convert pathway topology to gene network. *BMC Bioinformatics* 2012;13:20.
23. Leiserson MD, Vandin F, Wu HT, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat Genet* 2015;47:106–4.
24. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;26:841–2.
25. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 2010;26:589–95.
26. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;20:1297–303.
27. Robinson JT, Thorvaldsdottir H, Winckler W, et al. Integrative genomics viewer. *Nat Biotechnol* 2011;29:24–6.
28. Carvill GL, Heavin SB, Yendle SC, et al. Targeted resequencing in epileptic encephalopathies identifies de novo mutations in CHD2 and SYNGAP1. *Nat Genet* 2013;45:825–30.
29. Rodriguez D, Bretones G, Quesada V, et al. Mutations in CHD2 cause defective association with active chromatin in chronic lymphocytic leukemia. *Blood* 2015;126:195–202.
30. Reisman D, Glaros S, Thompson EA. The SWI/SNF complex and cancer. *Oncogene* 2009;28:1653–68.
31. Kuboki Y, Yamashita S, Niwa T, et al. Comprehensive analyses using next-generation sequencing and immunohistochemistry enable precise treatment in advanced gastric cancer. *Ann Oncol* 2016;27:127–33.
32. Owen JD, Ruest PJ, Fry DW, et al. Induced focal adhesion kinase (FAK) expression in FAK-null cells enhances cell spreading and migration requiring both auto- and activation loop phosphorylation sites and inhibits adhesion-dependent tyrosine phosphorylation of Pyk2. *Mol Cell Biol* 1999;19:4806–18.
33. Lee S, Qiao J, Paul P, et al. FAK is a critical regulator of neuroblastoma liver metastasis. *Oncotarget* 2012;3(12):1576–87.
34. Decristofaro MF, Betz BL, Rorie CJ, et al. Characterization of SWI/SNF protein expression in human breast cancer cell lines and other malignancies. *J Cell Physiol* 2001;186:136–45.
35. Schneppenheim R, Fruhwald MC, Gesk S, et al. Germline nonsense mutation and somatic inactivation of SMARCA4/BRG1 in a family with rhabdoid tumor predisposition syndrome. *Am J Hum Genet* 2010;86:279–84.
36. Wilson BG, Roberts CW. SWI/SNF nucleosome remodellers and cancer. *Nat Rev Cancer* 2011;11:481–92.
37. Robinson G, Parker M, Kranenburg TA, et al. Novel mutations target distinct subgroups of medulloblastoma. *Nature* 2012;488:43–8.
38. Sentani K, Oue N, Kondo H, et al. Increased expression but not genetic alteration of BRG1, a component of the SWI/SNF complex, is associated with the advanced stage of human gastric carcinomas. *Pathobiology* 2001;69:315–20.
39. Bai J, Mei PJ, Liu H, et al. BRG1 expression is increased in human glioma and controls glioma cell proliferation, migration and invasion in vitro. *J Cancer Res Clin Oncol* 2012;138:991–8.
40. Jubierre L, Soriano A, Planells-Ferrer L, et al. BRG1/SMARCA4 is essential for neuroblastoma cell viability through modulation of cell death and survival pathways. *Oncogene* 2016;35:5179–90.
41. Matsubara D, Kishaba Y, Ishikawa S, et al. Lung cancer with loss of BRG1/BRM, shows epithelial mesenchymal transition phenotype and distinct histologic and genetic features. *Cancer Sci* 2013;104:266–73.
42. Wu S, Ge Y, Huang L, et al. BRG1, the ATPase subunit of SWI/SNF chromatin remodeling complex, interacts with HDAC2 to modulate telomerase expression in human cancer cells. *Cell Cycle* 2014;13:2869–78.
43. Maida Y, Yasukawa M, Okamoto N, et al. Involvement of telomerase reverse transcriptase in heterochromatin maintenance. *Mol Cell Biol* 2014;34:1576–93.
44. Peifer M, Hertwig F, Roels F, et al. Telomerase activation by genomic rearrangements in high-risk neuroblastoma. *Nature* 2015;526:700–4.
45. Tonini GP. Growth, progression and chromosome instability of Neuroblastoma: a new scenario of tumorigenesis? *BMC Cancer* 2017;17:20.
46. Defferrari R, Mazzocco K, Ambros IM, et al. Influence of segmental chromosome abnormalities on survival in children over the age of 12 months with unresectable localised peripheral neuroblastic tumours without MYCN amplification. *Br J Cancer* 2015;112:290–5.
47. Eleveld TF, Oldridge DA, Bernard V, et al. Relapsed neuroblastomas show frequent RAS-MAPK pathway mutations. *Nat Genet* 2015;47:864–71.
48. Schramm A, Koster J, Assenov Y, et al. Mutational dynamics between primary and relapse neuroblastomas. *Nat Genet* 2015;47:872–7.
49. Padovan-Merhar OM, Raman P, Ostrovskaya I, et al. Enrichment of Targetable Mutations in the Relapsed Neuroblastoma Genome. *PLoS Genet* 2016;12:e1006501.

REFERENCES

- Chen CI, Bergsagel PL, Paul H, Xu W, Lau A, Dave N *et al.* Single-agent lenalidomide in the treatment of previously untreated chronic lymphocytic leukemia. *J Clin Oncol* 2011; **29**: 1175–1181.
- Badoux XC, Keating MJ, Wen S, Lee BN, Sivina M, Reuben J *et al.* Lenalidomide as initial therapy of elderly patients with chronic lymphocytic leukemia. *Blood* 2011; **118**: 3489–3498.
- Chanan-Khan A, Miller KC, Musail L, Lawrence D, Padmanabhan S, Takeshita K *et al.* Clinical efficacy of lenalidomide in patients with relapsed or refractory chronic lymphocytic leukemia: results of a phase II study. *J Clin Oncol* 2006; **24**: 5343–5349.
- Ferrajoli A, Lee BN, Schlette E, O'Brien SM, Gao H, Wen S *et al.* Lenalidomide induces complete and partial remissions in patients with relapsed and refractory chronic lymphocytic leukemia. *Blood* 2008; **111**: 5291–5297.
- Rai K, Peterson BL, Applebaum FR, Koltitz J, Elias L, Shepherd L *et al.* Fludarabine compared with chlorambucil as primary therapy for chronic lymphocytic leukemia. *N Engl J Med* 2000; **343**: 1750–1757.
- Eichhorst BF, Busch R, Stilgenbauer S, Stauch M, Bergmann MA, Ritgen M *et al.* First-line therapy with fludarabine compared with chlorambucil does not result in a

- major benefit for elderly patients with advanced chronic lymphocytic leukemia. *Blood* 2009; **114**: 3382–3391.
- Catovsky D, Richards S, Matutes E, Oscier D, Dyer MJ, Bezares RF *et al.* Assessment of fludarabine plus cyclophosphamide for patients with chronic lymphocytic leukaemia (the LRF CLL4 Trial): a randomized controlled trial. *Lancet* 2007; **370**: 230–23.



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

© The Author(s) 2017

Supplementary Information accompanies this paper on the Leukemia website (<http://www.nature.com/leu>)

OPEN

Genomic landscape characterization of large granular lymphocyte leukemia with a systems genetics approach

Leukemia (2017) **31**, 1243–1246; doi:10.1038/leu.2017.49

Large granular lymphocyte (LGL) leukemia is a rare clonal disease characterized by a persistent increase in the number of CD8+ cytotoxic T cells or CD16/56+ natural killer (NK) cells. It is associated with recurrent infections, severe cytopenias and autoimmune diseases. JAK/STAT pathway activation, deregulation of pro-apoptotic pathways (sphingolipid and FAS/FAS ligand) and activation of pro-survival signaling pathways (PI3K/AKT and RAS) are known hallmarks of LGL leukemia. Activating somatic *STAT3* mutations have been reported in the SH2 domain (30–70% of cases),^{1–3} and in the DNA-binding or coiled-coil domain (2%).⁴ *STAT5B* mutations are more rare, but typical of CD4+ T-LGL leukemia cases.^{5–7} The JAK/STAT pathway can also be activated by non-mutational mechanisms such as increased interleukin-6 (IL-6) secretion and epigenetic inactivation of JAK-STAT pathway inhibitors.⁸ Indeed, aberrant *STAT* signaling is observed in almost all LGL leukemia patients irrespective of the presence of JAK/STAT mutations.⁹

To characterize the genomic landscape of LGL leukemia, we performed whole-exome sequencing (Supplementary Methods and Supplementary Figure 1) from 19 paired tumor-control samples derived from untreated LGL leukemia patients including conventional CD8+ ($n=13$) T-cell cases, and more rare CD4+ or CD4+CD8+ T-cell cases ($n=3$), and NK LGL leukemias ($n=3$; Supplementary Table 1). Eleven *STAT*-mutation-negative patients were included for identification of new driver mutations. All sequenced samples were highly purified sorted cell populations (either CD8+ or CD4+ T cells or NK cells), and T-cell receptor Vbeta analysis confirmed monoclonal expansions in the tumor fractions of T-cell cases (see Supplementary Methods and Supplementary Table 1). The average sequencing coverage in the tumor samples was 32x (Supplementary Figure 2). Both the coverage and the number of raw called variants were similar in tumor and

control samples. After selecting high confidence variants (see Supplementary Methods), and filtering out variants already described in human populations single nucleotide polymorphism database and/or with allele frequency higher than 5% in exome aggregation consortium exomes, 28 508 somatic variants in 16 518 genes were identified in the whole cohort with a high prevalence of C>T and G>A transversions (Supplementary Figure 3A). Next, among high confidence and rare variants, we selected 370 variants in 347 genes with a strong predicted functional impact (Supplementary Methods and Supplementary Table 2). The observed differences in numbers of somatic mutations (range 5–40, average 20) and genes involved (range 4–41, 19) per patient were not because of coverage differences (Supplementary Figure 3B). A slight tendency toward more mutated genes per patient in *STAT*-mutation-positive (22.9 in average) versus negative patients (18.4 in average) was noticed. Sanger sequencing validations of somatic variants were obtained in 14 genes (Supplementary Table 3 and Supplementary Figure 4) being recurrent or prioritized according to functional criteria and/or connections emerged by integrated pathway-derived networks. The positions of the mutations in protein domains of selected genes are shown in Supplementary Figure 5.

In addition to *STAT3* (all in CD8+ T-LGL) and *STAT5B* (CD4+ and CD8+ cases) mutations (in 8/19 patients, 42%), 14 other genes had recurrent mutations including transcriptional/epigenetic regulator, tumor suppressor and cell proliferation genes (Figure 1a and 2a). *KMT2D* has been linked to lymphomagenesis¹⁰ and found to be frequently mutated in other cancers. Mutations of *PCLO*, a calcium sensor-regulating cAMP-induced exocytosis, have been previously reported in diffuse large B-cell lymphoma. *FAT4* is an upstream regulator of stem cell genes both during development and cancer, functioning as a tumor growth suppressor via activation of Hippo signaling. It was previously found recurrently mutated in human cancers, including leukemias. Also the other recurrently mutated gene, *ARL13B*, is linked to Hippo signaling. It encodes a small

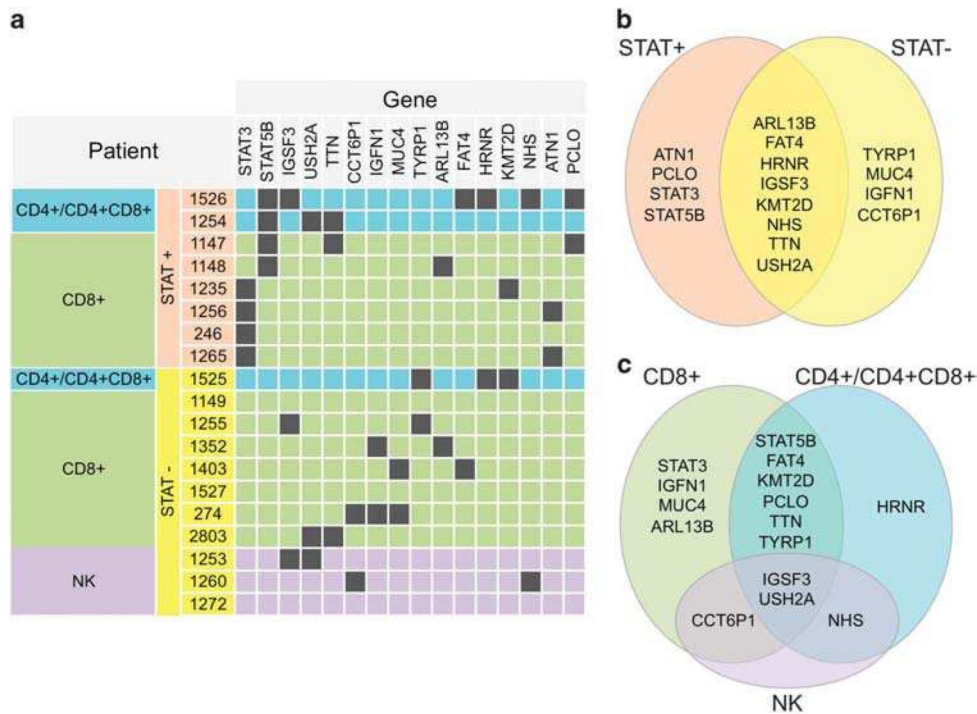


Figure 1. Recurrent somatic mutations in LGL leukemia patients. (a) The table indicates the genes that carry somatic variants in more than one patient, with a color code showing *STAT3* and *STAT5B* status and classification of patients. (b) Recurrently mutated gene sets found only in *STAT*-mutation-negative patients (*STAT*⁻), only in *STAT*-mutation-positive patients (*STAT*⁺) or in both groups. (c) Recurrently mutated genes that are found only in one or are shared among patient classes (*CD8*⁺, *CD4*⁺/*CD4*⁺*CD8*⁺ and *NK*⁺).

GTPase of primary cilia whose role in cell cycle control has recently been recognized, and they crosstalk with several signaling pathways including Hippo. *ARL13B* and *FAT4* genes were mutated in a mutually exclusive way. Additional non-recurrent somatic mutations of *YAP1* and of its inhibitor *AMOTL1* point toward an involvement of Hippo signaling deregulation in LGL leukemia.

When comparing the mutation profile between three different phenotypic LGL subgroups, qualitative and quantitative differences were observed, although the clinical characteristics of patients did not markedly differ (see details in Supplementary Results and Supplementary Table 1). Interestingly, higher mutation burden was observed in *CD4*⁺ T-LGL leukemia cases (Figure 2b). As the sequencing depth across samples did not vary significantly (Supplementary Figure 3), the differences in mutation load likely reflect a different natural history of the LGL phenotypes. Cytomegalovirus-derived stimulation and restricted usage of T-cell receptor $V\beta$ has been associated with *CD4*⁺ T-LGL cases,¹¹ and this could relate to the higher number of mutations. In the *CD4*⁺ group, only *STAT5B* and *HRNR* genes had recurrent mutations (Figure 1b). *HRNR* is a calcium-binding protein involved in hematopoietic progenitor cell differentiation, and it is mutated, amplified or overexpressed in many cancers. In *NK* LGL leukemias (all *STAT*-mutation-negative), 31 genes harbored somatic mutations including several ‘cancer genes’ such as *KRAS*, *PTK2*, *NOTCH2*, *CDC25B*, *HRASLS*, *RAB12*, *PTPRT* and *LRBA*.

Next, a custom knowledge-based ‘systems genetic’ approach, reminiscent of strategies recently implemented to interpret genome-wide transcriptome deregulation in cancer,^{12,13} provided the functional prioritization of mutated genes. As mutations hitting different genes can drive a similar phenotype in different patients and concur to it if present in the same patient, we reconstructed a pathway-derived meta-network depicting direct interactions and functional relations between genes somatically mutated in LGL leukemias. We identified 119 KEGG and 426 Reactome pathway-derived networks, each including at least one of the 347 previously prioritized mutated genes associated to high

confidence, rare and high-impact variants. The union of all pathway-derived networks generated a meta-network with 118 (34%) mutated genes, giving a non-redundant representation of functional relations, based on direct interactions between somatically mutated genes. Remarkably, 47 mutated genes were directly connected to at least another mutated gene in 18 multigene components (groups of genes whose products directly interact, that is, encode proteins taking part in the same molecular complex or regulating each other). Considering co-participation of mutated genes in pathways including *STAT* genes as additional functional link, seven multigene components connected by direct relations and three isolated genes converged into a component of 26 genes. In this reconstructed LGL leukemia network (Figure 2c and Supplementary Figure 6), 61 somatically mutated genes (occurring in many cases only in one sample) preferentially fall into a limited number of highly connected pathways, and in this manner collectively form a functional module hit by somatic mutations in LGL leukemia. The largest network component included 24 mutated genes either directly linked to *STAT* genes, to their neighbors and/or participating in pathways including *STAT* genes (Figure 2c). Beyond JAK-*STAT* signaling, the ‘*STAT*-related component’ included genes intervening in several other connected paths such as acute and chronic myeloid leukemia, ErbB, HIF-1, insulin, T-cell receptor and VEGF signaling pathways. In 16 out of 19 patients, at least one gene of this group was mutated with some patients showing more than one hit in the gene group. For instance, one *STAT*-mutation-negative *CD4*⁺ patient presented with mutated alleles in three genes of the component (*CD40LG*, *F8* and *PLA2G4C*). The similar variant allele frequency values of the validated variants support their co-presence in the dominant LGL leukemic clone (Supplementary Table 3). Altogether, 8 of 11 *STAT*-mutation-negative patients carried validated somatic mutations in at least one of the ‘*STAT*-related component’ genes, such as in *FLT3*, *KRAS*, *ADCY3*, *ANGPT2* and *PTK2*. These mutated genes also connect the *STAT* component to the MAPK-Ras-ERK (Figure 2c) pathway and to the IL-15, all known to be deregulated in LGL

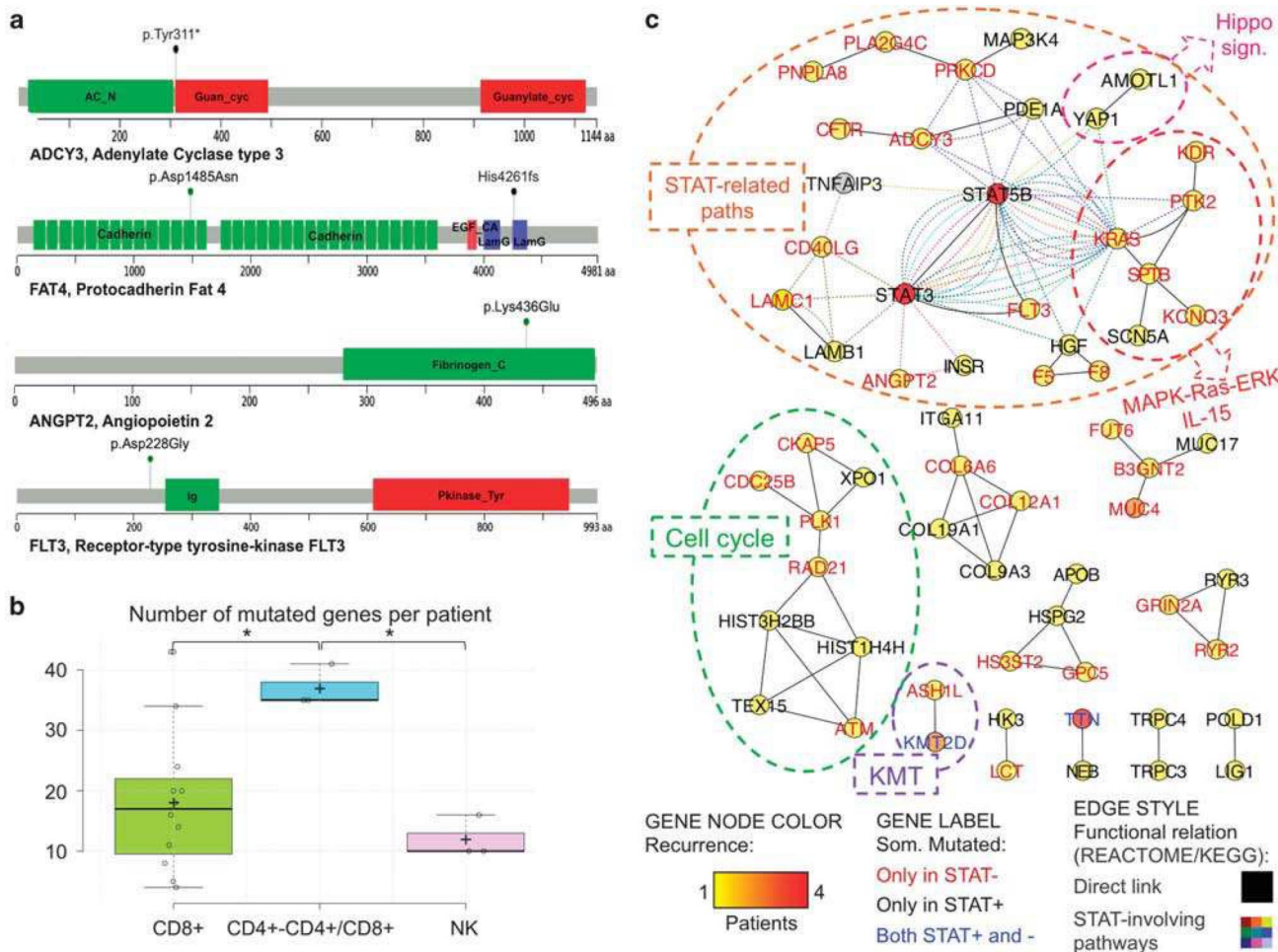


Figure 2. (a) Impact of selected somatic variants to protein products. Lollipop plots show the type and the position of somatic variants of four selected genes in relation to the protein sequence and domain structure (see Supplementary Figure 5 for an extended version of the figure including additional genes). The *ADCY3* Tyr311* variant induces a very premature stop preventing the synthesis of the protein region including Guanylate cyclase, ATP and Mg²⁺ domains; *FAT4* presents two variants, the high-impact missense variant Asp1485Asn in the Cadherin 14 domain and the frameshift variant Hys4261fs inducing a stop codon before Laminin G-like domain truncating the protein before the EGF-like 6 domain and the C terminal; *ANGPT2* presents the high-impact missense variant Lys463Glu in Fibrinogen C-terminal domain implicated in protein-protein interactions, and *FLT3* shows a high-impact Asp228Gly variant. (b) Number of mutations per patient in each class. Normal distribution of values was confirmed with the Shapiro–Wilk test ($P=0.099$). Both analysis of variance ($P=0.009$) and pairwise Tukey s.d. *post hoc* tests (P -values 0.010 and 0.019 in the comparisons of CD4+/-CD4+/CD8+ with NK and CD8+, respectively) confirmed the statistical significance of the observed difference. (c) LGL leukemia mutation network. The network shows the functional relations of genes somatically mutated in LGL leukemia patients, according to the integration of KEGG and Reactome pathway topology (see the text and Supplementary Methods for details on the pathway-derived network reconstruction procedure); network nodes represent somatically mutated genes, node color indicates recurrence (according to the legend heat color scale) in the cohort, node label indicates the gene Symbol (different label colors indicate genes that are mutated only in STAT-mutation-positive (STAT+), only in STAT-mutation-negative (STAT-) or in both patient groups, as shown in the legend); genes are connected with black solid lines if they are directly connected in KEGG- and/or Reactome-derived networks or with colored dashed lines if they participate in pathways including STAT3 and/or with STAT5B (see Supplementary Figure 6 for a detailed version of the network).

leukemia.¹⁴ For example, PTK2 is a non-receptor protein-tyrosine kinase, which is highly expressed in T cells and it regulates several processes, including cell cycle progression, cell proliferation and apoptosis, activation of numerous pathways such as PI3K/AKT signaling MAPK/ERK and MAP kinase signaling cascades. Also the mutated *ANGPT2* is linked to PI3K-AKT and RAS signaling pathways that it antagonizes. *ANGPT2* is expressed in lymphocytes and controls T-cell proliferation. *ANGPT2* and other angiogenic factors are reportedly involved in chronic lymphocytic leukemia where they exert pro-survival effects. Other STAT-connected genes were receptors such as *CD40LG* (modulates B-cell function, regulates immune system and participates in

STAT3 as well as in IL and NFAT signaling pathways) and *FLT3* (a class III receptor tyrosine kinase that promotes the phosphorylation of various proteins and kinases in the PI3K/AKT/mTOR, RAS and JAK/STAT signaling pathways). Interestingly, *CD40LG* was annotated in the same KEGG pathways as *TNFAIP3* (Figure 2c), which is a negative regulator of NF- κ B signaling and known tumor suppressor gene, and was recently found to be mutated in 8% of T-LGL leukemia patients.¹⁵ Other relevant variants confirmed in STAT-mutation-negative patients and connected to the STAT pathway were *KRAS* and the kinase *KDR/VEGFR2*.

Other components (and pathways) not directly linked to the main lesions were also of interest. Nine genes were linked to cell

cycle regulation, and include the *CDC25b* gene and *ATM*, which is involved in apoptosis and *P53* signaling (Figure 2c). Furthermore, the epigenetic node included the recurrently mutated *KMT2D*, which is connected to *ASH1L*. Both are histone methyltransferases involved in epigenetic regulation of gene expression programs and are part of the ASCOM complex, involved in transcriptional co-activation. The networks of genes mutated in individual CD8+ and CD4+ or NK LGL leukemia patients and in each patient subgroup are presented in the Supplementary Figures 7–9.

To conclude, with the systems genetic approach, we were able to map individual mutations found in LGL leukemia patients in novel functional modules. The central role of JAK-STAT network was further highlighted, and our data provide important new insights of the activation of this pathway in those LGL leukemias that do not carry STAT mutations.

CONFLICT OF INTEREST

SM has received honoraria and research funding from Novartis, Pfizer and Bristol-Myers Squibb and research funding from Ariad (none of these related to this project). The remaining authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We would like to thank the personnel at the Hematology Research Unit Helsinki and FIMM for their expert clinical and technical assistance. We thank also Dr Geertruy te Kronnie and Silvia Bresolin for useful discussion during development of the network-based method. This work was supported by the European Research Council (project M-IMM), Academy of Finland, the Finnish Cancer Institute, the Finnish Cancer Societies, the Signe and Ane Gyllenberg Foundation, Sigrid Juselius Foundation, Instrumentarium Science Foundation, State funding for university-level health research in Finland, Swedish Cultural Foundation, Blood Disease Foundation, the Finnish Cultural Foundation. SB received funding from Fondazione Cassa di Risparmio di Padova e Rovigo Progetti di Eccellenza 2011/2012 (<http://www.fondazioneclariparo.net/english-version/>), Ministero dell'Istruzione, dell'Università e della Ricerca (<http://www.istruzione.it/>), PRIN 2010/115 (2010NYKNS7_002) and FIRB 2011 (RBAP11CZLK) and from 'Special Program Molecular Clinical Oncology 5 × 1000' to Associazione Italiana per la Ricerca sul Cancro (<http://www.airc.it/english/>) and from the University of Padova. MH received funding from the German Research Foundation (FOR1961; HE3552/4-2).

AUTHOR CONTRIBUTIONS

AC, EIA, SM and StB designed the study, coordinated the project, analyzed the data and wrote the paper. AC, AB and StB designed and performed the bioinformatics analysis. EIA, VRG and SaB validated mutations. SM and EA provided sequence data. MC, JM and MH provided patient samples and clinical data. All authors read and approved the final manuscript.

A Coppe^{1,5}, El Andersson^{2,5}, A Binatti¹, VR Gasparini^{1,2}, S Bortoluzzi^{1,2}, M Clemente³, M Herling⁴, J Maciejewski³, S Mustjoki^{2,6} and S Bortoluzzi^{1,6}

¹Department of Molecular Medicine, University of Padova, Padova, Italy;

²Hematology Research Unit Helsinki, Department of Clinical Chemistry and Hematology, University of Helsinki and Helsinki University Hospital Comprehensive Cancer Center, Helsinki, Finland;

³Department of Translational Hematology and Oncology Research, Taussig Cancer Institute, Cleveland Clinic, Cleveland, OH, USA and

⁴Laboratory of Lymphocyte Signaling and Oncoproteome, CECAD, Cologne University, Cologne, Germany
E-mail: stefania.bortoluzzi@unipd.it

⁵These authors contributed equally to this work.
⁶Co-senior authors.

REFERENCES

- Koskela HL, Eldfors S, Ellonen P, van Adrichem AJ, Kuusanmaki H, Andersson El *et al*. Somatic STAT3 mutations in large granular lymphocytic leukemia. *N Engl J Med* 2012; **366**: 1905–1913.
- Fasan A, Kern W, Grossmann V, Haferlach C, Haferlach T, Schnittger S. STAT3 mutations are highly specific for large granular lymphocytic leukemia. *Leukemia* 2013; **27**: 1598–1600.
- Jerez A, Clemente MJ, Makishima H, Koskela H, Leblanc F, Peng NgK *et al*. STAT3 mutations unify the pathogenesis of chronic lymphoproliferative disorders of NK cells and T-cell large granular lymphocyte leukemia. *Blood* 2012; **120**: 3048–3057.
- Andersson E, Kuusanmaki H, Bortoluzzi S, Lagstrom S, Parsons A, Rajala H *et al*. Activating somatic mutations outside the SH2-domain of STAT3 in LGL leukemia. *Leukemia* 2016; **30**: 1204–1208.
- Andersson E, Tanahashi T, Sekiguchi N, Gasparini VR, Bortoluzzi S, Kawakami T *et al*. High incidence of activating STAT5B mutations in CD4-positive T-cell large granular lymphocyte leukemia. *Blood* 2016; **128**: 2465–2468.
- Rajala HL, Eldfors S, Kuusanmaki H, van Adrichem AJ, Olson T, Lagstrom S *et al*. Discovery of somatic STAT5b mutations in large granular lymphocytic leukemia. *Blood* 2013; **121**: 4541–4550.
- Rajala HL, Porkka K, Maciejewski JP, Loughran Jr TP, Mustjoki S. Uncovering the pathogenesis of large granular lymphocytic leukemia—novel STAT3 and STAT5b mutations. *Ann Med* 2014; **46**: 114–122.
- Teramo A, Gattazzo C, Passeri F, Lico A, Tasca G, Cabrelle A *et al*. Intrinsic and extrinsic mechanisms contribute to maintain the JAK/STAT pathway aberrantly activated in T-type large granular lymphocyte leukemia. *Blood* 2013; **121**: 3843–3854, S1.
- Epling-Burnette PK, Liu JH, Catlett-Falcone R, Turkson J, Oshiro M, Kothapalli R *et al*. Inhibition of STAT3 signaling leads to apoptosis of leukemic large granular lymphocytes and decreased Mcl-1 expression. *J Clin Invest* 2001; **107**: 351–362.
- Zhang J, Dominguez-Sola D, Hussein S, Lee JE, Holmes AB, Bansal M *et al*. Disruption of KMT2D perturbs germinal center B cell development and promotes lymphomagenesis. *Nat Med* 2015; **21**: 1190–1198.
- Rodriguez-Caballero A, Garcia-Montero AC, Barcena P, Almeida J, Ruiz-Cabello F, Taberero MD *et al*. Expanded cells in monoclonal TCR-alpha-beta+/CD4+/NKa+/CD8-/-dim T-LGL lymphocytosis recognize hCMV antigens. *Blood* 2008; **112**: 4609–4616.
- Calura E, Pizzini S, Bisognin A, Coppe A, Sales G, Gaffo E *et al*. A data-driven network model of primary myelofibrosis: transcriptional and post-transcriptional alterations in CD34+ cells. *Blood Cancer J* 2016; **6**: e439.
- Calura E, Bisognin A, Manzoni M, Todoerti K, Taiana E, Sales G *et al*. Disentangling the microRNA regulatory milieu in multiple myeloma: integrative genomics analysis outlines mixed miRNA-TF circuits and pathway-derived networks modulated in t(4;14) patients. *Oncotarget* 2016; **7**: 2367–2378.
- Leblanc F, Zhang D, Liu X, Loughran TP. Large granular lymphocyte leukemia: from dysregulated pathways to therapeutic targets. *Future Oncol* 2012; **8**: 787–801.
- Johansson P, Bergmann A, Rahmann S, Wohlers I, Scholtysik R, Przekopowicz M *et al*. Recurrent alterations of TNFAIP3 (A20) in T-cell large granular lymphocytic leukemia. *Int J Cancer* 2016; **138**: 121–124.



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/4.0/>

© The Author(s) 2017

Supplementary Information accompanies this paper on the Leukemia website (<http://www.nature.com/leu>)



blood[®]

2016 128: 2465-2468

doi:10.1182/blood-2016-06-724856 originally published
online October 3, 2016

High incidence of activating *STAT5B* mutations in CD4-positive T-cell large granular lymphocyte leukemia

Emma I. Andersson, Takahiro Tanahashi, Nodoka Sekiguchi, Vanessa Rebecca Gasparini, Sabrina Bortoluzzi, Toru Kawakami, Kazuyuki Matsuda, Takeki Mitsui, Samuli Eldfors, Stefania Bortoluzzi, Alessandro Coppe, Andrea Binatti, Sonja Lagström, Pekka Ellonen, Noriyasu Fukushima, Sayaka Nishina, Noriko Senoo, Hitoshi Sakai, Hideyuki Nakazawa, Yok-Lam Kwong, Thomas P. Loughran, Jaroslaw P. Maciejewski, Satu Mustjoki and Fumihiko Ishida

Updated information and services can be found at:

<http://www.bloodjournal.org/content/128/20/2465.full.html>

Articles on similar topics can be found in the following Blood collections

[Lymphoid Neoplasia](#) (2531 articles)

Information about reproducing this article in parts or in its entirety may be found online at:

http://www.bloodjournal.org/site/misc/rights.xhtml#repub_requests

Information about ordering reprints may be found online at:

<http://www.bloodjournal.org/site/misc/rights.xhtml#reprints>

Information about subscriptions and ASH membership may be found online at:

<http://www.bloodjournal.org/site/subscriptions/index.xhtml>

18. Badalian-Very G, Vergilio JA, Degar BA, et al. Recurrent BRAF mutations in Langerhans cell histiocytosis. *Blood*. 2010;116(11):1919-1923.
19. Chakraborty R, Hampton OA, Shen X, et al. Mutually exclusive recurrent somatic mutations in MAP2K1 and BRAF support a central role for ERK activation in LCH pathogenesis. *Blood*. 2014;124(19):3007-3015.
20. Brown NA, Furtado LV, Betz BL, et al. High prevalence of somatic MAP2K1 mutations in BRAF V600E-negative Langerhans cell histiocytosis. *Blood*. 2014;124(10):1655-1658.
21. Oberholzer PA, Kee D, Dziunycz P, et al. RAS mutations are associated with the development of cutaneous squamous cell tumors in patients treated with RAF inhibitors. *J Clin Oncol*. 2012;30(3):316-321.
22. da Rocha Dias S, Salmonson T, van Zwieten-Boot B, et al. The European Medicines Agency review of vemurafenib (Zelboraf®) for the treatment of adult patients with BRAF V600 mutation-positive unresectable or metastatic melanoma: summary of the scientific assessment of the Committee for Medicinal Products for Human Use. *Eur J Cancer*. 2013;49(7):1654-1661.

DOI 10.1182/blood-2016-06-721993

© 2016 by The American Society of Hematology

To the editor:

High incidence of activating *STAT5B* mutations in CD4-positive T-cell large granular lymphocyte leukemia

Emma I. Andersson,^{1,*} Takahiro Tanahashi,^{2,*} Nodoka Sekiguchi,^{3,4} Vanessa Rebecca Gasparini,¹ Sabrina Bortoluzzi,¹ Toru Kawakami,³ Kazuyuki Matsuda,⁵ Takeki Mitsui,⁶ Samuli Eldfors,⁷ Stefania Bortoluzzi,⁸ Alessandro Coppe,⁸ Andrea Binatti,⁸ Sonja Lagström,⁷ Pekka Ellonen,⁷ Noriyasu Fukushima,⁹ Sayaka Nishina,³ Noriko Senoo,³ Hitoshi Sakai,⁴ Hideyuki Nakazawa,³ Yok-Lam Kwong,¹⁰ Thomas P. Loughran,¹¹ Jaroslaw P. Maciejewski,¹² Satu Mustjoki,^{1,13,†} and Fumihiko Ishida^{2,3,14,†}

¹Hematology Research Unit Helsinki, Department of Clinical Chemistry and Hematology, University of Helsinki, Helsinki, Finland; ²Department of Clinical Laboratory Investigation, Graduate School of Medicine, Shinshu University, Matsumoto, Japan; ³Division of Hematology, Department of Internal Medicine, and ⁴Department of Comprehensive Cancer Therapy, Shinshu University School of Medicine, Matsumoto, Japan; ⁵Department of Laboratory Medicine, Shinshu University Hospital, Matsumoto, Japan; ⁶Department of Medicine and Clinical Sciences, Gunma University School of Medicine, Maebashi, Japan; ⁷Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland; ⁸Computational Genomics Laboratory, Department of Molecular Medicine, University of Padova, Padova, Italy; ⁹Department of Medical Science Technology, School of Health Sciences at Fukuoka, International University of Health and Welfare, Fukuoka, Japan; ¹⁰Department of Medicine, Queen Mary Hospital, Hong Kong, China; ¹¹University of Virginia Cancer Center, University of Virginia, Charlottesville, VA; ¹²Department of Translational Hematology and Oncology Research, Taussig Cancer Institute, Cleveland Clinic, Cleveland, OH; ¹³Helsinki University Hospital Comprehensive Cancer Center, Helsinki, Finland; and ¹⁴Department of Biomedical Laboratory Sciences, Shinshu University School of Medicine, Matsumoto, Japan

Large granular lymphocyte (LGL) leukemia is a group of chronic lymphoproliferative disorders of cytotoxic T or natural killer (NK) cells frequently complicated with cytopenia and autoimmune phenomena.^{1,2} In the current World Health Organization (WHO) classification, T-LGL leukemia and chronic lymphoproliferative disorder of NK cells (CLPD-NK) are included in this category.³

Recurrent somatic mutations in the Src homology 2 (SH2) domain of the signal transducer and activator of transcription 3 (*STAT3*) gene have been found in T-LGL leukemia and CLPD-NK,^{4,5} leading to constitutive activation of *STAT3* and dysregulation of genes downstream of *STAT3*. More recently, mutations outside the SH2 domain have been discovered in T-LGL leukemia.⁶ Activating mutations in the SH2 domain of the *STAT5B* gene were also identified in 2% of LGL leukemia patients,⁷ which further underlines the importance of the JAK/STAT signaling pathway in LGL leukemia.

The majority of T-LGL leukemia cases present with a clonal expansion of the CD8⁺ LGLs. However, in a small percentage of cases, the tumor cells have a CD4⁺ phenotype.⁸⁻¹⁰ Cytomegalovirus-derived stimulation and restricted use of the T-cell receptor (TCR)-Vβ region has been associated with CD4⁺ T-LGL cases,¹¹ but this rare disease entity still remains poorly described. To further elucidate the pathogenesis of this rare subgroup of T-LGL leukemia, we explored the mutational landscape of CD4⁺ cases using exome and targeted amplicon sequencing. Patients diagnosed with T-LGL leukemia and CLPD-NK were recruited. The diagnostic criteria were based on the WHO classifications of 2008. Three patient cohorts (described in

detail in the supplemental Appendix, available on the *Blood* Web site) were included in this study.

Exome sequencing was performed on 3 CD4⁺ T-LGL leukemia patients' sorted tumor (CD4⁺ or CD4⁺CD8⁺ T cells) and control (CD4⁻) fractions. The exome was captured with Nimblegen SeqCap EZ Exome Library v2.0, and sequencing was performed with the Illumina HiSeq2000 sequencing platform. Candidate somatic mutations were identified with a bioinformatics pipeline described earlier,⁴ as well as a novel pipeline described in more detail in the supplemental Appendix. Through exome sequencing, we were able to identify novel somatic missense mutations in the transactivation domain of *STAT5B* in 2 CD4⁺ T-LGL leukemia patients. Patient 1 had a Q706L mutation at a variant allele frequency (VAF) of 45% in the CD4⁺CD8⁺ tumor fraction. Patient 2 displayed an S715F mutation (VAF, 36%) in the CD4⁺ fraction (Figure 1A). Only wild-type (WT) *STAT5B* was observed in the CD4⁻ fractions, confirming that the mutations were somatic. The third patient with CD4⁺ T-LGL leukemia did not show any mutations in *STAT5B* or *STAT3* genes, but mutations in members of the protein tyrosine phosphatase family (*PTPN14*, *PTPN23*) regulating cell proliferation and tumor suppressor *MLL2* were observed (supplemental Table 3).

To study the functional properties of the novel variants, we generated *STAT5B* expression vectors for WT, Q706L, and S715F mutations and previously described activating N642H mutation.⁷ The transcriptional activity of the mutants was studied with luciferase reporter assays with and without interferon-α stimulation, and the phosphorylation status was analyzed by western blotting. In HeLa cells,

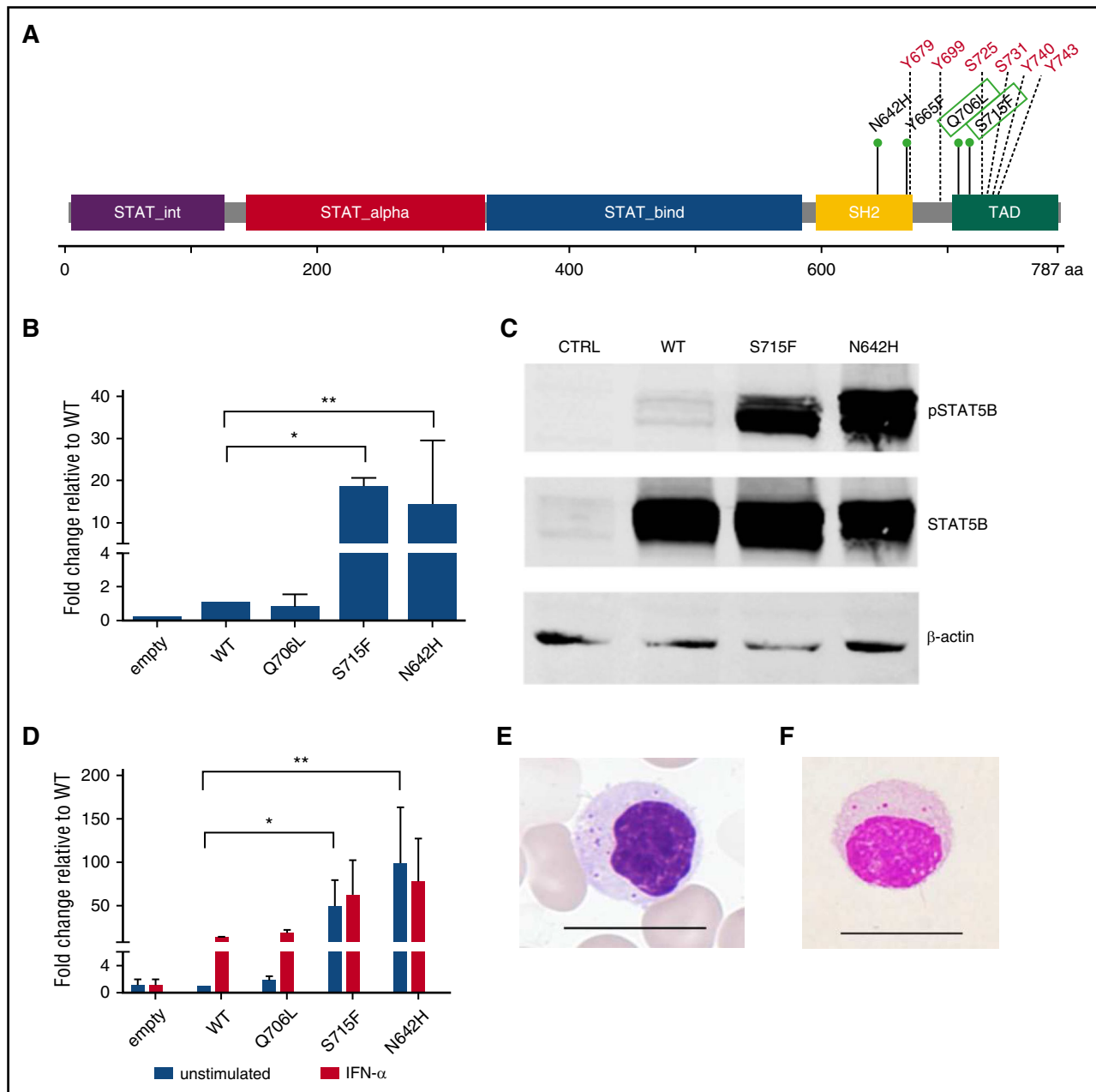


Figure 1. *STAT5B* mutation characterization. (A) Linear representation of the *STAT5B* protein structure. Previously known LGL leukemia mutations in *STAT5B* are marked in the SH2 domain, whereas the novel Q706L and S715F mutations in the transactivation domains are marked with green boxes. Multiple tyrosine and serine phosphorylation sites are marked in red. (B) *STAT5B* reporter assay results. Mutated *STAT5B* constructs (pCMV6-XL6 *STAT5B*) were generated through site-directed mutagenesis followed by transfection and expression of WT and mutated *STAT5B* (Q706L, S715F, N642H) in HeLa cells together with a *STAT5B* reporter. Dual-reporter luciferase assay was used to determine activation and phosphorylation of mutated *STAT5B*. The experiment was repeated 3 times. Columns represent mean of the fold-change activity. Error bars indicate the standard error of the mean (SEM), and the statistical significance was calculated with a 1-way analysis of variance (ANOVA; * $P < .05$, ** $P < .001$). (C) To investigate the phosphorylation status of the variants, HeLa cells transfected with the abovementioned variants were analyzed by western blot with a phospho $STAT5$ (Tyr694) specific antibody. Protein lysates of the different variants were separated on an sodium dodecyl sulfate-polyacrylamide gel electrophoresis gel and transferred to a nitrocellulose membrane. $STAT5$ protein levels of the different variants were used to normalize for the transfection efficacy. β -Actin was used as a loading control. (D) Transfected HeLa cells were stimulated with 100 ng/mL interferon- α for 6 hours. A dual-reporter luciferase assay was used to determine activation and phosphorylation of mutated *STAT5B*. The experiment was repeated 2 times. Columns represent mean of the fold-change activity. Error bars indicate the SEM, and the statistical significance was calculated with a 1-way ANOVA (* $P < .05$, ** $P < .001$). (E) Typical morphology of a representative LGL cell in a *STAT5B* mutated T-LGL patient. Scale bar, 15 μ m. (F) Morphology of lymphocyte expressing CD4, CD56, and TCR $\alpha\beta$ in a healthy individual. CD4 $^+$ CD56 $^+$ and TCR $\alpha\beta$ -type lymphocytes were sorted by the FACS method and stained with Wright-Giemsa stain. A representative cell is shown. Scale bar, 15 μ m.

the mutated *STAT5B* S715F construct significantly enhanced the transcription of the cotransfected *STAT5* reporter (18-fold compared with WT *STAT5B*) similarly to the N642H mutation (Figure 1B), whereas the Q706L mutation activation was equal to WT. In the western blot analysis, S715F and N642H mutations showed significantly

increased phosphorylation compared with WT *STAT5B* (Figure 1C), whereas no increased phosphorylation was observed with the Q706L mutation. The location of the novel S715F mutation in a serine phosphorylation site is likely to increase the phosphorylation of *STAT5B*. Stimulation with interferon- α revealed that the Q706L mutation behaved

Table 1. Clinical features of CD4⁺ T-LGL leukemia patients

	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5	Patient 6	Patient 7	Patient 8	Patient 9	Patient 10	Patient 11
STAT5b mutation (% VAF)	Q706L (45)	S715F (36)	N642H (25)	N642H (46)	Y665F (31)	N642H (27)	None	None	None	None	None
Vbeta expansion (CD4 ⁺ population)	Vb.13.1: 98%	Vb.8: 86%	NA	NA	NA	NA	Vb.13.1: 78%	NA	NA	NA	NA
Age (years)	61	70	74	79	82	66	80	67	58	70	79
Sex	M	F	M	F	M	M	F	F	F	F	F
WBC count (10 ⁹ /L)	8.5	10.2	9.0	8.7	13.9	9.4	8.7	6.8	5.5	6.1	8.2
Neutrophil (%)*	40	16	12	5	51	32	33	35	32	42	26
LGL (%)*	52	72	71	91	39	63	57	44	54	55	69
Hb (g/L)	134	124	119	126	155	141	135	142	73	135	120
Platelets (10 ⁹ /L)	399	204	144	186	245	265	241	143	200	229	156
Other neoplasias	None	None	None	None	None	None	None	None	None	None	None
Other diseases	Diabetes	None	None	Gastrointestinal hemorrhage	None	Lung cancer	Osteoarthritis, hypothyroidism	None	None	None	None
Observation period	5 years	7 years	14 years	6 months	3 years	2 years	3 years	12 years	6 years	12 years	15 months
Outcome	Alive	Alive	Death	Alive	Alive	Alive	Alive	Alive	Alive	Alive	Alive

F, female; Hb, hemoglobin; LGL, large granular lymphocyte; M, male; VAF, variant allele frequency; WBC, white blood cell.

*Neutrophil and LGL percentage from whole white blood cell population. From patients 1, 2, and 3, germline DNA was available for sequencing to confirm the somatic nature of the STAT5b mutations.

as the WT, whereas stimulation was not able to further increase the transcriptional activity of the S715F and N642H mutants (Figure 1D).

To elucidate whether *STAT5B* mutations are more prevalent in CD4⁺ T-LGL leukemia cases, deep amplicon sequencing was used for screening of the SH2 and transactivation domains of *STAT5B* in CD4⁺ (n = 8), *STAT3*-mutated CD8⁺ (n = 37) and nonmutated CD8⁺ (n = 58) T-LGL leukemia patients. Targeted *STAT5B* amplicon sequencing covering exons 14 to 19 was done with an in-house-developed deep amplicon sequencing panel using the Illumina Miseq platform.⁷ The data were analyzed with a bioinformatics pipeline described previously.¹² A variant was called when the variant base frequency was 0.5% of all reads covering a given a position. Additionally, the same regions were screened with Sanger sequencing in Japanese and Chinese LGL leukemia cohorts consisting of CD8⁺ and CLPD-NK cases (n = 57). None of the patients with CD8⁺ T-LGL leukemia or CLPD-NK had *STAT5B* mutations. In contrast, 4 of 8 CD4⁺ T-LGL leukemia cases had *STAT5B* mutations. Of the 4 patients with *STAT5B* mutations, 3 possessed the earlier described N642H mutation and 1 the Y665F mutation. Sanger sequencing-negative patients and healthy controls (n = 50) were also screened with allele-specific PCR for N642H and Y665F mutations, but no additional mutations were found. Altogether, the *STAT5B* mutation frequency in CD4⁺ T-LGL leukemia patients in our cohort was 55% (6 of 11 patients). This is significantly higher than in the previous study (2%) of 211 CD8⁺ T- and NK-cell LGL leukemia cases where *STAT5B* SH2 domain mutations were initially discovered.⁷ Most of the *STAT5B* mutations found in CD4⁺ T-LGL leukemia have also been seen in various T-cell neoplasms, including $\gamma\delta$ hepatosplenic T-cell lymphoma,¹³ T-cell acute lymphoblastic leukemia,^{14,15} T-cell prolymphocytic leukemia,¹⁶ type II enteropathy-associated T-cell lymphoma,¹⁷ and extranodal NK/T-cell lymphoma,¹⁸ suggesting that these are shared with other T-cell malignancies. The analyses of *STAT5* target genes with chromatin immunoprecipitation sequencing have shown that *STAT5B* is a key factor in T-cell development, binding to molecules such as *DOCK8*, *SNX9*, *FOXP3*, and *IL2RA*.¹⁹ Together these results suggest that the *STAT5B* pathway plays a central role in the development of T-cell neoplasms.

In contrast to other more aggressive T-cell malignancies with *STAT5B* mutations, the disease course in our CD4⁺ T-LGL leukemia

cohort was indolent, and none of the patients with *STAT5B* mutations needed therapy during the observation time (median follow-up, 4 years). Rheumatoid arthritis (RA) is commonly associated with CD8⁺ T-LGL leukemia, and especially patients with multiple *STAT3* mutations more often have RA.¹² In our cohort, none of the 11 cases with CD4⁺ T-LGL leukemia suffered from RA. Two patients showed neutropenia and 1 patient had anemia (Table 1).

All *STAT5B* mutated CD4⁺ T-LGL cases possessed a TCR $\alpha\beta$ T-cell phenotype with CD16⁺CD56⁺ and CD57⁺ (Figure 1E). Two cases were CD8⁻, 2 were weakly positive for CD8, and 2 were clearly positive for CD8 (supplemental Table 4). This is in accordance with the earlier reports⁸⁻¹⁰ of monoclonal CD4⁺ T-LGL cells, which have shown expression of TCR $\alpha\beta$, variable levels of CD8, and a typical cytotoxic (granzyme B⁺, CD56⁺, CD57⁺, CD11b^{+/+}) and activated/memory T-cell (CD2⁺bright, CD7^{-/+dim}, CD11a⁺bright, CD28⁻, CD62L⁻HLA-DR⁺) phenotype. Interestingly, all 6 patients with *STAT5B* mutations had large monoclonal TCR-V β expansions where the mutations were located, whereas significant proportions of *STAT3* mutations in CD8⁺ T-LGL leukemia and CLPD-NK are detected in small subclones.

Because the CD4⁺CD56⁺TCR $\alpha\beta$ ⁺ immunophenotypes recognized on *STAT5B*-mutated T-LGL leukemia cells have been poorly defined, we also investigated whether normal lymphocytes with similar phenotypic features exist in peripheral blood of healthy subjects. Among 27 healthy controls, the median percentage of CD4⁺CD56⁺TCR $\alpha\beta$ ⁺ T cells in lymphocytes was 0.2, and it varied from less than 0.02% to 6.5% (supplemental Figure 2). Fluorescence-activated cell sorter (FACS)-sorted CD4⁺CD56⁺TCR $\alpha\beta$ cells possessed LGL morphology with cytoplasmic azurophilic granules (Figure 1F; N = 3). Thus, phenotypically similar cells as observed in CD4⁺ T-LGL leukemia cases can also be observed in healthy individuals in small quantities. However, deep amplicon sequencing of sorted CD4⁺CD56⁺ cells from 5 healthy subjects revealed no mutations in the SH2 or transactivation domains of *STAT5B*.

In conclusion, activating *STAT5B* mutations can be found in the majority (55%) of CD4⁺ T-LGL leukemia cases, whereas among patients with CD8⁺ T-LGL leukemia or CLPD-NK, these are very rare. *STAT5B* mutations can be considered as a novel diagnostic marker for this specific disease subtype.

*E.I.A. and T.T. contributed equally to this work.

†S.M. and F.I. are joint senior authors.

The online version of this article contains a data supplement.

Acknowledgments: Personnel at the Hematology Research Unit Helsinki and Institute for Molecular Medicine Finland are acknowledged for their expert clinical and technical assistance.

This work was supported by the European Research Council (M-IMM), Academy of Finland, the Finnish Cancer Institute, the Finnish Cancer Societies, the Signe and Ane Gyllenberg Foundation, Sigrid Juselius Foundation, Instrumentarium Science Foundation, Biocentrum Helsinki, state funding for university-level health research in Finland, Swedish Cultural Foundation, Blood Disease Foundation, the Finnish Cultural Foundation, and Grant-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology of Japan. Work of J.P.M. was supported in part by National Institutes of Health National Cancer Institute grants 2K24HL077522, R01 CA127264A, and R01AI085578. Work of T.P.L. was supported by National Cancer Institute grants R01 CA098472 and R01 CA178393. Work of Stefania Bortoluzzi was supported by Cassa di Risparmio di Padova e Rovigo Foundation (Excellence projects 2011/2012), Italian Ministry of Education, Universities and Research (PRIN 2010/11 2010NYKNS7_002), and University of Padova. A.B. is recipient of a fellowship of the Program in Biosciences of the University of Padova.

Contribution: E.I.A., T.T., S.M., and F.I. designed the study, coordinated the project, analyzed the data, and wrote the paper; E.I.A., T.T., T.K., S.L., K.M., and P.E. performed sequence analysis and validated mutations; E.I.A., V.R.G., and Sabrina Bortoluzzi designed and performed the functional experiments; S.E., Stefania Bortoluzzi, A.C., and A.B. designed and performed the bioinformatics analysis; N.S., T.M., N.F., S.N., N.S., H.S., H.N., Y-L.K., T.P.L., and J.P.M. provided patient samples; and all authors read and approved the final manuscript.

Conflict-of-interest disclosure: S.M. has received honoraria and research funding from Novartis, Pfizer, and Bristol-Myers Squibb and research funding from Ariad. F.I. has received research funding from Bristol-Myers Squibb. The remaining authors declare no competing financial interests.

ORCID profiles: S.M., 0000-0002-0816-8241.

Correspondence: Satu Mustjoki, Hematology Research Unit Helsinki, University of Helsinki, Haartmaninkatu 8,00290 Helsinki, Finland; e-mail: satu.mustjoki@helsinki.fi; or Fumihiko Ishida, Department of Biomedical Laboratory Sciences, Shinshu University School of Medicine, 3-1-1, Matsumoto, Nagano 3908621, Japan; e-mail: fumishi@shinshu-u.ac.jp.

References

- Loughran TP Jr. Clonal diseases of large granular lymphocytes. *Blood*. 1993; 82(1):1-14.
- Zhang D, Loughran TP Jr. Large granular lymphocytic leukemia: molecular pathogenesis, clinical manifestations, and treatment. *Hematol Am Soc Hematol Educ Program*. 2012;2012:652-659.
- Sabattini E, Bacci F, Sagromoso C, Pileri SA. WHO classification of tumours of haematopoietic and lymphoid tissues in 2008: an overview. *Pathologica*. 2010; 102(3):83-87.
- Koskela HL, Eldfors S, Eilonen P, et al. Somatic STAT3 mutations in large granular lymphocytic leukemia. *N Engl J Med*. 2012;366(20):1905-1913.
- Jerez A, Clemente MJ, Makishima H, et al. STAT3 mutations unify the pathogenesis of chronic lymphoproliferative disorders of NK cells and T-cell large granular lymphocyte leukemia. *Blood*. 2012;120(15):3048-3057.
- Andersson EI, Rajala HL, Eldfors S, et al. Novel somatic mutations in large granular lymphocytic leukemia affecting the STAT-pathway and T-cell activation. *Blood Cancer J*. 2013;3:e168.
- Rajala HL, Eldfors S, Kuusanmäki H, et al. Discovery of somatic STAT5b mutations in large granular lymphocytic leukemia. *Blood*. 2013;121(22):4541-4550.
- Lima M, Almeida J, Dos Anjos Teixeira M, et al. TCRalpha+beta+/CD4+ large granular lymphocytosis: a new clonal T-cell lymphoproliferative disorder. *Am J Pathol*. 2003;163(2):763-771.
- Rodríguez-Caballero A, García-Montero AC, Bárcena P, et al. Expanded cells in monoclonal TCR-alpha+beta+/CD4+/NKa+/CD8-/- dim T-LGL lymphocytosis recognize hCMV antigens. *Blood*. 2008;112(12):4609-4616.
- Olteanu H, Karandikar NJ, Eshoa C, Kroft SH. Laboratory findings in CD4(+) large granular lymphocytoses. *Int J Lab Hematol*. 2010;32(1 Pt 1):e9-e16.
- Garrido P, Ruiz-Cabello F, Bárcena P, et al. Monoclonal TCR-Vbeta13.1+/CD4+/NKa+/CD8-/- dim T-LGL lymphocytosis: evidence for an antigen-driven chronic T-cell stimulation origin. *Blood*. 2007;109(11):4890-4898.
- Rajala HL, Olson T, Clemente MJ, et al. The analysis of clonal diversity and therapy responses using STAT3 mutations as a molecular marker in large granular lymphocytic leukemia. *Haematologica*. 2015;100(1):91-99.
- Nicolae A, Xi L, Pittaluga S, et al. Frequent STAT5B mutations in $\gamma\delta$ hepatosplenic T-cell lymphomas. *Leukemia*. 2014;28(11):2244-2248.
- Kontro M, Kuusanmäki H, Eldfors S, et al. Novel activating STAT5B mutations as putative drivers of T-cell acute lymphoblastic leukemia. *Leukemia*. 2014; 28(8):1738-1742.
- Bandapalli OR, Schuessle S, Kunz JB, et al. The activating STAT5B N642H mutation is a common abnormality in pediatric T-cell acute lymphoblastic leukemia and confers a higher risk of relapse. *Haematologica*. 2014;99(10):e188-e192.
- Kiel MJ, Velusamy T, Rolland D, et al. Integrated genomic sequencing reveals mutational landscape of T-cell prolymphocytic leukemia. *Blood*. 2014;124(9):1460-1472.
- Nairismägi ML, Tan J, Lim JQ, et al. JAK-STAT and G-protein-coupled receptor signaling pathways are frequently altered in epitheliotropic intestinal T-cell lymphoma. *Leukemia*. 2016;30(6):1311-1319.
- Küçük C, Jiang B, Hu X, et al. Activating mutations of STAT5B and STAT3 in lymphomas derived from $\gamma\delta$ -T or NK cells. *Nat Commun*. 2015;6:6025.
- Kanai T, Seki S, Jenks JA, et al. Identification of STAT5A and STAT5B target genes in human T cells. *PLoS One*. 2014;9(1):e86790.

DOI 10.1182/blood-2016-06-724856

© 2016 by The American Society of Hematology

To the editor:

Cardiac involvement in Erdheim-Chester disease: an MRI study

Davide Gianfreda,¹ Alessandro A. Palumbo,² Enrica Rossi,^{2,3} Lorenzo Buttarelli,² Gaia Manari,¹ Chiara Martini,² Massimo De Filippo,² and Augusto Vaglio¹

¹Nephrology Unit, ²Radiology Unit, Parma University Hospital, Parma, Italy; and ³Department of Imaging, Bambin Gesù Children's Hospital, Roma, Italy

Erdheim-Chester disease (ECD) is a rare non-Langerhans cell histiocytosis (<1000 cases reported in the literature), characterized by tissue infiltration by CD68⁺ CD1a⁻ "foamy" histiocytes. ECD commonly causes long bone osteosclerosis, retroperitoneal (periaortic and perirenal) fibrosis, central nervous system (CNS) lesions, but also involves the lung, the skin, and various endocrine axes.¹ Cardiovascular

manifestations are also common (~40% of the cases) and include infiltration of the myocardium (eg, pseudotumoral atrial masses), the pericardium (eg, pericarditis sometimes complicated by tamponade), and the aorta, with the typical aspect of "coated aorta."^{2,3} Patients with ECD with cardiovascular involvement are reported to have a poorer prognosis^{1,4,5} and are therefore usually treated aggressively, but