PhD degree in Systems Medicine (curriculum in Molecular Oncology)

European School of Molecular Medicine (SEMM),

University of Milan and University of Naples “Federico II”

Comprehensive whole genome sequencing unravels the complex genomic landscape of Neuroblastoma

Settore disciplinare: MED/03 Genetica Medica

Candidate: D’Alterio Giuseppe

Tutor: Prof. Mario Capasso

Dipartimento di Medicina Molecolare
e Biotecnologie Mediche, university
of Naples Federico II;
CEINGE Biotecnologie Avanzate s.r.l

PhD Coordinator: Prof. Saverio Minucci

Anno accademico 2022-2023

Table of contents

List of Abbreviations ***3***

Figures Index ***5***

Abstract ***7***

1. Introduction ***8***

1.1. Neuroblastoma **8**

1.1.1. *Biology of Neuroblastoma* 9

1.1.2. Clinical classification of Neuroblastoma 11

1.1.3. Sporadic and Familiar forms of NBL 13

1.1.4. Genetics of Neuroblastoma 13

1.1.4.1. Genetic predisposition to Neuroblastoma 13

1.1.4.2. Genetics and Genomics of Neuroblastoma 16

1.2. NBL in the post-genomic era **18**

1.2.1. Genome Wide Association Studies and the genetic predisposition to Neuroblastoma 18

1.2.2. The Next Generation Sequencing and Neuroblastoma 20

1.3. *Germline predisposition to tumor phenotypes* **28**

2. Aim ***33***

3. Methods ***34***

3.1. NLB Samples and Pipelines **34**

3.1.1. Data description and publicly available files 34

3.2. Variant calling and processing pipelines **36**

3.2.1. *Germline SNVs calling* 36

3.2.2. Somatic SNVs calling 38

3.2.3. Somatic CNAs calling 39

3.2.4. Somatic SVs calling 40

3.3. RNA-seq processing **41**

3.4. Genomic profiling and downstream analyses **42**

3.4.1. Tumor mutational burden and somatic SNVs prioritization 42

3.4.2. Focal and numerical CNAs 42

3.4.3. SV profiling 43

3.4.4. *Scores of genome instability* 45

3.4.5. SBS Mutational signatures 47

3.4.6. Differential gene-expression analysis 47

3.5. Germline correlation to somatic SV phenotype **48**

3.6. Statistical analysis and graphs **49**

4. Results ***50***

4.1. NBL patients characterization **50**

4.2. Genomic profiling of NBL samples **53**

4.2.1. Somatic SNVs 53

4.2.2. Copy Number profiling 61

4.2.3. Genomic Rearrangement profiling 69

4.3. *NBL patients showed different phenotypes based on the presence of an SV* .. **76**

4.3.1. SV group showed an overall increased degree of genetic instability compared to no-SV group 77
4.3.2. SV group was enriched in mutations of SBS18 79
4.3.3. SV group over-expressed DNA-repair related genes and under-expressed neuronal function and differentiation and synapse plasticity genes 83

4.4. Double-strand break repair genes were enriched in germline PVs in SV group **86**

5. Discussion ***91***

6. Conclusions ***96***

7. References ***97***

List of Abbreviations

8-oxoG: 8-oxo-guanosine

ACMG: American College of Medical Genetics

ALT: Alternative Lengthening of Telomeres

ASI: Allele Specific Imbalance

BAM: Binary Alignment Mapping

CGC: Cancer Gene Census

CNA: Copy Number Alteration

COSMIC: Catalogue of Somatic Mutations in Cancer

DGE: Differential Gene Expression

DM: Double Minute

DR: Discordant Reads

DSB: Double Strand Breaks

EGA: European Genome-phenome Archive

EMT: Epithelial-to-Mesenchymal Transition

FDR: False Discovery Rate

FPKM: Fragment Per Kilobase per Million

GWAS: Genome Wide Association Studies

HR: Homologous Recombination

INRG: International Neuroblastoma Risk Group

INSS: International Neuroblastoma Staging System

LoF: Loss of Function

LOH: Loss Of Heterozygosity

LST: Large-State Transitions

MAF: Minor Allele Frequency

MHC-I: class I Mayor Histocompatibility Complex

MQ: Mapping Quality

NBL: Neuroblastoma

NGS: Next Generation Sequencing

OR: Odds Ratio

ORA: Over Representation Analysis

OS: Overall Survival

P/LP: Pathogenic or Likely Pathogenic

PCA: Principal Component Analysis

PV: Pathogenic Variant

QC: Quality Control

QD: Quality by Depth

RNA-seq: RNA-sequencing

ROS: Reactive Oxygen Species

RTK: Receptor Tyrosine-Kinase

SBS: Single Base Substitutions

SNP: Single Nucleotide Polymorphisms

SNV: Small Nucleotide Variant

SV: Structural Variant

TARGET: Therapeutically Applicable Research to Generate Effective Treatments

TF: Transcription Factor

TMB: Tumor Mutational Burden

TMM: Telomere Maintenance Mechanism

TSG: Tumor-Suppressor Gene

TSV: Tab Separated Values

VAF: Variant Allele Frequency

VCF: Variant Calling Format

WES: Whole-Exome Sequencing

WGS: Whole-Genome Sequencing

Figures Index

Figure 1. Incidence of NBL in 15 countries of the 6 continents. 8
Figure 2. Pathogenicity mechanism of NBL development. 10
Figure 3. Frequency and penetrance of NBL-predisposing genes. 15
Figure 4. Segmental and numerical CNAs are associated to a decreased and an increased OS probability. 17
Figure 5. Manhattan plot showing the results of the first GWAS of NBL. 19
Figure 6. A first glance to the genomic landscape of NBL. 21
Figure 7. SBS18 causes driver mutations in NBL. 24
Figure 8. NBL shows 2 defined cell identities plus a third mixed phenotype. 26
Figure 9. Schematic depiction of somatic LOH for ASI. 29
Figure 10. Examples of germline-to-somatic correlations. 31
Figure 11. Germline SNVs calling and processing. 37
Figure 12. somatic SNVs calling and processing 38
*Fig*ure 13. Somatic CNA calling. 39
Figure 14. Somatic SVs calling. 40
Figure 15. RNA-seq pipeline. 41
Figure 16. SVs shared between EGA and TARGET tumors. 44
Figure 17. Depiction of genetic instability scores computation. 46
Figure 18. Workflow of pathway enrichment analysis of genes with germline P/PL SNVs. 48
Figure 19. Co-occurrence among clinical markers in the two NBL cohorts. 51
Figure 20. Inferred ethnicity of NBL samples. 52
Figure 21. Distribution of TMB across NBL clinical markers. 54
Figure 22. OS and EFS probability across median-divided TMB groups in TARGET samples. 55
Figure 23. OS and EFS probability across median-divided TMB groups in EGA samples. 56
Figure 24. Somatic SNVs in genes annotated as TSGs, oncogenes or both in Cosmic CGC v98 genes in NBL samples. 58
Figure 25. Frequency of genes affected by prioritized somatic SNVs shared between TARGET and EGA. 59
Figure 26. SKI proto-oncogene expression is increased in low to intermediate risk group. 60
Figure 27. Copy number profiling of TARGET and EGA NBL samples 62

Figure 28. Numerical CNAs are associated with good prognosis, conversely to focal CNAs. 63
Figure 29. Regions of gain/amplification and loss across the genome in the two datasets 64
Figure 30. Figure 30. Landscape of focal and numerical CNAs of NBL samples. 65
Figure 31. Correlation between focal and numerical CNAs and risk groups. 66
Figure 32. Frequency of low to intermediate-associated numerical CNAs across risk groups. 67
Figure 33. Aneuploidy of at least one of chromosome 1, 2, 3, 4, 7, 11, 12, 17 and 19 predicts survival in samples from EGA and TARGET datasets. 68
Figure 34. Distribution of SVs across NBL datasets and risk groups. 70
Figure 35. The presence of at least an SV predicted less overall survival in a multivariate Cox proportioned hazard regression model in TARGET dataset. 71
Figure 36.The presence of at least an SV predicted less overall survival in a univariate model in EGA dataset. 72
Figure 37. TERT rearrangements in NBL samples. 73
Figure 38. Recurrent SVs in NBL. 74
Figure 39. Genes hit by recurrent SVs. 75
Figure 40. Genome instability is increased in SV group. 78
Figure 41. Relative activity of SBS signature in SV and no-SV group. 80
Figure 42. Absolute activity of SBS1, 5, 18 and 40 in SV and no-SV groups. 81
Figure 43. The enrichment of samples with SBS18 activity remained significative in a multivariate analysis using mitochondrial gene expression, MYCN status and 17q gain. . 82
Figure 44. ORA results on SV group over and under-expressed genes. 84
Figure 45. DEGs remained under and over represented in SV group when considering only the low to intermediate risk tumors. 85
Figure 46. Mean number of germline PVs per SV group 87
Figure 47. SV group was enriched in PVs in genes of DNA IR-damage and cellular response via ATR. 87
Figure 48. The expected and observed distribution of germline PVs in WP4016 genes was comparable across datasets and ethnicity. 90

Abstract

Neuroblastoma (NBL) is a pediatric malignancy characterized by a broad spectrum of clinical outcomes, where the 5-year survival probability can shrink from 95% to 50% between the low and high-risk tumors, respectively. This heterogeneity reflects profound genetic and phenotypical differences among NBL tumors. In the last decades, thanks to the introduction of GWAS and NGS, several step forwards have been made to unravel the complexity of the genome of this tumor, providing – in many cases – a direct link between genomic alterations and clinical features, such as the onset of metastases, the probability of relapse or the patients' survival. Nonetheless, the current knowledge does not exhaustively explain the clinical heterogeneity, and a comprehensive analysis of NBL mutational landscape is still lacking.

In this dissertation we used publicly available Whole Genome Sequencing (WGS) data from two NBL databases – TARGET (N = 136) and EGA (N = 180) – to provide a full compendium of NBL genomic alterations (including somatic point mutations and structural variants), with a main focus on genomic rearrangements (translocations and inversions) which remains to date poorly investigated. Furthermore, we assessed the contribution of germline Small Nucleotide Variants (SNVs) to the genomic instability of NBL, a well-established marker of poor prognosis.

We found unreported point mutations in 3 cancer-related genes, *ESR1*, *MYH9* and *SKI*, the latter under-expressed in high-risk tumors. We report an increased survival probability in samples with specific numerical Copy Number Alterations (CNAs) (whole gain of chromosomes 1, 2, 7, 12 and 17 and whole loss of chromosomes 3, 4, 11 and 19). Our analysis of genomic rearrangements revealed 5 novel and recurrent ($\geq$ 3% of samples in both datasets) translocations (t(17q-19p), t(14q-17q), t(4p-17q), t(2p-3p), t(1p-2p)) enriched in high-risk patients, whose breakpoints affected genes related to synapse plasticity and neuronal differentiation. Finally, we observed that tumors carrying at least a genomic rearrangement also showed features of genetic instability, specific mutational signatures and a defined gene-expression pattern. In these patients we observed an enrichment of pathogenic or likely pathogenic (P/LP) germline SNVs in homologous-recombination pathway, whose deficiency in tumor is causally linked to genomic instability.

In conclusion, the results of this dissertation may improve the clinical stratification of NBL, help the development of novel personalized therapies and finally increase the knowledge about the genetic predisposition of this tumor.

1. Introduction

1.1. Neuroblastoma

Neuroblastoma (NBL) is a pediatric malignancy and represents the most common solid extracranial tumor diagnosed in children under 5 years of age[1]. Its incidence depends both on geographical area and age, with a global estimated incidence of 4.1-15.8 and of 0.4-1.0 per 1,000,000 people in children (0-14 years old) and adolescents (15–19 years old), respectively[2] (**Figure 1**).

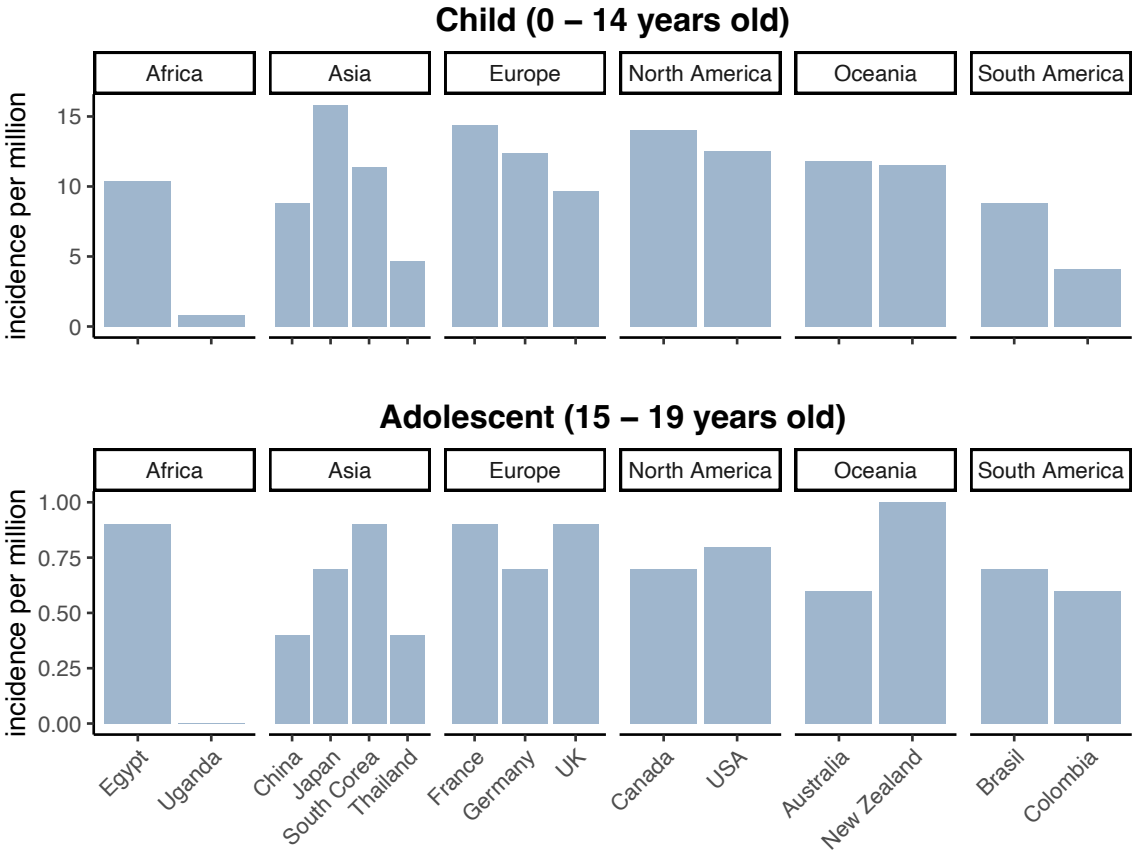


Figure 1. Incidence of NBL in 15 countries of the 6 continents.

Data for bar plot was retrieved by Okawa et al, 2022[2].

1.1.1. *Biology of Neuroblastoma*

NBL is the most representative and clinically relevant of a group of pediatric which originates from the neural crest cells during the early stages of fetal development, collectively known as peripheral neuroblastic tumors[3,4]. Neural crest cells are a group of multipotent stem cells located posteriorly to the neural tube that give rise to specialized cell belonging to sympathetic nervous system, including peripheral neurons, enteric neurons, glia, melanocytes, Schwann cells, cells of the craniofacial skeleton and adrenal medulla[5]. Although NBL can develop from all the precursor cells of sympathetic nervous system, it usually arises in chromaffin cells of adrenal gland[6].

A mechanism of onset of NBL has been proposed by Marshal and colleagues: under the effect of bone morphogenetic proteins and thanks to the transcription activity of *MYCN*, neural crest cells migrate forward in primary sympathetic ganglia, whereby differentiate in cells of the sympathetic ganglia or chromaffin cells. Genetic alteration of key NBL-associated genes – such as *MYCN*, whose role as driver gene will be discussed in this dissertation – initiate the tumor development, which usually is finally accomplished by further acquired genetic lesions and eventually with tumor progression, invasion and metastasis[7] (Figure 2).

Beyond the schematic depiction provided above, is becoming clearer and clearer that NBL is a malignancy characterized by an elevated extent of intra- and inter-tumor heterogeneity, and that the mechanism of onset, development, metastasizing and relapse differ from tumor to tumor[8]. It is well known that NBL of different patients can show profound differences in mutational landscape, gene expression and mutational patterns[9–11], which – as we will see in the course of this dissertation – are closely linked to the clinical outcome, the survival rate and the relapse probability. At the same time, within the same tumor it is possible to detect a wide spectrum of cell populations, which have been proven to be able to switch their phenotype thus affecting the efficacy of therapies[12].

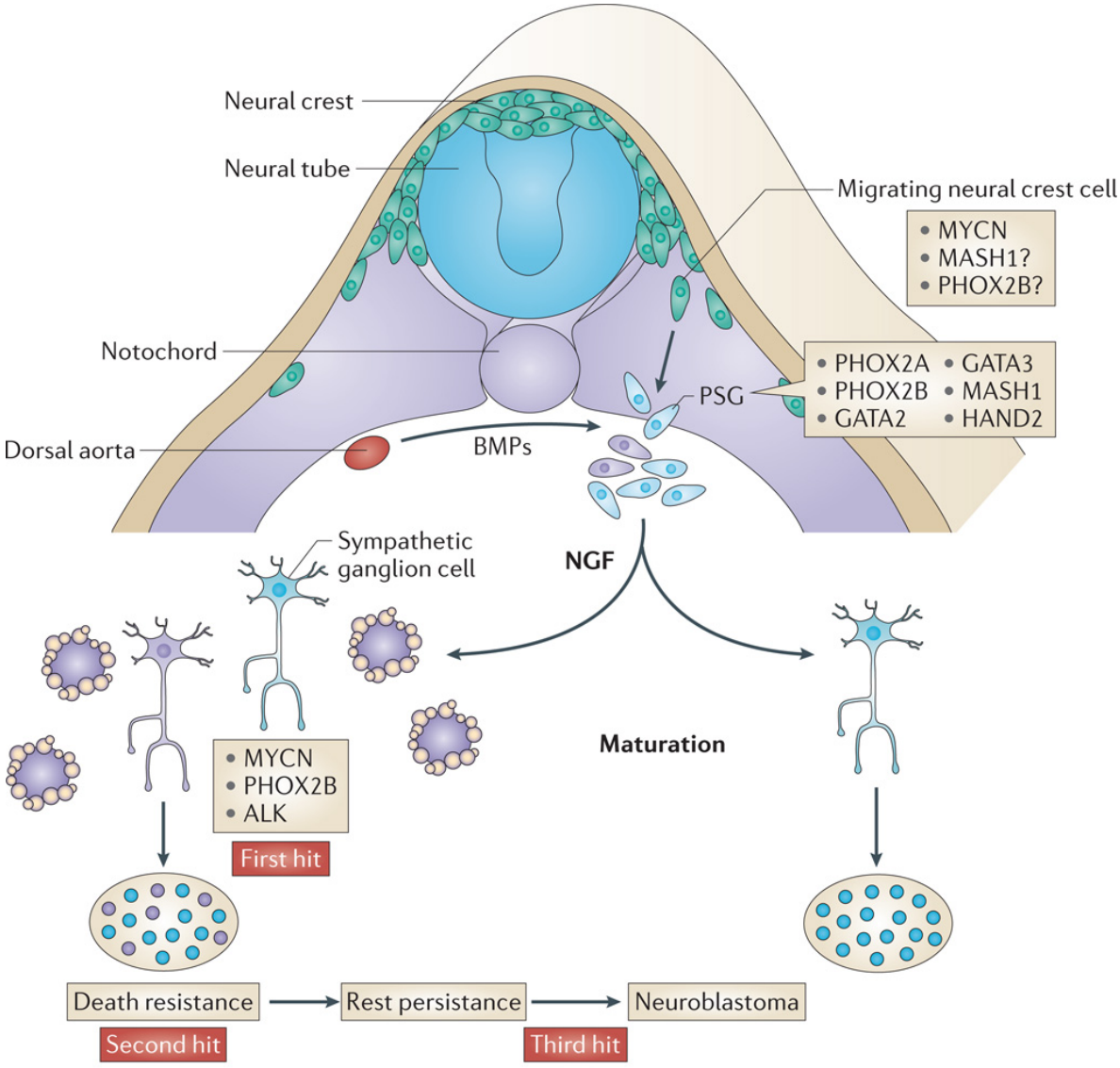


Figure 2. Pathogenicity mechanism of NBL development.

NBL derives from neural crest cells, a group of embryonal stem cells which give rise to several cell of peripheral nervous system lineage. These cells are located posteriorly to the neural tube and migrate forwards under the influence of bone morphogenetic proteins (BMPs), which promote the transcriptional activity of several transcription factors for a first differentiation in precursor cells resident in precursor sympathetic ganglia (PSG). Upon stimulus by nerve growth factor (NGF), they can maturate in mature sympathetic ganglion cells or in chromaffin cells of adrenal gland. Genomic lesion in NBL-associated genes, such as *MYCN*, *PHOX2B* or *ALK* lead to tumor initiation, which progresses thanks to the occurrence other driving mutations. *Image adapted by Marshal et al., 2014*[7].

1.1.2. Clinical classification of Neuroblastoma

As mentioned in the previous paragraph, the biological intermixture of NBL manifests itself in a high degree of clinical heterogeneity. Indeed, the 5-years Overall Survival (OS) probability overcomes the 95% in low-risk patients, but it shrinks to less than 50% for high-risk subtypes[13]. Several factors can influence the clinical outcome of NBL, such as the age at diagnosis (with children older than 18 moths having a poorer prognosis[14]) and the *MYCN* oncogene status (amplification of *MYCN* is associated to high-risk subtypes[15]). In the past decades several efforts have been made to provide a clinical classification of NBL, with a view to differentially manage and treat patients based on their clinical profiles. One of the first classification criteria of NBL was the International Neuroblastoma Staging System (INSS), which is based on the localization, resectability and dissemination of the tumor[16] (Table 1). More than a decade later was introduced the International Neuroblastoma Risk Group (INRG) classification system. This system provides a pre-treatment risk group stratification by the integration of clinical features of NBL, such as stage, age at diagnosis, *MYCN* status and grade of differentiation[14] (Table 2).

INSS Stage	Description
1	Localized tumor, grossly resected, no lymph node involvement
2A	Unilateral tumor, incomplete gross excision, negative lymph nodes
2B	Unilateral tumor with positive ipsilateral lymph nodes
3	Tumor infiltrating across midline or unilateral tumor with contralateral lymph nodes or midline tumor with bilateral lymph nodes
4	Distant metastatic disease
4S	Localized primary tumor as defined by stage 1 or 2 in patient under 12 months with dissemination limited to the liver, skin, and/or bone marrow (<10% involvement)

Table1. International Neuroblastoma Staging System.
Adapted from Sokol et al., 2019[17].

INRG Stage	Age (months)	Grade of tumor differentiation	MYCN status	11q gain	Ploidy	Pretreatment Risk group
L1			non-amplified			very low
			amplified			high
L2	< 18		non-amplified	No		low
				Yes		Intermediate
	≥ 18	Differentiating	non-amplified	No		Low
				Yes		
		Poorly differentiated or undifferentiated	non-amplified			Intermediate
			amplified			high
M	<18		non-amplified		Hyperdiploid	low
	<12		non-amplified		Diploid	intermediate
	12-18		non-amplified		Diploid	intermediate
	<18		amplified			high
	≥ 18					high
MS	<18		non-amplified	No		very low
				Yes		high
			amplified			high

Table 2. INRG criteria of risk stratification. This table shows the clinical, biological and genetic feature used to classify NBL patients according to the INRG. *Adapted from Luksch et al, 2016*[3].

1.1.3. Sporadic and Familiar forms of NBL

NBL can be divided in sporadic or familiar NBL. The latter accounts only for 1-2% of total cases, is transmitted in an autosomal dominant fashion and is caused by inherited DNA variants in tumor-associated genes like *PHOX2B*, which is found mutated in about 10% of familiar NBL[3,18]. Regarding the sporadic form, although in most cases no specific etiology seems to be linked to the development of NBL, several risk (and protective) factors have been discovered throughout the last decades, which can be divided in genetic and non-genetic risk factors. While the first ones will be considered later, in this paragraph we will list briefly the non-genetic risk factors. A well-established one is represented by fetal exposure to alcohol[19]: indeed, alcohol exposure is associated to the risk of developing NBL with odds ratios (OR) up to 12.0[20]. Parental occupation represents another risk factor. It has been shown that the mother's exposure to electromagnetic fields or volatile hydrocarbons increases the risk of NBL in offspring with an OR of 1.5[21]. The assumption of drugs like anti-hypertensives, codeine or oral contraceptives during the pregnancy is also associated to the risk of developing NBL[20,22,23]. Protective factors have also been described to reduce the risk of NBL. The most protective factor is represented by the intake of vitamins like folate – whose role in neuronal development is well established - during pregnancy, which can reduce the risk on NBL up to 60%[24].

1.1.4. Genetics of Neuroblastoma

NBL is one of the genetically most heterogeneous and characterized tumors[3]. The following paragraphs list and describe the main genetic and genomic alterations that have been found throughout the years, including predisposition genes, somatic mutations and recurrent Structural Variants (SVs). The paragraph 2 will explain how these alterations have been shed to light, focusing on the importance that had in this sense the Genome-Wide Association Studies (GWAS) and the Next Generation Sequencing (NGS).

1.1.4.1.Genetic predisposition to Neuroblastoma

As stated in the paragraph 1.3, NBL can be sporadic or familiar. The first identified gene causative of familiar NBL is *PHOX2B*, which account for almost 10% of cases[18]. It encodes for a pivotal transcription factor (TF) involved in chromaffin cells differentiation from neural crest cells[25] (see Figure 1), and is also a disease-causing gene of other congenital malformations of neural crest origin closely related to NBL[26]. The second identified gene was *ALK*, an oncogene firstly identified as a partner of a translocation in anaplastic large cell lymphoma. This gene encodes a receptor tyrosine-kinase (RTK) involved in sympathetic nervous system development[27].

Since its discovery as a NBL predisposition gene, germline DNA variants in *ALK* have been found in almost all cases of familiar NBL[27,28]. Germline variants causative of NBL mostly comprise point mutation in the tyrosine-kinase domain, such as the R1275Q, G1128A, F1174* and F1245*[29,30]. However, apart from *PHOX2B* and *ALK*, no other gene has been associated to familiar NBL.

The sporadic NBL can also be predisposed by predisposition genes whose both low-penetrance common and high-penetrance rare germline variants are associated with the risk of developing NBL (Figure 3). For instance, about 2% and 8% of sporadic NBL patients are estimated to carry rare high penetrant germline Pathogenic Variants (PVs) in *PHOX2B* and *ALK* genes, respectively[30–32]. Another gene associated with susceptibility to NBL is *BARD1* – which encodes the binding partner of *BRCA1* – that may serve both as an oncogene or a Tumor Suppressor Gene (TSG)[33]. As detailly described in paragraph 1.2.1, the link between this gene and NBL was first seen in GWAS[34] which first revealed the association between *BARD1* common SNPs and the risk of NBL. Some years later, thanks to introduction of NGS, it has been possible to detect also the association between rare germline *BARD1* variants with NBL[9]. Very recently, a study demonstrated how both common and rare germline PVs variants in *BARD1* are causally linked to chromosomal instability in NBL[35]. The importance of this breakthrough relies in the strong positive association between chromosomal instability phenotype and poor outcome[19]. As will be discussed in paragraph 1.2.2, during the last decade high-throughput screening methods allowed to reveal other genes associated with NBL, generally carrying high-penetrance rare variants. The majority of these genes are involved in DNA replication, DNA repair and in maintenance of genome stability.

Nonetheless, despite the enormous advances, to date the full spectrum or rare NBL predisposing variants has yet to be defined.

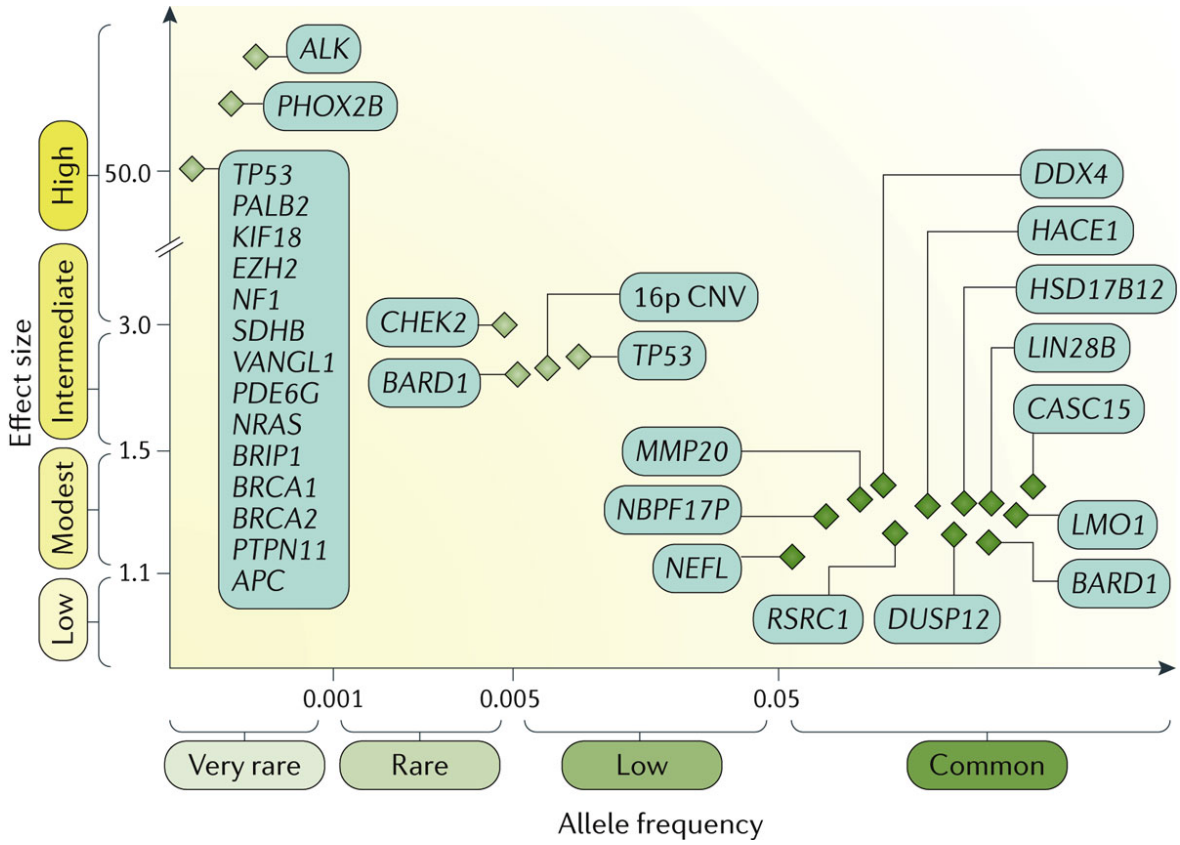


Figure 3. Frequency and penetrance of NBL-predisposing genes.

GWAS and NGS analyses allowed to detect several common and rare variants in NBL-predisposing genes, respectively. Allele frequency and penetrance (Effect size) are roughly inversely proportional to each other. *Adapted by Matthay et al., 2016*[6].

1.1.4.2. Genetics and Genomics of Neuroblastoma

In line with the majority of pediatric malignancies, NBL is characterized by a low degree of Tumor Mutational Burden (TMB) – generally defined as the number of non-synonymous mutation for megabase[36]. It is also characterized by a low recurrence of point mutations in driver genes, but on the other hand shows recurrent and typical large-scale genomic aberrations, many of which useful for risk stratification and application of personalized therapies[37]. Among the genetic alterations, the amplification of *MYCN* is one of the most frequent, with an estimated frequency of 20-30% in high-risk subtypes[38]. As discussed in the previous paragraph, *MYCN* amplification is one of the markers for clinical stratification, being *per se* sufficient to classify a tumor as high-risk according to INRG pre-treatment criteria (see Table 2). This gene maps in a region at 2p24.3 cytoband, and its amplification generally occurs through the formation of extrachromosomal and autonomous replicating regions called Double Minutes (DMs)[39]. *MYCN* is a TF which belongs to the *MYC* family of oncogenes. *MYCN* is defined as a master TF in NBL, as it regulates the transcriptional activity of hundreds of genes involved in different cell functions. For instance, *MYCN* indirectly guide the cell-cycle progression by the transcriptional activation of at least 10 kinases and TFs involved in cell division[40]. It also promotes the transcription activity of other TFs to establish regulatory circuitries and cell identity[41,42]. One of the main target of *MYCN* is represented by the *TERT* gene[43], whose *MYCN* increases the transcriptional activity. *TERT* encodes for the reverse transcriptase of the telomerase complex, and its hyper-activation maintains the telomere elongation mechanism in an active state[44]. *TERT* oncogene can also be affected by Gain of Function (GoF) mechanisms in NBL, generally with a mutual exclusivity with *MYCN* amplification[45]. Mutations of *TERT* usually involve genomic rearrangements (such as chromosome translocations or inversions) or amplification events involving the *TERT* locus (a ~600kb region at 5p15.33 cytoband), although it can also be activated by point mutations[46].

Point mutations occur at low recurrence in NBL, with the exception of *ALK*, which is the most affected gene at somatic level, being mutated in 6-17% of patients[47]. To date, this gene is considered an important therapeutic target, as patients with *ALK* mutations can benefit from therapies based on small molecules that specifically inhibit the tyrosine-kinase activity of the protein[48,49]. Another frequently mutated gene is represented by the TSG *ATRX*, a helicase required for the deposition of H3.3 histone at telomers, maintaining these sites in a quiescent status[50]. Loss of Function (LoF) mutations in *ATRX* – which ranges from point mutations to focal deletions[51] – trigger a mechanism of telomere maintenance called Alternative Lengthening of Telomeres (ALT), so that such mutations are thus mutually exclusive with *MYCN* amplification and *TERT* rearrangements[45]. Finally, in NBL are

recurrent mutations in the RAS-MAPK pathway – which acts downstream the RTK Alk – especially in relapsed forms of NLB[52].

Among other recurrent genomic aberrations in NBL, beside *MYCN* amplification and *TERT* rearrangements, we can find numerical or segmental CNAs, such as 17q gain, 1p loss, 11q deletion, 3p loss and chromosome 7 gain[53–56]. Of, note, some of these alterations can predict survival or are associated to the clinical outcome of patients (**Figure 4**). Generally, segmental CNAs – reflective of chromosomal instability – are associated to bad prognosis, while the presence of numerical CNAs correlates to low-risk tumors.

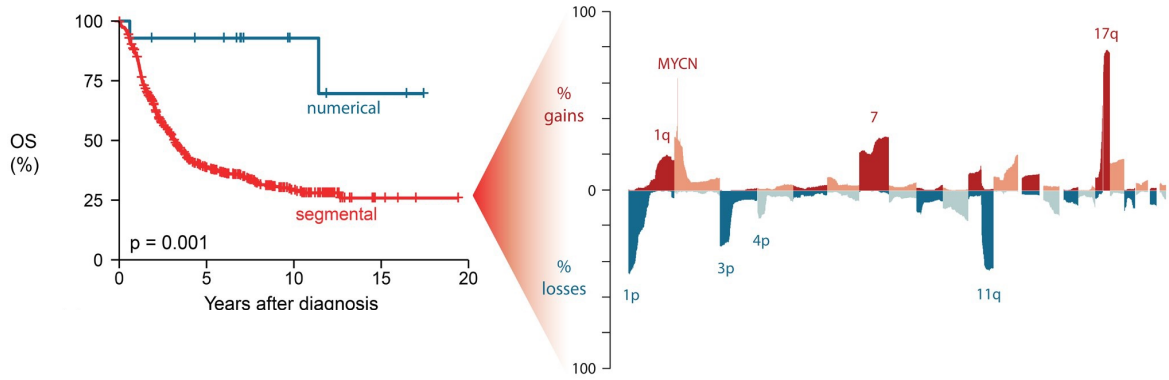


Figure 4. Segmental and numerical CNAs are associated to a decreased and an increased OS probability.

The figure, adapted from a study led on more than 500 NBL samples, shows how patients with segmental CNAs are less likely to survive compared to samples with numerical CNAs, with a 20-years OS probability less than 25%. To the right are shown recurrent segmental CNAs typical of NBL. *Adapted from Depuydt et al, 2018*[57].

1.2.NBL in the post-genomic era

As stated in the previous paragraph, the genomics of NBL is one of the most studied among both children and adult tumors[3]. The great majority of knowledge about genomic culprits of NBL, as well as its genetic predisposition, has been achieved thanks *i*) to introduction of high-throughput technologies that allowed the production of large amount of genomic data and *ii*) to the possibility to analyze these data by mean of specific informatic and bioinformatic tools. Of note, since the introduction of these techniques, a huge volume of data has been stored in genomic databases, allowing for the availability of genomic information that can be used by the worldwide scientific community[58].

Two techniques, above all, have revolutionized the field of Genetics, and have – and are being – extensively applied for the study of the genetics, the genomics and – in a broader sense – the biology of NBL: the GWAS and the NGS.

1.2.1. The GWAS and the genetic predisposition to Neuroblastoma

The GWAS are genetic case-controls studies that test for the statistical association between common Single Nucleotide Polymorphisms (SNPs) and complex diseases – such as hypertension, diabetes and cancer[59–61] – in an unbiased and genome-wide fashion. Schematically, GWAS is based on the SNP-array technique, which uses oligonucleotides (also known as probes) to interrogate up to almost one million common SNPs in the genome[62] allowing to genotype each common SNPs of a single individual; after a step of genotyping, the number of minor alleles or of specific genotypes at each SNP *locus* in of subjects affected by a specific condition (cases) is compared to the ones of healthy individuals (controls) in a large-cohort case-control study; the final step is the association analysis, that is the individuation of SNPs whose number is significantly higher – or enriched – in cases compared to controls (and vice versa) to identify predisposition *loci*[63]. Since the last 15 years, GWAS led to the identification of many predisposition *loci* in NBL. The first GWAS for NBL was performed in 2008 on a cohort of more than 1000 European-descendent American patients and 2000 controls, and the results replicated on a cohort of similar sample size. The authors identified a predisposition *locus* on chromosome 6 (6p22). In particular, 3 SNPs in linkage *disequilibrium* mapping in this region were associated to the risk of developing NBL, and were also enriched in patients with poor outcome[64]. A study led on high-risk NBL group of the same cohort, highlighted common SNPs in *BARD1 locus* (whose role has been discussed in paragraph 1.1.4.1) was enriched in this subgroup[34]. In the following years, other world-wide GWAS led on different populations identified other predisposition *loci* involving genes – such as *LMO1*, *LIN28B, DDX4* and *IL31RA* among the

others (see Figure 3) – whose role in NBL biology and pathogenesis have been subsequently elucidated[65–69].

Nevertheless, despite their relevance in the field of Human Genetics, GWAS present some limitations. One above all is the impossibility to study the contribution of rare variants to specific phenotypes[70].

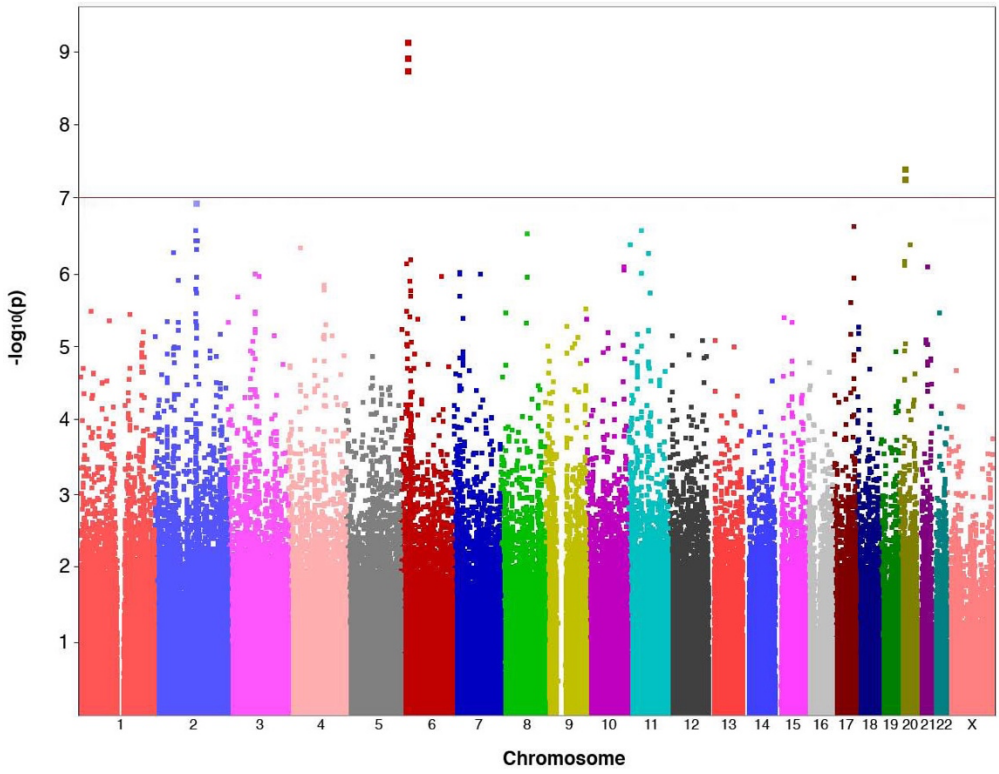


Figure 5. Manhattan plot showing the results of the first GWAS of NBL.

The graph, known as Manhattan plot, shows the significance level of each SNP expressed as -log10(P-value) resulting from a GWAS performed between comparison of more than 1000 NBL patients and 2000 healthy controls. Each dot represents a SNP differentially colored based on its chromosome. The horizontal red line indicates the genome-wide significance threshold, estimated based on the number of SNPs interrogated in the analysis (~500.000). Three LD SNP (rs6939340, rs4712653, rs9295536) on 6p22 locus (red) and two independent SNPs (rs3790171 and rs7272481) on 20p11 (gray) locus showed significance levels above the threshold, although only the first three retained statistical significance when accounting for population stratification. Adapted from Maris et al, 2008[64].

1.2.2. The NGS and Neuroblastoma

The NGS, also known as second-generation sequencing, is a sequencing technique introduced in the first half of 2000'. Compared to the classic Sanger sequencing technique, it allows the sequencing of a large amount of nucleic acid (DNA and RNA) from a biological sample thanks to the massive parallelization of the sequencing reactions[71]. The data obtained from these reactions can be stored, analyzed and processed by specific bioinformatic programs[72–74]. Since its introduction, several NGS-based techniques have been developed, such as the Whole-Exome Sequencing[75] (WES), used to identify mutations and rare SNVs in coding regions of the genome[76,77], the Whole-Genome Sequencing (WGS)[78], used to detect SNVs in non-coding regulatory DNA elements[79] or to assess the presence of SVs[80] and the RNA-sequencing (RNA-seq)[81], useful for the assessment of gene-expression pattern of biological samples[82].

Several NGS-based studies have provided important insights in the Genomic and Genetics of NBL. One of the first of these studies was carried out in 2013 on a cohort of 240 high risk NBL. By the integration of WGS an WES data obtained from different sequencing platforms the authors provided a first compendium of the genetic landscape of high-risk NBL. They also detected driver mutations in key NBL oncogenes and TSGs, some of which have been disclosed previously such as *ALK*, *MYCN*, *ATRX* and *NRAS*. Furthermore, comparing the study cohort with almost 2000 healthy individuals of the same ancestry, the authors observed an enrichment of germline rare PVs in genes like *ALK*, *CHEK2*, *PINK1*, *TP53*, *PALB2* and *BARD1* in which germline PVs – with the solely exception of the first one[28,83] – had not yet been reported in NBL[9].

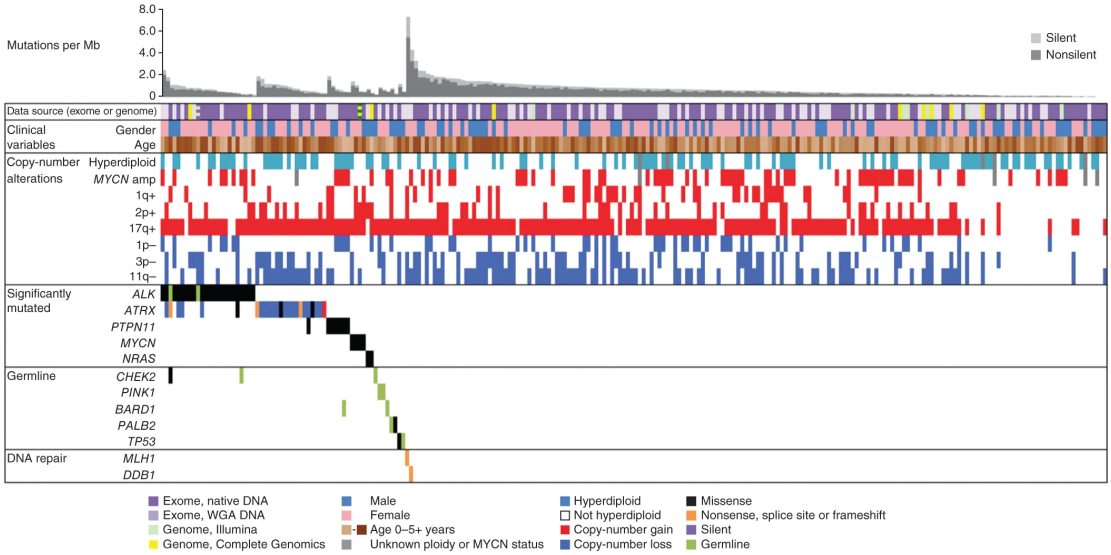


Figure 6. A first glance to the genomic landscape of NBL.

The heatmap shows the first genomic profile of NBL samples (columns) obtained by the mean of WGS and WES. The upper bar plot shows the TMB expressed as somatic mutations per Mb; The rows, from top to bottom, show the data source, the clinical parameters, the sex, the CNAs, the significantly mutated genes at somatic level and the genes affected by germline PVs, respectively. *Adapted from Pugh et al, 2013*[9].

Other subsequent studies based on DNA sequencing have increased the knowledge about genetic features of NBL. A study that integrated 56 WGS and RNA-seq data identified recurrent genomic rearrangements at the locus 5p15.33, in the close proximity of the *TERT* oncogene. the occurrence of *TERT* rearrangement was mutually exclusive with *MYCN* amplification, and samples with both *TERT* rearrangement or *MYCN* amplification showed increased *TERT* expression. The mutual exclusivity underlies the biological role of *MYC* oncogenes family and *TERT*[43,44], converging to a common Telomere Maintenance Mechanism (TMM)[45]. In line with these findings and given its biological function as a repressor of *TERT*[50], *ATRX* LoF mutations were subsequently found to be mutually exclusive with both *MYCN* amplification and *TERT* rearrangements[84], although giving rise to mechanisms of ALT[85]. Another study led on WGS data performed on 23 NBL samples showed that the RAS-MAPK pathway is frequently altered at various levels in relapsed tumors, with GoF mutations that can affect RTKs like *FGFR1* and *ALK*, the GTPase family of RAS (such as *KRAS* and *NRAS*), Mitotic Activating Protein Kinases (MAPK) like *BRAF* and also LoF mutations in TSGs which regulate the pathway such as the Guanosine Exchange Factor *NF1*[52]. A recent WES-based study led by our group provided further insights into genetic predisposition of NBL. In particular, by comparing coding germline SNVs of almost 700 cases and more than 800 healthy controls, we observed an enrichment of germline PVs in genes of Homologous Recombination (HR) pathway such as *BRCA1* and *RAD51C* (the latter never since reported as NBL predisposition gene) and also an enrichment of germline predicted PVs in genes involved in neural tube differentiation and in genes associated to neurodevelopmental disorders[86]. The NGS and in particular WGS can be also used to detect the presence and the role of genomic rearrangements in cancer[87]. In NBL, an analysis performed on WGS an SNP-array data suggested that SVs affect genes involved in neuronal differentiation. In detail, this study reported the gene *SHANK2* as the more frequently disrupted by SVs[88]. Indeed, this gene is located in a locus (11q13) involved in a known NBL translocation t(11q-17q) whose frequency on a medium-to-large cohort had not been yet estimated before[89]. Nonetheless, the role and the full spectrum of genomic rearrangements in NBL remains so far unraveled, mainly due to the difficulty to study these events with the NGS short-reads techniques[87].

WES and especially WGS are also used for the identification of specific tumor mutational patterns known as mutational signatures. Based on the kind of mutations (CNAs, SVs, SNVs) a plethora mutational signatures can be extracted from tumors. The most characterized of these patterns are the single base substitutions (SBSs) signatures. In brief, SBSs signatures are computed by comparing the percentage of SNVs of a tumor in of all the 96 possible tri-nucleotide contests (5'-NpC>ApN-3'; 5'-NpC>GpN-3'; 5'-NpC>TpN-3';

5’-NpT>ApN-3’; 5’-NpT>CpN-3’; 5’-NpT>GpN-3’, where N is any of A, C, G or T) with cancer tri-nucleotide profiles deposited in public databases. Currently, the most interrogated database is the Catalogue of Somatic Mutations in Cancer (COSMIC), which stores 86 SBSs signatures (to November 2023), many of which with specific etiological explanation[90]. The importance of the identification of SBSs signatures relies in their causal link to distinct mutational processes that are promoted by etiological agents – i.e., tobacco smoking or reactive oxygen species (ROS) – or underlie distinct tumor features – such as HR or Mismatch Repair deficiencies[91]. in recent years the presence of Cosmic SBSs signatures has been investigated in NBL. Although with slight differences, literature data report recurrent and typical signatures that are overall shared among NBL samples, such as SBS1, SBS5, SBS18 and SBS40[11,92,93]. The SBS1 is characterized by C>T substitution in the 5’-NpCpG-3’ context (where N is any of A, C, T, G) due to the deamination of a 5-metylcytosine to thymine, and has been proved to correlate with the advanced age at diagnosis[94]. The SBS5 also seems to be correlated with age at diagnosis, although to date its etiology remains unclear[91]. The SBS40 has not been yet linked to any phenomenon. Finally, the SBS18, which is probably the more distinctive of NBL, is characterized by C>A substitution[91]. The occurrence of this substitution is ascribed to the production of ROS which promote the formation of the 8-oxoguanine (8-oxoG), a modified base that preferentially matches an adenine instead of a cytosine[95]. Notably, some of the SBSs signatures have been associated to specific NBL characteristics. For instance, Brady et al. demonstrated that SBS18 is more active in tumors with *MYCN* amplification, 17q gain and high mitochondrial genes expression. The same group demonstrated how the majority of driver mutations in genes such as *ALK*, *ATRX*, *PTPN11*, *NF1* and *NRAS* belong to signature SBS18 (Figure 7)[11].

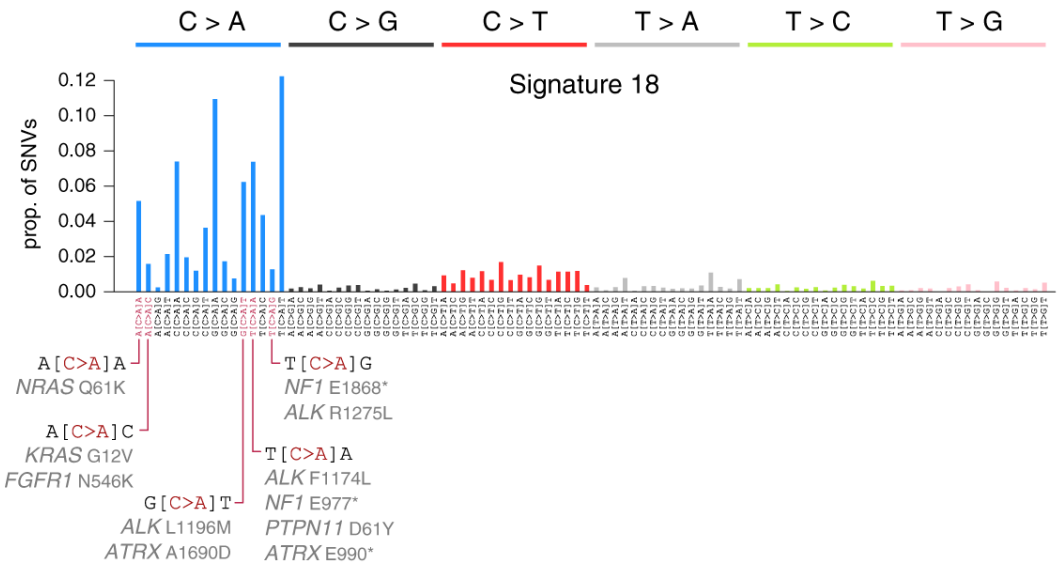


Figure 7. SBS18 causes driver mutations in NBL.

On a cohort of 205 WGS of NBL samples, Brady et al, demonstrated that many recurrent driver NBL mutations are caused by the signature SBS18, whose profile is depicted in the figure. Adapted by Brady et al, 2020[11].

If on the one hand DNA sequencing techniques like WES and WGS helped researchers to improve the knowledge about genetic predisposition and genomic features of NBL, on the other one RNA-seq provided important insights on the biology of NBL, highlighting and unravelling its complex heterogeneity. For instance, by integrating experiments of Chromatin Immuno-precipitation sequencing and RNA-seq on NBL and human neural crest cell lines it has been proven that the majority of NBLs belong to two different types of cell identities with different gene-expression profiles (Figure 8). One, more differentiated referred to as sympathetic noradrenergic (or simply noradrenergic), is defined by core regulatory circuitries promoted by key TFs of sympathetic development (see Figure 2), including *GATA2/3*, *HAND1/2* and *PHOX2A/B*. The other one, more similar to Neural crest cells and thus named NCC-like or mesenchymal, is defined by core regulatory circuitries composed by members of the AP-1 complex such as *FOSL1/2*, *RUNX1/2* and *IRF1/2/3*[41,96]. Beside from showing different gene-expression profiles and core regulatory circuitries, these two cell identities are characterized by different response to treatments[96,97].

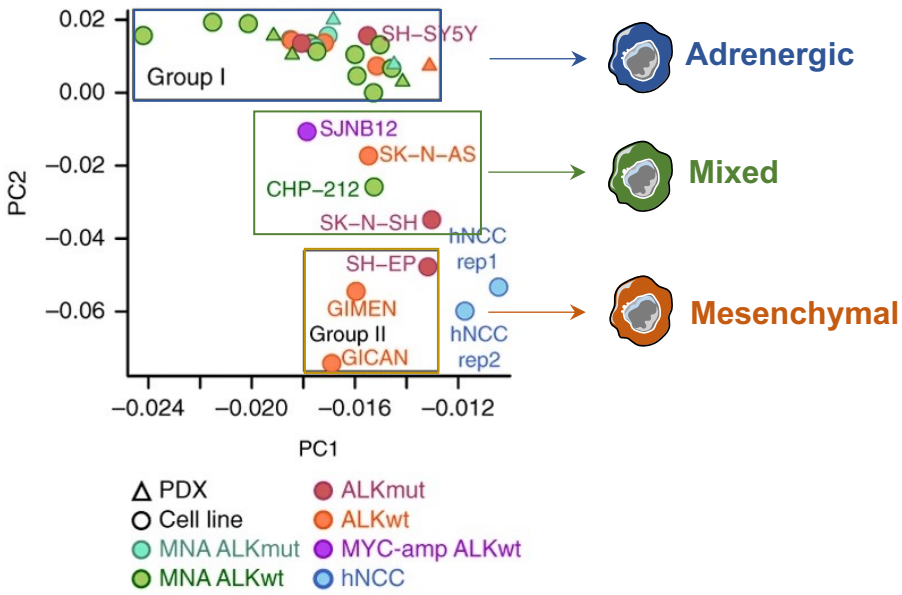


Figure 8. NBL shows 2 defined cell identities plus a third mixed phenotype.

Principal Component Analysis based on super-enhancer of 25 NBL cell line and 2 human neural crest cell lines showed two well defined clusters reflective of two cell identities, one defined adrenergic (blue) and the other mesenchymal (Orange). These two phenotypes are sustained by different core regulatory circuitries and show significant biological differences, as discussed in the main text. 4 cell lines showed a mixed phenotype between the mesenchymal and adrenergic one (green). Adapted from Boeva et al., 2017[41].

To summarize, the introduction of NGS allowed the discovery of genetic, genomic and biological characteristic of NBL, often providing important implications concerning patients' outcome and response to therapy. DNA sequencing techniques – including WGS and WES – provided, throughout the years – a comprehensive vision of genetic and genomic alterations of NBL, helped to identify key oncogenes and TSGs and to discover biological processes that withstand the occurrence of driver mutations in such genes. Other techniques – such as RNA-seq and ChIP-seq – have been used to unravel the biological heterogeneity of NBL, helping to identify distinct NBL identities with different phenotypical features.

However, in spite of all the advances that have been made, the current knowledge of NBL only partially explain the clinical and biological heterogeneity of NBL, and a comprehensive characterization for a proper risk stratification is still lacking.

1.3. *Germline predisposition to tumor phenotypes*

As discussed in the previous section, NGS analyses have been extensively used to detect and study somatic genetic and genomic alterations. Approaches to identify somatic mutations are generally based on a comparison of the germline background of an individual and his/her tumor counterpart, with the first only serving as a mere reference for the detection of putative cancer-driver mutations[98]. In this scheme, the role of germline inherited variants in the onset, development and biology of tumor remains uninvestigated.

It is known for decades, however, the role that germline variants can have in the development of cancer. The first link between germline variants and tumor predisposition was described in 1971 by Alfred Knudson on patients affected by retinoblastoma. He observed that patients with hereditary retinoblastoma, a form of the tumor which is caused by heterozygous germline PVs in *RB* gene, developed a bilateral form of tumor, conversely to patients affected by the sporadic type, and that tumor cells from hereditary forms lost the wild type (wt) allele of *RB* gene, in a mechanism went down to history as Loss of Heterozygosity (LOH)[99]. LOH promote tumorigenesis by removing the proliferation brake provided by the wt allele, creating a CN mismatch at TSG *loci* known as Allele Specific Imbalance (ASI) (Figure 9).

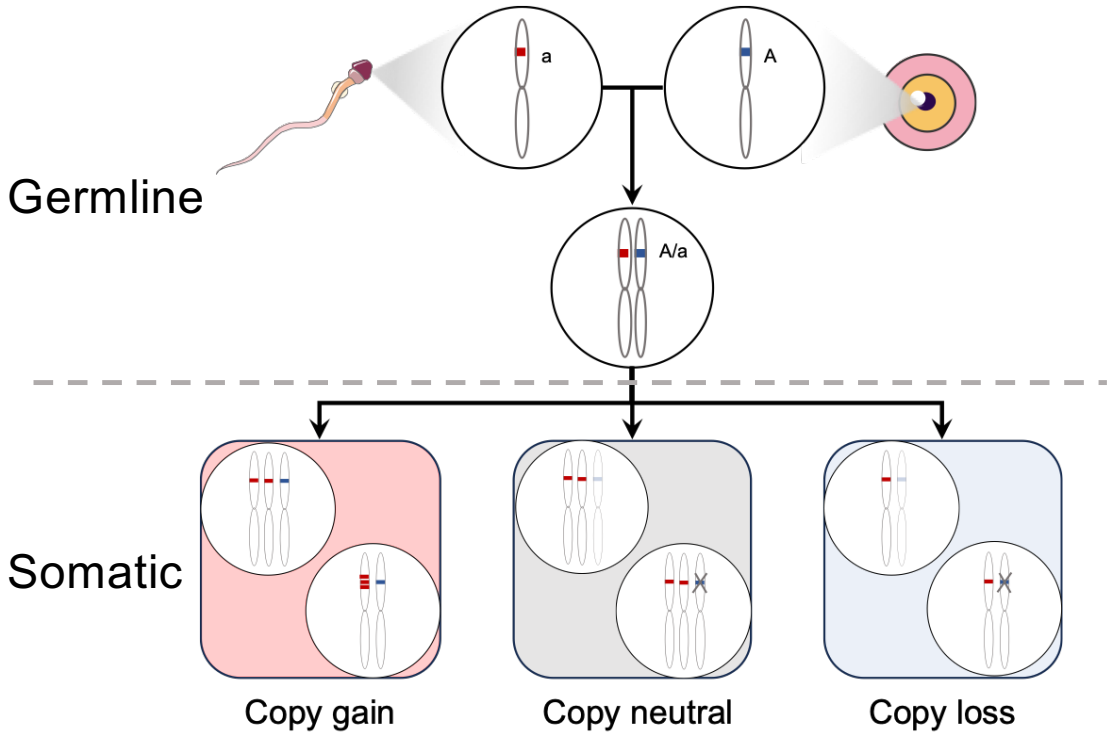


Figure 9. Schematic depiction of somatic LOH for ASI.

The inheritance of a PV in TSG (a allele, red) strongly predispose to ASI at somatic level, leading to LOH of that allele and in turn cancer development. ASI can occur through a copy number gain of the recessive allele (copy gain), a copy number gain of the recessive allele accompanied by a copy loss of the wt allele (copy neutral) or by the loss of the wt allele (A allele, copy loss). This mechanism explains the reason why cancer-predisposing syndromes caused by PVs in TSG are inherited in a dominant fashion.

This mechanism had subsequently been proposed as the cancer-driver force cancers arising from tumor predisposition syndromes, including NBL[100,101]. Thanks to the introduction of GWAS it was clear that not only rare PVs, but also common SNPs could be linked to ASI. Analyses of polymorphic risk *loci* detected by GWAS from different tumors, including colorectal cancer, glioblastoma, cutaneous squamous cell carcinoma, glioblastoma and myeloproliferative neoplasms, showed that risk allele was preferentially gained at somatic level to the detriment of the non-risk allele[102–108].

The mechanism of ASI described above is defined in *cis*, as it involves the somatic CN imbalance of the same gene, or the same *locus*, of the germline variant. However, some studies indicate that germline variation can act *in trans* to predispose mutations of different and apparently unlinked genes, in a model known as Germline Variant by Somatic Mutation (GxM) Association[108,109]. A pan-cancer GWAS conducted on almost 6000 individuals from The Cancer Genome Atlas database showed that germline SNPs can be associated to somatic mutations of genes at distant *loci*[110]. In this study, the authors identified 28 predisposition *loci* whose polymorphisms were associated to an increased probability of somatic mutation of 20 distally located cancer-related genes. The results of this study implemented the knowledge about gene networks in cancer. For instance, a SNP 19p13.13 *locus* were associated to a 4-fold increased probability of somatic *PTEN* phosphatase. *PTEN* is a TSG that encodes a phospholipid phosphatase that dephosphorylates the phosphatydylinositol-3-phosphate to shut down the oncogenic pathway of PI3K/AKT/mTOR[111]. The 19p13.13 locus contains an activator of the mTOR pathway (*GNA11*[112]) whose expression is increased in presence of the risk allele. The authors concluded that the risk SNP increased the expression of *GNA11*, which in turn facilitated the LoF of *PTEN*, a TSG that serves as a brake in the PI3K/AKT/mTOR pathway. Alongside other similar findings, the results of this study suggest that genetic background of oncologic patients can provide a fertile ground for the development of defined somatic events, including mutations of specific genes or the onset of a particular tumor phenotype.

Other studies, on the other hand, have also described how germline DNA variation can influence a broad spectrum of somatic characteristics that range outside the genetics. For instance, it is well known that breast cancer patients with germline PVs of *BRCA1* and *BRCA2* develop an aggressive basal-like triple negative subtype[113]. Different GWAS studies demonstrated he onset of different histological subtypes of breast cancer can be predisposed by inherited germline polymorphisms[114–116]. Other somatic features, such as mutational signatures, can be associated to germline SNVs. A study led on more than 1000 individuals affected by medulloblastoma reported that patients who carry rare germline variants of *BRCA2* and *PALB2*, two genes involved in HR[117], showed increased activities of COSMIC

signatures SBS3 and SBS8[118], two signatures associated with HR deficiency[94,119]. In the same study authors detected a higher frequency of chromothripsis in patients with PVs in *TP53*, confirming what had been previously reported[120]. Another WES/WGS-based analysis on more than 300 colorectal cancer individuals from almost 200 families showed that individuals with rare and predicted LoF mutations in *MDB4*, a glycosylase involved in Base Excision Repair, showed a somatic enrichment of C>T transitions typical of the signature SBS1[94], resulting from a lack of repair of G-T mismatches by *MDB4*[121].

Finally, inherited germline variation in genes involved in the adaptive immune response can determine the occurrence of defined tumor driver mutations. This is the case, for instance, of the *HLA* haplotype, where a study conducted on 9000 TCGA tumors that analyzed a subset of 1000 cancer driver mutations showed that class I Mayor Histocompatibility Complex (MHC-I) – encoded by the *HLA* – variation predisposes to the onset of specific and well-known somatic driver mutations. For instance, they demonstrated that a weak MHC-I-antigen interaction is associated with the $BRAF^{V600E}$ mutation, while a strong interaction predisposes to the $IDH1^{R132C}$ mutation[122] (Figure 10).

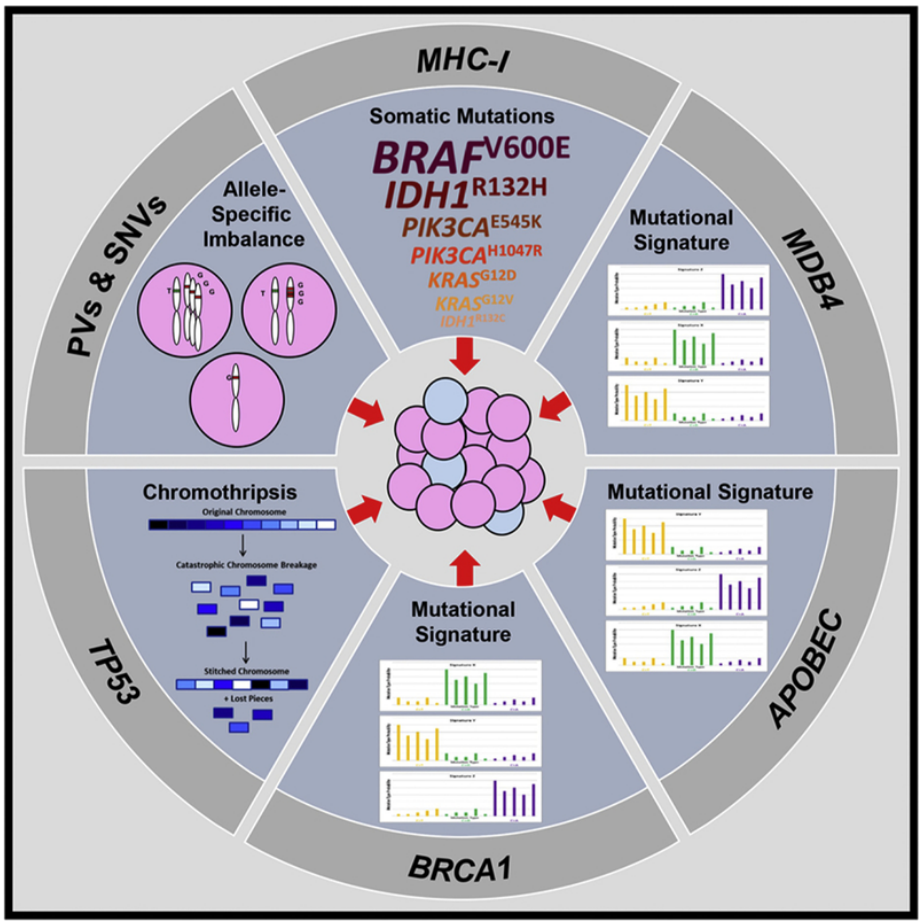


Figure 10. Examples of germline-to-somatic correlations.

The figure provides some examples of the germline-to-somatic correlation discussed in the main text. Germline variants in the genes listed in the outer layer of the circle are causative – in some tumors – to

the somatic feature depicted in corresponding inner wedge. All the correlations depicted have been disclosed in the main text. *Adapted from Ramroop et al, 2019*[108].

2. *Aim*

NBL is a pediatric neoplasm characterized by an elevated degree of genetical and biological heterogeneity, which reflects its difference in the outcome of affected patients. Indeed, the 5-years survival probability shrinks up to the 50% between low and high-risk tumors. Thanks to the introduction of GWAS and especially the NGS, In the last decades several aspects of the biology and of the genetics of NBL have been shed to light, often being linked to the clinical outcome, improving the clinical stratification of patients. For instance, it is now well established that genetic instability and SVs are strongly predictive of a poor survival probability. However, concerning the genetics of NBL, a comprehensive vision of NBL landscape, including a compendium of all the somatic alterations of this tumor, is still lacking, leaving holes for a correct patient stratification. Furthermore, although genetic predisposition to NBL has been long studied and several germline DNA variants – both common and rare – have been identified to predispose for this tumor, still few is known about the role of germline variation in the predisposition of defined somatic phenotypes.

Given these assumptions, the aims of this dissertation are to *i)* provide a complete characterization of somatic genetic landscape of NBL, with a main focus on SVs which remain so far poorly investigated, with a view to find unreported links between genomic alterations and clinics; *ii)* investigate genetic predisposition to the development of SVs and, in a broader sense, to genomic instability.

To address our goals, we leveraged publicly available WGS and RNA-seq data from two independent cohorts of NBL. Using both available programs and *in-house* scripts, we set up pipelines for detecting, filtering, processing and analyzing somatic SNVs, CNAs, SVs and germline SNVs.

3. Methods

3.1. NLB Samples and Pipelines

As introduced in the *Results* section, in this dissertation we used WGS data from two publicly available NBL databases – the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) and the European Genome-phenome Archive (EGA). In detail, from TARGET database (https://www.cancer.gov/ccg/research/genome-sequencing/target) we used data from a group of 136 matched normal-tumor samples belonging to the dbGaP study phs000467.v23.p8[9], which included NBL patients of the Children Hospital of Philadelphia. Patients, tumors and files belonging to this dataset will be referred to as the TARGET dataset, for simplicity. From the second database (https://ega.crg.eu), which included NBL patients from Germany involved in clinical trials from 1991 and 2018 of the Society for Pediatric Oncology and Hematology, we used WGS data of 180 NBL paired normal and tumor samples from 2 studies: EGAS00001001308[45] (N=55) and EGAS00001004349[85] (N=125). For simplicity, patients, tumors and files belonging to this dataset are referred to as the EGA dataset.

3.1.1. Data description and publicly available files

In this study we leveraged WGS data for analysis of germline and somatic SNVs, somatic CNAs and somatic genomic rearrangement, hereby referred as SVs for simplicity. For a subset of patients, we also processed and analyzed RNA-seq data for gene-expression analysis. Data availability and file formats of WGS and RNA-seq data differed between the two datasets: *i*) for what concerns WGS data, from TARGET database we downloaded individual (i.e. per sample) Variant Calling Format (VCF) files of somatic and germline SNVs, Tab Separated Values (TSV) files of 2kb-windows normalized relative coverage for detecting somatic CNAs and TSV files of "high-confidence SV calls" for somatic SVs, which included SVs with at least 10 supporting discordant reads (DR) (i.e. split-reads or reads whose one of the pair maps to a distant genomic location) with a frequency less than 1% in dbVar database[123] and not annotated as "germline" in 1000 Genomes SV Project[124]. All the files described above have been produced through the Cancer Pipeline v2.0 by the Complete Genomics (https://www.completegenomics.com) using GRCh37 as reference genome; *ii)* from EGA dataset, we downloaded two Binary Alignment Mapping (BAM) files per patient (one relative to the normal and the other to the tumor sample) which were previously obtained[45,85] with BWA-MEM versions 0.6.1 or 0.7.8 (https://github.com/lh3/bwa) using GRCh37 as reference genome, whose duplicates were marked Sambamba v 0.6.5[125] and filtered using SAMtools v0.1.19[126]. RNA-seq of TARGET samples were available as matrix of Fragment Per Kilobase per Million (FPKM) where each

cell indicated the FPKM value of each gene (rows) in each sample (columns) (https://target-data.nci.nih.gov/Controlled/NBL/WGS/CGI/). Raw-sequence (FASTQ) files of RNA-seq of EGA samples were downloaded from EGA database under the accession numbers EGAS00001001308 (N=54) and EGAS00001004349 (N=86).

Files from TARGET are publicly available under controlled access in the dbGaP database[127] and has been queried using the graphic-user interface of the Globus Connect Personal client (https://www.globus.org/globus-connect-personal). Files from EGAS00001001308 and EGAS00001004349 are publicly available under the controlled access in the EGA database and have been downloaded using the command-line interface of PyEGA client (https://github.com/EGA-archive/ega-download-client). Table 3 summarizes the main features of the two cohorts of samples.

		TARGET	**EGA**
General information	# Patients	136	180
	Study ID	phs000467.v23.p8	EGAS00001001308, EGAS00001004349
	Client	Globus Connect Personal	PyEGA
	Reference Genome	GRCh37	GRCh37
Sequencing	Type of sequencing	Paired end	Paired end
	Platform	Illumina HiSeq 2500	Illumina HiSeq 2000; Illumina patterned flowcell
	average read length	~71bp	~101bp
Data availability	mapping file (BAM)	No	Yes
	germline SNVs file	Yes	No
	somatic SNVs file	Yes	No
	somatic coverage	Yes	No
	somatic SVs	Yes	No
	RNA-seq	Yes (89 of samples)	Yes (140 samples)
Clinical features availability (%)	Risk classification	100%	84.4%
	INSS stage	100%	84.4%
	Age at diagnosis	100%	84.4%
	MYCN amplification	100%	84.4%
	OS information	100%	75.5%
	EFS information	100%	75.5%

Table 3. Summary characteristics of samples in TARGET and EGA datasets

3.2. Variant calling and processing pipelines

Downloaded files were subsequently processed to obtain good quality and easy-to-manipulate files containing information about germline and somatic SNVs, somatic CNAs and somatic SVs. In order to reduce batch effects, where possible, downstream processing was equally performed in the two cohorts.

3.2.1. *Germline SNVs calling*

Individual germline SNVs of TARGET samples were selected from VCF files of joint normal-tumor calling using BCFtools v1.10.2[128], whilst for EGA samples individual germline SNVs were singularly called with Strelka v2.9.10[129] and filtered with BCFtools v1.10.2 to include "PASS" variants. In detail, upon germline variant calling, Strelka labels as "PASS" SNVs *i*) that have a locus depth ≥ 3, *ii*) which can be properly genotyped or *iii*) whose genotype is consistent with chromosome ploidy. Subsequently, resulting VCF files from both datasets were annotated using Annovar[130] and SNVs classified according the American College of Medical Genetics (ACMG) classification with Tapes[131]. Finally, we applied Quality Control (QC) filters including variants with at least 8 read covering the alternate allele, a Quality by Depth (QD) equal or greater than 3 and a Mapping Quality (MQ) of at least 30 (the last 2 QC criteria only applied to EGA samples whose such information was available). We then selected only coding variants (exonic or splicing according to RefSeq database[132]) which served as input for downstream variant prioritization steps (see paragraph 3.5) (Figure 11).

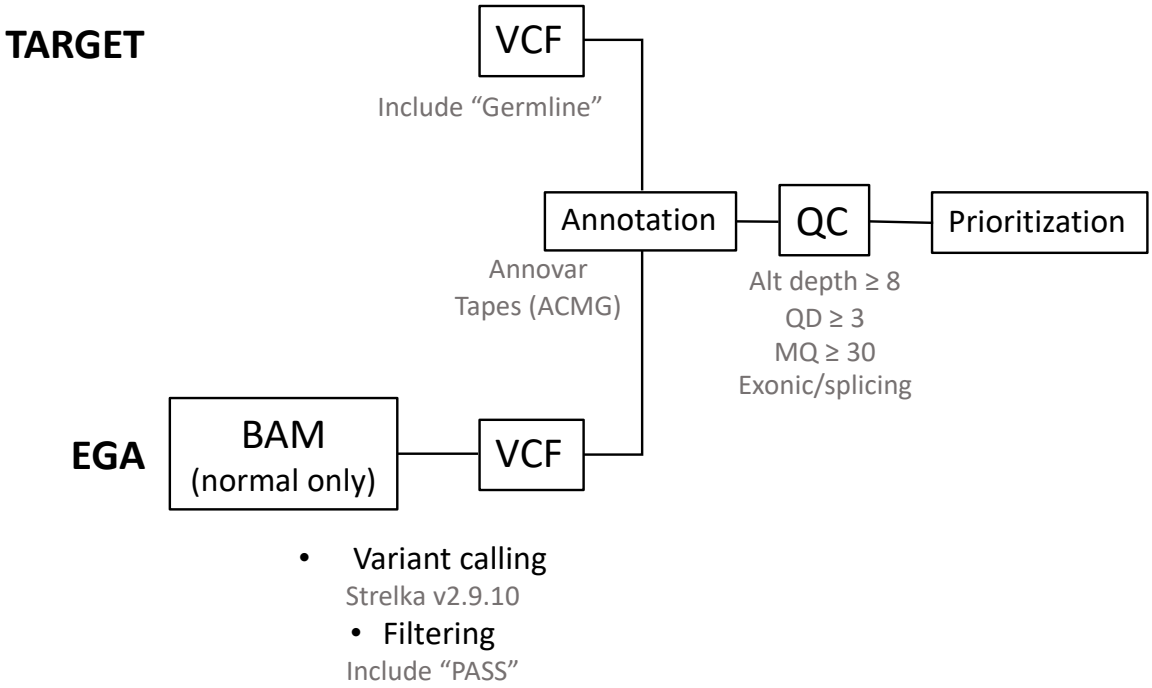


Figure 11. Germline SNVs calling and processing.

The workflow depicts the main steps through which we obtained QC-filtered exonic germline SNV in TARGET (top) and EGA (bottom) dataset. Details of each step are provided in the main text. Alt depth: depth of the alternate allele; QD: Quality by Depth; MQ: Mapping Quality.

3.2.2. Somatic SNVs calling

Somatic SNVs for each sample in the TARGET dataset were obtaining filtering in "somatic" SNVs from VCF file of joint normal-tumor calling using Bcftools v1.10.2. EGA somatic SNVs were called using Strelka v2.9.10 in a matched normal-tumor fashion and selected those labeled as "PASS" with Bcftools v1.10.2. In detail, Strelka considers as "PASS" somatic SNVs mapping in a locus whose tumor depth is not greater than threefold the depth of the same locus in its normal counterpart. Somatic variants from the two datasets were uniformly annotated using Annovar for information about rarity and pathogenicity. Finally, we applied QC filters including variants with at least 5 read covering the alternate allele, a Variant Allele Frequency (VAF) of at least the 5% and a mapping quality of at least 30 (the last QC criterium only applied to EGA samples as such information was unavailable in TARGET cohort). Resulting somatic QC-filtered files were used to compute TMB and parallelly given as input for downstream variant prioritization steps (see paragraph 3.4.1) (Figure 12).

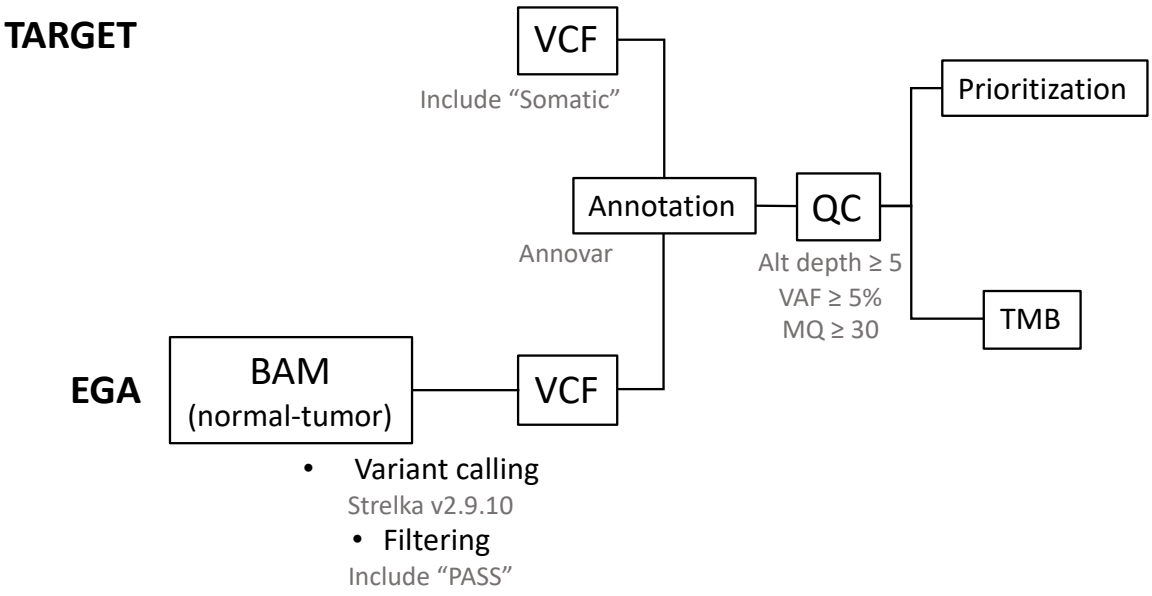
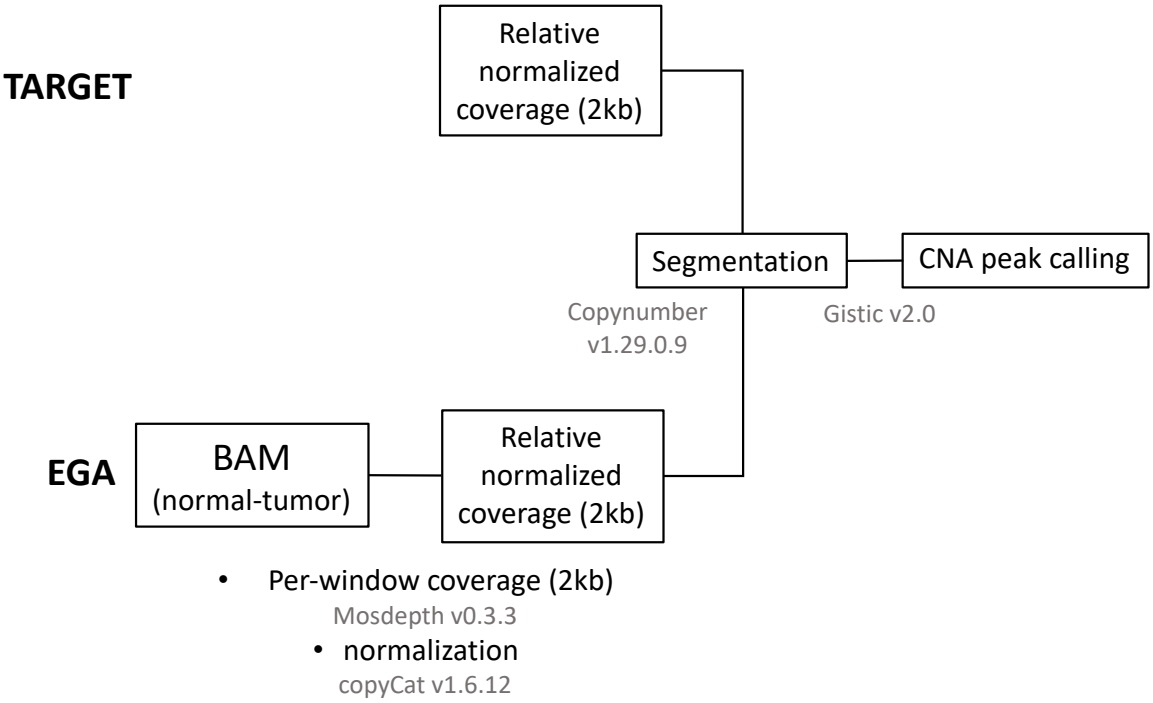


Figure 12. somatic SNVs calling and processing.

The scheme depicts the workflow for somatic SNVs calling and filtering in TARGET (top) and EGA (bottom) datasets. Details of each step are provided in the main text.

3.2.3. Somatic CNAs calling

For inferring CNAs and, more in general, somatic CN profile of NBL samples, we first computed the per-window (2kb) mean coverage of normal and tumor sample for each patient of the EGA dataset using Mosdepth v0.3.3[133] which were then normalized for CG content and mappability through CopyCat v1.6.12[134]. Subsequently, normalized tumor and normal coverage profiles were compared to each other to obtain a coverage file of relative (tumor vs normal) CN. The value of CN ratio was ln-transformed for downstream analyses (hereby referred to as logR). As specified in paragraph 1.1, somatic normalized logR per 2kb windows files were available in the TARGET dataset. We performed CN segmentation with the Bioconductor package Copynumber v1.29.0.9[135]. In detail, to achieve a smoother CN profile we applied a piecewise constant fitting algorithm setting as 2 the minimum number of contiguous bins (k) to form a segment and 1000 as discontinuity penalization (gamma). logR thresholds of deep loss, loss, gain and amplification were set to -0.69, -0.29, 0.22, and 1.39, which corresponded to a CN value of 0.5, 1.5, 2.5 and 8 in diploid regions, respectively. Segmentation files of each dataset were merged and given as input to Gistic v2.0[136] with default parameters to retrieve the significance and the boundaries of common amplification/deletion regions (Figure 13).



***Fig*ure 13. Somatic CNA calling.**

The scheme depicts the workflow for somatic CNAs calling and filtering in TARGET (top) and EGA (bottom) datasets. Details of each step are provided in the main text.

3.2.4. Somatic SVs calling

Somatic SVs of EGA samples were called in a matched normal-tumor fashion with Manta v1.6.0[137] and SVs defined as "PASS" and with a number of DR of at least 15 were filtered in with Bcftools v1.10.2. Briefly, Manta labels as "PASS" somatic SVs with an adequate MQ which are not present in the normal counterpart. EGA and TARGET somatic SVs were then filtered according to their type to include only chromosome translocations (TRA) and inversions (INV). TARGET SVs were also labeled as complex if the SV-calling pipeline was not able to define a specific variant type. To better characterize these variants, we used *in-house* R scripts to retrieve the CN status of the two breakpoints of a complex variant. A complex variant was included if the two breakpoints mapped to different chromosomes (defined as TRA) or the CN status of both breakpoints were equal to 2 (defined as INV) resulting good-quality SV files served for downstream analyses (see section 3.4.3) (Figure 14).

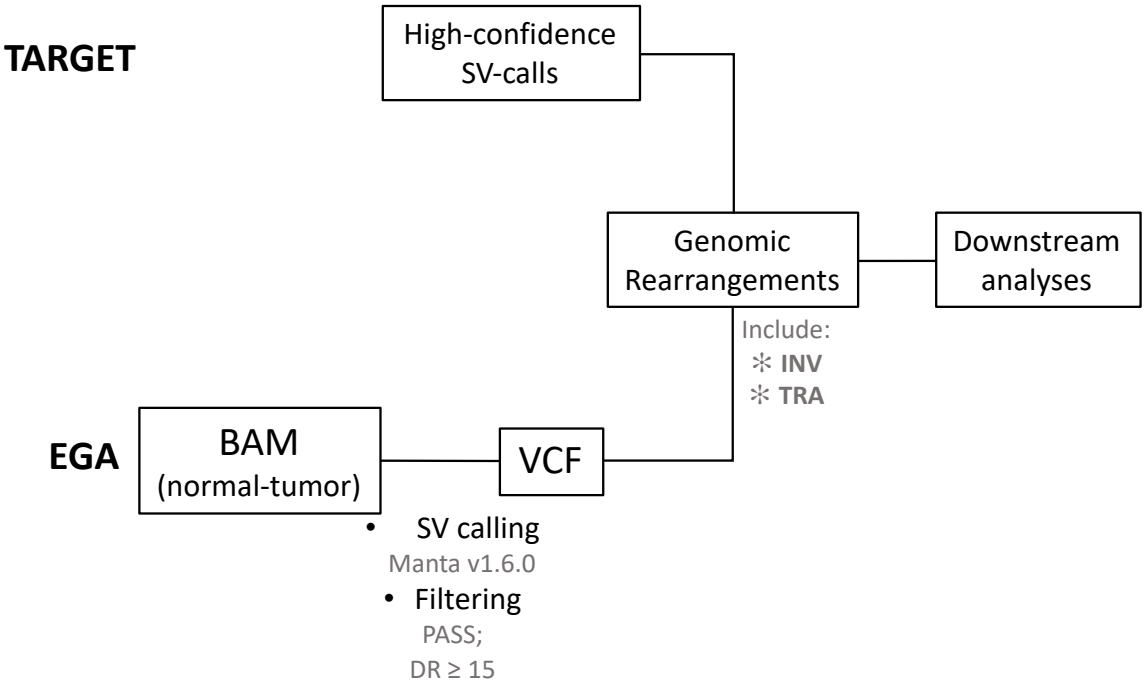


Figure 14. Somatic SVs calling.

Schematic depiction of the workflow for somatic SVs (including TRA and INV) calling and filtering in TARGET (top) and EGA (bottom) datasets. Each step is described detailly in the main text. DR: Discordant Reads; INV: Inversions, TRA: Translocations.

3.3. RNA-seq processing

For the TARGET cohort a genes-to-sample (G×S) FPKM matrix was available, which was used as such for downstream analysis. Paired-end FASTQ files of RNA-seq of each EGA sample were mapped against GRCh37 reference genome with STAR v2.7.10a aligner[138]. Raw read counts per gene were obtained with FeatureCounts v2.0.1[139] setting parameters as previously described[85] using the ENSEMBL GRCh37 transcriptome as a reference[140]. Individual raw-counts files were joined to obtain a G×S raw-count matrix and row-counts were subsequently normalized to FPKM using the DGEobj.utils R package v1.0.6[141] to produce a G×S FPKM matrix. TARGET and EGA G×S FPKM matrices served for downstream analyses of Differential Gene Expression (DGE) (see paragraph 3.4.6) (Figure 15).

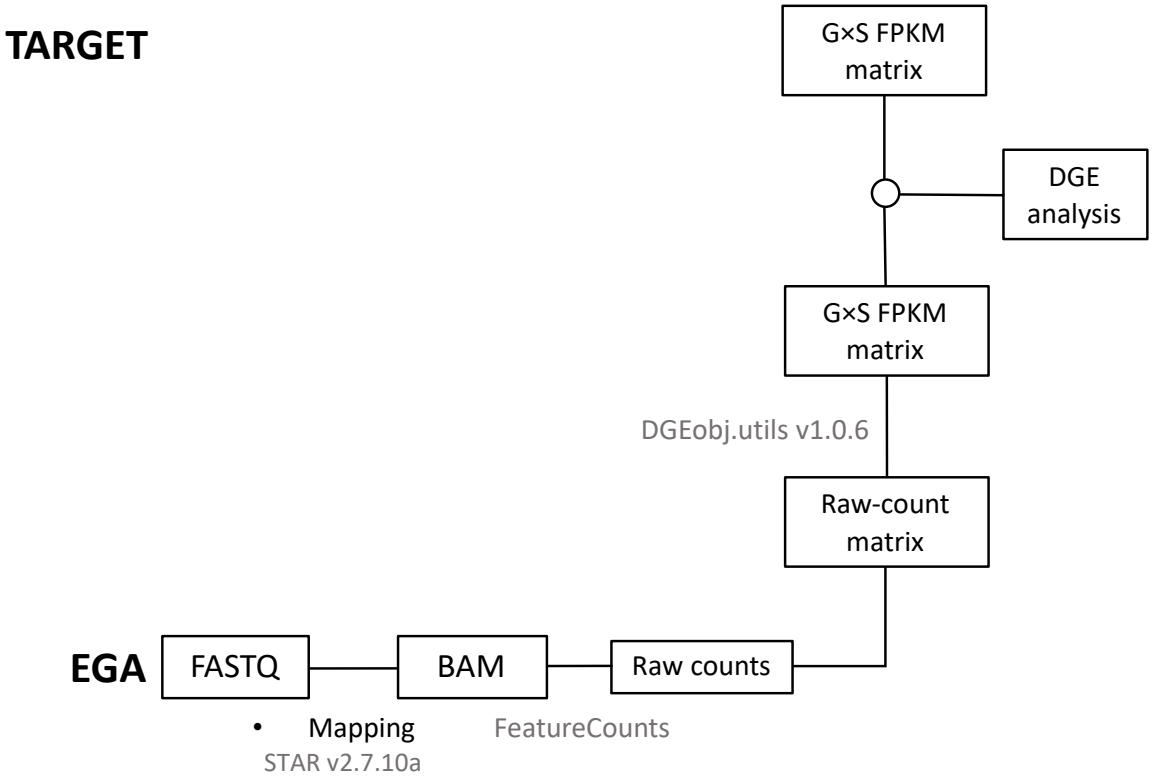


Figure 15. RNA-seq pipeline.

Scheme of the pipeline applied to process RNA-seq data of TARGET (top) and EGA (bottom) samples. Each step is detailly disclosed in the main text. G×S: gene-to-sample; DGE: Differential Gene Expression.

3.4. Genomic profiling and downstream analyses

Germline, somatic and RNA-seq files produced as described in the previous paragraph have been used as input to perform genomic profiling and a series of statistical analyses whose results are summarized in the *Results* section. The following paragraphs provide a comprehensive description of the methods (criteria, statistical tests and programs) applied to achieve such results.

3.4.1. Tumor mutational burden and somatic SNVs prioritization

From QC-filtered somatic SNVs (see Figure 12) we computed TMB for each sample counting the number of whole-genome non-synonymous SNVs for megabase, as previously described[142]. For prioritization, we selected somatic SNVs *i*) annotated as P/LP according to ClinVar[143] (updated to July 2023), *ii*) annotated as LoF (frameshift insertions, frameshift deletions, splicing and stop-gain SNVs) according to RefSeq[132] or *iii*) predicted as supporting pathogenic by at least 2 pathogenicity predictors among CADD v1.6 (CADD score ≥ 25.6), REVEL (REVEL score ≥ 0.685) and M-CAP v1.3 (M-CAP score ≥ 0.29), as suggested by published guidelines[144]. To detect putative cancer-causing SNVs we selected variants in 566 genes annotated as TSG, oncogene or both (dual role) according to COSMIC Cancer Gene Census (CGC) database v98[145]. The list of cancer-associated genes of CGC was downloaded at https://cancer.sanger.ac.uk/census.

3.4.2. Focal and numerical CNAs

CNAs are defined as genomic regions of loss or gain of genetic material compared to the ploidy of a karyotype of a cell. Generally, in NBL while numerical CNAs are referred to the loss or a gain of an entire chromosome, focal CNAs involve only a moiety of a chromosome, usually confined on a specific arm[2]/15/24 6:29:00 PM. For each chromosome with a CNA called with Gistic 2.0, CNAs were classified as numerical (or whole-chromosome) or focal (or segmental) if the number of altered bases in a specific direction (gain or loss) was greater or lower than the 85% the length of that chromosome, respectively (see Figure 28). The correlation of the number of focal and numerical CNAs with the risk group was independently assessed for each dataset through a Firth's logistic regression to account for low numbers and zero values in a particular risk group[146]. Resulting standard errors and effect sizes were collected and used to perform an inverse variance based meta-analysis to compute the weighted correlation of each CNA with clinical risk. In detail, let β_T and β_E be the effect sizes of each arm relative to the analysis performed on TARGET and EGA dataset, respectively, and SE_T and SE_E the corresponding standard errors. The weights of the two analyses were calculated as follows:

(1) $w_T = 1/SE_T^2$ and $w_E = 1/SE_E^2$

where w_T and w_E are the weights of TARGET and EGA cohorts, respectively. The weighted effect size and standard error was calculated as follows:

(2) $\beta = (\beta_T w_T + \beta_E w_E)/(w_{T+} w_E)$

and

(3) $SE = 1/(w_{T+} w_E)$

where β and SE are the weighted effect size and standard error, respectively. Finally, the z-score for each arm was computed as follows:

(4) $Z = \beta/SE.$

Given the z-scores the P-values are calculated as follows:

(5) $P = 2\Phi(|-Z|)$

where Φ is a factor that allows to fit z-score in a normal distribution.

3.4.3. SV profiling

Although in Genetics SVs include all the large-scale genomic aberrations including deletions, duplication, amplification, large insertions, inversions (INV) and translocations (TRA), in this dissertation we refer to as SVs only the last two. Gene annotation of SVs was performed with *in-house* R scripts by querying RefSeq database, and included SVs whose one of the two breakpoints mapped within the gene body (intron or exon). For detecting *TERT* rearrangements, we selected SVs in the *TERT* locus, involving 300kb upstream and downstream the gene body[46]. We empirically defined recurrent SVs those occurring in at least 4 and 6 samples in TARGET (2.94%) and EGA (3.33%) datasets (Figure 16).

For downstream analyses, we divided the datasets in two group of samples based on the presence or the absence of at least an SV – referred to as SV group and no-SV group, respectively.

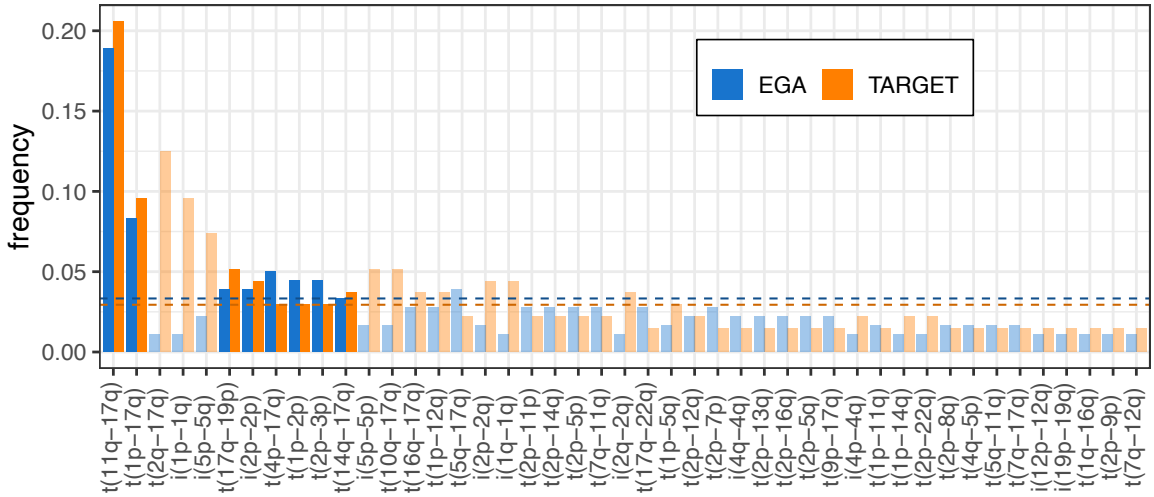


Figure 16. SVs shared between EGA and TARGET tumors.

Bar plot showing the frequency of shared SVs in EGA (blue bars) and TARGET (orange bars). The orange and blue dotted lines represent the TARGET (0.294) and EGA (0.333) thresholds for an SVs to be "recurrent", respectively. Shaded bars indicate SVs below these thresholds.

3.4.4. *Scores of genome instability*

We inferred the status of tumor genome instability by computing – for each sample – 5 scores from somatic SNVs, CNAs and SVs. In detail we assigned to NBL samples a TMB, the number of segmental/numerical CNAs, the number of Large-State Transitions (LST) and the SV burden (Figure 17). Note that in this case the number of numerical and focal CNAs were calculated based on the fraction of bases involved in a CNA on each chromosome arm (≤ 90% for focal and >90% for numerical CNAs), rather than on each chromosome, as described in literature[147]. The computation of LST was performed similarly as described by Taylor et al.[147]: an LST was called where two contiguous segments were larger than 5Mb and their absolute logR difference was greater than 0.304 (corresponding to $\Delta CN = 0.4$ in diploid regions). The SV burden was computed similarly as described by Lopez et al.[88]: for each sample, we counted the sum of all of SVs (deletions, duplications, translocations and inversions) and divided it per 10Mb.

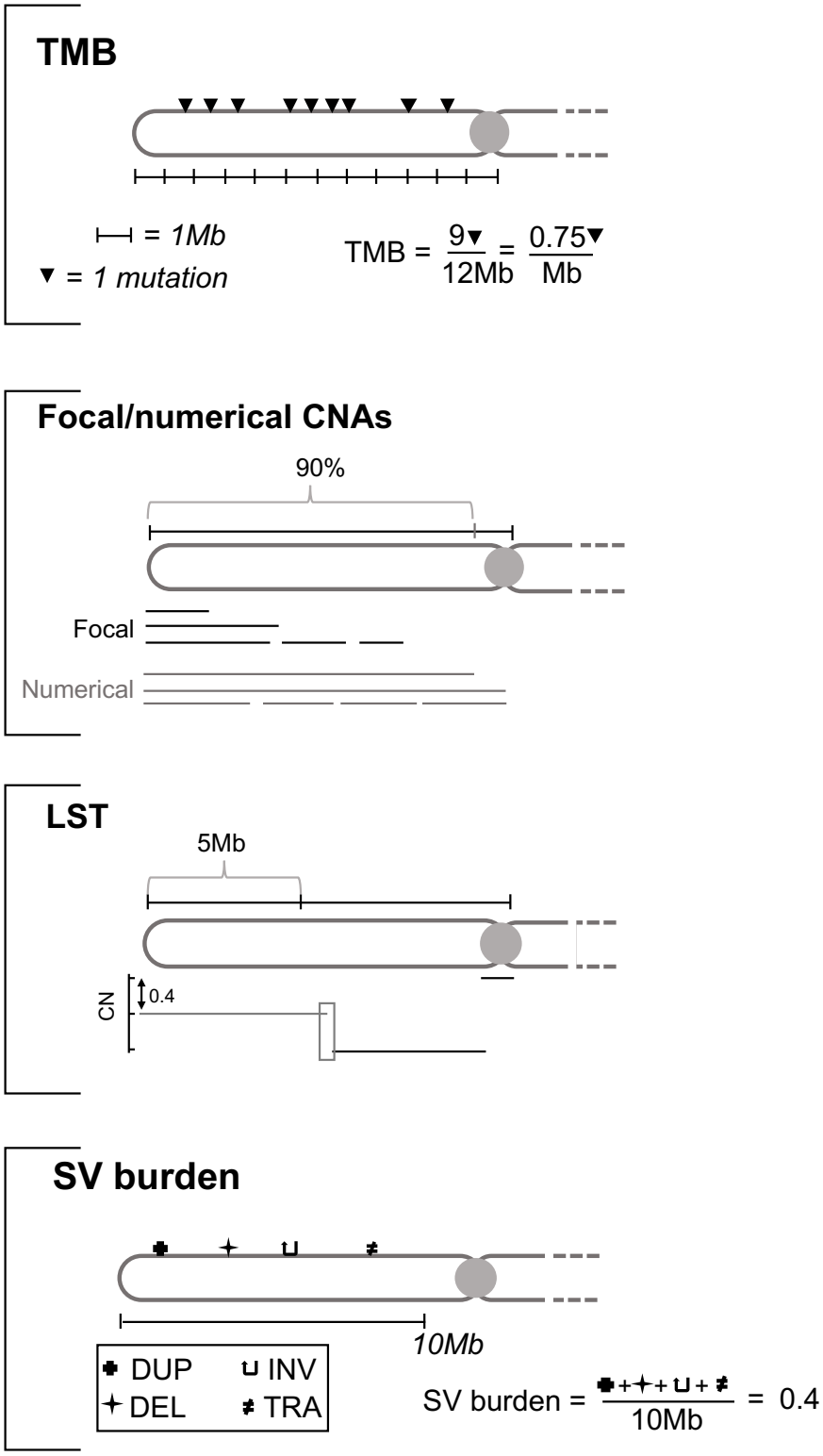


Figure 17. Depiction of genetic instability scores computation.

The figure provides schematic examples of calculation of (from top to bottom) TMB, number of focal and numerical CNAs, LST and SV burden, considering only a chromosome arm. More details are provided in the main text. DUP: duplications; DEL: deletions.

3.4.5. SBS Mutational signatures

Tumor mutational signatures are characteristic patterns of mutations recurrently occurring in different cancers[148]. SBS signatures are a subgroup of tumor mutational signatures relative to the profile of all the nucleotide substitutions of a tumor in their tri-nucleotide context, which can reflect the type of tumor, the exposure to an etiological agent or a specific mutational process[149]. We inferred somatic SBS mutational signatures from high-quality somatic SNVs (see Figure 11) using SigProfilerExtractorR v1.1.13 R package[150] setting a minimum average stability of 0.95. A total of 4 and 8 *de novo* signatures were extracted from TARGET and EGA dataset. Resulting *de novo* SBS (4 and 8 for TARGET and EGA datasets, respectively) were deconvoluted into Cosmic SBS signatures and filtered for cosine similarity (≥ 0.85, see Table 5). active SBSs were defined as SBSs whose mutations contributed to at least the 5% of the total mutations of a sample. Absolute Activity (AA) of each signature was computed for each group as follows

$$\mathrm{AA_i} = \left(\sum_{k=1}^{m} \mathrm{Mut_k}\right) \quad (6)$$

where AAi is the AA of the *i*th signature, $\mathrm{Mut_k}$ the number of mutations of the *i*th signature of the *k*th sample and m the total number of samples in which the *i*th signature was active.

Relative Activity (RA) of each signature was computed for each group as follows:

$$\mathrm{RA_i} = \left(\mathrm{AA_i} / \sum_{j=1}^{n} \mathrm{AA_j}\right) \quad (7)$$

where $\mathrm{RA_i}$ is the relative activity and of the *i*th SBS signature and n the total number of active signatures. A multivariate analysis was performed to assess the distribution of SBS18 in SV and no-SV groups using the *MYCN* amplification, the 17q gain and the expression of 1158 mitochondrial genes (listed in the MitoCarta v2.0 database[151]) as covariates, which are known to correlate with SBS18 in NBL[152].

3.4.6. Differential gene-expression analysis

Prior to perform differential gene-expression (DGE) analysis between SV and no SV groups, we discarded genes with an FPKM value of 0 in more than 10% of samples. DGE was assessed through a logistic regression, which has been shown to deflate the rate of type I errors with respect to other statistical methods[153]. Genes were defined as DE if, in both datasets, they were under or over-expressed in the SV group compared to the no-SV group with a p-value less than 0.01 and an absolute fold change of 1.5. Under and Over expressed genes were given as input to WebGestaltR v0.4.6 R package[154] for Over Representation Analysis (ORA), setting as functional database "Gene Ontology Biological Process (GO:BP) noRedundant" and "genome protein-coding" as reference gene set. We considered as enriched only those GO:BP terms with an FDR<0.1.

3.5. Germline correlation to somatic SV phenotype

Germline QC-filtered SNVs prioritization was performed by selecting only rare (MAF ≤ 1% in gnomAD database v2.1.1 global population[155]) P/LP variants according to ACMG criteria. To perform pathway enrichment analysis, we listed genes that were affected by at least a P/LP variant in SV and in no-SV group, which subsequently served as input for WebGestaltR v0.4.6[154]. Over Representation Analysis (ORA) was performed setting "WikiPathway cancer[156]" as functional database and "genome protein-coding" as reference gene set (Figure 18).

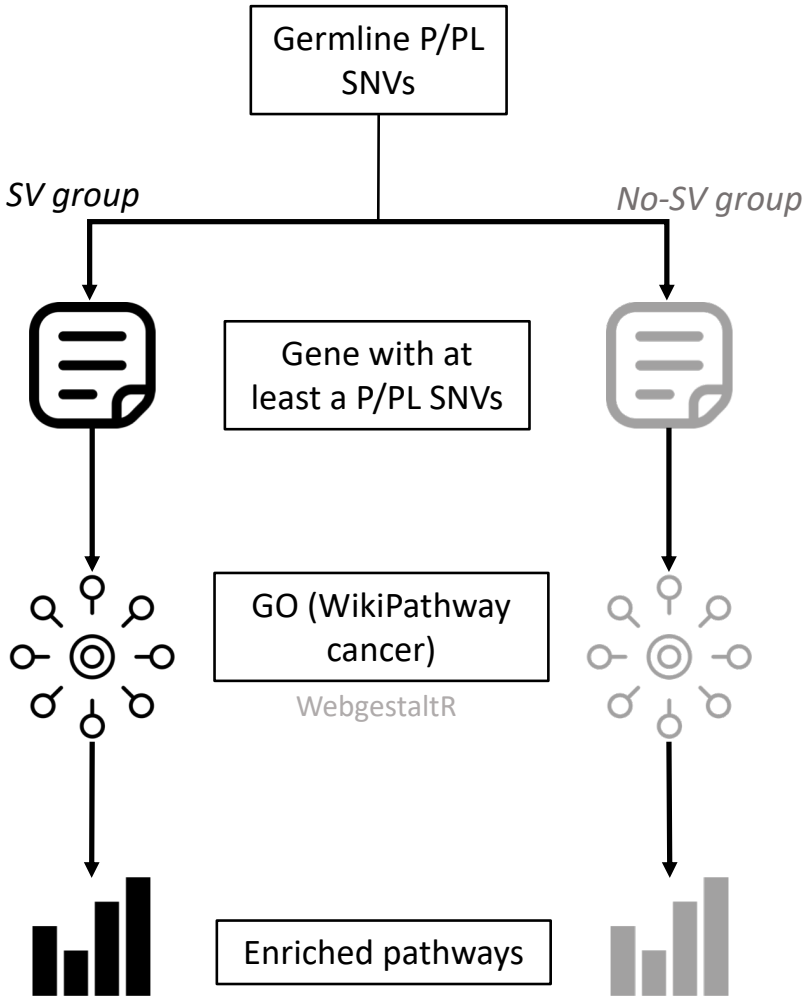


Figure 18. Workflow of pathway enrichment analysis of genes with germline P/PL SNVs.

The scheme above figuratively describes the steps from germline P/LP SNVs selection to the pathway enrichment analysis of genes with at least one P/LP variant in SV and in no-SV group. Further details are provided in the main text.

Ethnicity of NBL samples was deduced using Peddy v0.4.8[157], which takes germline SNVs as input and infers ancestry by the mean of a machine learning model trained on almost 24,000 common (MAF ≥ 4%) bi-allelic SNPs of 2,504 individuals from 1000 Genomes (1000G) database[158] with known ancestry.

3.6. Statistical analysis and graphs

All the statistical analyses were performed with R software v4.2.2[159]. Mann-Whitney U tests, Fisher's exact tests, Chi-squared tests, multivariate and univariate logistic regressions were performed with base R. Firth's corrected logistic regressions were executed with Logistf v1.24.1 R package[160] Cox proportional-hazards models for OS and Event Free Survival (EFS) analyses were performed with Survival v3.5.3 R package[161]. Upset plots, onco-prints and heatmaps were drawn using ComplexHeatmap v2.14.0 Bioconductor package[162]. Other graphs, including box plots, violin plots, bar plots and segments were drawn with ggplot2 v3.4.1 package[163] or with base R.

4. *Results*

4.1. NBL patients characterization

In this dissertation we leveraged publicly available WGS data from 2 NBL studies: one represented by dbGaP study phs000467.v23.p8[9], which included 136 NBL patients from the TARGET database, herein indicated as the TARGET dataset for simplicity; the second included 180 NBL patients from two studies (EGAS00001001308[164] and EGAS00001004349[85]) whose data are deposited in the EGA database, referred to as the EGA dataset.

Firstly, we characterized patients of the two datasets according to their clinical characteristics. Table 4 summarizes the clinical status of NBL patients used in this dissertation. In both datasets, most of the subjects involved were males (TARGET=84; EGA=102). The majority of TARGET patients showed features of poor prognosis (INSS Stage 4, Age ≥ 18 months old, *MYCN* amplification and high-risk), while EGA patients were more balanced in this sense. Of note, 28 (15.6%) of EGA samples missed clinical information.

		TARGET	EGA	TOT
Sex	Male	84 (61.8%)	102 (56.7%)	186 (58.9%)
	Female	52 (32.9%)	78 (43.3%)	130 (41.1%)
INSS Stage	4	106 (77.9%)	88 (48.9%)	194 (61.4%)
	1,2,3,4S	30 (22.4%)	64 (35.6%)	94 (29.7%)
	N/A	0 (0%)	28 (15.6%)	28 (8.9%)
Age at diagnosis	≥18mo	104 (76.5%)	116 (64.4%)	220 (69.6%)
	<18mo	32 (23.5%)	36 (20%)	68 (21.5%)
	N/A	0 (0%)	28 (15.6%)	28 (8.9%)
MYCN status	Amplified	32 (23.5%)	41 (22.8%)	73 (23.1%)
	Not amplified	104 (76.5%)	111 (61.7%)	215 (68%)
	N/A	0 (0%)	28 (15.6%)	28 (8.9%)
Risk	High	107 (78.7%)	98 (54.4%)	205 (64.9%)
	Intermediate/Low	29 (21.3%)	54 (30%)	83 (26.3%)
	N/A	0 (0%)	28 (15.6%)	28 (8.9%)

Table 4. Clinical characteristics of NBL Samples. The table provides a summary of the clinical characteristics of NBL samples of this study.

Some clinical markers showed a strong correlation to each other (Figure 19). In TARGET dataset, the INSS stage 4 strongly co-occurred with the age at diagnosis, with 96.2% of stage 4 patients having an age at diagnosis greater than 18 months (two-tailed Fisher's test $P = 1.6 \times 10^{-22}$). A similar trend was also observed in EGA dataset, where the percentage of stage 4 patients with an age at diagnosis greater than 1.5 years was 87.5% ($P = 2.0 \times 10^{-4}$). In both cohorts, we observed that the occurrence of *MYCN* amplification was rather independent from both stage and age at diagnosis ($P > 0.05$).

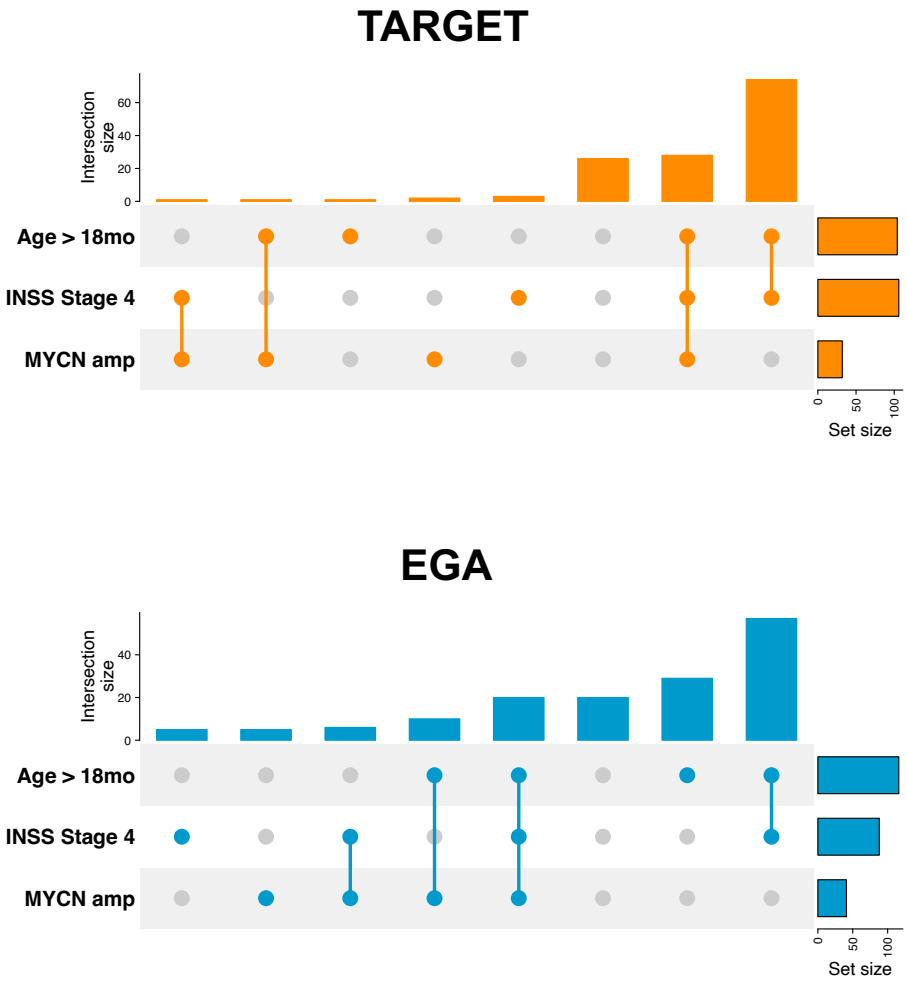
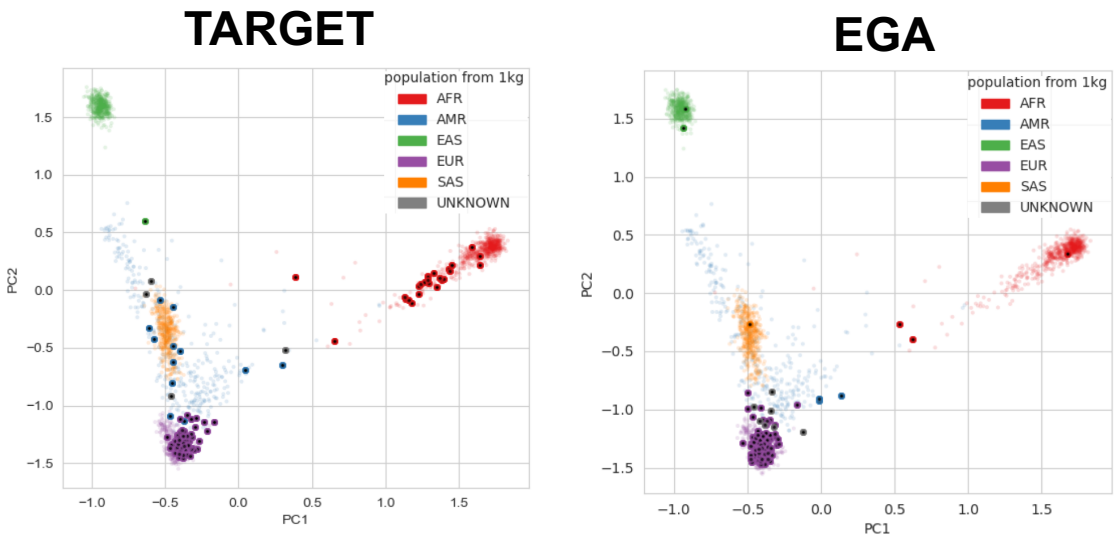


Figure 19. Co-occurrence among clinical markers in the two NBL cohorts.
Upset plot showing the number of samples with age ≥18 months at diagnosis, INSS stage 4 and *MYCN* amplification (bar plots on the right) in TARGET (top) and EGA (bottom) datasets. Bar plot on the top show the number of samples in which the clinical markers – flagged with a dot – co-occur.

We then performed ancestry analysis to account for population stratification. As expected, the great majority of samples belonged to European ethnicity (TARGET=94, 69.1%; EGA=164, 91.1%). A greater extent of population stratification was observed in the TARGET cohort, with a relevant proportion of African (25, 18.4%) and Latino-American (12, 8.8%) ancestries, consistently with the USA origin of this dataset[165] (Figure 20).



Ancestry	*EGA*	*TARGET*
AFR	3	25
AMR	3	12
EAS	2	1
EUR	164	94
SAS	1	0
UNKNOWN	7	4

Figure 20. Inferred ethnicity of NBL samples.

The upper panels show the results of a Principal Component Analysis (PCA) to predict ancestry of TARGET (left) and EGA (right) samples. On the x and the y axes are shown the value of the first and the second PC, respectively. PC values resulting from dimensional reduction of common SNPs of NBL samples were projected on the values of those of 1000G individuals. The table in the lower panel show the number of individuals for each predicted ancestry.

AFR: African; AMR: Latino-American; EAS: East-Asian; EUR: Non-Finnish European; SAS: South-East-Asians.

4.2.Genomic profiling of NBL samples

In order to characterize NBL samples in the two cohorts, we assessed the somatic genomic landscape of the patients in the two datasets. In detail, we analyzed the somatic small DNA variants, comprehensive of SNVs and Indels (hereby referred to as somatic SNVs for simplicity), the CNAs and large-scale genomic rearrangements comprehensive of chromosome inversions and translocations. As one of the aims of this thesis is the characterization of genomic rearrangements in NBL, we put a main focus on the latter.

4.2.1. Somatic SNVs

As detailly described in *Methods*, somatic SNVs of NBL samples have been obtained differently for the TARGET and the EGA datasets. from the publicly available WGS VCF files of TARGET individuals, we selected SNVs that were called in the somatic but not in the corresponding germline sample, while somatic SNVs of EGA patients were called through matched normal-tumor variant calling pipeline. In both datasets, we excluded variants with a VAF lower than 5% and retained only SNVs with an alternative allele depth greater or equal than 5. First of all, we computed the genomic TMB as described in *Methods*. Of note, the TMB distribution was comparable between the datasets (μ_{TARGET} = 1.01 mutations/MB; μ_{EGA} = 1.06 mutations/MB; two-sided Mann-Whitney U test P = 0.2). We then assessed the correlation of TMB with 4 NBL clinical markers (Risk group, INSS stage, Age at diagnosis and *MYCN* status) (Figure 21). As already observed and described in literature[166], in both datasets, we observed a strong correlation of TMB with the high-risk subtype, the stage 4 and the ≥ 18 months at diagnosis patients, but no correlation with the *MYCN* gene status. We then prompted to assess the correlation of the TMB with the OS and EFS probability. TARGET OS and EFS showed not significant correlation with TMB, both in a univariate analysis (P_{OS} = 0.12, OR_{OS} = 1.82; P_{EFS} =0.227, OR_{EFS} = 1.61) and in a multivariate Cox proportional-hazards model adjusted for INSS stage, Age at diagnosis and MYCN status (P_{OS} = 0.7, HR_{OS} = 1.12; P_{EFS} =0.98, OR_{EFS} = 1.01) (Figure 22). Conversely, In EGA dataset TMB was associated to a lesser OS and EFS probability, but only in a univariate model (P_{OS} = 0.017, OR_{OS} = 2.35; P_{EFS} =1.71×10^{-3}, OR_{EFS} = 3.17), but was not predictive of OS or EFS in the multivariate analysis (P_{OS} = 0.27, HR_{OS} = 1.37; P_{EFS} =0.11, OR_{EFS} = 1.46) (Figure 23). Altogether, our data confirm the association of TMB with poor prognosis, although it was not predictive of EFS and OS in any of the two datasets when correcting for covariates, although this result could be influenced by the co-occurrence of poor prognosis markers[167] (see Figure 18).

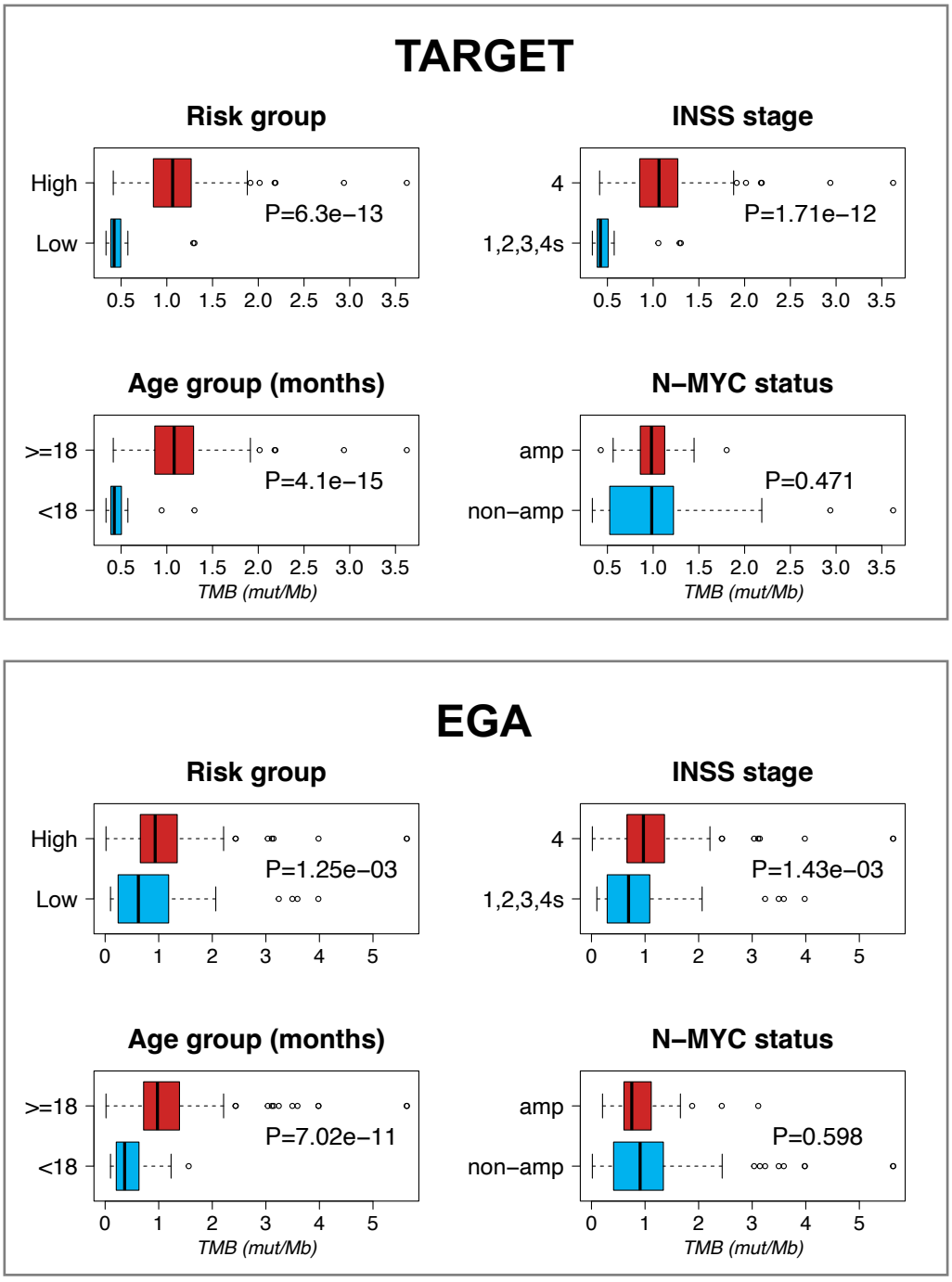


Figure 21. Distribution of TMB across NBL clinical markers.

Box plots showing the distribution of TMB across the clinical markers of NBL (clockwise: Risk groups, INSS stage *MYCN* status and Age group) in TARGET (top) and EGA (bottom) datasets. P-values were computed through a two-sided Mann-Whitney U test.

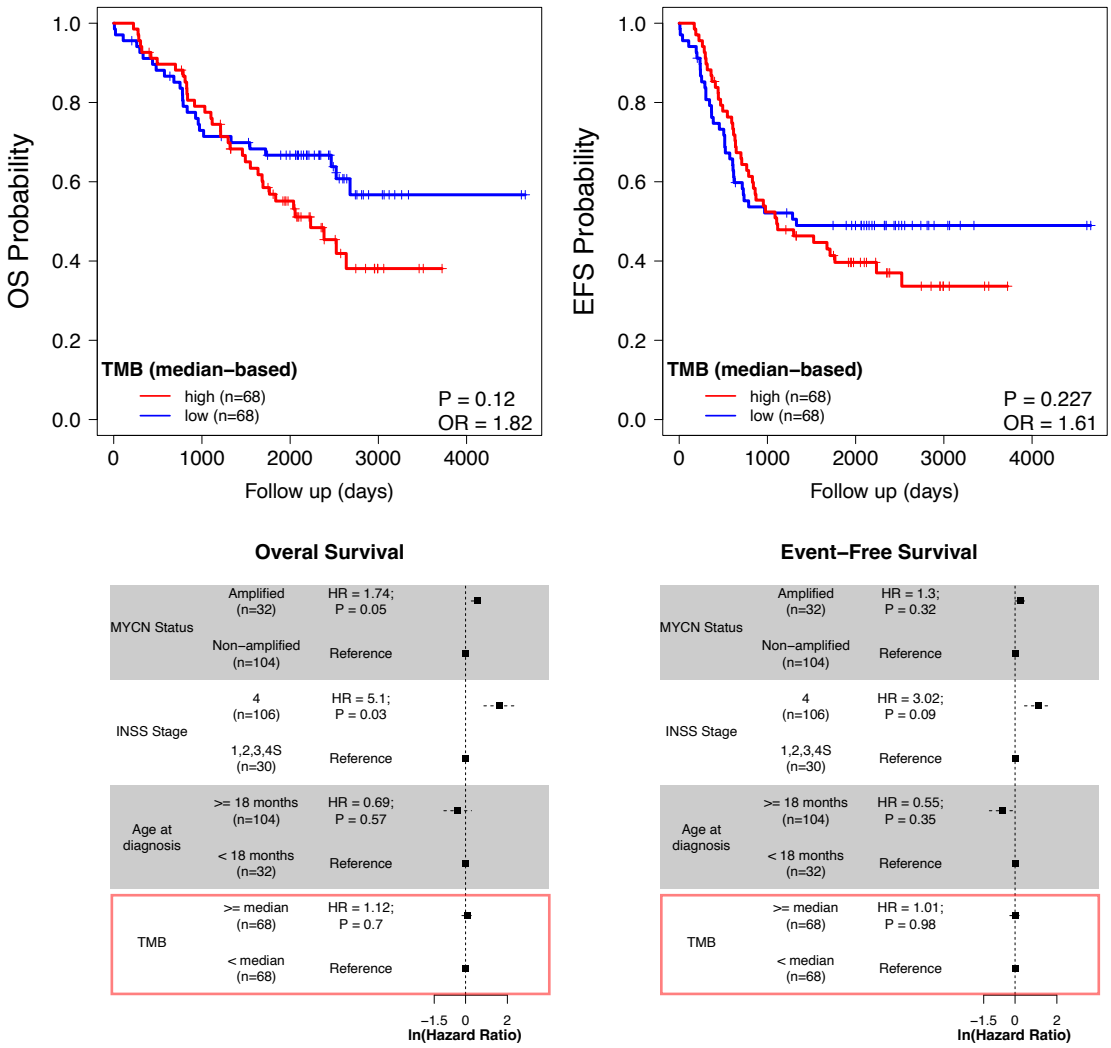


Figure 22. OS and EFS probability across median-divided TMB groups in TARGET samples.
The Kaplan-Meier curves show the OS (left) and EFS (right) probability in the TARGET dataset. The Odds Ratio (OR) and the P-value (P) in the graphs were computed by a two-tailed Fisher's exact test. The tables below the curves show the results of a multivariate Cox proportioned-hazard regression model adjusting for *MYCN* status, INSS stage and the age at diagnosis. Squared dots and dashed lines represent the estimates and the standard errors, respectively. Red box highlights the contribution of the TMB to the OS (left) and EFS (right) probabilities in this model.

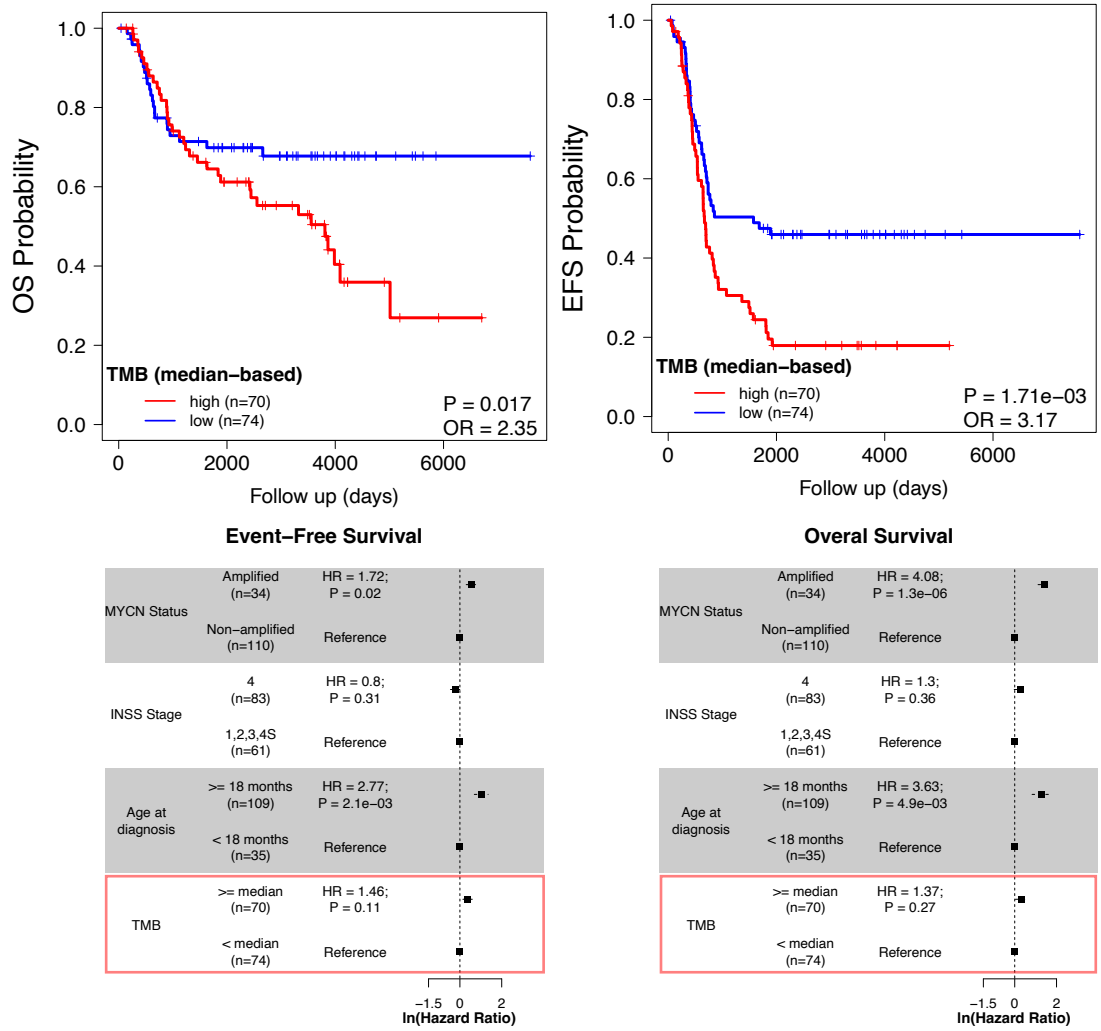


Figure 23. OS and EFS probability across median-divided TMB groups in EGA samples.

The Kaplan-Meier curves show the OS (left) and EFS (right) probability in the EGA dataset. The Odds Ratio (OR) and the P-value (P) in the graphs were computed by a two-tailed Fisher's test. The tables below the curves show the results of a multivariate Cox proportioned-hazard regression model adjusting for MYCN status, INSS stage and the age at diagnosis. Squared dots and dashed lines represent the estimates and the standard errors, respectively. Red box highlights the contribution of the TMB to the OS (left) and EFS (right) probabilities in this model.

Subsequently, we prompted to identify driver mutations in NBL samples. To this end, we first prioritized somatic SNVs, selecting only non-synonymous coding SNVs annotated as pathogenic or likely pathogenic according to ClinVar (updated to July 2023), LoF mutations (frameshift indels, splicing and nonsense SNVs) with a CADD v1.6 score equal or greater than 25 or missense mutations with at least two out of three pathogenicity prediction scores (REVEL, M-CAP and CADD) above established thresholds (see *Methods*), as suggested by published guidelines[144]. We then selected variants falling in 566 genes annotated as TSG, oncogene or both (dual role) according to the COSMIC CGC database v98[145]. We found a total of 53 ($\mu = 0.39$) and 91 ($\mu=0.50$) prioritized somatic SNVs in TARGET and EGA samples, respectively (Figure 24), which was comparable between the two datasets (Mann-Whitney U test's $P = 0.39$). We detected key NBL driver mutations shared between the two dataset, including variants in *ALK* such as R1275Q (5 in TARGET and 9 in EGA), F1174L (3 in TARGET and 4 in EGA), R1275L (2 in TARGET and 1 in EGA) and F1245V (1 in TARGET and 1 in EGA)[168], the *NRAS* Q61K mutation (2 in TARGET and 5 in EGA)[169], the *KRAS* G12D mutation (1 in TARGET and 1 in EGA)[52] and the N546K mutation in *FGFR1* gene (1 in TARGET and 3 in EGA)[52]. A total of 10 genes was affected by point mutations in both datasets (Figure 25). As expected, in both datasets, *ALK* was by far the most frequently mutated gene. Mutations in genes with a certain recurrence and an established role in NBL such as *ATRX*, *NRAS*, *KRAS*, *FGFR1* and *PTPN11* were also observed in both cohorts[52,85,170,171]. *TP53* gene was mutated in 4 deceased patients (2 in TARGET and 2 in EGA samples). This observation is in line with literature, as *TP53* mutations – although rare – are known to be associated with poor prognosis in NBL[172]. We also found mutations in *MYH9* gene in 3 samples (1 in TARGET and 2 in EGA samples). This gene has a well-established role in many adult cancers, where it can act both as an oncogene and as a TSG[173], but its function in NBL have not been hitherto investigated. Other 3 samples carried mutations in the *SKI* proto-oncogene (2 non-sense SNV in EGA and 1 missense SNV in TARGET). As *MYH9,* also *SKI* has a dual role in cancer and is mutated in several adult tumors[174]. However, given its role as inhibitor of TGF-β pathway – which promotes the Epithelial-to-Mesenchymal Transition (EMT) in NBL[175] – a role as TSG is more likely in NBL. Finally, in 2 patients we reported 2 missense variants (one for each dataset) in *ESR1*, a TSG important for neuronal differentiation repressed by *MYCN* in NBL[176]. In order to better elucidate the role of these genes in NBL, we assessed their gene expression on a subset of 89 and 140 samples from TARGET and EGA set, respectively, whose RNA-seq data was available (see *Methods*). In both cohorts, we observed an over-expression of *SKI* in the low to intermediate-risk group, even when correcting for the loss of *SKI* (1p36) loss, a recurrently lost region in high-risk NBL[177] (multivariate logistic regression $P < 0.001$, Figure 26).

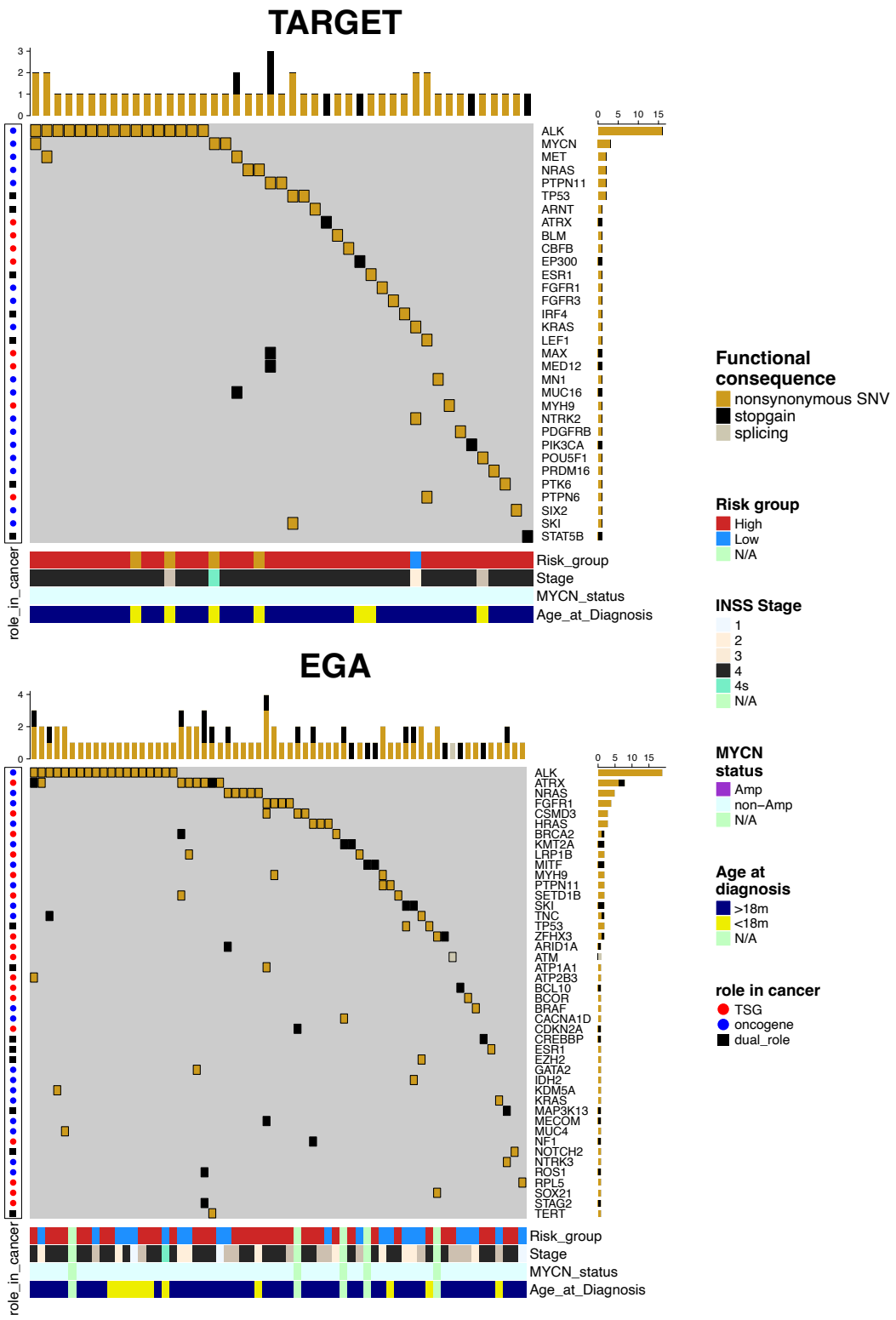


Figure 24. Somatic SNVs in genes annotated as TSGs, oncogenes or both in Cosmic CGC v98 genes in NBL samples.

The two oncoprints show the somatic mutations in CGC v98 genes annotated as TSGs, oncogenes or both. Cells are colored based on the functional consequence of the SNV. Top, right, bottom and left oncoprint annotations show the number of mutations in each sample, the number of mutations in each gene, clinical features of samples (INRG risk group, INSS stage MYCN status and age at diagnosis) and role in cancer (TSG, oncogene or dual role) according to CGC v98, respectively.

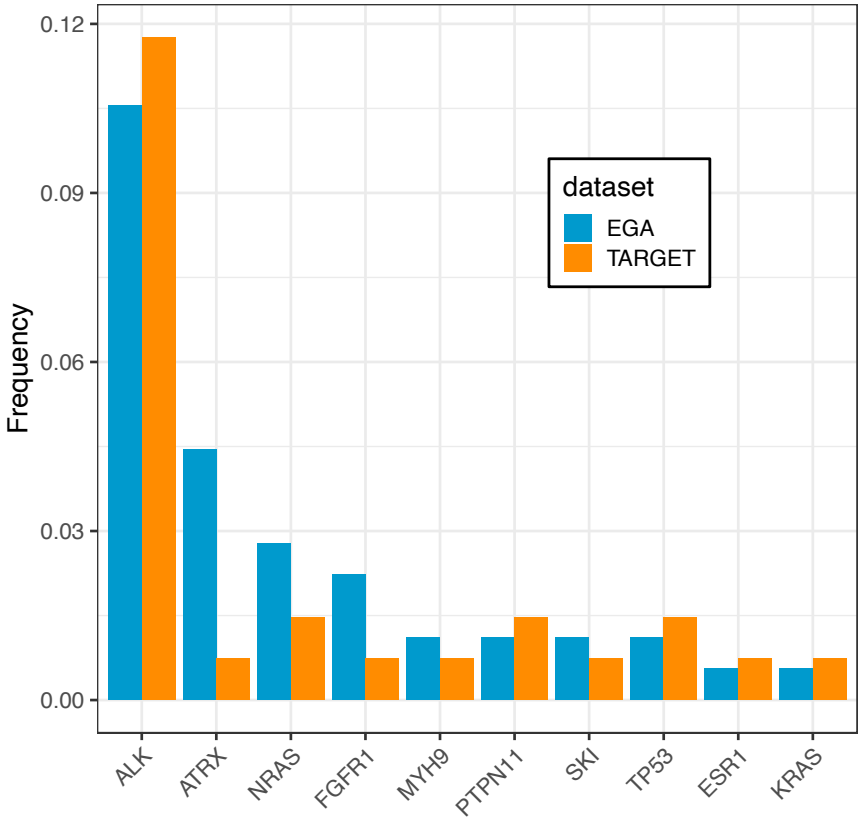


Figure 25. Frequency of genes affected by prioritized somatic SNVs shared between TARGET and EGA.

Bar plot showing the mutation rate in both dataset of 10 genes affected by prioritized SNVs (see main text and Methods) in at least one patient of both datasets.

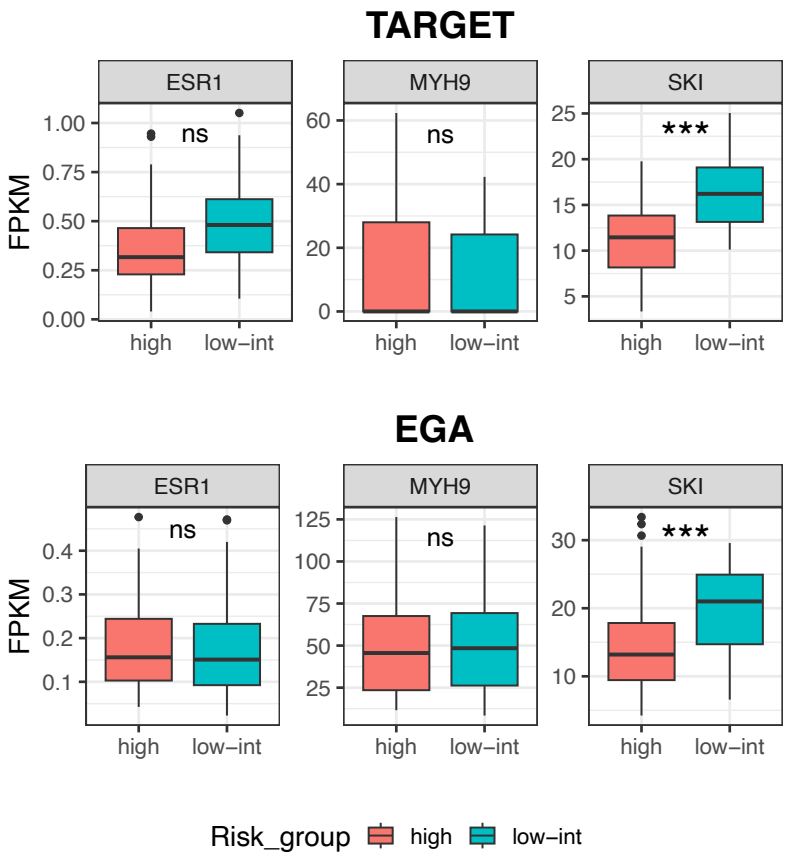


Figure 26. SKI proto-oncogene expression is increased in low to intermediate risk group.

Boxplot showing the distribution of *ESR1*, *MYH9* and *SKI* expression (expressed as Fragment per Kilobase per Million or FPKM) across risk groups (high vs low to intermediate) in TARGET (top) and EGA (bottom) cohorts. *SKI* is the only differentially expressed gene, being over-expressed in the low to intermediate (low-int) risk group, even when correcting for 1p loss. P-values were achieved through a multivariate logistic regression. ***: p<0.001; ns: P>0.05.

4.2.2. Copy Number profiling

As stated in *Introduction*, NBL is characterized by typical and recurrent CNAs, many of which are directly linked to patients' outcome[57]. In order to profile the CN status of NBL samples, we performed a CN analysis on somatic tissues and evaluated the correlation of CNAs with clinical status. Briefly, for the TARGET samples, we used the already publicly available files of somatic relative coverage (see *Methods*) of genomic windows of 2Kb, where to each window corresponded the natural logarithm (ln) of copy number ratio (hereby referred as logR) under a diploid model and normalized for CG-content and mappability. For samples in EGA dataset, from BAM files we *i*) computed the coverage of normal and tumor samples in windows of 2Kb, and *ii*) normalized for CG-content and mappability. The logR of tumor samples was computed using the normalized-coverage of its respective normal counterpart. Using as input the relative and normalized coverage files, we applied a segmentation algorithm to retrieve regions of contiguous CN status whose boundaries and statistical significance was assessed using Gistic v2.0[136]. We defined as "gain" and "loss" segments with logR $\geq$ 0.223 and logR $\leq$ -0.288, corresponding to CN $\geq$ 2.5 and CN $\leq$ 1.5 in diploid regions, respectively. The resulting CN profile of somatic samples of the two datasets was comparable (cosine similarity = 0.95) and the frequency of CNAs consistent with literature data[178] (Figure 27).

As it is well known that aneuploid (or numerical) and segmental (or focal) CNAs are generally associated to a better and a worse prognosis, respectively[57], we checked the distribution of aneuploid and focal CNAs and evaluated their correlation with clinics. We defined numerical (also referred to as whole-chromosome) or focal CNAs whether more or less than 85% of bases of a chromosome was involved in the same type of CNA (loss or gain), respectively (Figure 28). As expected, patients with poor prognosis markers were enriched in segmental CNAs, whilst aneuploidies associated with good prognosis[179] (Figure 28). We then assessed the presence of segmental or whole-arm CNAs on chromosome arms of each NBL samples. in order to exclude false positive observations, we considered only CNAs whose regions of gain or loss overlapped between TARGET and EGA (Figure 29). A total of 15 CNAs was shared between the two datasets, whose 8 were defined as gain or amplification (defined as CN $\geq$ 8 in diploid regions) and 7 as loss (1p loss, 1q gain, 2p gain, 3p loss, 4p loss, 5p gain, 7q gain, 9p loss, 11q gain, 11q loss, 12q gain, 14q gain, 16q loss, 17q gain and 19p loss) (Figures 29 and 30). These arms are characterized by recurrent NBL CNAs, both focal and numerical[178].

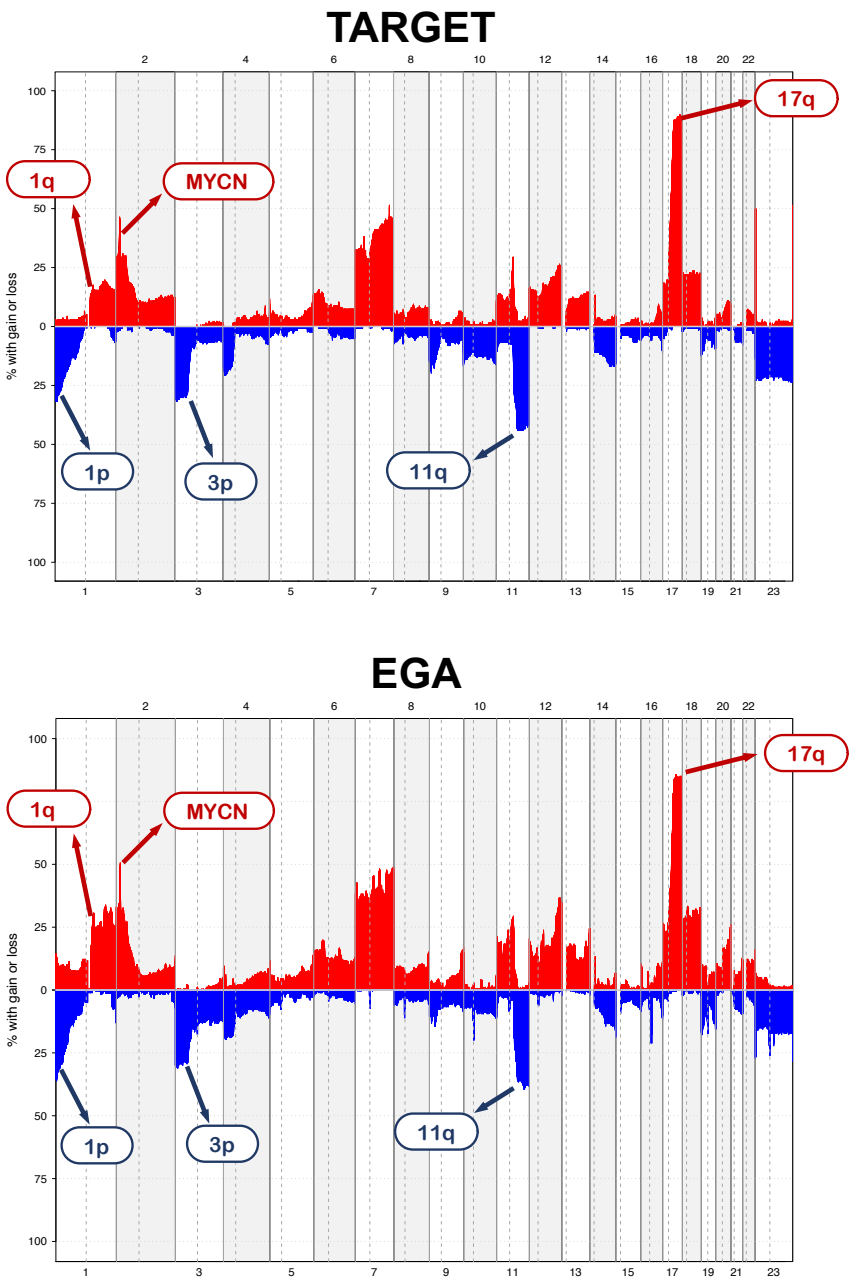


Figure 27. Copy number profiling of TARGET and EGA NBL samples

CN profile of TARGET (top) and EGA (bottom) NBL tumor samples. Y and x-axes represent the percentage of samples with gain (red) or loss (blue) and the chromosomes (23 stands for chromosome X), sorted by genomic position. Recurrent NBL loss and gain are shown and labelled in blue and red, respectively.

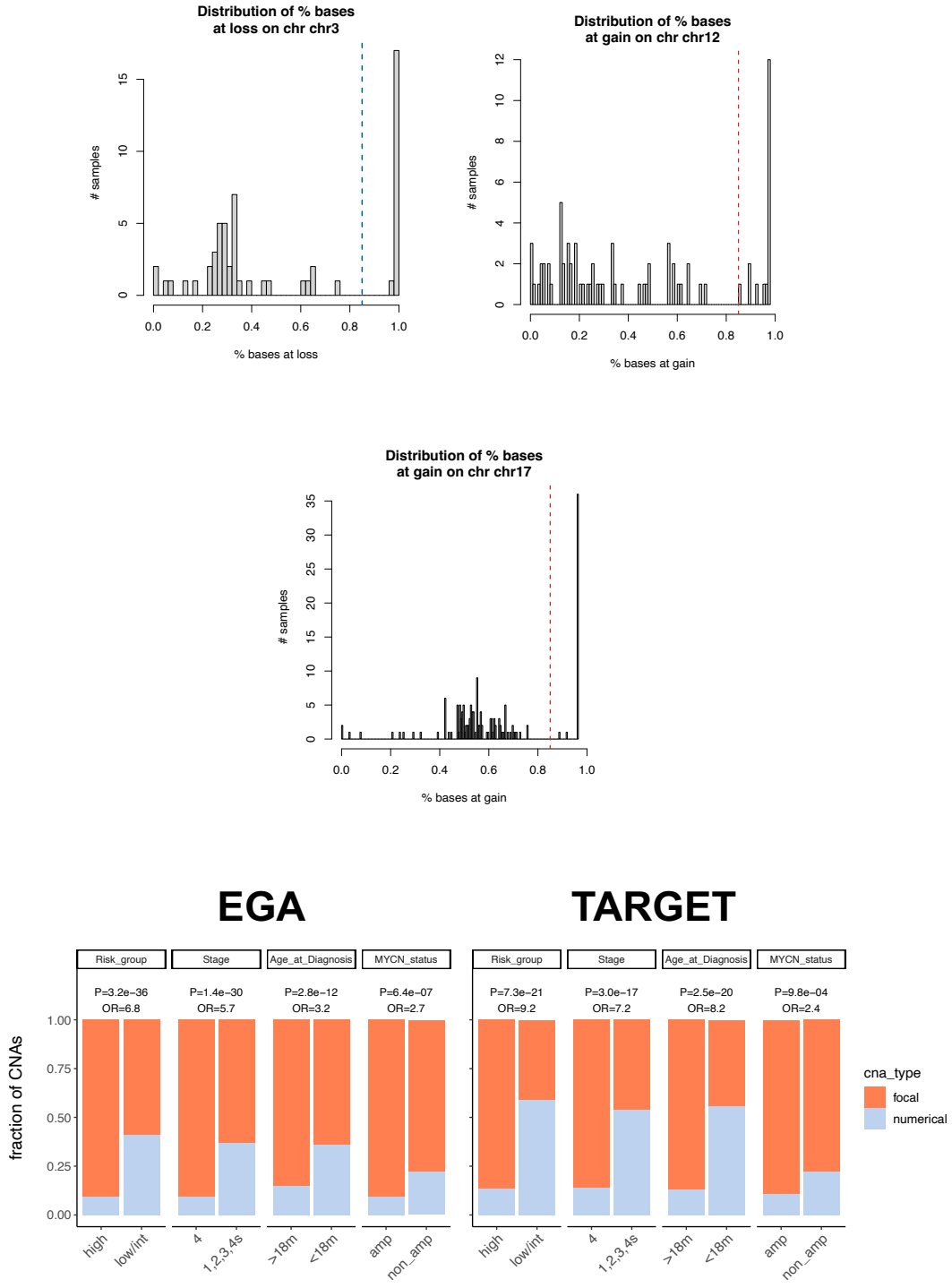


Figure 28. Numerical CNAs are associated with good prognosis, conversely to focal CNAs.

The histograms (top) show the distribution of the percentage of altered bases (gain or loss) for three chromosomes (chr3, chr12 and chr17) across the EGA samples, taken as an example. Arms with ≥ 85% of altered bases (red and blue dashed lines for gain and loss, respectively) were labeled as affected by numerical (or aneuploid) CNAs, otherwise as segmental CNAs. The bottom bar plots show the frequency of aneuploid CNAs (light blue) and segmental CNAs (orange) across NBL clinical markers (Risk, INSS stage, age at diagnosis and *MYCN* status. P-values and OR were assessed through a two-sided Fisher's exact test.

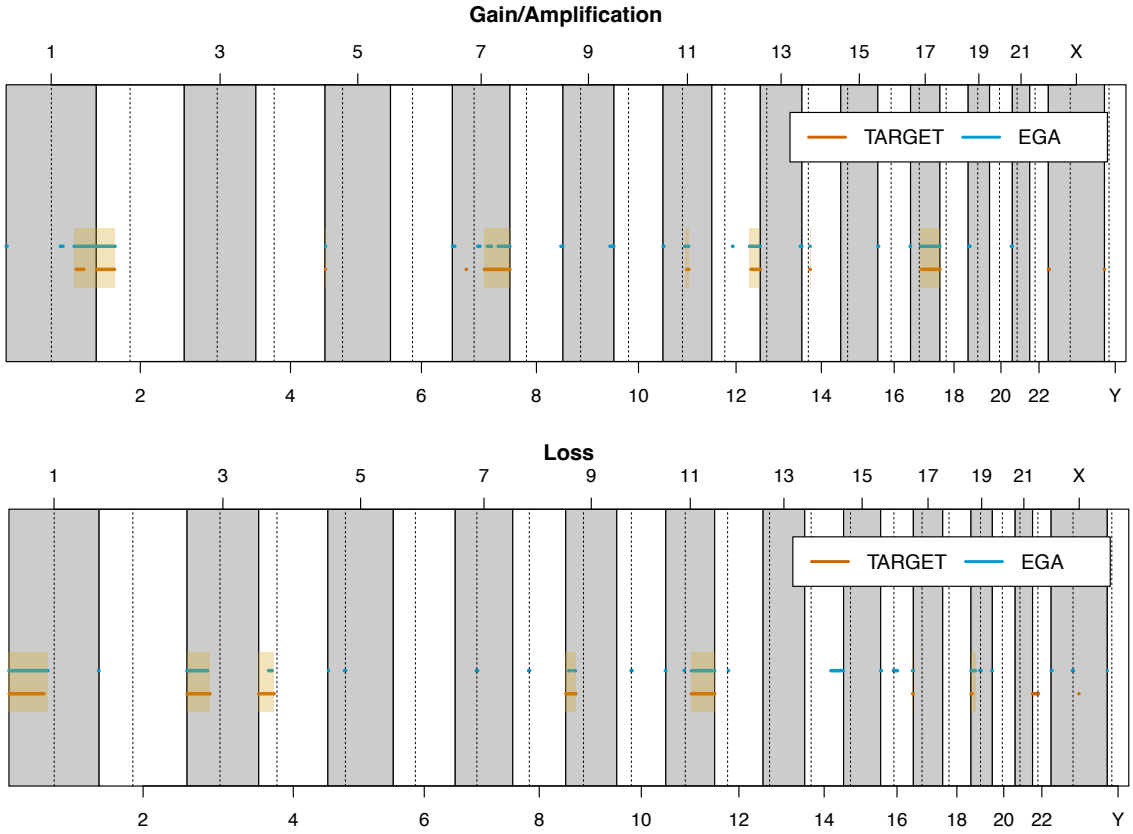


Figure 29. Regions of gain/amplification and loss across the genome in the two datasets

The figure shows the significant gain/amplification (top) and loss (bottom) regions found with Gistic2.0 in TARGET (orange segments) and EGA (blue segments) datasets. The golden-shaded rectangles highlight the 15 shared regions, 8 of gain/amplification (mapping on 1q, 2p, 5p, 7q, 11q, 12q, 14q, 17q) and 7 of loss (on 1p, 3p, 4p, 9p, 11q, 16q, 19p).

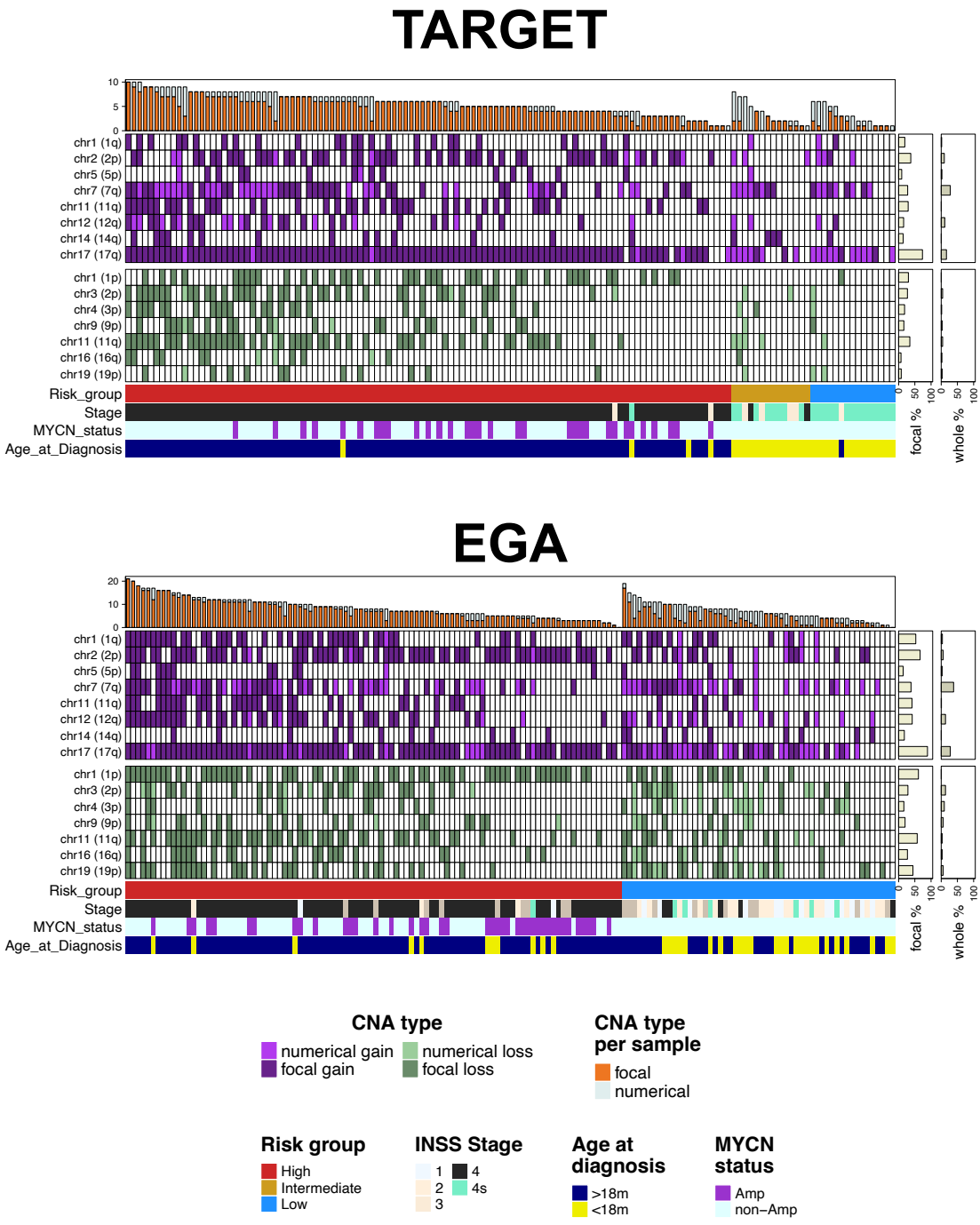


Figure 30. Landscape of focal and numerical CNAs of NBL samples.

Heatmaps showing the numerical and the focal CNAs across NBL samples of TARGET (top) and EGA (bottom) datasets, alongside the percentage of samples with focal or numerical CNAs (right), clinical characteristics of samples (bottom) and the number of focal (orange) and numerical (steely gray) CNAs per sample (top). For graphical representation, samples with unknown clinical features were removed from the EGA heatmap. Chromosome arms in brackets show the location of the respective focal CNAs.

To assess the correlation with clinics, each numerical and focal CNA was tested for enrichment in the two risk groups (high and low to intermediate). In detail, we performed a Firth's logistic regression to account for CNAs imbalance in the two risk groups and for the absence of some CNAs in a particular risk group (for instance, focal 11q gain was absent in low to intermediate tumors in TARGET dataset, see Figure 30)[146]. To increase the statistical power of the analysis, we finally performed an inverse variance-weighted meta-analysis using standard error and estimate parameters of the analyses carried out on single datasets[180] (Figure 31).

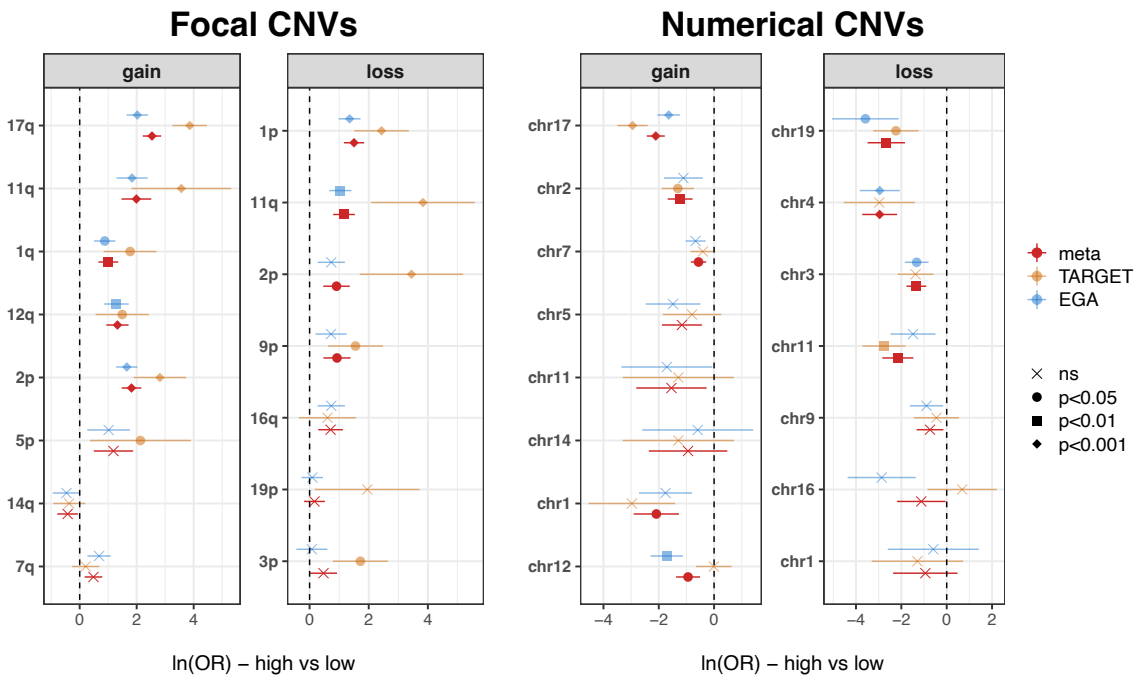


Figure 31. Correlation between focal and numerical CNAs and INRG risk groups.
Forest plot showing the results of a Firth's corrected logistic regression between focal (left) and numerical (right) CNAs and risk groups (high versus low to intermediate). P-value ranges are depicted in the legend on the right.

Consistently with literature data, we found that the high-risk subtype was enriched in focal gain/amplifications on 1q, 2p, 11q, 12p and 17q, as well as focal deletions on 1p, 3p, 4p, 9p and 11q[15,181–184]. On the other hand, low-intermediate risk group was enriched in numerical gain on chromosomes 1, 2, 7, 11 and 17, and in numerical loss on chromosomes 3, 4, 11 and 19. While it is known that aneuploidies of chromosomes 3, 4, 7 and 17 correlate to a better outcome[181,182,185], for the other numerical CNAs (chr1 gain, chr2 gain, chr11 loss, chr12 gain and chr19 loss) no association has been hitherto established. The frequency of these numerical alterations across risk groups is shown in Figure 32.

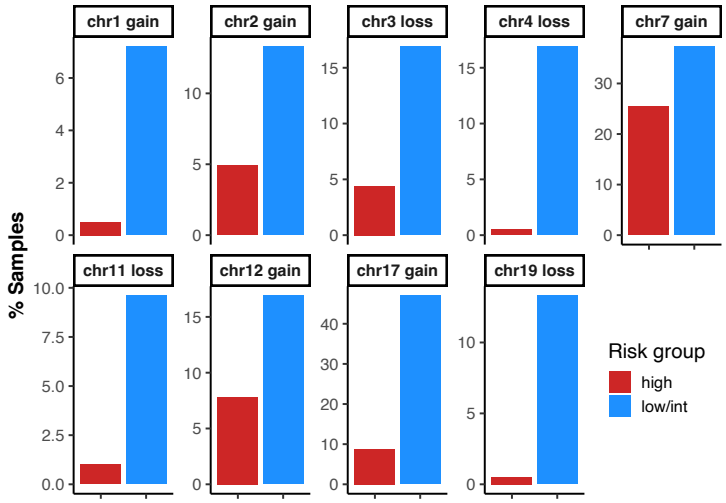


Figure 32. Frequency of low to intermediate-associated numerical CNAs across risk groups.
Bar plot showing numerical gain of chromosomes 1, 2, 7, 12 and 12 and numerical loss of chromosomes 3, 4, 11 and 19, which correlated with low to intermediate (low/int) risk group in a meta-analysis (see Figure 31). Note that frequencies are relative to the combination of the two datasets.

Interestingly, the presence of at least one of these aneuploidies predicted survival, both alone and in a multivariate model with age at diagnosis, INSS stage, *MYCN* status and the dataset belonging as covariates (Figure 33). Altogether, these data suggest that the presence of specific numerical CNAs predict good outcome, and therefore that their detection might be implemented for a better clinical stratification of NBL.

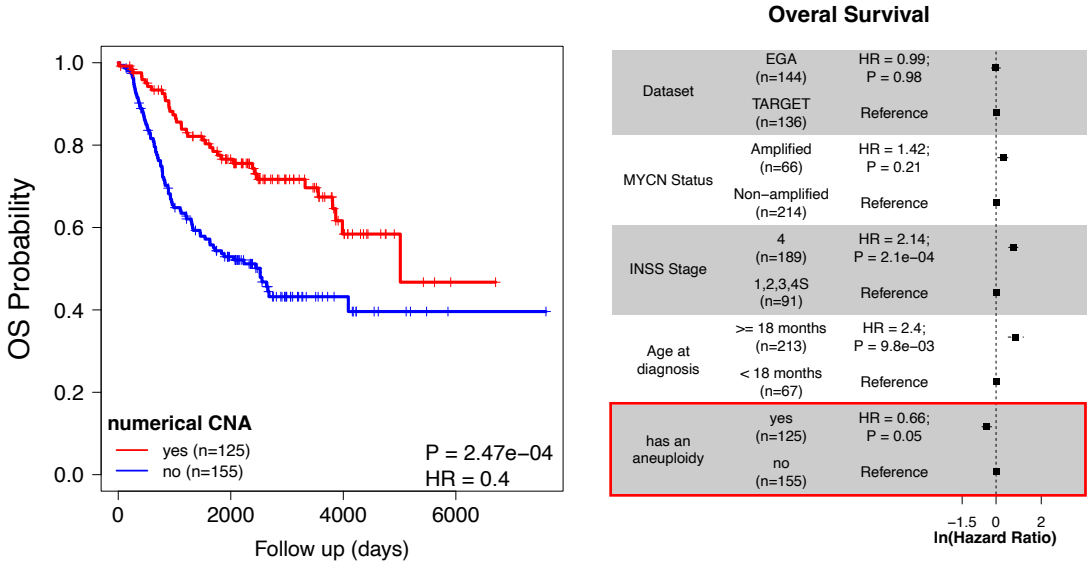


Figure 33. Aneuploidy of at least one of chromosome 1, 2, 3, 4, 7, 11, 12, 17 and 19 predicts survival in samples from EGA and TARGET datasets.

The Kaplan-Meier graph on the left shows the OS probability of NBL samples divided according the presence (red line) or the absence (blue line) of at least one of the numerical CNAs listed above in a univariate model, while the table on the right shows the results of a multivariate Cox proportioned-hazard regression model adjusting for *MYCN* status, INSS stage and the age at diagnosis. Squared dots and dashed lines represent the estimates and the standard errors, respectively. Red box highlights the contribution of the presence of these numerical CNAs to the OS probability in this model.

4.2.3. Genomic Rearrangement profiling

Genomic rearrangements are defined as mutations that change the karyotypic profile of a cell. They can be divided in 4 group of SVs according to their effect on karyotype: deletions, duplications, inversion and translocations. The latter can be further divided into balanced or unbalanced translocations, based on if they are coupled to a gain or loss of genomic material or are CN neutral, respectively. While the presence and clinical and biological implications of large-scale deletions and duplications have been extensively investigated in NBL – and also dissected in the previous paragraph – only few studies have been carried out so far to unravel the role of inversions and translocations[88,93,186]. To shed light on the role of inversions and translocations (hereby collectively referred to as SVs for simplicity) in NBL we analyzed frequency, clinical implications and biological role of SVs in samples from TARGET and EGA cohorts. Concerning the first dataset, we retrieved SVs from publicly available files of "High confidence SV calls" (see *Methods*) and selected only inversions, translocations and complex variants that involved two different chromosomes (labeled as translocations) or CN neutral complex variants on the same chromosome. Finally, we discarded SVs mapping in short arm of acrocentric chromosomes (13p, 14p, 15p, 21p and 22p) and on Y chromosome due to the low mappability of these regions. SVs in EGA dataset was called using Manta[137], low quality SVs discarded and afterward we selected only SVs supported by a conspicuous number of reads (see *Methods*). The distribution of SVs in the two datasets was comparable (Figure 34). In TARGET dataset we found a total of 469 (μ=3.45) translocations and 296 inversions (μ=2.35), for a total of 765 (μ=5.80) SVs; In EGA dataset we found a total of 732 (μ=4.07) translocations and 466 inversions (μ=2.59), for a total of 1,198 (μ=6.65) SVs. As expected, in both datasets high risk tumors were enriched in SVs (Figure 34). Moreover, survival analysis showed that the presence of a least an SV (N_{TARGET} = 112 (82.35%), N_{EGA} = 141 (78.3%)) was able to predict OS and EFS survival in a multivariate Cox proportioned-hazard regression model in TARGET dataset, but only in a univariate analysis in EGA cohort (Figures 35 and 36), although the result could be influenced by the co-linearity with clinical markers like the age at diagnosis and the INSS stage 4. *TERT* rearrangements – defined as inversions/translocations with one of the two breakpoints mapping 300kb upstream and downstream *TERT* (see *Methods*) – were detected in 15.9% and 9.2% of TARGET and EGA samples, respectively (Figure 37).

We then prompted to assess the frequency of translocations and inversions to identify novel low-frequency recurrent SVs. As expected, the three most frequent SVs were t(11q-17q), t(1p-17q), and *TERT* rearrangements, alterations well-characterized in NBL[46,88,186]. This analysis highlighted other novel unreported SVs, which were found in at least 4 and 6 samples in TARGET and EGA datasets, respectively, corresponding to a frequency of about