FABIO PIRES DE SOUZA SANTOS

Meta-análise de dados de sequenciamento completo de exoma em pacientes com neoplasias mieloproliferativas e mielodisplásicas/mieloproliferativas Filadélfia-negativo

Tese apresentada à Faculdade de Medicina da Universidade de São Paulo para obtenção do título de Doutor em Ciências

Programa de Oncologia

Orientador: Prof. Dr. Israel Bendit

São Paulo

2019

Dados Internacionais de Catalogação na Publicação (CIP)

Preparada pela Biblioteca da
Faculdade de Medicina da Universidade de São Paulo

Santos, Fabio Pires de Souza
Meta-análise de dados de sequenciamento completo de exoma em pacientes com neoplasias mieloproliferativas e mielodisplásicas / mieloproliferativas Filadélfia-negativo / Fabio Pires de Souza Santos. -- São Paulo, 2019.
Tese(doutorado)--Faculdade de Medicina da Universidade de São Paulo.
Programa de Oncologia.
Orientador: Israel Bendit.

Descritores: 1.Leucemia mielóide 2.Policitemia 3.Trombocitemia 4.Mutação genética 5.Mielofibrose 6.Exoma

USP/FM/DBD-427/19

Responsável: Erinalva da Conceição Batista, CRB-8 6755

FABIO PIRES DE SOUZA SANTOS

Meta-analysis of whole exome sequencing data in patients with Philadelphia-negative myeloproliferative and myelodysplastic/myeloproliferative neoplasms

Thesis presented to the Faculdade de Medicina, Universidade de São Paulo to obtain the degree of Doctor in Science

Graduate Program in Oncology

Advisor: Prof. Dr. Israel Bendit

São Paulo

2019

Santos FPS. *Meta-analysis of whole exome sequencing data in patients with Philadelphia-negative myeloproliferative and myelodysplastic/myeloproliferative neoplasms* [thesis]. São Paulo: “Faculdade de Medicina, Universidade de São Paulo”; 2019.

Aprovado em:

Banca Examinadora

Prof. Dr. _____
Instituição: _____
Julgamento: _____

Prof. Dr. _____
Instituição: _____
Julgamento: _____

Prof. Dr. _____
Instituição: _____
Julgamento: _____

Para Gui e Laurinha

AGRADECIMENTOS

A todos os pacientes que participaram do projeto de sequenciamento de neoplasias mieloides (Projeto LMA Brasil).

Ao Prof. Dr. Israel Bendit, por toda atenção, orientação e amizade, durante esses anos, incentivando e apoiando minhas escolhas.

Aos amigos da Hematologia do Hospital Israelita Albert Einstein por terem contribuído com a inclusão de pacientes durante esse estudo, em especial Guilherme Fleury Perini, Ricardo Helman, Iracema Esteves e Breno Gusmão.

Às pessoas que de alguma forma ajudaram na aprovação e condução do Projeto LMA Brasil, principalmente aos meus amigos: Paulo Campregher, Renato Puga, Tarcila Datoguia, Raquel Paiva, Isabel Bello, Sandra Nakashima, Giulliana Rangel, Bianca Lisboa e Welbert Pereira.

Aos professores do Leukemia Department, University of Texas M.D. Anderson Cancer Center, principalmente ao Dr. Jorge Cortes, que me recebeu como *fellow* em 2008, e ao Dr. Srdan Verstovsek, que me conduziu para o mundo das neoplasias mieloproliferativas Filadélfia-negativo.

Ao Dr. Nelson Hamerschlak, por todos os anos de convivência no Hospital Israelita Albert Einstein, que muito contribuíram para o meu crescimento profissional.

À minha família, meus irmãos Ricardo e Julia, meus pais, José Carlos e Zuleima, pelo amor e apoio em todas as etapas da minha vida.

À minha amada esposa Erika, por todo o carinho e amor nesses anos, e por ter me ajudado e dado suporte nessa jornada.

Aos meus filhos Guilherme e Laura, grandes amores da minha vida, fonte de coragem e inspiração.

'It is possible that these various conditions – "myeloproliferative disorders" – are all somewhat variable manifestations of proliferative activity of the bone marrow cells, perhaps due to a hitherto undiscovered stimulus. This may affect the marrow cells diffusely of irregularly with the result that various syndromes, either clear-cut or transitional, result' (DAMESHEK, 1951)

RESUMO

Santos FPS. *Meta-análise de dados de sequenciamento completo de exoma em pacientes com neoplasias mieloproliferativas e mielodisplásicas/mieloproliferativas Filadélfia-negativo* [tese]. São Paulo: “Faculdade de Medicina, Universidade de São Paulo”; 2019.

As neoplasias mieloproliferativas (NMP) e mielodisplásicas / mieloproliferativas (SMD/NMP) Filadélfia-negativo (‘Philadelphia’ [Ph]-negativo) são neoplasias mieloides crônicas que apresentam diversas mutações oncogênicas. Estudos recentes, utilizando técnicas de sequenciamento de última geração, descreveram as alterações mais comumente encontradas nessas neoplasias. A hipótese do presente estudo é que a análise dos dados de sequenciamento genômico de uma grande coorte destes pacientes poderia revelar novos oncogenes e as principais diferenças no perfil molecular destas neoplasias. Para tanto, foram analisados dados de sequenciamento de exoma total (WES; ‘Whole Exome Sequencing’) de 403 pacientes com diagnóstico de NMP (N=303) e SMD/NMP (N=100) Ph-negativo. A coorte incluía 124 pacientes brasileiros que realizaram a coleta de amostra e sequenciamento, cujos dados foram combinados com dados de 279 pacientes extraídos de estudos publicados na literatura médica. Testes estatísticos foram utilizados para determinar os genes mais frequentemente mutados nestas doenças, principais padrões de mutação, combinação de mutações entre genes e análise de heterogeneidade clonal. Modelos estatísticos de regressão logística e de Cox foram desenvolvidos para classificação e determinação da sobrevida dos pacientes com base em alterações genéticas. Foram identificados 54 oncogenes para estas doenças com base em dados de WES, incluindo 17 genes nunca previamente descritos como sendo oncogenes nestas neoplasias. A maioria dos 54 genes pertence a uma de 7 vias biológicas distintas, com papéis relevantes na oncogênese destas doenças. Dezenove genes apresentaram distribuição diferente entre NMPs e SMD/NMP, sugerindo que eles contribuem para o fenótipo da doença. Analisando as combinações de genes, cinco pares de genes e 6 tríades de genes, tem-se uma prevalência distinta entre NMPs e SMD/NMPs. As principais vias biológicas alteradas também apresentam distribuição distinta entre as diferentes doenças, assim como nos genes encontrados no topo da hierarquia clonal. Um modelo de regressão logística, baseado apenas nas alterações genéticas, conseguiu determinar, com elevada acurácia, o diagnóstico dos pacientes. Mutações do gene *NRAS* ou em genes de *splicing* de mRNA, foram fatores independentes associados com menor sobrevida. Pacientes com NMPs e SMD/NMPs apresentaram perfis relacionados, porém distintos, de mutações que auxiliam no diagnóstico diferencial e na estratificação prognóstica. Estudos futuros, empregando-se algoritmos de aprendizado por máquinas, poderão aperfeiçoar esses resultados e levar a uma classificação molecular destas doenças.

Descritores: Leucemia mielóide; Policitemia; Trombocitemia; Mutação genética; Mielofibrose; Exoma.

ABSTRACT

Santos FPS. *Meta-analysis of whole exome sequencing data in patients with Philadelphia-negative myeloproliferative and myelodysplastic/myeloproliferative neoplasms* [thesis]. São Paulo: “Faculdade de Medicina, Universidade de São Paulo”; 2019.

Philadelphia-negative (Ph-negative) myeloproliferative neoplasms (MPN) and myelodysplastic/myeloproliferative neoplasms (MDS/MPN) are related chronic myeloid disorders that present with several distinct oncogenic mutations. Recent studies utilizing next-generation sequencing technology have described the most frequent genetic abnormalities in these disorders. The hypothesis of the present study is that the analysis of a large cohort of such patients could reveal novel oncogenic drivers and the key differences in the molecular profile of these neoplasms. To this end, whole exome sequencing (WES) data from 403 patients with either MPNs (N=303) or MDS/MPN was analyzed. The cohort included 124 Brazilian patients who had sample collection and sequencing, and whose data was combined with data from 279 patients collected from studies published in the medical literature. Statistical tests were used to determine the most frequently mutated genes in these disorders, patterns of mutations, combinatorial mutational analysis between genes and clonal heterogeneity analysis. Logistic regression and proportional Cox Hazards model were fitted to classify and estimate survival based on genomic features. A total of 54 oncogenes were identified using WES data, including 17 genes not previously reported as being mutated in these neoplasms. Most of the 54 genes belonged to one of 7 distinct biological groups with relevant roles in the oncogenesis of these disorders. Nineteen genes had different distributions among MPNs and MDS/MPNs. Analyzing gene combinations, there were 5 gene pairs and 6 gene triads that had a distinct prevalence among MPNs and MDS/MPNs. The main biological pathways also had different distribution among the diseases, as well genes that presented within the top of the clonal hierarchy. A logistic regression model based solely on genetic abnormalities could classify with high accuracy patient’s diagnosis, and mutations of gene *NRAS* and genes associated with mRNA splicing were independent predictors of decreased survival. Patients with MPNs and MDS/MPNs present with related but distinct mutational profiles that can be used in differential diagnosis and prognostic stratification. Future studies employing machine learning algorithms can improve on these results and lead to a molecular classification of these disorders.

Descriptors: Leukemia, myeloid; Polycythemia; Thrombocythemia; Genetic mutation; Myelofibrosis; Exome.

Summary

Figures

Tables

-

INTRODUCTION

1. Introduction

1.1 Chronic Myeloid neoplasms

Myeloid neoplasms are neoplastic diseases that lead to clonal hematopoiesis and disturb production of normal blood cells, leading to either cytopenias, an increase in blood cells (e.g. leukocytosis) or a combination of both (ARBER; ORAZI; HASSERJIAN; THIELE *et al.*, 2016). Myeloid neoplasms can be further classified into either acute diseases (Acute Myeloid Leukemia, AML), wherein the normal differentiation of the hematopoietic stem cell is altered, and there is an accumulation of malignant myeloid blasts, and chronic myeloid neoplasms, disorders in which there is a preservation of the normal hematopoietic differentiation to a variable degree (ARBER; ORAZI; HASSERJIAN; THIELE *et al.*, 2016). Despite these differences, it is considered that all myeloid neoplasms are neoplastic diseases, and the disease-initiating cell in most cases, as gleaned from animal models, is the hematopoietic stem cell or a hematopoietic progenitor cell that gained self-renewal ability through genetic changes (HUNTLY; SHIGEMATSU; DEGUCHI; LEE *et al.*, 2004; KOSCHMIEDER; GOTTGENS; ZHANG; IWASAKI-ARAI *et al.*, 2005; SCHEMIONEK; ELLING; STEIDL; BAUMER *et al.*, 2010).

According to the 2016 World Health Organization (WHO) classification of hematopoietic and lymphoid neoplams, there are 5 main groups of chronic myeloid neoplasms (i.e. non-AML myeloid neoplasms): Myeloproliferative Neoplasms (MPNs), Myelodysplastic / Myeloproliferative Neoplasms (MDS/MPNs), Myelodysplastic Syndromes (MDSs),

Myeloid/lymphoid neoplasms associated with eosinophilia and rearrangement of *PDGFRA, PDGFRB*, or *FGFR1* or with *PCM1-JAK2* and Mastocytosis (considered until the previous WHO classification as a form of MPN) (ARBER; ORAZI; HASSERJIAN; THIELE *et al.*, 2016). These 5 subgroups present with distinct disease phenotypes. The last two subgroups are very uncommon disorders and will not be further discussed.

MPNs are characterized by increased cellular proliferation in the bone marrow, an increased risk of thrombosis, development of extramedullary hematopoiesis and propensity to evolve into acute myeloid leukemia (AML) (ARBER; ORAZI; HASSERJIAN; THIELE *et al.*, 2016; RUMI; CAZZOLA, 2017). Furthermore, MPNs can be classified as Philadelphia-positive (Chronic Myeloid Leukemia, CML) or Philadelphia-negative (Ph-negative MPNs) (ARBER; ORAZI; HASSERJIAN; THIELE *et al.*, 2016). The most common forms of Ph-negative MPNs include polycythemia vera (PV), essential thrombocythemia (ET) and primary myelofibrosis (PMF) (ARBER; ORAZI; HASSERJIAN; THIELE *et al.*, 2016; RUMI; CAZZOLA, 2017).

MDS are clonal hematopoietic stem cell disorders characterized by ineffective hematopoiesis, dysplastic morphology changes in blood cells, peripheral blood cytopenias and increased risk of evolution to AML (STEENSMA; BENNETT, 2006). Furthermore, MDS can be classified into one of 11 distinct subtypes, based on the type and degree of cytopenias, presence of increased blasts, presence of ring sideroblasts and presence of specific cytogenetic changes (ARBER; ORAZI; HASSERJIAN; THIELE *et al.*, 2016).

Myelodysplastic/myeloproliferative neoplasms (MDS/MPNs) are interface disorders between MPNs and myelodysplastic syndromes, characterized by both proliferative (e.g. leukocytosis, splenomegaly) and dysplastic (e.g. bone marrow dysplasia, cytopenias) features (ARBER; ORAZI; HASSERJIAN; THIELE *et al.*, 2016). The most common form of MDS/MPN in adults is chronic myelomonocytic leukemia (CMML), that can present as dysplastic or proliferative subtype (ARBER; ORAZI; HASSERJIAN; THIELE *et al.*, 2016; BALL; LIST; PADRON, 2016) and is characterized by an increased number monocytes in the peripheral blood.

The main 3 groups of chronic myeloid neoplasms are thus characterized by increased marrow cellularity and either an increased in blood cell production (MPNs), a decrease in blood cell production (MDS) or a combination of both proliferative and dysplastic features (MDS/MPNs). At the center of the pathogenesis of chronic myeloid neoplasms are the genomic changes that lead to the development of these disorders.

1.2 Genetic changes and the pathogenesis of the chronic myeloid neoplasms

Recurrent genetic abnormalities are responsible for the pathogenesis of all subtypes of chronic myeloid neoplasms, including MPNs, MDSs and MDS/MPNs (BALL; LIST; PADRON, 2016; PAPAEMMANUIL; GERSTUNG; MALCOVATI; TAURO *et al.*, 2013; VAINCHENKER; KRALOVICS, 2017). The first abnormalities to be described were karyotype changes in bone marrow cells, such as t(9;22)(q34;q11) in CML del(5q) in MDS and other recurring cytogenetic changes (NOWELL; HUNGERFORD, 1960; ROWLEY,

1973; VAN DEN BERGHE; CASSIMAN; DAVID; FRYNS *et al.*, 1974). These initial observations were followed by descriptions of how certain genetic changes were associated with a specific disease phenotype – e.g., the t(9;22) leads to the common phenotype of CML, with an increased white blood cell count, presence of immature granulocytes in peripheral blood, increased in eosinophil and basophil count. Similarly, isolated del(5q) is associated with a specific form of MDS, characterized by anemia, higher incidence in women and hypolobated megakaryocytes (DALEY; VAN ETTEN; BALTIMORE, 1990; VAN DEN BERGHE; CASSIMAN; DAVID; FRYNS *et al.*, 1974). These initial findings were suggestive of a highly specific genetic-phenotype correlation, with each disease having a specific genetic change that could be used for disease classification. This is reflected in the 2001 WHO classification, where CML was defined as a MPN characterized by the t(9;22)(q34;q11) (VARDIMAN; HARRIS; BRUNNING, 2002).

In the last 20 years, a plethora of genetic changes were described in patients with chronic and acute myeloid neoplasms. While some rare abnormalities are specific for a given disease phenotype (e.g. *PDGFRA* abnormalities and myeloid/lymphoid neoplasms with eosinophilia) (COOLS; DEANGELO; GOTLIB; STOVER *et al.*, 2003), in most cases there is no specific genetic finding. *TET2* mutations, for example, can occur at distinct frequencies in patients with MDS, MDS/MPNs, MPNs and AML (DELHOMMEAU; DUPONT; DELLA VALLE; JAMES *et al.*, 2009; KOSMIDER; GELSI-BOYER; CIUDAD; RACOEUR *et al.*, 2009; TEFFERI; PARDANANI; LIM; ABDEL-WAHAB *et al.*, 2009). In some cases, there is an association between a specific pathway of genes being mutated (i.e. gene mutations that lead to similar biological consequences) and one subtype of

myeloid neoplasm. For example, gene mutations leading to activation of the JAK-STAT pathway (e.g. mutations in *JAK2, MPL, CALR*) are frequently found in patients with Ph-negative MPNs (PV, ET and MF) (BAXTER; SCOTT; CAMPBELL; EAST *et al.*, 2005; JAMES; UGO; LE COUEDIC; STAERK *et al.*, 2005; KLAMPFL; GISSLINGER; HARUTYUNYAN; NIVARTHI *et al.*, 2013; KRALOVICS; PASSAMONTI; BUSER; TEO *et al.*, 2005; LEVINE; WADLEIGH; COOLS; EBERT *et al.*, 2005; NANGALIA; MASSIE; BAXTER; NICE *et al.*, 2013). Although rare, these mutations can also be found in patients with CMML, AML and other more uncommon disease phenotypes (JELINEK; OKI; GHARIBYAN; BUESO-RAMOS *et al.*, 2005; LEVINE; LORIAUX; HUNTLY; LOH *et al.*, 2005).

The complexity of genomic findings in chronic myeloid neoplasms is further underscored by the finding that certain combinations of genes are more frequently in some disease phenotypes. For example, the presence of both mutations of the gene involved in DNA hidroxymethylation *TET2* and the mRNA splicing factor gene *SRSF2* are a hallmark of CMML (KOSMIDER; GELSI-BOYER; CIUDAD; RACOEUR *et al.*, 2009; MAKISHIMA; VISCONTE; SAKAGUCHI; JANKOWSKA *et al.*, 2012; MALCOVATI; PAPAEMMANUIL; AMBAGLIO; ELENA *et al.*, 2014). The order of gene mutations also appear to play a role, as best exemplified by the paper by ORTMANN et al, where it was demonstrated that the development of activating *JAK2* mutations followed by inactivating *TET2* mutations more often leads to the appearance of PV, while the opposite finding leads in equal probability to the development of ET and PV (ORTMANN; KENT; NANGALIA; SILBER *et al.*, 2015).

It appears that, with a few exceptions, most genes are not specific to any disease entity, since similar genes are found to be recurrently mutated in all myeloid malignancies

(CANCER GENOME ATLAS RESEARCH; LEY; MILLER; DING *et al.*, 2013; GRINFELD; NANGALIA; BAXTER; WEDGE *et al.*, 2018; LUNDBERG; KAROW; NIENHOLD; LOOSER *et al.*, 2014; PAPAEMMANUIL; GERSTUNG; BULLINGER; GAIDZIK *et al.*, 2016; PAPAEMMANUIL; GERSTUNG; MALCOVATI; TAURO *et al.*, 2013). One hypothesis to explain this phenomenon is that disease phenotype is the result of a complex interaction between the specific gene that is mutated, the combination of mutations in distinct genes, the biological pathway that is affected by the mutations and the order of mutation occurrence (i.e. clonal mutations vs subclonal mutations) (MALCOVATI; PAPAEMMANUIL; AMBAGLIO; ELENA *et al.*, 2014; ORTMANN; KENT; NANGALIA; SILBER *et al.*, 2015). Determination of patterns of mutation occurrence, altered pathways and order of mutation acquisition and their association with each disease phenotype may thus facilitate interpretation of genetic sequencing studies of myeloid neoplasms (that have become more commonly used in clinical practice) and form the basis of a molecular classification of these neoplasms. Since there are probably several distinct genetic variables that associate with a specific phenotype, use of statistical models to best predict the probability of a specific disease phenotype may be an improvement over a simpler paradigm of associating one gene mutation with one disease.

1.3 Next-generation sequencing and its application in cancer genomics

In the last decade, development of massive sequencing of whole exomes (WES; whole exome sequencing) has revolutionized the field of cancer genomics, leading to studies detailing the genetic profile of patients with several types of neoplasms (BERGER;

LAWRENCE; DEMICHELIS; DRIER *et al.*, 2011; CANCER GENOME ATLAS RESEARCH; ALBERT EINSTEIN COLLEGE OF; ANALYTICAL BIOLOGICAL; BARRETOS CANCER *et al.*, 2017; CANCER GENOME ATLAS RESEARCH; LEY; MILLER; DING *et al.*, 2013; FRASER; SABELNYKOVA; YAMAGUCHI; HEISLER *et al.*, 2017; GEORGE; LIM; JANG; CUN *et al.*, 2015; SONG; LI; OU; GAO *et al.*, 2014). The use of WES and whole genome sequencing (WGS) formed the basis for two consortiums tasked with detailing the genomic landscape of several tumor types: The Cancer Genome Atlas (conducted mostly in the USA) and the International Cancer Genome Consortium (led mostly by European Institutions) (INTERNATIONAL CANCER GENOME; HUDSON; ANDERSON; ARTEZ *et al.*, 2010; WANG; JENSEN; ZENKLUSEN, 2016).

Analysis of data from WES sequencing allows for the identification of recurrently mutated genes that comprise the most likely candidates for being ‘driver’ genes (i.e. genes that lead to oncogenesis and neoplasm development). The search for recurrently mutated driver genes is complicated by the presence of background ‘passenger’ mutations, that play no role in oncogenesis and are a function of the gene size, gene expression level and timing of gene replication cell cycle (KOREN; POLAK; NEMESH; MICHAELSON *et al.*, 2012; LAWRENCE; STOJANOV; POLAK; KRYUKOV *et al.*, 2013; PLEASANCE; CHEETHAM; STEPHENS; MCBRIDE *et al.*, 2010). Several computer algorithms have been developed that seek to detect the most likely candidate for being gene drivers among the several dozen genes that are found to harbor somatic mutations in a given tumor analyzed by WES (GONZALEZ-PEREZ; PEREZ-LLAMAS; DEU-PONS; TAMBORERO *et al.*, 2013; LAWRENCE; STOJANOV; POLAK; KRYUKOV *et al.*, 2013).

Statistical analysis of WES data also allows for the determination of patterns of mutated genes that co-occurrence or are mutually exclusive. The first is important for determining pairs of genes that are often found to be mutated together in the same sample and may reveal synergy between gene mutations that may underlie differences in disease phenotype (as best exemplified by the already mentioned combination of *TET2* and *SRSF2* mutations in CMML) (MALCOVATI; PAPAEMMANUIL; AMBAGLIO; ELENA *et al.*, 2014). The second finding (mutual exclusive genes) is important for identification of putative biological pathways with a role in oncogenesis, such as JAK-STAT pathway mutations. This occurs because mutated genes that belong to the same pathway have a tendency for not being present in the same patient (unless they belong to distinct clones) (BABUR; GONEN; AKSOY; SCHULTZ *et al.*, 2015; LEISERSON; BLOKH; SHARAN; RAPHAEL, 2013). This can be important particularly when the wild-type gene has no known function in the biological pathway being considered, since in these cases querying gene function databases would not reveal the association. An example of this would be the relation between mutated *CALR* and activation of the JAK-STAT pathway. The wild-type CALR protein product does not lead to JAK-STAT pathway activation. However, mutated CALR protein (as found in MF/ET) causes JAK-STAT activation, and thus mutant *CALR* is mutually exclusive with mutations in the other JAK-STAT genes *JAK2* and *MPL* mutations (ELF; ABDELFATTAH; CHEN; PERALES-PATON *et al.*, 2016; KLAMPFL; GISSLINGER; HARUTYUNYAN; NIVARTHI *et al.*, 2013). Finally, analysis of WES data can reveal clues about the clonal architecture of selected cases with enough mutations for this analysis (MILLER; WHITE; DEES; GRIFFITH *et al.*, 2014). It is important to mention that cohort size is critical for analysis of WES data, particularly for the saturation

in the detection of driver genes (LAWRENCE; STOJANOV; MERMEL; ROBINSON *et al.*, 2014).

There have been several studies using WES that detailed the genetic profile of myeloid neoplasms. The largest studies have focused on patients with MDS and AML (CANCER GENOME ATLAS RESEARCH; LEY; MILLER; DING *et al.*, 2013; MAKISHIMA; YOSHIZATO; YOSHIDA; SEKERES *et al.*, 2017). Some studies have evaluated the genomic landscape through WES of MPNs and MDS/MPNs (CABAGNOLS; FAVALE; PASQUIER; MESSAOUDI *et al.*, 2016; KLAMPFL; GISSLINGER; HARUTYUNYAN; NIVARTHI *et al.*, 2013; MASON; KHORASHAD; TANTRAVAHI; KELLEY *et al.*, 2016; MERLEVEDE; DROIN; QIN; MELDI *et al.*, 2016; NANGALIA; MASSIE; BAXTER; NICE *et al.*, 2013; PIAZZA; VALLETTA; WINKELMANN; REDAELLI *et al.*, 2013; WANG; SWIERCZEK; DRUMMOND; HICKMAN *et al.*, 2014). However, most WES studies published in MPNs and MDS/MPNs have a small (<100) sample size, except for the study published by NANGALIA et al that recruited 151 cases. Recently, GRINFELD et al published the genetic analysis of a large cohort of patients with MPNs (GRINFELD; NANGALIA; BAXTER; WEDGE *et al.*, 2018). This study, despite its large sample size, suffers from two major limitations. The first is that the authors used a selected panel 69 genes, thus limiting discovery of mutated genes to known driver genes in these diseases. Second, the cohort, while comprising the 3 most common types of Ph-negative MPNs (ET, MF and PV) is heavily tilted towards ET patients, that comprise more than 50% of the cohort. Patients with MDS/MPNs were not included in the trial, and as mentioned, they may harbor similar mutations as patients with Ph-negative MPNs. To better discern the specific gene drivers found in these diseases and determine the most important

genetic predictors of disease phenotype it is crucial to analyze a large cohort of patients analyzed by an unbiased genomewide sequencing methodology such as WES. To overcome the cost limitations of conducting a single large study, combining data from several distinct studies increases sample size and statistical power, and this allows discovery of associations not seen with smaller cohorts.

The main hypothesis of the present study is that using genomic data from a large dataset of WES studies it is possible to discern statistical associations between mutated genes that predict for a specific disease diagnosis and/or subtype in MPNs and MDS/MPNs. These associations between genotype and phenotype will be evaluated at several distinct levels, mainly:

1. Association of a driver gene mutation with a disease phenotype
2. Association of a combination of driver genes mutations with a disease phenotype
3. Association of a biological pathway (based on mutated driver genes) with disease phenotype
4. Association of clonal dominance of driver genes with disease phenotype

OBJECTIVES

2. Objectives

2.1 Primary Objective

The primary objective is to describe the list of most likely driver genes to be mutated in MPNs and MDS/MPNs and correlate these findings with disease phenotype

2.2 Secondary Objectives

The secondary objectives are:

1. To annotate putative driver genes and determine the biological pathways that are found to be activated/disrupted due to drive gene mutation and correlate these findings with disease phenotype

2. To determine patterns of association between mutated genes, including co-occurrences of mutations and mutual exclusivity of mutated genes, and correlate these findings with disease phenotype

3. To determine, when possible, the clonally dominant driver genes and associate these findings with disease phenotype

4. To fit a logistic regression model to classify patients into a disease subtype based solely on the pattern of mutated driver genes

5. To determine the impact, if any, of mutated driver genes on survival outcomes of patients with MDS/MPNs and MPNs

METHODS

3. Methods

3.1 Cases

A combined cohort of 403 patients was analyzed in this study. It includes WES data from 124 Brazilian patients. Patients were recruited for study participation among 14 different centers in Brazil. The study was approved at each institution Institutional Review Board before first sample collection. Written informed consent was obtained from all patients. Samples were collected locally at each center and were shipped to the coordinating center in Sao Paulo, Brazil (Hospital Israelita Albert Einstein, HIAE) for sample processing and DNA extraction. Inclusion criteria for the Brazilian cohort included having a diagnosis of Ph-negative MPN or MDS/MPN, no prior history of transformation to AML, no prior history of hematopoietic stem cell transplantation, and willingness to collect a normal skin biopsy to serve as a matched DNA control. Patients with systemic mastocytosis and myeloid/lymphoid neoplasms with eosinophilia and *FGFR1/PDGFRA/PDGFRB* rearrangements were not included.

Regarding published data, studies were identified in which patients' samples were analyzed by WES, and which made available sample-level data on all individual mutations that were identified. Words used for search strategy on Pubmed were: "whole", "exome", "sequencing", 'essential thrombocythemia', 'polycythemia vera', 'myelofibrosis', 'chronic myelomonocytic leukemia', 'atypical chronic myeloid leukemia', 'myeloproliferative', 'myeloproliferative/myelodysplastic'. The search was conducted on April 2017. Targeted resequencing studies and case reports were not included. There were 9 potential studies from which 6 studies were selected, comprising 279 patients: (1) 151 patients with Ph-

negative MPNs (mainly ET, PV and MF) published by NANGALIA et al (NANGALIA; MASSIE; BAXTER; NICE *et al.*, 2013); (2) 49 patients with CMML published by MERLEVEDE et al (MERLEVEDE; DROIN; QIN; MELDI *et al.*, 2016); (3) 21 patients with CMML published by Mason et al (MASON; KHORASHAD; TANTRAVAHI; KELLEY *et al.*, 2016); (4) 10 patients with ET published by CABAGNOLS et al (CABAGNOLS; FAVALE; PASQUIER; MESSAOUDI *et al.*, 2016); (5) 15 patients with atypical chronic myeloid leukemia (aCML) published by Piazza et al (PIAZZA; VALLETTA; WINKELMANN; REDAELLI *et al.*, 2013); (6) 33 patients with PV published by Wang et al (WANG; SWIERCZEK; DRUMMOND; HICKMAN *et al.*, 2014). Data on individual mutations was extracted from supplemental data from the studies.

3.2 Sequencing and mutation detection of the Brazilian Cohort

DNA was extracted from electromagnetically sorted CD66b-granulocytes (Stem Cell Technologies, Vancouver, Canada) and a paired normal skin biopsy (representing control DNA), followed by exome enrichment using SureSelect Human All Exon V4 (Agilent Technologies, Santa Clara, CA). Prepared libraries were then sequenced with 100 bp paired end reads on a Illumina HiSeq 2000 (San Diego, CA). Sequencing was performed at the core sequencing facility of the Columbia Genome Center in New York, NY on a fee-for-service basis. Bioinformatics and subsequent analysis were performed locally at HIAE after downloading raw (FASTQ) data from the Columbia Genome Center server.

FASTQ reads were aligned to human genome build 37 using BWA (v.0.7.15) (LI; DURBIN, 2009). Picard (v.2.14.0) was used for PCR duplicate removal (INSTITUTE), and

tools from the Genome Analysis Toolkit (GATK, v.3.7) were utilized for indel realignment and base quality recalibration (DEPRISTO; BANKS; POPLIN; GARIMELLA *et al.*, 2011). Somatic variants were called combining the output of SomaticSniper (v.1.0.5.0), Mutect (v.1.1.5) and Pindel (v.0.2.5b8) (CIBULSKIS; LAWRENCE; CARTER; SIVACHENKO *et al.*, 2013; LARSON; HARRIS; CHEN; KOBOLDT *et al.*, 2012; YE; SCHULZ; LONG; APWEILER *et al.*, 2009). The combined output of these 3 mutation callers was annotated using Annovar (v.2016-02-01) (WANG; LI; HAKONARSON, 2010) and filtered using the following in-house criteria: minimum coverage at both tumor/germline samples ≥ 8 reads; fraction of reads supporting variant allele (variant allele fraction [VAF]) ≤10% in germline; ratio of tumor VAF:germline VAF >2. Remaining calls were further filtered by removing synonymous mutations, non-coding mutations located more than 6 bases from splice junction, mutations found with a frequency ≥1% at either the 1000 Genome (GENOMES PROJECT; AUTON; BROOKS; DURBIN *et al.*, 2015) or the NHLBI GO Exome Sequencing Project databases (Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA (URL: http://evs.gs.washington.edu/EVS/) [April, 2017]), and mutations reported at dbSNP138 (SHERRY; WARD; KHOLODOV; BAKER *et al.*, 2001) that were not previously associated with a malignant phenotype. Review of unfiltered calls for genes that commonly present with large deletions (e.g. *CALR*) was undertaken to ensure that no mutations were lost due to stringent filtering. Mutations were validated through either Sanger sequencing or using a custom, targeted Haloplex HS sequencing panel.

3.3 MAF File Creation

Data on chromosome, start position, end position, reference allele, variant allele, sample identifier, and diagnosis was extracted from the final mutation call file from the Brazilian cohort and from supplementary data files from published papers and chromosome positions were adjusted to the 1-base notation when needed. Mutations were annotated using the Oncotator Web API through the R package 'maftools', and a MAF (Mutation Annotation Format file) was generated for each cohort (MAYAKONDA; LIN; ASSENOV; PLASS *et al.*, 2018). All MAF files that were generated were merged in a single MAF file that was used for all subsequent analysis.

3.4 Bioinformatic and statistical analysis

3.4.1 Driver Mutation Prediction

Softwares MutSigCV, version 1.2, through the GenePattern portal (www.genepattern.com) from Broad Institute and IntOGen (www.intogen.org) were used for prediction of driver genes (GONZALEZ-PEREZ; PEREZ-LLAMAS; DEU-PONS; TAMBORERO *et al.*, 2013; LAWRENCE; STOJANOV; POLAK; KRYUKOV *et al.*, 2013). For MutSigCV, the coverage file utilized was a territory file for the reference human exome that directed MutSigCV to assume full coverage, the covariate file used was the same file used in the MutSigCV publication (LAWRENCE; STOJANOV; POLAK; KRYUKOV *et al.*, 2013), and a dictionary file supplied by Broad was used for category and effect discovery.

All files are publicly available on the GenePattern servers from Broad Institute. For the IntOgen software, parameter OncodriveFM genes threshold was set at 2 (minimum number of samples a given gene must have to be analyzed by OncodriveFM) and OncodriveCLUST genes threshold was set at 5 (minimum number of mutated samples a gene must have to be analyzed by OncodriveCLUST). Genes considered as being significantly mutated were genes with q-value <0.1 and/or genes with a mutational hotspot that were present in the results of either one of the algorithms. The platform DAVID (http://david.ncifcrf.gov) version 6.8 was used for functional annotation of the list of driver genes generated.

3.4.2 Determination of patterns of co-occurrence and mutual exclusivity among gene mutations

Patterns of co-occurrence and mutual exclusivity among pairs of mutated genes were determined for driver genes with a frequency ≥2%. Pairwise analysis utilized the SELECT algorithm, that was previously published (MINA; RAYNAUD; TAVERNARI; BATTISTELLO *et al.*, 2017). Analysis were conducted on the whole cohort, only on MPNs and only on MDS/MPNs. Valid interactions were those considered as passing the false discovery rate filter of the SELECT algorithm (MINA; RAYNAUD; TAVERNARI; BATTISTELLO *et al.*, 2017). For co-mutated patterns among 3 or more genes, the Apriori market basket analysis algorithm was used (MORGAN; NI; MIRANKER; IYER, 2007). This algorithm is a machine learning method that can discover associations between variables in a dataset. It is frequently used by large retailers (e.g. Amazon) to uncover association between different items (i.e. discover relationships between items that people

buy). For this analysis, every mutated gene was considered as an ‘item’, and association between three or more mutated genes as a ‘transaction’ (e.g. patient who has mutation in gene ‘X’ and ‘Y’ frequently has mutation in gene ‘Z’). Parameters used were support of 0.1% (i.e. percentage of transactions that contained the items in the association rule), confidence of 80% (i.e. confidence that the association exists) and count of 3 (i.e. at least 3 samples had the association rule).

3.4.3 Determination of Association between Disease Phenotype and Mutated Genes or Pathways

Fisher tests were used to determine statistical significance of measured differences in distribution among disease diagnosis/subtypes and mutated genes, pairs/triads of genes and biological pathways. The logarithm of the Odds Ratio (OR) was used to measure the strength of the association between disease phenotype and mutated genes. P-values were adjusted for false discovery rate (FDR) using the Benjamini-Hochberg method that calculated Q-values (BENJAMINI; HOCHBERG, 1995).

3.4.4 Analysis of clonal heterogeneity

Clonal heterogeneity was assessed using variant allele fraction (VAF) of mutations. In order to account for sites of loss of heterozygosis, mutations localized in chromosome X or Y and mutations with a VAF ≥80% were removed from the analysis. The cutpoint of 80% for VAF was chosen since a Kernel Density Plot of mutation VAF of all samples showed a peak above 80%, which suggests that those comprised mutations localized in either sexual chromosomes or in sites of loss of heterozygosity. Since no copy number

data was available for most of the samples, we decided to exclude those mutations from the analysis of clonal heterogeneity. For the remaining mutations, clonal heterogeneity was inferred through a parametric finite mixture model using the R package 'maftools', which clusters mutations with a similar VAF and classifies them in single clones (JARA; HANSON; QUINTANA; MULLER *et al.*, 2011; MAYAKONDA; LIN; ASSENOV; PLASS *et al.*, 2018). Mutations localized in the clone with the highest VAF were considered 'clonal' mutations, and mutations localized in other clones with lower VAF were considered 'subclonal' mutations. The proportion in which a driver gene was found to be clonal or subclonal across the samples was evaluated for each sample affected by a given driver. Gene drivers were classified as early vs late drivers based on the degree of enrichment for being classified as clonal (early driver gene, q-value<0.2) vs being classified as subclonal (late driver gene, q-value<0.2) employing a binomial test (where the number of successes was defined as the number of times the gene was classified as clonal, the number of trials the number of cases with the mutated gene, with a probability of success of 50%).

3.4.5 Logistic regression model

Separate logistic regression models were fit to determine variables associated with each specific diagnosis (only the more common diagnosis PV, MF, CMML and ET were included) and subtype (i.e. MPN, MDS/MPN). The cohort was split 70/30 into a training (N=283) and validation set (N=120). The training set was used for model development. Variables to be included in the final model were selected using LASSO (Least Absolute Shrinkage and Selection Operator) (TIBSHIRANI, 1996). Models were then fitted on the

validation data, and fitted probabilities were estimated for each case. Those cases with fitted probabilities greater than 0.5 were considered as “positive” cases, and this information was used together with the actual diagnosis to build a confusion matrix and estimate accuracy using the following formula:

$$\frac{True\ Positives+True\ Negatives}{True\ Positives+True\ Negatives+False\ Positives+False\ Negatives}.$$

3.4.6 Survival Analysis

Survival outcomes were solely analyzed on the cohort of Brazilian patients, since this was the only cohort that the author had access to survival data. Overall survival (OS) was estimated using the Kaplan-Meier method, and OS was defined as the time from sample collection for exome sequencing until death from any cause, and patients who were alive at last follow-up were censored (KAPLAN; MEIER, 1958). The size effect of mutations on OS was estimated using the Hazard Ratio (HR), and the HR was calculated with a Cox proportional hazards regression model (COX, 1972). Separate models were fit for each genetic driver that was found, if it was found to be altered in at least 2 cases in the Brazilian cohort. Since different disease diagnosis can influence survival outcomes, the impact of a given genetic abnormality on survival was adjusted for the subtype of neoplasm that the patient presented with (MPN vs MDS/MPN). The resulting p-value of each covariate on each individual Cox-model was adjusted for multiple testing using the FDR method of Benjamini-Hochberg (BENJAMINI; HOCHBERG, 1995).

3.4.7 Software used for analysis

R (version 3.4.2) was used for all analysis and plotting graphics.

RESULTS

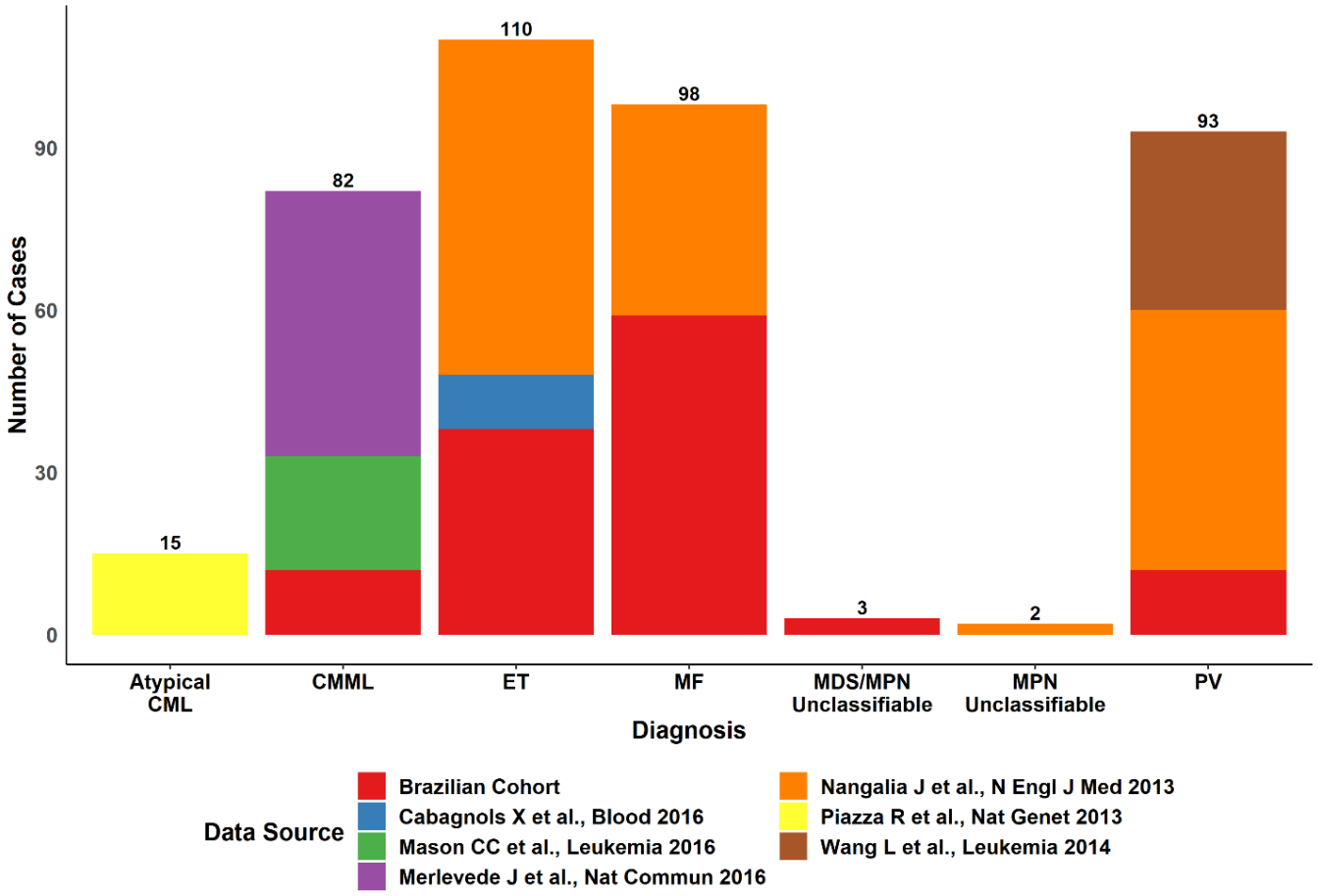
4. Results

4.1 Unbiased candidate driver gene discovery in MPNs and MDS/MPNs

Samples collected from 124 patients with a diagnosis of Ph-negative MPNs and/or MDS/MPNs were analyzed through WES. Samples were classified as tumor samples (CD66b-selected granulocytes) and matched germline samples (skin biopsy). Mean read depth for tumor and germline samples was 142x and 68x, respectively. After mutation calling and filtering, a mean ± s.d. number of 30.9 ± 57.1 silent and non-silent somatic mutations (including single nucleotide variants and insertions and deletions) were detected per case. The median number of mutations per case was 16.5 (range 0-341). One patient had no somatic mutations detected.

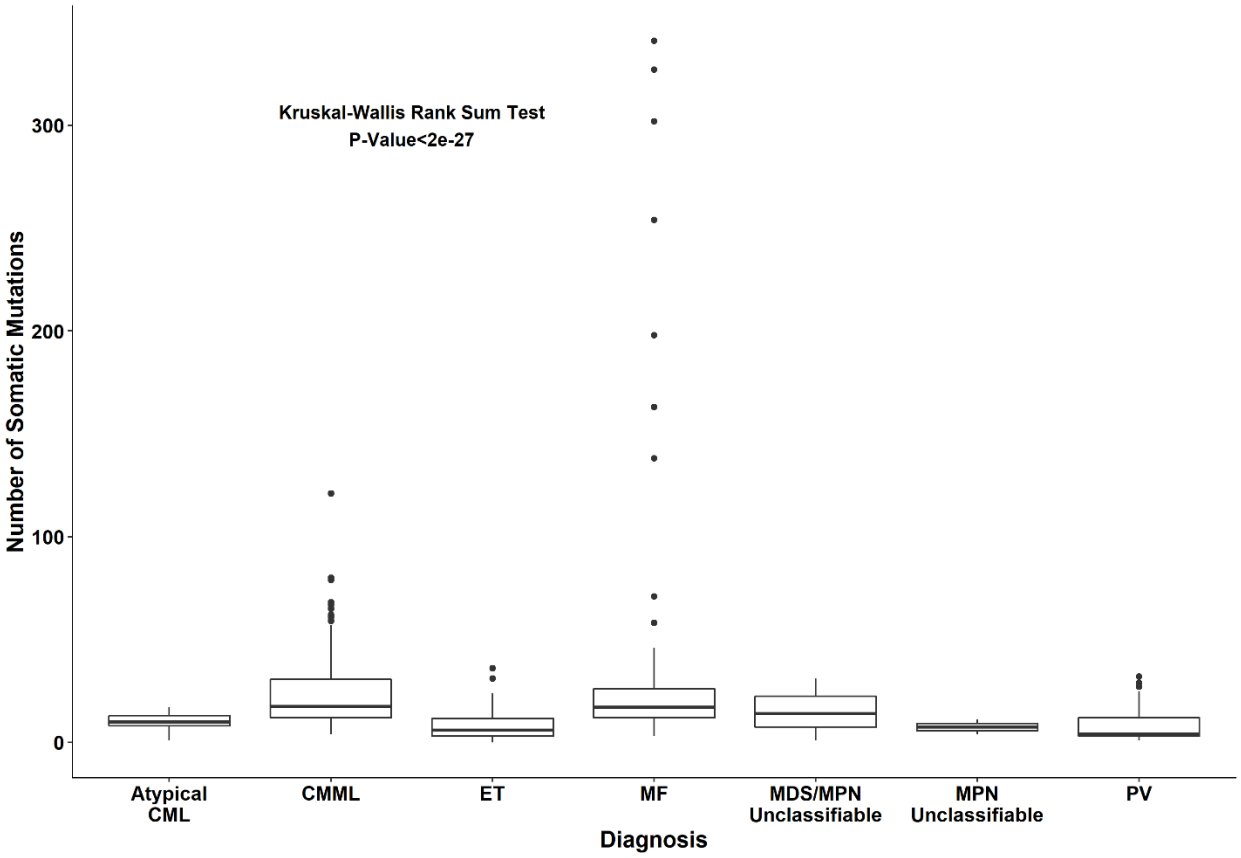
To infer candidate driver genes in MPNs and MDS/MPNs, two different algorithms, MutSigCV and IntOgen, were utilized. In order to maximize statistical sensitivity, data from the Brazilian cohort was combined with previously published data on 279 patients, for a total cohort of 403 patients. Diagnosis and data source are summarized in Figure 1.

Figure 1 - Disease diagnosis breakdown and data source utilized in the project



Two patients in the final cohort (one from the Brazilian cohort and the other from the Nangalia cohort) had no somatic mutations detected. Both patients had a diagnosis of ET. This final cohort had a mean mutation number of 18.4 ± 35 mutations per case, or 0.36 mutations/Megabase (Mb; considering an exome size of 51 Mb). The median mutation number of 11 mutations (range 0-341). The number of somatic mutations varied across different diagnosis, with a p-value<1.9e27 by the Kruskal-Wallis test (Figure 2).

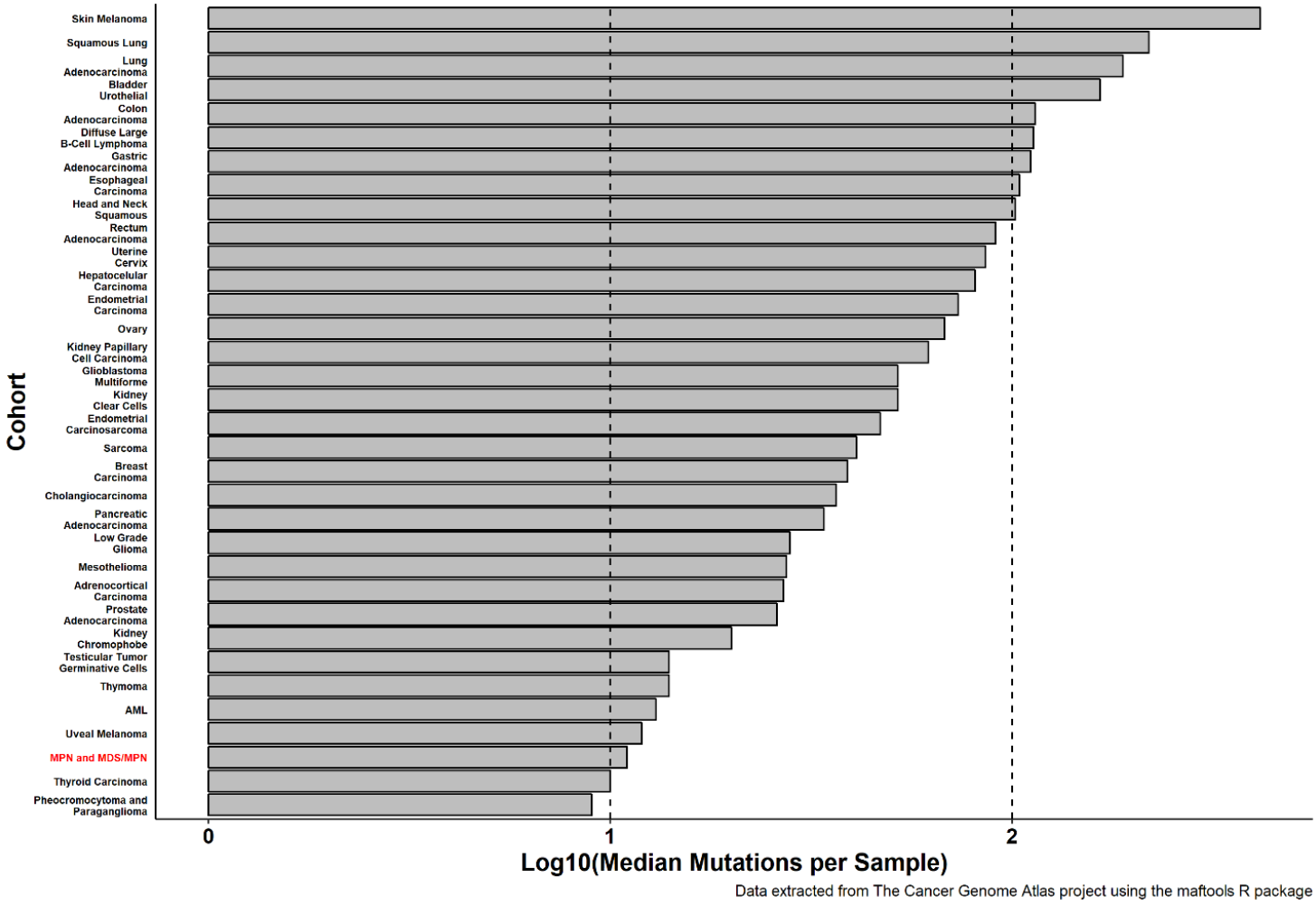
Figure 2 - Comparison of Number of Mutations across different diagnosis of MPNs and MDS/MPNs



Comparing the number of mutations in this cohort with the number of mutations in other neoplasms, based on data from "The Cancer Genome Atlas" project, that are publicly available, patients with MDS and MDS/MPNs have a lower number of somatic mutations than most other neoplasms that were studied on the "The Cancer Genome Atlas" (Figure 3). Using the statistical power calculation for tumor sequencing studies at http://www.tumorportal.org/power, and considering the mutation rate as 0.36 mutations/Mb it can be expected that a cohort with 403 patients would saturate (i.e. detect all genes) candidate driver gene discovery present in 5% of patients, would provide a

power of 94% to detect genes mutated in 3% of cases, and 67% power to detect genes mutated in 2% of cases.

Figure 3 - Comparison of somatic mutation burden between the study cohort and cohorts from other tumor types analyzed on "The Cancer Genome Atlas" project



The combined analysis of MutSigCV and IntOgen identified 54 genes as being statistically significantly mutated in MPNs and MDS/MPNs and likely drivers in disease pathogenesis

(Figure 4). A mutation in at least one of these genes was found in 94.5% of patients. Seventeen genes (32%) were not previously reported to be mutated in these neoplasms. Median number of driver genes per sample was 2 (range 0-8), with 22 patients having no identified driver genes. There was no statistical difference in the distribution of mutated driver genes in the Brazilian cohort compared to the data extracted from the literature (Table 1).

Figure 4 - Number of mutations by gene and diagnosis among 54 putative driver genes

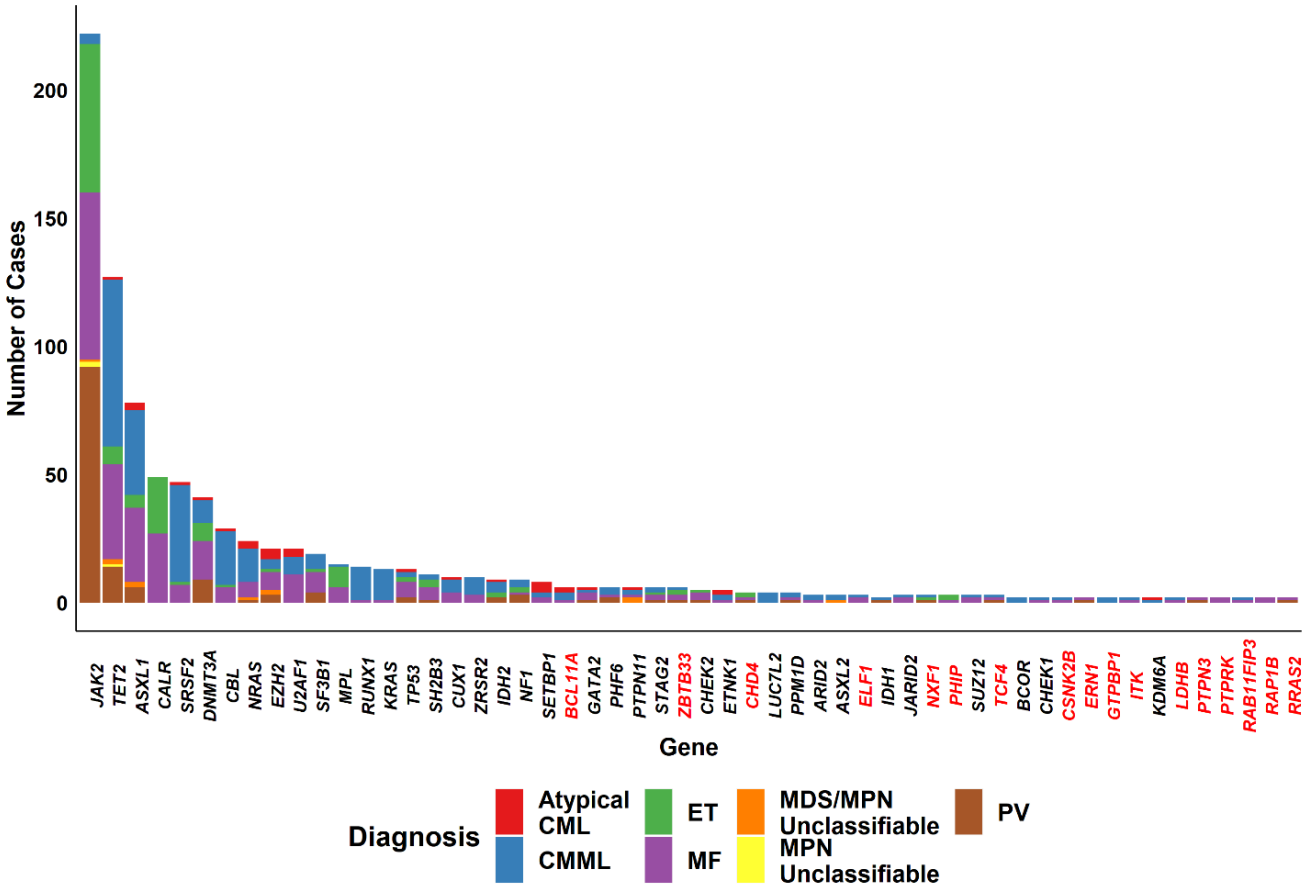


Table 1 – Fisher test results for comparison of prevalence of mutated driver genes in the Brazilian cohort with the international cohort from published studies

Gene	P-Value	Q-Value	Gene	P-Value	Q-Value	Gene	P-Value	Q-Value
ARID2	1.000	1.000	*KRAS*	0.116	0.654	*TP53*	0.121	0.654
ASXL1	0.073	0.631	*LDHB*	0.520	0.871	*U2AF1*	0.455	0.871
ASXL2	1.000	1.000	*LUC7L2*	1.000	1.000	*ZBTB33*	0.074	0.631
BCL11A	0.671	0.980	*MPL*	0.565	0.871	*ZRSR2*	0.256	0.871
BCOR	1.000	1.000	*NF1*	0.444	0.871			
CALR	0.008	0.312	*NRAS*	0.647	0.971			
CBL	0.360	0.871	*NXF1*	0.556	0.871			
CHD4	1.000	1.000	*PHF6*	1.000	1.000			
CHEK1	0.094	0.631	*PHIP*	1.000	1.000			
CHEK2	1.000	1.000	*PPM1D*	1.000	1.000			
CSNK2B	0.520	0.871	*PTPN11*	0.376	0.871			
CUX1	0.465	0.871	*PTPN3*	0.520	0.871			
DNMT3A	0.468	0.871	*PTPRK*	0.094	0.631			
ELF1	1.000	1.000	*RAB11FIP3*	0.520	0.871			
ERN1	0.094	0.631	*RAP1B*	0.520	0.871			
ETNK1	1.000	1.000	*RRAS2*	0.520	0.871			
EZH2	0.420	0.871	*RUNX1*	0.549	0.871			
GATA2	0.376	0.871	*SETBP1*	1.000	1.000			
GTPBP1	1.000	1.000	*SF3B1*	0.309	0.871			
IDH1	1.000	1.000	*SH2B3*	0.040	0.631			
IDH2	0.286	0.871	*SRSF2*	0.178	0.870			
ITK	1.000	1.000	*STAG2*	0.012	0.312			
JAK2	0.193	0.870	*SUZ12*	0.224	0.871			
JARID2	0.520	0.871	*TCF4*	0.556	0.871			
KDM6A	1.000	1.000	*TET2*	0.798	1.000			

Figure 5 - Number of Driver Genes per Disease Diagnosis

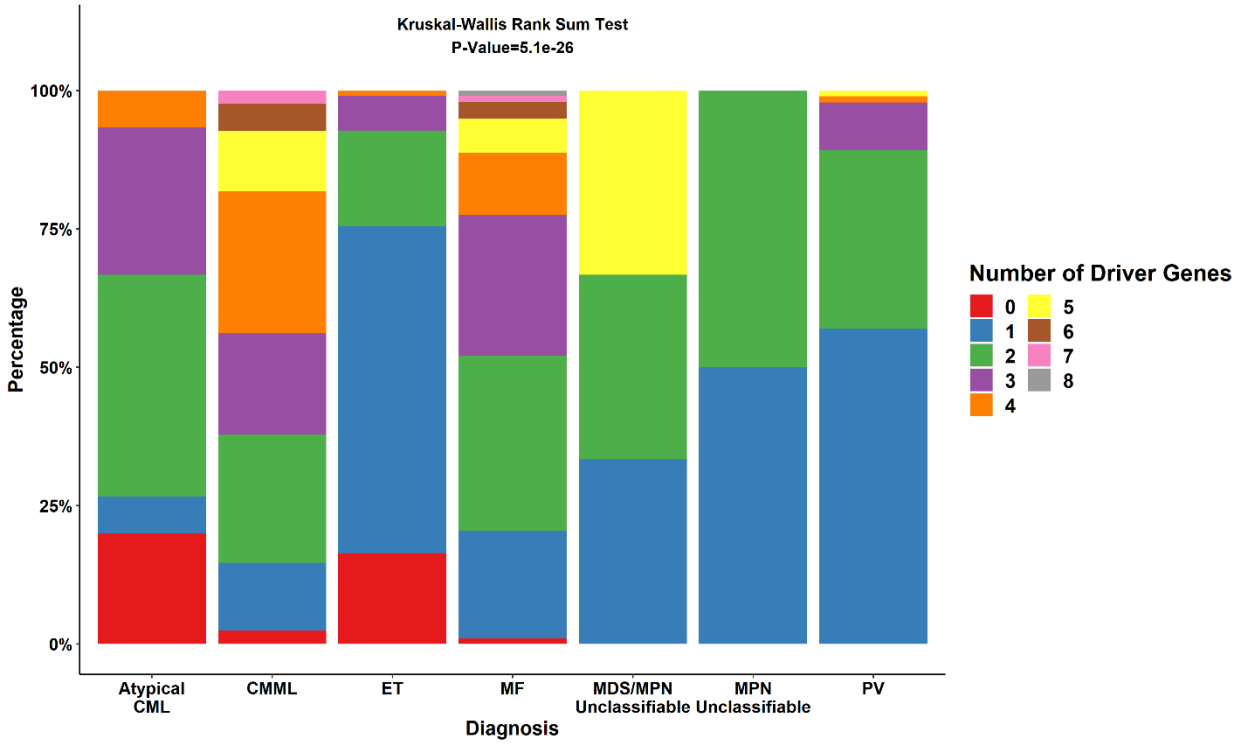
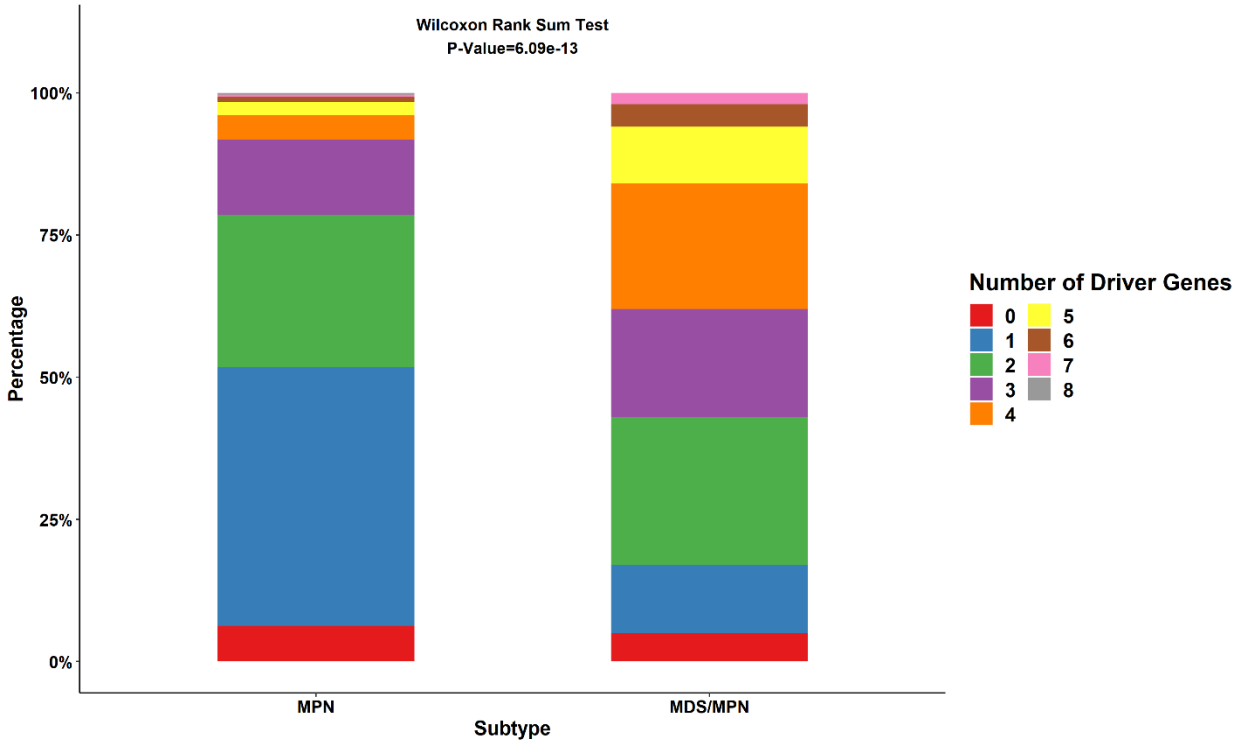


Figure 6 - Number of Driver Genes per Disease Subtype



The number of driver mutations found varied by diagnosis and by disease subtype, with a higher number of driver mutations found in MF, CMML, atypical CML and in patients with MDS/MPNs vs patients with MPNs (Figures 5 and 6). Nineteen genes had a differential frequency among the various disease diagnosis and subtypes, indicating that the presence of mutations in these genes may contribute to the disease phenotype. For example, *JAK2* and *CALR* occurred almost exclusively in patients with Ph-negative MPNs, while *RUNX1* and *KRAS* were more frequently mutated in patients with MDS/MPNs (Figures 7 and 8).

Figure 7 - Distribution of Driver Genes among different Disease Diagnosis

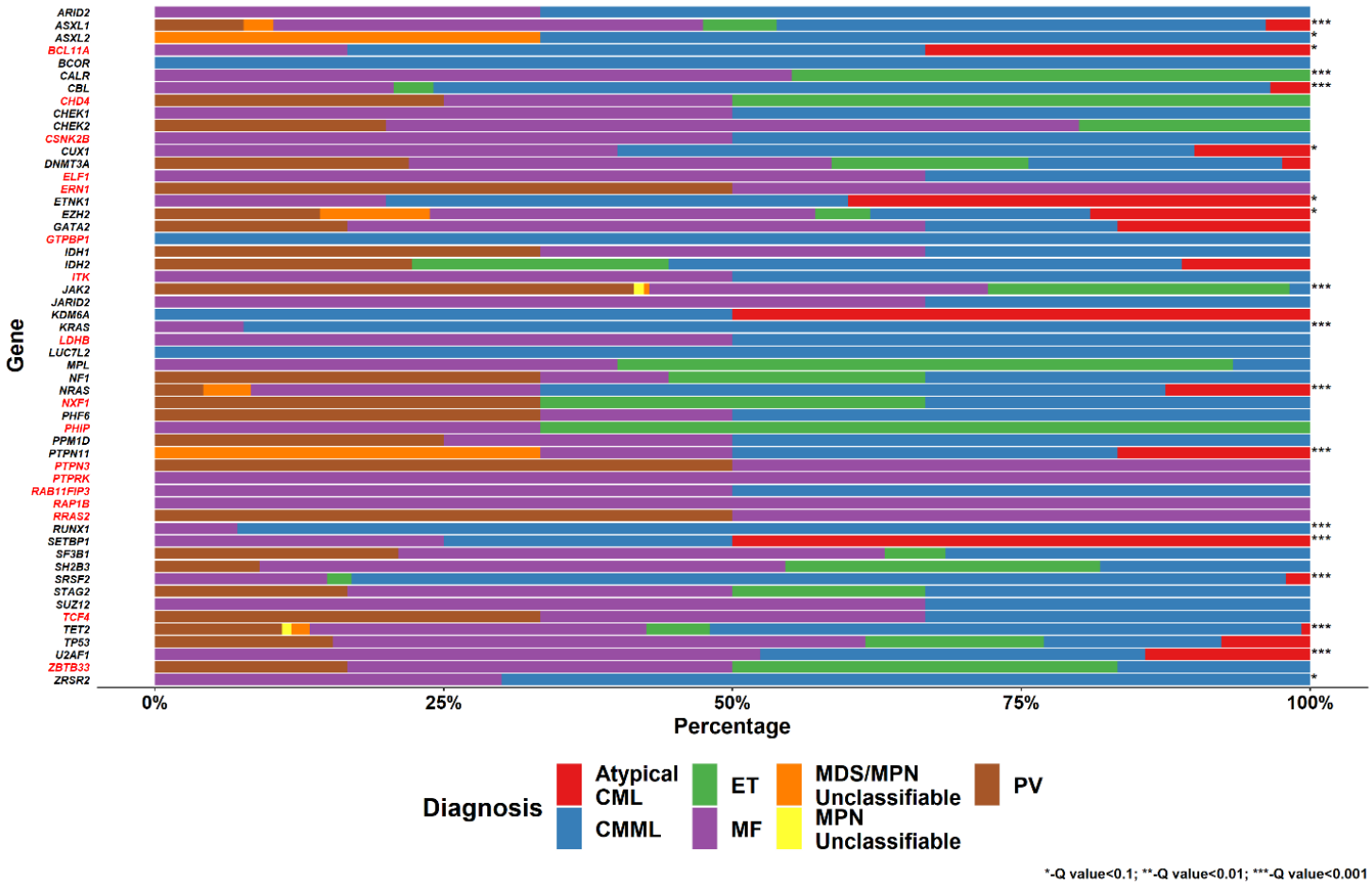
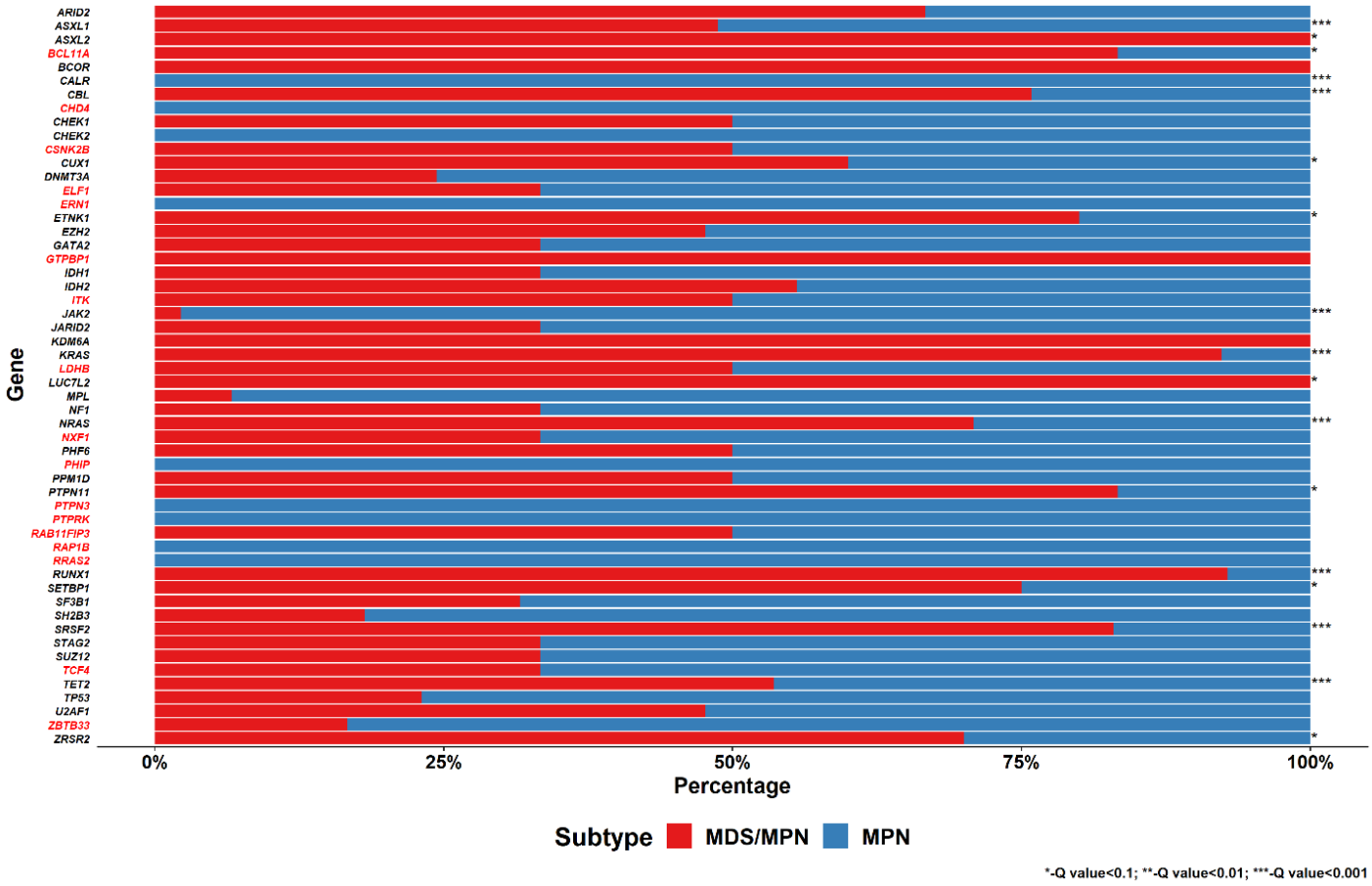


Figure 8 - Distribution of Driver Genes among different Disease Subtypes



4.2 Driver Gene Function Annotation

Through functional annotation of the list of 54 driver genes utilizing the web platform DAVID and the published medical literature, 41 of the 54 putative driver genes (75.9%) could be grouped into 7 distinct groups representing biological pathways or biological function that are altered in these neoplasms, including: JAK-STAT (*JAK2, CALR, MPL,*

SH2B3), DNA methylation (*TET2, DNMT3A, IDH1, IDH2*), histone modification (*ASXL1, ASXL2, ARID2, EZH2, SUZ12, JARID2, SETBP1, CHD4, KDM6A*), mRNA splicing (*SRSF2, SF3B1, ZRSR2, U2AF1*), TP53 activation and regulation (*TP53, PPM1D, CHEK1, CHEK2, CSNK2B*), RAS pathway (*NRAS, KRAS, PTPN11, CBL, NF1, RRAS2, RAP1B*) and DNA-binding proteins that regulate transcription (*BCOR, CUX1, GATA2, PHF6, RUNX1, BCL11A, ZBTB33, TCF4*).

4.2.1 Mutations in genes that activate the JAK-STAT pathway

A total of 282 patients (70%) had mutations in genes that lead to constitutive activation of the cytokine signaling JAK-STAT intracellular pathway. There were 33 distinct non-silent mutations in 4 genes (*JAK2, CALR, MPL, SH2B3*). The most common mutation was the p.V617F mutation in the *JAK2* gene, found in 217 patients (53.8%), followed by insertions/deletions in exon of the *CALR* gene, found in 48 cases (11.9%). The *MPL* gene was mutated in 14 patients (3.47%), and in 11 cases the mutation was in the tryptophan residue W515, in 1 case in serine residue S505, and the remaining 2 cases in non-canonical sites (p.S204P and p.R592Q). One patient presented with two *MPL* gene mutations, both on residue W515, being one the canonical mutation p.W515R and the other the nonsense mutation p.W515*. The negative regulator of the JAK-STAT pathway gene *SH2B3* was altered in 11 patients (2.73%), including 4 cases with frameshift insertions/deletions, 5 nonsynonymous mutations and 2 splice-site mutations. Mutations in JAK-STAT related genes were found almost exclusively in patients with MPN (274 of 282 patients, 92.7%; OR=106.13, 95% CI 45.9-278) (Figure 9). There was high rate of

mutual exclusivity among genes in this pathway, with only 9 of the 282 (3.1%) patients presenting with mutations in 2 genes of the same pathway, and in 8 of these 9 cases the combination comprised the *JAK2* p.V617F mutation with another mutation in the *SH2B3* gene (Figure 10).

Figure 9 - Diagnosis breakdown among patients with JAK-STAT genes mutations

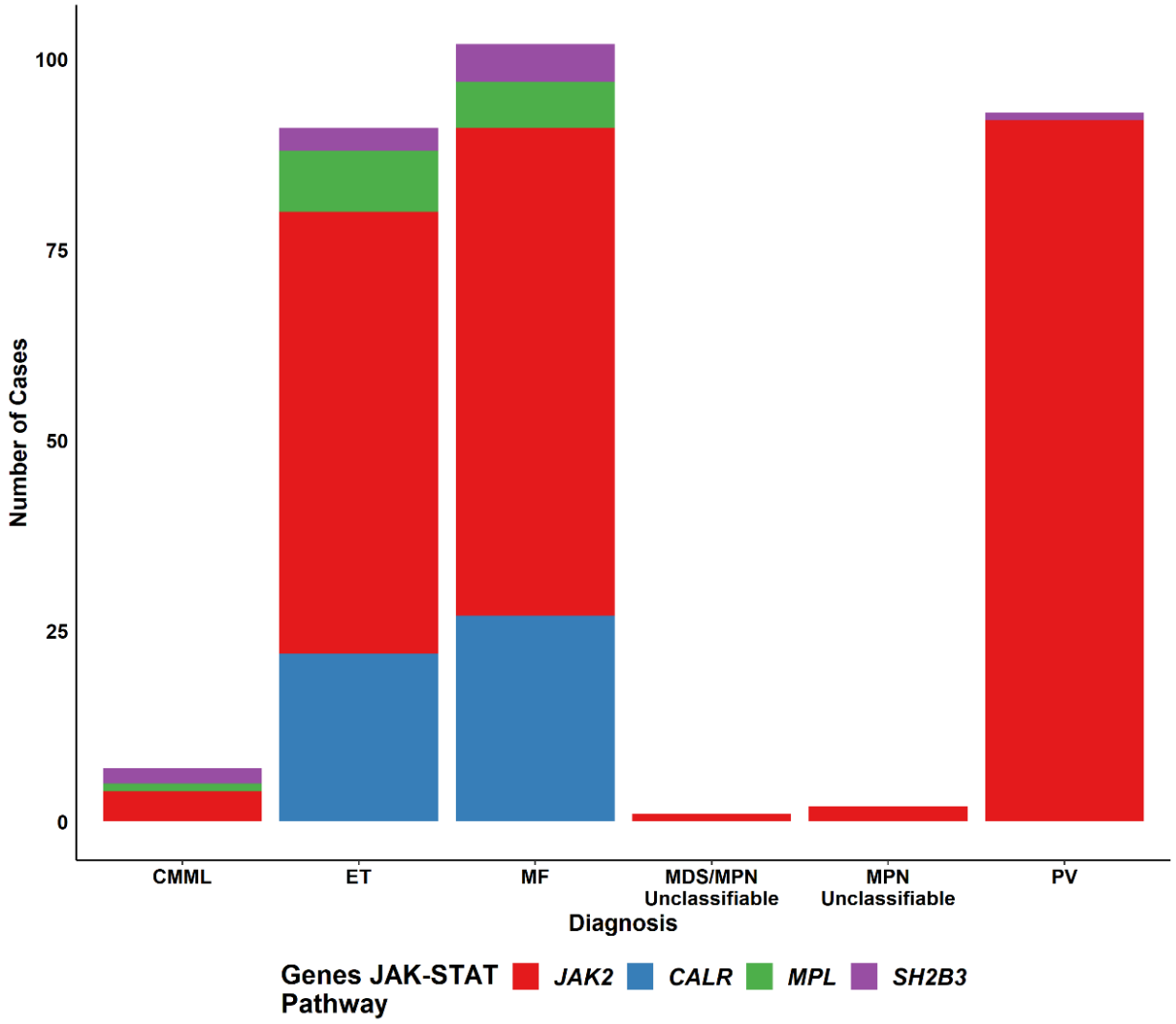
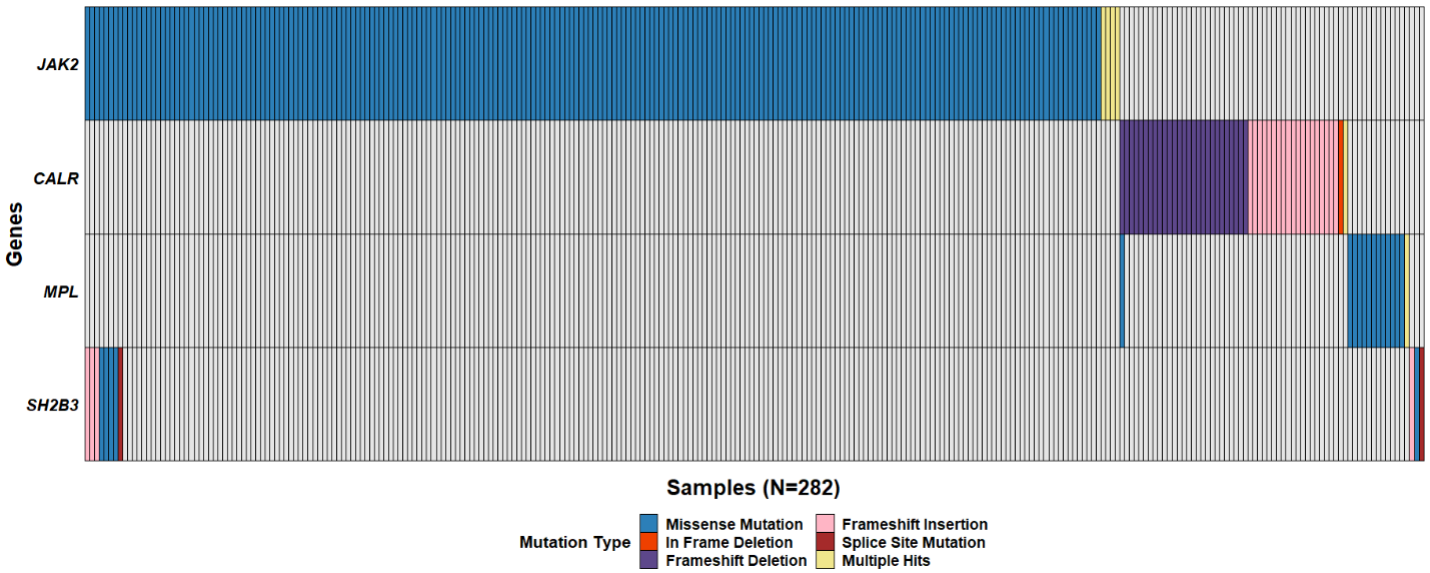


Figure 10 - Mutations in JAK-STAT pathway genes



4.2.2 Mutations in genes that alter DNA methylation

A total of 132 patients (32.7%) carried 140 distinct non-silent mutations in the following 4 genes that alter DNA methylation when mutated: *TET2, DNMT3A, IDH1* and *IDH2*. *TET2* was the most commonly mutated gene, with 91 patients (22.6%) having a *TET2* mutation. *TET2* mutations most often (77.6% of the mutations) led to generation of a premature stop codon and protein truncation, consistent with loss of function mutations. *TET2* mutations were more common in MDS/MPN patients (46% vs. 15.5%; OR=4.61, 95% CI 2.71-7.88) (Figure 11). There were 33 patients who had 2 mutations on *TET2*, the most frequent driver gene in the whole cohort to harbor double mutations. Among *TET2*-mutated patients, double-mutants were more common in patients with MDS/MPNs compared to MPNs (OR=2.96, 95% CI 1.13-8.11). There was no difference in the

prevalence of *DNMT3A, IDH1* and *IDH2* mutations among patients with MDS/MPN and MPNs. Among the 132 patients with DNA methylation-related driver gene mutations, most (N=123, 93%) only carried mutations in one gene, indicating a high degree of mutual exclusivity. Of the 9 patients who carried mutations in 2 distinct genes of this group, in 8 cases the combination of mutations was among the *TET2* and the *DNMT3A* gene, and in 1 case among the *IDH2* and *DNMT3A* gene (Figure 12).

Figure 11 - Diagnosis breakdown among patients with mutations in genes related to DNA methylation

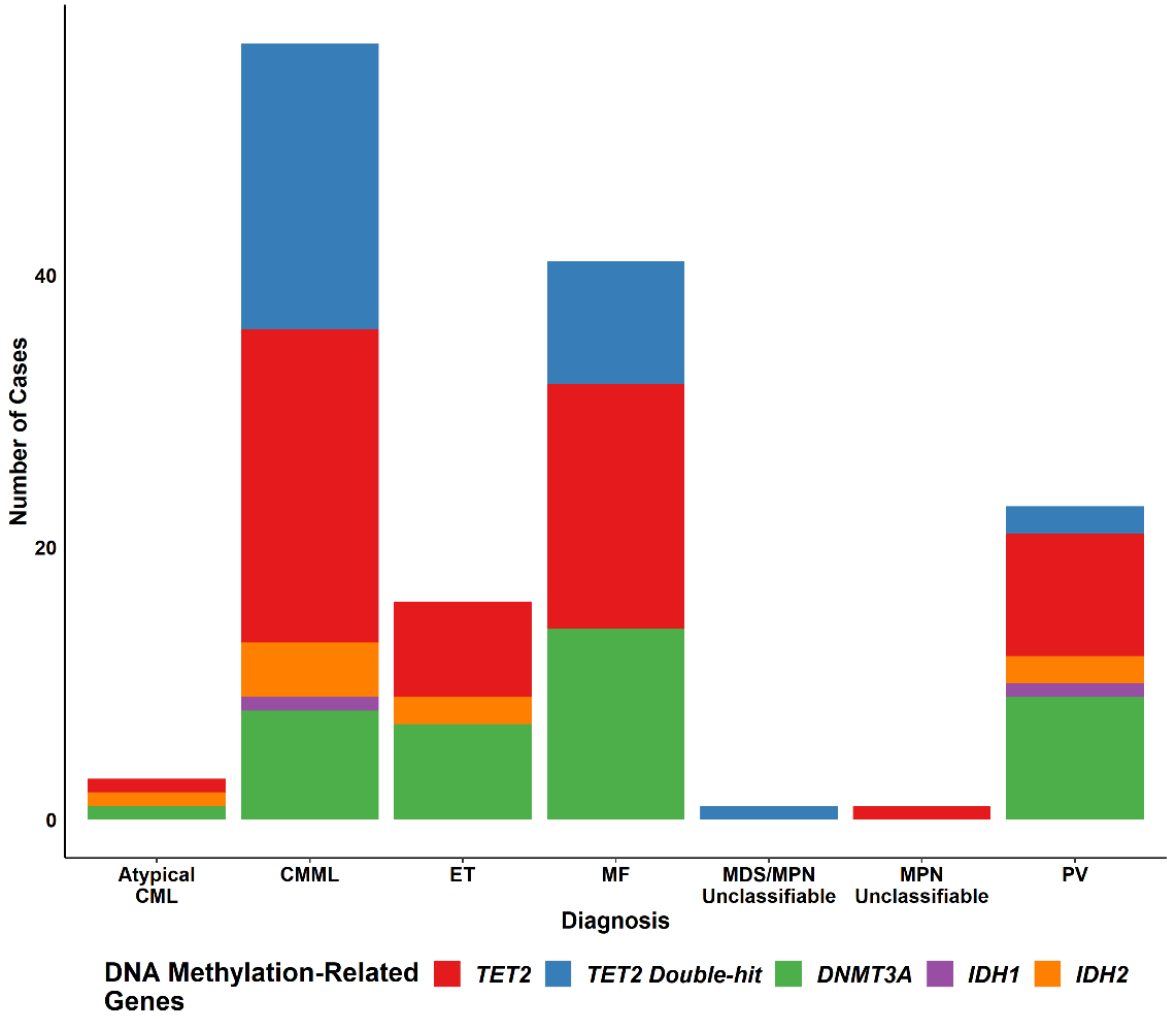
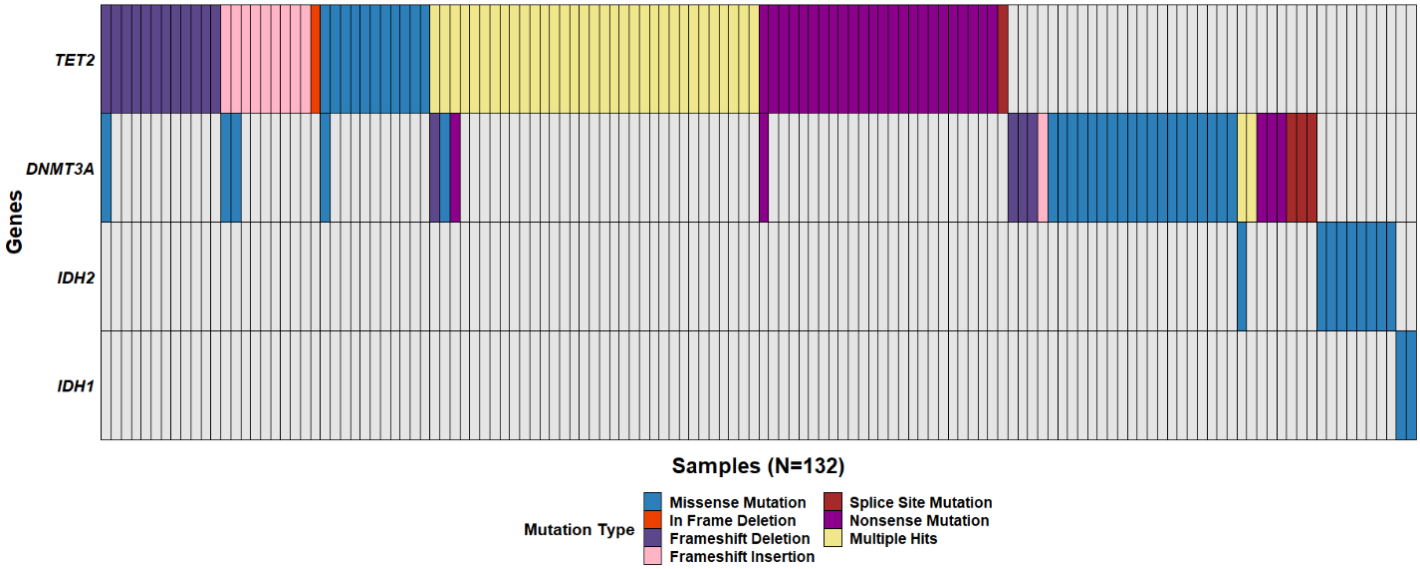


Figure 12 - Mutations in DNA methylation related genes



4.2.3 Mutations in genes that modify histones

Mutations were found in nine genes (*ASXL1, ASXL2, EZH2, JARID2, SUZ12, SETBP1, KDM6A, ARID2, CHD4*) that lead to altered histone modification and epigenetic abnormalities. There were 79 distinct mutations found in 98 patients (24%) and the most commonly mutated genes of this group were *ASXL1* (N=76), *EZH2* (N=16) and *SETBP1* (N=8). There were 4 patients with missense mutations in the gene encoding the chromodomain helicase DNA-binding protein 4 (*CHD4*). *CHD4* encodes a protein that is involved in ATP-dependent chromatin remodeling, and somatic missense mutations in the *CHD4* gene have been reported in patients with serous endometrial tumors.(LE GALLO; O'HARA; RUDD; URICK *et al.*, 2012) Three patients harbored mutations in the *ARID2* gene, another gene responsible for chromatin remodeling that has been previously

reported to be mutated in myeloid malignancies.(SAKAI; HOSONO; PRZYCHODZEN; CARRAWAY *et al.*, 2014) Mutations in histone modifying genes were found in all diagnosis, but were more common in MDS/MPNs compared to MPNs (OR=5.63, 95% CI 3.32-9.63) (Figure 13). Among the 2 most commonly mutated genes (*ASXL1* and *EZH2*), there were 8 cases that harbored mutations in both genes. This corresponds to 61% of *EZH2*-mutated patients, suggesting that despite both genes being associated with histone modifying activity, they may play distinct or synergic roles in disease pathogenesis (Figure 14).

Figure 13 - Diagnosis breakdown among patients with mutations in genes related to histone modification

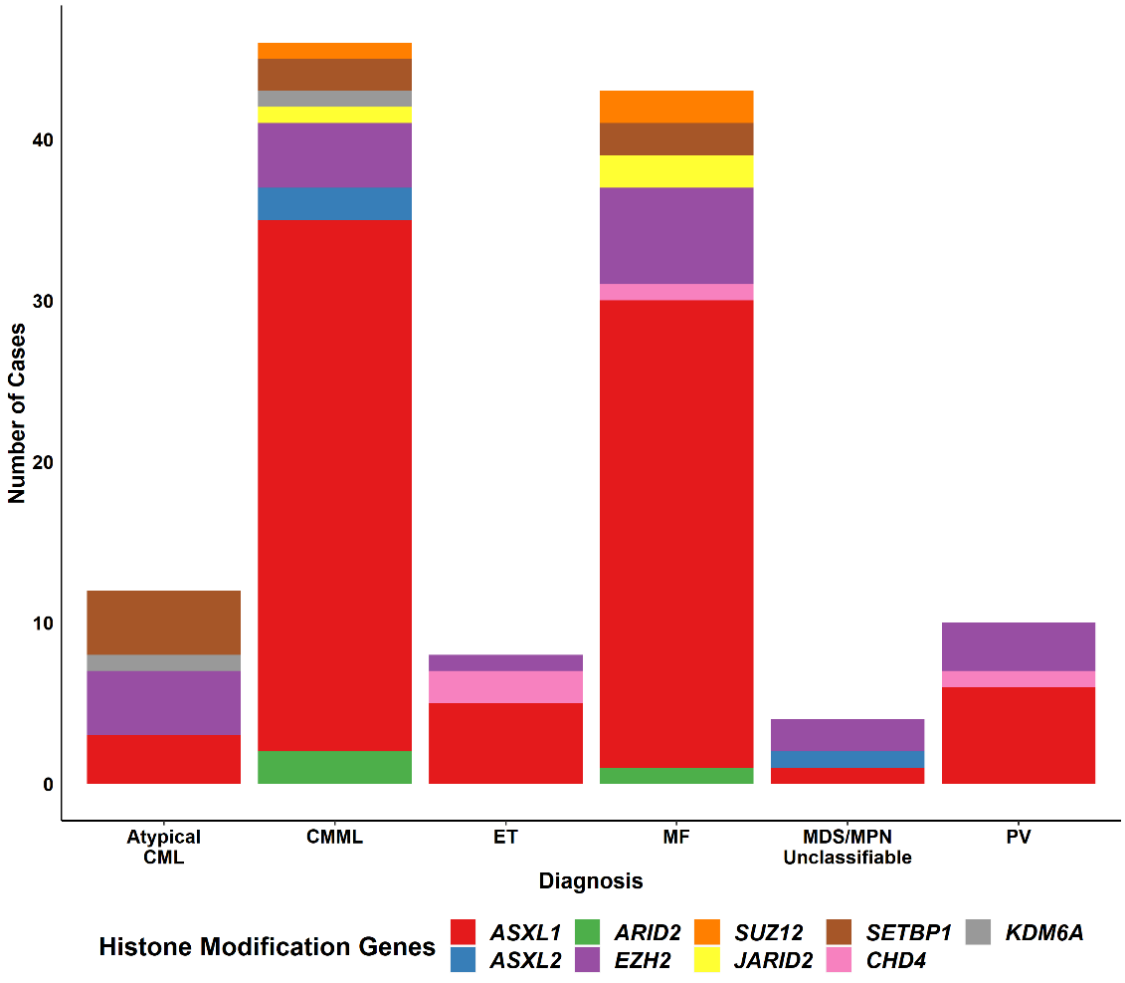
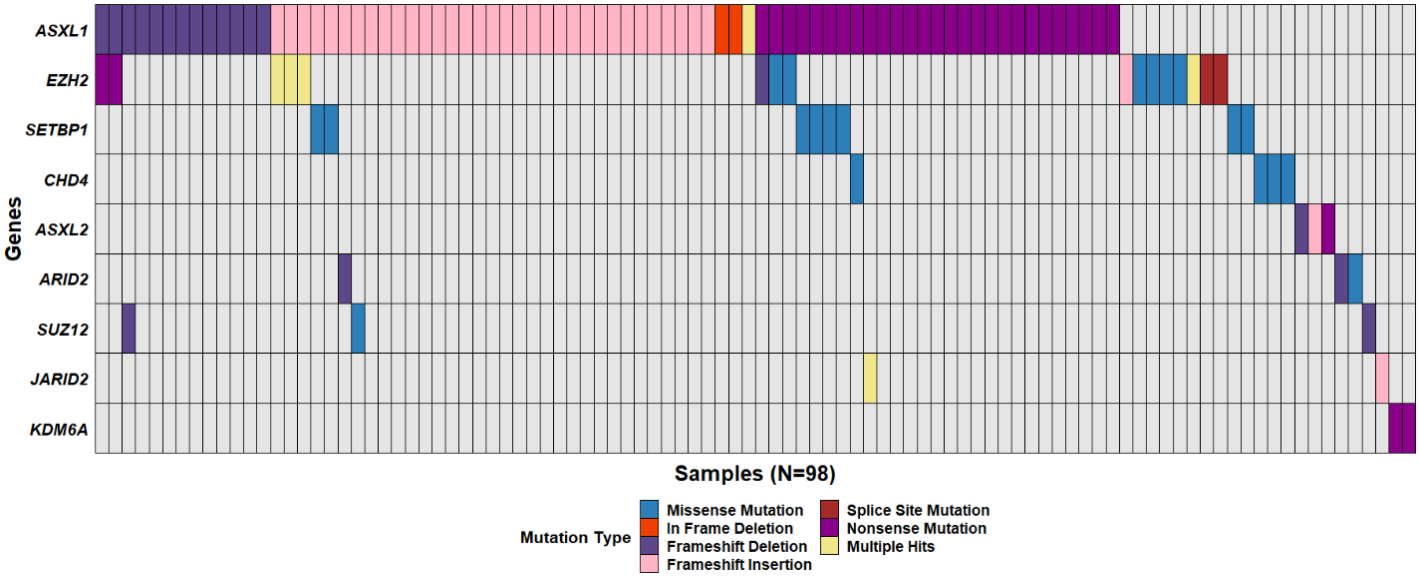


Figure 14 - Mutations in histone modification related genes



4.2.4 Mutations in genes that participate in mRNA splicing

The mRNA splicing machinery comprehends a diverse set of proteins, several of which have been known to be altered in myeloid malignancies. There were 94 patients who presented with 29 distinct non-silent somatic mutations in 4 different splicing-related genes (*SRSF2, SF3B1, U2AF1, ZRSR2*) (Figure 15). Mutations in the mRNA splicing gene *SRSF2* and *ZRSR2* were more common in MDS/MPNs than in MPNs (OR=23.29, 95% CI 10.10-60.75 [SRSF2]; OR=9.53, 95% CI 1.67-98.1). There were no significant differences in the prevalence of mutations in the other 2 genes when comparing MPNs to MDS/MPNs: *U2AF1* (OR=2.61, 95% CI 0.92-7.19), *SF3B1* (OR=1.42, 95% CI 0.43-4.14). This difference in prevalence of prevalence mRNA splicing gene mutations among MPNs and MDS/MPNs was largely explained by a very high prevalence of mutations in these

genes in patients with CMML: 57 patients (69.1%) with this neoplasm presented with mutations in genes of this group, most commonly *SRSF2* (38 of 57 patients, 66%). Among the 94 patients with splicing-related gene mutations, no patient presented with mutations in 2 genes of this pathway (Figure 16). This suggests that these genes play similar roles in oncogenesis, or that mutations in 2 or more of these genes are lethal to the cell. There were 3 patients with mutations in the splicing related gene *LUC7L2*. However, these 3 patients also harbored mutations in other splicing genes (*SRSF2, U2AF1* and *ZRSR2*, one case each). This suggests that *LUC7L2* gene mutations may not play a significant role in disease pathogenesis, but this needs further exploration with functional studies.

Figure 15 - Diagnosis breakdown among patients with mutations in mRNA splicing related genes

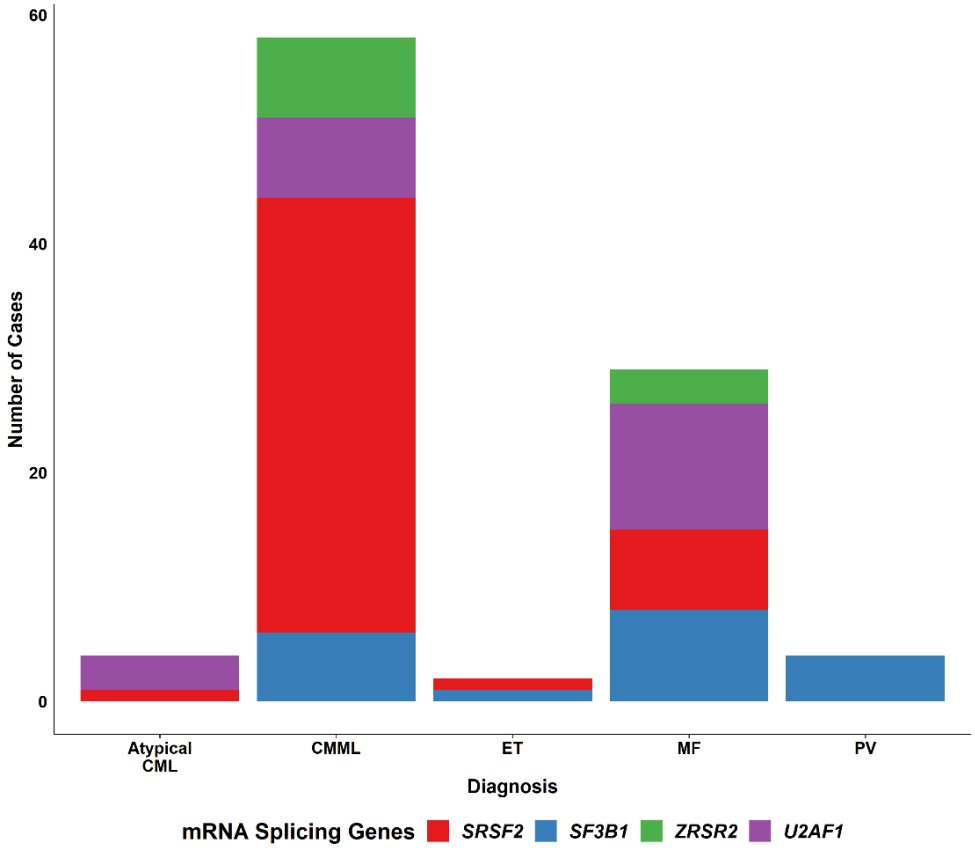
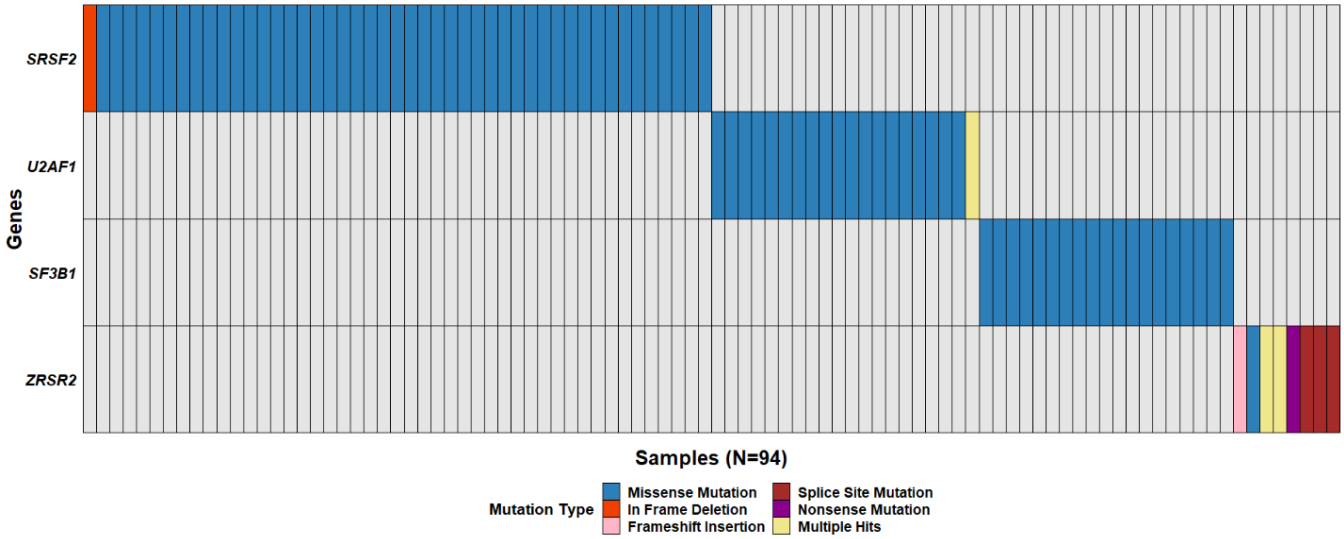


Figure 16 - Mutations in mRNA splicing related genes



4.2.5 Mutations in genes of the RAS pathway

Gene mutations that activate the RAS-RAF-MAPK-ERK pathway are known to be found in patients with MDS/MPNs such as CMML (KOHLMANN; GROSSMANN; KLEIN; SCHINDELA *et al.*, 2010). Eighty non-silent mutations were found in the following 5 genes known to activate the RAS signaling pathway (*NRAS, KRAS, CBL, PTPN11, NF1*) in 65 patients (16.1% of the cohort). There were 4 additional mutations in two other genes of the RAS superfamily (*RRAS2* and *RASP1B*). Thus, mutations that activate RAS or similar genes were found in 16.8% of patients. These mutations were found in most subtypes of disease, not being exclusive to patients with CMML (Figure 17). Most mutations found in the RAS pathway were mutually exclusive, although in some cases the same patient could harbor mutations in two different RAS pathway genes (Figure 18). In 9 patients

there were multiple mutations in the same gene, most commonly in the *CBL* gene (N=6), which is expected since it is a tumor suppressor gene.

Figure 17 - Diagnosis breakdown among patients with mutations in RAS pathway genes

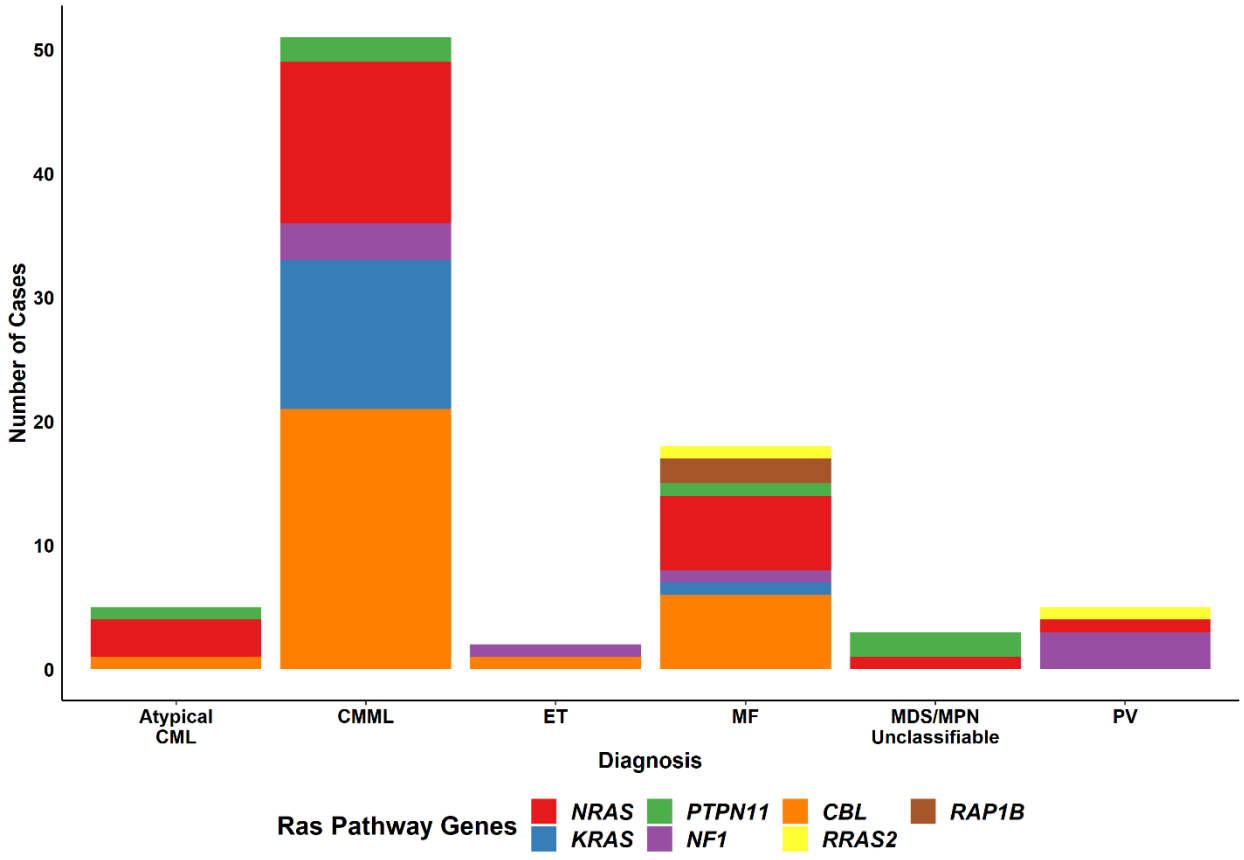
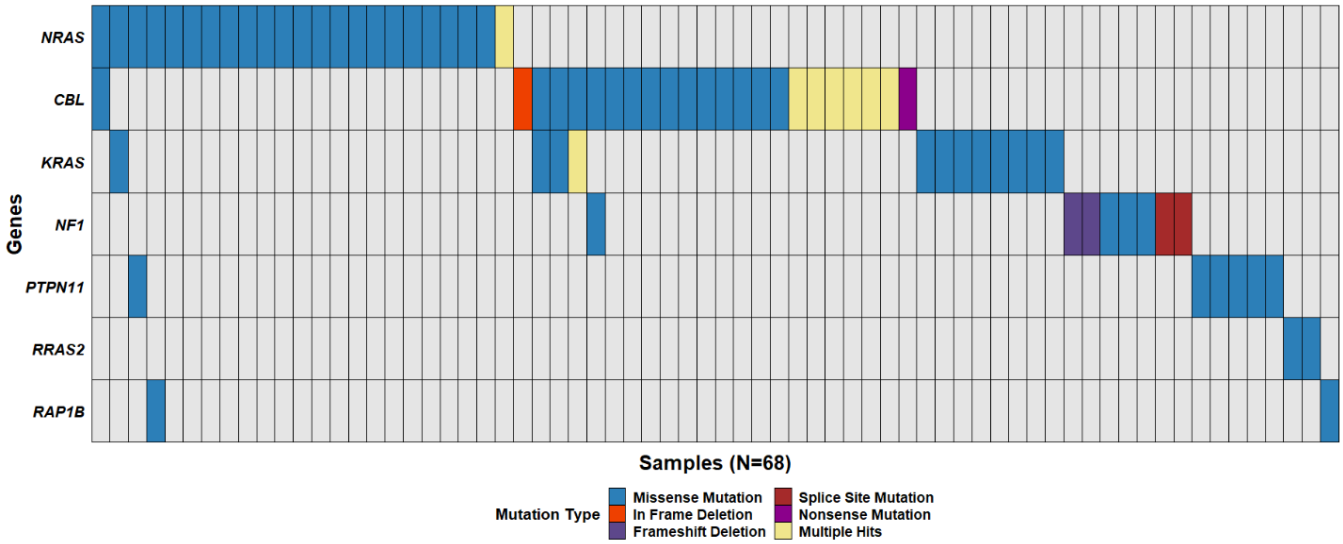


Figure 18 - Mutations in RAS pathway genes



4.2.6 Mutations in Hematopoietic Transcription Factors

There were 46 patients (11.4%) with mutations in genes that encode DNA-binding proteins that regulate transcription. Patients with MDS/MPNs more frequently presented with these mutations (OR 6.12, 95% CI 3.07-12.46), with CMML being the most common diagnosis (24 of 46 patients, 52%). The most commonly mutated genes were *CUX1* and *RUNX1* (N=9 each), followed by *GATA2, PHF6, BCL11A* and *ZBTB33* (N=6 each). All genes could be found to be mutated both in patients with MDS/MPN and MPNs, except for *RUNX1*, that was found to be mutated solely in CMML (Figure 19). Three of 9 patients with *CUX1* mutations presented with an elevated VAF (>90%) suggesting the presence of uniparental disomy (UPD) of the long arm of chromosome 7 (ALY; RAMDZAN; NAGATA; BALASUBRAMANIAN *et al.*, 2019). One of these patients also carried a high VAF mutation in the *EZH2* gene, also located in the long arm of chromosome, another indication of UPD of 7q. Among 35 patients, only 2 patients presented with 2 mutations

in distinct hematopoietic transcription factor genes (Figure 20). In both cases it was a combination of *RUNX1* mutation with another gene (*GATA2* and *ZBTB33*, one each).

Figure 19 - Diagnosis breakdown among patients with mutations in transcription factor genes

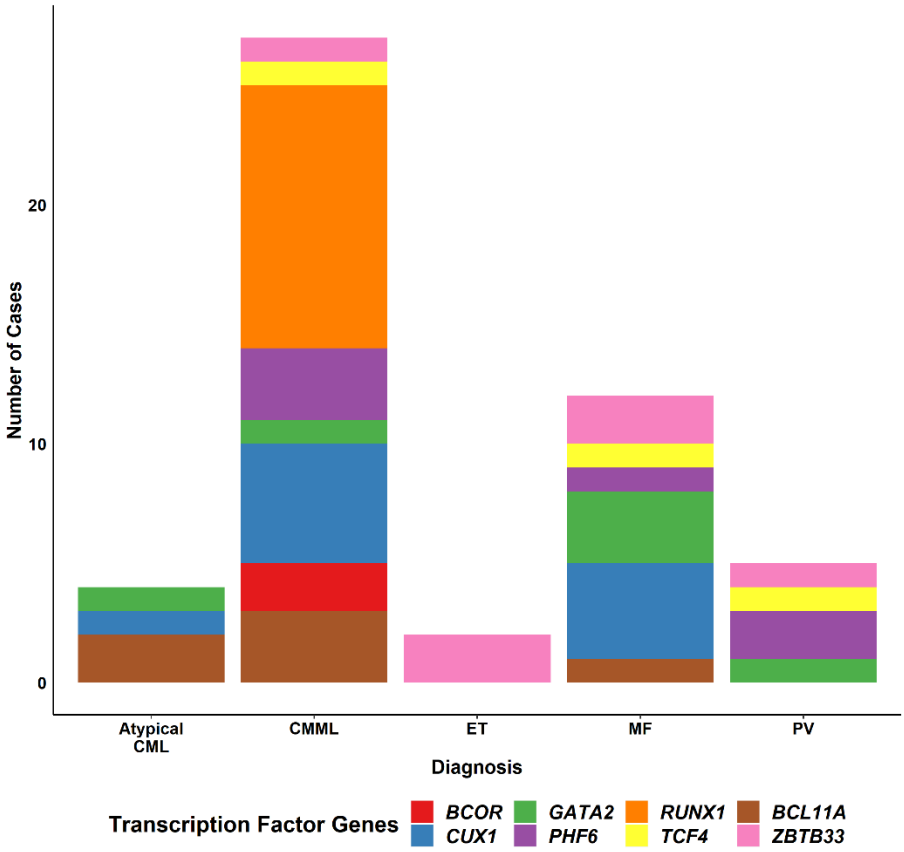
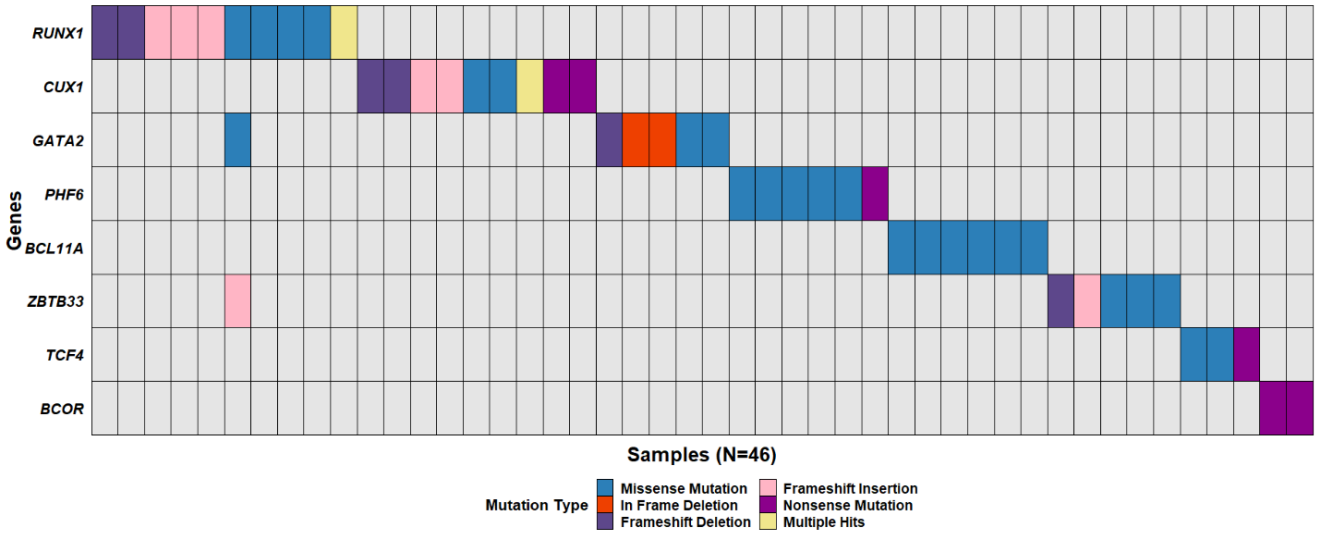


Figure 20 - Mutations in transcription factor genes



4.2.7 Mutations in genes related to TP53

Mutations that lead to *TP53* inactivation or reduced activity are found in several types of cancer (KANDOTH; MCLELLAN; VANDIN; YE *et al.*, 2013). In 25 patients (6.2% of the cohort), there were mutations in the following 4 genes that may alter TP53 function or participate in TP53 activation: *TP53, PPM1D, CHEK1* and *CHEK2*. There were 4 cases (CMML=2, MF=1, PV=1) with *PPM1D* mutations, all nonsense or truncating frameshift indels in the amino-terminal region, that have been shown to lead to increase phosphatase activity and reduced TP53 function.(DUDGEON; SHREERAM; TANOUE; MAZUR *et al.*, 2013; KAHN; MILLER; SILVER; SELLAR *et al.*, 2018; KLEIBLOVA; SHALTIEL; BENADA; SEVCIK *et al.*, 2013) There were 2 cases with truncating *CHEK1* mutations (CMML=1, MF=1) and 4 cases with nonsilent missense *CHEK2* mutations

(MF=2, PV=1, ET=1). Two patients (CMML=1 and MF=1) had frameshift truncating mutations in the regulatory β subunit of casein kinase 2 (*CSNK2B*) affecting residues Y188 and R178. There was no significant difference in the prevalence of TP53-related gene mutations in MDS/MPNs compared to MPNs (OR=1.19, 95% CI 0.40-3.10, Figure 21). Genes were also not found in the same patients suggesting that they play a similar role in disease pathogenesis (Figure 22).

Figure 21 - Diagnosis breakdown among patients with mutations in TP53 related genes

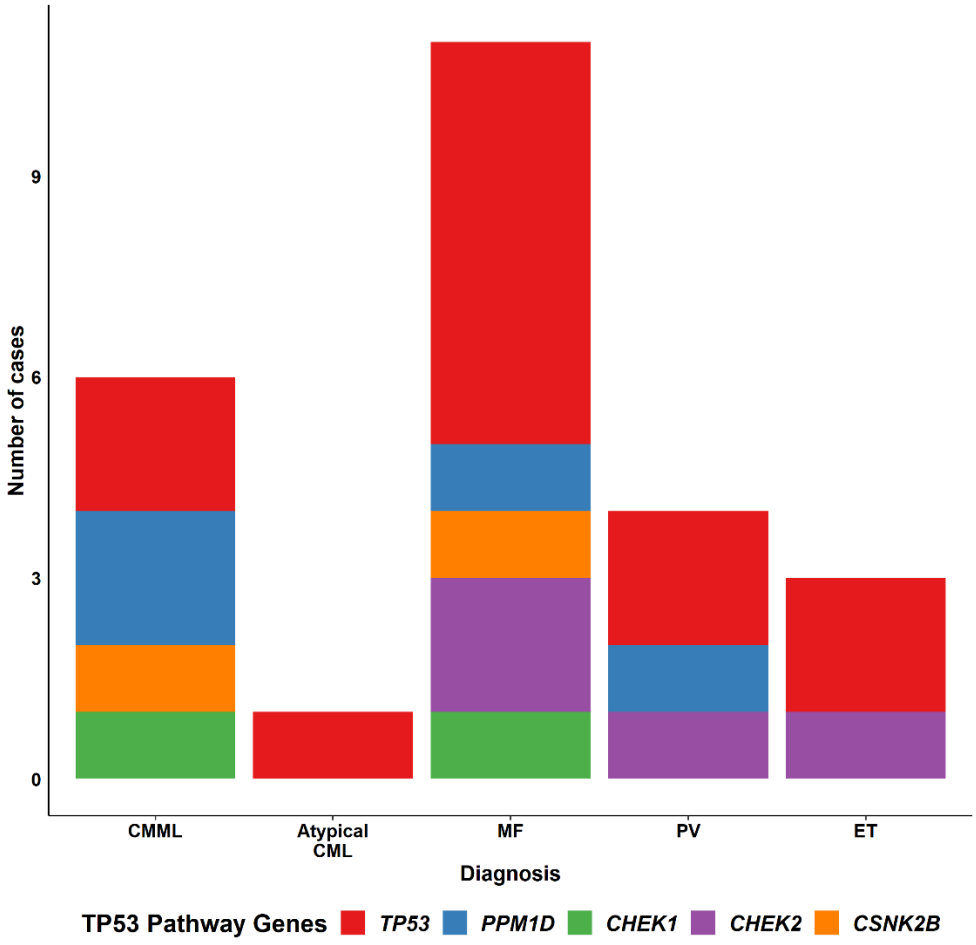
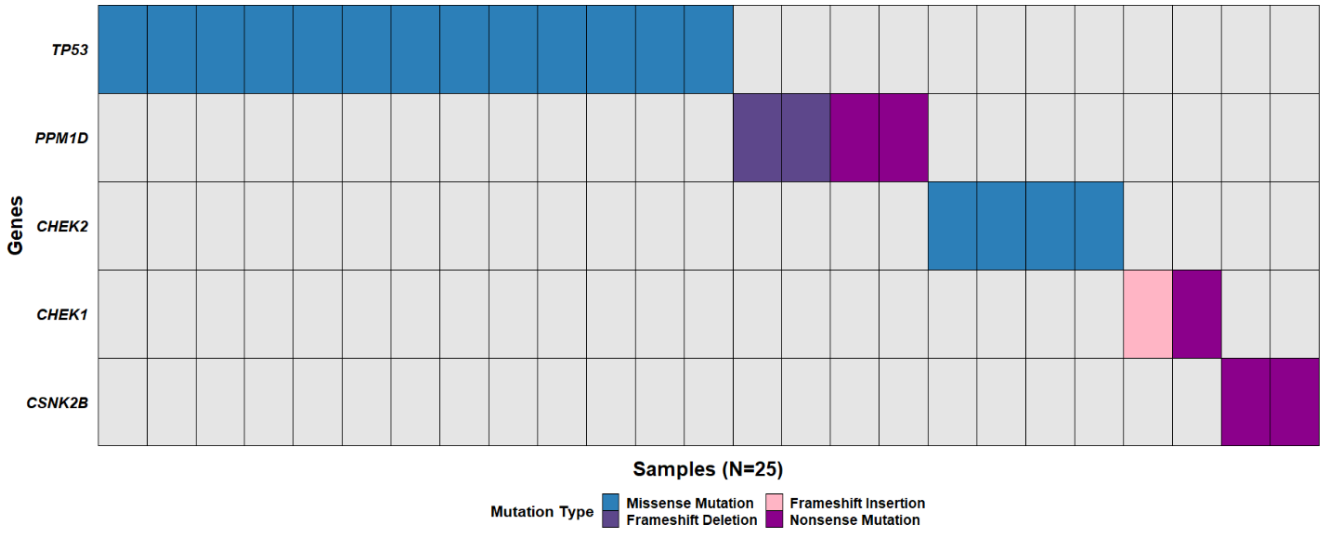


Figure 22 - Mutations in TP53 related genes



4.2.8 Other genes

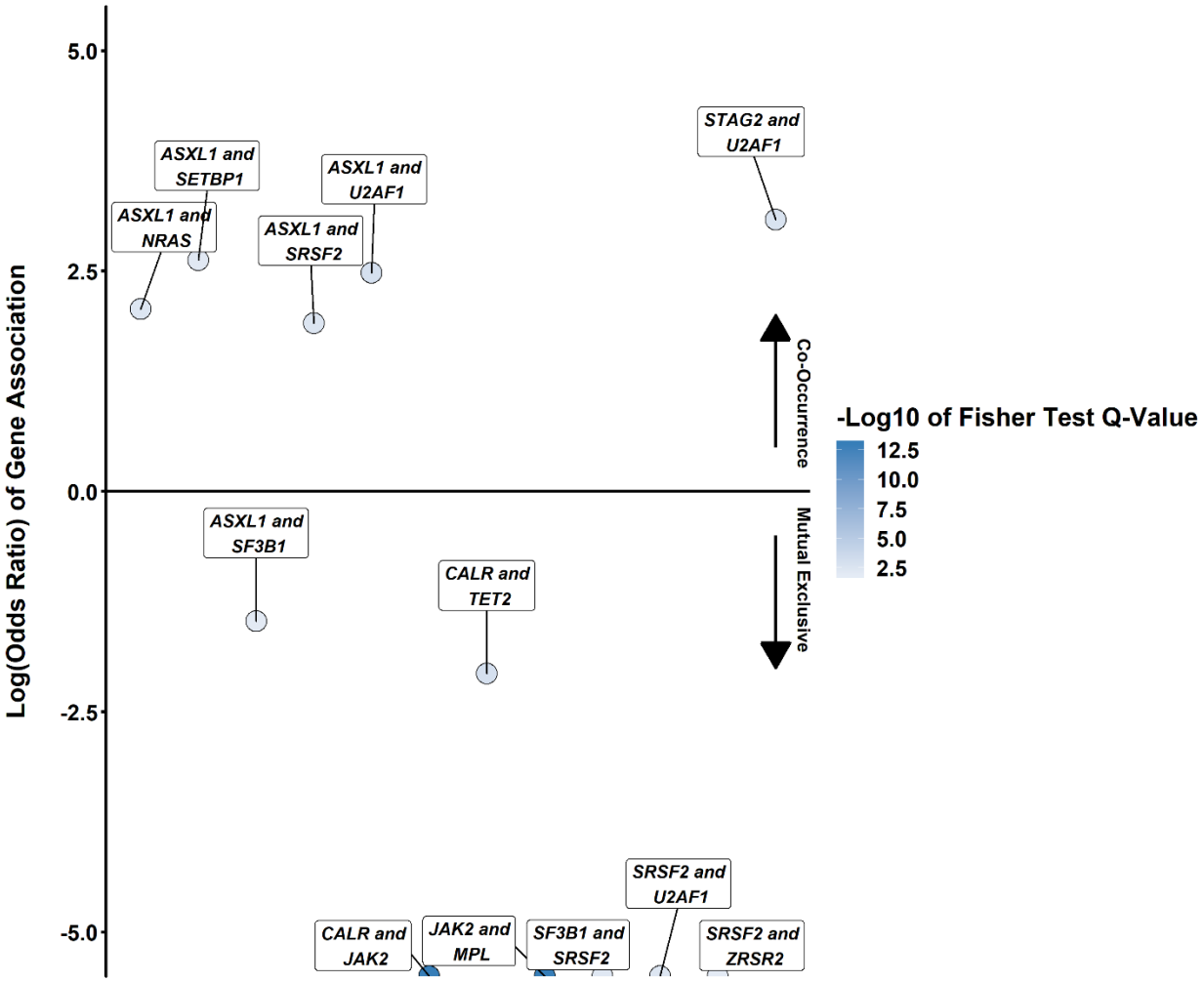
There were other novel recurrently mutated driver genes that are known to be mutated in myeloid malignancies but were not grouped in a single pathway. One of these genes is *STAG2*, that is part of the cohesion complex genes, and was found to be mutated in 6 cases in the cohort, including 4 patients with MPNs and 2 patients with CMML (FISHER; MCNULTY; BURKE; CRISPINO *et al.*, 2017). The other gene is *ETNK1*, the gene that encodes the ethanolamine kinase enzyme. A hotspot mutation in codons H243-N244 is known to be present in MDS/MPNs (GAMBACORTI-PASSERINI; DONADONI; PARMIANI; PIROLA *et al.*, 2015). Indeed, 5 patients (2 with atypical CML, 2 with CMML and 1 with MF) harbored missense mutations in these codons. The other 11 putative driver genes that are previously not known to be mutated in myeloid neoplasms and are not part of a recurrent pathway, were found in 2-3 samples at most. This suggests that

these genes may be false-positives of the driver-detection algorithms (MutSigCV and IntOgen) but further study of larger cohorts would be necessary to confirm this.

4.3 Patterns of mutual co-occurrence and mutual co-exclusivity in gene mutations

After determining which genes and pathways were more commonly mutated in these neoplasms, the next step was to define whether there were patterns of mutations co-occurrence and mutual co-exclusivity among the mutated genes. Utilizing the SELECT algorithm, there were 12 pairs of mutated genes with a statistically significant correlation (Figure 23). There was a significant degree of mutual exclusivity among genes of the JAK-STAT pathway (*JAK2, CALR* and *MPL*) and among genes responsible for mRNA splicing (*SF3B1, SRSF2, U2AF1, ZRSR2*). Novel correlations included positive correlations among *ASXL1/NRAS* (q=0.004), *ASXL1/SETBP1* (q=0.007), *ASXL1/SRSF2* (q=0.003), *ASXL1/U2AF1* (q=0.001) and *STAG2/U2AF1* (q=0.002). Novel negative correlations were seen among genes *ASXL1/SF3B1* (q=0.01) and *CALR/TET2* (q=0.001).

Figure 23 - Gene Pairs with a significant positive or negative association- Entire Dataset



Since some of these interactions may reflect the underlying disease diagnosis, the analysis was repeated considering only MPNs or MDS/MPNs as separate datasets. In the MPN-only dataset, there were 11 significant interactions, whereas with the MDS/MPN dataset (due to smaller numbers), only 3 significant correlations were seen. Results are summarized in Figure 24 and Figure 25.

Figure 24 - Gene Pairs with a significant positive or negative association- MPN Dataset

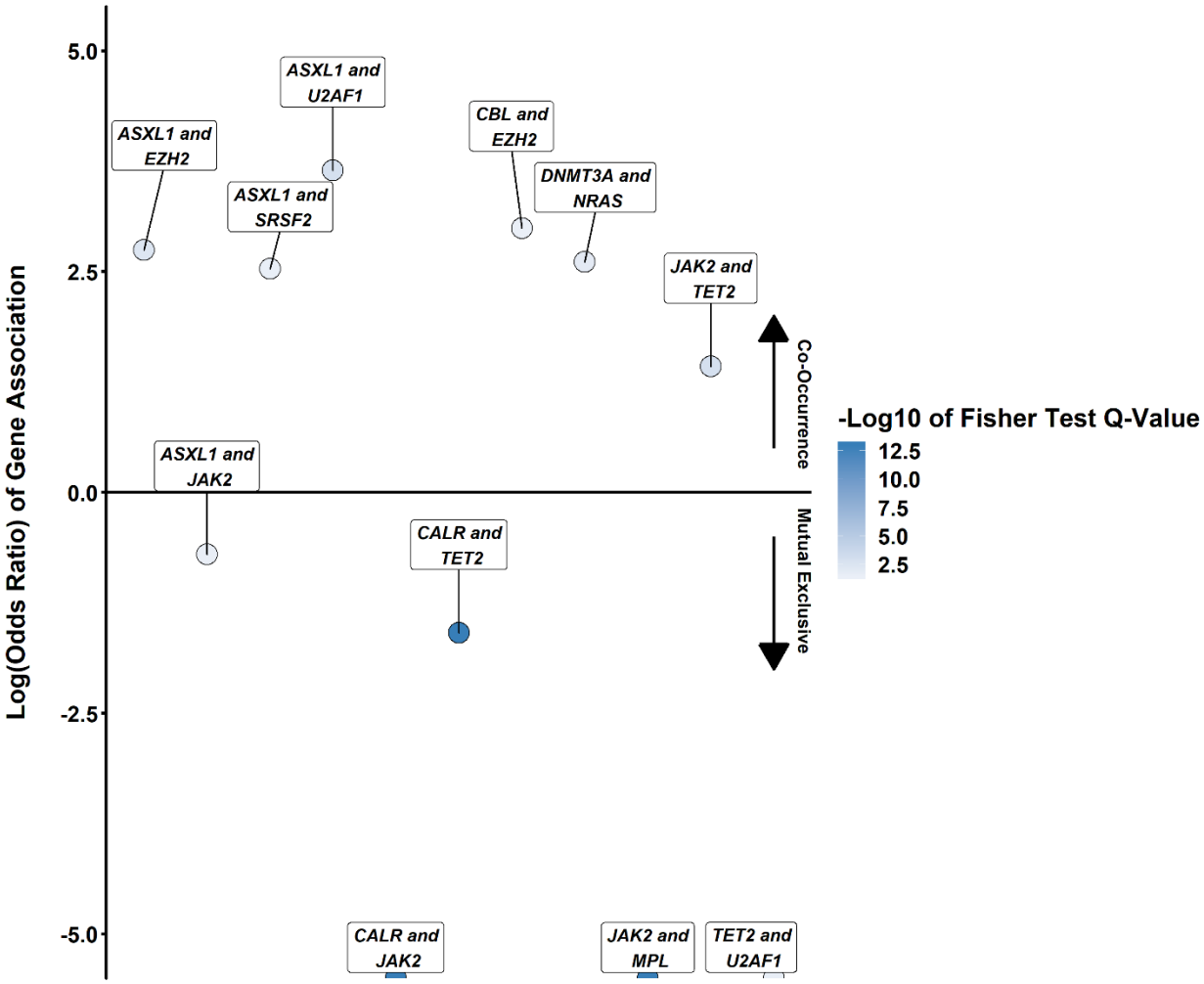
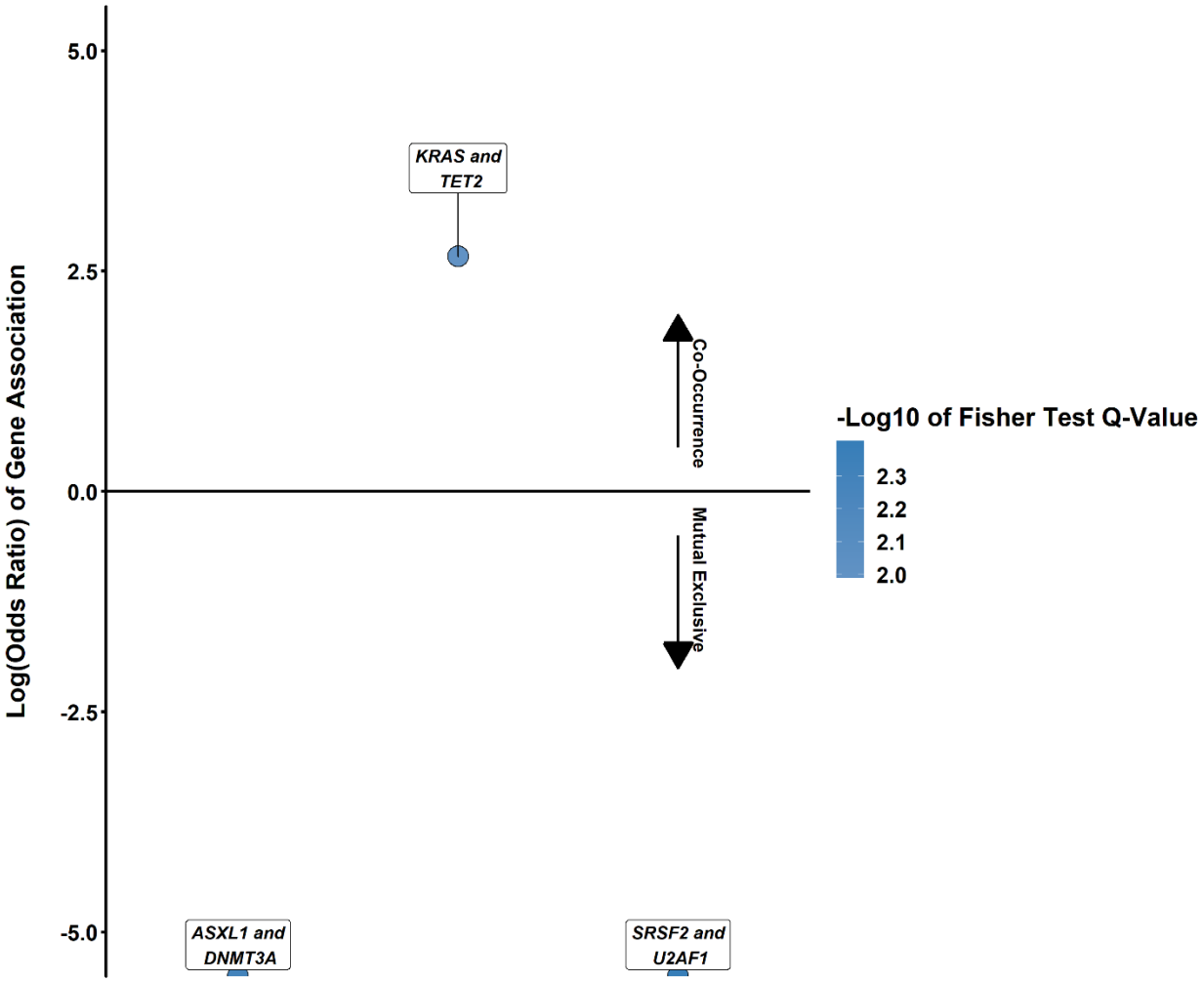


Figure 25 - Gene Pairs with a significant positive or negative association- MDS/MPN Dataset



Analysis of groups of genes with a positive correlation may lead to discovery of molecular subtypes of MPNs and MDS/MPNs. Based on this hypothesis, the 10 gene pairs with a positive correlation were analyzed to determine the distribution of disease diagnosis and subtypes. Q-values were calculated with Fisher Exact Tests. As seen in Figure 26, 7 gene pairs (*'ASXL1-EZH2', 'ASXL1-NRAS', 'ASXL1-SETBP1', 'ASXL1-SRSF2', 'ASXL1-U2AF1', 'JAK2-TET2'* and *'KRAS-TET2')* had a differential frequency among the various

distinct diagnosis. Similarly, 5 gene pairs had a differential distribution among the subtypes of MPNs and MDS/MPNs (Figure 27).

Figure 26 - Diagnosis and Pairwise Gene Co-Occurrences

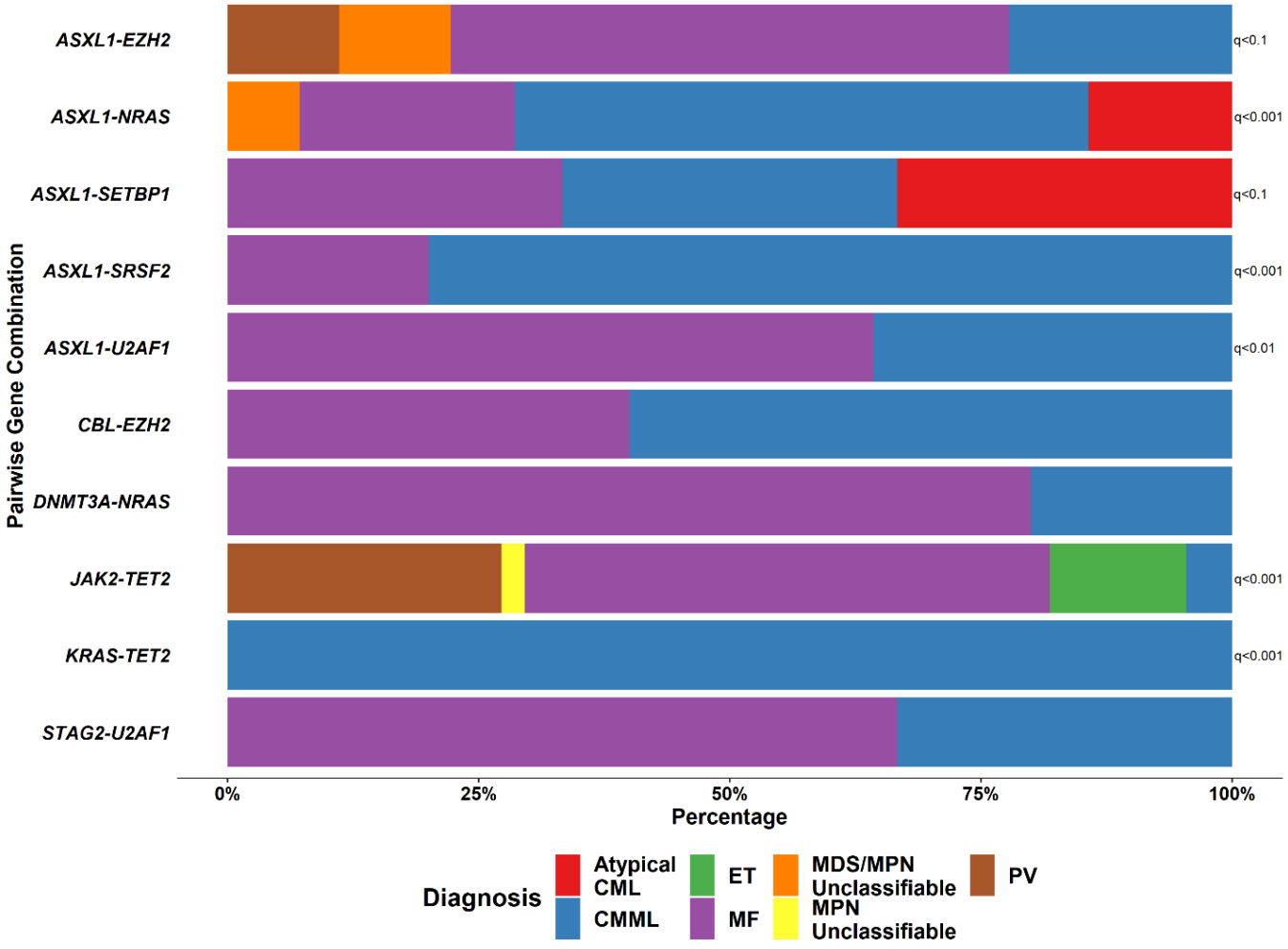
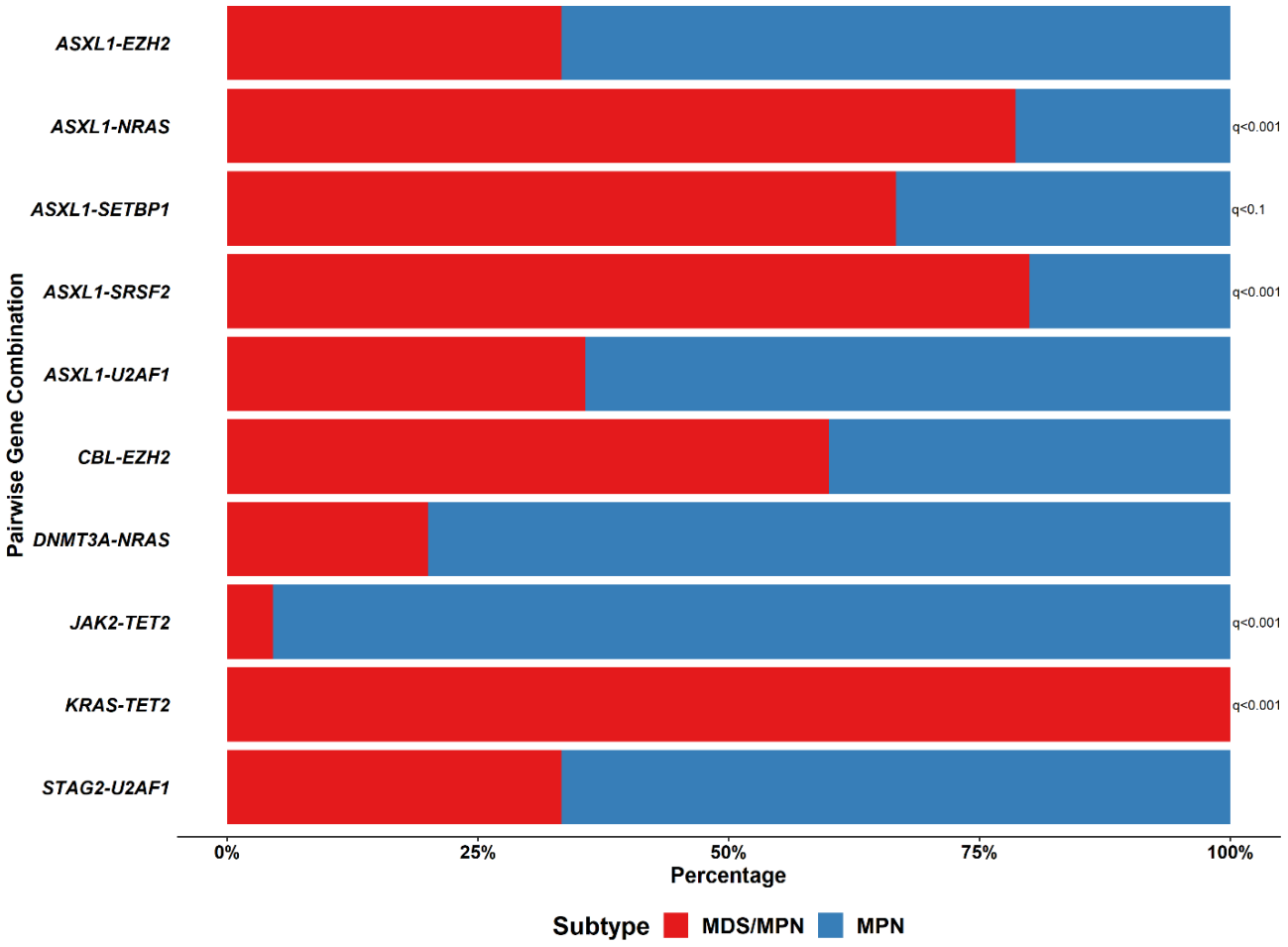


Figure 27 - Subtype and Pairwise Gene Co-Occurrences



Gene sets comprising 3 or more genes can also be seen more frequently than expected by chance. The Apriori Market Basket analysis method was used to discover association rules between 3 or more mutated genes in the dataset. There were 19 gene sets that fulfilled association rules, including 1 set with 4 distinct genes. Among these 19 gene sets, there were five sets that had a statistically significant distinct distribution among the various disease diagnosis (Figure 28) and 6 gene sets that had a distinct distribution

among disease subtypes (Figure 29). The most frequently involved genes were *TET2* (present in 13 of 19 gene sets) and *ASXL1* (present in 12 of 19 gene sets).

Figure 28 - Diagnosis and combinations of 3 or more genes

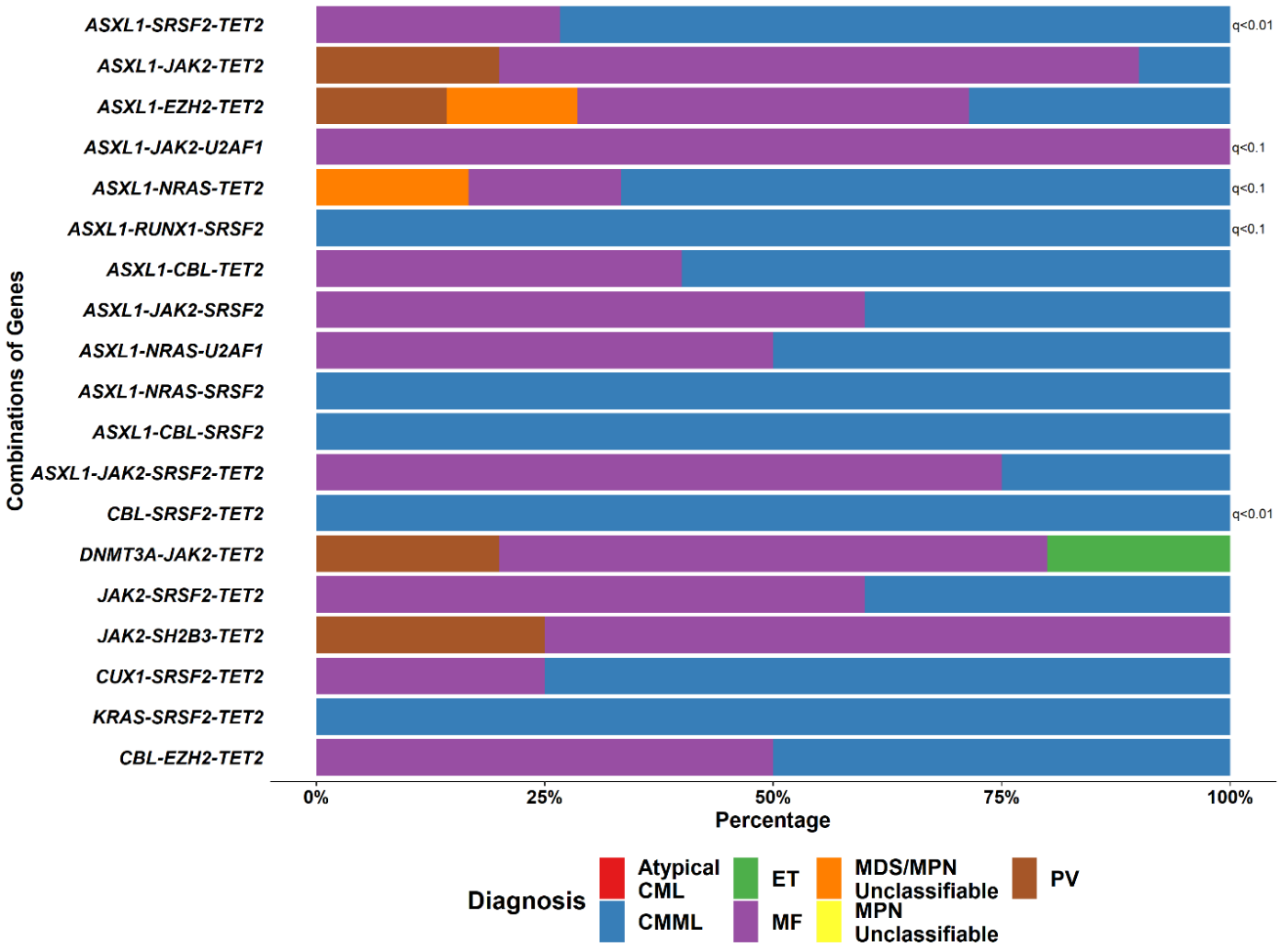
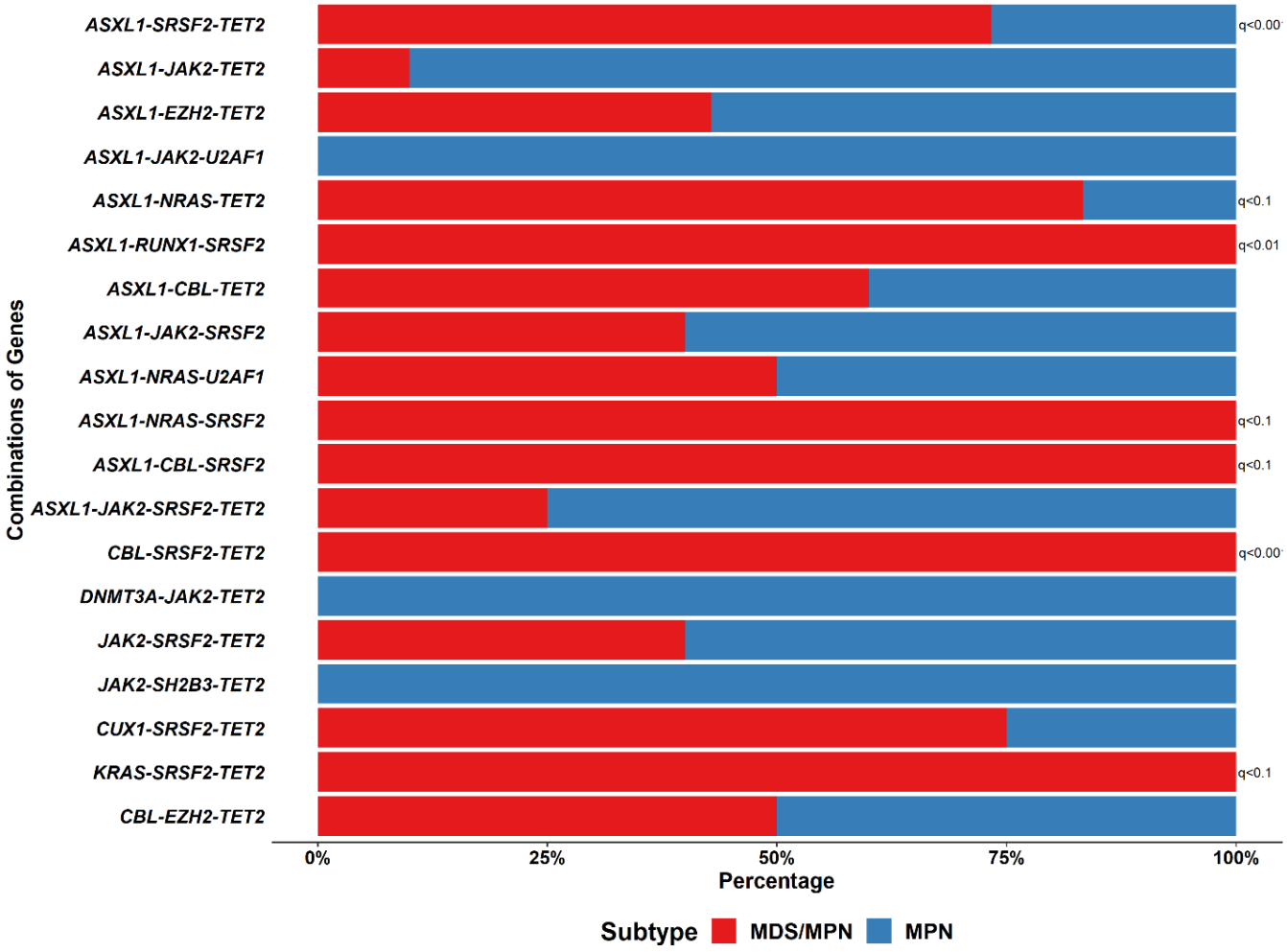


Figure 29 - Subtype and combinations of 3 or more genes

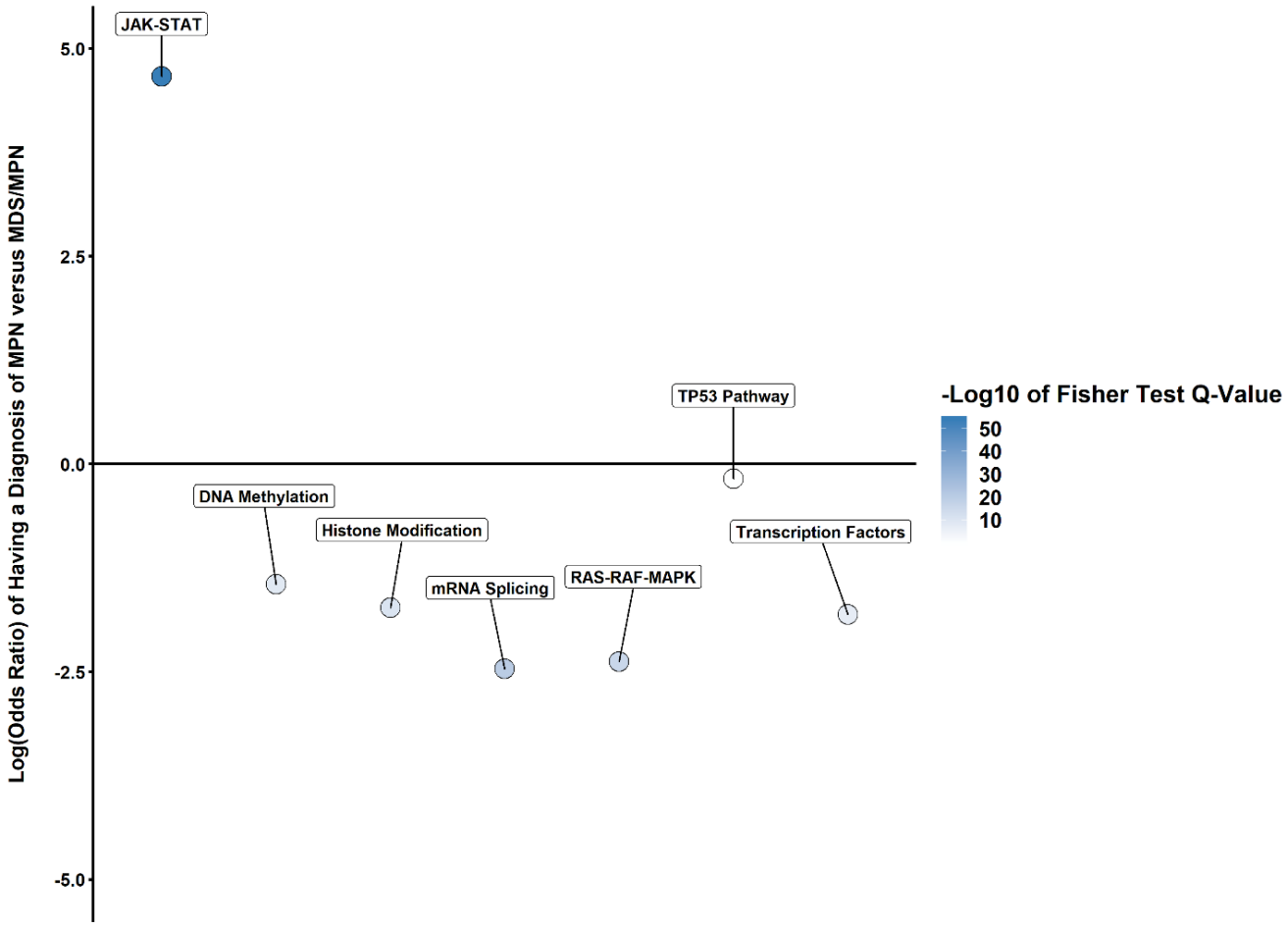


4.4 Analysis of Recurrently Mutated Pathways and Disease Phenotype

Mutations in genes belonging to the same biological pathway may reflect on the disease phenotype and influence the disease category and diagnosis. To analyze this, based on the 7 biological pathways described previously, Fisher tests were employed to determine the measure of association and statistical significance of having a mutation in a specific

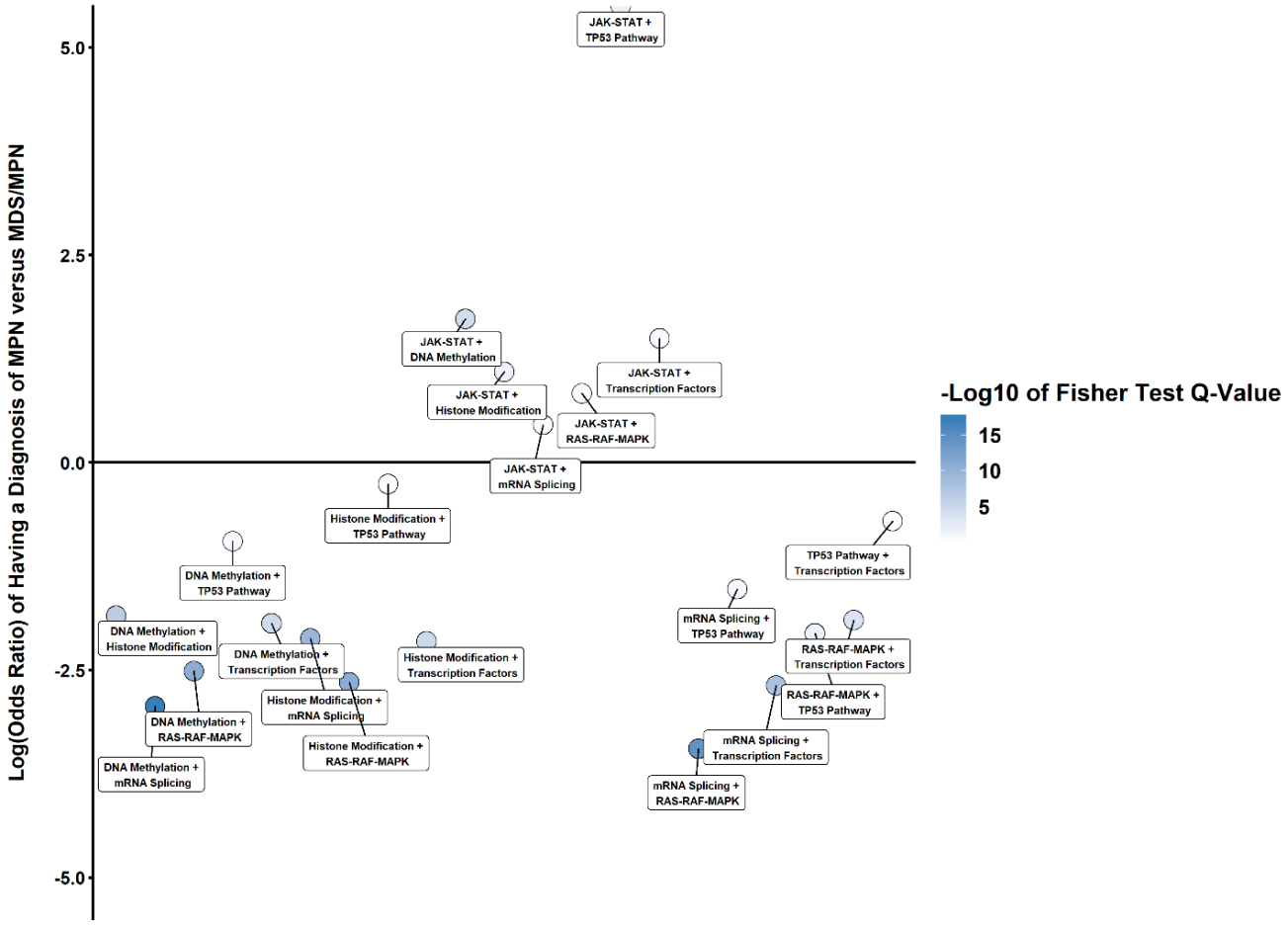
pathway and being diagnosed with MPN or MDS/MPN. Figure 30 shows that the presence of mutations in JAK-STAT pathway was associated with a strong association and a high log(OR) of being diagnosed with MPN compared to MDS/MPNs. Conversely, mutations in the other pathways were associated with a diagnosis of MDS/MPN, with a lower log(OR), particularly mutations in mRNA splicing genes and mutations in genes of the RAS pathway. The strength of the measure of association was not as high as with the JAK-STAT mutations for a diagnosis of MPN, reflecting that mutations in these pathways can be found in both disease groups, albeit more commonly in patients with MDS/MPNs.

Figure 30 - Odds Ratio of having MPN vs MDS/MPNs based on the Biological Pathway that is mutated



The next step was to repeat the analysis, but this time considering pairs of mutations in different pathways. This could determine which combination of biological pathways occurred more frequently in patients with MPNs compared to patients with MDS/MPNs (Figure 31). Like the single pathway analysis, mutations in the JAK-STAT pathway lead to a predominant MPN phenotype, regardless of the accompanying mutated pathway. The strength of the association, as measured by the log(OR) was much lower, perhaps a reflection that combinations of JAK-STAT pathway mutations with other biological pathways are not as common as sole JAK-STAT mutations and thus the association is weaker. Regarding the diagnosis of MDS/MPNs, the strongest associations were between mutations in DNA methylation genes/mRNA splicing genes and *RAS* pathway genes/mRNA splicing genes.

Figure 31 - Odds Ratio of having MPN vs MDS/MPNs based on the Combination of Biological Pathways that are mutated



4.5 Analysis of Clonal Heterogeneity

A parametric finite mixture model was used to assign mutations to individual clones. With this model, a clonal hierarchy could be established for 317 patients (79% of the cohort). The median number of clones per patient was 2 (range 1-5 clones). Some diseases were associated with more clones than others (P=0.000147), but there was no difference in the median number of clones among MPNs and MDS/MPNs (P=0.20) (Figures 32-33).

Figure 32 - Number of Clones and Disease Diagnosis

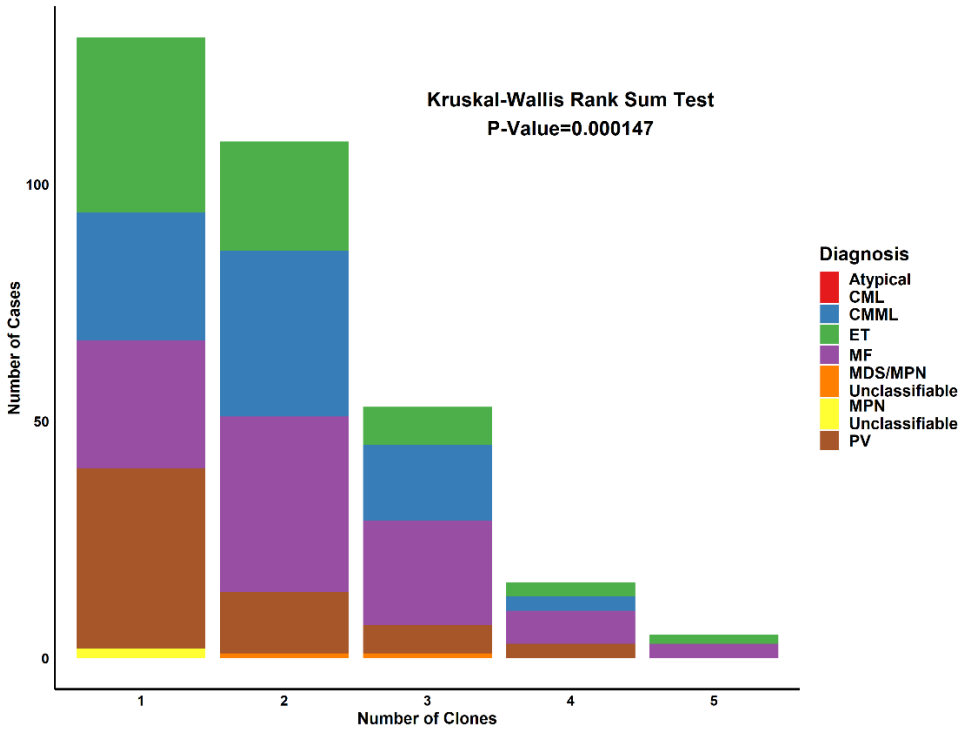
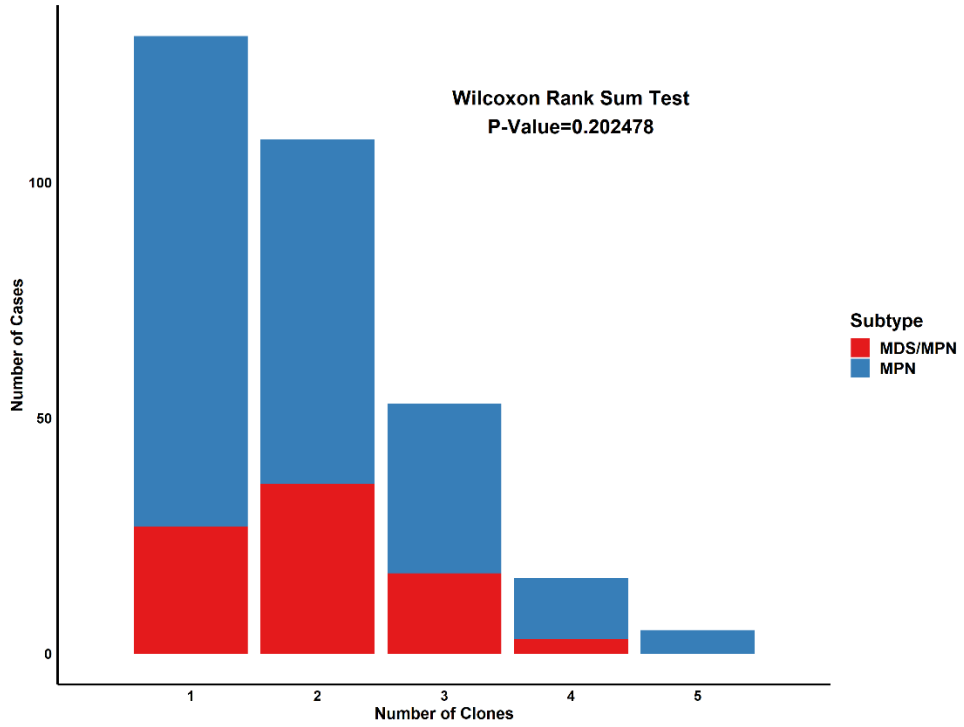


Figure 33 - Number of Clones and Disease Subtype



Considering a mutated gene as ‘clonal’ if it was found in the primary clone, and subclonal if it was found in any other clone, the clonal/subclonal profile of the 48 genes that are not located in sex chromosomes was analyzed. Results for the whole cohort are summarized in Figure 34, demonstrating that some genes more often present as clonal drivers when to other. To determine the statistical significance of this finding, the frequency of clonal drivers was modelled as a binomial process. Results are shown in Figures 35 and 36. In MPNs, genes *JAK2* and *CALR* are were the ones with a statistically significant higher probability of being clonal, while in the case of MDS/MPNs, the genes that were more likely to be found in the major clone were *SRSF2* and *TET2*.

Figure 34 - Clonality status of driver genes

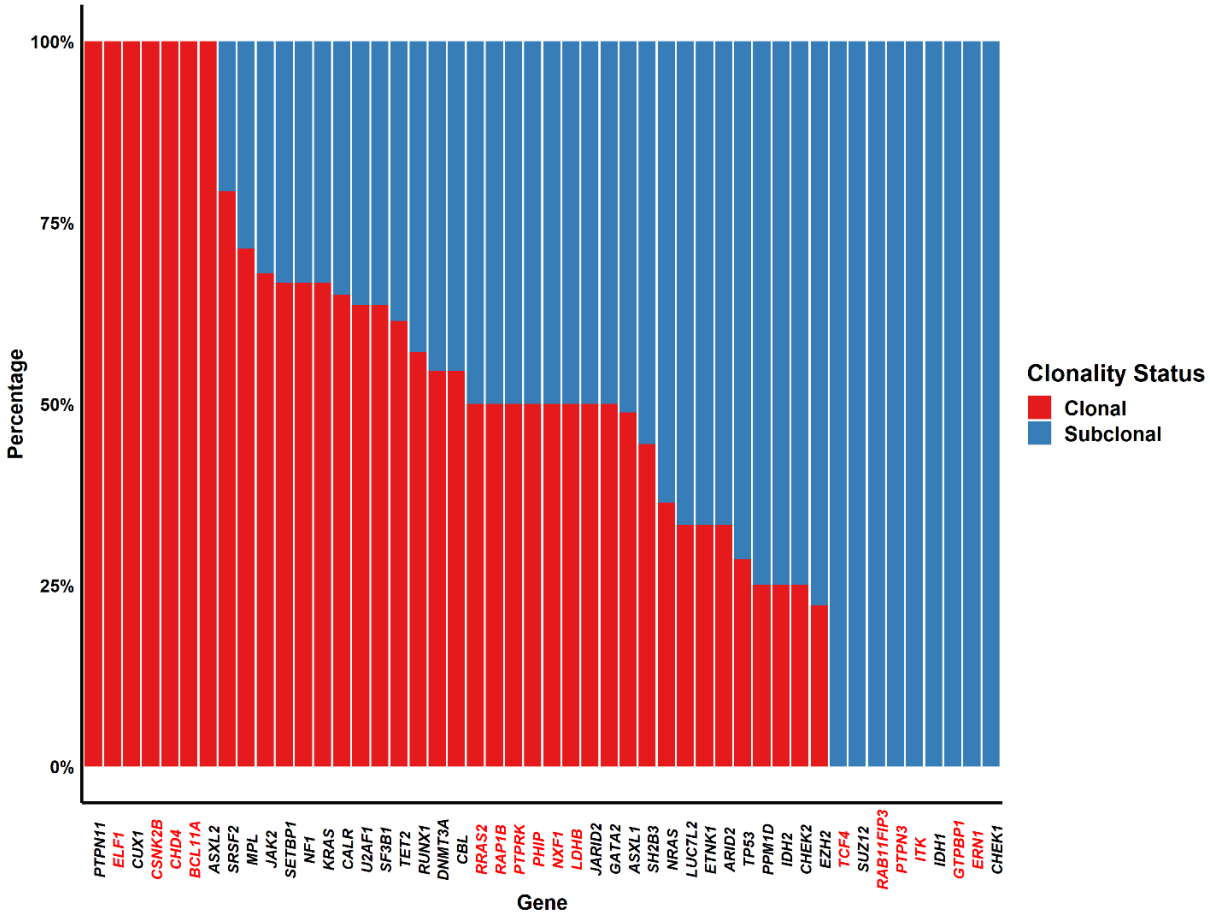


Figure 35 - Probability of Driver Gene Being Clonal and Significance by the Binomial Test-MPNs

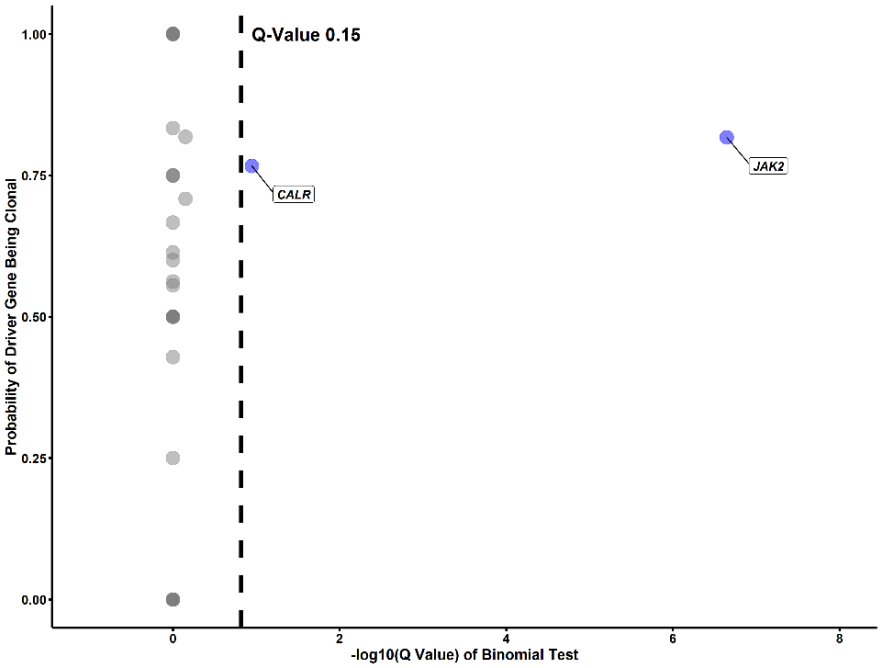
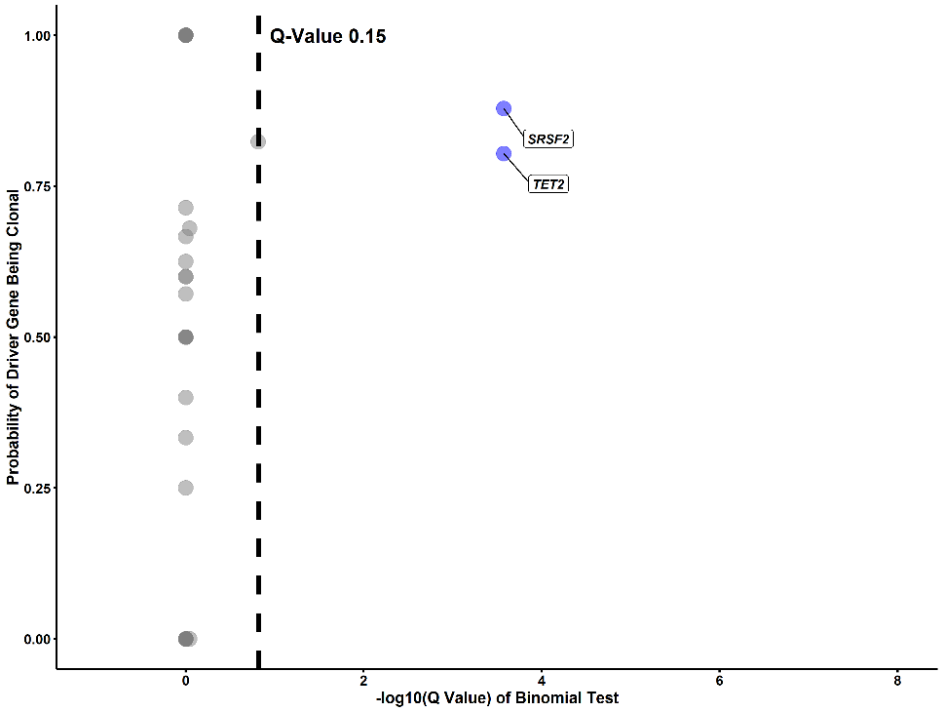


Figure 36 - Probability of Driver Gene Being Clonal and Significance by the Binomial Test-MDS/MPNs



4.6 Logistic regression model for disease classification based solely on genetic features

The previous results have demonstrated that different profiles of mutated genes, pairs/triads of genes, clonally dominant genes and altered biological pathways are associated with distinct phenotypes of myeloid neoplasms. However, it is not clear what are the most important genetic predictors of disease phenotype. Thus, a logistic regression was fitted for disease classification, considering the diagnosis of each disorder as the dependent variable, and putative independent variables as driver genes, gene combinations and known biological pathways. For model training and validation, the cohort was split into a training and validation set. The training set was used for model development, and variables were selected by LASSO. For model validation, each model was applied to the validation cohort and accuracy was calculated separately based on fitted probabilities for each model. Results on accuracy are shown in Figure 37. In 4 out of 5 models, the accuracy for the distinction of one type of disease was greater than 75%, and in the PV model it was 71.4%.

Figure 37 - Accuracy of logistic regression models for disease classification in Validation Cohort

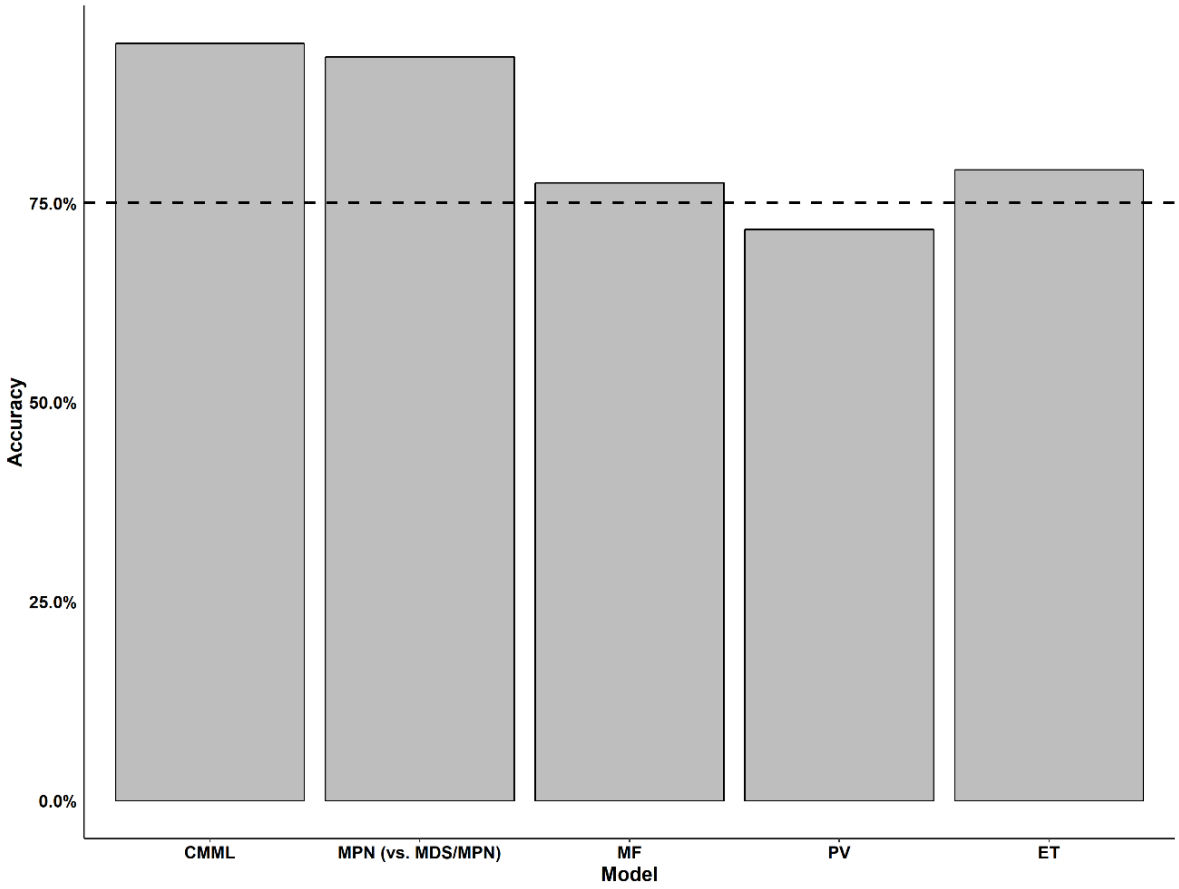


Table 2 shows the genetic covariates that predict for each disease diagnosis, with corresponding values of coefficients that were estimated by each individual model. For example, regarding CMML, the most relevant genetic abnormalities appear to be mutations in mRNA splicing genes (particularly *SRSF2*), mutations in genes that regulate DNA methylation and the combination thereof of these 2 groups of mutations. For ET the most important features were the absence of mutations in several biological pathways and a lower JAK2 allele burden.

Table 2 – Coefficients of Logistic Regression Model to Classify Patients Disease Phenotype

Disease	Covariates	Coefficient
MPN (vs. MDS/MPN)	*SRSF2* Mutation	-0.84
	JAK-STAT Pathway Mutation	2.67
	mRNA Splicing Pathway Mutation	-0.76
	RAS Pathway Mutation	-0.35
MDS/MPN (vs. MPN)	*SRSF2* Mutation	0.91
	JAK-STAT Pathway Mutation	-2.76
	mRNA Splicing Pathway Mutation	0.82
	RAS Pathway Mutation	0.43
CMML (vs. Others)	*SRSF2* Mutation	0.33
	Clonal *SRSF2* Mutation	0.78
	TET2 Mutation	0.36
	JAK-STAT Pathway Mutation	-2.23
	mRNA Splicing Pathway Mutation	0.94
	DNA Methylation and mRNA Splicing Pathway Mutation	0.10
	RAS Pathway + mRNA Splicing Pathway Mutation	0.38
ET (vs. Others)	*JAK2* Mutation Allele Burden	-0.01
	DNA Methylation Pathway Mutation	-0.15
	Histone Modification Pathway Mutation	-0.18
	mRNA Splicing Pathway Mutation	-1.07
	RAS Pathway Mutation	-0.69
	Hematopoietic Transcription Factors Mutation	-0.34

MF (vs. Others)	*CALR* Clonal Mutation	1.24
	JAK2 Mutation Allele Burden	0.009
	JAK-STAT Pathway + DNA Methylation Pathway Mutation	0.02
	JAK-STAT Pathway + mRNA Splicing Pathway Mutation	0.84
PV (vs. Others)	*JAK2* Mutation	1.99
	TET2 Mutation	-0.03
	mRNA Splicing Pathway Mutation	-0.03
	JAK-STAT Pathway + mRNA Splicing Pathway Mutation	-0.63

4.7 Survival Analysis

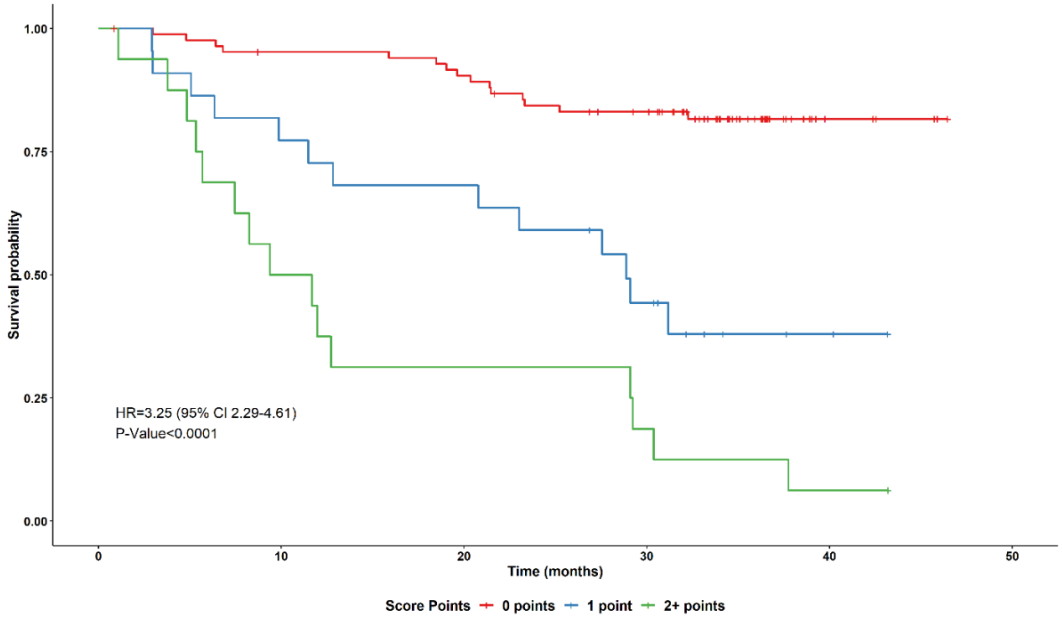
The OS outcomes of the 124 Brazilian patients were analyzed, comparing survival outcomes according to the presence of a specific genetic abnormality, either a mutated gene or pathway. In a bivariate model, after adjusting for the disease subtype (MPN or MDS/MPN), the following genetic abnormalities were found to be predictors of OS in the cohort: mutations in *ASXL1* (HR 2.08, 95% CI 1.08-3.98)*, NRAS* (HR 5.70, 95% CI 2.42-13.46)*, U2AF1* (HR 6.03, 95% CI 2.48-14.67) and mutations in genes associated with DNA methylation (HR 1.95, 95% CI 1.06-3.58), histone modification (HR 2.18, 95% CI 1.15-4.15), mRNA splicing (HR 3.54, 95% CI 1.81-6.93) and RAS pathway activation (HR 2.46, 95% CI 1.18-5.12). In a multivariate proportional hazards Cox model, the following covariates were independent predictors of survival after adjusting for disease subtype: *NRAS* mutations and mutations in genes related to mRNA splicing (Table 3).

Table 3 – Multivariate Cox Proportional Hazards Model

Covariate	Hazard Ratio (95% CI)	P-value
***NRAS* mutation**	6.56 (2.77-15.50)	1.80e-05
mRNA splicing mutations	3.72 (1.92-7.19)	9.19e-05
MDS/MPN (vs. MPN)	3.14 (1.54-6.38)	0.00152

Based on these results, a prognostic score was established, with 1 point being given for mRNA splicing mutations and MDS/MPN subtype, and 2 points being given for *NRAS* mutations. Patients were divided into three tiers, with zero points (N=85, 68.5% [95% CI 59-76%]), 1 point (N=22, 17.7% [95% CI 11-25%]) and 2+ points (N=16, 12.9% [95% CI 7-20%]). Median OS for each group was not reached, 28.9 months and 10.5 months, respectively, with a HR of 3.25 (95% CI 2.29-4.61) for a one step increase in category and two-tailed p-value <0.0001 (Figure 38).

Figure 38 - Overall Survival based on risk score categories



DISCUSSION

5. Discussion

In this article, WES data from a Brazilian cohort of patients with MDS/MPNs and MPNs and published papers analyzing the same subgroup of malignancies was combined for analysis and determination of driver genes that are mutated and responsible for disease pathogenesis, and to define which specific genes are responsible for a specific disease phenotype.

The first finding is that most patients with these neoplasms, either MPNs or MDS/MPNs have a low mutational burden when compared to solid tumors as comparison to data from TCGA reveals (LAWRENCE; STOJANOV; POLAK; KRYUKOV *et al.*, 2013). While this may reflect the more initial stage in oncogenesis of these disorders (when compared to AML), it also may reflect the high prevalence of mutations in epigenetic regulator genes, such as genes that influence DNA methylation and histone modifications. Mutations in epigenetic regulators can influence the function of several thousand genes at the same time and are frequently found in other neoplasms with low mutational burden, such as pediatric rhabdoid tumors. Rhabdoid tumors are neoplasms with a very low mutational burden and are characterized by biallelic loss of the *SMARCB1* gene, a member of the chromatin remodeling complex SWI/SNF (LEE; STEWART; CARTER; AMBROGIO *et al.*, 2012).

It is known that mutational burden in hematopoietic stem cells increases exponentially with aging (WELCH; LEY; LINK; MILLER *et al.*, 2012). Thus, it is to be expected that most mutations found in a malignant myeloid clone are unrelated to disease pathogenesis and could be considered “passenger” mutations. To detect “driver” mutations (mutations

associated with oncogenesis), one needs to detect mutations that have either a high mutational rate or are clustered around a mutational hotspot. To this end, the WES data was analyzed using computer algorithms (MutSigCV and IntOgen) that detect genes with a higher than expected mutation rate based on the background mutation rate and/or genes with mutational hotspots (GONZALEZ-PEREZ; PEREZ-LLAMAS; DEU-PONS; TAMBORERO *et al.*, 2013; LAWRENCE; STOJANOV; POLAK; KRYUKOV *et al.*, 2013). This is an unbiased and proven methodology for identification of putative oncogenes, and several other large-scale genome projects in cancer samples have employed similar approaches (KANDOTH; MCLELLAN; VANDIN; YE *et al.*, 2013; LAWRENCE; STOJANOV; MERMEL; ROBINSON *et al.*, 2014). The fact that a large fraction of the identified driver genes in the cohort (N=37, 68%) were already known drivers from the literature underscores the robustness of this approach. Using a large cohort of patients increased statistical power and allowed the identification of several novel putative driver genes (LAWRENCE; STOJANOV; MERMEL; ROBINSON *et al.*, 2014).

Identified driver genes were then analyzed using DAVID to determine recurrent biological pathways that are altered in these neoplasms. Most of the driver genes could be grouped into one or pathway based on DAVID results, and the final list of pathways that was used was based on a combination of results from DAVID and previously published articles. One key feature of most biological pathways that play a role in oncogenesis is the phenomenon of mutual exclusiveness, wherein several genes of the same pathway do not tend to be found mutated in the same patient, due to a redundancy in oncogenic mechanism and/or increased cellular toxicity (BABUR; GONEN; AKSOY; SCHULTZ *et al.*, 2015). There are computational algorithms that employ this approach in order to

determine novel biological pathways in cancer samples (PULIDO-TAMAYO; WEYTJENS; DE MAEYER; MARCHAL, 2016). While such approach was not used in the present manuscript, in the final list of pathways that was used there was a high rate of mutual exclusiveness among genes belonging to the same pathway, suggesting that this approach was valid for pathway determination.

The list of novel putative driver genes needs to be the subject of further experimental studies to determine their role in disease pathogenesis. As an example, in the study cohort there were several patients who harbored mutations in genes that are associated with TP53, a known tumor suppressor gene. Besides TP53, there were mutations in *CSNK2B, PPM1D, CHEK1 and CHEK2.* These genes were mutually exclusive from each other, suggesting, as already mentioned, a pathogenic role. *CHEK1* and *CHEK2* encode serine-threonine kinases Chk1 and Chk2 that are phosphorylated by ATM and ATR and initiate the DNA damage response (BARTEK; LUKAS, 2003; GATEI; SLOPER; SORENSEN; SYLJUASEN *et al.*, 2003; SORENSEN; SYLJUASEN; FALCK; SCHROEDER *et al.*, 2003). Mutations in *CHEK2* in MPN were first reported by NANGALIA et al (NANGALIA; MASSIE; BAXTER; NICE *et al.*, 2013), and data from studies conducted in solid tumors suggest that suppression of DNA damage response may enhance cell killing by genotoxic drugs (SMITH; THO; XU; GILLESPIE, 2010). Similarly, there were also loss-of-function mutations in the *CSNK2B* gene that encodes the kinase CK2. CK2 is part of an enzymatic complex responsible for phosphorylation of the TP53 protein at residue S392 in response to DNA damage from ultraviolet irradiation, and this leads to increased TP53 activity.(KELLER; ZENG; WANG; ZHANG *et al.*, 2001) It has been previously shown that loss of the C-terminal region of the β-subunit of CK2,

comprising amino acids residues 171-215, leads to downregulation of catalytic activity of the enzyme (SARNO; MARIN; BOSCHETTI; PAGANO *et al.*, 2000). It is possible that the *CSNK2B* gene mutations that were found decrease TP53 activation through this mechanism. Thus, further efforts in experimental studies in the lab are needed to confirm the oncogenic role of the novel drivers that were detected.

One objective of the study was to expand on the knowledge of what are the molecular basis underlying the clinical differences among MPNs and MDS/MPNs. The results showed that, as expected, patients with these disorders share a number of commonly mutated genes, similar to what has been reported previously, and no gene was found to be specific for a single disease or subtype (LUNDBERG; KAROW; NIENHOLD; LOOSER *et al.*, 2014; MASON; KHORASHAD; TANTRAVAHI; KELLEY *et al.*, 2016; NANGALIA; MASSIE; BAXTER; NICE *et al.*, 2013; PATEL; PRZYCHODZEN; THOTA; RADIVOYEVITCH *et al.*, 2017). Some genes and pathways were found to be more commonly mutated/altered in one disease type compared to the other. More specifically, the presence of JAK-STAT pathway mutations was much more common in patients with MPNs compared to patients with MDS/MPNs (RUMI; PIETRA; PASCUTTO; GUGLIELMELLI *et al.*, 2014). On the other side of the spectrum, mutations in mRNA-splicing genes, particularly when combined with either RAS-pathway mutations or mutations in genes related to DNA methylation was found more commonly in MDS/MPNs (MASON; KHORASHAD; TANTRAVAHI; KELLEY *et al.*, 2016). Similarly, combinations of specific genes, either in doublets or in triplets was found to be associated with specific subtypes of MPNs and MDS/MPNs. The absolute number of cases with these genes doublets or triplets was small in most cases, and the strength of the association between

these genes and a specific should be confirmed in other cohorts. These results suggest that there is no specific disease pathway that leads to the development of either MPN or MDS/MPN, and that rather the disease phenotype is a complex product of the interaction among different gene(s) or combination of genes, and the biological pathways that are altered after acquisition of these gene mutations.

One key finding of the study is that JAK-STAT pathway mutations are a strongly associated with having a diagnosis of a Ph-negative MPNs, with a very large measure of association (log(OR) near 5.0). For patients with MDS/MPN, there was no similarly large measure of association in the other biological pathways. Some patients with Ph-negative MPNs (MF, PV and ET) are known to be “triple-negative” and harbor no mutation in the triad of JAK-STAT activation genes (*JAK2, MPL* and *CALR*). In the recent study by GRINFELD et al, most patients who were negative for the canonical mutations in *JAK2, MPL* and *CALR* presented with noncanonical mutations in these genes, more often in the *MPL* gene (GRINFELD; NANGALIA; BAXTER; WEDGE *et al.*, 2018). It appears that most, if not all, patients with Ph-negative MPNs have gene mutations that activate JAK-STAT signaling. This begs the question of whether there exist some patients with Ph-negative MPNs who are truly “triple-negative” and do not have JAK-STAT activating mutations, or if mutations that activate this pathway are an essential condition for the appearance of Ph-negative MPNs. The results of the present study, when considered together with the results presented by GRINFELD et al, suggest the second answer. This is of greater relevance nowadays due to the existence of medications that inhibit JAK-STAT signaling and are marketed for patients with MF and PV, regardless of mutational status. The consideration that JAK-STAT activating mutations are essential for

development of Ph-negative would demand that the presence of one of these mutations is necessary for the use of JAK2 inhibitors.

Some mutations tend to be acquired earlier in disease pathogenesis, while others usually occur in subclones. Previously, ORTMANN et al have demonstrated that acquisition of *JAK2* mutations prior to *TET2* mutations is associated with distinct phenotypes, including a higher propensity to develop PV compared to ET, and a higher risk of thrombosis (ORTMANN; KENT; NANGALIA; SILBER *et al.*, 2015). In the present study, the order of mutations acquisition was inferred by clustering mutations based on VAF using a Gaussian mixture model and determining the prevalence of clonal or subclonal mutations among each disease subtype. Clonal *JAK2* and *CALR* mutations were more common in patients with MPNs, while clonal *SRSF2* and *TET2* mutations were more common in MDS/MPN, similar to what has been reported by Patel et al. in patients with CMML (PATEL; PRZYCHODZEN; THOTA; RADIVOYEVITCH *et al.*, 2017). The mutation that is present in the major clone appears to be an important variable of disease, phenotype, as in the logistic regression classification model, clonal *SRSF2* mutations were an independent factor associated with CMML and clonal *CALR* mutations were associated with MF. These results do not imply that other genes cannot be found in major clones in these disorders, but merely that most cases of a specific disorder tend to harbor clonally dominant mutations in such genes. These results may provide basis for other studies employing a similar approach to the one used by ORTMANN et al. to determine the biological impact of acquiring these mutations in different orders.

Regression models and machine learning algorithms when applied to genomics data can predict tumor cell type and classification based on tumor-specific signatures (AMAR;

IZRAELI; SHAMIR, 2017; CIRIELLO; MILLER; AKSOY; SENBABAOGLU *et al.*, 2013; HAO; LUO; KRAWCZYK; WEI *et al.*, 2017; KANG; LI; CHEN; ZHOU *et al.*, 2017; KHAN; WEI; RINGNER; SAAL *et al.*, 2001; RAMASWAMY; TAMAYO; RIFKIN; MUKHERJEE *et al.*, 2001; SOH; SZCZUREK; SAKOPARNIG; BEERENWINKEL, 2017; TIBSHIRANI; HASTIE; NARASIMHAN; CHU, 2002). Initial studies focused on analysis of gene expression data (KHAN; WEI; RINGNER; SAAL *et al.*, 2001; RAMASWAMY; TAMAYO; RIFKIN; MUKHERJEE *et al.*, 2001; TIBSHIRANI; HASTIE; NARASIMHAN; CHU, 2002), with more recent papers employing epigenetic signatures and/or integrating several distinct genomic datasets (CIRIELLO; MILLER; AKSOY; SENBABAOGLU *et al.*, 2013; HAO; LUO; KRAWCZYK; WEI *et al.*, 2017; KANG; LI; CHEN; ZHOU *et al.*, 2017). Previous publications have shown that somatic mutation data alone is sufficient for cancer type classification (AMAR; IZRAELI; SHAMIR, 2017; SOH; SZCZUREK; SAKOPARNIG; BEERENWINKEL, 2017). The potential of this approach in MPNs and MDS/MPNs was demonstrated by employing a logistic model for disease classification utilizing solely molecular point mutations that could accurately classify 75-90% of cases of these neoplasms. While most cases of MPN do not present as a major diagnostic challenge, the diagnosis of MPN / MDS is not always clear and the model may improve the diagnostic accuracy of such neoplasms. It is possible that inclusion of additional genomic information (e.g. gene expression, DNA methylation) and/or clinical data, and use of machine learning algorithms could improve the efficacy of such models. This could represent an important step forward in accurately diagnosing patients with these disorders. In a recent publication, a genomic classification with prognostic relevance has been developed for patients with MPN (GRINFELD; NANGALIA; BAXTER; WEDGE *et al.*, 2018). While MPN

was the focus of that work, herein the focus was on the pattern of genomic differences between MPN and MPN / MDS. In addition, most patients of that study were evaluated with a 69 gene panel, while all patients in the present study had an exome sequencing performed, what empowers our study for the identification of putative novel rare gene drivers.

Somatic gene mutations also play a role in survival prognostication, and the various driver mutations and biological pathways that were analyzed in the cohort were used to model survival outcomes. For this analysis, only Brazilian patients were used, since the data on survival was not available for the other datasets. This imposes a limitation on the findings, since the number of evaluated patients and events is much smaller (N=124 and N=43), which diminishes the statistical power of the analysis and limits the number of variables that can be used in the Cox model due to the risk of model overfitting. Nonetheless, after adjusting for multiple testing and for disease subtype (either MPNs or MDS/MPNs), two genetic abnormalities were found to be independent predictors of survival: *NRAS* mutations and mutations in genes associated with mRNA splicing. In the large study by GRINFELD et al, similar findings were reported, where *NRAS* mutations were associated with worse survival in patients with PV and ET (GRINFELD; NANGALIA; BAXTER; WEDGE *et al.*, 2018). These results that novel genomics based prognostic models for patients with myeloid neoplasms should consider the role of *RAS* mutations in the outcome of these disorders.

CONCLUSION

6. Conclusion

In conclusion, in this study a large dataset of WES sequencing data from patients with MPNs and MDS/MPNs was analyzed. An unbiased analysis to discover driver genes found 54 genes with a statistically significant rate of mutations greater than expected based on gene size and background mutational rate. Among these 54 driver genes, 38 were genes that were previously reported, and 17 are novel driver genes that need to be further studied in the laboratory to confirm their pathogenicity in MPNs and MDS/MPNs. Nineteen of these 54 genes had a different prevalence among the diverse disease categories, suggesting that these may contribute to disease phenotype.

Annotating these genes for their function using the DAVID platform and the published literature, 75% of them could be grouped into one of 7 biological pathways that play a role in the pathogenesis of MPNs and MDS/MPNs. Several of the novel driver genes were also grouped into one of these pathways and were present in a mutually exclusive manner with the other genes in the pathway, suggesting that they are relevant for oncogenesis. Furthermore, the pattern of biological pathways that was found to be mutated was associated with disease phenotype, with JAK-STAT gene mutations being associated with Ph-negative MPNs and mutations in RAS pathway, Mrna *splicing* and DNA methylation genes being associated with MDS/MPNs.

Genes were found to co-occur and be mutually exclusive with each other more often than expected in several pairs of genes, and the presence of co-occurring gene pairs/triads was also associated with a specific disease phenotype. Similarly, the pattern of gene clonality (i.e. which gene was found to be mutated in the dominant clone) was also

predictive of disease phenotype, with clonal *JAK2* and *CALR* being associated with MPNs, and clonal *TET2* and *SRSF2* being associated with MDS/MPNs.

To determine which were the main genetic factors determining disease phenotype, a logistic regression model was fit and could determine diagnosis and classification with high accuracy, considering only data from genetic sequencing studies. Finally, some of these genomic changes were found to be predictive of decreased survival.

The results of the present study support the idea that underlying the many distinct diseases that are currently diagnosed and classified as chronic myeloid neoplasms there is a select group of recurrently mutated genes. Disease phenotype appears to be a consequence of the combinations of mutations in specific genes, their order of appearance and which biological pathway they belong to. Statistical models for disease classification and prognostication based solely on gene mutations will improve with the inclusion of additional data generated from large cohort sequencing studies. Future efforts should be directed at the integration of large amounts of genetic data with clinical features to develop machine learning algorithms for disease classification and prognostication, and this study represents a small step in this endeavor.

ACKNOWLEDGEMENTS

7. Acknowledgements

This research was supported by a Federal research grant from the Brazilian Ministry of Health to F.P.S.S. (PROADI-SUS SIPAR no. 25000179520/2011-36) and a grant from AMIGOH to F.P.S.S.

REFERENCES

8. References

ALY, M.; RAMDZAN, Z. M.; NAGATA, Y.; BALASUBRAMANIAN, S. K. *et al.* Distinct clinical and biological implications of CUX1 in myeloid neoplasms. **Blood Adv**, 3, n. 14, p. 2164-2178, Jul 23 2019.

AMAR, D.; IZRAELI, S.; SHAMIR, R. Utilizing somatic mutation data from numerous studies for cancer research: proof of concept and applications. **Oncogene**, 36, n. 24, p. 3375-3383, Jun 15 2017.

ARBER, D. A.; ORAZI, A.; HASSERJIAN, R.; THIELE, J. *et al.* The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. **Blood**, 127, n. 20, p. 2391-2405, May 19 2016.

BABUR, O.; GONEN, M.; AKSOY, B. A.; SCHULTZ, N. *et al.* Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations. **Genome Biol**, 16, p. 45, Feb 26 2015.

BALL, M.; LIST, A. F.; PADRON, E. When clinical heterogeneity exceeds genetic heterogeneity: thinking outside the genomic box in chronic myelomonocytic leukemia. **Blood**, 128, n. 20, p. 2381-2387, Nov 17 2016.

BARTEK, J.; LUKAS, J. Chk1 and Chk2 kinases in checkpoint control and cancer. **Cancer Cell**, 3, n. 5, p. 421-429, May 2003.

BAXTER, E. J.; SCOTT, L. M.; CAMPBELL, P. J.; EAST, C. *et al.* Acquired mutation of the tyrosine kinase JAK2 in human myeloproliferative disorders. **Lancet**, 365, n. 9464, p. 1054-1061, Mar 19-25 2005.

BENJAMINI, Y.; HOCHBERG, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. **Journal of the Royal Statistical Society. Series B (Methodological)**, 57, n. 1, p. 289-300, 1995.

BERGER, M. F.; LAWRENCE, M. S.; DEMICHELIS, F.; DRIER, Y. *et al.* The genomic complexity of primary human prostate cancer. **Nature**, 470, n. 7333, p. 214-220, Feb 10 2011.

CABAGNOLS, X.; FAVALE, F.; PASQUIER, F.; MESSAOUDI, K. *et al.* Presence of atypical thrombopoietin receptor (MPL) mutations in triple-negative essential thrombocythemia patients. **Blood**, 127, n. 3, p. 333-342, Jan 21 2016.

CANCER GENOME ATLAS RESEARCH, N.; ALBERT EINSTEIN COLLEGE OF, M.; ANALYTICAL BIOLOGICAL, S.; BARRETOS CANCER, H. *et al.* Integrated genomic and molecular characterization of cervical cancer. **Nature**, 543, n. 7645, p. 378-384, Mar 16 2017.

CANCER GENOME ATLAS RESEARCH, N.; LEY, T. J.; MILLER, C.; DING, L. *et al.* Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. **N Engl J Med**, 368, n. 22, p. 2059-2074, May 30 2013.

CIBULSKIS, K.; LAWRENCE, M. S.; CARTER, S. L.; SIVACHENKO, A. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. **Nat Biotechnol**, 31, n. 3, p. 213-219, Mar 2013.

CIRIELLO, G.; MILLER, M. L.; AKSOY, B. A.; SENBABAOGLU, Y. *et al.* Emerging landscape of oncogenic signatures across human cancers. **Nat Genet**, 45, n. 10, p. 1127-1133, Oct 2013.

COOLS, J.; DEANGELO, D. J.; GOTLIB, J.; STOVER, E. H. *et al.* A tyrosine kinase created by fusion of the PDGFRA and FIP1L1 genes as a therapeutic target of imatinib in idiopathic hypereosinophilic syndrome. **N Engl J Med**, 348, n. 13, p. 1201-1214, Mar 27 2003.

COX, D. R. Regression Models and Life-Tables. **Journal of the Royal Statistical Society. Series B (Methodological)**, 34, n. 2, p. 187-220, 1972.

DALEY, G. Q.; VAN ETTEN, R. A.; BALTIMORE, D. Induction of chronic myelogenous leukemia in mice by the P210bcr/abl gene of the Philadelphia chromosome. **Science**, 247, n. 4944, p. 824-830, Feb 16 1990.

DAMESHEK, W. Some speculations on the myeloproliferative syndromes. **Blood**, 6, n. 4, p. 372-375, Apr 1951.

DELHOMMEAU, F.; DUPONT, S.; DELLA VALLE, V.; JAMES, C. *et al.* Mutation in TET2 in myeloid cancers. **N Engl J Med**, 360, n. 22, p. 2289-2301, May 28 2009.

DEPRISTO, M. A.; BANKS, E.; POPLIN, R.; GARIMELLA, K. V. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. **Nat Genet**, 43, n. 5, p. 491-498, May 2011.

DUDGEON, C.; SHREERAM, S.; TANOUE, K.; MAZUR, S. J. *et al.* Genetic variants and mutations of PPM1D control the response to DNA damage. **Cell Cycle**, 12, n. 16, p. 2656-2664, Aug 15 2013.

ELF, S.; ABDELFATTAH, N. S.; CHEN, E.; PERALES-PATON, J. *et al.* Mutant Calreticulin Requires Both Its Mutant C-terminus and the Thrombopoietin Receptor for Oncogenic Transformation. **Cancer Discov**, 6, n. 4, p. 368-381, Apr 2016.

Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA (URL: http://evs.gs.washington.edu/EVS/) [April, 2017].

FISHER, J. B.; MCNULTY, M.; BURKE, M. J.; CRISPINO, J. D. *et al.* Cohesin Mutations in Myeloid Malignancies. **Trends Cancer**, 3, n. 4, p. 282-293, Apr 2017.

FRASER, M.; SABELNYKOVA, V. Y.; YAMAGUCHI, T. N.; HEISLER, L. E. *et al.* Genomic hallmarks of localized, non-indolent prostate cancer. **Nature**, 541, n. 7637, p. 359-364, Jan 19 2017.

GAMBACORTI-PASSERINI, C. B.; DONADONI, C.; PARMIANI, A.; PIROLA, A. *et al.* Recurrent ETNK1 mutations in atypical chronic myeloid leukemia. **Blood**, 125, n. 3, p. 499-503, Jan 15 2015.

GATEI, M.; SLOPER, K.; SORENSEN, C.; SYLJUASEN, R. *et al.* Ataxia-telangiectasia-mutated (ATM) and NBS1-dependent phosphorylation of Chk1 on Ser-317 in response to ionizing radiation. **J Biol Chem**, 278, n. 17, p. 14806-14811, Apr 25 2003.

GENOMES PROJECT, C.; AUTON, A.; BROOKS, L. D.; DURBIN, R. M. *et al.* A global reference for human genetic variation. **Nature**, 526, n. 7571, p. 68-74, Oct 1 2015.

GEORGE, J.; LIM, J. S.; JANG, S. J.; CUN, Y. *et al.* Comprehensive genomic profiles of small cell lung cancer. **Nature**, 524, n. 7563, p. 47-53, Aug 6 2015.

GONZALEZ-PEREZ, A.; PEREZ-LLAMAS, C.; DEU-PONS, J.; TAMBORERO, D. *et al.* IntOGen-mutations identifies cancer drivers across tumor types. **Nat Methods**, 10, n. 11, p. 1081-1082, Nov 2013.

GRINFELD, J.; NANGALIA, J.; BAXTER, E. J.; WEDGE, D. C. *et al.* Classification and Personalized Prognosis in Myeloproliferative Neoplasms. **N Engl J Med**, 379, n. 15, p. 1416-1430, Oct 11 2018.

HAO, X.; LUO, H.; KRAWCZYK, M.; WEI, W. *et al.* DNA methylation markers for diagnosis and prognosis of common cancers. **Proc Natl Acad Sci U S A**, 114, n. 28, p. 7414-7419, Jul 11 2017.

HUNTLY, B. J.; SHIGEMATSU, H.; DEGUCHI, K.; LEE, B. H. *et al.* MOZ-TIF2, but not BCR-ABL, confers properties of leukemic stem cells to committed murine hematopoietic progenitors. **Cancer Cell**, 6, n. 6, p. 587-596, Dec 2004.

INSTITUTE, B. **Picard Tools**. Versão v.0.2.5b8. http://broadinstitute.github.io/picard/.

INTERNATIONAL CANCER GENOME, C.; HUDSON, T. J.; ANDERSON, W.; ARTEZ, A. *et al.* International network of cancer genome projects. **Nature**, 464, n. 7291, p. 993-998, Apr 15 2010.

JAMES, C.; UGO, V.; LE COUEDIC, J. P.; STAERK, J. *et al.* A unique clonal JAK2 mutation leading to constitutive signalling causes polycythaemia vera. **Nature**, 434, n. 7037, p. 1144-1148, Apr 28 2005.

JARA, A.; HANSON, T. E.; QUINTANA, F. A.; MULLER, P. *et al.* DPpackage: Bayesian Non- and Semi-parametric Modelling in R. **J Stat Softw**, 40, n. 5, p. 1-30, Apr 1 2011.

JELINEK, J.; OKI, Y.; GHARIBYAN, V.; BUESO-RAMOS, C. *et al.* JAK2 mutation 1849G>T is rare in acute leukemias but can be found in CMML, Philadelphia chromosome-negative CML, and megakaryocytic leukemia. **Blood**, 106, n. 10, p. 3370-3373, Nov 15 2005.

KAHN, J. D.; MILLER, P. G.; SILVER, A. J.; SELLAR, R. S. *et al.* PPM1D-truncating mutations confer resistance to chemotherapy and sensitivity to PPM1D inhibition in hematopoietic cells. **Blood**, 132, n. 11, p. 1095-1105, Sep 13 2018.

KANDOTH, C.; MCLELLAN, M. D.; VANDIN, F.; YE, K. *et al.* Mutational landscape and significance across 12 major cancer types. **Nature**, 502, n. 7471, p. 333-339, Oct 17 2013.

KANG, S.; LI, Q.; CHEN, Q.; ZHOU, Y. *et al.* CancerLocator: non-invasive cancer diagnosis and tissue-of-origin prediction using methylation profiles of cell-free DNA. **Genome Biol**, 18, n. 1, p. 53, Mar 24 2017.

KAPLAN, E. L.; MEIER, P. Nonparametric Estimation from Incomplete Observations. **Journal of the American Statistical Association**, 53, n. 282, p. 457-481, 1958.

KELLER, D. M.; ZENG, X.; WANG, Y.; ZHANG, Q. H. *et al.* A DNA damage-induced p53 serine 392 kinase complex contains CK2, hSpt16, and SSRP1. **Mol Cell**, 7, n. 2, p. 283-292, Feb 2001.

KHAN, J.; WEI, J. S.; RINGNER, M.; SAAL, L. H. *et al.* Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. **Nat Med**, 7, n. 6, p. 673-679, Jun 2001.

KLAMPFL, T.; GISSLINGER, H.; HARUTYUNYAN, A. S.; NIVARTHI, H. *et al.* Somatic mutations of calreticulin in myeloproliferative neoplasms. **N Engl J Med**, 369, n. 25, p. 2379-2390, Dec 19 2013.

KLEIBLOVA, P.; SHALTIEL, I. A.; BENADA, J.; SEVCIK, J. *et al.* Gain-of-function mutations of PPM1D/Wip1 impair the p53-dependent G1 checkpoint. **J Cell Biol**, 201, n. 4, p. 511-521, May 13 2013.

KOHLMANN, A.; GROSSMANN, V.; KLEIN, H. U.; SCHINDELA, S. *et al.* Next-generation sequencing technology reveals a characteristic pattern of molecular mutations in 72.8% of chronic myelomonocytic leukemia by detecting frequent alterations in TET2, CBL, RAS, and RUNX1. **J Clin Oncol**, 28, n. 24, p. 3858-3865, Aug 20 2010.

KOREN, A.; POLAK, P.; NEMESH, J.; MICHAELSON, J. J. *et al.* Differential relationship of DNA replication timing to different forms of human mutation and variation. **Am J Hum Genet**, 91, n. 6, p. 1033-1040, Dec 7 2012.

KOSCHMIEDER, S.; GOTTGENS, B.; ZHANG, P.; IWASAKI-ARAI, J. *et al.* Inducible chronic phase of myeloid leukemia with expansion of hematopoietic stem cells in a transgenic model of BCR-ABL leukemogenesis. **Blood**, 105, n. 1, p. 324-334, Jan 1 2005.

KOSMIDER, O.; GELSI-BOYER, V.; CIUDAD, M.; RACOEUR, C. *et al.* TET2 gene mutation is a frequent and adverse event in chronic myelomonocytic leukemia. **Haematologica**, 94, n. 12, p. 1676-1681, Dec 2009.

KRALOVICS, R.; PASSAMONTI, F.; BUSER, A. S.; TEO, S. S. *et al.* A gain-of-function mutation of JAK2 in myeloproliferative disorders. **N Engl J Med**, 352, n. 17, p. 1779-1790, Apr 28 2005.

LARSON, D. E.; HARRIS, C. C.; CHEN, K.; KOBOLDT, D. C. *et al.* SomaticSniper: identification of somatic point mutations in whole genome sequencing data. **Bioinformatics**, 28, n. 3, p. 311-317, Feb 1 2012.

LAWRENCE, M. S.; STOJANOV, P.; MERMEL, C. H.; ROBINSON, J. T. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. **Nature**, 505, n. 7484, p. 495-501, Jan 23 2014.

LAWRENCE, M. S.; STOJANOV, P.; POLAK, P.; KRYUKOV, G. V. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. **Nature**, 499, n. 7457, p. 214-218, Jul 11 2013.

LE GALLO, M.; O'HARA, A. J.; RUDD, M. L.; URICK, M. E. *et al.* Exome sequencing of serous endometrial tumors identifies recurrent somatic mutations in chromatin-remodeling and ubiquitin ligase complex genes. **Nat Genet**, 44, n. 12, p. 1310-1315, Dec 2012.

LEE, R. S.; STEWART, C.; CARTER, S. L.; AMBROGIO, L. *et al.* A remarkably simple genome underlies highly malignant pediatric rhabdoid cancers. **J Clin Invest**, 122, n. 8, p. 2983-2988, Aug 2012.

LEISERSON, M. D.; BLOKH, D.; SHARAN, R.; RAPHAEL, B. J. Simultaneous identification of multiple driver pathways in cancer. **PLoS Comput Biol**, 9, n. 5, p. e1003054, 2013.

LEVINE, R. L.; LORIAUX, M.; HUNTLY, B. J.; LOH, M. L. *et al.* The JAK2V617F activating mutation occurs in chronic myelomonocytic leukemia and acute myeloid leukemia, but not in acute lymphoblastic leukemia or chronic lymphocytic leukemia. **Blood**, 106, n. 10, p. 3377-3379, Nov 15 2005.

LEVINE, R. L.; WADLEIGH, M.; COOLS, J.; EBERT, B. L. *et al.* Activating mutation in the tyrosine kinase JAK2 in polycythemia vera, essential thrombocythemia, and myeloid metaplasia with myelofibrosis. **Cancer Cell**, 7, n. 4, p. 387-397, Apr 2005.

LI, H.; DURBIN, R. Fast and accurate short read alignment with Burrows-Wheeler transform. **Bioinformatics**, 25, n. 14, p. 1754-1760, Jul 15 2009.

LUNDBERG, P.; KAROW, A.; NIENHOLD, R.; LOOSER, R. *et al.* Clonal evolution and clinical correlates of somatic mutations in myeloproliferative neoplasms. **Blood**, 123, n. 14, p. 2220-2228, Apr 3 2014.

MAKISHIMA, H.; VISCONTE, V.; SAKAGUCHI, H.; JANKOWSKA, A. M. *et al.* Mutations in the spliceosome machinery, a novel and ubiquitous pathway in leukemogenesis. **Blood**, 119, n. 14, p. 3203-3210, Apr 5 2012.

MAKISHIMA, H.; YOSHIZATO, T.; YOSHIDA, K.; SEKERES, M. A. *et al.* Dynamics of clonal evolution in myelodysplastic syndromes. **Nat Genet**, 49, n. 2, p. 204-212, Feb 2017.

MALCOVATI, L.; PAPAEMMANUIL, E.; AMBAGLIO, I.; ELENA, C. *et al.* Driver somatic mutations identify distinct disease entities within myeloid neoplasms with myelodysplasia. **Blood**, 124, n. 9, p. 1513-1521, Aug 28 2014.

MASON, C. C.; KHORASHAD, J. S.; TANTRAVAHI, S. K.; KELLEY, T. W. *et al.* Age-related mutations and chronic myelomonocytic leukemia. **Leukemia**, 30, n. 4, p. 906-913, Apr 2016.

MAYAKONDA, A.; LIN, D. C.; ASSENOV, Y.; PLASS, C. *et al.* Maftools: efficient and comprehensive analysis of somatic variants in cancer. **Genome Res**, 28, n. 11, p. 1747-1756, Nov 2018.

MERLEVEDE, J.; DROIN, N.; QIN, T.; MELDI, K. *et al.* Mutation allele burden remains unchanged in chronic myelomonocytic leukaemia responding to hypomethylating agents. **Nat Commun**, 7, p. 10767, Feb 24 2016.

MILLER, C. A.; WHITE, B. S.; DEES, N. D.; GRIFFITH, M. *et al.* SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. **PLoS Comput Biol**, 10, n. 8, p. e1003665, Aug 2014.

MINA, M.; RAYNAUD, F.; TAVERNARI, D.; BATTISTELLO, E. *et al.* Conditional Selection of Genomic Alterations Dictates Cancer Evolution and Oncogenic Dependencies. **Cancer Cell**, 32, n. 2, p. 155-168 e156, Aug 14 2017.

MORGAN, X. C.; NI, S.; MIRANKER, D. P.; IYER, V. R. Predicting combinatorial binding of transcription factors to regulatory elements in the human genome by association rule mining. **BMC Bioinformatics**, 8, p. 445, Nov 15 2007.

NANGALIA, J.; MASSIE, C. E.; BAXTER, E. J.; NICE, F. L. *et al.* Somatic CALR mutations in myeloproliferative neoplasms with nonmutated JAK2. **N Engl J Med**, 369, n. 25, p. 2391-2405, Dec 19 2013.

NOWELL, P. C.; HUNGERFORD, D. A. A Minute Chromosome in Human Chronic Granulocytic Leukemia. **Science**, 132, n. 3438, p. 1497, Nov 18 1960.

ORTMANN, C. A.; KENT, D. G.; NANGALIA, J.; SILBER, Y. *et al.* Effect of mutation order on myeloproliferative neoplasms. **N Engl J Med**, 372, n. 7, p. 601-612, Feb 12 2015.

PAPAEMMANUIL, E.; GERSTUNG, M.; BULLINGER, L.; GAIDZIK, V. I. *et al.* Genomic Classification and Prognosis in Acute Myeloid Leukemia. **N Engl J Med**, 374, n. 23, p. 2209-2221, Jun 9 2016.

PAPAEMMANUIL, E.; GERSTUNG, M.; MALCOVATI, L.; TAURO, S. *et al.* Clinical and biological implications of driver mutations in myelodysplastic syndromes. **Blood**, 122, n. 22, p. 3616-3627; quiz 3699, Nov 21 2013.

PATEL, B. J.; PRZYCHODZEN, B.; THOTA, S.; RADIVOYEVITCH, T. *et al.* Genomic determinants of chronic myelomonocytic leukemia. **Leukemia**, 31, n. 12, p. 2815-2823, Dec 2017.

PIAZZA, R.; VALLETTA, S.; WINKELMANN, N.; REDAELLI, S. *et al.* Recurrent SETBP1 mutations in atypical chronic myeloid leukemia. **Nat Genet**, 45, n. 1, p. 18-24, Jan 2013.

PLEASANCE, E. D.; CHEETHAM, R. K.; STEPHENS, P. J.; MCBRIDE, D. J. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. **Nature**, 463, n. 7278, p. 191-196, Jan 14 2010.

PULIDO-TAMAYO, S.; WEYTJENS, B.; DE MAEYER, D.; MARCHAL, K. SSA-ME Detection of cancer driver genes using mutual exclusivity by small subnetwork analysis. **Sci Rep**, 6, p. 36257, Nov 3 2016.

RAMASWAMY, S.; TAMAYO, P.; RIFKIN, R.; MUKHERJEE, S. *et al.* Multiclass cancer diagnosis using tumor gene expression signatures. **Proc Natl Acad Sci U S A**, 98, n. 26, p. 15149-15154, Dec 18 2001.

ROWLEY, J. D. Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. **Nature**, 243, n. 5405, p. 290-293, Jun 1 1973.

RUMI, E.; CAZZOLA, M. Diagnosis, risk stratification, and response evaluation in classical myeloproliferative neoplasms. **Blood**, 129, n. 6, p. 680-692, Feb 9 2017.

RUMI, E.; PIETRA, D.; PASCUTTO, C.; GUGLIELMELLI, P. *et al.* Clinical effect of driver mutations of JAK2, CALR, or MPL in primary myelofibrosis. **Blood**, 124, n. 7, p. 1062-1069, Aug 14 2014.

SAKAI, H.; HOSONO, N.; PRZYCHODZEN, B. P.; CARRAWAY, H. E. *et al.* Multiple Mechanisms Leading to ARID2 defects in Myeloid Neoplasms. **Blood**, 124, p. 4610-4610, 2014-12-06 00:00:00 2014.

SARNO, S.; MARIN, O.; BOSCHETTI, M.; PAGANO, M. A. *et al.* Cooperative modulation of protein kinase CK2 by separate domains of its regulatory beta-subunit. **Biochemistry**, 39, n. 40, p. 12324-12329, Oct 10 2000.

SCHEMIONEK, M.; ELLING, C.; STEIDL, U.; BAUMER, N. *et al.* BCR-ABL enhances differentiation of long-term repopulating hematopoietic stem cells. **Blood**, 115, n. 16, p. 3185-3195, Apr 22 2010.

SHERRY, S. T.; WARD, M. H.; KHOLODOV, M.; BAKER, J. *et al.* dbSNP: the NCBI database of genetic variation. **Nucleic Acids Res**, 29, n. 1, p. 308-311, Jan 1 2001.

SMITH, J.; THO, L. M.; XU, N.; GILLESPIE, D. A. The ATM-Chk2 and ATR-Chk1 pathways in DNA damage signaling and cancer. **Adv Cancer Res**, 108, p. 73-112, 2010.

SOH, K. P.; SZCZUREK, E.; SAKOPARNIG, T.; BEERENWINKEL, N. Predicting cancer type from tumour DNA signatures. **Genome Med**, 9, n. 1, p. 104, Nov 28 2017.

SONG, Y.; LI, L.; OU, Y.; GAO, Z. *et al.* Identification of genomic alterations in oesophageal squamous cell cancer. **Nature**, 509, n. 7498, p. 91-95, May 1 2014.

SORENSEN, C. S.; SYLJUASEN, R. G.; FALCK, J.; SCHROEDER, T. *et al.* Chk1 regulates the S phase checkpoint by coupling the physiological turnover and ionizing radiation-induced accelerated proteolysis of Cdc25A. **Cancer Cell**, 3, n. 3, p. 247-258, Mar 2003.

STEENSMA, D. P.; BENNETT, J. M. The myelodysplastic syndromes: diagnosis and treatment. **Mayo Clin Proc**, 81, n. 1, p. 104-130, Jan 2006.

TEFFERI, A.; PARDANANI, A.; LIM, K. H.; ABDEL-WAHAB, O. *et al.* TET2 mutations and their clinical correlates in polycythemia vera, essential thrombocythemia and myelofibrosis. **Leukemia**, 23, n. 5, p. 905-911, May 2009.

TIBSHIRANI, R. Regression Shrinkage and Selection via the Lasso. **Journal of the Royal Statistical Society. Series B (Methodological)**, 58, n. 1, p. 267-288, 1996.

TIBSHIRANI, R.; HASTIE, T.; NARASIMHAN, B.; CHU, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. **Proc Natl Acad Sci U S A**, 99, n. 10, p. 6567-6572, May 14 2002.

VAINCHENKER, W.; KRALOVICS, R. Genetic basis and molecular pathophysiology of classical myeloproliferative neoplasms. **Blood**, 129, n. 6, p. 667-679, Feb 9 2017.

VAN DEN BERGHE, H.; CASSIMAN, J. J.; DAVID, G.; FRYNS, J. P. *et al.* Distinct haematological disorder with deletion of long arm of no. 5 chromosome. **Nature**, 251, n. 5474, p. 437-438, Oct 4 1974.

VARDIMAN, J. W.; HARRIS, N. L.; BRUNNING, R. D. The World Health Organization (WHO) classification of the myeloid neoplasms. **Blood**, 100, n. 7, p. 2292-2302, Oct 1 2002.

WANG, K.; LI, M.; HAKONARSON, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. **Nucleic Acids Res**, 38, n. 16, p. e164, Sep 2010.

WANG, L.; SWIERCZEK, S. I.; DRUMMOND, J.; HICKMAN, K. *et al.* Whole-exome sequencing of polycythemia vera revealed novel driver genes and somatic mutation shared by T cells and granulocytes. **Leukemia**, 28, n. 4, p. 935-938, Apr 2014.

WANG, Z.; JENSEN, M. A.; ZENKLUSEN, J. C. A Practical Guide to The Cancer Genome Atlas (TCGA). **Methods Mol Biol**, 1418, p. 111-141, 2016.

WELCH, J. S.; LEY, T. J.; LINK, D. C.; MILLER, C. A. *et al.* The origin and evolution of mutations in acute myeloid leukemia. **Cell**, 150, n. 2, p. 264-278, Jul 20 2012.

YE, K.; SCHULZ, M. H.; LONG, Q.; APWEILER, R. *et al.* Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. **Bioinformatics**, 25, n. 21, p. 2865-2871, Nov 1 2009.