UNIVERSITA' DEGLI STUDI DI PAVIA

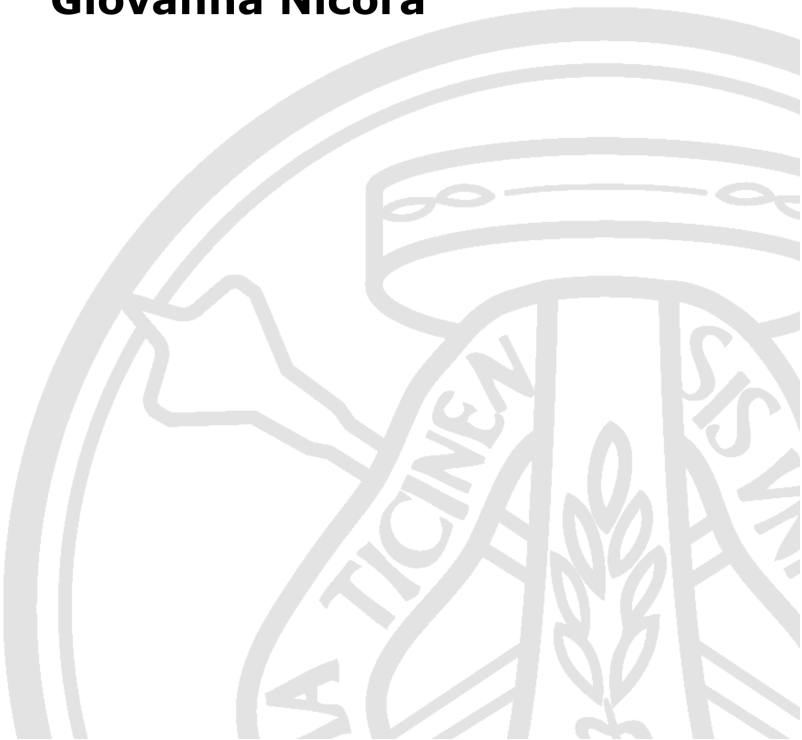
FACOLTA' DI INGEGNERIA
DIPARTIMENTO DI INGEGNERIA INDUSTRIALE E DELL'INFORMAZIONE

DOTTORATO DI RICERCA IN TECNOLOGIE PER LA SALUTE, BIOINGEGNERIA E BIOINFORMATICA
XXXIII CICLO - 2020

Artificial Intelligence strategies for Genomic Variant Interpretation in Hematological Cancer

PhD Thesis by

Giovanna Nicora



Advisor:
Prof. Riccardo Bellazzi

PhD Program Chair:
Prof. Silvana Quaglini

A mio zio Stefano

Abstract (English)

Advances in high-throughput sequencing technologies is leading to the definition of new health care paradigms based on patient's genomics information. Precision Medicine categorizes individuals in homogeneous subpopulations with respect to treatment response or disease prognosis, thus driving clinicians towards an improvement in patient care quality. Given the intrinsic genomics nature of cancer, cancer care would greatly benefit from the realization of Precision Medicine initiatives. Computational tools are essential for the analysis of the huge amount of genomics data produced by sequencing technologies. Bioinformatics pipelines and Artificial Intelligence tools can extract knowledge encoded in patient's genome, assisting clinicians in the application of Precision Medicine's concepts.
Genomic variant interpretation represents a central step in this process. In cancer, the classification of somatic variants reveals possible diagnostic, prognostic and therapeutic biomarkers. Mutations in relevant genes can drive the progression of cancer towards different stages. Variant interpretation should also identify putative pathogenic mutations able to drive cancer. This type of somatic variants is therefore known as driver. Driver identification represents a first step in the discovery of new actionable variants harboring clinically relevant information, such as prognostic indicators.
The research activity described in this thesis aims at supporting the implementation of Precision Medicine in hematological cancer, within the Rete

Ematologica Lombarda (REL) project.
Two complementary tools for automatic somatic variant interpretation are developed. To support the identification of driver variants, a pipeline is projected for somatic oncogenicity prediction based on Machine Learning techniques. The pipeline includes variant annotation according to publicly available tools and databases, and subsequent Machine Learning prediction. Within the proposed approach, unclassified variants detected in tumor samples from large sequencing studies are exploited for unsupervised dimensionality reduction. An incremental model is then trained on pan-cancer transformed data for prediction, allowing its partial re-training when new classified variants are available. This approach addresses issues common in variant interpretation tools and in the application of AI-based tools in the medical context. By including unlabeled variants in the classification process in a semi-supervised fashion, the huge amount of available data is efficiently exploited by Deep Learning Autoencoder. Autoencoder is a type of neural network able to learn a new features representation of the input data in an unsupervised manner. The Autoencoder is then used to obtain new representations of the labeled dataset, with a reduced number of meta-features, with the aim to reduce redundancy and extract the relevant information. The new representations are in turn exploited to train and test a Random Forest Machine Learning algorithm. The Random Forest will also allow future updates of the model when new labeled variants will be available. Our knowledge of the functional role of genomic variants is always evolving, thus it is essential to be able to efficiently update the developed tools. Moreover, by incrementally training the Random Forest with driver variants associated with Myelodysplastic syndromes, a cancer type specific model is developed from the underlying pan-cancer population. To identify possible prediction failures due to dataset shift and unrepresentative data in the training set, it is developed an approach to assess the reliability, or trustfulness, of Machine Learning prediction. A new reliability measure is computed based on the similarity of a new instance (in this case, a genomic variant) to the training set. In particular, it is evaluated whether this example would be selected as informative by an instance selection method, in comparison with the available training set.

The second approach developed to support variant interpretation is a Rule Based Expert System that implements standard guidelines for somatic clinical significance assessment. Such guidelines have been proposed in the last years by working groups including the Association for Molecular Pathology (AMP), the American Society of Clinical Oncology (ASCO) and the College of American Pathologists (CAP). They are implemented worldwide to standardize the identification of therapeutic, diagnostic and prognostic biomarkers within somatic variants. The guidelines consist of a set rules combining knowledge extracted from professional guidelines, clinical trials, public databases, sequencing studies. After knowledge base collection, these guidelines are implemented in a Rule Based Expert System. Expert Systems are Artificial Intelligence systems that emulate expert human reasoning process over a set of rules and knowledge from a specific domain. This approach is able to automatically classify somatic variants into clinically relevant groups defined by the guidelines, while also allowing the user to follow the reasoning process that lead to the final reported classification. In the last section of the thesis, the presence of genomic mutations in cancer genes is studied in the context of Myelodysplastic syndromes progression. Myelodysplastic syndromes are heterogenous diseases that can evolve towards Acute Myeloid Leukemia. Disease progression occurs in subsequent stages defined by a prognostic score. This prognostic score is calculated from clinical data and it is exploited in the clinical setting to understand patient's prognosis. Yet, genomic mutations can drive the progression towards Acute Myeloid Leukemia. To model the influence of mutated genes, patients clinical and genomics data should be collected over time. However, longitudinal analysis requires long-term data collection and curation, which can be time demanding, expensive and sometimes unfeasible. In this context, a clinical decision support framework is proposed that combines the simulation of disease progression from cross-sectional data with a Markov model, that exploits continuous-time transition probabilities derived from Cox regression. Trajectories between patients at different disease stages are stochastically built according to a measure of patient similarity, computed with a matrix tri-factorization technique. Such trajectories are seen as realizations drawn from the stochastic process driving the transitions between

the disease stages. Eventually, Markov models applied to the resulting longitudinal dataset highlight potentially relevant information. Patients' trajectories are computed across increasing and subsequent levels of risks of developing Acute Myeloid Leukemia, and a Cox model is applied to the simulated longitudinal dataset to assess whether genomic characteristics could be associated with a higher or lower probability of disease progression. The learned parameters of such Cox model are used to calculate the transition probabilities of a continuous-time Markov model that describes the patients' evolution across stages. Results are in most cases confirmed by previous studies, thus demonstrating that simulated longitudinal data represent a valuable resource to investigate disease progression of cancer patients.

The approaches developed in this thesis can support the implementation of Precision Medicine in the hematological context. The variant interpretation tools can suggest possible oncogenic mutations, as well as matching treatment. Prognostic indicators can be discovered from Markov and Cox modeling applied to simulated longitudinal datasets.

Abstract (Italian)

Negli ultimi anni stiamo assistendo a enormi sviluppi tecnologici nel campo del sequenziamento genomico. Il sequenziamento *high-throughput* del genoma del paziente permette ai medici di utilizzare le informazioni genomiche per definire i processi di cura. La Medicina di Precisione identifica gruppi di individui omogenei rispetto alla risposta al trattamento o alla prognosi, tenendo conto delle caratteristiche cliniche e genomiche di ogni paziente. In questo modo, ogni paziente riceve la cura più adatta alle sue caratteristiche. La Medicina di Precisione può essere applicata alla cura di tutti i tipi di patologie. In particolar modo, la cura delle neoplasie potrà beneficiare molto dall'applicazione della Medicina di Precisione e del sequenziamento genomico, in quanto il cancro è una patologia dovuta ad alterazioni genomiche. Per poter analizzare la grande quantità di dati genomici prodotta dalle tecnologie di sequenziamento, è necessario sviluppare approcci computazionali, come pipelines bioinformatiche e programmi di Intelligenza Artificiale, che possano estrarre la conoscenza racchiusa nel genoma del paziente. I software basati su questi approcci sono fondamentali per permettere ai medici di applicare i concetti propri della Medicina di Precisione.
L'interpretazione delle varianti genomiche rappresenta un passaggio centrale in questo processo. Nel cancro, la classificazione di varianti somatiche rivela biomarcatori diagnostici, prognostici e terapeutici. Ad esempio, la

progressione della patologia può essere dovuta alla presenza di mutazioni in geni particolari. Nel processo di interpretazione delle varianti, è importante identificare le mutazioni patogeniche che hanno permesso al cancro di svilupparsi. Questo tipo di varianti somatiche è noto come "*driver*". L'identificazione delle varianti "*driver*" è un primo passo nella scoperta di nuovi varianti "*actionable*", ossia quelle varianti contengono informazioni rilevanti dal punto di vista clinico, come per esempio indicatori prognostici. L'attività di ricerca descritta in questa tesi ha lo scopo di supportare l'implementazione della Medicina di Precisione nelle patologie neoplastiche ematologiche, all'interno del progetto Rete Ematologica Lombarda (REL). In particolare, sono stati sviluppati due programmi complementari per l'interpretazione delle varianti somatiche. Il primo tool ha lo scopo di identificare le varianti oncogeniche (*driver*), attraverso l'applicazione di tecniche di apprendimento automatico (Machine Learning). La pipeline sviluppata include l'annotazione delle varianti a partire da software e database disponibili nella comunità scientifica, e la successiva predizione della patogenicità di ogni variante tramite Machine Learning. Nell'approccio proposto, le varianti non ancora classificate, ma presenti in campioni tumorali analizzati da ampi studi di sequenziamento, sono utilizzate per ridurre la dimensionalità del problema in maniera non supervisionata. Un approccio di tipo incremental learning, allenato su varianti pan-cancer, permetterà un ulteriore allenamento parziale quando nuove varianti classificate saranno disponibili. Il tool proposto affronta diversi problemi sia nello sviluppo di software per l'interpretazione delle varianti, sia nell'applicazione di metodologie di Intelligenza Artificiale in contesti medici. Attraverso l'inclusione di varianti non classificate all'interno del processo di classificazione, l'informazione statistica contenuta nella grande quantità di varianti non classificate è stata sfruttata da architetture di Deep Learning come gli Autoencoder. Gli Autoencoder sono un tipo di rete neurale capaci di imparare una nuova rappresentazione degli attributi nei dati di input con un approccio non supervisionato. L'Autoencoder è usato per ottenere una nuova rappresentazione del dataset di varianti classificate per il training, con un ridotto numero di meta-attributi, riducendo la ridondanza e estraendo informazioni rilevanti. La nuova rappresentazione dei dati è usata per allenare e testare un al-

goritmo di Machine Learning di tipo "Random Forest". Questo algoritmo permetterà aggiornamenti del modello quando nuove varianti classificate saranno rese disponibili: la nostra conoscenza sul ruolo delle varianti genomiche in contesto tumorale è sempre in evoluzione, ed è perciò importante poter aggiornare gli strumenti sviluppati in maniera efficiente. Inoltre, attraverso l'allenamento incrementale della Random Forest con varianti driver associate a sindromi mielodisplastiche, è stato sviluppato un modello specifico per una singola tipologia di cancro. Infine, è stato sviluppato un approccio che determina l'affidabilità delle predizioni effettuate dal modello di Machine Learning, permettendo di identificare possibili errori nelle predizioni. Questi errori possono essere dovuti al fatto che il dataset usato per l'allenamento non è sufficientemente rappresentativo della popolazione di varianti che dovranno essere classificate. Per questo motivo, la nuova misura di affidabilità è calcolata a partire dalla similarità di una nuova variante al dataset di allenamento. In particolare, è stato valutato se questo nuovo esempio di variante sarebbe stato selezionato come informativo da un metodo di selezione, confrontandolo con il dataset di allenamento disponibile.
Il secondo approccio sviluppato è un Sistema Esperto basato su regole, che implementa le linee guida per la determinazione della significatività clinica delle varianti somatiche. Queste linee guida sono state proposte negli ultimi anni da diverse associazioni, che includono la Association for Molecular Pathology (AMP), la American Society of Clinical Oncology (ASCO) e il College of American Pathologists (CAP). Queste linee guida sono state implementate a livello mondiale per standardizzare l'identificazione di biomarcatori terapeutici, diagnostici e prognostici all'interno delle mutazioni somatiche. Le linee guida sono costituite da un set di regole che combinano la conoscenza estratta da linee guida professionali, trial clinici, database pubblici e studi di sequenziamento. Dopo la raccolta della base di conoscenza, che comprende le varianti note per essere biomarcatori e presenti in diversi database, le linee guida sono state implementate in un Sistema Esperto basato su regole. I Sistemi Esperti sono sistemi di Intelligenza Artificiale che simulano il processo di ragionamento di un esperto umano in base a un set di regole e alla conoscenza in un dominio specifico. Questo approccio è in grado di classificare automaticamente varianti somatiche nei gruppi di

rilevanza clinica definiti dalle linee guida, permettendo anche all'utente di seguire il ragionamento effettuato dal Sistema Esperto, e che ha portato a una determinata classificazione della variante.
Nell'ultima sezione della tesi, si è valutato come la presenza di mutazioni genomiche in determinati geni di interesse possa influire nella progressione delle sindromi mielodisplastiche. Le sindromi mielodisplastiche sono patologie eterogenee che possono portare allo sviluppo della Leucemia Mieloide Acuta. La progressione della patologia avviene in stadi sequenziali definiti da uno score prognostico, calcolato a partire dai dati clinici del paziente, ed è usato nella pratica clinica per comprendere la prognosi del paziente. Tuttavia, esistono mutazioni genomiche che possono causare la progressione in Leucemia Mieloide Acuta. Al fine di modellizzare l'influenza dei geni mutati nel corso della progressione della malattia, i dati clinici e genomici dei pazienti dovrebbero essere raccolti nel tempo. Tuttavia, le analisi longitudinali richiedono la raccolta dei dati per lunghi periodi, rendendo il processo costoso e talvolta irrealizzabile. È stato quindi proposto un approccio per il supporto delle decisioni cliniche che combina la simulazione della progressione della patologia da dati cross-sectional attraverso un modello di Markov. Le traiettorie tra pazienti a diversi stadi sono costruite stocasticamente secondo una misura di similarità, calcolata tramite tecniche di tri-fattorizzazione. Queste traiettorie sono viste come realizzazioni estratte dal processo stocastico che determina la progressione attraverso diversi stadi della patologia. Infine, modelli di Markov sono applicati al risultante dataset longitudinale simulato per determinare se le caratteristiche genomiche possono essere associate a una maggiore o minore probabilità di progressione della patologia. I parametri del modello di Cox sono usati per calcolare la probabilità di transizione di un modello di Markov a tempo continuo, che descrive l'evoluzione del paziente attraverso i diversi stadi. I risultati ottenuti sono confermati da studi precedenti nella maggior parte dei casi, dimostrando come i dati longitudinali simulati possano rappresentare una valida risorsa per studiare la progressione della patologia.
Le metodologie proposte in questa tesi possono essere di supporto alla implementazione della Medicina di Precisione in un contesto ematologico. Gli approcci sviluppati per l'interpretazione delle varianti somatiche possono in-

dicare possibili terapie, oltre ad individuare possibili varianti oncogeniche. La simulazione della progressione della patologia, combinato con i modelli di Markov e Cox, può permettere inoltre l'identificazione di nuovi indicatori prognostici.

List of Abbreviations

ACMG American College of Medical Genetics and Genomics

AE Autoencoder

AI Artificial Intelligence

AML Acute Myeloid Leukemia

AMP Association for Molecular Pathology

ASCO American Society of Clinical Oncology

CAP College of American Pathologists

DL Deep Learning

ES Expert System

F1 F score

INDEL Insertion/deletion variant

IPSSR International Prognostic Scoring System Revided

MCC Matthews Correlation Coefficient

MDS Myelodysplastic syndromes

ML Machine Learning

MLVD Minimun Variant Level Data

NGS Next Generation Sequencing

PCA Principal Component Analysis

PM Precision Medicine

REL Rete Ematologica Lombarda

RF Random Forest

SNV Single Nucleotide Variants

SVM Support Vector Machine

VUS Variant of Uncertain Significance

Contents

List of Abbreviations **xi**

1 Introduction **1**
- 1.1 Precision Medicine and the Genomics Revolution 1
- 1.2 The role of genomics in hematological cancer 3
- 1.3 Somatic and germline variant interpretation: principles and tool 5
  - 1.3.1 Variant Interpretation Principles 5
  - 1.3.2 Cancer Variant Repositories and Datasets 12
  - 1.3.3 Guidelines-based approaches 16
  - 1.3.4 Data-driven approaches 19
- 1.4 Thesis Outline 22

2 Semi-Supervised Machine Learning approaches for somatic variant pathogenicity assessment **25**
- 2.1 Variant classification problem 25
- 2.2 Data collection and Pre-processing 27
- 2.3 Dimensionality Reduction from Unlabelled data 34
- 2.4 Incremental Learning with Random Forests 40
- 2.5 Comparison with state-of-the-art 50

2.6 Graphical User Interface 53
2.7 Reliability estimation 54
2.7.1 An approach for the determining the trustfulness of Machine Learning predictions on new unseen instance 54
2.7.2 Reliability of the Semi-Supervised Learner 62

3 Implementation of a Rule-based Expert System for somatic variant interpretation in clinical setting 67
3.1 Standard guidelines for somatic variant interpretation . . . 67
3.2 Expert System Implementation 70
3.2.1 Preprocessing: Knowledge Base collection 70
3.2.2 ES Implementation 72
3.3 Case study: application on data from Myelodysplastic syndromes patients 74

4 Genomic variant in prognostic models: Estimation of risk progression in Myelodysplastic syndromes 77
4.1 Risk stratification in Myelodysplastic syndromes 77
4.2 Stochastic Simulation of Longitudinal dataset 82
4.2.1 Stage survival probability computation 83
4.2.2 Patient similarity 84
4.2.3 Progression algorithm 85
4.3 Risk progression estimation based on genomic profiles by Cox and Markov Model 90

5 Conclusions 99
5.1 Somatic Variant Pathogenicity Prediction 100
5.1.1 Feature Transformation 101
5.1.2 Machine Learning model 102
5.2 Somatic Variant Interpretation according to standard guidelines 105
5.3 Risk Progression Prediction in Myelodysplastic syndromes . 107

Appendix 108

A First Appendix **109**

B Second Appendix **113**
B.1 Matrix Trifactorization 113
B.2 Progression Algorithm 115
B.3 Cox and Markov model 117

Bibliography **123**

Publications **185**

List of Figures

1.1 Somatic variant acquisition over time 6
1.2 Cancer Variant Interpretation work ow in Precision Oncology 13

2.1 Spearman correlation coefficient between features of different groups. 33
2.2 Illustration of a general autoencoder architecture 35
2.3 First 2 PCA components for the labelled train and test set 37
2.4 Spearman Correlation on Meta-features extracted from different autoencoders 39
2.5 Mean ROC AUC score on 10-fold cross validation (with standard deviation) and ROC AUC score computed on the test set. 45
2.6 Mean Average Precision score on 10-fold cross validation (with standard deviation) and Average Precision score computed on the test set. 46
2.7 RF performances when the number of meta-features from autoencoders varies 48
2.8 RF Performance on the Test Set 49
2.9 Workflow for the development of the Machine Learning classifier 51

2.10 Confusion Matrix for the developed model and CanDra . . 52
2.11 Confusion Matrix for the developed model and CScape-somatic 53
2.12 Workflow for the classification of new variants 55
2.13 GUI for the variant annotation and prediction of somatic variant pathogenicity 56
2.14 Inner and Outer borders for a two-dimensional problem . . 59
2.15 SVM Results on the simulated dataset 63
2.16 Histogram of reliability measures in the test set 64
2.17 Difference in performance across the complete Test Set, the Reliable subset of the Test set, and the unreliable subset . 65

3.1 Clinical significance Tiers (image from Li et al.9, [1] 69
3.2 Distribution of Therapeutic, Diagnostic and Prognostic SNV and indel biomarkers in the six databases 71
3.3 Upset plot showing intersections among different cancer databases 72
3.4 Output of the Analysis made with the Expert System . . . 73

4.1 Driver and Damaging mutations in the cohort 80
4.2 Percentages of trajectories of different lengths 89
4.3 Markov Model for the description of MDS progression across IPSSR stages 92
4.4 Resulting Markov Model 93
4.5 Survival Curves 93

A.1 Single-Layer Autoencoders Training and Test Loss and R^2 on Unlabeled data 110
A.2 Deep Autoencoders Training and Test Loss and R^2 on Unlabeled data 111
A.3 Training Loss and R2 of an Autoencoder with 4 layers and 35 nodes in the hidden dimension 112

B.1 Consesus matrix representing MDS patient similarities . . . 116

List of Tables

1.1 Repositories that gather cancer variant interpretation . . . 15
1.2 Papers proving interpretation of somatic variants 16
1.3 Tools for Germline Variant Interpretation according to ACMG/AMP guidelines 18
1.4 Tools for somatic clinical significance assessment 19
1.5 Data-driven approaches for the determination of somatic variants driver/passenger status 22

2.1 List of annotation features that characterize each variant . 30
2.2 First 15 genes in terms of number of driver mutations . . . 32
2.3 Results on Reliable and Unreliable subsets of the Simulated dataset 62

3.1 Cancer-specific database information. 71

4.1 Clinical manifestation of 921 MDS patients 79
4.2 Example of simulated longitudinal trajectories 89
4.3 Survival time median (in years) in different IPSSR stages . 94

B.1 "Very Low" to "Low" transition relevant genes 118
B.2 "Low" to "Intermediate" transition relevant genes 119

B.3 “Intermediate” to “High” transition relevent genes 120
B.4 “High” to “Very High” relevant genes 121

Chapter 1

Introduction

1.1 Precision Medicine and the Genomics Revolution

Precision medicine (PM) represents a promising paradigm in health. Its aim is to tailor treatment and care based on patient's specific characteristics. PM categorizes individuals in homogeneous subpopulations with respect to treatment response or disease prognosis, thus driving clinicians towards an improvement in patient care quality [2, 3, 4]. To some extent, the principles of PM have always been adopted in medicine. Patients' characteristics, such as blood pressure or blood glucose concentration, have always been used by physicians to categorize patients to some degree [5]. Yet, it is in the very last years that more specific markers are exploited for patients' categorization. These markers can include molecular information extracted from -omics data. With the term "omics" we refer to the quantification of biological molecules, such as the genome, the transcriptome and the proteome, detected in patients' cells. The molecular characterization of each patient allows us to uncover patient characteristics at an unprecedented level of detail. It is with these extraordinary particulars that researchers

have been able to investigate new frontiers in therapeutic, prognostic and diagnostic procedures [6].
A landmark of this ongoing process is the Human Genome Project, started in 1990 and completed in 2003. It provided for the first time the map of the human DNA [7]. Its completion opened the road for studying the genomics underling a wide range of diseases. Since then, medical genetics were considered to be studying only unusual cases of mendelian (inherited) diseases [8]. Now the genetic basis of different types of disorders have been explained and these insights are exploited in clinical practice to implement PM.
This "revolution" has been possible thanks to advances in high-throughput sequencing technologies, that occurred during and after the Human Genome Project [9]. In particular, the development of Next Generation Sequencing Technologies (NGS) led to decreasing costs and increasing speed in patient's genome analysis [10]. In 2014, the barrier of 1,000 $ per genome (about 840 €) was broken [11]. Yet, a recent study highlights that this cost is still underestimated. By considering the "complete" cost of whole genome sequencing (e.g. including staff's salary and consumables), the average cost reported is around £7,000 (7,500 €) [12]. Another multi-center study evaluated the complete costs of target gene panels sequencing, both for germline and somatic genetics. Target panels focus on a subset of genes of interest, without sequencing the entire genome. The reported analysis shows that the mean cost for a target panel is around 600 € [13]. Given the actual costs of NGS sequencing and data collection, approaches for data simulation represent an interesting field of research in bioinformatics.
Another fundamental aspect that made the Precision Medicine Revolution possible is data analysis [14]. Informatics and big data techniques have been able to process and identify patterns in NGS data. Findings from data analysis provide information able to guide clinical decision making [15]. As we will see, Artificial Intelligence (AI) techniques are increasingly applied to gain information from health data [16, 17, 18]. AI is a field of computer science that includes different techniques whose aim is to reproduce human reasoning. A subfield of AI, known as Machine Learning, (ML) is able to learn hidden patterns and rules from a large amount of

data. ML techniques can be applied to a wide range of fields. In medicine, ML are used to predict the risk of disease onset and to analyse X-ray images [19]. In genomics, ML is widely used to mine the huge amount of data from high-throughput sequencing studies [20]. For instance, ML techniques integrates multi-omics data to find patient subgroups or predict therapeutic/prognostic outcome [21].
Nowadays, PM initiatives are intrinsically collaborative. Ideally, data and information are shared among clinicians, patients, providers, clinical laboratories and researchers [3]. Along with data analysis, bioinformatics pipelines and architectures are essential not only to record and manage information, but also to implement this collaborative framework [22].

1.2 The role of genomics in hematological cancer

Cancer is one of the leading causes of death worldwide. In 2020, the estimate number of deaths in Europe is estimated to be 1.3 million, while 2.7 million new cases are expected [23]. Yet, these estimates do not account for the recent pandemic caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). It is likely that we will observe an increase in the number of avoidable cancer deaths, as a consequence of diagnostic delays during the outbreak [24].
Different types of cancer exist, with different prevalence in men and women. The most common cancer in male individuals is prostate cancer, while breast cancer is the most frequent neoplasm in women and the second with the highest death rate [23, 25]. Hematological malignancies occur when cancer affects blood, bone marrow and lymph nodes. Example of such disorders are Myelodysplastic syndromes, Leukemia, Myeloma, Myeloproliferative disease, Lymphoma. Yet, hematological diseases are highly heterogenous and their categorization may be not standardized in different countries [26].
Cancer, independently from its type, has an intrinsic genomic nature. Its development is due to the acquisition of genomic alterations that confer to the cell neoplastic characteristics. Such characteristics include the ca-

pacity to grow and divide at an abnormally high rate. This leads to the development and progression of cancer clones that can invade other tissues. The DNA of each individual's cells undergoes alteration along lifetime, due to age and exposure to external factors, such as UV radiation. Many of these alterations do not confer any advantage, while some of them will be able to drive cancer progression [27]. Despite cancer is developed stochastically, inherited DNA alterations can confer a higher risk for cancer development. From 5% to 10% of diagnosed tumors are due to hereditary cancer predisposition syndromes [28]. A more detailed explanation of such genomic alterations will be provided in the next paragraph. Due to cancer genomic nature, Precision Medicine initiatives in this context are widely implemented and can be referred to as Precision Oncology. Moreover, the introduction of genomics information in cancer care has caused a paradigm change: therapies are selected based on the individual's genome rather than the cancer types (such as the tissue location) [29]. From here, the importance of the so called *off-label* therapies, i.e. therapies that were developed for a different type of cancer, but that can be prescribed in the presence of the same genomic alterations. For instance, patients with BRAF V600E mutation can undergo the same drug regimen, even if they have different cancer types.

Several Precision Oncology initiatives in the context of hematological malignancies have been launched in these last years. In Denmark, clinicians and researchers developed a workflow for hematological patient care within the Aalborg University Hospital. This workflow is based on both DNA and RNA sequencing to detect possible molecular biomarkers [30]. Efforts have been made to characterize Multiple Myeloma response to target therapies [31]. In Italy, in particular in Lombardia, the "Rete Ematologica Lombarda" ("Hematologic Network") is a Precision Medicine initiative recently born to connect various hematological centers. This network will allow collaborative research and sharing between different hospitals and research centers.

The aim of this thesis is to support the genomic profiling of hematological cancers, in particular myelodysplastic syndromes and Acute Myeloid Leukemia, through the implementation of bioinformatics pipelines and AI methodologies. In particular, the proposed pipelines will be focused on

cancer genomic variant interpretation.

1.3 Somatic and germline variant interpretation: principles and tool

1.3.1 Variant Interpretation Principles

Diseases, like neurological disorders, cardiovascular diseases and cancer, are due to genomic alterations. Such alterations, known as mutations, are more technically known as variants. Genomic variants could be single nucleotide change, short insertion or deletion, copy number alterations or more complex rearrangements. Two types of different variants are detected after NGS sequencing: germline and somatic variants.
Germline variants are inherited. Their presence is detected in all the individual's cells. Germline variants can be pathogenic i.e. they can be responsible for the development of inherited disorders. Such disorders include cardiovascular diseases or neurological disorders. On the contrary, benign germline variants are also at the basis of individuals heterogeneity. Yet, the potential pathogenicity of many germline variants is still far from being understood. This latter group of variants is known as "VUS", an acronym that stands for "Variant of Uncertain Significance" [32].
Somatic variants appear in one or few cells along lifetime (Figure 1.1). When somatic variants occur in genes involved in cell life regulation, they could turn a subset of cells into a neoplasm. Somatic variants driving cancer progression are known as "Driver". Those that do not confer any advantage to the cell are called "Passenger" [33]. Moreover, a subset of driver variants may be "actionable". Actionable variants are biomarkers, i.e. show significant diagnostic, therapeutic or prognostic implication. A subset of actionable variants can be druggable, i.e. target for specific therapies [34]. Yet, also germline variants can have a role in cancer development. In fact, their presence is associated with a higher susceptibility to cancer development [1, 32, 27].

The human DNA is composed of 3 billion nucleotides. Data from large

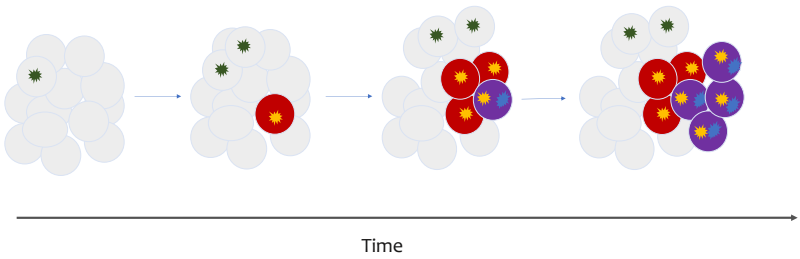


Figure 1.1: Somatic cells acquire mutations during time. These mutations are propagated to cell clones that originate from the same cell. Some of these mutations are passenger, i.e. they do not confer any advantage to the cell. In the image, at the beginning, a cell on the left acquires a passenger mutation (in green). This mutation is propagated to the cell that originates in step 2 through cell division (mitosis). Since the mutation is benign, these cells will not become a neoplasm, and they will undergo a normal cell cycle. Instead, at step 2, a cell acquires a driver mutation (yellow mutation in the red cell), which is able to confer advantage. The cell will develop without control, giving origin to a tumor clone (red cells at step 3). New mutations can also be acquired within the tumor cells (blue mutation in purple cell at step 3). This event will create another tumor subclone with different genomic landscape.

sequencing studies show that the human DNA can contain about 11,000 non-synonymous nucleic acid substitution and hundreds of deletions and insertion, compared to the reference genome. Yet, on average, only 50 to 100 variants in a genome were previously associated with inherited disorders [35]. In 2017 half of variants in public repositories is still reported as VUS [36]. Also, cancer genes vary in the composition of driver and passenger mutations. However, in general the number of driver variants is much lower than the number of passengers in a tumor sample [37, 38]. Still in 2017, it is estimated that half of driver variants occur in cancer genes yet to be discovered [37]. To categorize variants in benign/pathogenic (passenger/driver), variant interpretation should be performed by clinicians.

Variant interpretation is the process of understanding the role of genomic variants during disease development, progression and response to therapy. Cancer variant interpretation includes pathogenicity assessment and clinical significance comprehension. Pathogenicity assessment classifies variants into the different germline and somatic categories mentioned above. Clinical significance assesses whether a somatic variant is a biomarker. Therapeutic biomarkers are associated with response to a particular therapy (actionable variants). Other biomarkers can reveal a particular prognosis, such a more severe outcome of the disease [39, 1]. From one hand, variant interpretation should be able to determine whether any variant detected is pathogenic or not (even if this variant was not previously seen in other patients). On the other hand, variant interpretation should determine whether the patient harbors at list a biomarker. This step relies on current knowledge about biomarkers and supports clinical decisions, by suggesting tailored therapies or follow-up. The two aspects of cancer variant interpretation are clearly connected. By operating on the bench side, pathogenicity assessment can identify new potential biomarkers that could be exploited on the bed side [40]. Over the past 5 years, about 20% of all drugs approved by the Food and Drug Administration (FDA) are specific for Precision Medicine treatments [41]. Clinical trials investigating the power of biomarkers represent important steps in Precision Oncology initiative [42, 43, 22].

As the amount of genomics data that can be collected increases with falling

sequencing costs, variant interpretation represents the actual bottleneck to the complete usage of genomics information within a Precision Medicine clinical practice [39, 44]. Functional variant interpretation represents the definitive approach for the assessment of variant pathogenicity, but it is time and cost-demanding to be used in clinical setting [45, 29]. Large whole genome sequencing studies have provided an unprecedented amount of omics data. Omics depict the genomic landscape of several cancer types. Yet, only for a subset of these data functional validation is available [45]. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium is an international collaboration that identified common pattern of somatic mutations in more than 2,600 tumor samples. Raw data can be downloaded and exploited for analysis. The proposed driver detection is based on known driver events and a ranking approach. This approach is based on variant's recurrence, estimated functional consequence and expected pattern of drivers in that genomic region [46]. An extensive functional interpretation is lacking. However, efforts have been made to provide interpretation of publicly available variants. In FASMIC database, more than 1,000 VUS variants from TCGA have been functionally validated and interpreted [45]. Within the MTB-Report R package, researchers matched TCGA samples with three databases of actionable variants to find therapeutic biomarkers [47]. Another massive project is represented by the AACR Project Genie. This project was launched in 2015 and it contains more than 44,000 tumor cases, matched with clinical information [48].
In clinical setting, variant interpretation has been usually performed by annotating variants and by querying public repositories to find evidence for previous interpretations of detected variants. **Variant annotation** is the characterization of each genomic variant with a set of features that can provide useful information to determine the role of the variant itself. Important information is its location along the genome, such as the gene and the intron/exon in which the variant occurs. Various in silico prediction tools are used to determine the potential damaging impact of the variant on the gene product [49]. These tools can give clues about variant role, but they should not be used to assess pathogenicity, since a disruptive variant may not be implicated in a disease [32]): studies have shown that up to

one third of benign variations could be predicted as disease causing [50] and that even state-of-the-art in silico prediction tools should not be used to infer pathogenicity [51]. Other essential resources to complete variant annotation are public repositories gathering previous interpretation of variants and population allele frequencies (see [52] for a comprehensive review of germline and somatic resources). Several bioinformatics pipelines and web resources are available to perform variant annotation [53], as well as public databases that collect previous interpretation of known variants. As we will see later, these repositories are essential for the implementation of standard guidelines as well for the development of data-driven approaches. For this reason, the next section will review current known repositories.
Variant annotation alone could not be simply translated in variant interpretation. First of all, clinicians need to draw a definitive interpretation from the plethora of information collected for each single variant during the annotation process. Given the number of variants to be considered, evaluating such information for each variant is time-consuming and error prone. Secondly, variant interpretation should be standardized across laboratories. The same variant should not be classified differently from different laboratories. Yet, without standard rules clinicians could draw conflicting interpretations. Guidelines have been proposed throughout recent years to standardize variant interpretation among worldwide laboratories. These guidelines consist of rules that combine information collected during variant annotation [54, 1, 55, 32, 56]. Germline guidelines proposed by the ACMG/AMP working group, and somatic guidelines defined by the AMP/ASCO/CAP have become the standard in clinical practice [57]. Germline and somatic guidelines differ especially for the scope of the interpretation output itself: while the ACMG/AMP guidelines focus on germline variant pathogenicity, the AMP/ASCO/CAP guidelines interpret the *clinical significance* of a somatic variants, i.e. whether the variant is a therapeutic/diagnostic/prognostic biomarker. These guidelines will be further analysed in Chapter 3. Similar to the the AMP/ASCO/CAP working group, the European Society for Medical Oncology (ESMO) proposed their guidelines for clinical significance of somatic variants in 2018 [55]. Germline ACMG/AMP guidelines are applied to a wide range of inherited disorders,

and they are not cancer-specific. However, these guidelines are applied to this context as well, to detect susceptibility to cancer. Disease and gene-specific refinements of the ACMG/AMP guidelines have been proposed by several working group. In cancer, adaptations have been proposed to be applied to the cancer gene TP53 [58] (associated with a broad set of cancer types), PTEN [59] (thyroid, breast and endometrial cancer), CDH1 (gastric cancer) [60]), RUNX1 (associated with myeloid malignancies) [61]. The ClinGen working group also suggested a system to integrate somatic data when classifying germline predisposition variants [62].
Despite the great importance and adoption of standard guidelines in clinical settings, they much rely on current knowledge about interpreted variants. This aspect makes difficult the discovery on new rare pathogenic/driver variants [63, 64].
This is especially true for the widely adopted AMP/ASCO/CAP somatic guidelines for clinical significance interpretation [1]. In fact, their main assumption is whether a variant has already been considered as biomarker by previous studies. Evaluation of the clinical significance represents a fundamental process in the implementation and application of Precision Oncology, since it is able to suggest possible tailored treatment or follow-up based on patient's genomic profile. However, as mentioned above, we are still at the beginning of the process of elucidating the functional roles of genomic variants. Studies have shown that less 10% of patients harbour already known actionable variants [29]. When a rare and unknown variant is found in a patient, it is imperative to determine its pathogenicity. For this reason, guidelines for somatic pathogenicity interpretation are needed. Notably, the Belgian ComPerMed has suggested a rule-based scoring system workflow for somatic variant pathogenicity interpretation in both solid haematological cancer [54]), while the Spanish MDS group proposed guidelines for somatic pathogenicity assessment in myelodysplastic syndromes and chronic myelomonocytic leukaemia [65]. Very recently, the Variant Interpretation Consortium provides a draft for oncogenicity assessment using an approach similar to the ACMG/AMP pathogenicity guidelines [66].
Computational approaches and Machine Learning (ML) represent an opportunity for variant interpretation. In fact, these methodologies allow us

to exploit the huge amount of data that have been collected in these years [67], by highlighting patterns and hidden information that cannot be captured by a simple set of rules [39] and that can predict the oncogenicity of new unseen variants. Issues in the application of ML in variant interpretation, and in general in medicine, remains. Some ML algorithms are often seen as "black-box" models, where it is difficult to understand why a particular prediction has been made [16, 68]. This aspect could hamper the integration of opaque ML predictions in clinical recommender systems. Research towards interpretability and explainability of ML has become an important topic, especially in the light of new European regulations [69]. Yet, some researchers pointed out that *"opaque decisions are more common in medicine than critics realize"* [70].
A major issue is the availability of labelled data (i.e. variants known to be pathogenic or benign). Since supervised ML models work by fitting available data, the more data we have the more we can be statistically reliable. However, we can fail when predicting outcome of rare and underrepresented events [39, 16]. The success of a ML algorithm in classification depends on the robustness of the training set and on the completeness of the features that describe each instance to be classified (in our case, genomic variant). The training set is a subset of instances (variants) with known class (pathogenic or benign) coming from an underlying true population. This true population could be actually unknown or inaccessible. The set of variants known to be benign (passenger) or pathogenic (driver) are therefore seen as a reference for ML algorithms. Such approaches will learn patterns from the reference dataset. These patterns can then be invoked for the classification of new and previously unseen variants. It is crucial that such reference set is as much as possible representative of the entire domain of interest. However, ML systems inherently suffer from dataset shifts and poor generalization ability across different population [71, 16]. Therefore, approaches and metrics to assess the trustfulness, or reliability, of a ML model on a new unseen example could help in the actual application of ML in clinical routine [72, 73]. Within the genomics context, it has been observed that in some cases variants that are used in the training phase to determine model parameter are also encountered in the evaluation

set, thus leading to overfitting. This is a well-known issue in the Machine Learning community that must be carefully address when dealing with genomic variant datasets built from public resources, since such repositories tend to include each other. For instance, when using data from large sequencing studies such as the TCGA or GENIE, it is imperative to assess the overlapping between different datasets, otherwise it may happen that the same variant is shared between the training and the test set. In the genomic domain, the lack of a statistically representative set of variants is a serious risk, as it often occurs that variants from the same gene are jointly reported as pathogenic or benign. This phenomenon is called *"circularity"*: tools affected by this circularity will have good results in pathogenic prediction on genes already known to have pathogenic variants but will fail to detect novel risk genes. Another subtle consequence of the circularity occurs when multiple tools are combined since it is more likely that variants in the evaluation set appear in the training set of (at least) one of these tools [74].
Along with the expansion of NGS sequencing in clinical setting, software and tools are demanded to process the huge amount of data. Tools for variant interpretation could be guidelines-based, if they implement automatically existing guidelines, or could be data-driven, if they rely on statistical and/or machine learning algorithm to infer pathogenicity. Both types of tools are important for the implementation of Precision Oncology in clinical setting (Fig. 1.2), and they will be further analysed in paragraph 1.3.3 and 1.3.4.

1.3.2 Cancer Variant Repositories and Datasets

The implementation of tools for variant interpretation strongly relies on knowledge sharing through public repositories. Important information could be gathered also from published works. In this section, the most important repositories and sources of knowledge for cancer variant interpretation are listed.
Perhaps the most known database in variant interpretation, ClinVar (available at https://www.ncbi.nlm.nih.gov/clinvar/) gathers interpretation of

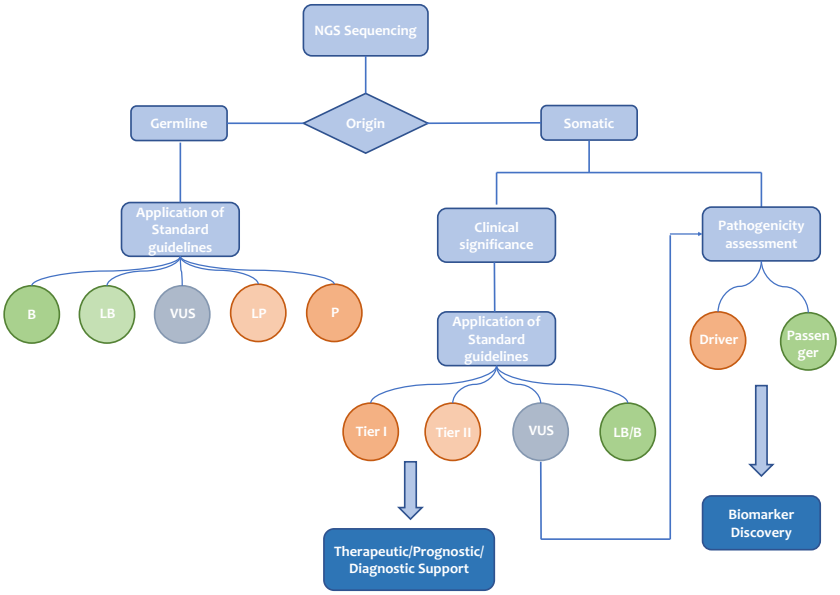


Figure 1.2: Cancer Variant Interpretation workflow in Precision Oncology. After NGS sequencing, the origin of each variant (somatic or germline) is determined. If the variant is germline, the application of ACMG/AMP guidelines allows to determine its pathogenicity. If the variant is still VUS, prioritization approaches could be used to further evaluate the pathogenicity. If the variant is somatic, the AMP/ASCO/CAP guidelines allows to interpret its actionability. If the variant is in Tier I or Tier II, this result could be used to decide patient care. Pathogenicity assessment through computational approaches can help suggesting variants to be studied in the context of biomarker discovery.

829,989 unique variations (at August 3rd 2020). However, it is known that somatic variants, that have such an important role in cancer, are not well-represented in ClinVar [75].
Table 1.1 provides a list of web resources and repositories that gather cancer-related variant interpretation. The majority is focused on clinical significance of somatic variants rather than oncogenicity (driver or passenger classification).
Repositories collecting clinical significance interpretation are mainly focused on therapeutic biomarker, i.e. variants that are associated with response or resistance to drugs. Some of them will be further explained in Chapter 3. The majority of these repositories are pan-cancer, while the IARC database is focused on germline and somatic variants occurring in the TP53 gene. TP53 encodes for a tumor suppression protein and it is one of the most mutated gene across all cancer types [76]. The HemOnc.org is not specifically a precision medicine knowledge base, but contains general information related to hematologic disorders, among which genomic biomarkers for therapeutic purpose [77]. As we will explore in Chapter 3, when developing an automatic tool for variant interpretation, the different repositories are joined together in order to exploit as much information as possible. Databases integration may not be straightforward. Different databases often use different terms to indicate the same concept (such as the same therapeutic level that can be reported as "Response" or "Sensitive"). Also, the variants themselves could be reported with different nomenclatures (genomic coordinates in VCF format or HGVS nomenclature). The lack of a standardization followed by all the repositories can lead to misinterpretation or redundancy. Also, data can be accessed in different ways: through API or by download, while some repositories do not allow any download and information is only available through the web resources.

Published papers describing functional interpretation from sequencing studies provide useful dataset, often reported in their Supplementary Materials. Table 1.2 details few relevant papers for somatic variant interpretation. Functional interpretation of variant(s) can be described within the article's text. This fact can lead to difficulties in retrieving complete infor-

Table 1.1: Repositories that gather cancer variant interpretation

Repository	URL	Reference	Cancer Type	Variant classification
CIViC	https://civicdb.org/home	[78]	Pan-cancer	Biomarker
COSMIC	https://cancer.sanger.ac.uk/cosmic	[79]	Pan-cancer	Biomarker (Therapeutic)
CGI	https://www.cancergenomeinterpreter.org/biomarkers	[80]	Pan-cancer	Biomarker (Therapeutic)
OncoKB	http://oncokb.org/	[81]	Pan-Cancer	Biomarker (Therapeutic)
DEPO	http://depo-dinglab.ddns.net/	[82]	Pan-Cancer	Biomarker (Therapeutic)
DOCM	http://docm.info/	[83]	Pan-Cancer	Biomarker (Diagnostic)
IARC TP53 Database	https://p53.iarc.fr/	[76]	TP53-related cancer	Biomarker (Prognostic)
HemOnc.org	https://www.hemonc.org/wiki/Main_Page	[77]	Hematological cancer	Biomarker
Jackson Laboratory Clinical Knowledge Base	https://ckb.jax.org/	[84]	Pan-Cancer	Biomarker (Therapeutic)
PMKB	https://pmkb.weill.cornell.edu	[85]	Pan-Cancer	Biomarker
MyCancerGenome	https://www.mycancergenome.org/content/biomarkers/		Pan-Cancer	Biomarker
CanDL	https://candl.osu.edu/search/BTK	[86]	Pan-Cancer	Biomarker
DGIdb	http://www.dgidb.org		Pan-Cancer	Biomarker (Therapeutic)
dbCPM	http://bioinfo.ahu.edu.cn:8080/dbCPM/	[87]	Pan-Cancer	Oncogenicity
dbCID	http://bioinfo.ahu.edu.cn:8080/dbCID/	[88]	Pan-Cancer	Oncogenicity
FASMIC	https://ibl.mdanderson.org/fasmic/#!/	[45]	Pan-cancer	Oncogenicity

Table 1.2: Papers proving interpretation of somatic variants

Title	Cancer Type	Reference	Variant classification
Clinical and biological implications of driver mutations in myelodysplastic syndromes	Myelodysplastic syndromes	[91]	Driver/Passenger
Widespread genetic heterogeneity in multiple myeloma: implications for targeted therapy	Multiple myeloma	[92]	Driver/Passenger
Benchmarking mutation effect prediction algorithms using functionally validated cancer-related missense mutations	Pan-cancer	[93]	Driver/Passenger

mation, also due to the use of different non-standard formats for variant reporting. For this reason, the development of methods that can extract genomic variant information from literature can be extremely useful [89, 90].

Along with public studies and databases, also several commercial knowledge bases have been developed in the last years [94, 95].
Comparison of decision support tools highlights that clinical interpretation of somatic variants is still suffering from discrepancies. Despite the introduction of standards, patient clinical care can still be arbitrary, depending on the particular software used for the analysis [95].
Moreover, it is worth to note that variant interpretation could be an ongoing process and it can adapt with gaining knowledge. Reinterpretation of known variants in public repositories are often suggested [96]. It is therefore imperative to periodically update tools that rely on such repositories and versioning variant interpretation in time.

1.3.3 Guidelines-based approaches

As it was mentioned earlier, in the last years the standard proposed by the ACMG/AMP for germline variant interpretation [32] and by the AMP/ASCO/CAP for somatic clinical significance [1] have been adopted worldwide. Given their complexity and the amount of genomic data that need to be processed for each patient, the actual application of such guide-

lines in clinical practice requires the development of automatic tools implementing them, thus speeding the analysis and reducing errors. Here it is provided a review of the recent literature about automatic implementation of standard guidelines, both for germline and somatic variants.

Germline Guidelines-based tools

The ACMG/AMP guidelines for germline variant interpretation are focused on inherited disorders. Cancer, as we already said, starts from stochastically events, but some germline variants can increase the probability of cancer development. Therefore, for hereditary cancer, germline variant interpretation in family members can drive follow-up also for non-affected individuals.
Tools implementing the ACMG/AMP guidelines take as input the list of variants to be classified (usually in VCF format), annotate the variants and provide as output the classification for each variant according the ACMG/AMP rules. A variant will be interpreted as Pathogenic, Likely pathogenic, Benign, Likely benign or "VUS" in case of insufficient or contradictory evidence. Table 1.3 lists some freely available software and web tools for germline variant interpretation according to the ACMG/AMP guidelines.

Among the published tools, only PathoMAN is specifically developed for hereditary cancer.
An issue in the application of ACMG/AMP guidelines is the evaluation of VUS variants [63, 104]. To prioritize VUS variants, CharGer [98] and CardioVAI [100] (that was developed by our laboratory at University of Pavia), introduce scores proportional to the number of ACMG/AMP criteria that were triggered by VUS variants. Also machine learning could be used to solve VUS interpretation.

Somatic Guidelines-based tools

AMP/ASCO/CAP variant interpretation classifies variants into 4 tiers of different therapeutic, prognostic and diagnostic clinical significance. Im-

Table 1.3: Tools for Germline Variant Interpretation according to ACMG/AMP guidelines

Tool	**Publication Year**	**Link**	**Disease-specific**
Intervar [97]	2018	Local installation: https://github.com/WGLab/InterVar	No
CharGer [98]	2018	Local installation: https://github.com/ding-lab/CharGer	No
CardioClassifier [99]	2018	Web resources: https://www.cardioclassifier.org/	Cardiovascular diseases
CardioVAI [100]	2018	Web resources: http://cardiovai.engenome.com	Cardiovascular diseases
Tapes [101]	2019	Local installation: https://github.com/a-xavier/tapes	No
PathoMAN [102]	2019	Web resources at: https://pathoman.mskcc.org	Cancer
CancerSIGVAR [94]	2020 (pre-print)		Cancer
VIP-HL [103]	2020 (pre-print)	http://hearing.genetics.bgi.com/	Hearing loss

Table 1.4: Tools for somatic clinical significance assessment

Tool	**Publication Year**	**Link**
SMART [112]	2018	Web resources: https://smart-cancer-navigator.github.io/home
VIC [113])	2019	Local installation: https://github.com/HGLab/VIC/
OncoPDDS [114]	2020	Web resource: https://oncopdss.capitalbiobigdata.com
MTBP [108]	2020	Web resource: https://mtbp.org/MTBfaqPublic.php

plementations mainly differ in the number of knowledge bases that are integrated. In Chapter 3, an implementation of such guidelines based on AI principles will be shown.
Several groups relies on the AMP/ASCO/CAP guidelines integration in their clinical routine workflows [105, 106, 107, 108]. Also public repositories are adopting the AMP/ASCO/CAP terminology and tier-based classification system [109]. Along with guidelines-based tools, other approaches do not implement AMP/ASCO/CAP guidelines, but provides software for variant analysis, integrating several repositories, such as the Personal Cancer Genome Reporter [110], or AMLVaran, specifically developed for Acute Myeloid Leukaemia genomics data [111]. In Table 1.4 tools and web resources using standard guidelines are reported.

1.3.4 Data-driven approaches

Data-driven tools for germline variant pathogenicity assessment

Variant pathogenicity in clinical practice is widely performed using the standard guidelines proposed by the ACMG/AMP [57]. However, a variant classified as VUS (i.e. uncertain) hampers genetic diagnostic and opens the opportunity for the discovery of new pathogenic/benign variants. In this

context, data-driven approaches can suggest a possible classification, or a possible ranking of VUS variants. Such prioritization helps clinicians to focus on the most likely pathogenic (or most likely benign) variants. As we said earlier, some guidelines-based tools convert the ACMG/AMP criteria into a scoring system [100, 98] that can provide prioritization. Machine Learning (ML) methodologies have been trained to distinguish pathogenic and germline variants. Class probabilities provided as output of ML tools can be easily exploited to prioritize variants. Logistic regression and Random Forest showed high performance in distinguishing cancer predisposition variants in 24 cancer genes [115]. ClinPred combines Random Forest and Gradient Boosting prediction in an ensemble approach. DOCM germline variants were classified, reporting high performance in pathogenic detection [116]. Also MISTIC, that combines prediction from a Random Forest and a Logistic Regression, has been validated on DOCM variants [117]. In [118] authors proposed a tree-based approach and provided systematic comparison with several tools and machine learning models on ClinVar data. Other computational tools of interest are reported in this review [119].

Data-driven tools for somatic variants oncogenicity assessment

In this section, we focus on data-driven approaches for the prediction of driver and passenger mutations, given their importance in cancer. As we said earlier, many cancer genes are yet to be discovered. For this reason, several computational tools aim at discovering new cancer driver genes [120, 121, 122, 123].
It is however clear that, also within a driver gene, not all the mutations will be oncogenic. Data-driven tools that detect driver events at mutation resolutions apply different statistical and/or machine learning approaches to the big amount of data generated from sequencing studies. Among the tools using statistically-derived metrics, MutaGene implements a probabilistic model that estimate the background rate of mutability to rank somatic mutations [124]. Fun-Seq2 combines different annotation features, such as evolutionary conservation, into a weighted scoring system to prioritize variants [125]. Fathmm-cancer (Functional Analysis Through Hidden Markov

Model) models protein domain in humans through Hidden Markov Model and then estimates the functional impact of driver mutations or cancer germline variants [126]. Other approaches combine different tools for a final prediction. CTAT (Combined Tool Adjusted Total) combines different prioritization tools in a single prediction by using the first Principal Component from Principal Component Analysis [127]. TransFIC (Transformed Functional Impact Score for Cancer) uses ontologies and functional scores [128].
ML-based tools usually characterize variants with annotation features, some of which could be gathered from large pre-annotated resources, such as dbNSFP [129]. For instance, the Hematological Predictor of Pathogenicity (HePPy) is a Random Forest trained on features from dbNSFP to distinguish pathogenic somatic variants from germline benign in hematological setting [130]. CHASMplus is also a Random Forest, trained on data from TCGA. Since TCGA variants are not labelled, CHASMplus introduce a semi-supervised approach: labels for training data (driver or passenger) are assigned based on clusters assumptions [131]. CScape-somatic is a Support Vector Machine (SVM) whose results have been assessed using leave-one-chromosome-out cross-validation (LOCO-CV) [132], avoiding gene circularity. CanDrA is also a SVM trained on 96 structural, evolutionary and gene features computed by over 10 other functional prediction algorithms [133]. Table 1.5 reports a list of relevant data-driven approaches.

Recently, a comprehensive comparison of different computational tools to assess their ability in driver prediction has been published [135]. As authors underline, it could be difficult to evaluate the relative performance of such algorithms, for different reasons. First of all, authors tend to choose favourable benchmark datasets to prove the validity and utility of their work. Also, computational approaches often used in cancer research were actually developed for other scope.

Table 1.5: Data-driven approaches for the determination of somatic variants driver/passenger status

Tool	Year	Availability	Cancer type	Mutation type
CHASMPlus [131]	2019	https://github.com/KarchinLab/CHASMplus	yes	missense
MutaGene [124]	2019	https://www.ncbi.nlm.nih.gov/research/mutagene/gene	yes	Missense, nonsense, silent
CScape-somatic [132]	2020	http://CScape-somatic.biocompute.org.uk/	no	Both coding and non-coding
QuaDMutEx [134]	2017	https://github.com/bokhariy/QuaDMutEx		
CanDrA [133]	2013	http://bioinformatics.mdanderson.org/main/CanDrA.	yes	missense
FunSeq2 [125]	2014	http://funseq2.gersteinlab.org	yes	Non-coding
transFIC [128]	2012	http://bg.upf.edu/transfic	no	Non-synonymous
Fathmm-cancer [126]	2013	http://fathmm.biocompute.org.uk	no	coding

1.4 Thesis Outline

As pointed out in the previous section, variant interpretation is a fundamental step within a Precision Medicine workflow. In order to support the implementation of Precision Medicine within the "REL" project, this thesis proposes different computational tools and methodologies to classify cancer variants, both in the context of the oncogenicity and the clinical significance. In the last chapter, it is shown a method based on Cox and Markov modeling to evaluate the influence of genomic variants in cancer progression from simulated data.
The main goal of this thesis is the development and application of AI-based tools for cancer genomic variant interpretation.

The dissertation is organized as follows:

In **Chapter 2** a Semi-Supervised Machine Learning approach for the

oncogenicity interpretation of somatic variants is proposed. The methodology allows an efficient inclusion of unlabeled genomic data from huge sequencing projects. Moreover, the ML algorithm used allows for future updates of the model when new classified variants will be available, or when a cancer-specific model needs to be developed from few variants.

Chapter 3 introduces a guidelines-based AI tool for the clinical significance assessment of somatic variants. The developed tool integrates different repositories and databases to suggest therapeutic, diagnostic and prognostic biomarkers detected in patient's genome, according to the AMP/ASCO/CAP guidelines.

Chapter 4 shows an approach for the simulation of longitudinal genomics and clinical data based on a cohort of patients with Myelodysplastic syndromes. This simulation allows us to define a prognostic model in which mutations in specific cancer genes may determine the progression towards a more severe outcome.

Finally, **Chapter 5** draws the overall conclusions and possible future directions.

Chapter 2

Semi-Supervised Machine Learning approaches for somatic variant pathogenicity assessment

2.1 Variant classification problem

Determining the oncogenicity of a somatic variant is crucial to gain insights into cancer development and to guide possible research on new biomarkers. As we saw earlier, the great advances in high throughput sequencing technologies, combined with national and international research projects, are producing a huge amount of genomics data. The landscape of mutations in different types of cancer has been made available [46, 48]. Yet, we are still far from a comprehensive functional classification of every possible somatic variant. Given the availability of labelled somatic variants, Machine Learning tools have been trained on these variants to detect new driver events (see 1.3.4).

Two problems arise when developing such tools. First of all, standard supervised ML methods rely on the availability of comprehensive labelled datasets, which should be as much as possible representative of the true underling (and unknown) population. This means that the huge amount of not-labelled sequencing data that has been produced (and will be produced) cannot be exploited to train these models. In this context, semi-supervised learning could be applied. Semi-supervised learning techniques combine unlabeled data with labelled data to improve model performance. The assumptions is that there is some structure in the underlying distribution of data that the unlabeled instances will help elucidate [136]. So far, CHASMplus proposes a semi-supervised approach for driver/passenger detection. Its framework consists of two training steps. First, authors labelled TCGA mutations as driver or passenger based on the occurrence of the mutation in a known driver gene predicted to be significantly mutated for the selected cancer type. Then the labelled dataset is used for Random Forest training [131]. However, by labelling variants based on known cancer genes, bias could be introduced, and the algorithm can fail to accurately classify variants that occur in not known driver genes.
The second aspect that should be addressed when developing such tools is that our knowledge of the functional mechanisms that lead to pathogenicity is always evolving. Retrospective studies re-classify variants detected with genetic testing in past years, with the current and most updated knowledge [137]. A recent study retrospectively reclassified hereditary cancer variant detected in 20 years in 1,9 million patients. Among them, about 1,3 million patients harbor a variant that has been reclassified [138]. Reclassification of variants poses ethical and practical challenges within laboratories [139, 140]. As knowledge evolves, computational tools that have been trained with outdated information should be re-trained or at least re-validated. A possible solution could be the implementation of incremental learning algorithms, that are able to incorporate new knowledge without a complete re-training of its parameters. When new classified instances are added over time to the knowledge (training data), incremental ML techniques do not have to completely re-train. Instead they can learn gradually from new data, starting from their current state [141].

In this section of the thesis, it is shown the development of a tool for somatic oncogenicity prediction based on machine learning techniques. The developed pipeline includes variant annotation and subsequent ML prediction. Within the proposed approach, unlabeled variants are exploited for unsupervised dimensionality reduction. An incremental model is then trained on transformed data for prediction, allowing its partial re-train when new classified variants are available.

2.2 Data collection and Pre-processing

In order to train ML algorithms, we used a labelled dataset already available within our research group, that was collected from public resources. Such dataset gathers of more than 84,000 somatic variants known to be Driver or Passenger in different types of cancer. The sources of knowledge from which the labelled dataset is built are the following:

- The published work by **Martelotto et al**. [93], where authors compared different prediction algorithms. About 1,000 of the variants in 15 cancer genes exploited for benchmarking have been functionally validated, and they are therefore known to be driver or passenger.
- The population allele frequency database dbSNP [142]: this repository collects common polymorphisms, and common somatic variants reported in **dbSNP** in cancer genes are selected as passenger variants.
- The **Cancer Genome Interpreter (CGI)** includes driver and passenger variants from different tumor samples [80]. About 1600 variants from the CGI are included in the labelled dataset.

The total number of driver variants collected is 976, while 83,718 instances in the dataset are Passenger. Therefore, the percentage of drivers in the dataset is 1,15%. This great imbalance in the dataset can reflect also the imbalance in tumor samples. Variants occur in 1397 different genes, but only 73 of them have both driver and passenger variants.
Additionally, it was collected an unlabeled dataset of variations detected in

tumor samples. In particular, about 80,000 unclassified genomic variants from the GENIE project [48] were downloaded (in June 2017) to populate the unlabeled dataset.
To characterize each variant with a set of features than can be then used by Machine Learning to derive classification rules, an automated variant annotation pipeline was developed, starting from public tools and repositories. In particular, the Ensembl Variant Effect Predictor (VEP) command line tool was used, since it provides an extensive collection of genomic annotations, such as *in silico* prediction and allele frequencies, and also allows for further custom annotation [143]. VEP is widely used in the bioinformatics community. For instance, the DECIPHER database reports variants from more than 30,000 patients annotated with VEP [144].
Table 2.1 reports the list of features that were selected to annotate each variant. Annotation features can be grouped in different annotation categories:

- ***In silico* predictions**: prediction of the damaging impact of the variant on the gene products. Tools included are mainly ML-based and they usually provide probabilities as output. Some of them are specific for the prediction of particular types of variants. For instance, the MPC, CAROL, SIFT and Polyphen-2 scores reflect the deleteriousness of missense variants [145, 146, 147, 148]. Therefore, for other types of variations all these scores will be null and set to zero. CAROL is also an ensemble tool, meaning that it combines and integrates prediction from other tools, in particular SIFT and PolyPhen-2. One tool (LofTool) is operating at the gene level, providing a measure of tolerance of a particular gene with respect to LOF mutations [149]. Therefore, if the gene where the variant occurs has lower LofTool score, its susceptibility to disease in the presence of LOF variants increases. All the in silico predictions features are included in the VEP command line tool or extracted using VEP plugins.

- **Conservation scores** identify regions along the genome where variation is expected to be more deleterious, even in the absence of specific

annotations of functional elements. Four different scores are included in the annotation: GERP [150], BLOSUM62, PhyloP (46 and 100 way) [151].

- **Gene constraints**: these features are reported at the gene level, therefore all variants in the same gene will have the same values for these features. Genes constraints measure the degree of tolerance of a gene with respect to different types of mutations, such as LOF, missense, nonsense. They have been calculated from ExAC data [152] and the can be downloaded from the ExAC FTP server. ExAC gene constraints scores are used also by the MPC score for damaging prediction [147]. ExAC scores from the downloaded file are added into the annotation pipeline after VEP annotation takes place.

- **Cancer distribution**: this set of gene-based features are extracted from Supplementary Material of Kumar et al. [122], and are added to the annotation pipeline after VEP annotation step.

- **Population allele frequencies** of the variant reported in different databases available in VEP (such as gnomAD).

- **Splicing impact features**: predictions of damaging impact on the splice site.

- **Location**: features related to the location of the variant along the genome. For instance, it is reported the gene in which the variant occurs (not included as ML features), the effect of the variant on the transcript (string) or the protein region affected by the variant.

Some features, such as the effect on the transcript, are actually categories. To deal with categorical attributes, a variation of one hot encoding has been used. For each possible value that the categorical variable can have, we will have a single "transcript effect" feature. Since the variant can occur in different transcript according to different splicing, each "transcript effect" feature is calculated as the percentage of transcripts in which the variant has that particular effect. The total number of ML features after

Table 2.1: List of annotation features that characterize each variant

Category	Feature	Description	Possible values	Source
In silico predictions	ADA score	dbscSNV ADA score [153]	[0-1] -	VEP
	RF score	dbscSNV RF score [153]	[0-1] -	VEP
	LofTool	LoFtool score for gene [149]	[0-1] -	VEP
	MPC	MPC score for missense deleteriousness prediction [147]	[0-5] -	VEP
	CADD_PHRED, CADD_RAW	Scores from the Combined Annotation-Dependent Depletion (CADD) [154]	[0-1] -	VEP
	FATHMM_MKL_C FATHMM_MKL_NC	FATHMM scores for coding variants [155]	[0-1] -	VEP
	SIFT score	Prediction of protein function alteration due to amino acid substitution with SIFT (Sorting Intolerant From Tolerant) [148]	[0-1] -	VEP
	PolyPhen score	Prediction of the amino acid substitution impact by PholyPhen-2 [145]	[0-1] -	VEP
	CAROL score	Impact of missense mutations [146]	[0-1] -	VEP
Conservation scores	GERP	GERP conservation score	[0-1] -	VEP
	PhyloP46	PhyloP46 conservation score [151]	[0-1] -	VEP (custom)
	PhyloP100	PhyloP100 conservation score [151]	[0-1] -	VEP (custom)
	BLOSUM62	BLOSUM62 substitution score for the reference and alternative amino acids	[0-1] -	VEP
Genes Constraints	ExAC_lof_z, ExAC_syn_z, ExAC_mis_z, ExAC_pLI, ExAC_pNull, ExAC_pREC	Z-scores and probability of gene intolerance to different types of variants from ExAC data [152]	[0-1] -	ExAC
Cancer distribution	Patient distribution, cancer type distribution, unaffected residue	Features extracted by the Kumar et al. study, that statistically quantified patient distribution and cancer types with different mutated genes [122]	[0-1] -	[122]
Population allele frequency	AF, AFR_AF, AMR_AF, EAS_AF, EUR_AF, SAS_AF, AA_AF, EA_AF, gnomAD scores	Allele frequencies in different population from 1000 Genome Projects, ESP and gnomAD	[0-1] -	VEP
Splicing impact	SpliceRegion	Granular prediction of splicing effect	Splice region type (string)	VEP plugin
	Max_Ent_scan_alt, max_ent_scan_ref, max_ent_scan_diff,	MaxEntScan alternate, reference and difference sequence score	[0-1] -	VEP plugin
Location	Gene	Gene in which the variant occurs	Gene symbol	VEP
	Transcripts	Transcripts in which the variant occurs	Transcript symbol	VEP
	Effect on transcripts	Consequence of the variant on the transcript	Sequence Ontology terms	VEP
	Uniprot domain	Uniprot domain in which the variant occurs	Domain name according to Uniprot	VEP (custom)
	Repeat Mask	Whether the variant occurs in a repeated region		VEP (custom)
	TSS distance	Distance from the transcription start site	Integer	VEP

this preprocessing is 110.
In order to select a subset of data for testing, the entire dataset is divided based on different genes to avoid the circularity issue [74]. This step is extremely important to avoid overfitting, especially when some features applied to a variant are actually gene-based, such as ExAC gene constraints or the LofTool prediction. Therefore, if a gene is selected as training gene, no variants occurring in this gene will appear also in the test set. Moreover, the partition of genes is selected so that the proportion of driver and passenger variants reflects the proportion in the entire dataset. In Table 2.2 the first 15 genes in terms of the number of driver mutations are reported. The last column reports whether the gene was selected for training or for testing. As we can see, TP53 is the only gene in our cohort with a higher percentage of driver variants compared to passengers. TP53 encodes a tumor suppressor protein containing transcriptional activation, DNA binding and oligomerization domains. The encoded protein regulates expression of target genes, and provokes cell cycle arrest, apoptosis, senescence, DNA repair, or changes in metabolism. Mutations in this gene are associated with a variety of human cancers, including hereditary cancers [156].

After splitting training and test set, we evaluate possible correlation among annotation features in the training set. We calculate the Spearman correlation coefficient [157] for each set of features inside the same functional group (such as population allele frequency or in silico prediction tools). Even before calculating the correlation coefficient we can predict that our dataset can be correlated: for instance, CAROL prediction tools include SIFT scores, or the feature corresponding to the missense effect of the variant may be positively correlated with scores predicting the damaging impact of missense variants only. Figure 2.1 reports heatmap of significant correlations between different annotation scores. We can see that, among in silico prediction tools, the adascores and the rfscores are highly positively correlated, as well as Sift, Polyphen, Carol and MPC. Not surprisingly, CAROL has high correlation with SIFT and PolyPhen since it is built from those scores. Positive correlation is reported also in population allele frequency from ExAC and gnomAD (Fig 2.1b), as well as in conservation scores for PhyloP100 and GERP (Fig 2.1c). Gene constraints metrics are

Table 2.2: First 15 genes in terms of number of driver mutations

GENE	# **Var**	# **Driver**	# **Passenger**	% **DRIVER**	% **PASSENGER**	**Training Gene**
TP53	896	642	254	71,65	28,34	True
ERBB2	264	36	228	13,63	86,363	False
EGFR	2331	34	2297	1,45	98,54	False
PIK3CA	784	31	753	3,95	96,04	False
BRAF	2004	27	1977	1,347	98,65	False
KIT	981	26	955	2,65	97,34	False
KRAS	495	24	471	4,84	95,15	False
BRCA1	754	20	734	2,65	97,34	False
BRCA2	854	12	842	1,4	98,59	False
DICER1	763	11	752	1,44	98,55	False
ESR1	4846	10	4836	0,2	99,79	False
NF1	2658	9	2649	0,33	99,66	True
SF3B1	422	7	415	1,65	98,34	False
PHF6	420	6	414	1,42	98,57	False

both positively and negatively correlated (Fig. 2.1e). In silico prediction scores from CADD and conservational scores GERP and PhyloP100 are only moderately correlated (Fig 2.1d), while poor correlation is reported between in silico pathogenicity scores and population allele frequencies (Fig 2.1f). The LofTool scores are actually not correlated with the ExAC gene constraints in our dataset (data not shown).

Eventually, it is investigated whether there is any strong correlation between features and the desired outcome (the oncogenicity class). Among in silico prediction tools, only PolyPhen, CAROL and MPC have a strong correlation (around 0.75) with the positive class (driver). Among different effects on the transcripts, missense variants have a strong correlation (0.75) with the outcome. Uniprot domain of "Natural variant" has a very strong correlation (0.87) with the outcome. All the other groups of features (such as conservation and gene constraints scores) have weak correlation.

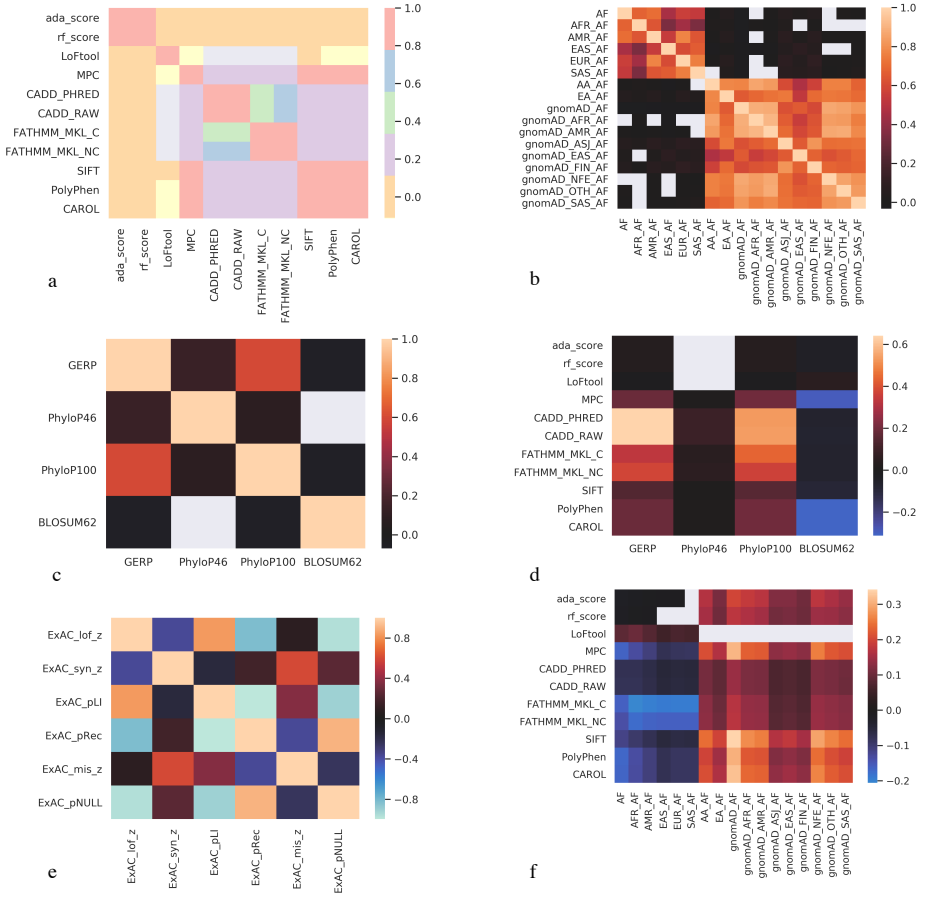


Figure 2.1: Spearman correlation coefficient between features of different groups. Grey areas correspond to not statistically significant correlations (95% confidence) a) Correlation between in silico prediction scores. b) correlation between population allele frequency in different databases. c) Correlation between conservation scores. d) Correlation between in silico prediction scores and conservation scores. e) Conservation between gene constraints score from ExAC. f) Correlation between in silico prediction scores and population allele frequencies.

2.3 Dimensionality Reduction from Unlabelled data

As we see from the correlation analysis, some features are highly correlated. Correlation can be frequent in bioinformatics since many tools tend to exploit previously developed approaches that may also be included in different analysis, possibly leading to circularity [74]. However, many ML models, such as Logistic Regression and Random Forest, become unstable in presence of correlated features [158]. A possible solution could be the transformation of the features in a particular space, such that in the new space the transformed features are actually uncorrelated. Usually, this new space has lower dimension, and it should retain statistical properties and useful information in the data.
A popular data mining technique to perform feature transformation is Principal Component Analysis (PCA). PCA finds a linear subspace in which most of the variability in the data is maintained [159]. In the new coordinate system, the greatest variance by the scalar projection of the data lies in the first component, i.e. the first coordinate. The second greatest variance lies in the second component and so on. The components are also uncorrelated.
Other feature transformation approaches can rely on Deep Learning. Deep neural networks are able to transform features within their layers [160]. In particular, autoencoder architectures (AE) can be used to project features into a non-linear subspace that can retain useful information in data, discarding noise and correlation [161]. AEs are specifically designed with the purpose of learning new features. They consist of two symmetric parts: an encoder and a decoder. The hidden layer between the encoder and the encoder is called *code* (Fig. 2.2). The objective of the AE is to recover a representation of the input data which includes as much information as the original data. The output will have the same number of features of the input data, but, if correctly trained, the AE is able to reconstruct also corrupted or noisy inputs. From the hidden layer of the AE it is possible to extract a lower dimensional representation of the input. This new encoded space can be used for data visualization, as well as to train Machine Learning models [162].

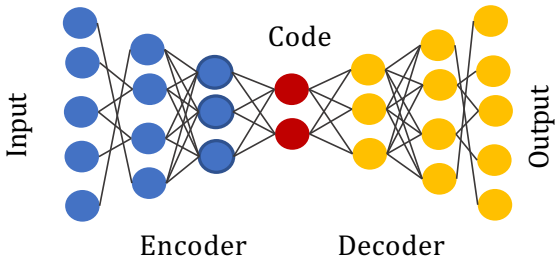


Figure 2.2: Illustration of a general autoencoder architecture

Both the techniques (PCA and AE) are intrinsically unsupervised: unsupervised learning is a subfields of machine learning that aims at finding hidden pattern in data without a pre-existence label. In our case, such approaches will not use the information on the oncogenicity class, but only the intrinsic properties of the data. For this reason, the amount of unlabeled data available from GENIE could be used to train these models. Even if GENIE variants are not classified, they have been reported in tumor patients. Therefore, the assumption is that GENIE data may represent the underlying population from which useful information can be extracted. Secondly, since DL algorithms require a huge amount of data for training, the use of all the available data, even if not classified, could be helpful.
We evaluated whether different AE architectures are able to learn a new representation of the data that can help ML models to learn a better separation of the classes. The AEs will be trained on unlabeled data. The trained AE will be used to project labeled data into the new learned representation. By doing so, we are avoiding the possible information leakage that may occur by training both the AEs and the supervised model with the labeled data. Moreover, if ML prediction will be improved compared with the case of not-transformed annotation features, we will be able to efficiently use the unlabeled information in a semi-supervised fashion.

Unlabeled GENIE variants are divided into 70% training and 30% test for autoencoder training. Moreover, the unlabeled training set is used as reference for data normalization in 0-1 ranges, both for the unlabeled test set and the entire labeled set. The PCA was applied through the scikit-learn implementation in Python, while the autoencoders were developed with Keras. All code to reproduce the results are publicly available on github (https://github.com/GiovannaNicora/semi_supervised_learning-somatic_variant_classification).
The PCA was fitted on the unlabeled training set. The first component is able to explain 19% of the variance, the second 11% and the third 8%. As we can see from Fig. 2.3, if we plot the first two components, driver variants (in red) seem to be located in a particular portion of the 2-dimensional space, especially in the training set. However, passenger variants (in blue) are also detected near driver variants, even if the majority is located in the right portion. This analysis reveals that in this representation a clear separation of the two classes is not discovered.

The first 15 components of the PCA explain almost 80% of the variance in the dataset. Therefore, these 15 components were exploited as features for ML.
Secondly, several autoencoder architectures were implemented and trained. The different AEs differ in the number of layers and the number of nodes in the code (hidden) layer, that will determine the final dimensionality of the dataset. The tested architectures are the following:

- DeepAEs: for each AE, we determine the number of nodes in hidden dimension as one among 2, 5, 10, 20, 30, 40, 50, 60, 70, 80 and 90. From the starting number of features (110), the encoder and decoder layers are five nodes smaller than the previous layer, until the code layer is reached. Therefore, the number of layers in the encoder (or decoder) varies with the number of nodes in the hidden code layer.

- Three-Layer AEs: for each AE, the code layer (with number nodes equal to 2, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90 and 100) is directly connected to the input and output layer.

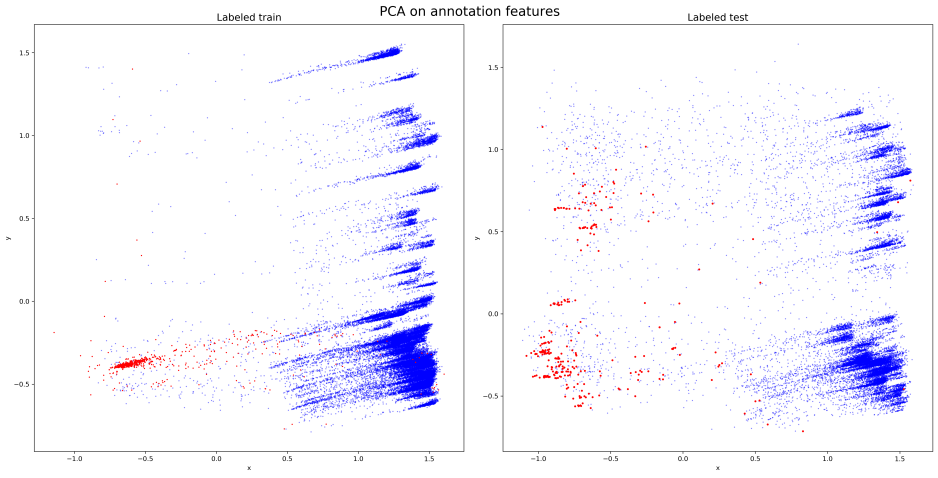


Figure 2.3: First 2 PCA components for the labelled train and test set. Red points are driver variants, while blue points represent passengers.

- An autoencoder with 2 layers in the encoder (and decoder) section, and 35 nodes in the hidden dimension. The first layer in the encoder (and the last layer of the decoder) has 75 nodes, the second layer of the encoder (and the first layer of the decoder) has 50 nodes.

Autoencoders are trained on the unlabeled training set and the unlabeled test set is used to evaluate predictions. In the Appendix B, results in terms of loss and R2 are reported for each architecture, along 50 training epochs. As we can see from Fig. A.1 and Fig. A.2, the three-layer architecture reaches higher performances in the reconstruction of the test set. Moreover, lower dimensions of the hidden layers are associated with a decrease in performance, both for the three-layers AEs and for the DeepAEs. For the three-layer AE, architectures with at least 20 nodes in the code report high performance. Instead, Deeper Networks present sudden decreases in different epochs. Also in this case, architectures with higher number of hidden nodes (and layers) have generally better performance in input reconstruction. The Deep AE with 4 layers and 35 nodes in the code shows good performance with only 50 epochs of training. Since the autoencoders should be able to produce uncorrelated features, we calculated the Spearman correlation on the meta-features predicted from the unlabeled testing dataset for each architecture. In Fig. 2.4, two representative examples are reported for the AE with 50 nodes in the hidden layer and for the DeepAE with 90 hidden nodes. The heatmaps show the correlation between each pair of meta-features, while in the histograms it is shown the number of times a particular value of correlation is detected. Only statistically significant (with 95% confidence) correlations are shown. As we can see, features are poorly correlated, and the majority of correlation is around 0 for both architectures. Yet, several correlations for meta-features from the deeper networks are actually not significant (grey cells in the heatmaps).

It is worth to note that our ultimate goal is not input reconstruction, but dimensionality reduction. We will therefore derive the ability of AEs in reducing the dimension to a meaningful space by evaluating the performance of ML trained on the transformed features. Such transformed features are extracted from the hidden layer of the different architectures,

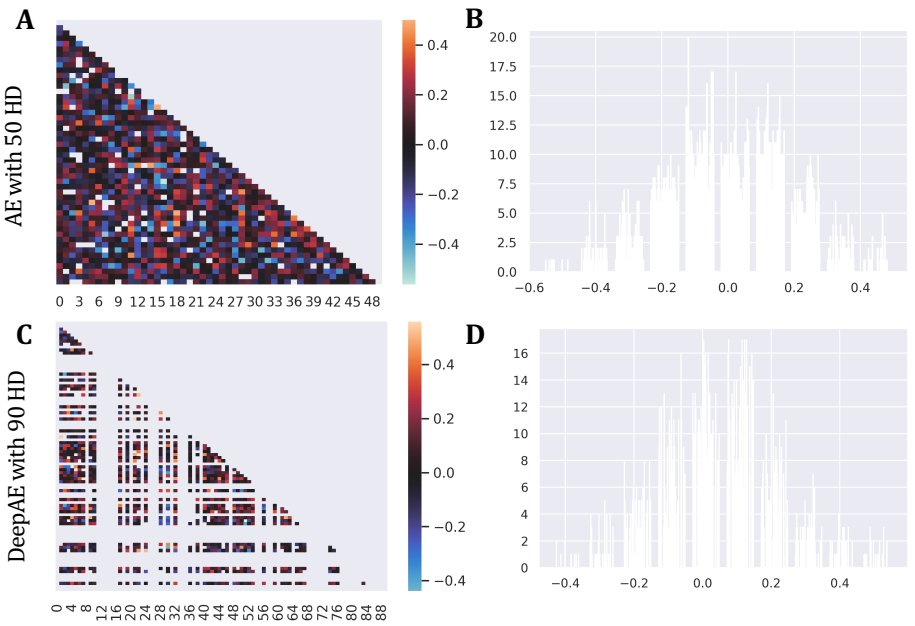


Figure 2.4: A) Heatmap showing Spearman correlation coefficient between meta-features predicted from the AE with 50 nodes in the hidden layer. Grey areas correspond to not statistically significant correlations (95% confidence). B) Histogram of the distribution of significant correlations for meta-features predicted from the AE with 50 nodes. C) Heatmap showing Spearman correlation coefficient between meta-features predicted from the DeepAE with 90 nodes in the hidden layer. Grey areas correspond to not statistically significant correlations (95% confidence). D) Histogram of the distribution of significant correlations for meta-features predicted from the DeepAE with 90 nodes.

and we refer to them as "meta-features", as opposed to the "raw" annotation features, collected from the VEP-based annotation pipeline and provided as input to the AEs.

2.4 Incremental Learning with Random Forests

After features collection and pre-processing, which included features transformation, a ML model to be used for the actual classification was defined. The model has been trained on the labelled training set described above. Moreover, future updates will be accomplished without a complete re-training, in an incremental learning manner. Therefore, the model will not be "frozen" to a particular state, but it will change over time when new labelled data will be available.
To achieve this in a simple and straightforward way within the Python scikit-learn package, the Random Forest algorithm was exploited.
Random Forests (RF) are a popular ML technique that combine several single decision trees into an ensemble classifier. The predicted class is the most frequently predicted within the underlying trees. A single decision tree is a classifier made of different subsequent nodes, that start from a root node and develops in branches, just like a tree. At each node, the data are binary splitted according to the value of a particular feature, so that the new partition set have a higher level of purity with respect to one of the classes. Purity can be measured for instance using the entropy. At the end of the learning process, the tree can generate an IF-THEN classification rule, by following the path that goes from the root node to the leaves.
A Random Forest is made of different decision trees, each trained with a bagging approach, i.e. a sample Z of size N is selected with replacement (bootstrap) from the training set. The single tree is trained on Z with only a subset of randomly selected features. Compared to decision trees, RF have shown to be less prone to overfitting the training data [163].
We used RF for incremental learning in this way. First, a RF with an initial pool of T trees is trained on the training set. The hyperparameter T is set empirically, taking into consideration that a higher number of trees

reduces the variance (overfitting), but increases both training and prediction time. We trained different RFs on the different data representation discussed above: "raw" annotation features, PCA components, and meta-features predicted by the AE and DeepAE. The best RF in terms of balance between precision and sensitivity is selected for deployment. When a new set of labelled data is available, the RF is not re-trained. Instead, new trees are added to the initial pool. The number of such new trees can be set empirically, or such that the proportion between the number of labelled data and number of trees will be the same of the initial training set on the initial number of trees. The re-training will take place only if the number of new labelled data is sufficient to add at least one tree to the RF. During the re-training, the initial pool of trees will not be modified. The pseudo-code

of the Incremental Random Forest is shown below.

Algorithm 1: Incremental Random Forest

Result: Trained Random Forest
Input: Labeled Training Set with **N** instances;
Number of Decision Trees T

1. **for** (*t=1 to T*) {

   (a) Draw a bootstrap sample **Z** of size **N** from the training data.

   (b) Grow a random-forest tree D_t to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.

      i. Select m variables at random from the p variables

      ii. Pick the best variable/split-point among the m.

      iii. Split the node into two daughter nodes.

   }

2. Output the ensemble of trees T_t with $t = 1 : t$.

To classify a new point variant x:
Let $\hat{C}_t(x)$ be the class prediction of bth random forest, then
$\hat{C}_t(x)$ = majority vote $\hat{C}_t(x)$ with $t = 1 : T$

When new Labeled data of size S is available:

- Set the desire number of new trees n_t or calculate it as $n_t = \frac{T \cdot S}{N}$ OR set a number of trees empirically (to weight more or less the new set of training instances)
- if $n_t >= 1$ repeat step 1 and step 2 for the new n_t trees. Add the new trained trees to the previous ones.

Since our dataset is highly imbalanced, results are evaluated in terms of metrics such as the Matthew Correlation Coefficient and the Precision and

Recall Area Under the Curve (PRC AUC) [164, 165]. Data unbalancing is a major issue in ML application, especially in medicine and bioinformatics. Traditional ML approaches usually perform poorly on imbalance dataset, since they tend to minimize error rates of class labels. By doing so, good performances are achieved in the majority class, while errors in the minority class do not impact greatly the overall performance on the dataset. Yet, in many applications, the cost of misclassification for the minority class is much higher. If we think about the medical domain, misclassifying a cancer patient as healthy has a higher cost than misclassifying a healthy patient as unhealthy [166]. In our specific case, we want to ensure that our algorithm is able to find as many positive (driver) variants, while ideally have a low number of False Positive (passenger variants predicted as driver). Therefore, we would like that our algorithm is sensitive and precise. The sensibility, or recall, is the ability to detect driver variants. Precision will ensure us that a small subset of the great number of passenger variants will be reported as driver, while detecting as many driver variants as possible. Data unbalancing in bioinformatics has been faced by Schubach et al., for predicting non-coding variants. Their approach involves under-sampling of negative class and oversampling of the positive class [167]. Overs-sampling and under-sampling are widely used to solve the problem of data unbalancing [168]. Yet, when over-sampling is done, overfitting can arise if applied before data partitioning for evaluation [169]. Moreover, since in our case the proportion of driver and passenger in the dataset is similar to the proportion in real tumour samples, over-sampling and under-sampling were not exploited. Instead, it was adjusted the classification threshold to achieve high precision and recall. The probability threshold for classification is usually 0,5, which works well when the classes are balanced, and the misclassification cost is the same. Therefore, moving the 0,5 threshold for imbalanced data could be a more effective approach [166]. In our case, the best threshold will be the one that maximize the harmonic mean between the precision and the recall, i.e. the F1 score. By doing so, we are ensuring that a good balance between precision and recall is achieved.

Here, results of Random Forests with 100 trees trained on the different data

representations are reported. On the training set, a 10-fold cross validation is performed to evaluate the mean and standard deviation of predictions and to select the best threshold for classification in the test set. K-fold cross validation is a well-known validation technique. It consists of splitting the data into K different sets. Subsequently, the k-th fold is used for testing the ML algorithm trained on the remaining folds [163]. As mentioned above, the training set and test set contain two different set of genes to avoid the circularity problem. Instead, in the 10-fold cross validation, variants are selected randomly to be part of the training or test folds. For this reason, results on the 10-fold cross validation can face the circularity issue. Figure 2.5 and Figure 2.6 report the results on the training set obtained with the 10-fold cross validation and on the test set. Metrics reported are the Area Under the Curve of the Receiving Operating Curve (ROC AUC) and the Average Precision Score. They are computed by varying the classification threshold and evaluate sensitivity and specificity (for the ROC) and the precision and sensitivity (for the Average Precision Score) at each different threshold. Performance on the training set with cross validation are high, both for AE-based model, for the PCA-based features, and for the raw annotation features. Instead, if we compare the metrics on the test set without circularity, the performances drop. These results confirm that, even if the number of features that are gene-dependant is not high (see Data Preprocessing Paragraph), circularity can be a serious issue that can lead to a dramatic decrease in performance when a model trained on variants coming from only a subset of genes is deployed. Moreover, the use of ROC AUC for evaluating performance on unbalanced datasets is confirmed to be a non-optimal choice: if we would focus only on Figure 2.5, despite the difference in performance between the cross validation and the test set, we will report high performance for almost all the RFs across the different data representation.

Since the number of different AEs architecture tested is high, we can focus on the best performing RFs in Figure 2.7, and then compare the bests with the RF trained with the annotation features and the RF trained with the PCA components. In this way, we will qualitatively evaluate whether feature transformation is able to extract informative patterns that

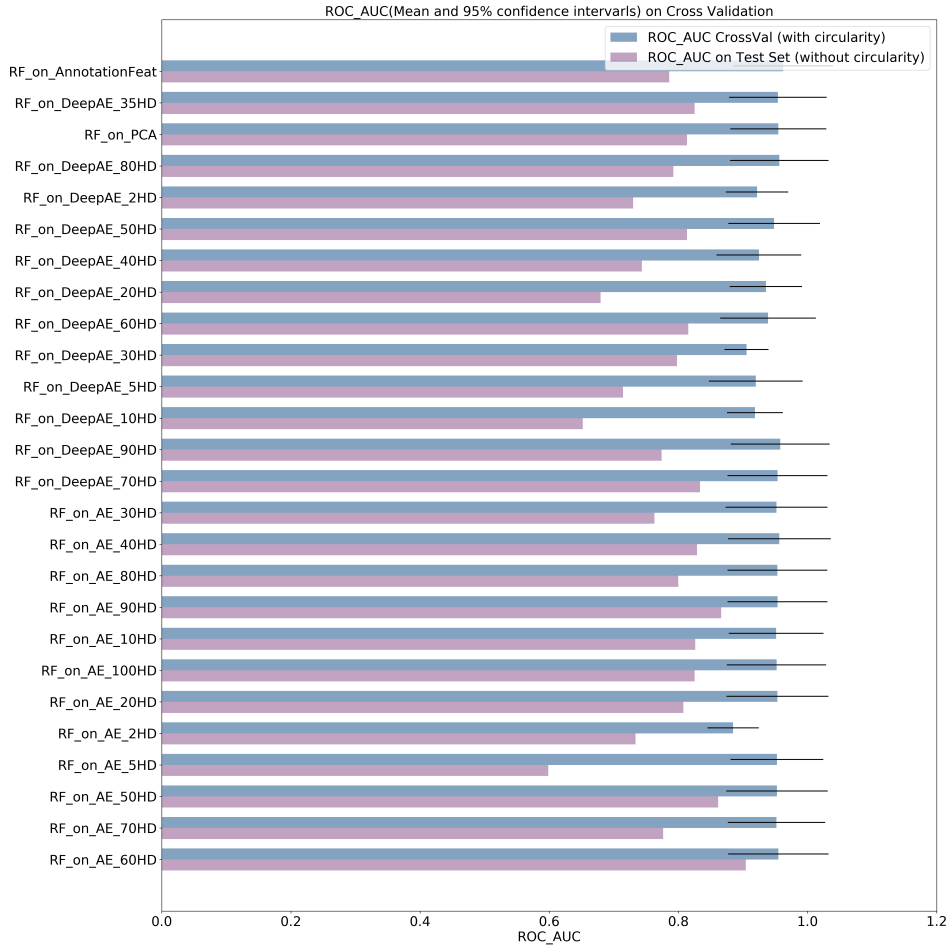


Figure 2.5: Mean ROC AUC score on 10-fold cross validation (with standard deviation) and ROC AUC score computed on the test set.

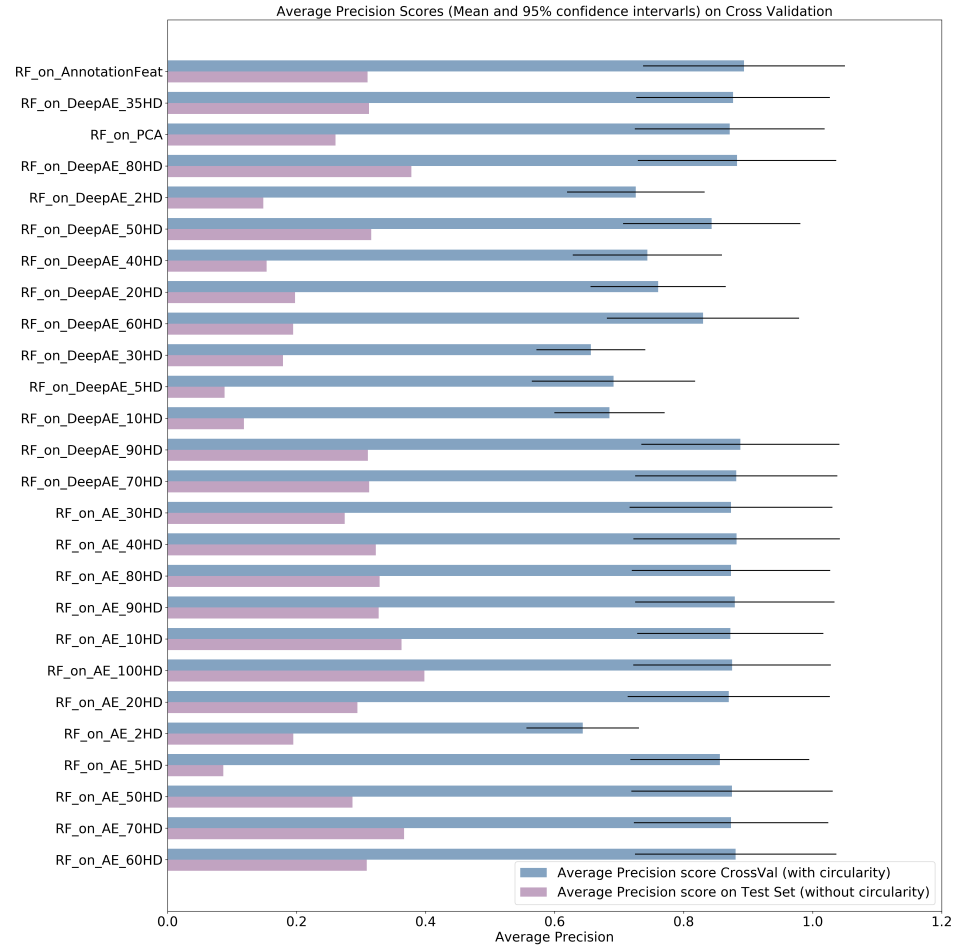


Figure 2.6: Mean Average Precision score on 10-fold cross validation (with standard deviation) and Average Precision score computed on the test set.

are hidden in correlated and sparse data.

Figure 2.7 shows different metrics on the test set, as the number of hidden nodes (on the x-axis) increases for the different DeepAE and AE networks. When recall (sensitivity) increases, precision can dramatically decrease, as the case of the AE with 60 hidden nodes. In this case, the recall is more than 80%, while the precision is lower than 40%. The poor representation on the unlabeled data by the autoencoders with few nodes in the hidden layers is translated also in poor performances for the RF, that therefore is not able to extract meaningful information from the lower dimensional meta-features. The RF trained on the DeepAE with 80 hidden nodes, as well as the RF trained on the AEs with 70 and 80 hidden nodes have balanced precision and recall, around 60%. Results of these three RFs are compared with the RFs trained with PCA components and with annotation features in Figure 2.8. Specificity, i.e. the ability to detect passenger mutations, is high for all the models. The models trained with the PCA components and with the 90 meta-features have high sensitivity, although precision is low, around 40%. The model trained with the annotation features is more balanced, but performances are lower (around 55%). The RF trained with 70 meta-features has the highest precision (65%), with recall equal to 55%. Instead, the RF trained on the 80 meta-features extracted from the DeepAE has slightly lower precision but also slightly higher recall. The RF trained with 100 meta-features has the higher recall (65%). Therefore, the models trained with the meta-features are able to maintain a good balancing between precision and recall, while having higher overall performances compared to the "raw" model with annotation features and also to the model trained with the components from the linear PCA transformation.

Considering that within a cancer sample hundreds of variants can be detected, the desired selected the model will be the one with the highest precision. If many False Positive are reported (i.e. passenger variants classified as drivers), the suggested number of putative driver variants to be studied by clinicians increases, making difficult and time-demanding the evaluation process. The final model selected is therefore the RF trained with meta-features extracted from an autoencoder with a single hidden layer of 70

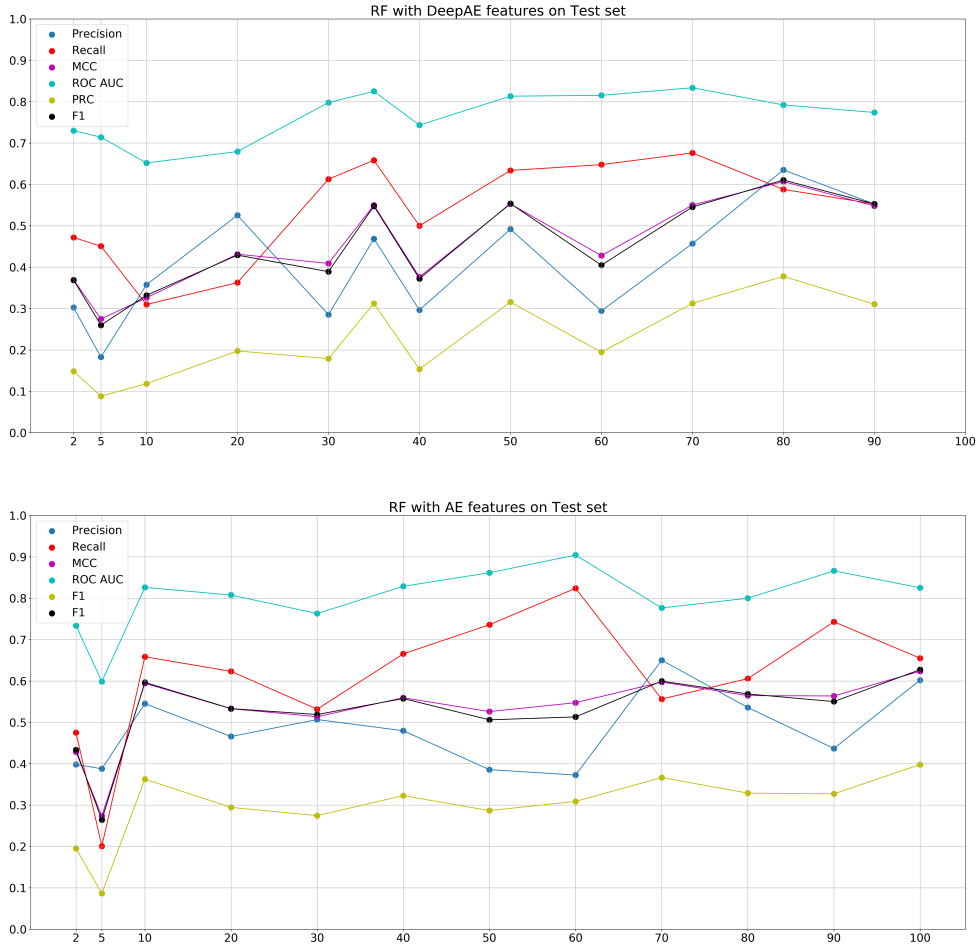


Figure 2.7: RF performances (y axis) when the number of nodes in the DeepAE and AE networks varies (x axis)

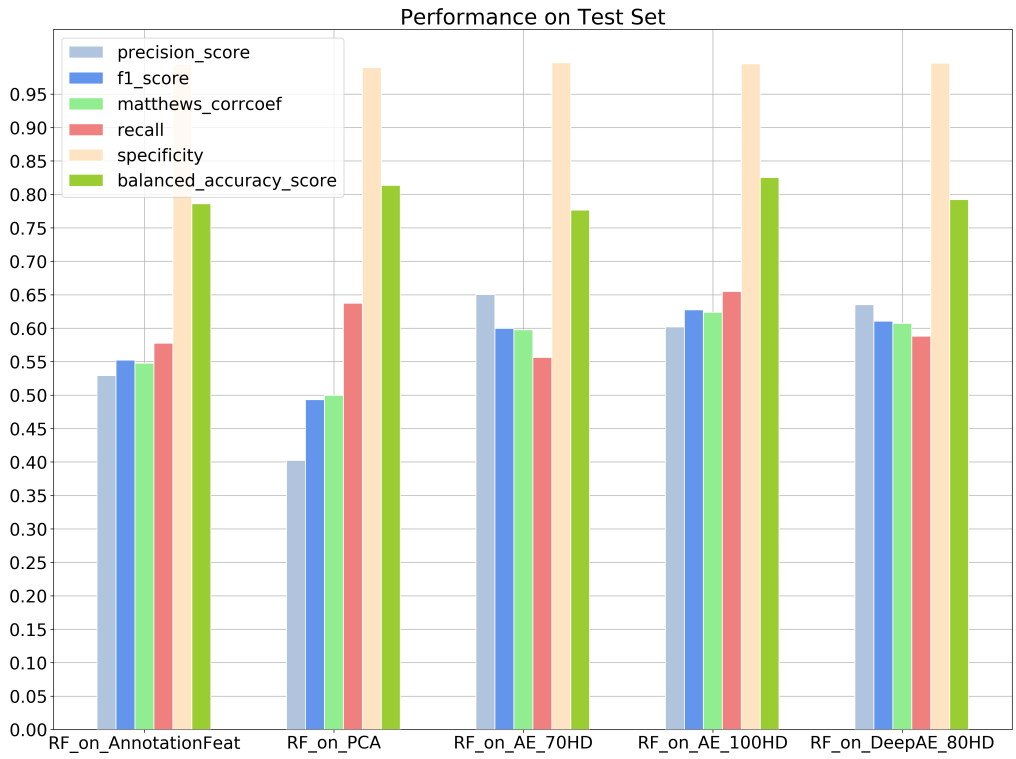


Figure 2.8: Performance on Test Set of the RF trained with the annotation features, RF trained with PCA and RFs trained with meta-features extracted from autoencoders with 70 and 100 hidden dimensions, and the deep autoencoder with 80 hidden nodes.

nodes. The probability threshold for driver prediction, calculated with a 10-fold cross validation described above to maximize the F1 score, is 0,31. Therefore, if a variant has a predicted probability for the driver class that is greater or equal to 0,31, then the class predicted is driver.
Figure 2.9 shows the complete workflow followed for the development of the binary classifier for oncogenicity prediction.

To simulate the situation when new variants are available to update the model, a dataset of variants associated with Myelodysplastic syndromes is used. Myelodysplastic syndromes (MDS) are clonal hematopoietic disorders that can lead to Acute Myeloid Leukemia. MDS will be better studied in Chapter 4. Here, a set of 744 variants detected in 310 patients and validated as oncogenic [170] is used for partial re-training. The RF trained on AE meta-features correctly identifies 240 variants as driver, while the remaining 503 are predicted as passenger. For the incremental implementation of RF, the algorithm 1 is followed: the trees that were previously trained were "frozen", while new trees to be trained with the new set of instances are added. According to the proportion of trees and training samples, only 2 trees would be added. To give more importance to this set of driver variants, the number of new trees that were actually trained was 10. The incremental trained RF is now able to correctly classify 294 driver variants. Still 449 variants are still reported as Passenger, due to the fact that the previously trained 100 trees still have a high weight in the decision process. On the test set, the incremental trained RF correctly classifies 32 more variants as driver, but precision decreases to 55% due to a higher number of False Positive.

2.5 Comparison with state-of-the-art

In this section, the developed semi-supervised machine learning pipeline is compared with state-of-the-art tools for driver and passenger status prediction. In particular, predictions on the test set is compared with CanDrA and CScape-somatic.

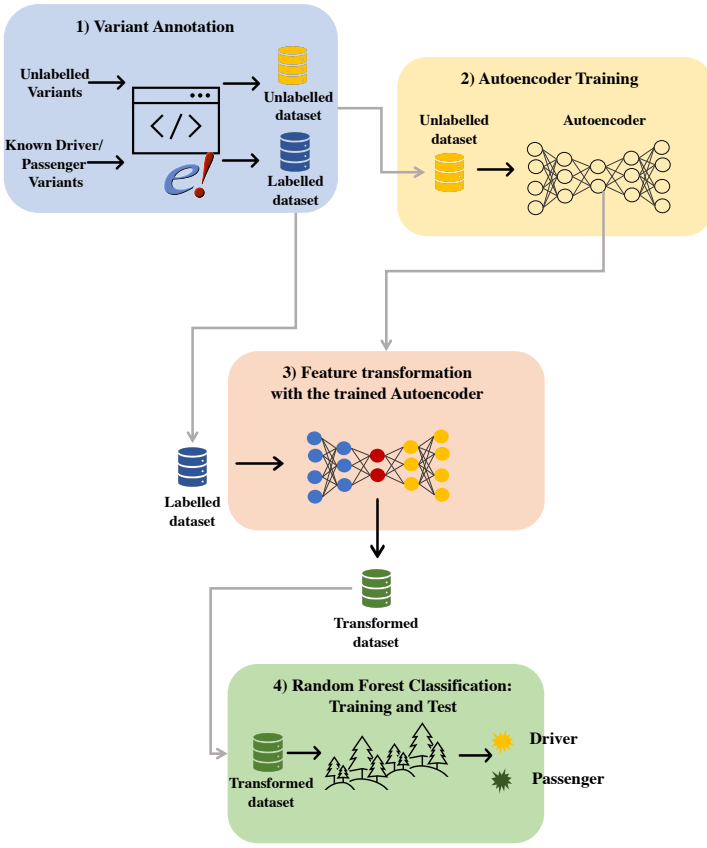


Figure 2.9: Workflow ML classifier development. Starting from the list of unlabelled variants (from GENIE [48]) and the list of labelled variants (training and test set), the annotation pipeline characterized each variant with a list of relevant genomics features (unlabelled and labelled dataset). The unlabelled dataset is exploited to train an autoencoder for feature transformation. This step transforms the annotation features into a set of uncorrelated and informative meta-features. The trained autoencoder transforms the labelled dataset that is used to train and test a Random Forest for classification.

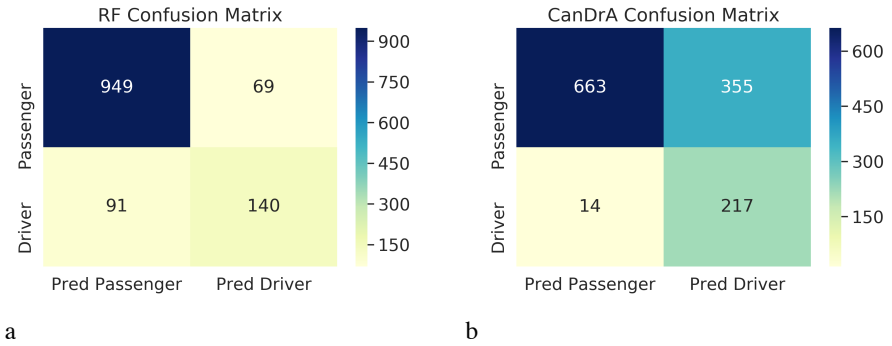
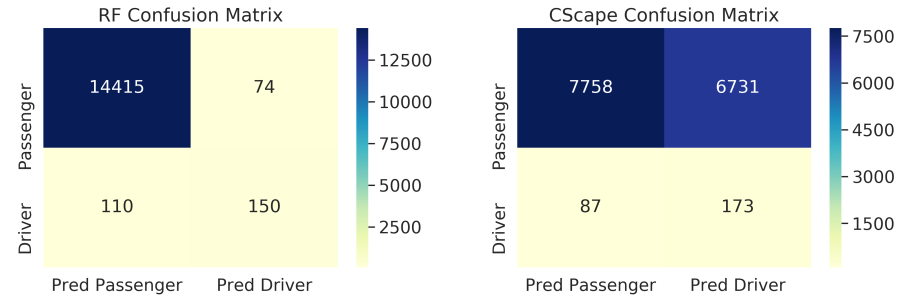


Figure 2.10: Confusion Matrix of prediction on missense variants in the Test Set for the Semi-Supervised Random Forest trained on meta-features from AE with 70 hidden dimension (a) and for CanDrA (b)

CanDrA was developed to classify missense somatic mutations into driver and passenger, based on 95 structural and evolutionary features and in silico predictions. The underlying ML model is a Support Vector Machine, trained on data from TCGA, COSMIC and CCLE [133]. Moreover, CanDrA allows for pan-cancer prediction as well as cancer type predictions. CanDrA (version +) has been downloaded and installed locally to predict oncogenicity of test data. Among test variants, 1550 are missense variants that are predicted by CanDrA. 14% of these variations are driver. On this missense subset, CanDrA has 93% specificity, while the RF with meta-features extracted from the Autoencoder with 70 hidden nodes has only 60% of specificity. Yet, the precision of the RF is 66%, while CanDrA has a precision of 37%. As we can see in Figure 2.10, CanDrA detects a higher number of driver variants, but also the number of False Positive (passenger variants predicted to be driver) is more than five times higher in comparison with the RF.

CScape-somatic, recently published, is still a Support Vector Machine trained on COSMIC and dbSNP data to predict driver and passenger mutations both in the coding and non-coding part of the genome [132]. CScape-



a b

Figure 2.11: Confusion Matrix of prediction on missense variants in the Test Set for the Semi-Supervised Random Forest trained on meta-features from AE with 70 hidden dimension (a) and for CScape-somatic (b))

somatic is used from the web tool available at http://cscape-somatic.biocompute.org.uk/. CScape-somatic classified 14,749 variants in the Test Set, with a recall of 66% but a really low precision (around 2%). On the same data, the RF has a specificity of 57% and a precision of 67%. The number of drivers detected is higher, but also in this case the number of False Positive is dramatically high (6731 compared with 74 False Positive predicted by the RF) (Figure 2.11).

2.6 Graphical User Interface

To support an easy usage of the developed pipeline, a Graphical User Interface (GUI) has been built, using the TKinter Python package. The GUI allows user to predict the oncogenicity of a somatic variant by entering the genomic coordinate or a VCF file with the list of variants to be classified. Then, all the steps described above, from variant annotation to prediction, are performed, and relevant information are shown in the GUI window. Therefore, the GUI (1) integrates the VEP-based annotation pipeline, (2) load the trained autoencoder (single layer, with 70 nodes in

the hidden dimension), (3) load the trained Random Forest model. The complete workflow for the classification of new variants is exemplified in Figure 2.12. The user can predict the oncogenicity of a variant by typing the genomic coordinate in GrCh37 assembly. When annotation and subsequent classification is performed, the GUI will show the final predicted class, as well the predicted probability and the reliability, or trustfulness, of the prediction. This measurement is computed with an approach that will be further explained in the next paragraph. Other useful information shown are the complete annotation information, such as the population allele frequency of the variant and its location. Moreover, a link to the UCSC Genome Browser will show the variant along its chromosome along with useful genomic information [171].
The GUI will also allow possible updates of the model in the incremental fashion described above.
In Figure 2.13, classification of a single missense variant is reported: the variant is located in the DNMT3A gene and it is driver for Myelodysplastic syndromes patients [172]. The developed model correctly identifies the variant as driver, with a reliability of 100%. User can also navigate the annotation features reported in the interactive table: for each transcript in which the variant occurs, the effect on the transcript is reported, as well as allele frequencies, *in silico* predictions, protein location, possible PUBMED identifiers of papers studying that variant, identifier of the variant in other relevant databases, such as COSMIC.

2.7 Reliability estimation

2.7.1 An approach for the determining the trustfulness of Machine Learning predictions on new unseen instance

The majority of ML studies are still restricted to the research stage [16]. Along with logistical and regulatory difficulties in clinical deployment, ML systems inherently suffer from dataset shifts and poor generalization ability across different populations [71, 16]. The diffusion of benchmark datasets representative of true patient populations is essential to develop reliable

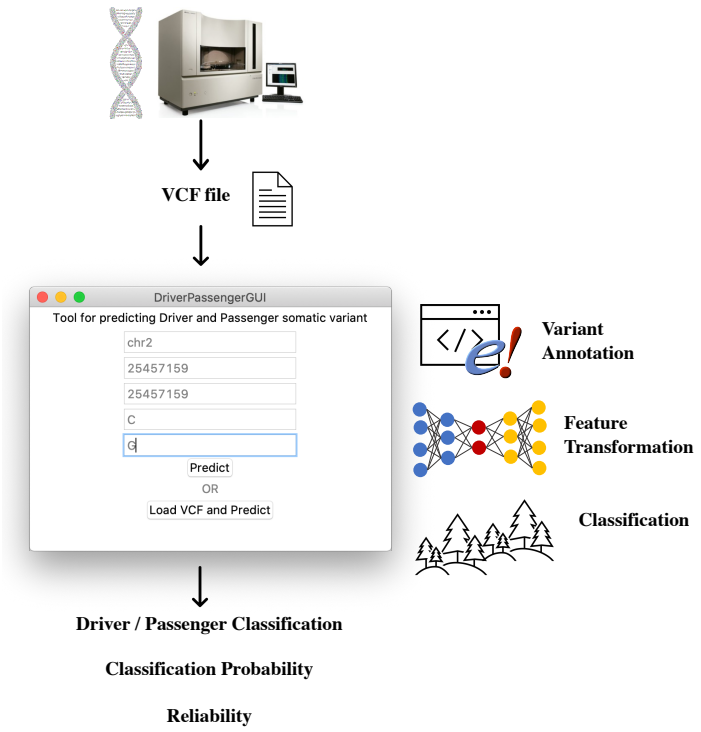


Figure 2.12: Workflow for the classification of new variants. After NGS sequencing, a VCF file with the list of variants to be classified is provided as input for the graphical user interface. Variants are annotated and their annotation features are transformed by the trained autoencoder. The transformed features are then fed into the trained Random Forest. The GUI reports the binary classification and the classification probability predicted by the Random Forest. Trustfullness (i.e. reliability) of the prediction is also reported.

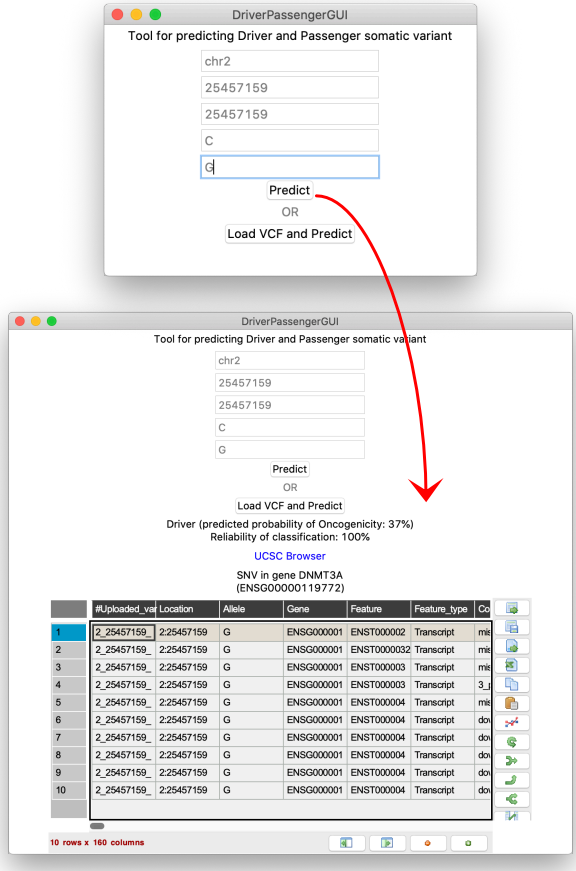


Figure 2.13: GUI for the variant annotation and prediction of somatic variant pathogenicity. User can predict a single variant by writing the genomic coordinates in GrCh37 assembly, or it can predict variants in a VCF file by uploading it. Once the user has entered the variant(s) to be classified, variant annotation according to the VEP-based pipeline is performed. The annotation table for each variant resulting from our pipeline is shown as interactive data-frame thanks to the *pandastable* package. Annotation features are projected in the 70-dimensional space of the autoencoder hidden layer, and then new set of meta-features is fed into the trained Random Forest to predict the driver/passenger oncogenicity class. Predicted class, as long as the predicted probability for driver classification, is shown. Note that the predicted class is driver if the predicted probability is equal or greater than 0.31. Reliability (or trustfulness) of the prediction is shown as well.

and fair AI algorithm to be applied in the clinics. Yet, research in best practice for collection and sharing of meaningful data is still ongoing [173]. Especially in genomics, where data collection and interpretation could be highly expensive, benchmark datasets are limited. Approaches and metrics to assess the trustfulness, or reliability, of a ML model on a new unseen example could help in the actual application of ML in routine clinical practice, in particular when ML algorithms are trained on not representative or shifting data.

Reliability in ML systems could be ensured a priori, by preventing failure, and/or by monitoring and identifying failures when they occur [174]. While the first case deals with dataset preprocessing and model selection, the identification of failure is a monitoring procedure that takes place over time, and it is mainly focused in assessing whether the current trained model is adequate to classify a specific new instance.

Algorithms applying Bayesian inference naturally provide a measure of uncertainty from the posterior distribution, while more recently a new method computes the reliability of a trained model, provided that we have access to the gradient of its loss function [175].

A model-free method to measure reliability of a new instance explicitly compares the class probability distributions predicted from (1) an available trained model and (2) a second model trained on the union between the training set and the new example [176]. The idea is that if these distributions are different, then the new instance is adding information to the training set. Therefore, its prediction made by the first model will not be considered to be reliable. This method requires the re-training of the model for the comparison, and therefore it could be computationally expensive, depending on the amount of the training data and the selected ML algorithm.

In this section, it is proposed a methodology to identify failures of a trained model which is independent from the ML algorithm applied and does not require the re-training of the model. Our framework is based on a previously published approach for training instance selection. Instance selection methods are exploited to remove from the training set the non-useful instances, thus speeding runtime training while not affecting classification

performance [177]. In particular, concepts introduced by the "patterns by ordered projections" algorithm (POP) [178] are incorporated, such as the concept of "border" example, i.e. the instance that is the nearest to an example of the opposite class for a given attribute. In POP instance selection, border examples are those retained in the training set. In this framework, POP is not explicitly used to select training examples, but rather to assess whether a new unseen example would be selected as "border" in comparison with the training set: if so, the new element may not come from the training set population distribution, and therefore we cannot trust the model prediction.
Once training border examples are detected according to the POP algorithm, a ML algorithm, such as Support Vector Machine, Logistic Regression or Random Forest, is trained on the entire training set. As in a typical ML pipeline, performance metrics such as accuracy, sensitivity and specificity are computed on test set predictions from the trained model. At this point, the reliability for test examples is calculated by comparing each test instance to training borders: for a test instance x, the number of times (i.e. attributes) for which x would become a border instance is recorded. In Figure 2.14B, T1 and T2 instances are compared with inner and outer borders. The number of times T1 would be a border is $m_1 = 1$, while for T2 $m_2 = 2$.

The reliability is computed as:

$$rel(x) = 1 - \frac{m_x}{m} \tag{2.1}$$

where m is the total number of attributes and m_x is the number of attributes for which x would be a border instance. In Fig. 2.14b, $rel(T1) = 0.5$, while $rel(T2) = 0$. Intuitively, if a test instance "falls" in an attribute region where no training examples are detected, then we cannot trust the model prediction on it. This is translated in the reliability score by computing the fraction between the number of attributes for which the test instance falls outside the known borders and the total number of attributes.
As a benchmark, an approach based on borders in a multidimensional space

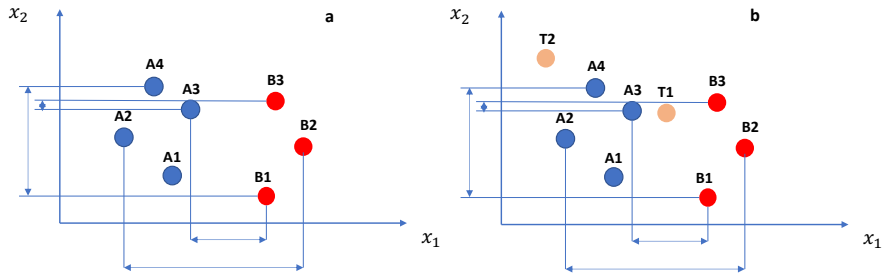


Figure 2.14: a) Examples of inner and outer border examples on two attributes. A1-A4 belongs to class A, B1-B3 to class B. For attribute x_1, A3 and B1 are inner borders, A2 and B2 outer borders. For attribute x_2, A3 and B3 are inner borders, while A4 and B1 are outer borders. b) The test example T1would be inner border for attribute x_1, while the T2 example would be outer example for both for x_1 and x_2 attribute.

is tested to understand whether it could be accurate in identifying unreliable examples. Multidimensional training borders are defined as those training instances which are nearest or more distant to opposite class examples, with distances computed not on each single attribute, but on the entire features set. Therefore, a proper distance metric must be chosen. For instance, if we compute the Euclidean distance on the examples of Fig.1, outer borders will be A2 and B2, while A3 and B3 are inner borders. In this situation, the reliability is a dichotomic variable: a new unseen example is unreliable if it falls outside the borders, otherwise it is reliable.
To summarize, the pipeline is the following:

1. Given a training set, border instances are detected for each attribute independently.

2. A ML algorithm is trained on the training set

3. When a new unseen example must be classified, it is compared to the training borders, and its reliability is computed.

4. We trust the new instance prediction based on its reliability.

In order to evaluate the validity of the method, a binary unbalanced dataset of 6000 samples, with 2 attributes, is simulated. For each class, samples can be drawn from two different clusters of normally distributed points. The left red cluster in class 1 is then "hidden", and a balanced dataset of about 600 samples is selected, for division in training and test set. In this way, the scenario where the real population distribution is not known is simulated and the ML model (in this case, we exploited a Support Vector Machine) is trained on a subset which is only in part representative. Examples excluded from the training and test set are included in a validation set, where classification reliability is evaluated in comparison with training borders. the number of samples (6,000) is chosen empirically to have well-populated datasets for training and testing.
First, results obtained on the validation set from the simulated dataset with the proposed (1) attribute-by-attribute border selection approach (equation 1), with those obtained with the benchmark (2) multidimensional-borders

approach are evaluated. In this latter case, distance between objects is calculated with different metrics, such Euclidean distance or Mahalanobis. Here, reported results are those obtained with the Minkowski distance (r=1), since it showed better performances on this simulated dataset. The unbalanced validation dataset contains 905 red samples and 4475 blue samples. Reliabilities are computed for each validation instance with the two methods. For (1) approach, all those samples with reliability equal to 1 are considered reliablie.
The attribute-by-attribute border selection approach identified 377 unreliable instances out of 5380 validation examples, while the second approach identified 470 unreliable instances. Performance metrics on each reliable/unreliable set are computed. Accuracy on unreliable instances from method (2) is about 76%, while on unreliable instances from method (1) is 55%. The proposed approach (the attribute-by-attribute reliability measure) was able to find more instances for which classification was wrong and therefore unreliable. The following results refer to the attribute-by-attribute border selection method.
As we can see in Table 2.7.1, the percentage of correctly classified examples in the reliable subset of the validation set is much greater than the same percentage in the unreliable set (88.5% against 55%). Therefore, we are able to select a group of instances from an unseen population (the validation set), where we are more confident to perform a correct classification with the available trained SVM.

Fig. 2.15 shows results in terms of accuracy, precision and Matthews Correlation Coefficient (MCC)(Akosa, 2017) on the simulated dataset. On the balanced test set (186 examples) the SVM has 80% of accuracy, 86% of precision in identifying red examples and 60% of MCC. In Fig. 2.15, results on the test set (in blue), on the entire (unbalanced) validation set (in orange), on the reliable validation samples (in green) and on the unreliable validation instances (red bar) are compared. As we can see, performances in terms of accuracy, precision and MCC drop for unreliable examples. The 95% confidence interval was computed for precision, using the entire validation set. The precision mean falls with 95% of confidence in the

Table 2.3: Number and percentage of correctly classified examples (true positive and true negative) and number of incorrectly classified examples (false negative and false positive) in the validation set of simulated data. The first column ("Validation") shows results on the entire set, while the other two columns refer to the reliable and unreliable subsets detected in the validation set.

	Validation	**Reliable Validation**	**Unreliable Validation**
Correctly classified	86 % (4636/5380)	88.5 % (4429/5003)	55 % (207/377)
Incorrectly classified	14 % (744/5380)	11.5 % (574/5003)	45 % (170/377)

interval between 0.55 and 0.57. The precision on unreliable instances falls outside this interval.
The codes for calculating the developed reliability measure are publicly available (https://github.com/GiovannaNicora/reliability/)

2.7.2 Reliability of the Semi-Supervised Learner

Once selected the final model for deployment, it is evaluated whether the suggested framework for identifying failures could work on our genomic dataset. To do so, the procedure describe above was applied. First, training borders are identified on the training set with 70 meta-features extracted from the autoencoder. Then, the test set is compared to the training borders, and the reliability measure is computed for each test variants. The distribution of reliabilities in the test set is shown in Figure 2.16. The majority of test variants has low reliability (around 0,3), but more than 8,000 variants have reliability close to one. Given this distribution, the reliability threshold was set to 0,5: if an example has reliability measure greater or equal this threshold, it is considered reliable, otherwise it is considered unreliable.

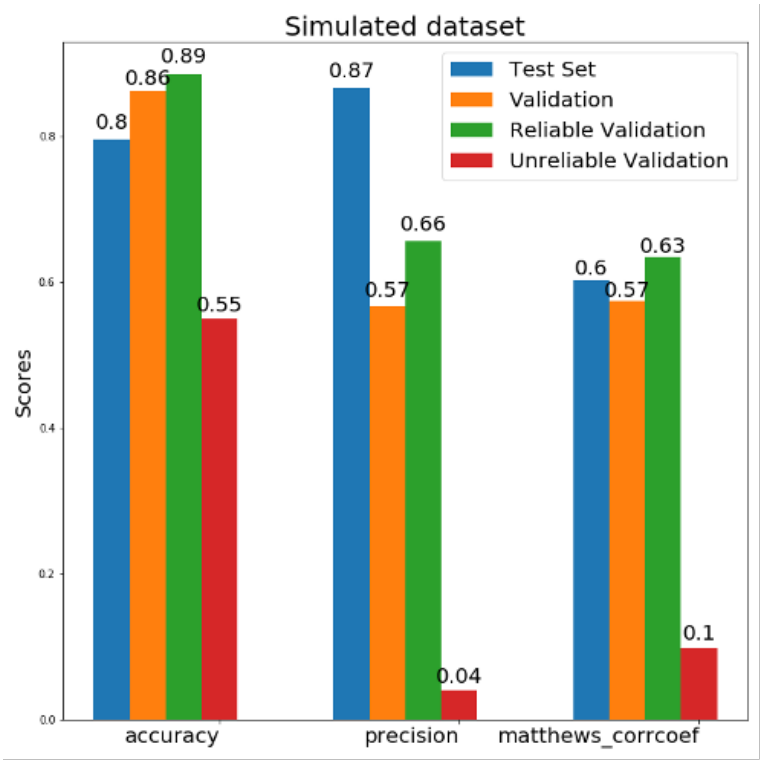


Figure 2.15: Results of a Support Vector Machine on the simulated dataset in terms of accuracy, precision and MCC.

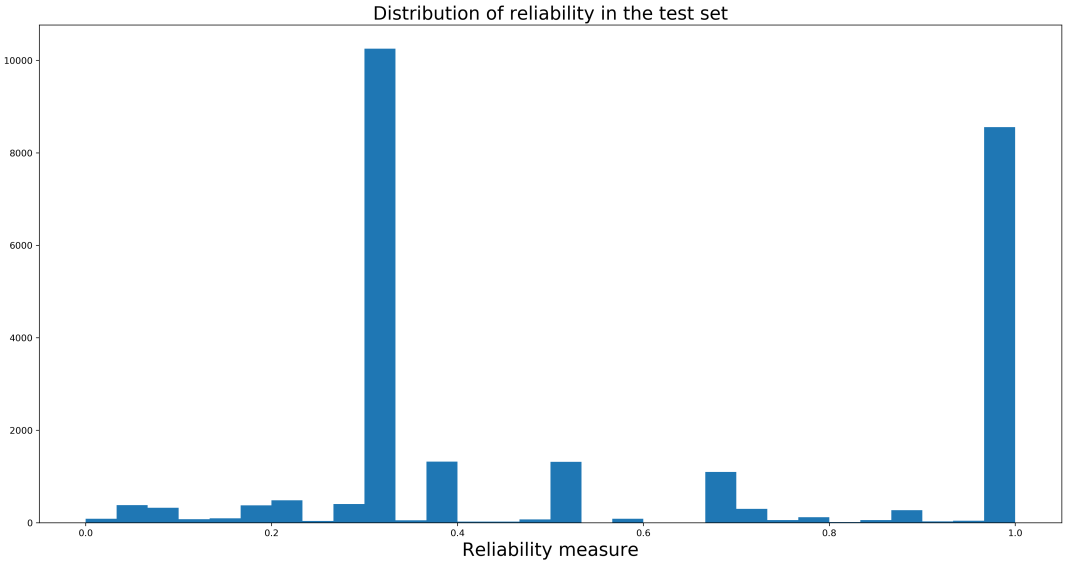


Figure 2.16: Histogram of reliability measures in the test set

In Figure 2.17, precision, recall, MCC and F score are reported for the entire test set, for the reliable variants (11,970 variants) and for the unreliable variants (14,022).

As we can see, on the reliable subset of the test set, our model has slightly higher performance. Instead, on the unreliable test set, all the metrics decrease. On the MDS driver variants, using the same reliability threshold set on the test data (0,5), 344 variants are considered as reliable, while 400 are unreliable. The recall on the reliable set of the RF (not

Figure 2.17: Difference in performance across the complete Test Set, the Reliable subset of the Test set, and the unreliable subset

incrementally trained on the MDS variants) is 48% while on the unreliable variants it drops to 18%.

Chapter 3

Implementation of a Rule-based Expert System for somatic variant interpretation in clinical setting

3.1 Standard guidelines for somatic variant interpretation

In 2017, Li et al. published guidelines for interpreting somatic variants according to their clinical significance [1]. These guidelines were developed by a working group formed by three different associations: the ASCO (American Society of Clinical Oncology), the AMP (Association for Molecular Pathology) and the CAP (College of American Pathologists). These guidelines aim at becoming the standard for clinical significance interpretation in clinical settings. As we saw earlier, a variant

has a clinical significance when it has already been observed as a therapeutic/prognostic/diagnostic biomarker for a particular cancer type. A therapeutic (or predictive) biomarker is a variant that predicts response or resistance to specific therapies. Examples of cancer therapies are Molecular Target Agents (MTA), drugs able to disrupt oncoproteins produced by driver-mutated genes [179]. Their use represents a safe and promising alternative to chemotherapy, given their specificity in targeting tumor cells only. An example of this drug is vemurafenib. Response to vemurafenib treatment is predicted positively by the presence of BRAF V600E mutation in patients with melanoma. Yet, caution must be taken when using target drugs, since drug resistance development is frequent in cancer [180]. It may occur that tumor clones without the targeted mutation survive. Given the heterogeneity of the tumor even within the same patient, combination of different MTA and/or chemotherapy is frequent [181].

Prognostic biomarkers are variants associated with a more favorable or adverse disease progression, typically referred to the survival time. For instance, mutations in the Core binding factor protein are associated with favorable prognosis in patients with Acute Myeloid Leukemia [182]. Diagnostic biomarkers, instead, allow the early detection of cancer as well as secondary prevention [183]. For instance, the gene fusion PML-RARA is diagnostic for promyelocyt leukemia. Biomarkers could span the different categories: for instance, the PML-RARA fusion is both diagnostic and predictive of response to target molecules [1]. The ASCO/AMP/CAP guidelines combine annotation features and previous interpretations of variants to determine their clinical significance. In particular, they defined 4 levels (from Level A to level D) of clinical and experimental evidence that can be assigned to each somatic variant. Each level has a definition with respect to the type of significance (therapeutic, diagnostic or prognostic). For instance, Level A Therapeutic is applied to variant that predicts response/resistance to therapy according to the FDA or professional guidelines for a specific type of tumor. Level A Prognostic includes all variants already defined as prognostic by professional guidelines for the specific type of tumor. The strength of the evidence decreases across the levels. Level B applies to biomarkers not yet included in professional guidelines but with

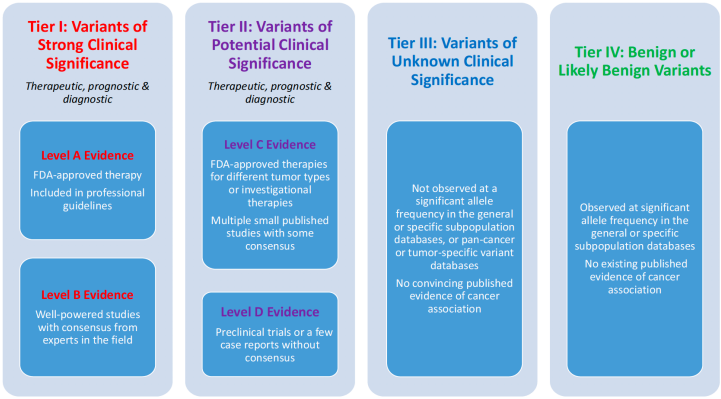


Figure 3.1: Clinical significance Tiers (image from Li et al.9, [1]

well-powered studies supporting the biomarker definition. Level C includes biomarkers approved by FDA/professional guidelines for a different type of tumor or that serve as inclusion for clinical trials. Level D gathers biomarkers whose effect is reported in preclinical studies. The different Levels are combined as shown in Figure 3.1 (from [1]) to classify the variant into 4 different Tiers.

Different issues can hamper the actual application of such guidelines in clinical practice. First of all, they require the integration of different sources of knowledge (such as professional guidelines and databases), whose systematic access could be difficult. Moreover, terminology and variant nomenclature across repositories widely vary. A recent study reported low agreement in variant classification according to these guidelines performed by different laboratories. Barriers perceived are the complexity of these guidelines and the lack of familiarity, as well as discordance between clinical significance and oncogenic relevance [184]. For this reason, the development of automatic tools that implements the ASCO/AMP/CAP guidelines are needed.

3.2 Expert System Implementation

We developed a Rule-based Expert System (ES) able to automatically interpret somatic variant according to AMP/ASCO/CAP guidelines. ESs are Artificial Intelligence systems that emulate expert human reasoning process over a set of rules and knowledge from a specific domain [185]. In our case, rules are represented by the AMP/ASCO/CAP guidelines, while the domain knowledge needs to be gathered from several public omics-resources. After knowledge base collection, the ES is implemented in a Python program thanks to the PyKnow library, which creates an environment to define Rules and fire them against Facts. ES architecture will allow future updates of the Rules, avoiding complex alteration of the application code. The ES receives as input a list of genomic variants, it performs inference, and then provides as output a JSON file for each variant, reporting variant annotation and AMP/ASCO/CAP interpretation, according to MVLD. Thanks to output files, the ES allows user to follow the reasoning process that lead to the final classification.

3.2.1 Preprocessing: Knowledge Base collection

We collected information about known biomarkers from 6 different cancer-specific databases among those that we discussed in Chapter 1. These repositories provide evidence about variants' clinical impact, public literature references, clinical trials and professional guidelines. Information about cancer-specific databases is listed in Table 3.1.

We developed automatic pipelines that extract relevant information and standardize nomenclature from each resource. In fact, each database has different terminologies: for instance, OncoKB Therapeutic levels are "Resistance" or "Response", while in DEPO the same concept is represented by "Resistant" and "Sensitive". Moreover, we standardized cancer representation to Disease Ontology terms, and we select single nucleotide variations and indels.

As we can see from Figure 3.2, the majority of variants are therapeutic biomarkers. DOCM, with diagnostic biomarkers, and CGI, with predictive

Table 3.1: Cancer-specific database information.

Database	URL	Type of evidence
CGI [80]	https://www.cancergenomeinterpreter.org/biomarkers	Therapeutic
CiVIC [78]	https://civicdb.org/home	Therapeutic, Diagnostic, Prognostic
OncoKB [81]	http://oncokb.org/	Therapeutic
DEPO [82](	http://depo-dinglab.ddns.net/	Therapeutic
DOCM [83]	http://docm.info/	Diagnostic
COSMIC (Resistance Mutation) [79]	https://cancer.sanger.ac.uk/cosmic/download	Therapeutic

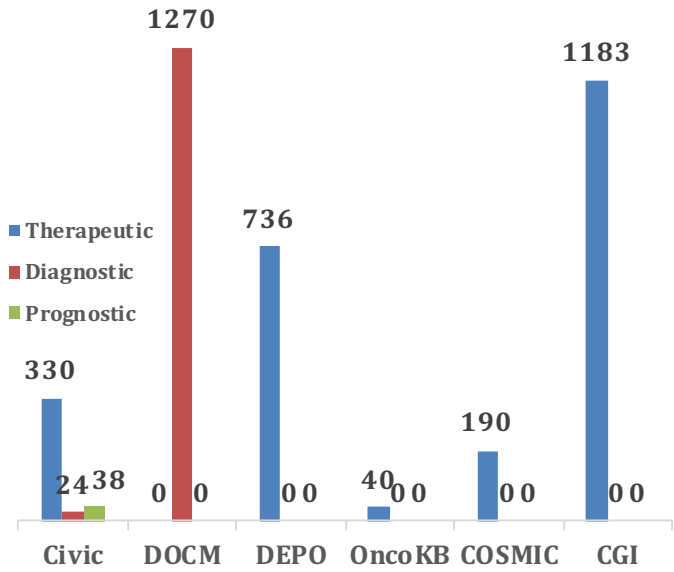


Figure 3.2: Distribution of Therapeutic, Diagnostic and Prognostic SNV and indel biomarkers in the six databases

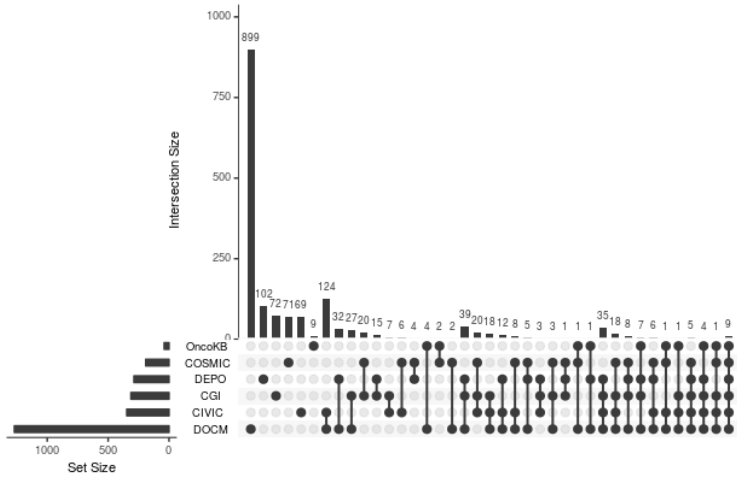


Figure 3.3: Upset plot showing intersections among different cancer databases

biomarkers, have the highest number of variants. A variant can be reported for different significance groups (for instance, a variant can be both a therapeutic and diagnostic biomarkers). A total number of 1681 of variants are gathered into the knowledge base. As we can see from Figure 3.3, the overlapping among the repository is high. For instance, in OncoKB, all variants except 9 are reported also in other databases.

3.2.2 ES Implementation

The ES is implemented in a Python program. Input files are the following: an annotation tab-delimited file with the lists of genomic coordinates of somatic variants that need to be classified and a tab-delimited file for each collected omics-resource, resulting from our preprocessing pipeline. Data are organized into an Object-oriented model.
Rules, representing AMP/ASCO/CAP guidelines, are defined through PyKnow. For instance, the final rule for "Tier I" classification is composed

CHROM	START	STOP	REF	ALT	GENE	PATIENT_PHENOTYPE	AMP_ASCO_CAP_CLASSIFI	AMP_ASCO_OBSERVED_	CLINICAL_UTILITY	CLINICAL_SIGNIFICANCE
9	133750263	133750263	C	T	ABL1	DOID_8552-chronic myeloid l	Strong Clinical Significance-I	A	CGI	Therapeutic	Resistance|Imatinib
12	25398284	25398284	C	T	KRAS	DOID_285-hairy cell leukemia	Strong Clinical Significance-I	B	CIVIC,Docm	Therapeutic	Resistance|Melphalan,Ve
12	25398284	25398284	C	T	KRAS	DOID_2394-ovarian cancer	Strong Clinical Significance-I	B	CIVIC,Docm	Therapeutic	Resistance|Melphalan,Ve
12	25398284	25398284	C	T	KRAS	DOID_9256-colorectal cancer	Strong Clinical Significance-I	B	CIVIC,Docm	Therapeutic	Resistance|Melphalan,Ve
12	25398284	25398284	C	T	KRAS	DOID_1324-lung cancer	Strong Clinical Significance-I	B	CIVIC,Docm	Therapeutic	Resistance|Melphalan,Ve
12	25398284	25398284	C	T	KRAS	DOID_1909-melanoma	Strong Clinical Significance-I	B	CIVIC,Docm	Therapeutic	Resistance|Melphalan,Ve
12	25398284	25398284	C	T	KRAS	DOID_9538-multiple myelom	Strong Clinical Significance-I	B	CIVIC,Docm	Therapeutic	Resistance|Melphalan,Ve
12	25398284	25398284	C	T	KRAS	DOID_1795-tumor of exocrine	Strong Clinical Significance-I	B	CIVIC,Docm	Prognostic	Poor Outcome
12	25398284	25398284	C	T	KRAS	DOID_3587-pancreatic ductal	Strong Clinical Significance-I	B	CIVIC,Docm	Prognostic	Poor Outcome
12	25398284	25398284	C	T	KRAS	DOID_1793-pancreatic cancer	Strong Clinical Significance-I	B	CIVIC,Docm	Prognostic	Poor Outcome
12	25398284	25398284	C	T	KRAS	DOID_1324-lung cancer	Strong Clinical Significance-I	B	CIVIC,Docm	Diagnostic	Positive
12	25398284	25398284	C	T	KRAS	DOID_9256-colorectal cancer	Strong Clinical Significance-I	B	CIVIC,Docm	Prognostic	Positive
21	44514777	44514777	T	G	U2AF1	DOID_9119-acute myeloid leu	Strong Clinical Significance-I	B	CIVIC,Docm	Diagnostic	Positive
21	44514777	44514777	T	G		DOID_0050908-myelodysplas	Strong Clinical Significance-I	B	CIVIC,Docm	Prognostic	Poor Outcome
21	44524456	44524456	G	A	U2AF1	DOID_0050908-myelodysplas	Strong Clinical Significance-I	B	CIVIC,Docm	Prognostic	Poor Outcome
21	44524456	44524456	G	A	U2AF1	DOID_9119-acute myeloid leu	Strong Clinical Significance-I	B	CIVIC,Docm	Diagnostic	Positive
2	209113113	209113113	G	A	IDH1	DOID_9119-acute myeloid leu	Strong Clinical Significance-I	B	CIVIC,Docm	Diagnostic	Positive
17	7577539	7577539	G	A	TP53	DOID_1612-breast cancer	Strong Clinical Significance-I	B	CIVIC,Docm	Prognostic	Poor Outcome
12	25398284	25398284	C	A	KRAS	DOID_3908-non-small cell lur	Strong Clinical Significance-I	B	CIVIC,Docm	Prognostic	Poor Outcome
9	5073770	5073770	G	T	JAK2	DOID_4971-myelofibrosis	Strong Clinical Significance-I	A	CGI,Depo,CIV	Therapeutic	Response|Ruxolitinib
9	5073770	5073770	G	T	JAK2	DOID_2226-myeloproliferativ	Strong Clinical Significance-I	A	CGI,Depo,CIV	Therapeutic	Response|ruxolitinib
9	5073770	5073770	G	T	JAK2	DOID_8552-chronic myeloid l	Strong Clinical Significance-I	B	CGI,Depo,CIV	Diagnostic	Positive
9	5073770	5073770	G	T	JAK2	DOID_4960-bone marrow can	Strong Clinical Significance-I	B	CGI,Depo,CIV	Diagnostic	Positive

Figure 3.4: Output of the Analysis made with the Expert System

by three "sub-rules": one is related to the allele population frequency, the second to in silico prediction of damaging impact, and the last one checks if the variant is actually reported as a biomarker. The final rules could be therefore the following: (IF variant allele frequency <= 5 % in DbSNP, ExAC and Esp population databases THEN variant has low allele frequency) AND (IF PaPI, Dann and dbscSNA prediction score >= 0.8 THEN variant has damaging impact) AND (variant is reported in the knowledge base as "Therapeutic/Prognostic/Diagnostic" by FDA or professional guidelines) THEN variant is Tier I Therapeutic, Prognostic or Diagnostic. Rules could overlap since a variant could be interpreted as both Tier I Therapeutic and Tier I Prognostic, but it cannot be interpreted both as Tier I and Benign. After classification process, the ES provides as output a tab-delimited file with final classification for each input variant and a JSON file for each variant, containing information about variant annotation and classification, following the minimal variant level data (MVLD), a recently proposed framework to standardize cancer variants data for clinical utility [186]. Moreover, a tab delimited file with results is also provided (Fig. 3.4).

3.3 Case study: application on data from Myelodysplastic syndromes patients

We interpreted 884 variants found in a cohort of 310 patients with myelodysplastic syndromes (MDS). MDS are heterogeneous hematopoietic disorders whose progression could lead to Acute Myeloid Leukemia. The ES took 6.15 seconds to interpret all 884 variants. Among these, 8 variants were classified as "Strong Clinical Significance": 5 variants were reported as Diagnostic biomarkers, 5 as Prognostic and 3 as Therapeutic (3 variants are reported as both Diagnostic and Prognostic, while a variant has been observed as Therapeutic, Diagnostic and Prognostic). 27 variants were interpreted as "Potential Clinical Significance" (34 as Diagnostic, 1 as Prognostic and 11 as Therapeutic). The remaining variants are interpreted as Uncertain. The 35 classified variants occurred in 115 different patients, therefore 37 % of patients could benefit from the use of genomic information in the clinical care. For instance, a variant located in IDH2 gene (ENST00000330062:c.419G>A) is a Therapeutic biomarker of Potential Clinical significance reported both in DOCM and DEPO. Its presence is associated with response to IDH2 inhibitor molecules in patients with Acute Myeloid Leukemia [187]. Instead, the presence of mutations in ABL1 genes indicates resistance to Imatinib treatment [188]. Mutations in U2AF1 gene indicate poor outcome in patients with myelodysplastic syndromes, according to Civic and DOCM. Therefore, a mutation (NC_000021.8:g.44514777T>G) in our cohort is reported as Prognostic biomarker. We compared our classification of MDS variants with a previous study classifying mutations as oncogenic/possible oncogenic or uncertain, in 111 genes associated with MDS or closely related neoplasm [91]. We found that 225 variants in our cohort have been reported by this study as "oncogenic". Among that, we interpreted 7 as "Strong" and 34 as "Potential". Only one "Strong" variant is reported as uncertain by the previous study. Therefore, the 97 % of variants interpreted to have a clinical impact are reported as oncogenic. It is important to underline that these guidelines are not supposed to predict the pathogenicity of a variant, but they

provide a framework to evaluate the clinical impact of a variant according known studies. As an additional benchmarking towards similar tools, we run the Variant Interpretation for Cancer (VIC) on the MDS cohort. VIC implements the ASCO/ASCO/CAP guidelines to classify the clinical significance of variants annotated through ANNOVAR [113]. VIC knowledge base relies on COSMIC, CGI, PMKB and Civic. On our cohort, VIC does not find any variant with Strong Clinical Significance, but it classifies as Potential Clinical Significance 18 variants. Among these, our system detects 11 variants with Potential Clinical Significance and 7 Uncertain. However, the remaining 23 variants reported by our system as Potential biomarkers are classified by VIC as Uncertain. To conclude, our system efficiently integrates more repositories compared to a similar tool (VIC). This integration leads to a higher number of biomarkers detected. Yet, the percentage of known biomarkers in the cohort is still low (4 %), proving the need for further investigation of potential biomarkers.

Chapter 4

Genomic variant in prognostic models: Estimation of risk progression in Myelodysplastic syndromes

4.1 Risk stratification in Myelodysplastic syndromes

As we saw earlier, indicators of individual's diagnosis and prognosis can be found within the genomics profile. Variants in genes of interest can be the sign of a particular course of the disease. These prognostic biomarkers are used to decide the patient care and follow up. The stratification of patients based on prognostic scores that exploit clinical characteristics to reveal the progression of the disease is not new. Yet, the demand for the integration of genomic information in prognostic scoring system is high within the Precision Medicine context [189, 190, 191].

This section is dedicated to the modelling of Myelodysplastic syndromes (MDS) progression. MDS are heterogeneous clonal hematopoietic disorders associated with mutations and abnormalities in maturation and differentiation of hematopoietic cell lines [192]. Patients with MDS are characterized by different risks of Acute Myeloid Leukaemia (AML) development and genetics events are found to drive disease progression from MDS towards AML [193]. AML risk progression is usually evaluated in clinical practice according to the International Prognostic Scoring System Revised (IPSSR) score, suggested in 2012 by the International Working Group for Prognosis in MDS. IPSSR score combines some clinical features, such as the blast percentage value and bone marrow cytogenetics, to categorize MDS patients into one of the following groups, representing different risk progressions towards AML: "Very Low", "Low", "Intermediate", "High", "Very High" [194]. AML progression occurs subsequently into these five stages and the evolution could be naturally modelled with a continuous time Markov chain approach. A previous study investigated clinical features contribution in MDS progression and survival analysis designing a Markov model, with states corresponding to IPSSR stages, AML and death [195]. Markov models are also useful to evaluate possible outcomes of allogenic hematopoietic stem cell transplantation (HSCT), the only curative treatment for MDS, and estimate overall survival [196]. However, none of the aforementioned works explore the role of genomic events, such as the manifestation of driver mutations, in MDS progression.

Here, it is reported the analysis of clinical and genomic data from MDS patients provided by the IRCCS Fondazione Policlinico San Matteo of Pavia, a University Hospital in Italy. This investigation was approved by the Ethics Committee of the Fondazione Istituto di Ricovero e Cura a Carattere Scientifico (IRCCS) Policlinico San Matteo, Pavia (Italy) and informed consent was obtained from all subjects. All procedures were carried out in accordance with the ethical standards of the Declaration of Helsinki. Each patient belongs to one of the five risk groups (Very Low; Low; Intermediate; High; Very High) in accordance with the International Prognostic Scoring System Revised (IPSSR) [194] (see Table 4.1).

Table 4.1: Clinical manifestation of 921 MDS patients

Variables	**Total Patients**
Gender	
Male	572 (62%)
Female	349 (38%)
Median Age (year)	67 (39-91)
IPSS-R	
Very low	244 (26%)
Low	338 (37%)
Intermediate	168 (18%)
High	122 (13%)
Very High	49 (0.06%)

The clinical dataset records 627 features of 921 patients. Examples of clinical features are the peripheral blood and the bone marrow blasts percentages, hemoglobin, age, comorbidities and hematopoietic stem cells dysplasia.
NGS analysis on a panel of 44 genes associated with MDS was performed on 310 patients. About 1,144 mutations were detected, 534 of which were reported as driver mutations in a previous work [91]. The remaining passenger or uncertain variations were not included in the following analysis.
Figure 4.1a shows the percentage of driver mutations in different IPSSR stages in the 25 most mutated genes of our cohort: mutations in ASXL1, SFRS2, STAG2 and U2AF1 are detected in all five IPSSR stages, while mutations in JAK2 are reported only for Very Low and High patients. MLL, also known as KMT2A, and MPL are found in early stages of the disease.

Patients share very few driver mutations, even if driver mutations are found in the same gene. Considering the patients-mutation matrix, a value of 0 indicates that a particular mutation is not found in a given patient,

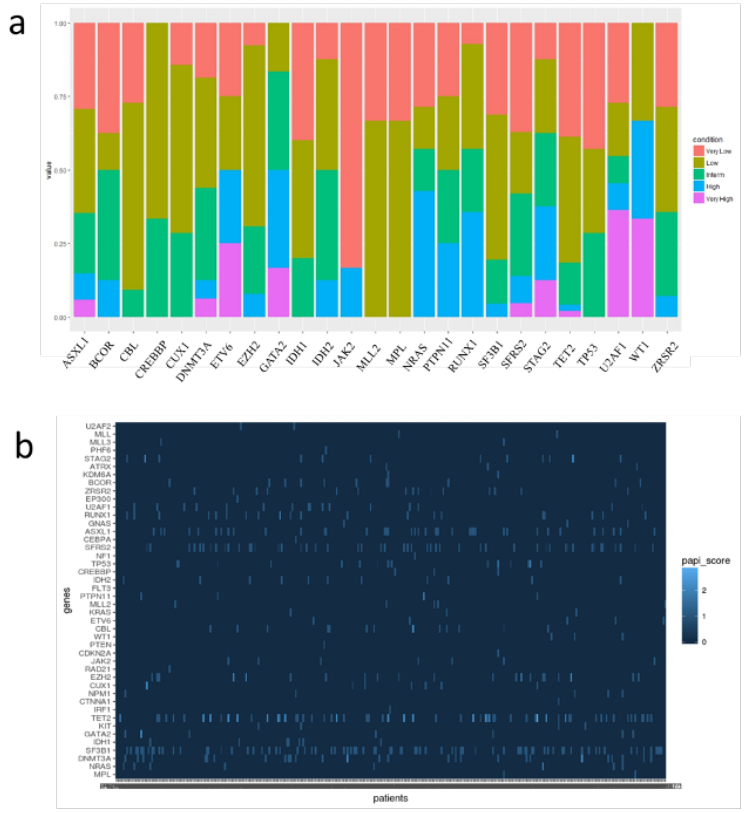


Figure 4.1: a) For each gene, the percentage of driver mutations detected in different IPSSR levels is shown. b) PaPI score calculated for each gene in each patient: a lighter cell shows that in a particular patient the gene has a higher rate of damaging mutations

while a value of 1 means that the patient harbors the mutation. To assess the sparsity of the patients-mutations matrix, we calculated a sparsity index as the number of 0 elements divided by the entire number of elements in the matrix. It was obtained a very high sparsity index (0.99867). For this reason, mutations occurring on the same gene were gathered together and 44 variables were obtained (Figure 4.1b). For a given patient, each feature value is the sum of the scores measuring the potential damage of mutations on a given gene. In order to evaluate the potential damage effect of mutations, the PaPI score was computed. PaPI is based on a machine-learning approach, and it estimates the human coding variants probability to alter their protein-related function. The higher is the PaPI score, the higher is the probability that the variant has a damaging effect [197].
Combining clinical and genomic data, a dataset of 671 features and 921 patients was obtained. The aim is to quantify the effect of mutated genes in disease progression, by using Markov and Cox modeling. In order to build a Cox regression model, the variables number was decreased, according to the heuristic rule stating that, in a multiple logistic regression model, at most one variable every 10 events should be considered [198]. In our model, the event represents the transition between subsequent IPSSR risk levels or towards death. Since we can potentially observe 921 events, we conservatively decided to include at most 88 variables.
All the 44 genomic features were retained, since they were not considered at all in a previous landmark study on IPSSR [195]. Then seventeen clinical variables have been selected on the basis of the literature confirmation of their role in the MDS progression [199, 200, 201, 202]. Eventually, the remaining 27 variables were selected among those with the highest variability, according to the coefficient of variation, calculated as the ratio between the standard deviation and the mean [203].

4.2 Stochastic Simulation of Longitudinal dataset

The huge amount of clinical and genomics data collected can be analyzed with statistical or Machine Learning approaches to extract information that may guide decisions in clinical settings, with particular focus on disease progression. In particular, when such progression can be represented as a sequence of **states** or **disease stages**, such as in MDS, it is possible to combine Cox regression and Markov models to describe patients' evolution. Cox Models identify prognostic or treatment factors that could be associated with differences in survival or progression towards different health states [204], while Markov Models are often used to describe disease evolution into a set of finite health state [205], such as different stages of cancer progression [206] or HIV infection [207]. These techniques need longitudinal data (i.e. collected over time) from a cohort of patients, who need to be carefully followed and monitored, potentially for a long-time span.
Longitudinal analysis is thus time and cost demanding, and therefore many studies only cover a small window within disease progression, giving a snapshot of patients health status that does not provide information about long-term progression [208]. To model disease progression from cross-sectional studies, Li et. al. [209] proposed an algorithm that builds trajectories through cross-sectional data starting from healthy cases to diseased, across a number of underlying stages [209]. The algorithm automatically finds different disease states along trajectories through a Hidden Markov Model, and it can be applied to different types of disorders, from breast cancer to Parkinson. A suitable adaptation of this algorithm, designed to jointly model genomics and clinical data, and its coupling with Cox regression, seems an interesting option to make use of cross-sectional data of patients in different disease stages.
In order to study MDS evolution from cross-sectional data, we have developed a method that leverages patient similarity to simulate disease progression across predefined disease stages, such as IPSSR levels in MDS. The result is a simulated longitudinal dataset that can be exploited for follow-up statistical approaches such as Cox and Markov models.
In brief, each patient in a given stage has a probability of progression to

the following stage defined by the mean survival probability in that stage. If a patient evolves, he/she "becomes" one of the patients of the subsequent stage with a probability proportional to their similarity through a Roulette Wheel algorithm. A couple of patients linked by the simulation strategy becomes a single macro-patient. This procedure is applied iteratively for the following IPSSR stages until patients in Very High stage are evaluated. Potentially, macro-patients evolving in all the five IPSSR stages can be simulated. Ten thousand Monte Carlo simulations are performed, and the longest most frequent trajectories are selected. The progression simulation algorithm is shown in Algorithm 2. The following sections detail the algorithm and the pre-processing steps needed to obtain the input variables.

4.2.1 Stage survival probability computation

In order to assign to each patient a probability of progression towards the following stage, the mean survival years for each IPSSR stage was calculated from MDS survival curves published in literature [194]). In particular, the survival function for each stage **s** can be defined as

$$S(t) = e^{-\lambda t} \tag{4.1}$$

where $h(t)$ is the function that rates the risk of death at t time. If we set $h(t)$ equal to the constant λ, $S(t)$ could be calculated as the exponential function.
The mean of an exponentially distributed random value is given by $\frac{1}{\lambda}$ while its median is $\frac{ln(2)}{\lambda}$.
Therefore, the mean survival years (m_{sy}) is computed from the corresponding median by diving the latter by the natural logarithm of 2. The survival curve value at $t = m_{sy}$ is taken as the mean survival probability in $s(P_s)$ or equivalently as the mean probability of being in stage s.

4.2.2 Patient similarity

The proposed method performs simulations on the basis of the similarity between patients. To this end, it is important to take into account the diverse sources of available information, which range from clinical to genomics data. Several approaches can be exploited to efficiently combine large sets of data from multiple sources [210].
A joint matrix tri-factorization algorithm, recently applied by our group on AML data [211], was used for patient similarity computation. Matrix tri-factorization is a knowledge-based method where relation among concepts, such as genes or patients, are organized into relational matrices. The algorithm allows to deal with data sparsity by interpolating missing data and reveals unknown interactions underlying initial data, such as patient similarities, through matrix decomposition. For our purpose, the following concepts (objects) were considered: Patients, Clinical Data, Mutations, Genes, Diagnosis. The relational built matrices were: Patients-Clinical Data, Patients-Mutations, Patients-Diagnosis, Patients-Genes, Genes-Mutations. For instance, in Patients-Mutations matrix, a value of 1 in (i, j) position denotes that patient i harbors a particular mutation j. The Matrix Tri-factorization algorithm (implemented in Matlab) was run with ranks empirically set at 200 for each concept. Ranks are crucial parameters, since they define the dimension of latent factors revealing hidden structure in the data, but there is no consensus about the selection method on these variables [211]. For a detailed explanation of tri-factorization technique see Vitali et al, [211] and Appendix B. The output of the algorithm is the consensus matrix M, where the position (i, j) represents the measure of similarity between patient i and patient j. The transition probability is then calculated as the ration between the similarity between the patients i and j and the sum of similarities between patient i and all the other n patients in the cohort:

$$P_t(i, j) = \frac{M(i, j)}{\sum_{z=0}^{n} M(i, z)} \tag{4.2}$$

Matrix trifactorization allows us to compute patient similarity despite the sparsity of our datasets, in particular in the genomic one, where only

310 of 921 patients are reported and for which few shared mutations are found.

4.2.3 Progression algorithm

The method developed to simulate longitudinal data from cross-sectional MDS patient data is explained in detail in the following (see also Algorithm 2). The procedure is applied independently for male and female patients. For each disease stage s (in our case defined by the IPSSR levels), we select patients in stage s and in $s+1$. For each patient p_i in stage s, the algorithm decides if a patient evolves by randomly selecting a number r between 0 and 1. If r is greater than the mean survival probability m_s, then p_i evolves in $s+1$. The patient p_j in stage $s+1$, in which p_i evolves, is obtained by sampling from the probability distribution $P(i,1) = \{P(i,1), ..., P(i,n)\}$, with a Roulette Wheel algorithm [212]. The macro-patient p_{ij} is therefore

created.

Algorithm 2: Progression simulations in different stages

linenosize= **Result:** Simulated Longitudinal Dataset L
Input: Male/Female cross sectional dataset D;
Mean survival probability for each disease stage m_s;
Transition probability matrix P_t; Disease stages list S;
initialization: z=0, N=10.000;
while $z < N$ **do**
  initialize longitudinal dataset L_z as empty ;
  for (*s in S-1*) {
    Select mean survival probability m_s;
    for (p_i *in s*) {
      if p_i *dies in s* **then**
        Add p_i to L_z;
        continue
      else
        Randomly select r $\epsilon[0, 1]$;
        if $r > r$ **then**
          Associate each patient p_j in *s+1*
          with an area proportional to $P_t(i, j)$;
          Sample a random number R;
          Select the p_j patient
          in s+1 whose area contains R;
          Create the macro-patient p_{ij}
          that evolves from stage s to s+1 ;
          Add two rows in L_z:
          one with data from patient p_i in s stage,
          and one from patient p_j in stage s+1
        else
          Add p_i with his/her variables to L_z;
        end
      end
    }
    for (p_j *in s+1*) {
      if p_j *is in more then one macro-patient* **then**
        Select the p_i patient with the
        maximum transition probability to p_j
      end
    }
  }
end
Select the longest most frequent trajectories in L_z
with z from 1 to 10000 to populate L

The whole process is repeated in a 10,000 Monte Carlo simulation and for each patient the longest trajectory (the one that spans more IPSSR

stages) is selected. If more trajectories have the same length, the most frequent one is selected. The number of Monte Carlo simulations has been set in order to ensure that the simulation converges to the final absorbing state.
Once the longitudinal trajectories are computed, some adjustments need to be performed. For instance, it can happen that a patient p_i in stage s is associated with the patient p_j in stage $s+1$ which was actually diagnosed and followed-up before in time. Therefore, it necessary to adjust diagnosis and follow-up time, required for Markov and Cox modeling.
To get realistic trajectories, we sample a time T from the survival curve in s and we add it to the diagnosis date in s: we set the result as the new diagnosis date in $s+1$. If the new diagnosis date is later than the last follow up date in s (FU_s), we set the new diagnosis date equal to (FU_s). Moreover, we modify the last follow up date in $s+1$ $((FU_s)+1)$ by adding to it a number of days equal to the initial difference between $(FU_s)+1$ and

its corresponding diagnosis date (see Algorithm 3).

Algorithm 3: Adjusting diagnosis dates

Result: Simulated Longitudinal Dataset L with correct dates
Input: Male/Female simulated dataset L;
Survival probability curves for each disease stage P_s;
Disease stages list S;
for (*s in S-1*) {
  Select from L rows corresponding to
  macro-patients in stage s and *s+1* (L_s);
  for (p_i *in* L_s) {
    Compute the difference in days d between
    the p_i diagnosis date in stage s and the p_i last follow up
     date in s+1;
    Sample a time T from the survival curve;
    Sample a random number N between 0 and 1;
  }
  while $N >= P_s(T)$ **do**
    Sample another T;
    $date_{diagnosis} = T * 365 days$
  end
  Set $date_{diagnosis}$ as the new diagnosis date for p_i in *s+1*
}

The resulting simulated longitudinal dataset consists of 671 trajectories among different IPSS-R levels of risk: 262 of these involve women, while the remaining 409 are men. 163 men and 87 women do not get worse and remain in their initial IPSS-R levels of risk.

Table 4.2 shows the simulated evolution of the macro-patient 791_248_1391_252 from Low stage (patient 791) to Very High (patient 252).

The proposed algorithm is a probabilistic framework that relies on similarity to infer patient trajectories among known clinical disease stages. Our approach could allow the exploration of temporal disease progression through the simulation of a longitudinal dataset from a cohort of cross-sectional data. This problem is not new in the biomedical fields, where

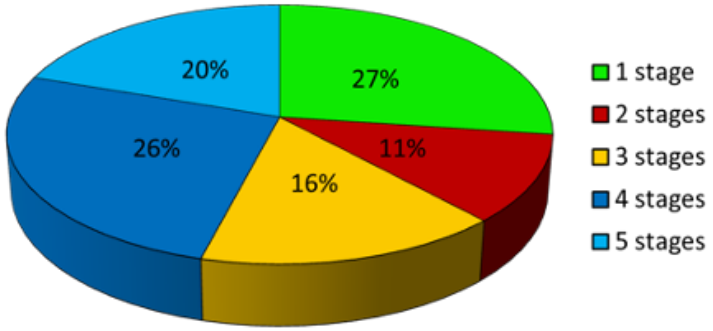


Figure 4.2: Percentages of trajectories of different lengths. 27% of patients remains in their initial stage ("1 stage"), 11% of patients evolves in two subsequent IPSSR stages ("2 stages", for instance a patient evolves from Intermediate to High), 16% of patients evolves across three stages, 26% across four stages and 20% of patients evolves from Very Low to Very High.

Table 4.2: Example of simulated longitudinal trajectories. The macro-patient 791_248_1391_252 is the result of the 791 patient's evolution from Low to Very High IPSS-R stage. He dies in the last stage

ID MACRO_PZ	**ID PZ**	**IPSS-R**	**Diagnosis Date**	F.U. Date	Death	Time (Years)	Status	Other variables
791_248_1391_252	791	Low	24/07/2003	01/05/2005	No	0.0003	2	...
791_248_1391_252	248	Intermediate	01/05/2005	01/05/2005	No	1.77	3	...
791_248_1391_252	1391	High	20/03/2007	09/05/2007	No	3.66	4	...
791_248_1391_252	252	Very High	19/08/2007	03/06/2008	Yes	4.07	5	...
791_248_1391_252	252	Very High	19/08/2007	03/06/2008	Yes	4.87	6	...

clinical trials are usually conducted within a defined time interval that covers just a small window within the disease process, which actually could span over a large period of time. These changes in patient's health status are reflected in clinical and genomics data. Li et al. identifies intermediate stages that lead from healthy status to the disease state in a cross-sectional cohort, by building pseudo time-series based upon Euclidean distances and temporal bootstrap. Hidden Markov Model trained on the pseudo-time series revealed transition tables between inferred disease stages, characterized by different symptoms in Glaucoma, Breast Cancer and Parkinson's disease [209].
Despite this procedure is able to find "hidden" disease states that could have a prognostic meaning, data from healthy patients are needed. In a Precision Medicine initiative, where also genomics data are reported, it could be cost demanding to generate such data for a healthy cohort. Moreover, for some diseases, such as myelodysplastic syndromes, clinicians already embed clinical information in prognostic scoring systems that may reflect the temporal nature of the disease. For instance, MDS patients are stratified into 5 different stages, representing the disease course towards Acute Myeloid Leukemia.
Our approach differs from previous works since it is suitable for simulating disease trajectories that span through diseased patients, when known clinical stages are defined. Moreover, in our approach patient similarity is computed through a data fusion approach that combines large amount of clinical and genomics data. However, the simulated trajectories were not validated, since true longitudinal data were missing. This limitation of our work highlights the need for longitudinal studies to be made available.

4.3 Risk progression estimation based on genomic profiles by Cox and Markov Model

The final aim of this procedure is to aid clinical decision making, suggesting prognostic factors that could influence disease progression. Statis-

tical methodologies could reveal such factors from longitudinal data. For instance, through the application of Markov modeling integrated with Cox Regression, we can find the mutated genes which are statistically significant for the disease progression, but also the probabilities of disease evolution through the different IPSSR stages. Moreover, coupling Cox regression coefficients with Markov transition intensities, we can model the effect of patient covariates on the progression probabilities. In fact, such transition intensities are computed with covariates coefficients estimated by the Cox Regression.
A Cox model is a multivariate regression method that allows exploring the relationship between the occurrence of an event and several explanatory variables. Typically, this type of model is adopted in survival analysis, in which the event to analyze is death [213].
In this work, two different Cox models were implemented to select the most significant features in MDS progression throughout different IPSSR levels and death, respectively. Death is considered as a distinct event compared to the worsening of the patient's clinical status. Both models were developed using the function *My.stepwise.coxph* included in the *My.stepwise* R package. This function allows to obtain the best candidate final proportional hazards model applying a stepwise variable selection procedure [214].
On the other hand, Markov models are multistate models describing a process that evolves in a probabilistic way and in which it is assumed that the next step depends only on the present state and not on the past events. Figure 4.3 represents the Markov model adopted to describe MDS evolution across IPSSR levels. Only transitions towards subsequent stages or death are allowed.

To develop a continuous-time Markov model, we used the *msm* R package applied to simulated longitudinal dataset that collects functions to computes the Markov model transition probabilities that best fits the data through maximum likelihood estimation. Cox and Markov results are listed in the Appendix B, while here a general discussion of results is provided.
Markov models describe in each time a process evolution, through transitions intensities and probabilities matrices ($\mathbf{Q}$ and $\mathbf{P}$, see Appendix B).

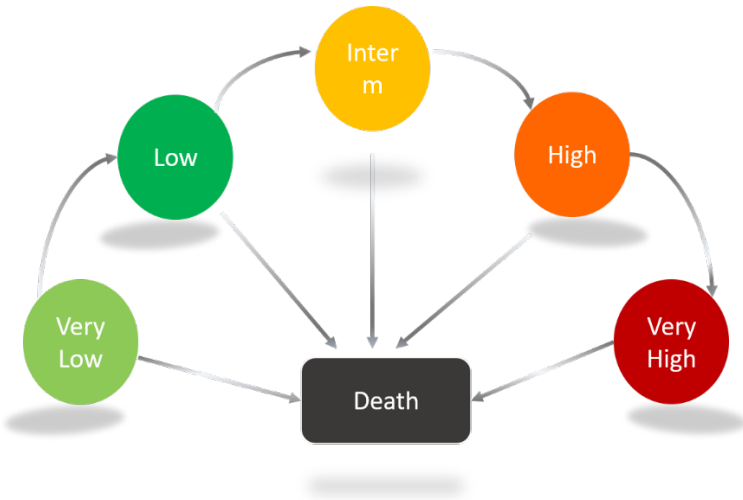


Figure 4.3: Markov Model for the description of MDS progression across IPSSR stages

By estimating transition intensities with the Cox Model, we can take into account significant covariates, i.e. genomic and clinical features, in the assessment of progression probability. Therefore, these quantities represent the probability of transition between stages given the presence of the significant covariates, for instance a mutated gene in the patient's genome. Transitions probabilities at 10 years after diagnosis are illustrated in Figure 4.4, that shows also some of the features that, according to our Cox models and literature, have an impact on progression towards higher stages or death. Covariates are coloured in red when they are associated with poor prognosis, while they are green when they have a protective role. Moreover, *msm* function allowed to estimate survival curves for patients in different IPSSR levels and the corresponding survival time medians. We compared both with the results of a recent similar study that analyses MDS progression without considering the genetic contribution [195] (Figure 4.5).

Both the curves for Very Low IPSSR level show a plateau for about the

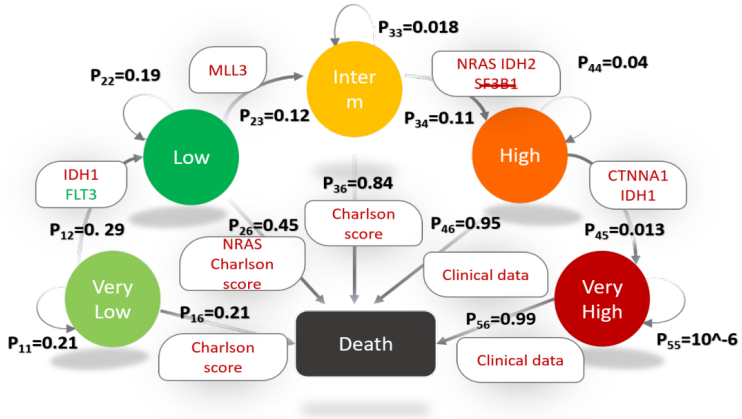


Figure 4.4: Resulting Markov Model. Transition probabilities at 10 years after diagnosis are shown. Some of the covariates selected by Cox models and confirmed by literature and experts as prognostic factors in transitions are colored in red when associated with poor prognosis and in green when protective

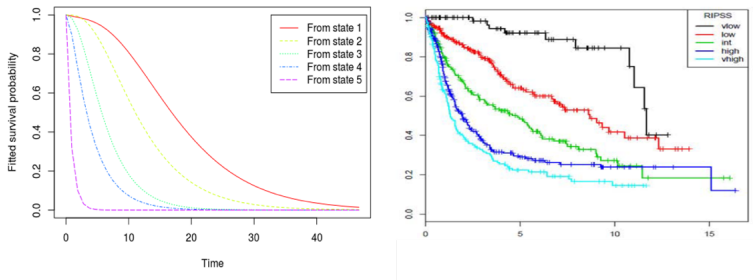


Figure 4.5: Survival curves computed with simulated longitudinal dataset and survival curves from Della Porta et. al [195]

Table 4.3: Survival time median (in years) in different IPSSR stages comparison between the Markov model (M) generated from the simulated dataset and the one implemented from Della Porta et al. [195]

	Very Low	**Low**	**Intermediate**	**High**	**Very High**
Survival median predicted from M	15.88	11.21	5.6	2.8	0.93
Survival median from Della Porta et al.	12	9	5	2	1.25

first ten years from diagnosis. However, according to our study, the survival probability for patients in Intermediate and High IPSSR levels tends to zero in about ten years. This temporal range decreases to a couple of years for Very High stage. These probabilities reach roughly a 0.2 value in the study with only clinical features. Taking into account genomic features increases the slope of survival curves. Table 4.3 below shows the survival time medians of both studies

Since true longitudinal data are not available for this cohort, we could not validate the simulated trajectories. At this step, our final goal is to assess whether significant insights into the molecular aspects of the disease could be obtained also from simulated data. In order to qualitatively evaluate our results, we performed an extensive literature search on myelodysplastic syndromes. We tried to find whether the significant covariates resulting from the Cox model (built on the simulated data) are confirmed or not by literature studies. To our knowledge, the influence of mutated genes in the transition between two known IPSSR stages (for instance from "Very Low" risk to "Low") has not been assessed, while many studies investigated the association between mutations and the presence of a more aggressive disease course, with the probability of leukemia evolution increased. Reminding that IPSSR states are actually different stages of the disease that becomes more aggressive during its evolution towards leukemia, we made a comparison between our results and the aforementioned studies.
According to Cox modelling, mutations in KDM6A, IDH1 and NPM1 in-

crease the probability of progression from the initial stage to the "Low" risk stages, while patients with FLT3 gene mutated are less likely to progress (Table B.1). KDM6A, also known as UTX, is a histone demethylase whose inactivation likely contributes to transcriptional repression or activation of distinct genes. Mutations in KDM6A were found in MDS patients, with an inactivating effect on tumor suppressor genes [215]. Even if the gene could be targeted by small inhibiting molecules, thus representing a therapeutic biomarker, its prognostic role in MDS is still unclear [216]. On the contrary, their negative impact on prognosis of patients with myeloma has been assessed, where mutations and deletions on KDM6A gene are associated with shorten overall survival [217].
IDH1 gene encodes the isocitrate dehydrogenase 1, an enzyme involved in the production of a molecule (NADPH) necessary for many cellular processes. IDH1 is reported to be significant also in the transition between "High" and "Very High" risk (see Table B.4). Also, IDH2, another gene involved in the same pathway, is associated with transition from "Intermediate" to "High" (Table B.3). Mutations in these genes are reported in different types of cancers, and some studies suggest that IDH1 mutations represent an inferior prognostic indicator, thus confirming our result [218, 219]. IDH2 mutated patients showed a significantly higher risk of developing AML and are associated with significantly worse of the overall survival [220]. Moreover, IDH mutations are found to be among the most common mutations related to AML [221], they are enriched in high-risk vs low-risk MDS and might drive the progression to high-risk MDS [222], thus confirming our result.
NPM1 gene's product is involved in ribosome biogenesis and it is upregulated in different types of cancer, while FLT3 gene stimulates signaling pathways that control important cellular processes such as proliferation and survival , in particular during hematopoiesis [223].
NPM1 mutations are rare in MDS, but they are common in AML [222]. Although the available evidence is scanty, MDS patients carrying NPM1 mutations show an unfavorable clinical course, consistent with the effect observed in our study. Moreover, MDS patients with NPM1 and FLT3 mutations share cytogenetic and mutational profiles similar to those in AML,

suggesting the evaluation of these mutations as prognostic biomarkers. Selected FLT3 mutations, namely internal tandem duplication and tyrosine kinase domain mutations, are frequent in AML and associated with poor prognosis. Mutations involving this gene in MDS are rare and their value remain to be clarified.
In the intermediate transition, only MLL3 is reported to provide a higher probability of progression (see Table B.2). MLL3, also known as KMT2C, is a member of the myeloid/lymphoid or mixed lineage leukemia (MLL) gene family and it is involved in transcriptional activation. Deletions of chromosome 7 and on the long arm of the chromosome 7 (7q) [7/del(7q)] are reported in MDS and AML and they are associated with poor diagnosis [224, 225]. Mutations in MLL3 are frequent and have a negative impact on progression-free survival also in breast cancer [226], but are rare in MDS and their prognostic impact is still not clear [227].
Several genes (Table B.3) seem to drive progression from "Intermediate" to "High" risk. CBL is a proto-oncogene whose protein product regulates transduction of signalling pathways. Mutations in CBL mainly occur as late events in MDS [192]. Patients with overlapping syndromes between MDS and myeloproliferative neoplasms (MPN) harbouring pathogenic CBL mutations are associated with poor prognosis [228], and they have been observed during AML transformation [229]. Interestingly, it has been recently suggested to include CBL mutations, along with IDH2, DNMT3A, ASXL1 and TP53 mutations within the International Prognostic Scoring System to improve MDS risk stratification, since these genes are independent biomarkers of shorter survival [230]. SF3B1 gene is involved in RNA splicing and it is recurrently mutated in MDS. This mutated gene is associated with a favourable prognostic value in low-risk disease, which is not retained in advanced stages, likely as a consequence of additional subclonal genetic lesions. Notably, these include gene rearrangements that are not captured by DNA sequencing-based platforms, as the one adopted in this study [231].TET2 gene encodes for a protein involved in myelopoiesis, therefore defects in this gene are associated with several myeloid disorders. In MDS, its prognostic role is still unclear, with some studies stating TET2 as a favourable prognostic factor [232], other works reporting

instead a poor overall survival [233]. CREB binding protein (CREBBP) interacts with DNA damage response and DNA repair proteins enhancing their functions. Moreover, CREBBP modulates the activity of poly(ADP-ribose) polymerase-1 (PARP1), a factor involved in transcriptional regulation. CREBBP is a known target of translocations events in acute leukaemia and so our Cox model rightly assigned a positive regression coefficient to driver mutations in this gene for Intermediate-High transitions [234].The NRAS gene codifies a protein called N-Ras that is primarily involved in regulating cell division. NRAS mutation are enriched in AML [222] and are associated with poor prognosis, particularly in lower MDS risk levels: mutations in this gene predict a shorter overall survival as confirmed by our Cox model. Moreover, the prognostic significance of NRAS mutations is independent from other risk factors, such as sex, age and mutations in 16 other genes.
MDS genomes show different global DNA methylation patterns compared to normal bone marrow cells suggesting that there may be methylation-specific gene alterations that contribute to these diseases. DNMT3A is a DNA methyltransferase enzyme that mediates methylation of CpG dinucleotides, and it is often involved in cancer onset. DNMT3A mutations are the most common drivers of pre-malignant clonal expansions referred to as age-related clonal hematopoiesis or clonal hematopoiesis of indeterminate potential [235, 236]. Progression into myeloid neoplasms is driven by additional subclonal mutations, which are usually the major determinant of disease phenotype and course, while the independent prognostic value of DNMT3A mutations remain to be clarified.
Eventually, 3 genes are reported to be significant in progression from "High" to "Very High" risk (Table B.4) CDKN2A gene encodes proteins that regulate two critical cell cycle regulatory pathways, the p53 pathway and the RB1 pathway. In MDS patients, an aberrant methylation is found during disease progression and it is associated with leukemic transformation and tends to shorten the overall survival [237]. Cox model wrongly computed a negative regression coefficient for driver mutations in this gene.
CTNNA1 gene is a tumour suppressor for its roles in inhibiting proliferation and promoting apoptosis. It is associated with MDS and its deletion

is likely to contribute to a poor diagnosis(Hemmat et al., 2014). Moreover, acquired epigenetic inactivation of CTNNA1 is associated with higher IPSSR risk in MDS and it is a component of leukaemia progression in patients with myeloid malignancies [238, 239]. Deletion in CTNNA1 seems to provide a growth advantage towards MDS and AML [240]. These evidences are confirmed by the very high Cox regression coefficient given to mutations in this gene.
Regarding transitions towards death, during low stages of the disease, Charlson score values mainly determine the likelihood of the transition. Charlson comorbidity index estimates relative risk of death for each comorbidity [241]. Therefore, in lower stages, our result suggests that death is mainly caused by other comorbidities and not by MDS. In higher stages, clinical features such as promonocytes percentage or karyotype are associated with poor prognosis. Notably, recent studies showed that founding genetic lesions driving pre-malignant or early malignant clonal expansions, are also associated with an increased risk of death not explained by the risk of developing a myeloid neoplasm, and mostly associated with increased cardiovascular morbidity and mortality, as well as with inflammation-driven diseases [242, 243] This evidence is supporting the notion of a direct connection between MDS and extra-hematologic comorbidities, as captured by our model. However, the contribution of specific gene mutations in this process remain to be clarified and deserve addition biological and clinical studies.

Chapter 5

Conclusions

Genomic screening programs can identify individual at risks, and therefore facilitate risk management and early diagnosis [244] in a Precision Medicine setting. Cancer, and in particular heterogenous disorders such as hematologic diseases, represents a suitable area of application of PM strategies. The assumption is that the genomic information, which can be now extracted from large cohorts of patients, will elucidate previously unseen mechanism of the diseases, that were obscure so far, and that may explain heterogenicity in treatment response and disease progression among different individuals. Relying on this new type of information, patient care can be conducted in a more precise and personalized way. For instance, genomic screening and tailored treatments, along with early diagnosis, have shown to increase survival probability in breast cancer [245]. However, another trial has shown no improvement in progression-free survival when molecular target agents are used [246]. We are facing an understandable hype in genomics-based medicine, due to the increasing number of data from sequencing projects and findings in this field. Yet, functionally-validated data to detect new actionable variants still need to be collected [247].
In this context, bioinformatics and data analysis play an essential role to mine the huge amount of sequencing data and to turn that data into vi-

able information and knowledge supporting clinical decision making [15, 246, 248]. This thesis focuses on computational tools for the interpretation of somatic variant and for the integration of mutational and clinical information for disease monitoring. The proposed approaches include Artificial Intelligence and Machine Learning methodologies and address different issues that hamper the effective application of such tools in daily clinical care, as well as the extraction of possible information also from uncomplete data.
The methodologies developed in this thesis examine the role of mutations in cancer development and progression, by using Artificial Intelligence and statistical methodologies. The design of the AI-based tools will allow future updates as new knowledge will be acquired, the identification of prediction failures and the interpretability of guidelines-based interpretation. Simulation of longitudinal data, starting from patient similarity computed with data fusion approaches, can be used to locate a patient into a pathway of disease progression for prognostic purpose. Overall, these tools can support the implementation of Precision Medicine in the hematological context, by suggesting possible oncogenic mutations, matching treatment and prognostic indicators.
The following paragraphs discuss results and outline possible limitations, as well as future directions, for each methodologies presented in this thesis.

5.1 Somatic Variant Pathogenicity Prediction

The distinction between passenger and driver mutations in tumor samples is a critical step to understand patients individual cancerogenesis. Thanks to the great availability of public cancer data, bioinformaticians have often trained ML models to classify somatic variants into driver or passenger. However, many known variants could not be used for training since their ground truth labels are missing. Semi-supervised learning approaches could allow to fully benefit from genomic sequencing data reported by public resources. Moreover, as our evidence about oncogenic variants will increase, ML algorithms for driver and passenger classification

should allow the incorporation of new knowledge. In this thesis, an incremental semi-supervised approach has been proposed. Unlabeled data and meta-features representation have been conveyed to improve performance of standard supervised methods. Unlabeled data from large sequencing studies are effectively assimilated to learn an informative representation from known genomic-annotation features of somatic variants by autoencoder networks. Labeled data, collected from different public resources, represents a pan-cancer dataset of point mutations and short indels, representative of the plethora of possible somatic mutations (coding, non-coding, missense, frameshift). Labeled data are exploited to train a Random Forest algorithm, that can allow incremental training when new examples will be available. A case study, with driver variants detected in Myelodysplastic syndromes patients, is performed to show incremental learning with different parameters, that can influence the learning rate of the algorithm on the new data. The incremental training with peculiar variants from a cancer type can allow the development of cancer-specific models, starting from the available pan-cancer background.

5.1.1 Feature Transformation

Several Deep Learning and Neural Networks architectures for data transformation have been tested. RFs trained on such transformations are compared with RF trained on "raw" annotation data and linearly transformed (PCA) components. Performance are compared in terms of several metrics. The most performing models are the RFs trained with meta-features extracted from AEs. The number of meta-features that seems to be more informative for the RFs is actually high (from 70 to 100). This result is confirming results from Wang et al., that showed increasing accuracy as the dimension of the data transformation increases [249]. Nevertheless, the usage of highly non-linear transformation of data hampers the explainability of the model. Methodologies to interpret AE meta-features in the light of variant annotation features are required and represent an interesting future direction of this work.

5.1.2 Machine Learning model

The best final Random Forest model for deployment is selected according to precision: in fact, given that NGS sequencing could detect hundreds of variants, it is imperative to have high precision (i.e. a lower number of False Positive while having a high number of True Positive), to suggest to clinicians a concise number of variants to study. It is worth to note, however, that in our approach the threshold for driver classification has been moved from 0,5 to have a good balance of precision and recall. Therefore, even if the final model is selected according to precision, recall (or specificity) will not decrease to undesired levels, as also the comparison with state-of-art tools showed. The division of training and test variants considers circularity issue, i.e. the possible information leakage due to gene-based information shared by multiple variants [74]. To do so, genes whose variants are selected for training will not be represented by any variant in the test set. The training and test set division also reflects the proportion of driver and passenger variants in the entire dataset, which ultimately reflects the actual proportion in many tumor samples. By comparing the results of a 10-fold cross validation on the training set and on the Test Set, it is shown that circularity is a serious issue that be must carefully addressed.
The Random Forest trained on Autoencoder meta-features is compared with other tools commonly used for driver and passenger classification. A further important step will be the comparison with the Standard Operating Procedure (SOP) very recently proposed by the Variant Interpretation for Cancer Consortium (VICC) Knowledge Curation and Interpretation Standards (KCIS) working group. These guidelines encode genomics information into criteria and rules similar to the ACMG/AMP germline pathogenicity assessment, to determine the oncogenicity of somatic variants [66]. Given their interpretability and the parallelisms with widely adopted germline guidelines, SOP aims at becoming the standard for variant oncogenicity prediction. A comparison with such guidelines may be important to quantify whether a completely data-driven approach, like the one proposed in this thesis, is able to capture nuances of variant interpretation not reported into the guidelines.

To identify possible failures in the prediction, a new reliability estimation approach is proposed. Reliability, or trustfulness, of ML prediction is an important topic to implement ML models in the medical context. Reliability assessment is essential to identify failures, possible bias and unfairness in the data, which are among the reasons that prevent the actual spread of AI-based tools in the clinical practice [16]. The methodology compares new variants to be classified with the most informative training examples according to the POP instance selection algorithm [178]. The reliability reflects to which degree the new variants is "distant" to the informative examples and detects a subset of unreliable variants for which the classification made by the Random Forest may be wrong at a higher rate.
To ease the usage of the developed framework, a Graphical User Interface (GUI) in Python is implemented. The GUI allows for (1) variant annotation based on a popular annotation tool (Ensembl-VEP) (2) variant interpretation in terms of pathogenicity (driver or passenger status) and (3) reliability assessment for prediction. Source code is available on github.
The proposed model for somatic variant interpretation is based on the assumption that cancer is driven by only a subset of variants (*driver*), while all the remaining (*passenger*) are considered benign and therefore not implicated in cancer development and progression [27, 33]. Based on this assumption, tools for somatic variant classification, including the one suggested in this thesis, provide a binary distinction between driver and passenger [135, 133, 132, 250]. Moreover, each variant is classified independently from each other. Yet, works are highlighting that cancer development and progression is a more multifaceted process, where also passenger mutations can have an active role by working in synergy with other mutations. This type of passenger mutations is called "latent driver". Latent drivers are inert, but when cooperating, they cause allosteric events that can turn a cell from normal to cancer, or that can push cancer to metastasize [251]. Mathematical modeling on COSMIC data confirms the suggested impact of passenger mutations on cancer progression [252], yet functional studies are still lacking. Following a "binary" distinction between passenger and driver variants, current tools are failing in the identification of cooperativity.

Incremental Model

In this thesis, incremental learning is obtained by adding new trees to the initial Random Forest, which is considered the "reference" model. This model, trained with pan-cancer data, shows balanced accuracy around 77 % on a test set, outperforming previously developed tools.
Ensemble strategies, including Random Forest, combine a collection of different models and for this reason they are seen as suitable methods to represent and mediate different information in parallel [253]. For instance, Lean++ incrementally adds weak neural network classifiers as new data for training are available. The inclusion of weak classifiers to the initial pool in an incremental manner efficiently deals with concept drifts that occur due to unforeseen changes in data distribution [254, 255]. In [254], Hoeffding trees represent the pool of weak classifiers. The voting weight of each tree is related to its Mean Squared Error. When new labeled data are available, the incremental procedure is carried out both by updating the voting weights of "old" trees based on new data misclassification, and by adding a new single tree trained on the new data. This last tree is seen as the "perfect" classifier since it is trained on the most recent data, and it will have a higher weight in the final classification. Yet, given this assumption, a tree should be added only when the collected chunk of data is representative and sufficient for training. To avoid a potential infinite increase in the number of weak classifiers, a maximum number of trees k is decided as model parameter. When the number of trees reaches k, and when a new trained tree is ready to be included in the initial pool, the least accurate classifier is removed. Incremental algorithms based on Random Forest are often devoted to the incorporation of new class [256] and they update a trained RF by increasing the number of splits at the leaf nodes [257, 258] rather than adding new trees. Instead, in this thesis new trees are added to the initial pool of the Random Forest, in a workflow more similar to [254, 255]. The new trees are trained on the new set of labelled data, while the initial pool is not updated. Once the new trees are added, a new unseen variant will be classified according to the majority class of the entire updated Random Forest. The number of trees to be added is an important parameter

of the algorithm that must assure a proper weighting of the information encoded in the different chunks of data. A possible approach could be the addition of a number of trees such that the proportion between the size of data and the number of trees is the same as in the initial Random Forest. Also, the maximum number of trees should be set as model parameter, and therefore a strategy for including and removing trees needs to be chosen. The predictive ability of a Random Forest is associated with the strength and diversity of its individual trees. The main issue is assessing whether each single tree, especially if trained on few data, has a sufficient prediction ability. Out-of-bag training data could be used to select the subset representative trees, as proposed by [259]. Yet, the decision of incrementally learning new trees (or other weak classifiers) should not be taken only on the basis of the availability of new labelled data, but it should be related with model diagnostics and monitoring over time. Especially in the clinical environment, it is essential to assess whether the performance of a classifier is deteriorating [260]. Model maintenance strategies can include the periodically re-training of the algorithm on new data, although this approach can neglect information learned from past associations, and can lead to model that lack generalizability [260, 261]. The incremental learning of Ensemble methods should be performed within model monitoring framework, to prevent possible deterioration and distortion of the classification.

5.2 Somatic Variant Interpretation according to standard guidelines

In a clinical setting, variant pathogenicity assessment should not be the only source of genomic information. Rather, standard guidelines for interpretation of the clinical significance should be used to find known biomarkers that can effectively drive patient care towards Precision Medicine. The actual implementation of interpretation guidelines in clinical practice calls for tools able to reason over a heterogeneous and always growing knowledge base. We collected and standardized data from 6 databases over 1500 mu-

tations known to have a clinical impact in cancer. Within the annotation process, we associated each variant with further information. Developed ETL pipelines will allow future update of the knowledge base. We then implemented an Expert System that reasons over the collected standardized knowledge base and automatically interprets somatic variant according to standard AMP/ASCO/CAP guidelines. ES architecture will allow future updates of the Rules, avoiding complex alteration of the application code. The ES receives as input a list of genomic variants, performs inference, and then provides as output a JSON file for each variant, reporting variant annotation and AMP/ASCO/CAP interpretation, according to MVLD. Thanks to output files, the ES allows user to follow the reasoning process that lead to the final classification. We interpreted more than 800 variants in patients with myelodysplastic syndromes, suggesting that almost half of the cohort carried variants of strong or potential clinical significance. This information could therefore help clinicians in clinical decision-making process. Future improvements will be the possibility to interpret also complex alteration and the development of a web tool where user could query the ES interpretations. Moreover, other databases could be included in the knowledge base.

The two different but complementary variant classification tools that have been developed spans through the variant interpretation process (Figure 1.2), by addressing (1) the pathogenicity classification of somatic variants and (2) the clinical significance interpretation of somatic variants. While the first approach may suggest putative oncogenic variants to be functionally validated to detect new biomarkers, the second guidelines-based approach translates current knowledge about biomarkers and actionable variants into insights that can guide decision making for a given patient. Both tools rely on data collection and information encoded in available omics resources. Yet, such databases may not be frequently updated. Versioning of the knowledge base will be essential to capture all the incoming information and to keep track of the variant interpretation path. An important aspect that should be addressed is the incorporation of knowledge not only from structured databases, but also from published works. In fact, findings from sequencing studies are usually reported as text in scientific papers, before

they included in databases and repositories. Text mining and Natural Language Processing approaches can support the automatic extraction of such knowledge from high number of papers published each year [262]. Another aspect of somatic variant interpretation that should be further analyzed is the impact of structural variations. The tools developed in this thesis aim at classifying somatic point mutations (SNVs) and short insertions and deletions (INDELs). Despite SNVs and INDELs characterize the majority of actionable variants in cancer, more complex structural variations, such as gene fusion, play an important role in different cancer types, and their interpretation should be addressed as well [263]. Moreover, to further elucidate the functional role of variations in cancer development, the integration of different omics data, such as transcriptomics and proteomics, along with the genomics, can be studied [21]. The analysis of sequencing data from Single-Cell technologies can inspect tumor samples at an unprecedented level of detail. Given the heterogenicity of cancer, a more effective Precision Medicine strategy may require insights from Single-Cell sequencing.

5.3 Risk Progression Prediction in Myelodysplastic syndromes

The last part of the thesis focuses on how knowledge from interpreted variants can be integrated with patient clinical characteristics to predict disease progression. Statistical approaches can reveal associations between particular mutations and disease progression, as long as that longitudinal data from a cohort of patients are provided. However, longitudinal data collection is time and cost demanding. A possible solution is the development of new approaches for the simulation of repeated observations in time. It is whether statistical analysis performed on simulated longitudinal dataset could support clinical decision making by providing meaningful insights in the progression of myeloid neoplasms. It is proposed a method for simulating trajectories of disease progression across known stages from cross-sectional data, taking into account patient similarity. The

simulated longitudinal dataset allows to apply well-known statistical approaches, i.e. Cox and Markov models, to suggest covariates that could play a role in disease progression and to assess survival probabilities in different stages. This strategy is applied to a cohort of patients with myelodysplastic syndromes, hematopoietic diseases that can evolve into Acute Myeloid Leukemia through different known prognostic stages. This framework could therefore be applied to any type of disease where different stages can be observed. In this work, the focus is in the contribution of mutated genes in MDS evolution. In most cases, the results are consistent with literature findings, and Cox analysis on simulated longitudinal dataset reveals that in the MDS cohort the KDM6A gene is a prognostic biomarker. Alteration in this gene have been associated with lower overall survival in different type of cancers, suggesting that this type of analysis can propose new potential biomarkers to be experimentally evaluated. It must be noted, however, that a systematic validation of the simulation algorithm with a real longitudinal dataset could not be provided. This limitation highlights the need for longitudinal studies to be made available to the research community. Despite these constraints, it is was possible to assess whether statistical approaches applied on simulated datasets can provide useful information about the prognostic impact of molecular alteration.

Appendix A

Autoencoder results

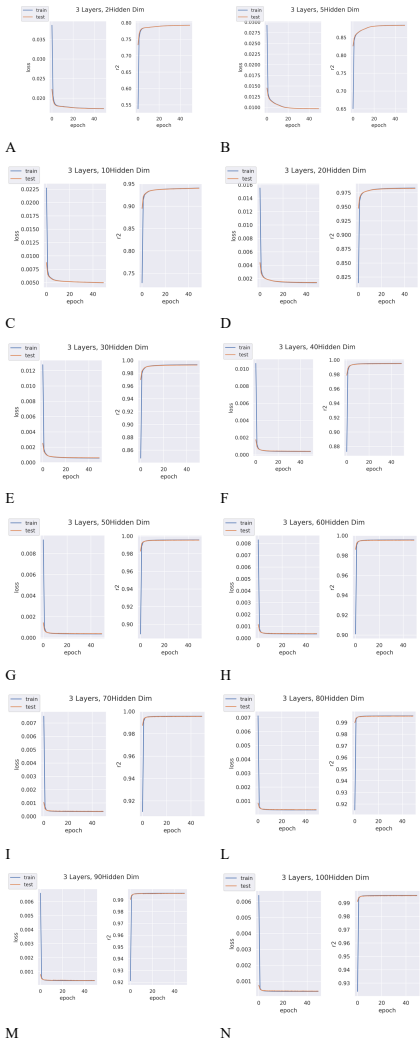


Figure A.1: Single-Layer Autoencoders Training and Test Loss and R^2 on Unlabeled data

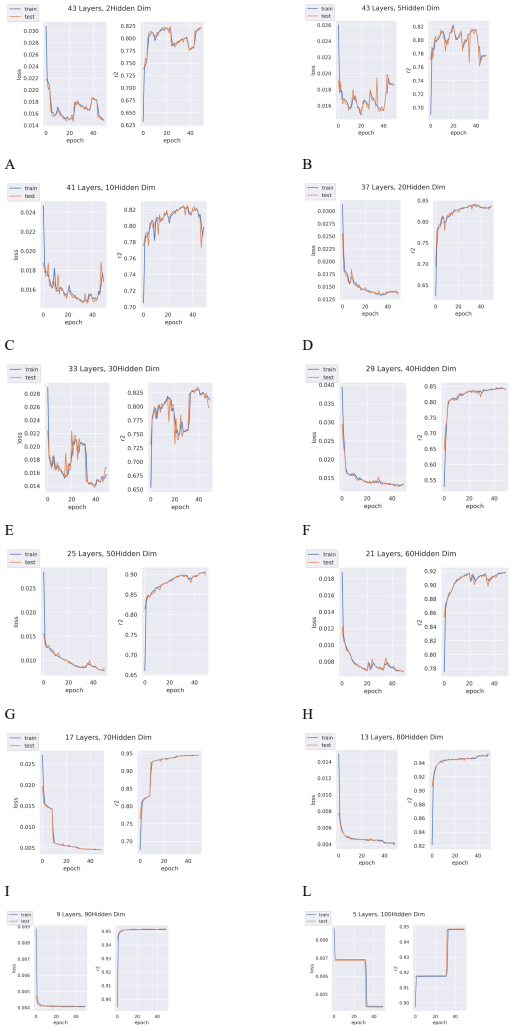


Figure A.2: Deep Autoencoders Training and Test Loss and R^2 on Unlabeled data

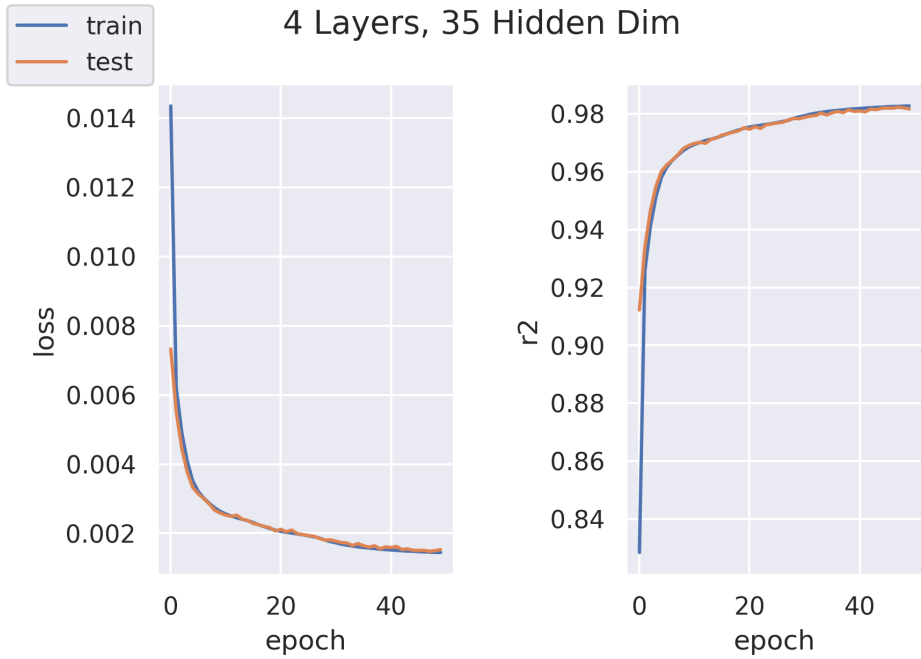


Figure A.3: Training Loss and R2 of an Autoencoder with 4 layers and 35 nodes in the hidden dimension

Appendix B

Cox and Markov Model Results from simulated dataset

B.1 Matrix Trifactorization

Matrix Trifactorization techniques are data fusion approaches that perform dimensionality reduction of large and sparse data sets. The computed latent variable could reveal hidden interactions in data. In Precision Medicine studies, this interaction has been associated with patient similarity.
The algorithm works as follows.
Given c different concepts, such as patients, genes, pathways, and given a data source that relates these concepts (for instance we can relate genes and pathway through KEGG database), we can represent each relationship with a relationship matrix $R_{ij} \epsilon \mathbb{R}^{n_i \times n_j}$, where n_i is the number of objects of type i, and n_j the number of objects of type j. The entire set of R_{ij} are represented in a block matrix R:

$$\mathrm{R}=\begin{pmatrix} R_{11} & R_{12} & \cdots & R_{1r} \\ R_{21} & R_{22} & \cdots & R_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ R_{r1} & R_{r2} & \cdots & R_{rr} \end{pmatrix}$$

(B.1)

R values are in a range between 0 and 1 and they represent the strength of the relationships between two objects.
We could also know the relationship between objects of the same type, such as co-expression of genes. This information is represented by a constraint matrix $\Theta_i \epsilon \mathbb{R}^{n_i \times n_i}$. The comprehensive constraint matrix Θ is therefore a diagonal block matrix, whose values could vary from -1 to 1, where -1 indicates perfect similarity and 1 full dissimilarity.

A crucial parameter is the *rank*, i.e. the dimension of the latent factor that we want to achieve through the matrix factorization. The rank k_i must be empirically set for each of the concepts. Then, after rank selection, two lower-rank block matrices, G and S, are defined:

$$\mathrm{S}=\begin{pmatrix} S_{11}^{k_1 \times k_1} & S_{12}^{k_1 \times k_2} & \cdots & S_{1r}^{k_1 \times k_r} \\ S_{21}^{k_2 \times k_1} & S_{22}^{k_2 \times k_2} & \cdots & S_{2r}^{k_2 \times k_r} \\ \vdots & \vdots & \ddots & \vdots \\ S_{r1}^{k_r \times k_1} & S_{r2}^{k_r \times k_2} & \cdots & S_{rr}^{k_r \times k_r} \end{pmatrix}$$

(B.2)

The two matrices are reconstructed by minimizing the following objective function:

where $||.||$ is the Frobenius norm and $tr()$ is the trace of the matrix.
G and S are randomly initialized and they are iteratively updated till convergence [264]. The algorithm above has been adapted in order to compute

the similarity between two objects of the same type [211], such as "patients". In particular, G_i is a $n_i k_i$, where the rows correspond to patients and the columns to groups. Therefore, we can cluster each patient in the group identified by the column with the maximum value in the patient row. Since the optimization strategy strongly depends on the initialization, i.e. the choice of the ranks k_i, which also correspond to the number of groups, the final consensus matrix C is obtained over n application (in our case, $n = 10$). The value $c_i j$ in C shows how many times the patient pi is clustered together with the patient p_j. For instance, if $c_i j$ is equal to 0.5, this means that the patient pi is grouped with p_j 5 times out of 10. The final consensus matrix is shown in Figure B.1.

B.2 Progression Algorithm

In order to assess the proper number of Monte Carlo simulations needed to build patient trajectories (see Algorithm 2) we relied on the mathematical theory behind Markov chains and their absorbing states [265]. Since our final step is to model MDS disease stages as a Markov process, we ensured that the simulation converges to the absorbing state, i.e. the fixed stages that once reached, the system will never leave. In our case, the final stage is the Very High IPSSR stage, therefore we are looking for trajectories that span through all intermediate stages up to the absorbing one. The distribution of Markov states at time t_n of the Markov chain is the vector v:
$v_n = v_0 * P^(t_n)$
Where P is the transition probability between different stages.
To estimate the number of iterations, we suppose that $P_i j$ between stage i and stage j is represented, in the worst case scenario, by the lowest transition probability between two patients in the two subsequent stages. According to our approach, a patient enconpasses two probabilistic steps to be part of a trajectory: first, he/she can evolve according to the mean transition probability in his/her stage [194]; if he/she evolves, the following patient in the trajectory is selected through a Roulette wheel with probabilities proportional to matrix trifactorization similarity (see Algorithm

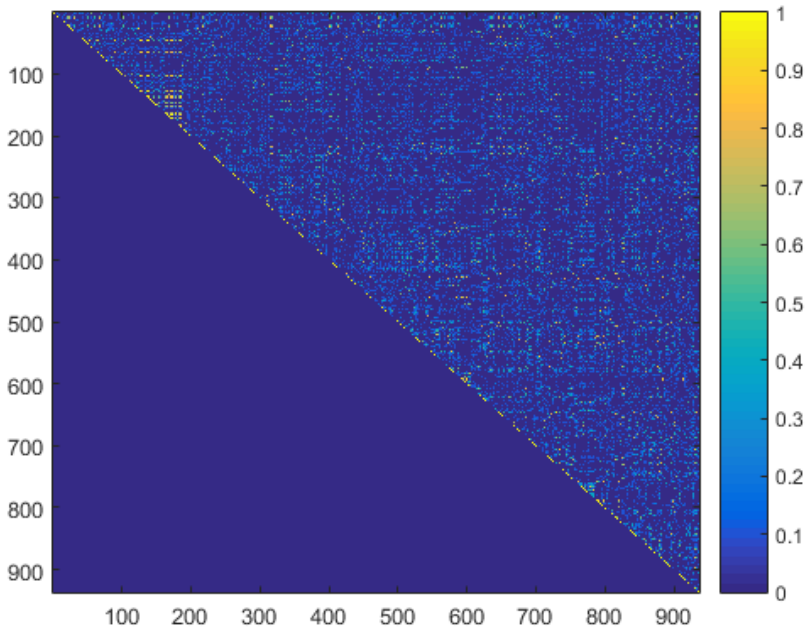


Figure B.1: Consesus matrix representing MDS patient similarities. Higher values (yellow cells) indicate higher similarity.

2). Therefore, in the worst case scenario, the probability of transition is the product of the mean transition probability from the literature[206] and the lowest similarity (corresponding to the more distant patient) computed with the matrix trifactorization. The computed P is equal to:

$$P = \begin{pmatrix} 0.9977 & 0.0023 & 0 & 0 & 0 \\ 0 & 0.9949 & 0.0051 & 0 & 0 \\ 0 & 0 & 0.9932 & 0.0068 & 0 \\ 0 & 0 & 0 & 0.9902 & 0.0098 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Starting from the lowest risk stage ($v_0 = [1,\ 0,\ 0,\ 0,\ 0]$) the final final distribution of the Markov states v_n at $t_n = 4,000$ is equal to $v_n = [0,\ 0,\ 0,\ 0,\ 0.999]$, while for $t_n = 5,000$ or higher, the Makov chain converges to the absorbing state ($v_n = [0,\ 0,\ 0,\ 0,\ 1]$). A conservative choice, since our final aim is to estimate P after longitudinal dataset simulation, is to set the number of iterations t_n to be 10^4.

B.3 Cox and Markov model

The results of Cox modelling are the list of genomic and clinical features that play a role in MDS progression across IPSSR stages.
Here, we mainly focus our attention on genomic features. In our dataset genomic covariates correspond to a numeric score that represents the damaging effect of somatic mutations in 44 genes associated with MDS. In the transition between "Very Low" and "Low" IPSSR levels, mutations on IDH1, NPM1 and KDM6A genes promote the progression of the disease, whereas driver mutations on FLT3 have a protective role (Table B.1).

In the transition between "Low" and "Intermediate" levels of risks, different clinical features have a significant prognostic impact. Only a gene was reported to provide a higher probability of progression (see Table B.2). Cox results on transition between "Intermediate" and "High" IPSSR levels of risks suggest that many genes have a prognostic impact in our cohort.

Table B.1: Genes where driver somatic mutations have an effect in "Very Low" to "Low" transition, according to Cox results. The References column lists previous published works that confirm or not our results. "Poor prognosis"" definition refers to covariates with positive β coefficients, indicating that mutations in that particular gene are associated with a higher probability of progression towards higher stages.

Gene	Effect	References
KDM6A	Poor prognosis	PMID: 27023522; PMID 27235425;
IDH1	Poor prognosis	PMID: 28873367; PMID: 20494930; PMID 27992414
NPM1	Poor prognosis	PMID: 16455956; PMID: 21173125; PMID 27992414; PMID 24220272
FLT3	Good prognosis	PMID: 23115106

Table B.2: Genes where driver somatic mutations have an effect in "Low" to "Intermediate" transition, according to Cox results. The References column lists previous published works that confirm or not our results. "Poor prognosis" definition refers to covariates with positive β coefficients, indicating that mutations in that particular gene are associated with a higher probability of progression towards higher stages.

Gene	Effect	References
MLL3	Poor prognosis	PMID: 24794707; PMID: 27610619; PMID 29615405

Mutations in DNMT3A seems to be protective against disease progression. In contrast, genomic variations in CBL, SF3B1, TET2, CREBB, NRAS and IDH2 are associated with poor prognosis.
In the transition between "High" and "Very High" IPSSR stages according to Cox results, driver mutations in CDKN2A, CTNNA1 and IDH1 genes are reported to provide a higher probability of progression and are associated with poor prognosis (Table B.4).

Regarding transitions towards death, during low stages of the disease, Charlson score values mainly determine the likelihood of the transition. We analysed the transitions probabilities among different IPSSR stages relatively to ten years after the diagnosis (Matrix 2). Transitions between non subsequent levels of risk subtend transitions between intermediate stages, since our Markov model allows evolutions only towards the subsequent stage.

$$P(t) = \begin{pmatrix} 0.211 & 0.288 & 0.127 & 0.132 & 0.0288 & 0.213 \\ 0 & 0.185 & 0.122 & 0.195 & 0.0469 & 0.451 \\ 0 & 0 & 0.0184 & 0.112 & 0.0316 & 0.838 \\ 0 & 0 & 0 & 0.0413 & 0.0125 & 0.946 \\ 0 & 0 & 0 & 0 & 1.51-06 & 0.999 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Table B.3: Genes where driver somatic mutations have an effect in "Intermediate" to "High" transition according to Cox results. The References column lists previous published works that confirm or not our results. "Poor prognosis" definition refers to covariates with positive β coefficients, indicating that mutations in that particular gene are associated with a higher probability of progression towards higher stages, while "Good prognosis" states that the covariate is protective.

Gene	Effect	References
CBL	Poor prognosis	PMID: 23010802;PMID 2131879
SF3B1	Poor prognosis	PMID: 25957392 (not confirming); PMID: 23160465
TET2	Poor prognosis	PMID: 19666869 (not confirming); PMID: 21714648
CREBB	Poor prognosis	PMID: 21390130;
NRAS	Poor prognosis	PMID: 23708912; PMID 27992414
IDH2	Poor prognosis	PMID: 29549529; PMID 27992414
DNMT3A	Good prognosis	PMID: 21415852 (not confirming)

Table B.4: Genes where driver somatic mutations have an effect in "High" to "Very High", transition according to Cox results. The References column lists previous published works that confirm or not our results. "Poor prognosis" definition refers to covariates with positive β coefficients, indicating that mutations in that particular gene are associated with a higher probability of progression towards higher stages, while "Good prognosis" states that the covariate is protective.

Gene	Effect	References
CDKN2A	Good prognosis	PMID: 22248274 (not confirming)
CTNNA1	Poor prognosis	PMID: 25177364; PMID: 25153418; PMID: 19826047; PMID 17159988
IDH1	Poor prognosis	PMID: 28873367; PMID: 20494930; PMID 27992414

Bibliography

[1] Marilyn M. Li, Michael Datto, Eric J. Duncavage, Shashikant Kulkarni, Neal I. Lindeman, Somak Roy, Apostolia M. Tsimberidou, Cindy L. Vnencak-Jones, Daynna J. Wolff, Anas Younes, and Marina N. Nikiforova. Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer: A Joint Consensus Recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. *The Journal of Molecular Diagnostics*, 19(1):4–23, January 2017.

[2] Francis S. Collins and Harold Varmus. A New Initiative on Precision Medicine. *New England Journal of Medicine*, 372(9):793–795, February 2015. Publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMp1500523.

[3] Geoffrey S Ginsburg and Kathryn A Phillips. Precision Medicine: From Science to Value. *Health affairs (Project Hope)*, 37(5):694–701, May 2018.

[4] National Research Council (US) Committee on A Framework for Developing a New Taxonomy of Disease. *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New*

Taxonomy of Disease. The National Academies Collection: Reports funded by National Institutes of Health. National Academies Press (US), Washington (DC), 2011.

[5] Jean-Louis Vincent. The coming era of precision medicine for intensive care. *Critical Care*, 21(3):314, December 2017.

[6] Rui Chen and Michael Snyder. Promise of personalized omics to precision medicine. *WIREs Systems Biology and Medicine*, 5(1):73–82, 2013. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wsbm.1198.

[7] Eric S. Lander, Lauren M. Linton, Bruce Birren, Chad Nusbaum, Michael C. Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, Roel Funke, Diane Gage, Katrina Harris, Andrew Heaford, John Howland, Lisa Kann, Jessica Lehoczky, Rosie LeVine, Paul McEwan, Kevin McKernan, James Meldrim, Jill P. Mesirov, Cher Miranda, William Morris, Jerome Naylor, Christina Raymond, Mark Rosetti, Ralph Santos, Andrew Sheridan, Carrie Sougnez, Nicole Stange-Thomann, Nikola Stojanovic, Aravind Subramanian, Dudley Wyman, Jane Rogers, John Sulston, Rachael Ainscough, Stephan Beck, David Bentley, John Burton, Christopher Clee, Nigel Carter, Alan Coulson, Rebecca Deadman, Panos Deloukas, Andrew Dunham, Ian Dunham, Richard Durbin, Lisa French, Darren Grafham, Simon Gregory, Tim Hubbard, Sean Humphray, Adrienne Hunt, Matthew Jones, Christine Lloyd, Amanda McMurray, Lucy Matthews, Simon Mercer, Sarah Milne, James C. Mullikin, Andrew Mungall, Robert Plumb, Mark Ross, Ratna Shownkeen, Sarah Sims, Robert H. Waterston, Richard K. Wilson, LaDeana W. Hillier, John D. McPherson, Marco A. Marra, Elaine R. Mardis, Lucinda A. Fulton, Asif T. Chinwalla, Kymberlie H. Pepin, Warren R. Gish, Stephanie L. Chissoe, Michael C. Wendl, Kim D. Delehaunty, Tracie L. Miner, Andrew Delehaunty, Jason B. Kramer, Lisa L. Cook, Robert S. Fulton, Douglas L. Johnson, Patrick J. Minx, Sandra W. Clifton, Trevor Hawkins, Elbert Branscomb, Paul Predki,

Paul Richardson, Sarah Wenning, Tom Slezak, Norman Doggett, Jan-Fang Cheng, Anne Olsen, Susan Lucas, Christopher Elkin, Edward Uberbacher, Marvin Frazier, Richard A. Gibbs, Donna M. Muzny, Steven E. Scherer, John B. Bouck, Erica J. Sodergren, Kim C. Worley, Catherine M. Rives, James H. Gorrell, Michael L. Metzker, Susan L. Naylor, Raju S. Kucherlapati, David L. Nelson, George M. Weinstock, Yoshiyuki Sakaki, Asao Fujiyama, Masahira Hattori, Tetsushi Yada, Atsushi Toyoda, Takehiko Itoh, Chiharu Kawagoe, Hidemi Watanabe, Yasushi Totoki, Todd Taylor, Jean Weissenbach, Roland Heilig, William Saurin, Francois Artiguenave, Philippe Brottier, Thomas Bruls, Eric Pelletier, Catherine Robert, Patrick Wincker, André Rosenthal, Matthias Platzer, Gerald Nyakatura, Stefan Taudien, Andreas Rump, Douglas R. Smith, Lynn Doucette-Stamm, Marc Rubenfield, Keith Weinstock, Hong Mei Lee, JoAnn Dubois, Huanming Yang, Jun Yu, Jian Wang, Guyang Huang, Jun Gu, Leroy Hood, Lee Rowen, Anup Madan, Shizen Qin, Ronald W. Davis, Nancy A. Federspiel, A. Pia Abola, Michael J. Proctor, Bruce A. Roe, Feng Chen, Huaqin Pan, Juliane Ramser, Hans Lehrach, Richard Reinhardt, W. Richard McCombie, Melissa de la Bastide, Neilay Dedhia, Helmut Blöcker, Klaus Hornischer, Gabriele Nordsiek, Richa Agarwala, L. Aravind, Jeffrey A. Bailey, Alex Bateman, Serafim Batzoglou, Ewan Birney, Peer Bork, Daniel G. Brown, Christopher B. Burge, Lorenzo Cerutti, Hsiu-Chuan Chen, Deanna Church, Michele Clamp, Richard R. Copley, Tobias Doerks, Sean R. Eddy, Evan E. Eichler, Terrence S. Furey, James Galagan, James G. R. Gilbert, Cyrus Harmon, Yoshihide Hayashizaki, David Haussler, Henning Hermjakob, Karsten Hokamp, Wonhee Jang, L. Steven Johnson, Thomas A. Jones, Simon Kasif, Arek Kaspryzk, Scot Kennedy, W. James Kent, Paul Kitts, Eugene V. Koonin, Ian Korf, David Kulp, Doron Lancet, Todd M. Lowe, Aoife McLysaght, Tarjei Mikkelsen, John V. Moran, Nicola Mulder, Victor J. Pollara, Chris P. Ponting, Greg Schuler, Jörg Schultz, Guy Slater, Arian F. A. Smit, Elia Stupka, Joseph Szustakowki, Danielle Thierry-Mieg, Jean Thierry-Mieg, Lukas Wagner, John Wallis, Raymond Wheeler, Alan Williams,

Yuri I. Wolf, Kenneth H. Wolfe, Shiaw-Pyng Yang, Ru-Fang Yeh, Francis Collins, Mark S. Guyer, Jane Peterson, Adam Felsenfeld, Kris A. Wetterstrand, Richard M. Myers, Jeremy Schmutz, Mark Dickson, Jane Grimwood, David R. Cox, Maynard V. Olson, Rajinder Kaul, Christopher Raymond, Nobuyoshi Shimizu, Kazuhiko Kawasaki, Shinsei Minoshima, Glen A. Evans, Maria Athanasiou, Roger Schultz, Aristides Patrinos, Michael J. Morgan, International Human Genome Sequencing Consortium, Center for Genome Research: Whitehead Institute for Biomedical Research, The Sanger Centre:, Washington University Genome Sequencing Center, US DOE Joint Genome Institute:, Baylor College of Medicine Human Genome Sequencing Center:, RIKEN Genomic Sciences Center:, Genoscope and CNRS UMR-8030:, Institute of Molecular Biotechnology: Department of Genome Analysis, GTC Sequencing Center:, Beijing Genomics Institute/Human Genome Center:, The Institute for Systems Biology: Multimegabase Sequencing Center, Stanford Genome Technology Center:, University of Oklahoma's Advanced Center for Genome Technology:, Max Planck Institute for Molecular Genetics:, Lita Annenberg Hazen Genome Center: Cold Spring Harbor Laboratory, GBF—German Research Centre for Biotechnology:, also includes individuals listed under other headings): *Genome Analysis Group (listed in alphabetical order, US National Institutes of Health: Scientific management: National Human Genome Research Institute, Stanford Human Genome Center:, University of Washington Genome Center:, Keio University School of Medicine: Department of Molecular Biology, University of Texas Southwestern Medical Center at Dallas:, US Department of Energy: Office of Science, and The Wellcome Trust:. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, February 2001. Number: 6822 Publisher: Nature Publishing Group.

[8] Francis S. Collins and Victor A. McKusick. Implications of the Human Genome Project for Medical Science. *JAMA*, 285(5):540–544, February 2001. Publisher: American Medical Association.

[9] Chandra Shekhar Pareek, Rafal Smoczynski, and Andrzej Tretyn. Sequencing technologies and genome sequencing. *Journal of Applied Genetics*, 52(4):413–435, November 2011.

[10] Elaine R. Mardis. The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, 24(3):133–141, March 2008.

[11] Erwin L. van Dijk, Hélène Auger, Yan Jaszczyszyn, and Claude Thermes. Ten years of next-generation sequencing technology. *Trends in Genetics*, 30(9):418–426, September 2014.

[12] Katharina Schwarze, James Buchanan, Jilles M. Fermont, Helene Dreau, Mark W. Tilley, John M. Taylor, Pavlos Antoniou, Samantha J. L. Knight, Carme Camps, Melissa M. Pentony, Erika M. Kvikstad, Steve Harris, Niko Popitsch, Alistair T. Pagnamenta, Anna Schuh, Jenny C. Taylor, and Sarah Wordsworth. The complete costs of genome sequencing: a microcosting study in cancer and rare diseases from a single center in the United Kingdom. *Genetics in Medicine*, 22(1):85–94, January 2020. Number: 1 Publisher: Nature Publishing Group.

[13] Patricia Marino, Rajae Touzani, Lionel Perrier, Etienne Rouleau, Dede Sika Kossi, Zou Zhaomin, Nathanaël Charrier, Nicolas Goardon, Claude Preudhomme, Isabelle Durand-Zaleski, Isabelle Borget, Sandrine Baffert, and NGSEco Group:. Cost of cancer diagnosis using next-generation sequencing targeted gene panels in routine practice: a nationwide French study. *European journal of human genetics: EJHG*, 26(3):314–323, 2018.

[14] Wallace J. Hopp, Jun Li, and Guihua Wang. Big Data and the Precision Medicine Revolution. *Production and Operations Management*, 27(9):1647–1664, 2018. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/poms.12891.

[15] Sowmiya Moorthie, Alison Hall, and Caroline F. Wright. Informatics and clinical genome sequencing: opening the black box. *Genetics in*

Medicine, 15(3):165–171, March 2013. Number: 3 Publisher: Nature Publishing Group.

[16] Christopher J. Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17(1):195, October 2019.

[17] Annie Y. S. Lau and Pascal Staccini. Artificial Intelligence in Health: New Opportunities, Challenges, and Practical Implications. *Yearbook of Medical Informatics*, 28(1):174–178, August 2019.

[18] Jia Xu, Pengwei Yang, Shang Xue, Bhuvan Sharma, Marta Sanchez-Martin, Fang Wang, Kirk A. Beaty, Elinor Dehan, and Baiju Parikh. Translating cancer genomics into precision medicine with artificial intelligence: applications, challenges and future perspectives. *Human Genetics*, 138(2):109–124, February 2019.

[19] Aliza Becker. Artificial intelligence in medicine: What is it doing for us today? *Health Policy and Technology*, 8(2):198–205, June 2019.

[20] Zodwa Dlamini, Flavia Zita Francies, Rodney Hull, and Rahaba Marima. Artificial intelligence (AI) and big data in cancer and precision oncology. *Computational and Structural Biotechnology Journal*, August 2020.

[21] Giovanna Nicora, Francesca Vitali, Arianna Dagliati, Nophar Geifman, and Riccardo Bellazzi. Integrated Multi-Omics Analyses in Oncology: A Review of Machine Learning Methods and Tools. *Frontiers in Oncology*, 10, 2020. Publisher: Frontiers.

[22] Nicolas Servant, Julien Roméjon, Pierre Gestraud, Philippe La Rosa, Georges Lucotte, Séverine Lair, Virginie Bernard, Bruno Zeitouni, Fanny Coffin, Gérôme Jules-Clément, Florent Yvon, Alban Lermine, Patrick Poullet, Stéphane Liva, Stuart Pook, Tatiana Popova, Camille Barette, François Prud'homme, Jean-Gabriel Dick, Maud Kamal,

Christophe Le Tourneau, Emmanuel Barillot, and Philippe Hupé. Bioinformatics for precision medicine in oncology: principles and application to the SHIVA clinical trial. *Frontiers in Genetics*, 5:152, 2014.

[23] G. Carioli, P. Bertuccio, P. Boffetta, F. Levi, C. La Vecchia, E. Negri, and M. Malvezzi. European cancer mortality predictions for the year 2020 with a focus on prostate cancer. *Annals of Oncology*, 31(5):650–658, May 2020. Publisher: Elsevier.

[24] Camille Maringe, James Spicer, Melanie Morris, Arnie Purushotham, Ellen Nolte, Richard Sullivan, Bernard Rachet, and Ajay Aggarwal. The impact of the COVID-19 pandemic on cancer deaths due to delays in diagnosis in England, UK: a national, population-based, modelling study. *The Lancet Oncology*, 21(8):1023–1034, August 2020. Publisher: Elsevier.

[25] Rebecca L. Siegel, Kimberly D. Miller, and Ahmedin Jemal. Cancer statistics, 2020. *CA: A Cancer Journal for Clinicians*, 70(1):7–30, 2020. _eprint: https://acsjournals.onlinelibrary.wiley.com/doi/pdf/10.3322/caac.21590.

[26] Milena Sant, Claudia Allemani, Carmen Tereanu, Roberta De Angelis, Riccardo Capocaccia, Otto Visser, Rafael Marcos-Gragera, Marc Maynadié, Arianna Simonetti, Jean-Michel Lutz, and Franco Berrino. Incidence of hematologic malignancies in Europe by morphologic subtype: results of the HAEMACARE project. *Blood*, 116(19):3724–3734, November 2010. Publisher: American Society of Hematology.

[27] Michael R. Stratton, Peter J. Campbell, and P. Andrew Futreal. The cancer genome. *Nature*, 458(7239):719–724, April 2009. Number: 7239 Publisher: Nature Publishing Group.

[28] Georgios N. Tsaousis, Eirini Papadopoulou, Angela Apessos, Konstantinos Agiannitopoulos, Georgia Pepe, Stavroula Kampouri, Nikolaos Diamantopoulos, Theofanis Floros, Rodoniki Iosifidou, Ourania

Katopodi, Anna Koumarianou, Christos Markopoulos, Konstantinos Papazisis, Vasileios Venizelos, Ioannis Xanthakis, Grigorios Xepapadakis, Eugeniu Banu, Dan Tudor Eniu, Serban Negru, Dana Lucia Stanculeanu, Andrei Ungureanu, Vahit Ozmen, Sualp Tansan, Mehmet Tekinel, Suayib Yalcin, and George Nasioulas. Analysis of hereditary cancer syndromes by using a panel of genes: novel and multiple pathogenic mutations. *BMC Cancer*, 19, June 2019.

[29] Ruth Nussinov, Hyunbum Jang, Chung-Jung Tsai, and Feixiong Cheng. Review: Precision medicine and driver mutations: Computational methods, functional assays and conformational principles for interpreting cancer drivers. *PLOS Computational Biology*, 15(3):e1006658, March 2019. Publisher: Public Library of Science.

[30] Julie S. Bødker, Mads Sønderkær, Charles Vesteghem, Alexander Schmitz, Rasmus F. Brøndum, Mia Sommer, Anne S. Rytter, Marlene M. Nielsen, Jakob Madsen, Paw Jensen, Inge S. Pedersen, Lykke Grubach, Marianne T. Severinsen, Anne S. Roug, Tarec C. El-Galaly, Karen Dybkær, and Martin Bøgsted. Development of a Precision Medicine Workflow in Hematological Cancers, Aalborg University Hospital, Denmark. *Cancers*, 12(2):312, February 2020. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

[31] Daniel Auclair, Sagar Lonial, Kenneth C. Anderson, and Shaji K. Kumar. Precision medicine in multiple myeloma: are we there yet? *Expert Review of Precision Medicine and Drug Development*, 4(2):51–53, March 2019. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/23808993.2019.1578172.

[32] Sue Richards, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W. Grody, Madhuri Hegde, Elaine Lyon, Elaine Spector, Karl Voelkerding, Heidi L. Rehm, and ACMG Laboratory Quality Assurance Committee. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and

the Association for Molecular Pathology. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, 17(5):405–424, May 2015.

[33] Bert Vogelstein, Nickolas Papadopoulos, Victor E. Velculescu, Shibin Zhou, Luis A. Diaz, and Kenneth W. Kinzler. Cancer Genome Landscapes. *Science (New York, N.Y.)*, 339(6127):1546–1558, March 2013.

[34] Janet E. Dancey, Philippe L. Bedard, Nicole Onetto, and Thomas J. Hudson. The genetic basis for cancer treatment decisions. *Cell*, 148(3):409–420, February 2012.

[35] Richard M. Durbin, David Altshuler, Richard M. Durbin, Gonçalo R. Abecasis, David R. Bentley, Aravinda Chakravarti, Andrew G. Clark, Francis S. Collins, Francisco M. De La Vega, Peter Donnelly, Michael Egholm, Paul Flicek, Stacey B. Gabriel, Richard A. Gibbs, Bartha M. Knoppers, Eric S. Lander, Hans Lehrach, Elaine R. Mardis, Gil A. McVean, Deborah A. Nickerson, Leena Peltonen, Alan J. Schafer, Stephen T. Sherry, Jun Wang, Richard K. Wilson, Richard A. Gibbs, David Deiros, Mike Metzker, Donna Muzny, Jeff Reid, David Wheeler, Jun Wang, Jingxiang Li, Min Jian, Guoqing Li, Ruiqiang Li, Huiqing Liang, Geng Tian, Bo Wang, Jian Wang, Wei Wang, Huanming Yang, Xiuqing Zhang, Huisong Zheng, Eric S. Lander, David Altshuler, Lauren Ambrogio, Toby Bloom, Kristian Cibulskis, Tim J. Fennell, Stacey B. Gabriel, David B. Jaffe, Erica Shefler, Carrie L. Sougnez, David R. Bentley, Niall Gormley, Sean Humphray, Zoya Kingsbury, Paula Kokko-Gonzales, Jennifer Stone, Kevin J. McKernan, Gina L. Costa, Jeffry K. Ichikawa, Clarence C. Lee, Ralf Sudbrak, Hans Lehrach, Tatiana A. Borodina, Andreas Dahl, Alexey N. Davydov, Peter Marquardt, Florian Mertes, Wilfiried Nietfeld, Philip Rosenstiel, Stefan Schreiber, Aleksey V. Soldatov, Bernd Timmermann, Marius Tolzmann, Michael Egholm, Jason Affourtit, Dana Ashworth, Said Attiya, Melissa Bachorski, Eli Buglione, Adam Burke, Amanda Caprio, Christopher Celone, Shauna Clark, David Conners, Brian Desany, Lisa Gu, Lorri Guccione, Kalvin Kao, An-

drew Kebbel, Jennifer Knowlton, Matthew Labrecque, Louise McDade, Craig Mealmaker, Melissa Minderman, Anne Nawrocki, Faheem Niazi, Kristen Pareja, Ravi Ramenani, David Riches, Wanmin Song, Cynthia Turcotte, Shally Wang, Elaine R. Mardis, Richard K. Wilson, David Dooling, Lucinda Fulton, Robert Fulton, George Weinstock, Richard M. Durbin, John Burton, David M. Carter, Carol Churcher, Alison Coffey, Anthony Cox, Aarno Palotie, Michael Quail, Tom Skelly, James Stalker, Harold P. Swerdlow, Daniel Turner, Anniek De Witte, Shane Giles, Richard A. Gibbs, David Wheeler, Matthew Bainbridge, Danny Challis, Aniko Sabo, Fuli Yu, Jin Yu, Jun Wang, Xiaodong Fang, Xiaosen Guo, Ruiqiang Li, Yingrui Li, Ruibang Luo, Shuaishuai Tai, Honglong Wu, Hancheng Zheng, Xiaole Zheng, Yan Zhou, Guoqing Li, Jian Wang, Huanming Yang, Gabor T. Marth, Erik P. Garrison, Weichun Huang, Amit Indap, Deniz Kural, Wan-Ping Lee, Wen Fung Leong, Aaron R. Quinlan, Chip Stewart, Michael P. Stromberg, Alistair N. Ward, Jiantao Wu, Charles Lee, Ryan E. Mills, Xinghua Shi, Mark J. Daly, Mark A. DePristo, David Altshuler, Aaron D. Ball, Eric Banks, Toby Bloom, Brian L. Browning, Kristian Cibulskis, Tim J. Fennell, Kiran V. Garimella, Sharon R. Grossman, Robert E. Handsaker, Matt Hanna, Chris Hartl, David B. Jaffe, Andrew M. Kernytsky, Joshua M. Korn, Heng Li, Jared R. Maguire, Steven A. McCarroll, Aaron McKenna, James C. Nemesh, Anthony A. Philippakis, Ryan E. Poplin, Alkes Price, Manuel A. Rivas, Pardis C. Sabeti, Stephen F. Schaffner, Erica Shefler, Ilya A. Shlyakhter, David N. Cooper, Edward V. Ball, Matthew Mort, Andrew D. Phillips, Peter D. Stenson, Jonathan Sebat, Vladimir Makarov, Kenny Ye, Seungtai C. Yoon, Carlos D. Bustamante, Laura Clarke, Paul Flicek, Fiona Cunningham, Javier Herrero, Stephen Keenen, Eugene Kulesha, Rasko Leinonen, William M. McLaren, Rajesh Radhakrishnan, Richard E. Smith, Vadim Zalunin, Xiangqun Zheng-Bradley, Jan O. Korbel, Adrian M. Stütz, Sean Humphray, Markus Bauer, R. Keira Cheetham, Tony Cox, Michael Eberle, Terena James, Scott Kahn, Lisa Murray, Aravinda Chakravarti, Kai Ye, Francisco M. De La Vega, Yutao Fu, Fiona C. L.

Hyland, Jonathan M. Manning, Stephen F. McLaughlin, Heather E. Peckham, Onur Sakarya, Yongming A. Sun, Eric F. Tsung, Mark A. Batzer, Miriam K. Konkel, Jerilyn A. Walker, Ralf Sudbrak, Marcus W. Albrecht, Vyacheslav S. Amstislavskiy, Ralf Herwig, Dimitri V. Parkhomchuk, Stephen T. Sherry, Richa Agarwala, Hoda M. Khouri, Aleksandr O. Morgulis, Justin E. Paschall, Lon D. Phan, Kirill E. Rotmistrovsky, Robert D. Sanders, Martin F. Shumway, Chunlin Xiao, Gil A. McVean, Adam Auton, Zamin Iqbal, Gerton Lunter, Jonathan L. Marchini, Loukas Moutsianas, Simon Myers, Afidalina Tumian, Brian Desany, James Knight, Roger Winer, David W. Craig, Steve M. Beckstrom-Sternberg, Alexis Christoforides, The 1000 Genomes Project Consortium, Corresponding author, Steering committee, Production group: Baylor College of Medicine, BGI-Shenzhen, Broad Institute of MIT and Harvard, Illumina, Life Technologies, Max Planck Institute for Molecular Genetics, Roche Applied Science, Washington University in St Louis, Wellcome Trust Sanger Institute, Analysis group: Agilent Technologies, Baylor College of Medicine, Boston College, Brigham and Women's Hospital, The Human Gene Mutation Database Cardiff University, Cold Spring Harbor Laboratory, Cornell and Stanford Universities, European Bioinformatics Institute, European Molecular Biology Laboratory, Johns Hopkins University, Leiden University Medical Center, Louisiana State University, US National Institutes of Health, Oxford University, and The Translational Genomics Research Institute. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, October 2010. Number: 7319 Publisher: Nature Publishing Group.

[36] Lea M. Starita, Nadav Ahituv, Maitreya J. Dunham, Jacob O. Kitzman, Frederick P. Roth, Georg Seelig, Jay Shendure, and Douglas M. Fowler. Variant Interpretation: Functional Assays to the Rescue. *The American Journal of Human Genetics*, 101(3):315–325, September 2017.

[37] Iñigo Martincorena, Keiran M. Raine, Moritz Gerstung, Kevin J. Dawson, Kerstin Haase, Peter Van Loo, Helen Davies, Michael R. Stratton, and Peter J. Campbell. Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell*, 171(5):1029–1041.e21, November 2017.

[38] Laura D. Wood, D. Williams Parsons, Siân Jones, Jimmy Lin, Tobias Sjöblom, Rebecca J. Leary, Dong Shen, Simina M. Boca, Thomas Barber, Janine Ptak, Natalie Silliman, Steve Szabo, Zoltan Dezso, Vadim Ustyanksky, Tatiana Nikolskaya, Yuri Nikolsky, Rachel Karchin, Paul A. Wilson, Joshua S. Kaminker, Zemin Zhang, Randal Croshaw, Joseph Willis, Dawn Dawson, Michail Shipitsin, James K. V. Willson, Saraswati Sukumar, Kornelia Polyak, Ben Ho Park, Charit L. Pethiyagoda, P. V. Krishna Pant, Dennis G. Ballinger, Andrew B. Sparks, James Hartigan, Douglas R. Smith, Erick Suh, Nickolas Papadopoulos, Phillip Buckhaults, Sanford D. Markowitz, Giovanni Parmigiani, Kenneth W. Kinzler, Victor E. Velculescu, and Bert Vogelstein. The genomic landscapes of human breast and colorectal cancers. *Science (New York, N.Y.)*, 318(5853):1108–1113, November 2007.

[39] Constance Baer, Wencke Walter, Stephan Hutter, Sven Twardziok, Manja Meggendorfer, Wolfgang Kern, Torsten Haferlach, and Claudia Haferlach. "Somatic" and "pathogenic" - is the classification strategy applicable in times of large-scale sequencing? *Haematologica*, 104(8):1515–1520, August 2019. Publisher: Haematologica Section: Perspective Article.

[40] Gayane Badalian-Very. Personalized medicine in hematology — A landmark from bench to bed. *Computational and Structural Biotechnology Journal*, 10(17):70–77, July 2014.

[41] Suanna Steeby Bruinooge, Shimere Sherwood, Stephen Grubbs, and Richard L. Schilsky. Determining If a Somatic Tumor Mutation Is Targetable and Options for Accessing Targeted Therapies. *Journal*

of Oncology Practice, 15(11):575–583, August 2019. Publisher: American Society of Clinical Oncology.

[42] Connie Lee Batlevi, Gunjan Shah, Christopher Forlenza, and Andrew Intlekofer. Using genomic data for selecting the treatment of lymphoma patients. *Current Opinion in Hematology*, 26(4):303–312, 2019.

[43] Daniela Senft, Mark D. M. Leiserson, Eytan Ruppin, and Ze'ev A. Ronai. Precision Oncology: The Road Ahead. *Trends in Molecular Medicine*, 23(10):874–898, 2017.

[44] Abhishek Niroula and Mauno Vihinen. Variation Interpretation Predictors: Principles, Types, Performance, and Choice. *Human Mutation*, 37(6):579–597, 2016. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/humu.22987.

[45] Patrick Kwok-Shing Ng, Jun Li, Kang Jin Jeong, Shan Shao, Hu Chen, Yiu Huen Tsang, Sohini Sengupta, Zixing Wang, Venkata Hemanjani Bhavana, Richard Tran, Stephanie Soewito, Darlan Conterno Minussi, Daniela Moreno, Kathleen Kong, Turgut Dogruluk, Hengyu Lu, Jianjiong Gao, Collin Tokheim, Daniel Cui Zhou, Amber M. Johnson, Jia Zeng, Carman Ka Man Ip, Zhenlin Ju, Matthew Wester, Shuangxing Yu, Yongsheng Li, Christopher P. Vellano, Nikolaus Schultz, Rachel Karchin, Li Ding, Yiling Lu, Lydia Wai Ting Cheung, Ken Chen, Kenna R. Shaw, Funda Meric-Bernstam, Kenneth L. Scott, Song Yi, Nidhi Sahni, Han Liang, and Gordon B. Mills. Systematic Functional Annotation of Somatic Mutations in Cancer. *Cancer Cell*, 33(3):450–462.e10, March 2018.

[46] Peter J. Campbell, Gad Getz, Jan O. Korbel, Joshua M. Stuart, Jennifer L. Jennings, Lincoln D. Stein, Marc D. Perry, Hardeep K. Nahal-Bose, B. F. Francis Ouellette, Constance H. Li, Esther Rheinbay, G. Petur Nielsen, Dennis C. Sgroi, Chin-Lee Wu, William C. Faquin, Vikram Deshpande, Paul C. Boutros, Alexander J. Lazar, Katherine A. Hoadley, David N. Louis, L. Jonathan Dursi, Christina K.

Yung, Matthew H. Bailey, Gordon Saksena, Keiran M. Raine, Ivo Buchhalter, Kortine Kleinheinz, Matthias Schlesner, Junjun Zhang, Wenyi Wang, David A. Wheeler, Li Ding, Jared T. Simpson, Brian D. O'Connor, Sergei Yakneen, Kyle Ellrott, Naoki Miyoshi, Adam P. Butler, Romina Royo, Solomon I. Shorser, Miguel Vazquez, Tobias Rausch, Grace Tiao, Sebastian M. Waszak, Bernardo Rodriguez-Martin, Suyash Shringarpure, Dai-Ying Wu, German M. Demidov, Olivier Delaneau, Shuto Hayashi, Seiya Imoto, Nina Habermann, Ayellet V. Segre, Erik Garrison, Andy Cafferkey, Eva G. Alvarez, José María Heredia-Genestar, Francesc Muyas, Oliver Drechsel, Alicia L. Bruzos, Javier Temes, Jorge Zamora, Adrian Baez-Ortega, Hyung-Lae Kim, R. Jay Mashl, Kai Ye, Anthony DiBiase, Kuanlin Huang, Ivica Letunic, Michael D. McLellan, Steven J. Newhouse, Tal Shmaya, Sushant Kumar, David C. Wedge, Mark H. Wright, Venkata D. Yellapantula, Mark Gerstein, Ekta Khurana, Tomas Marques-Bonet, Arcadi Navarro, Carlos D. Bustamante, Reiner Siebert, Hidewaki Nakagawa, Douglas F. Easton, Stephan Ossowski, Jose M. C. Tubio, Francisco M. De La Vega, Xavier Estivill, Denis Yuen, George L. Mihaiescu, Larsson Omberg, Vincent Ferretti, Radhakrishnan Sabarinathan, Oriol Pich, Abel Gonzalez-Perez, Amaro Taylor-Weiner, Matthew W. Fittall, Jonas Demeulemeester, Maxime Tarabichi, Nicola D. Roberts, Peter Van Loo, Isidro Cortés-Ciriano, Lara Urban, Peter Park, Bin Zhu, Esa Pitkänen, Yilong Li, Natalie Saini, Leszek J. Klimczak, Joachim Weischenfeldt, Nikos Sidiropoulos, Ludmil B. Alexandrov, Raquel Rabionet, Georgia Escaramis, Mattia Bosio, Aliaksei Z. Holik, Hana Susak, Aparna Prasad, Serap Erkek, Claudia Calabrese, Benjamin Raeder, Eoghan Harrington, Simon Mayes, Daniel Turner, Sissel Juul, Steven A. Roberts, Lei Song, Roelof Koster, Lisa Mirabello, Xing Hua, Tomas J. Tanskanen, Marta Tojo, Jieming Chen, Lauri A. Aaltonen, Gunnar Rätsch, Roland F. Schwarz, Atul J. Butte, Alvis Brazma, Stephen J. Chanock, Nilanjan Chatterjee, Oliver Stegle, Olivier Harismendy, G. Steven Bova, Dmitry A. Gordenin, David Haan, Lina Sieverling, Lars Feuerbach, Don Chalmers, Yann Joly, Bartha Knoppers, Fruzsina Molnár-

Gábor, Mark Phillips, Adrian Thorogood, David Townend, Mary Goldman, Nuno A. Fonseca, Qian Xiang, Brian Craft, Elena Piñeiro-Yáñez, Alfonso Muñoz, Robert Petryszak, Anja Füllgrabe, Fatima Al-Shahrour, Maria Keays, David Haussler, John Weinstein, Wolfgang Huber, Alfonso Valencia, Irene Papatheodorou, Jingchun Zhu, Yu Fan, David Torrents, Matthias Bieg, Ken Chen, Zechen Chong, Kristian Cibulskis, Roland Eils, Robert S. Fulton, Josep L. Gelpi, Santiago Gonzalez, Ivo G. Gut, Faraz Hach, Michael Heinold, Taobo Hu, Vincent Huang, Barbara Hutter, Natalie Jäger, Jongsun Jung, Yogesh Kumar, Christopher Lalansingh, Ignaty Leshchiner, Dimitri Livitz, Eric Z. Ma, Yosef E. Maruvka, Ana Milovanovic, Morten Muhlig Nielsen, Nagarajan Paramasivam, Jakob Skou Pedersen, Montserrat Puiggròs, S. Cenk Sahinalp, Iman Sarrafi, Chip Stewart, Miranda D. Stobbe, Jeremiah A. Wala, Jiayin Wang, Michael Wendl, Johannes Werner, Zhenggang Wu, Hong Xue, Takafumi N. Yamaguchi, Venkata Yellapantula, Brandi N. Davis-Dusenbery, Robert L. Grossman, Youngwook Kim, Michael C. Heinold, Jonathan Hinton, David R. Jones, Andrew Menzies, Lucy Stebbings, Julian M. Hess, Mara Rosenberg, Andrew J. Dunford, Manaswi Gupta, Marcin Imielinski, Matthew Meyerson, Rameen Beroukhim, Jüri Reimand, Priyanka Dhingra, Francesco Favero, Stefan Dentro, Jeff Wintersinger, Vasilisa Rudneva, Ji Wan Park, Eun Pyo Hong, Seong Gu Heo, André Kahles, Kjong-Van Lehmann, Cameron M. Soulette, Yuichi Shiraishi, Fenglin Liu, Yao He, Deniz Demircioğlu, Natalie R. Davidson, Liliana Greger, Siliang Li, Dongbing Liu, Stefan G. Stark, Fan Zhang, Samirkumar B. Amin, Peter Bailey, Aurélien Chateigner, Milana Frenkel-Morgenstern, Yong Hou, Matthew R. Huska, Helena Kilpinen, Fabien C. Lamaze, Chang Li, Xiaobo Li, Xinyue Li, Xingmin Liu, Maximillian G. Marin, Julia Markowski, Tannistha Nandi, Akinyemi I. Ojesina, Qiang Pan-Hammarström, Peter J. Park, Chandra Sekhar Pedamallu, Hong Su, Patrick Tan, Bin Tean Teh, Jian Wang, Heng Xiong, Chen Ye, Christina Yung, Xiuqing Zhang, Liangtao Zheng, Shida Zhu, Philip Awadalla, Chad J. Creighton, Kui Wu, Huanming Yang, Jonathan Göke, Zemin Zhang, Angela N. Brooks,

Matthew W. Fittall, Iñigo Martincorena, Carlota Rubio-Perez, Malene Juul, Steven Schumacher, Ofer Shapira, David Tamborero, Loris Mularoni, Henrik Hornshøj, Jordi Deu-Pons, Ferran Muiños, Johanna Bertl, Qianyun Guo, Abel Gonzalez-Perez, Qian Xiang, and The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature*, 578(7793):82–93, February 2020. Number: 7793 Publisher: Nature Publishing Group.

[47] Júlia Perera-Bel, Barbara Hutter, Christoph Heining, Annalen Bleckmann, Martina Fröhlich, Stefan Fröhling, Hanno Glimm, Benedikt Brors, and Tim Beißbarth. From somatic variants towards precision oncology: Evidence-driven reporting of treatment options in molecular tumor boards. *Genome Medicine*, 10(1):18, March 2018.

[48] The AACR Project GENIE Consortium. AACR Project GENIE: Powering Precision Medicine through an International Consortium. *Cancer Discovery*, 7(8):818–831, August 2017. Publisher: American Association for Cancer Research Section: Research Articles.

[49] Zhiqiang Hu, Changhua Yu, Mabel Furutsuki, Gaia Andreoletti, Melissa Ly, Roger Hoskins, Aashish N. Adhikari, and Steven E. Brenner. VIPdb, a genetic Variant Impact Predictor Database. *Human Mutation*, 40(9):1202–1214, 2019.

[50] Abhishek Niroula and Mauno Vihinen. How good are pathogenicity predictors in detecting benign variants? *PLOS Computational Biology*, 15(2):e1006481, February 2019.

[51] Corinna Ernst, Eric Hahnen, Christoph Engel, Michael Nothnagel, Jonas Weber, Rita K. Schmutzler, and Jan Hauke. Performance of in silico prediction tools for the classification of rare BRCA1/2 missense variants in clinical diagnostics. *BMC Medical Genomics*, 11, March 2018.

[52] Hsinyi Tsang, KanakaDurga Addepalli, and Sean R. Davis. Resources for Interpreting Variants in Precision Genomic Oncology Applications. *Frontiers in Oncology*, 7, 2017. Publisher: Frontiers.

[53] Graham Rs Ritchie and Paul Flicek. Computational approaches to interpreting genomic sequence variation. *Genome Medicine*, 6(10):87, 2014.

[54] Guy Froyen, Marie Le Mercier, Els Lierman, Karl Vandepoele, Friedel Nollet, Elke Boone, Joni Van der Meulen, Koen Jacobs, Suzan Lambin, Sara Vander Borght, Els Van Valckenborgh, Aline Antoniou, and Aline Hébrant. Standardization of Somatic Variant Classifications in Solid and Haematological Tumours by a Two-Level Approach of Biological and Clinical Classes: An Initiative of the Belgian ComPerMed Expert Panel. *Cancers*, 11(12):2030, December 2019. Number: 12 Publisher: Multidisciplinary Digital Publishing Institute.

[55] J. Mateo, D. Chakravarty, R. Dienstmann, S. Jezdic, A. Gonzalez-Perez, N. Lopez-Bigas, C. K. Y. Ng, P. L. Bedard, G. Tortora, J.-Y. Douillard, E. M. Van Allen, N. Schultz, C. Swanton, F. André, and L. Pusztai. A framework to rank genomic alterations as targets for cancer precision medicine: the ESMO Scale for Clinical Actionability of molecular Targets (ESCAT). *Annals of Oncology: Official Journal of the European Society for Medical Oncology*, 29(9):1895–1902, 2018.

[56] Mahadeo A. Sukhai, Kenneth J. Craddock, Mariam Thomas, Aaron R. Hansen, Tong Zhang, Lillian Siu, Philippe Bedard, Tracy L. Stockley, and Suzanne Kamel-Reid. A classification system for clinical relevance of somatic variants identified in molecular profiling of cancer. *Genetics in Medicine*, 18(2):128–136, February 2016. Number: 2 Publisher: Nature Publishing Group.

[57] Annie Niehaus, Danielle R. Azzariti, Steven M. Harrison, Marina T. DiStefano, Sarah E. Hemphill, Ozlem Senol-Cosar, and Heidi L. Rehm. A survey assessing adoption of the ACMG-AMP guidelines

for interpreting sequence variants and identification of areas for continued improvement. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, 21(8):1699–1701, 2019.

[58] Cristina Fortuno, Kristy Lee, Magali Olivier, Tina Pesaran, Phuong L. Mai, Kelvin C. de Andrade, Laura D. Attardi, Stephanie Crowley, D. Gareth Evans, Bing-Jian Feng, Ann Katherine Major Foreman, Megan N. Frone, Robert Huether, Paul A. James, Kelly McGoldrick, Jessica Mester, Bryce A. Seifert, Thomas P. Slavin, Leora Witkowski, Liying Zhang, Sharon E. Plon, Amanda B. Spurdle, and Sharon A. Savage. Specifications of the ACMG/AMP variant interpretation guidelines for germline TP53 variants. *medRxiv*, page 2020.04.25.20078931, May 2020. Publisher: Cold Spring Harbor Laboratory Press.

[59] Jessica L. Mester, Rajarshi Ghosh, Tina Pesaran, Robert Huether, Rachid Karam, Kathleen S. Hruska, Helio A. Costa, Katherine Lachlan, Joanne Ngeow, Jill Barnholtz-Sloan, Kaitlin Sesock, Felicia Hernandez, Liying Zhang, Laura Milko, Sharon E. Plon, Madhuri Hegde, and Charis Eng. Gene-specific criteria for PTEN variant curation: Recommendations from the ClinGen PTEN Expert Panel. *Human Mutation*, 39(11):1581–1592, 2018.

[60] Kristy Lee, Kate Krempely, Maegan E. Roberts, Michael J. Anderson, Fatima Carneiro, Elizabeth Chao, Katherine Dixon, Joana Figueiredo, Rajarshi Ghosh, David Huntsman, Pardeep Kaurah, Chimene Kesserwan, Tyler Landrith, Shuwei Li, Arjen R. Mensenkamp, Carla Oliveira, Carolina Pardo, Tina Pesaran, Matthew Richardson, Thomas P. Slavin, Amanda B. Spurdle, Mackenzie Trapp, Leora Witkowski, Charles S. Yi, Liying Zhang, Sharon E. Plon, Kasmintan A. Schrader, and Rachid Karam. Specifications of the ACMG/AMP variant curation guidelines for the analysis of germline CDH1 sequence variants. *Human Mutation*, 39(11):1553–1568, 2018.

[61] Xi Luo, Simone Feurstein, Shruthi Mohan, Christopher C. Porter, Sarah A. Jackson, Sioban Keel, Michael Chicka, Anna L. Brown, Chimene Kesserwan, Anupriya Agarwal, Minjie Luo, Zejuan Li, Justyne E. Ross, Panagiotis Baliakas, Daniel Pineda-Alvarez, Courtney D. DiNardo, Alison A. Bertuch, Nikita Mehta, Tom Vulliamy, Ying Wang, Kim E. Nichols, Luca Malcovati, Michael F. Walsh, Lesley H. Rawlings, Shannon K. McWeeney, Jean Soulier, Anna Raimbault, Mark J. Routbort, Liying Zhang, Gabriella Ryan, Nancy A. Speck, Sharon E. Plon, David Wu, and Lucy A. Godley. ClinGen Myeloid Malignancy Variant Curation Expert Panel recommendations for germline RUNX1 variants. *Blood Advances*, 3(20):2962–2979, 2019.

[62] Michael F. Walsh, Deborah I. Ritter, Chimene Kesserwan, Dmitriy Sonkin, Debyani Chakravarty, Elizabeth Chao, Rajarshi Ghosh, Yelena Kemel, Gang Wu, Kristy Lee, Shashikant Kulkarni, Dale Hedges, Diana Mandelker, Ozge Ceyhan-Birsoy, Minjie Luo, Michael Drazer, Liying Zhang, Kenneth Offit, and Sharon E. Plon. Integrating somatic variant data and biomarkers for germline variant classification in cancer predisposition genes. *Human Mutation*, 39(11):1542–1552, 2018.

[63] Jeffrey S. Bennett, Madison Bernhardt, Kim L. McBride, Shalini C. Reshmi, Erik Zmuda, Naomi J. Kertesz, Vidu Garg, Sara Fitzgerald-Butt, and Anna N. Kamp. Reclassification of Variants of Uncertain Significance in Children with Inherited Arrhythmia Syndromes is Predicted by Clinical Factors. *Pediatric Cardiology*, 40(8):1679–1687, December 2019.

[64] Morales Ana and Hershberger Ray E. Variants of Uncertain Significance. *Circulation: Genomic and Precision Medicine*, 11(6):e002169, June 2018. Publisher: American Heart Association.

[65] Laura Palomo, Mariam Ibáñez, María Abáigar, Iria Vázquez, Sara Álvarez, Marta Cabezón, Bárbara Tazón-Vega, Inmaculada Ra-

pado, Francisco Fuster-Tormo, José Cervera, Rocío Benito, María J. Larrayoz, Juan C. Cigudosa, Lurdes Zamora, David Valcárcel, María T. Cedena, Pamela Acha, Jesús M. Hernández-Sánchez, Marta Fernández-Mercado, Guillermo Sanz, Jesús M. Hernández-Rivas, María J. Calasanz, Francesc Solé, Esperanza Such, and Spanish Group of MDS (GESMD). Spanish Guidelines for the use of targeted deep sequencing in myelodysplastic syndromes and chronic myelomonocytic leukaemia. *British Journal of Haematology*, 188(5):605–622, March 2020.

[66] Peter Horak, Malachi Griffith, Arpad Danos, Beth A. Pitel, Subha Madhavan, Xuelu Liu, Jennifer Lee, Gordana Raca, Shirley Li, Alex H. Wagner, Shashikant Kulkarni, Obi L. Griffith, Debyani Chakravarty, and Dmitriy Sonkin. Abstract 5707: A standard operating procedure for the interpretation of oncogenicity/pathogenicity of somatic mutations. *Cancer Research*, 80(16 Supplement):5707–5707, August 2020. Publisher: American Association for Cancer Research Section: Tumor Biology.

[67] Lefteris Koumakis. Deep learning models in genomics; are we there yet? *Computational and Structural Biotechnology Journal*, June 2020.

[68] W. Nicholson Price. Big Data and Black-Box Medical Algorithms. *Science translational medicine*, 10(471), December 2018.

[69] Finale Doshi-Velez and Been Kim. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv:1702.08608 [cs, stat]*, March 2017. arXiv: 1702.08608.

[70] Alex John London. Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *The Hastings Center Report*, 49(1):15–21, January 2019.

[71] Eui Jin Hwang, Sunggyun Park, Kwang-Nam Jin, Jung Im Kim, So Young Choi, Jong Hyuk Lee, Jin Mo Goo, Jaehong Aum, Jae-Joon Yim, Julien G. Cohen, Gilbert R. Ferretti, Chang Min Park,

and DLAD Development and Evaluation Group. Development and Validation of a Deep Learning-Based Automated Detection Algorithm for Major Thoracic Diseases on Chest Radiographs. *JAMA network open*, 2(3):e191095, 2019.

[72] Jianxing He, Sally L. Baxter, Jie Xu, Jiming Xu, Xingtao Zhou, and Kang Zhang. The practical implementation of artificial intelligence technologies in medicine. *Nature Medicine*, 25(1):30–36, January 2019. Number: 1 Publisher: Nature Publishing Group.

[73] Jenna Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X. Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, Pilar N. Ossorio, Sonoo Thadaney-Israni, and Anna Goldenberg. Do no harm: a roadmap for responsible machine learning for health care. *Nature Medicine*, 25(9):1337–1340, September 2019. Number: 9 Publisher: Nature Publishing Group.

[74] Dominik G. Grimm, Chloé-Agathe Azencott, Fabian Aicheler, Udo Gieraths, Daniel G. MacArthur, Kaitlin E. Samocha, David N. Cooper, Peter D. Stenson, Mark J. Daly, Jordan W. Smoller, Laramie E. Duncan, and Karsten M. Borgwardt. The Evaluation of Tools Used to Predict the Impact of Missense Variants Is Hindered by Two Types of Circularity. *Human Mutation*, 36(5):513–523, 2015. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/humu.22768.

[75] Melissa J. Landrum, Shanmuga Chitipiralla, Garth R. Brown, Chao Chen, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Wonhee Jang, Kuljeet Kaur, Chunlei Liu, Vitaly Lyoshin, Zenith Maddipatla, Rama Maiti, Joseph Mitchell, Nuala O'Leary, George R. Riley, Wenyao Shi, George Zhou, Valerie Schneider, Donna Maglott, J. Bradley Holmes, and Brandi L. Kattman. ClinVar: improvements to accessing data. *Nucleic Acids Research*, 48(D1):D835–D844, January 2020. Publisher: Oxford Academic.

[76] Liacine Bouaoun, Dmitriy Sonkin, Maude Ardin, Monica Hollstein, Graham Byrnes, Jiri Zavadil, and Magali Olivier. TP53 Variations in Human Cancers: New Lessons from the IARC TP53 Database and Genomics Data. *Human Mutation*, 37(9):865–876, 2016.

[77] Jeremy L. Warner, Dmitry Dymshyts, Christian G. Reich, Michael J. Gurley, Harry Hochheiser, Zachary H. Moldwin, Rimma Belenkaya, Andrew E. Williams, and Peter C. Yang. HemOnc: A new standard vocabulary for chemotherapy regimen representation in the OMOP common data model. *Journal of Biomedical Informatics*, 96:103239, 2019.

[78] Malachi Griffith, Nicholas C Spies, Kilannin Krysiak, Joshua F McMichael, Adam C Coffman, Arpad M Danos, Benjamin J Ainscough, Cody A Ramirez, Damian T Rieke, Lynzey Kujan, Erica K Barnell, Alex H Wagner, Zachary L Skidmore, Amber Wollam, Connor J Liu, Martin R Jones, Rachel L Bilski, Robert Lesurf, Yan-Yang Feng, Nakul M Shah, Melika Bonakdar, Lee Trani, Matthew Matlock, Avinash Ramu, Katie M Campbell, Gregory C Spies, Aaron P Graubert, Karthik Gangavarapu, James M Eldred, David E Larson, Jason R Walker, Benjamin M Good, Chunlei Wu, Andrew I Su, Rodrigo Dienstmann, Adam A Margolin, David Tamborero, Nuria Lopez-Bigas, Steven J M Jones, Ron Bose, David H Spencer, Lukas D Wartman, Richard K Wilson, Elaine R Mardis, and Obi L Griffith. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nature genetics*, 49(2):170–174, January 2017.

[79] John G. Tate, Sally Bamford, Harry C. Jubb, Zbyslaw Sondka, David M. Beare, Nidhi Bindal, Harry Boutselakis, Charlotte G. Cole, Celestino Creatore, Elisabeth Dawson, Peter Fish, Bhavana Harsha, Charlie Hathaway, Steve C. Jupe, Chai Yin Kok, Kate Noble, Laura Ponting, Christopher C. Ramshaw, Claire E. Rye, Helen E. Speedy, Ray Stefancsik, Sam L. Thompson, Shicai Wang, Sari Ward, Peter J. Campbell, and Simon A. Forbes. COSMIC: the Catalogue Of Somatic

Mutations In Cancer. *Nucleic Acids Research*, 47(D1):D941–D947, January 2019. Publisher: Oxford Academic.

[80] David Tamborero, Carlota Rubio-Perez, Jordi Deu-Pons, Michael P. Schroeder, Ana Vivancos, Ana Rovira, Ignasi Tusquets, Joan Albanell, Jordi Rodon, Josep Tabernero, Carmen de Torres, Rodrigo Dienstmann, Abel Gonzalez-Perez, and Nuria Lopez-Bigas. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Medicine*, 10(1):25, 2018.

[81] Debyani Chakravarty, Jianjiong Gao, Sarah Phillips, Ritika Kundra, Hongxin Zhang, Jiaojiao Wang, Julia E. Rudolph, Rona Yaeger, Tara Soumerai, Moriah H. Nissan, Matthew T. Chang, Sarat Chandarlapaty, Tiffany A. Traina, Paul K. Paik, Alan L. Ho, Feras M. Hantash, Andrew Grupe, Shrujal S. Baxi, Margaret K. Callahan, Alexandra Snyder, Ping Chi, Daniel C. Danila, Mrinal Gounder, James J. Harding, Matthew D. Hellmann, Gopa Iyer, Yelena Y. Janjigian, Thomas Kaley, Douglas A. Levine, Maeve Lowery, Antonio Omuro, Michael A. Postow, Dana Rathkopf, Alexander N. Shoushtari, Neerav Shukla, Martin H. Voss, Ederlinda Paraiso, Ahmet Zehir, Michael F. Berger, Barry S. Taylor, Leonard B. Saltz, Gregory J. Riely, Marc Ladanyi, David M. Hyman, José Baselga, Paul Sabbatini, David B. Solit, and Nikolaus Schultz. OncoKB: A Precision Oncology Knowledge Base. *JCO Precision Oncology*, (1):1–16, May 2017. Publisher: American Society of Clinical Oncology.

[82] Sam Q. Sun, R. Jay Mashl, Sohini Sengupta, Adam D. Scott, Weihua Wang, Prag Batra, Liang-Bo Wang, Matthew A. Wyczalkowski, and Li Ding. Database of evidence for precision oncology portal. *Bioinformatics*, 34(24):4315–4317, December 2018. Publisher: Oxford Academic.

[83] Benjamin J. Ainscough, Malachi Griffith, Adam C Coffman, Alex H. Wagner, Jason Kunisaki, Mayank NK Choudhary, Joshua F. McMichael, Robert S. Fulton, Richard K. Wilson, Obi L. Griffith,

and Elaine R. Mardis. DoCM: a database of curated mutations in cancer. *Nature methods*, 13(10):806–807, September 2016.

[84] Sara E. Patterson, Cara M. Statz, Taofei Yin, and Susan M. Mockus. Utility of the JAX Clinical Knowledgebase in capture and assessment of complex genomic cancer data. *npj Precision Oncology*, 3(1):1–6, January 2019. Number: 1 Publisher: Nature Publishing Group.

[85] Linda Huang, Helen Fernandes, Hamid Zia, Peyman Tavassoli, Hanna Rennert, David Pisapia, Marcin Imielinski, Andrea Sboner, Mark A. Rubin, Michael Kluk, and Olivier Elemento. The cancer precision medicine knowledge base for structured clinical-grade mutations and interpretations. *Journal of the American Medical Informatics Association*, 24(3):513–519, May 2017. Publisher: Oxford Academic.

[86] Senthilkumar Damodaran, Jharna Miya, Esko Kautto, Eliot Zhu, Eric Samorodnitsky, Jharna Datta, Julie W. Reeser, and Sameek Roychowdhury. Cancer Driver Log (CanDL). *The Journal of Molecular Diagnostics : JMD*, 17(5):554–559, September 2015.

[87] Zhenyu Yue, Le Zhao, and Junfeng Xia. dbCPM: a manually curated database for exploring the cancer passenger mutations. *Briefings in Bioinformatics*, October 2018.

[88] Zhenyu Yue, Le Zhao, Na Cheng, Hua Yan, and Junfeng Xia. dbCID: a manually curated resource for exploring the driver indels in human cancer. *Briefings in Bioinformatics*, 20(5):1925–1933, 2019.

[89] Kyubum Lee, Chih-Hsuan Wei, and Zhiyong Lu. Recent advances of automated methods for searching and extracting genomic variant information from biomedical literature. *Briefings in Bioinformatics*.

[90] A. S. M. Ashique Mahmood, Shruti Rao, Peter McGarvey, Cathy Wu, Subha Madhavan, and K. Vijay-Shanker. eGARD: Extracting associations between genomic anomalies and drug responses from text. *PLOS ONE*, 12(12):e0189663, December 2017. Publisher: Public Library of Science.

[91] Elli Papaemmanuil, Moritz Gerstung, Luca Malcovati, Sudhir Tauro, Gunes Gundem, Peter Van Loo, Chris J. Yoon, Peter Ellis, David C. Wedge, Andrea Pellagatti, Adam Shlien, Michael John Groves, Simon A. Forbes, Keiran Raine, Jon Hinton, Laura J. Mudie, Stuart McLaren, Claire Hardy, Calli Latimer, Matteo G. Della Porta, Sarah O'Meara, Ilaria Ambaglio, Anna Galli, Adam P. Butler, Gunilla Walldin, Jon W. Teague, Lynn Quek, Alex Sternberg, Carlo Gambacorti-Passerini, Nicholas C. P. Cross, Anthony R. Green, Jacqueline Boultwood, Paresh Vyas, Eva Hellstrom-Lindberg, David Bowen, Mario Cazzola, Michael R. Stratton, and Peter J. Campbell. Clinical and biological implications of driver mutations in myelodysplastic syndromes. *Blood*, 122(22):3616–3627, November 2013.

[92] Jens G. Lohr, Petar Stojanov, Scott L. Carter, Peter Cruz-Gordillo, Michael S. Lawrence, Daniel Auclair, Carrie Sougnez, Birgit Knoechel, Joshua Gould, Gordon Saksena, Kristian Cibulskis, Aaron McKenna, Michael A. Chapman, Ravid Straussman, Joan Levy, Louise M. Perkins, Jonathan J. Keats, Steven E. Schumacher, Mara Rosenberg, Multiple Myeloma Research Consortium, Gad Getz, and Todd R. Golub. Widespread genetic heterogeneity in multiple myeloma: implications for targeted therapy. *Cancer Cell*, 25(1):91–101, January 2014.

[93] Luciano G. Martelotto, Charlotte Ky Ng, Maria R. De Filippo, Yan Zhang, Salvatore Piscuoglio, Raymond S. Lim, Ronglai Shen, Larry Norton, Jorge S. Reis-Filho, and Britta Weigelt. Benchmarking mutation effect prediction algorithms using functionally validated cancer-related missense mutations. *Genome Biology*, 15(10):484, October 2014.

[94] Hong Li, Shuixia Liu, Shuangying Wang, Quanlei Zeng, Yulan Chen, Ting Fang, Yi Zhang, Ying Zhou, Yu Zhang, Kaiyue Wang, Zhangwei Yan, Cuicui Qiang, Meng Xu, Xianghua Chai, Yuying Yuan, Ming Huang, Hongyun Zhang, and Yun Xiong. Cancer SIGVAR: A semi-automated interpretation tool for germline variants of hered-

itary cancer-related genes. *bioRxiv*, page 2020.04.15.042283, April 2020. Publisher: Cold Spring Harbor Laboratory Section: New Results.

[95] Samantha O. Perakis, Sabrina Weber, Qing Zhou, Ricarda Graf, Sabine Hojas, Jakob M. Riedl, Armin Gerger, Nadia Dandachi, Marija Balic, Gerald Hoefler, Ed Schuuring, Harry J. M. Groen, Jochen B. Geigl, Ellen Heitzer, and Michael R. Speicher. Comparison of three commercial decision support platforms for matching of next-generation sequencing results with therapies in patients with cancer. *ESMO Open*, 5(5):e000872, September 2020. Publisher: BMJ Publishing Group Limited Section: Original research.

[96] Jiale Xiang, Jiyun Yang, Lisha Chen, Qiang Chen, Haiyan Yang, Chengcheng Sun, Qing Zhou, and Zhiyu Peng. Reinterpretation of common pathogenic variants in ClinVar revealed a high proportion of downgrades. *Scientific Reports*, 10(1):331, January 2020. Number: 1 Publisher: Nature Publishing Group.

[97] Quan Li and Kai Wang. InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. *American Journal of Human Genetics*, 100(2):267–280, 2017.

[98] Adam D. Scott, Kuan-Lin Huang, Amila Weerasinghe, R. Jay Mashl, Qingsong Gao, Fernanda Martins Rodrigues, Matthew A. Wyczalkowski, and Li Ding. CharGer: clinical Characterization of Germline variants. *Bioinformatics (Oxford, England)*, 35(5):865–867, 2019.

[99] Nicola Whiffin, Roddy Walsh, Risha Govind, Matthew Edwards, Mian Ahmad, Xiaolei Zhang, Upasana Tayal, Rachel Buchan, William Midwinter, Alicja E. Wilk, Hanna Najgebauer, Catherine Francis, Sam Wilkinson, Thomas Monk, Laura Brett, Declan P. O'Regan, Sanjay K. Prasad, Deborah J. Morris-Rosendahl, Paul J. R. Barton, Elizabeth Edwards, James S. Ware, and Stuart A. Cook. CardioClassifier: disease- and gene-specific computational decision support for clinical genome interpretation. *Genetics in Medicine*,

20(10):1246–1254, October 2018. Number: 10 Publisher: Nature Publishing Group.

[100] Giovanna Nicora, Ivan Limongelli, Patrick Gambelli, Mirella Memmi, Alberto Malovini, Andrea Mazzanti, Carlo Napolitano, Silvia Priori, and Riccardo Bellazzi. CardioVAI: An automatic implementation of ACMG-AMP variant interpretation guidelines in the diagnosis of cardiovascular diseases. *Human Mutation*, 39(12):1835–1846, 2018.

[101] Alexandre Xavier, Rodney J. Scott, and Bente A. Talseth-Palmer. TAPES: A tool for assessment and prioritisation in exome studies. *PLOS Computational Biology*, 15(10):e1007453, October 2019. Publisher: Public Library of Science.

[102] Vignesh Ravichandran, Zarina Shameer, Yelena Kemel, Michael Walsh, Karen Cadoo, Steven Lipkin, Diana Mandelker, Liying Zhang, Zsofia Stadler, Mark Robson, Kenneth Offit, and Joseph Vijai. Toward automation of germline variant curation in clinical cancer genetics. *Genetics in Medicine*, 21(9):2116–2125, September 2019.

[103] Jiguang Peng, Jiale Xiang, Xiangqian Jin, Junhua Meng, Nana Song, Lisha Chen, Ahmad Abou Tayoun, and Zhiyu Peng. VIP-HL: Semi-automated ACMG/AMP variant interpretation platform for genetic hearing loss. *bioRxiv*, page 2020.08.10.243642, August 2020. Publisher: Cold Spring Harbor Laboratory Section: New Results.

[104] Volkan Okur and Wendy K. Chung. The impact of hereditary cancer gene panels on clinical care and lessons learned. *Cold Spring Harbor Molecular Case Studies*, 3(6), November 2017.

[105] Xiaoju Max Ma, Stephanie J. Yaung, Liu Xi, Christine Ju, John F. Palma, and Maximilian Schmid. Assessment of a highly curated somatic oncology mutation database to facilitate identification of clinically important variants in NGS results. *Journal of Clinical Oncology*, 37(15_suppl):e18086–e18086, May 2019. Publisher: American Society of Clinical Oncology.

[106] Qian Nie, Gregory Omerza, Harshpreet Chandok, Matthew Prego, Meng-Chang Hsiao, Bridgette Meyers, Andrew Hesse, Jasmina Uvalic, Melissa Soucy, Daniel Bergeron, Michael Peracchio, Shelbi Burns, Kevin Kelly, Shannon Rowe, Jens Rueter, and Honey V Reddi. Molecular profiling of gynecologic cancers for treatment and management of disease – demonstrating clinical significance using the AMP/ASCO/CAP guidelines for interpretation and reporting of somatic variants. *Cancer Genetics*, 242:25–34, April 2020.

[107] Bijal A. Parikh, Latisha Love-Gregory, Eric J. Duncavage, and Jonathan W. Heusel. Identification of challenges and a framework for implementation of the AMP/ASCO/CAP classification guidelines for reporting somatic variants. *Practical Laboratory Medicine*, page e00170, May 2020.

[108] David Tamborero, Rodrigo Dienstmann, Maan Haj Rachid, Jorrit Boekel, Richard Baird, Irene Brana, Luigi De Petris, Jeffrey Yachnin, Christophe Massard, Frans L. Opdam, Richard Schlenk, Claudio Vernieri, Elena Garralda, Michele Masucci, Xenia Villalobos, Elena Chavarria, Fabien Calvo, Stefan Fröhling, Alexander Eggermont, Giovanni Apolone, Emile E. Voest, Carlos Caldas, Josep Tabernero, Ingemar Ernberg, Jordi Rodon, and Janne Lehtiö. Support systems to guide clinical decision-making in precision oncology: The Cancer Core Europe Molecular Tumor Board Portal. *Nature Medicine*, 26(7):992–994, July 2020. Number: 7 Publisher: Nature Publishing Group.

[109] Arpad M. Danos, Kilannin Krysiak, Erica K. Barnell, Adam C. Coffman, Joshua F. McMichael, Susanna Kiwala, Nicholas C. Spies, Lana M. Sheta, Shahil P. Pema, Lynzey Kujan, Kaitlin A. Clark, Amber Z. Wollam, Shruti Rao, Deborah I. Ritter, Dmitriy Sonkin, Gordana Raca, Raymond H. Kim, Alex H. Wagner, Subha Madhavan, Malachi Griffith, and Obi L. Griffith. The CIViC knowledge model and standard operating procedures for curation and clinical

interpretation of variants in cancer. *bioRxiv*, page 700179, July 2019. Publisher: Cold Spring Harbor Laboratory Section: New Results.

[110] Sigve Nakken, Ghislain Fournous, Daniel Vodák, Lars Birger Aasheim, Ola Myklebost, and Eivind Hovig. Personal Cancer Genome Reporter: variant interpretation report for precision oncology. *Bioinformatics*, 34(10):1778–1780, May 2018. Publisher: Oxford Academic.

[111] Christian Wünsch, Henrik Banck, Carsten Müller-Tidow, and Martin Dugas. AMLVaran: a software approach to implement variant analysis of targeted NGS sequencing data in an oncological care setting. *BMC Medical Genomics*, 13(1):17, February 2020.

[112] Jeremy L. Warner, Ishaan Prasad, Makiah Bennett, Monica Arniella, Alicia Beeghly-Fadiel, Kenneth D. Mandl, and Gil Alterovitz. SMART Cancer Navigator: A Framework for Implementing ASCO Workshop Recommendations to Enable Precision Cancer Medicine. *JCO precision oncology*, 2018, 2018.

[113] Max M. He, Quan Li, Muqing Yan, Hui Cao, Yue Hu, Karen Y. He, Kajia Cao, Marilyn M. Li, and Kai Wang. Variant Interpretation for Cancer (VIC): a computational tool for assessing clinical impacts of somatic variants. *Genome Medicine*, 11(1):53, August 2019.

[114] Quan Xu, Jin-Cheng Zhai, Cai-Qin Huo, Yang Li, Xue-Jiao Dong, Dong-Fang Li, Ru-Dan Huang, Chuang Shen, Yu-Jun Chang, Xi-Ling Zeng, Fan-Lin Meng, Fang Yang, Wan-Ling Zhang, Sheng-Nan Zhang, Yi-Ming Zhou, and Zhi Zhang. OncoPDSS: an evidence-based clinical decision support system for oncology pharmacotherapy at the individual level. *BMC Cancer*, 20(1):740, August 2020.

[115] Carmen Lai, Anjali D. Zimmer, Robert O'Connor, Serra Kim, Ray Chan, Jeroen van den Akker, Alicia Y. Zhou, Scott Topper, and Gilad Mishne. LEAP: Using machine learning to support variant classification in a clinical setting. *Human Mutation*, 41(6):1079–1090, June 2020.

[116] Najmeh Alirezaie, Kristin D. Kernohan, Taila Hartley, Jacek Majewski, and Toby Dylan Hocking. ClinPred: Prediction Tool to Identify Disease-Relevant Nonsynonymous Single-Nucleotide Variants. *The American Journal of Human Genetics*, 103(4):474–483, October 2018.

[117] Kirsley Chennen, Thomas Weber, Xavière Lornage, Arnaud Kress, Johann Böhm, Julie Thompson, Jocelyn Laporte, and Olivier Poch. MISTIC: A prediction tool to reveal disease-relevant deleterious missense variants. *PLOS ONE*, 15(7):e0236962, July 2020. Publisher: Public Library of Science.

[118] Priscilla Machado do Nascimento, Inácio Gomes Medeiros, Raul Maia Falcão, Beatriz Stransky, and Jorge Estefano Santana de Souza. A decision tree to improve identification of pathogenic mutations in clinical practice. *BMC Medical Informatics and Decision Making*, 20(1):52, December 2020.

[119] Marwa S. Hassan, A. A. Shaalan, M. I. Dessouky, Abdelaziz E. Abdelnaiem, and Mahmoud ElHefnawi. A review study: Computational techniques for expecting the impact of non-synonymous single nucleotide variants in human diseases. *Gene*, 680:20–33, January 2019.

[120] Yi Han, Juze Yang, Xinyi Qian, Wei-Chung Cheng, Shu-Hsuan Liu, Xing Hua, Liyuan Zhou, Yaning Yang, Qingbiao Wu, Pengyuan Liu, and Yan Lu. DriverML: a machine learning algorithm for identifying driver genes in cancer sequencing studies. *Nucleic Acids Research*, 47(8):e45, May 2019.

[121] Jack P. Hou and Jian Ma. DawnRank: discovering personalized driver genes in cancer. *Genome Medicine*, 6(7):56, 2014.

[122] Runjun D. Kumar, Adam C. Searleman, S. Joshua Swamidass, Obi L. Griffith, and Ron Bose. Statistically identifying tumor suppressors and oncogenes from pan-cancer genome-sequencing data. *Bioinformatics (Oxford, England)*, 31(22):3561–3568, November 2015.

[123] Francisco Martínez-Jiménez, Ferran Muiños, Inés Sentís, Jordi Deu-Pons, Iker Reyes-Salazar, Claudia Arnedo-Pac, Loris Mularoni, Oriol Pich, Jose Bonet, Hanna Kranas, Abel Gonzalez-Perez, and Nuria Lopez-Bigas. A compendium of mutational cancer driver genes. *Nature Reviews Cancer*, pages 1–18, August 2020. Publisher: Nature Publishing Group.

[124] Anna-Leigh Brown, Minghui Li, Alexander Goncearenco, and Anna R. Panchenko. Finding driver mutations in cancer: Elucidating the role of background mutational processes. *PLOS Computational Biology*, 15(4):e1006981, April 2019. Publisher: Public Library of Science.

[125] Yao Fu, Zhu Liu, Shaoke Lou, Jason Bedford, Xinmeng Jasmine Mu, Kevin Y. Yip, Ekta Khurana, and Mark Gerstein. FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biology*, 15(10):480, October 2014.

[126] Hashem A. Shihab, Julian Gough, David N. Cooper, Ian N. M. Day, and Tom R. Gaunt. Predicting the functional consequences of cancer-associated amino acid substitutions. *Bioinformatics (Oxford, England)*, 29(12):1504–1510, June 2013.

[127] Matthew H. Bailey, Collin Tokheim, Eduard Porta-Pardo, Sohini Sengupta, Denis Bertrand, Amila Weerasinghe, Antonio Colaprico, Michael C. Wendl, Jaegil Kim, Brendan Reardon, Patrick Kwok-Shing Ng, Kang Jin Jeong, Song Cao, Zixing Wang, Jianjiong Gao, Qingsong Gao, Fang Wang, Eric Minwei Liu, Loris Mularoni, Carlota Rubio-Perez, Niranjan Nagarajan, Isidro Cortés-Ciriano, Daniel Cui Zhou, Wen-Wei Liang, Julian M. Hess, Venkata D. Yellapantula, David Tamborero, Abel Gonzalez-Perez, Chayaporn Suphavilai, Jia Yu Ko, Ekta Khurana, Peter J. Park, Eliezer M. Van Allen, Han Liang, Michael S. Lawrence, Adam Godzik, Nuria Lopez-Bigas, Josh Stuart, David Wheeler, Gad Getz, Ken Chen, Alexander J. Lazar, Gordon B. Mills, Rachel Karchin, Li Ding, Samantha J.

Caesar-Johnson, John A. Demchok, Ina Felau, Melpomeni Kasapi, Martin L. Ferguson, Carolyn M. Hutter, Heidi J. Sofia, Roy Tarnuzzer, Zhining Wang, Liming Yang, Jean C. Zenklusen, Jiashan (Julia) Zhang, Sudha Chudamani, Jia Liu, Laxmi Lolla, Rashi Naresh, Todd Pihl, Qiang Sun, Yunhu Wan, Ye Wu, Juok Cho, Timothy DeFreitas, Scott Frazer, Nils Gehlenborg, Gad Getz, David I. Heiman, Jaegil Kim, Michael S. Lawrence, Pei Lin, Sam Meier, Michael S. Noble, Gordon Saksena, Doug Voet, Hailei Zhang, Brady Bernard, Nyasha Chambwe, Varsha Dhankani, Theo Knijnenburg, Roger Kramer, Kalle Leinonen, Yuexin Liu, Michael Miller, Sheila Reynolds, Ilya Shmulevich, Vesteinn Thorsson, Wei Zhang, Rehan Akbani, Bradley M. Broom, Apurva M. Hegde, Zhenlin Ju, Rupa S. Kanchi, Anil Korkut, Jun Li, Han Liang, Shiyun Ling, Wenbin Liu, Yiling Lu, Gordon B. Mills, Kwok-Shing Ng, Arvind Rao, Michael Ryan, Jing Wang, John N. Weinstein, Jiexin Zhang, Adam Abeshouse, Joshua Armenia, Debyani Chakravarty, Walid K. Chatila, Ino de Bruijn, Jianjiong Gao, Benjamin E. Gross, Zachary J. Heins, Ritika Kundra, Konnor La, Marc Ladanyi, Augustin Luna, Moriah G. Nissan, Angelica Ochoa, Sarah M. Phillips, Ed Reznik, Francisco Sanchez-Vega, Chris Sander, Nikolaus Schultz, Robert Sheridan, S. Onur Sumer, Yichao Sun, Barry S. Taylor, Jioajiao Wang, Hongxin Zhang, Pavana Anur, Myron Peto, Paul Spellman, Christopher Benz, Joshua M. Stuart, Christopher K. Wong, Christina Yau, D. Neil Hayes, Joel S. Parker, Matthew D. Wilkerson, Adrian Ally, Miruna Balasundaram, Reanne Bowlby, Denise Brooks, Rebecca Carlsen, Eric Chuah, Noreen Dhalla, Robert Holt, Steven J.M. Jones, Katayoon Kasaian, Darlene Lee, Yussanne Ma, Marco A. Marra, Michael Mayo, Richard A. Moore, Andrew J. Mungall, Karen Mungall, A. Gordon Robertson, Sara Sadeghi, Jacqueline E. Schein, Payal Sipahimalani, Angela Tam, Nina Thiessen, Kane Tse, Tina Wong, Ashton C. Berger, Rameen Beroukhim, Andrew D. Cherniack, Carrie Cibulskis, Stacey B. Gabriel, Galen F. Gao, Gavin Ha, Matthew Meyerson, Steven E. Schumacher, Juliann Shih, Melanie H. Kucherlapati, Raju S. Kucherlapati, Stephen Baylin, Leslie Cope, Ludmila

Danilova, Moiz S. Bootwalla, Phillip H. Lai, Dennis T. Maglinte, David J. Van Den Berg, Daniel J. Weisenberger, J. Todd Auman, Saianand Balu, Tom Bodenheimer, Cheng Fan, Katherine A. Hoadley, Alan P. Hoyle, Stuart R. Jefferys, Corbin D. Jones, Shaowu Meng, Piotr A. Mieczkowski, Lisle E. Mose, Amy H. Perou, Charles M. Perou, Jeffrey Roach, Yan Shi, Janae V. Simons, Tara Skelly, Matthew G. Soloway, Donghui Tan, Umadevi Veluvolu, Huihui Fan, Toshinori Hinoue, Peter W. Laird, Hui Shen, Wanding Zhou, Michelle Bellair, Kyle Chang, Kyle Covington, Chad J. Creighton, Huyen Dinh, HarshaVardhan Doddapaneni, Lawrence A. Donehower, Jennifer Drummond, Richard A. Gibbs, Robert Glenn, Walker Hale, Yi Han, Jianhong Hu, Viktoriya Korchina, Sandra Lee, Lora Lewis, Wei Li, Xiuping Liu, Margaret Morgan, Donna Morton, Donna Muzny, Jireh Santibanez, Margi Sheth, Eve Shinbrot, Linghua Wang, Min Wang, David A. Wheeler, Liu Xi, Fengmei Zhao, Julian Hess, Elizabeth L. Appelbaum, Matthew Bailey, Matthew G. Cordes, Li Ding, Catrina C. Fronick, Lucinda A. Fulton, Robert S. Fulton, Cyriac Kandoth, Elaine R. Mardis, Michael D. McLellan, Christopher A. Miller, Heather K. Schmidt, Richard K. Wilson, Daniel Crain, Erin Curley, Johanna Gardner, Kevin Lau, David Mallery, Scott Morris, Joseph Paulauskis, Robert Penny, Candace Shelton, Troy Shelton, Mark Sherman, Eric Thompson, Peggy Yena, Jay Bowen, Julie M. Gastier-Foster, Mark Gerken, Kristen M. Leraas, Tara M. Lichtenberg, Nilsa C. Ramirez, Lisa Wise, Erik Zmuda, Niall Corcoran, Tony Costello, Christopher Hovens, Andre L. Carvalho, Ana C. de Carvalho, José H. Fregnani, Adhemar Longatto-Filho, Rui M. Reis, Cristovam Scapulatempo-Neto, Henrique C.S. Silveira, Daniel O. Vidal, Andrew Burnette, Jennifer Eschbacher, Beth Hermes, Ardene Noss, Rosy Singh, Matthew L. Anderson, Patricia D. Castro, Michael Ittmann, David Huntsman, Bernard Kohl, Xuan Le, Richard Thorp, Chris Andry, Elizabeth R. Duffy, Vladimir Lyadov, Oxana Paklina, Galiya Setdikova, Alexey Shabunin, Mikhail Tavobilov, Christopher McPherson, Ronald Warnick, Ross Berkowitz, Daniel Cramer, Colleen Feltmate, Neil Horowitz, Adam Kibel, Michael Muto, Chan-

drajit P. Raut, Andrei Malykh, Jill S. Barnholtz-Sloan, Wendi Barrett, Karen Devine, Jordonna Fulop, Quinn T. Ostrom, Kristen Shimmel, Yingli Wolinsky, Andrew E. Sloan, Agostino De Rose, Felice Giuliante, Marc Goodman, Beth Y. Karlan, Curt H. Hagedorn, John Eckman, Jodi Harr, Jerome Myers, Kelinda Tucker, Leigh Anne Zach, Brenda Deyarmin, Hai Hu, Leonid Kvecher, Caroline Larson, Richard J. Mural, Stella Somiari, Ales Vicha, Tomas Zelinka, Joseph Bennett, Mary Iacocca, Brenda Rabeno, Patricia Swanson, Mathieu Latour, Louis Lacombe, Bernard Têtu, Alain Bergeron, Mary McGraw, Susan M. Staugaitis, John Chabot, Hanina Hibshoosh, Antonia Sepulveda, Tao Su, Timothy Wang, Olga Potapova, Olga Voronina, Laurence Desjardins, Odette Mariani, Sergio Roman-Roman, Xavier Sastre, Marc-Henri Stern, Feixiong Cheng, Sabina Signoretti, Andrew Berchuck, Darell Bigner, Eric Lipp, Jeffrey Marks, Shannon McCall, Roger McLendon, Angeles Secord, Alexis Sharp, Madhusmita Behera, Daniel J. Brat, Amy Chen, Keith Delman, Seth Force, Fadlo Khuri, Kelly Magliocca, Shishir Maithel, Jeffrey J. Olson, Taofeek Owonikoko, Alan Pickens, Suresh Ramalingam, Dong M. Shin, Gabriel Sica, Erwin G. Van Meir, Hongzheng Zhang, Wil Eijckenboom, Ad Gillis, Esther Korpershoek, Leendert Looijenga, Wolter Oosterhuis, Hans Stoop, Kim E. van Kessel, Ellen C. Zwarthoff, Chiara Calatozzolo, Lucia Cuppini, Stefania Cuzzubbo, Francesco DiMeco, Gaetano Finocchiaro, Luca Mattei, Alessandro Perin, Bianca Pollo, Chu Chen, John Houck, Pawadee Lohavanichbutr, Arndt Hartmann, Christine Stoehr, Robert Stoehr, Helge Taubert, Sven Wach, Bernd Wullich, Witold Kycler, Dawid Murawa, Maciej Wiznerowicz, Ki Chung, W. Jeffrey Edenfield, Julie Martin, Eric Baudin, Glenn Bubley, Raphael Bueno, Assunta De Rienzo, William G. Richards, Steven Kalkanis, Tom Mikkelsen, Houtan Noushmehr, Lisa Scarpace, Nicolas Girard, Marta Aymerich, Elias Campo, Eva Giné, Armando López Guillermo, Nguyen Van Bang, Phan Thi Hanh, Bui Duc Phu, Yufang Tang, Howard Colman, Kimberley Evason, Peter R. Dottino, John A. Martignetti, Hani Gabra, Hartmut Juhl, Teniola Akeredolu, Serghei Stepa, Dave Hoon,

Keunsoo Ahn, Koo Jeong Kang, Felix Beuschlein, Anne Breggia, Michael Birrer, Debra Bell, Mitesh Borad, Alan H. Bryce, Erik Castle, Vishal Chandan, John Cheville, John A. Copland, Michael Farnell, Thomas Flotte, Nasra Giama, Thai Ho, Michael Kendrick, Jean-Pierre Kocher, Karla Kopp, Catherine Moser, David Nagorney, Daniel O'Brien, Brian Patrick O'Neill, Tushar Patel, Gloria Petersen, Florencia Que, Michael Rivera, Lewis Roberts, Robert Smallridge, Thomas Smyrk, Melissa Stanton, R. Houston Thompson, Michael Torbenson, Ju Dong Yang, Lizhi Zhang, Fadi Brimo, Jaffer A. Ajani, Ana Maria Angulo Gonzalez, Carmen Behrens, Jolanta Bondaruk, Russell Broaddus, Bogdan Czerniak, Bita Esmaeli, Junya Fujimoto, Jeffrey Gershenwald, Charles Guo, Alexander J. Lazar, Christopher Logothetis, Funda Meric-Bernstam, Cesar Moran, Lois Ramondetta, David Rice, Anil Sood, Pheroze Tamboli, Timothy Thompson, Patricia Troncoso, Anne Tsao, Ignacio Wistuba, Candace Carter, Lauren Haydu, Peter Hersey, Valerie Jakrot, Hojabr Kakavand, Richard Kefford, Kenneth Lee, Georgina Long, Graham Mann, Michael Quinn, Robyn Saw, Richard Scolyer, Kerwin Shannon, Andrew Spillane, Jonathan Stretch, Maria Synott, John Thompson, James Wilmott, Hikmat Al-Ahmadie, Timothy A. Chan, Ronald Ghossein, Anuradha Gopalan, Douglas A. Levine, Victor Reuter, Samuel Singer, Bhuvanesh Singh, Nguyen Viet Tien, Thomas Broudy, Cyrus Mirsaidi, Praveen Nair, Paul Drwiega, Judy Miller, Jennifer Smith, Howard Zaren, Joong-Won Park, Nguyen Phi Hung, Electron Kebebew, W. Marston Linehan, Adam R. Metwalli, Karel Pacak, Peter A. Pinto, Mark Schiffman, Laura S. Schmidt, Cathy D. Vocke, Nicolas Wentzensen, Robert Worrell, Hannah Yang, Marc Moncrieff, Chandra Goparaju, Jonathan Melamed, Harvey Pass, Natalia Botnariuc, Irina Caraman, Mircea Cernat, Inga Chemencedji, Adrian Clipca, Serghei Doruc, Ghenadie Gorincioi, Sergiu Mura, Maria Pirtac, Irina Stancul, Diana Tcaciuc, Monique Albert, Iakovina Alexopoulou, Angel Arnaout, John Bartlett, Jay Engel, Sebastien Gilbert, Jeremy Parfitt, Harman Sekhon, George Thomas, Doris M. Rassl, Robert C. Rintoul, Carlo Bifulco, Raina Tamakawa, Walter Urba, Nicholas Hayward,

Henri Timmers, Anna Antenucci, Francesco Facciolo, Gianluca Grazi, Mirella Marino, Roberta Merola, Ronald de Krijger, Anne-Paule Gimenez-Roqueplo, Alain Piché, Simone Chevalier, Ginette McKercher, Kivanc Birsoy, Gene Barnett, Cathy Brewer, Carol Farver, Theresa Naska, Nathan A. Pennell, Daniel Raymond, Cathy Schilero, Kathy Smolenski, Felicia Williams, Carl Morrison, Jeffrey A. Borgia, Michael J. Liptay, Mark Pool, Christopher W. Seder, Kerstin Junker, Larsson Omberg, Mikhail Dinkin, George Manikhas, Domenico Alvaro, Maria Consiglia Bragazzi, Vincenzo Cardinale, Guido Carpino, Eugenio Gaudio, David Chesla, Sandra Cottingham, Michael Dubina, Fedor Moiseenko, Renumathy Dhanasekaran, Karl-Friedrich Becker, Klaus-Peter Janssen, Julia Slotta-Huspenina, Mohamed H. Abdel-Rahman, Dina Aziz, Sue Bell, Colleen M. Cebulla, Amy Davis, Rebecca Duell, J. Bradley Elder, Joe Hilty, Bahavna Kumar, James Lang, Norman L. Lehman, Randy Mandt, Phuong Nguyen, Robert Pilarski, Karan Rai, Lynn Schoenfield, Kelly Senecal, Paul Wakely, Paul Hansen, Ronald Lechan, James Powers, Arthur Tischler, William E. Grizzle, Katherine C. Sexton, Alison Kastl, Joel Henderson, Sima Porten, Jens Waldmann, Martin Fassnacht, Sylvia L. Asa, Dirk Schadendorf, Marta Couce, Markus Graefen, Hartwig Huland, Guido Sauter, Thorsten Schlomm, Ronald Simon, Pierre Tennstedt, Oluwole Olabode, Mark Nelson, Oliver Bathe, Peter R. Carroll, June M. Chan, Philip Disaia, Pat Glenn, Robin K. Kelley, Charles N. Landen, Joanna Phillips, Michael Prados, Jeffry Simko, Karen Smith-McCune, Scott VandenBerg, Kevin Roggin, Ashley Fehrenbach, Ady Kendler, Suzanne Sifri, Ruth Steele, Antonio Jimeno, Francis Carey, Ian Forgie, Massimo Mannelli, Michael Carney, Brenda Hernandez, Benito Campos, Christel Herold-Mende, Christin Jungk, Andreas Unterberg, Andreas von Deimling, Aaron Bossler, Joseph Galbraith, Laura Jacobus, Michael Knudson, Tina Knutson, Deqin Ma, Mohammed Milhem, Rita Sigmund, Andrew K. Godwin, Rashna Madan, Howard G. Rosenthal, Clement Adebamowo, Sally N. Adebamowo, Alex Boussioutas, David Beer, Thomas Giordano, Anne-Marie Mes-Masson, Fred Saad, Therese Bocklage, Lisa

Landrum, Robert Mannel, Kathleen Moore, Katherine Moxley, Russel Postier, Joan Walker, Rosemary Zuna, Michael Feldman, Federico Valdivieso, Rajiv Dhir, James Luketich, Edna M. Mora Pinero, Mario Quintero-Aguilo, Carlos Gilberto Carlotti, Jose Sebastião Dos Santos, Rafael Kemp, Ajith Sankarankuty, Daniela Tirapelli, James Catto, Kathy Agnew, Elizabeth Swisher, Jenette Creaney, Bruce Robinson, Carl Simon Shelley, Eryn M. Godwin, Sara Kendall, Cassaundra Shipman, Carol Bradford, Thomas Carey, Andrea Haddad, Jeffey Moyer, Lisa Peterson, Mark Prince, Laura Rozek, Gregory Wolf, Rayleen Bowman, Kwun M. Fong, Ian Yang, Robert Korst, W. Kimryn Rathmell, J. Leigh Fantacone-Campbell, Jeffrey A. Hooke, Albert J. Kovatich, Craig D. Shriver, John DiPersio, Bettina Drake, Ramaswamy Govindan, Sharon Heath, Timothy Ley, Brian Van Tine, Peter Westervelt, Mark A. Rubin, Jung Il Lee, Natália D. Aredes, and Armaz Mariamidze. Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell*, 173(2):371–385.e18, April 2018.

[128] Abel Gonzalez-Perez, Jordi Deu-Pons, and Nuria Lopez-Bigas. Improving the prediction of the functional impact of cancer mutations by baseline tolerance transformation. *Genome Medicine*, 4(11):89, November 2012.

[129] Xiaoming Liu, Xueqiu Jian, and Eric Boerwinkle. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Human Mutation*, 32(8):894–899, August 2011.

[130] Stephan Hutter, Constance Baer, Wencke Walter, Wolfgang Kern, Claudia Haferlach, and Torsten Haferlach. A Novel Machine Learning Based in silico Pathogenicity Predictor for Missense Variants in a Hematological Setting. *Blood*, 134(Supplement_1):2090–2090, November 2019. Publisher: American Society of Hematology.

[131] Collin Tokheim and Rachel Karchin. CHASMplus Reveals the Scope of Somatic Missense Mutations Driving Human Cancers. *Cell Systems*, 9(1):9–23.e8, July 2019.

[132] Mark F. Rogers, Tom R. Gaunt, and Colin Campbell. CScape-somatic: distinguishing driver and passenger point mutations in the cancer genome. *Bioinformatics*, 36(12):3637–3644, June 2020. Publisher: Oxford Academic.

[133] Yong Mao, Han Chen, Han Liang, Funda Meric-Bernstam, Gordon B. Mills, and Ken Chen. CanDrA: Cancer-Specific Driver Missense Mutation Annotation with Optimized Features. *PLOS ONE*, 8(10):e77945, October 2013. Publisher: Public Library of Science.

[134] Yahya Bokhari and Tomasz Arodz. QuaDMutEx: quadratic driver mutation explorer. *BMC Bioinformatics*, 18(1):458, October 2017.

[135] Hu Chen, Jun Li, Yumeng Wang, Patrick Kwok-Shing Ng, Yiu Huen Tsang, Kenna R. Shaw, Gordon B. Mills, and Han Liang. Comprehensive assessment of computational algorithms in predicting cancer driver mutations. *Genome Biology*, 21(1):43, February 2020.

[136] O. Chapelle, B. Schlkopf, and Zien. *Semi-Supervised Learning*. The MIT Press. Publication Title: Semi-Supervised Learning.

[137] Jacqueline Mersch, Nichole Brown, Sara Pirzadeh-Miller, Erin Mundt, Hannah C. Cox, Krystal Brown, Melissa Aston, Lisa Esterling, Susan Manley, and Theodora Ross. Prevalence of Variant Reclassification Following Hereditary Cancer Genetic Testing. *JAMA*, 320(12):1266–1274, September 2018. Publisher: American Medical Association.

[138] Lisa Esterling, Ranjula Wijayatunge, Krystal Brown, Brian Morris, Elisha Hughes, Dmitry Pruss, Susan Manley, Karla R. Bowles, and Theodora S. Ross. Impact of a Cancer Gene Variant Reclassification Program Over a 20-Year Period. *JCO Precision Oncology*, (4):944–954, August 2020. Publisher: American Society of Clinical Oncology.

[139] J. M. Eggington, K. R. Bowles, K. Moyes, S. Manley, L. Esterling, S. Sizemore, E. Rosenthal, A. Theisen, J. Saam, C. Arnell, D. Pruss, J. Bennett, L. A. Burbidge, B. Roa, and R. J.

Wenstrup. A comprehensive laboratory-based program for classification of variants of uncertain significance in hereditary cancer genes. *Clinical Genetics*, 86(3):229–237, 2014. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cge.12315.

[140] Chloe Mighton, George S. Charames, Marina Wang, Kathleen-Rose Zakoor, Andrew Wong, Salma Shickh, Nicholas Watkins, Matthew S. Lebo, Yvonne Bombard, and Jordan Lerner-Ellis. Variant classification changes over time in BRCA1 and BRCA2. *Genetics in Medicine*, 21(10):2248–2254, October 2019. Number: 10 Publisher: Nature Publishing Group.

[141] Abdelhamid Bouchachia, Bogdan Gabrys, and Zoheir Sahel. Overview of Some Incremental Learning Algorithms. In *2007 IEEE International Fuzzy Systems Conference*, pages 1–6, July 2007. ISSN: 1098-7584.

[142] Elizabeth M. Smigielski, Karl Sirotkin, Minghong Ward, and Stephen T. Sherry. dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Research*, 28(1):352–355, January 2000.

[143] William McLaren, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R. S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The Ensembl Variant Effect Predictor. *Genome Biology*, 17(1):122, June 2016.

[144] Helen V. Firth, Shola M. Richards, A. Paul Bevan, Stephen Clayton, Manuel Corpas, Diana Rajan, Steven Van Vooren, Yves Moreau, Roger M. Pettett, and Nigel P. Carter. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *The American Journal of Human Genetics*, 84(4):524–533, April 2009.

[145] Ivan A. Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E. Ramensky, Anna Gerasimova, Peer Bork, Alexey S. Kondrashov, and

Shamil R. Sunyaev. A method and server for predicting damaging missense mutations. *Nature Methods*, 7(4):248–249, April 2010.

[146] Margarida C. Lopes, Chris Joyce, Graham R. S. Ritchie, Sally L. John, Fiona Cunningham, Jennifer Asimit, and Eleftheria Zeggini. A combined functional annotation score for non-synonymous variants. *Human Heredity*, 73(1):47–51, 2012.

[147] Kaitlin E. Samocha, Jack A. Kosmicki, Konrad J. Karczewski, Anne H. O'Donnell-Luria, Emma Pierce-Hoffman, Daniel G. MacArthur, Benjamin M. Neale, and Mark J. Daly. Regional missense constraint improves variant deleteriousness prediction. *bioRxiv*, page 148353, June 2017. Publisher: Cold Spring Harbor Laboratory Section: New Results.

[148] Robert Vaser, Swarnaseetha Adusumalli, Sim Ngak Leng, Mile Sikic, and Pauline C. Ng. SIFT missense predictions for genomes. *Nature Protocols*, 11(1):1–9, January 2016. Number: 1 Publisher: Nature Publishing Group.

[149] João Fadista, Nikolay Oskolkov, Ola Hansson, and Leif Groop. LoFtool: a gene intolerance score based on loss-of-function variants in 60 706 individuals. *Bioinformatics (Oxford, England)*, 33(4):471–474, 2017.

[150] Gregory M. Cooper, Eric A. Stone, George Asimenos, Eric D. Green, Serafim Batzoglou, and Arend Sidow. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Research*, 15(7):901–913, July 2005.

[151] Katherine S. Pollard, Melissa J. Hubisz, Kate R. Rosenbloom, and Adam Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, 20(1):110–121, January 2010. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor

Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[152] Monkol Lek, Konrad J. Karczewski, Eric V. Minikel, Kaitlin E. Samocha, Eric Banks, Timothy Fennell, Anne H. O'Donnell-Luria, James S. Ware, Andrew J. Hill, Beryl B. Cummings, Taru Tukiainen, Daniel P. Birnbaum, Jack A. Kosmicki, Laramie E. Duncan, Karol Estrada, Fengmei Zhao, James Zou, Emma Pierce-Hoffman, Joanne Berghout, David N. Cooper, Nicole Deflaux, Mark DePristo, Ron Do, Jason Flannick, Menachem Fromer, Laura Gauthier, Jackie Goldstein, Namrata Gupta, Daniel Howrigan, Adam Kiezun, Mitja I. Kurki, Ami Levy Moonshine, Pradeep Natarajan, Lorena Orozco, Gina M. Peloso, Ryan Poplin, Manuel A. Rivas, Valentin Ruano-Rubio, Samuel A. Rose, Douglas M. Ruderfer, Khalid Shakir, Peter D. Stenson, Christine Stevens, Brett P. Thomas, Grace Tiao, Maria T. Tusie-Luna, Ben Weisburd, Hong-Hee Won, Dongmei Yu, David M. Altshuler, Diego Ardissino, Michael Boehnke, John Danesh, Stacey Donnelly, Roberto Elosua, Jose C. Florez, Stacey B. Gabriel, Gad Getz, Stephen J. Glatt, Christina M. Hultman, Sekar Kathiresan, Markku Laakso, Steven McCarroll, Mark I. McCarthy, Dermot McGovern, Ruth McPherson, Benjamin M. Neale, Aarno Palotie, Shaun M. Purcell, Danish Saleheen, Jeremiah M. Scharf, Pamela Sklar, Patrick F. Sullivan, Jaakko Tuomilehto, Ming T. Tsuang, Hugh C. Watkins, James G. Wilson, Mark J. Daly, and Daniel G. MacArthur. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285–291, August 2016. Number: 7616 Publisher: Nature Publishing Group.

[153] Xueqiu Jian, Eric Boerwinkle, and Xiaoming Liu. In silico prediction of splice-altering single nucleotide variants in the human genome. *Nucleic Acids Research*, 42(22):13534–13544, December 2014. Publisher: Oxford Academic.

[154] Martin Kircher, Daniela M. Witten, Preti Jain, Brian J. O'Roak, Gregory M. Cooper, and Jay Shendure. A general framework for es-

timating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46(3):310–315, March 2014.

[155] Hashem A. Shihab, Julian Gough, Matthew Mort, David N. Cooper, Ian N. M. Day, and Tom R. Gaunt. Ranking non-synonymous single nucleotide polymorphisms based on disease concepts. *Human Genomics*, 8:11, June 2014.

[156] Magali Olivier, Monica Hollstein, and Pierre Hainaut. TP53 Mutations in Human Cancers: Origins, Consequences, and Clinical Use. *Cold Spring Harbor Perspectives in Biology*, 2(1), January 2010.

[157] Haldun Akoglu. User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, 18(3):91–93, August 2018.

[158] Laura Toloşi and Thomas Lengauer. Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics*, 27(14):1986–1994, July 2011. Publisher: Oxford Academic.

[159] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1):37–52, August 1987.

[160] G. E. Hinton and R. R. Salakhutdinov. Reducing the Dimensionality of Data with Neural Networks. *Science*, 313(5786):504–507, July 2006. Publisher: American Association for the Advancement of Science Section: Report.

[161] Yasi Wang, Hongxun Yao, and Sicheng Zhao. Auto-encoder based dimensionality reduction. *Neurocomputing*, 184:232–242, April 2016.

[162] David Charte, Francisco Charte, María J. del Jesus, and Francisco Herrera. An analysis on the use of autoencoders for representation learning: Fundamentals, learning task case studies, explainability and challenges. *Neurocomputing*, 404:93–107, September 2020.

[163] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[164] Davide Chicco and Giuseppe Jurman. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21, January 2020.

[165] Takaya Saito and Marc Rehmsmeier. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS ONE*, 10(3), March 2015.

[166] Quan Zou, Sifa Xie, Ziyu Lin, Meihong Wu, and Ying Ju. Finding the Best Classification Threshold in Imbalanced Classification. *Big Data Research*, 5:2–8, September 2016.

[167] Max Schubach, Matteo Re, Peter N. Robinson, and Giorgio Valentini. Imbalance-Aware Machine Learning for Predicting Rare and Common Disease-Associated Non-Coding Variants. *Scientific Reports*, 7(1):2959, June 2017. Number: 1 Publisher: Nature Publishing Group.

[168] Bee Wah Yap, Khatijahhusna Abd Rani, Hezlin Aryani Abd Rahman, Simon Fong, Zuraida Khairudin, and Nik Nik Abdullah. An Application of Oversampling, Undersampling, Bagging and Boosting in Handling Imbalanced Datasets. In Tutut Herawan, Mustafa Mat Deris, and Jemal Abawajy, editors, *Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013)*, Lecture Notes in Electrical Engineering, pages 13–22, Singapore, 2014. Springer.

[169] Gilles Vandewiele, Isabelle Dehaene, Olivier Janssens, Femke Ongenae, Femke De Backere, Filip De Turck, Kristien Roelens, Sofie Van Hoecke, and Thomas Demeester. A Critical Look at Studies Applying Over-Sampling on the TPEHGDB Dataset. In David Riaño, Szymon Wilk, and Annette ten Teije, editors, *Artificial Intelligence in*

Medicine, Lecture Notes in Computer Science, pages 355–364, Cham, 2019. Springer International Publishing.

[170] Elli Papaemmanuil, Moritz Gerstung, Luca Malcovati, Sudhir Tauro, Gunes Gundem, Peter Van Loo, Chris J. Yoon, Peter Ellis, David C. Wedge, Andrea Pellagatti, Adam Shlien, Michael John Groves, Simon A. Forbes, Keiran Raine, Jon Hinton, Laura J. Mudie, Stuart McLaren, Claire Hardy, Calli Latimer, Matteo G. Della Porta, Sarah O'Meara, Ilaria Ambaglio, Anna Galli, Adam P. Butler, Gunilla Walldin, Jon W. Teague, Lynn Quek, Alex Sternberg, Carlo Gambacorti-Passerini, Nicholas C. P. Cross, Anthony R. Green, Jacqueline Boultwood, Paresh Vyas, Eva Hellstrom-Lindberg, David Bowen, Mario Cazzola, Michael R. Stratton, and Peter J. Campbell. Clinical and biological implications of driver mutations in myelodysplastic syndromes. *Blood*, 122(22):3616–3627, November 2013.

[171] W. James Kent, Charles W. Sugnet, Terrence S. Furey, Krishna M. Roskin, Tom H. Pringle, Alan M. Zahler, and and David Haussler. The Human Genome Browser at UCSC. *Genome Research*, 12(6):996–1006, January 2002. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[172] Elli Papaemmanuil, Moritz Gerstung, Luca Malcovati, Sudhir Tauro, Gunes Gundem, Peter Van Loo, Chris J. Yoon, Peter Ellis, David C. Wedge, Andrea Pellagatti, Adam Shlien, Michael John Groves, Simon A. Forbes, Keiran Raine, Jon Hinton, Laura J. Mudie, Stuart McLaren, Claire Hardy, Calli Latimer, Matteo G. Della Porta, Sarah O'Meara, Ilaria Ambaglio, Anna Galli, Adam P. Butler, Gunilla Walldin, Jon W. Teague, Lynn Quek, Alex Sternberg, Carlo Gambacorti-Passerini, Nicholas C. P. Cross, Anthony R. Green, Jacqueline Boultwood, Paresh Vyas, Eva Hellstrom-Lindberg, David Bowen, Mario Cazzola, Michael R. Stratton, Peter J. Campbell, and Chronic Myeloid Disorders Working Group of the International Can-

cer Genome Consortium. Clinical and biological implications of driver mutations in myelodysplastic syndromes. *Blood*, 122(22):3616–3627; quiz 3699, November 2013. Number: 22.

[173] Trishan Panch, Tom J. Pollard, Heather Mattie, Emily Lindemer, Pearse A. Keane, and Leo Anthony Celi. "Yes, but will it work for my patients?" Driving clinically relevant research with benchmark datasets. *npj Digital Medicine*, 3(1):1–4, June 2020. Number: 1 Publisher: Nature Publishing Group.

[174] Suchi Saria and Adarsh Subbaswamy. Tutorial: Safe and Reliable Machine Learning. April 2019.

[175] Peter Schulam and Suchi Saria. Can You Trust This Prediction? Auditing Pointwise Reliability After Learning. January 2019.

[176] Matjaz Kukar and Igor Kononenko. Reliable Classifications with Machine Learning. In *Proceedings of the 13th European Conference on Machine Learning*, ECML '02, pages 219–231, Berlin, Heidelberg, August 2002. Springer-Verlag.

[177] J. Arturo Olvera-López, J. Ariel Carrasco-Ochoa, J. Francisco Martínez-Trinidad, and Josef Kittler. A review of instance selection methods. *Artificial Intelligence Review*, 34(2):133–143, August 2010.

[178] José C. Riquelme, Jesús S. Aguilar-Ruiz, and Miguel Toro. Finding representative patterns with ordered projections. *Pattern Recognition*, 36(4):1009–1018, April 2003.

[179] Xing Ke and Lisong Shen. Molecular targeted therapy of cancer: The progress and future prospect. *Frontiers in Laboratory Medicine*, 1(2):69–75, June 2017.

[180] D. B. Longley and P. G. Johnston. Molecular mechanisms of drug resistance. *The Journal of Pathology*, 205(2):275–292, 2005. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/path.1706.

[181] Feifei Li, Changqi Zhao, and Lili Wang. Molecular-targeted agents combination therapy for cancer: Developments and potentials. *International Journal of Cancer*, 134(6):1257–1269, 2014. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ijc.28261.

[182] Hartmut Döhner, Elihu Estey, David Grimwade, Sergio Amadori, Frederick R. Appelbaum, Thomas Büchner, Hervé Dombret, Benjamin L. Ebert, Pierre Fenaux, Richard A. Larson, Ross L. Levine, Francesco Lo-Coco, Tomoki Naoe, Dietger Niederwieser, Gert J. Ossenkoppele, Miguel Sanz, Jorge Sierra, Martin S. Tallman, Hwei-Fang Tien, Andrew H. Wei, Bob Löwenberg, and Clara D. Bloomfield. Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. *Blood*, 129(4):424–447, January 2017. Publisher: American Society of Hematology.

[183] Nicola Carlomagno, Paola Incollingo, Vincenzo Tammaro, Gaia Peluso, Niccolò Rupealta, Gaetano Chiacchio, Maria Laura Sandoval Sotelo, Gianluca Minieri, Antonio Pisani, Eleonora Riccio, Massimo Sabbatini, Umberto Marcello Bracale, Armando Calogero, Concetta Anna Dodaro, and Michele Santangelo. Diagnostic, predictive, prognostic, and therapeutic molecular biomarkers in third millennium: A breakthrough in gastric cancer. 2017.

[184] Deepika Sirohi, Robert L. Schmidt, Dara L. Aisner, Amir Behdad, Bryan L. Betz, Noah Brown, Joshua F. Coleman, Christopher L. Corless, Georgios Deftereos, Mark D. Ewalt, Helen Fernandes, Susan J. Hsiao, Mahesh M. Mansukhani, Sarah S. Murray, Nifang Niu, Lauren L. Ritterhouse, Carlos J. Suarez, Laura J. Tafe, John A. Thorson, Jeremy P. Segal, and Larissa V. Furtado. Multi-Institutional Evaluation of Interrater Agreement of Variant Classification Based on the 2017 Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer. *The Journal of molecular diagnostics: JMD*, 22(2):284–293, February 2020.

[185] Shu-Hsien Liao. Expert system methodologies and applications—a decade review from 1995 to 2004. *Expert Systems with Applications*, 28(1):93–103, January 2005.

[186] Deborah I. Ritter, Sameek Roychowdhury, Angshumoy Roy, Shruti Rao, Melissa J. Landrum, Dmitriy Sonkin, Mamatha Shekar, Caleb F. Davis, Reece K. Hart, Christine Micheel, Meredith Weaver, Eliezer M. Van Allen, Donald W. Parsons, Howard L. McLeod, Michael S. Watson, Sharon E. Plon, Shashikant Kulkarni, Subha Madhavan, and ClinGen Somatic Cancer Working Group. Somatic cancer variant curation and harmonization through consensus minimum variant level data. *Genome Medicine*, 8(1):117, 2016.

[187] Matthew T. Chang, Saurabh Asthana, Sizhi Paul Gao, Byron H. Lee, Jocelyn S. Chapman, Cyriac Kandoth, JianJiong Gao, Nicholas D. Socci, David B. Solit, Adam B. Olshen, Nikolaus Schultz, and Barry S. Taylor. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nature Biotechnology*, 34(2):155–163, February 2016.

[188] Simona Soverini, Andreas Hochhaus, Franck E. Nicolini, Franz Gruber, Thoralf Lange, Giuseppe Saglio, Fabrizio Pane, Martin C. Müller, Thomas Ernst, Gianantonio Rosti, Kimmo Porkka, Michele Baccarani, Nicholas C. P. Cross, and Giovanni Martinelli. BCR-ABL kinase domain mutation analysis in chronic myeloid leukemia patients treated with tyrosine kinase inhibitors: recommendations from an expert panel on behalf of European LeukemiaNet. *Blood*, 118(5):1208–1215, August 2011.

[189] Aziz Nazha. The MDS genomics-prognosis symbiosis. *Hematology: the American Society of Hematology Education Program*, 2018(1):270–276, November 2018.

[190] Hideyuki Shimizu and Keiichi I. Nakayama. A 23 gene–based molecular prognostic score precisely predicts overall survival of breast cancer patients. *EBioMedicine*, 46:150–159, August 2019.

[191] Ayalew Tefferi, Paola Guglielmelli, Maura Nicolosi, Francesco Mannelli, Mythri Mudireddy, Niccolo Bartalucci, Christy M. Finke, Terra L. Lasho, Curtis A. Hanson, Rhett P. Ketterling, Kebede H. Begna, Naseema Gangat, Animesh Pardanani, and Alessandro M. Vannucchi. GIPSS: genetically inspired prognostic scoring system for primary myelofibrosis. *Leukemia*, 32(7):1631–1642, July 2018. Number: 7 Publisher: Nature Publishing Group.

[192] Adam S. Sperling, Christopher J. Gibson, and Benjamin L. Ebert. The genetics of myelodysplastic syndrome: from clonal hematopoiesis to secondary leukemia. *Nature reviews. Cancer*, 17(1):5–19, January 2017. Number: 1.

[193] Daria V. Babushok, Monica Bessler, and Timothy S. Olson. Genetic predisposition to myelodysplastic syndrome and acute myeloid leukemia in children and young adults. *Leukemia & lymphoma*, 57(3):520–536, March 2016. Number: 3.

[194] Peter L. Greenberg, Heinz Tuechler, Julie Schanz, Guillermo Sanz, Guillermo Garcia-Manero, Francesc Solé, John M. Bennett, David Bowen, Pierre Fenaux, Francois Dreyfus, Hagop Kantarjian, Andrea Kuendgen, Alessandro Levis, Luca Malcovati, Mario Cazzola, Jaroslav Cermak, Christa Fonatsch, Michelle M. Le Beau, Marilyn L. Slovak, Otto Krieger, Michael Luebbert, Jaroslaw Maciejewski, Silvia M. M. Magalhaes, Yasushi Miyazaki, Michael Pfeilstöcker, Mikkael Sekeres, Wolfgang R. Sperr, Reinhard Stauder, Sudhir Tauro, Peter Valent, Teresa Vallespi, Arjan A. van de Loosdrecht, Ulrich Germing, and Detlef Haase. Revised International Prognostic Scoring System for Myelodysplastic Syndromes. *Blood*, 120(12):2454–2465, September 2012. Number: 12.

[195] M. G. Della Porta, C. H. Jackson, E. P. Alessandrino, M. Rossi, A. Bacigalupo, M. T. van Lint, M. Bernardi, B. Allione, A. Bosi, S. Guidi, V. Santini, L. Malcovati, M. Ubezio, C. Milanesi, E. Todisco, M. T. Voso, P. Musto, F. Onida, A. P. Iori, R. Cerretti, G. Grillo,

A. Molteni, P. Pioltelli, L. Borin, E. Angelucci, E. Oldani, S. Sica, C. Pascutto, V. Ferretti, A. Santoro, F. Bonifazi, M. Cazzola, and A. Rambaldi. Decision analysis of allogeneic hematopoietic stem cell transplantation for patients with myelodysplastic syndrome stratified according to the revised International Prognostic Scoring System. *Leukemia*, 31(11):2449–2457, November 2017. Number: 11.

[196] Emilio Paolo Alessandrino, Matteo G Della Porta, Luca Malcovati, Christopher H Jackson, Cristiana Pascutto, Andrea Bacigalupo, Maria Teresa van Lint, Michele Falda, Massimo Bernardi, Francesco Onida, Stefano Guidi, Anna Paola Iori, Raffaella Cerretti, Paola Marenco, Pietro Pioltelli, Emanuele Angelucci, Rosi Oneto, Francesco Ripamonti, Alessandro Rambaldi, Alberto Bosi, and Mario Cazzola. Optimal timing of allogeneic hematopoietic stem cell transplantation in patients with myelodysplastic syndrome. *American Journal of Hematology*, 88(7):581–588, July 2013. Number: 7.

[197] Ivan Limongelli, Simone Marini, and Riccardo Bellazzi. PaPI: pseudo amino acid composition to score human protein-coding variants. *BMC Bioinformatics*, 16:123, April 2015.

[198] P. Peduzzi, J. Concato, E. Kemper, T. R. Holford, and A. R. Feinstein. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, 49(12):1373–1379, December 1996. Number: 12.

[199] Moritz Gerstung, Elli Papaemmanuil, Inigo Martincorena, Lars Bullinger, Verena I. Gaidzik, Peter Paschka, Michael Heuser, Felicitas Thol, Niccolo Bolli, Peter Ganly, Arnold Ganser, Ultan McDermott, Konstanze Döhner, Richard F. Schlenk, Hartmut Döhner, and Peter J. Campbell. Precision oncology for acute myeloid leukemia using a knowledge bank approach. *Nature Genetics*, 49(3):332–340, March 2017. Number: 3.

[200] Pedro da Silva-Coelho, Leonie I. Kroeze, Kenichi Yoshida, Theresia N. Koorenhof-Scheele, Ruth Knops, Louis T. van de Locht, Aniek O.

de Graaf, Marion Massop, Sarah Sandmann, Martin Dugas, Marian J. Stevens-Kroef, Jaroslav Cermak, Yuichi Shiraishi, Kenichi Chiba, Hiroko Tanaka, Satoru Miyano, Theo de Witte, Nicole M. A. Blijlevens, Petra Muus, Gerwin Huls, Bert A. van der Reijden, Seishi Ogawa, and Joop H. Jansen. Clonal evolution in myelodysplastic syndromes. *Nature Communications*, 8:15099, April 2017.

[201] Jose F. Falantes, Cristina Calderón, Francisco J. Márquez Malaver, Dora Alonso, Antonio Martín Noya, Estrella Carrillo, María L. Martino, Isabel Montero, Jose González, Rocío Parody, Ildefonso Espigado, and Jose A. Pérez-Simón. Clinical prognostic factors for survival and risk of progression to acute myeloid leukemia in patients with myelodysplastic syndromes with $< 10\%$ marrow blasts and non-unfavorable cytogenetic categories. *Clinical Lymphoma, Myeloma & Leukemia*, 13(2):144–152, April 2013. Number: 2.

[202] Irina TRIANTAFYLLIDIS, Anca CIOBANU, Oana STANCA, and Anca Roxana LUPU. Prognostic Factors in Myelodysplastic Syndromes. *Mædica*, 7(4):295–302, December 2012. Number: 4.

[203] Cambridge dictionary statistics 4th edition | Statistics and probability: general interest, September 2018.

[204] Ritesh Singh and Keshab Mukhopadhyay. Survival analysis in clinical trials: Basics and must know areas. *Perspectives in Clinical Research*, 2(4):145–148, 2011. Number: 4.

[205] F. A. Sonnenberg and J. R. Beck. Markov models in medical decision making: a practical guide. *Medical Decision Making: An International Journal of the Society for Medical Decision Making*, 13(4):322–338, December 1993. Number: 4.

[206] Stephen W. Duffy, Hsiu-Hsi Chen, Laszlo Tabar, and Nicholas E. Day. Estimation of mean sojourn time in breast cancer screening using a Markov chain model of both entry to and exit from the preclinical

detectable phase. *Statistics in Medicine*, 14(14):1531–1543, July 1995. Number: 14.

[207] I. M. Longini, W. S. Clark, R. H. Byers, J. W. Ward, W. W. Darrow, G. F. Lemp, and H. W. Hethcote. Statistical analysis of the stages of HIV infection using a Markov model. *Statistics in Medicine*, 8(7):831–843, July 1989. Number: 7.

[208] Edward Joseph Caruana, Marius Roman, Jules Hernández-Sánchez, and Piergiorgio Solli. Longitudinal studies. *Journal of Thoracic Disease*, 7(11):E537–E540, November 2015. Number: 11.

[209] Yuanxi Li, Stephen Swift, and Allan Tucker. Modelling and analysing the dynamics of disease progression from cross-sectional studies. *Journal of Biomedical Informatics*, 46(2):266–274, April 2013. Number: 2.

[210] E. Parimbelli, S. Marini, L. Sacchi, and R. Bellazzi. Patient similarity for precision medicine: A systematic review. *Journal of Biomedical Informatics*, 83:87–96, 2018.

[211] F. Vitali, S. Marini, D. Pala, A. Demartini, S. Montoli, A. Zambelli, and R. Bellazzi. Patient similarity by joint matrix trifactorization to identify subgroups in acute myeloid leukemia. *JAMIA open*, 1(1):75–86, July 2018.

[212] David E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1st edition, 1989.

[213] Terry M. Therneau and Patricia M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer Science & Business Media, November 2013. Google-Books-ID: oj0mBQAAQBAJ.

[214] P. Hougaard. Multi-state models: a review. *Lifetime Data Analysis*, 5(3):239–264, September 1999. Number: 3.

[215] Anna M. Jankowska, Hadrian Szpurka, Venugopalan Cheriyath, Kwok Peng Ng, Zhenbo Hu, Michael A. McDevitt, Yogen Saunthararajah, and Jaroslaw P. Maciejewski. Consequences of UTX Dysfunction in Myelodysplastic Syndrome. *Blood*, 118(21):2427–2427, November 2011. Number: 21.

[216] Harinder Gill, Anskar Leung, Yok-Lam Kwong, Harinder Gill, Anskar Y. H. Leung, and Yok-Lam Kwong. Molecular and Cellular Mechanisms of Myelodysplastic Syndrome: Implications on Targeted Therapy. *International Journal of Molecular Sciences*, 17(4):440, March 2016. Number: 4.

[217] Charlotte Pawlyn, Martin F. Kaiser, Christoph Heuck, Lorenzo Melchor, Christopher P. Wardell, Alex Murison, Shweta S. Chavan, David C. Johnson, Dil B. Begum, Nasrin M. Dahir, Paula Z. Proszek, David A. Cairns, Eileen M. Boyle, John R. Jones, Gordon Cook, Mark T. Drayson, Roger G. Owen, Walter M. Gregory, Graham H. Jackson, Bart Barlogie, Faith E. Davies, Brian A. Walker, and Gareth J. Morgan. The Spectrum and Clinical Impact of Epigenetic Modifier Mutations in Myeloma. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, 22(23):5783–5794, December 2016. Number: 23.

[218] F. Thol, E. M. Weissinger, J. Krauter, K. Wagner, F. Damm, M. Wichmann, G. Göhring, C. Schumann, G. Bug, O. Ottmann, W. K. Hofmann, B. Schlegelberger, A. Ganser, and M. Heuser. IDH1 mutations in patients with myelodysplastic syndromes are associated with an unfavorable prognosis. *Haematologica*, 95(10):1668–1674, October 2010. Number: 10.

[219] Na Wang, Fei Wang, Ningning Shan, Xiaohui Sui, and Hongzhi Xu. IDH1 Mutation Is an Independent Inferior Prognostic Indicator for Patients with Myelodysplastic Syndromes. *Acta Haematologica*, 138(3):143–151, 2017. Number: 3.

[220] Peipei Lin, Yingwan Luo, Shuanghong Zhu, Dominic Maggio, Haiyang Yang, Chao Hu, Jinghan Wang, Hua Zhang, Yanling Ren, Xinping Zhou, Chen Mei, Liya Ma, Weilai Xu, Li Ye, Zhengping Zhuang, Jie Jin, and Hongyan Tong. Isocitrate dehydrogenase 2 mutations correlate with leukemic transformation and are predicted by 2-hydroxyglutarate in myelodysplastic syndromes. *Journal of Cancer Research and Clinical Oncology*, 144(6):1037–1047, June 2018. Number: 6.

[221] Wen-Chien Chou, Hsin-An Hou, Chien-Yuan Chen, Jih-Luh Tang, Ming Yao, Woei Tsay, Bor-Shen Ko, Shang-Ju Wu, Shang-Yi Huang, Szu-Chun Hsu, Yao-Chang Chen, Yen-Ning Huang, Yi-Chang Chang, Fen-Yu Lee, Ming-Chi Liu, Chia-Wen Liu, Mei-Hsuan Tseng, Chi-Fei Huang, and Hwei-Fang Tien. Distinct clinical and biologic characteristics in adult acute myeloid leukemia bearing the isocitrate dehydrogenase 1 mutation. *Blood*, 115(14):2749–2754, April 2010. Number: 14.

[222] Hideki Makishima, Tetsuichi Yoshizato, Kenichi Yoshida, Mikkael A. Sekeres, Tomas Radivoyevitch, Hiromichi Suzuki, Bartlomiej Przychodzen, Yasunobu Nagata, Manja Meggendorfer, Masashi Sanada, Yusuke Okuno, Cassandra Hirsch, Teodora Kuzmanovic, Yusuke Sato, Aiko Sato-Otsubo, Thomas LaFramboise, Naoko Hosono, Yuichi Shiraishi, Kenichi Chiba, Claudia Haferlach, Wolfgang Kern, Hiroko Tanaka, Yusuke Shiozawa, Inés Gómez-Seguí, Holleh D. Husseinzadeh, Swapna Thota, Kathryn M. Guinta, Brittney Dienes, Tsuyoshi Nakamaki, Shuichi Miyawaki, Yogen Saunthararajah, Shigeru Chiba, Satoru Miyano, Lee-Yung Shih, Torsten Haferlach, Seishi Ogawa, and Jaroslaw P. Maciejewski. Dynamics of clonal evolution in myelodysplastic syndromes. *Nature Genetics*, 49(2):204–212, February 2017. Number: 2.

[223] Graciele Burnatt, Marley Aparecida Licínio, Pâmela Cristina Gaspar, Arthur Schveitzer Ferreira, Manoela Lira Reis, Ana Carolina Rabello de Moraes, Thaís Cristine Marques Sincero, Maria Cláu-

dia Santos-Silva, Graciele Burnatt, Marley Aparecida Licínio, Pâmela Cristina Gaspar, Arthur Schveitzer Ferreira, Manoela Lira Reis, Ana Carolina Rabello de Moraes, Thaís Cristine Marques Sincero, and Maria Cláudia Santos-Silva. Analysis of the presence of FLT3 gene mutation and association with prognostic factors in adult and pediatric acute leukemia patients. *Brazilian Journal of Pharmaceutical Sciences*, 53(2), 2017. Number: 2.

[224] Kelly M. Arcipowski, Marinka Bulic, Sandeep Gurbuxani, and Jonathan D. Licht. Loss of Mll3 Catalytic Function Promotes Aberrant Myelopoiesis. *PLOS ONE*, 11(9):e0162515, September 2016. Number: 9.

[225] Chong Chen, Yu Liu, Amy R. Rappaport, Thomas Kitzing, Nikolaus Schultz, Zhen Zhao, Aditya S. Shroff, Ross A. Dickins, Christopher R. Vakoc, James E. Bradner, Wendy Stock, Michelle M. LeBeau, Kevin M. Shannon, Scott Kogan, Johannes Zuber, and Scott W. Lowe. MLL3 Is a Haploinsufficient 7q Tumor Suppressor in Acute Myeloid Leukemia. *Cancer cell*, 25(5):652–665, May 2014. Number: 5.

[226] Kinisha Gala, Qing Li, Amit Sinha, Pedram Razavi, Madeline Dorso, Francisco Sanchez-Vega, Young Rock Chung, Ronald Hendrickson, James J. Hsieh, Michael Berger, Nikolaus Schultz, Alessandro Pastore, Omar Abdel-Wahab, and Sarat Chandarlapaty. KMT2C mediates the estrogen dependence of breast cancer through regulation of ER enhancer function. *Oncogene*, 37(34):4692–4710, August 2018. Number: 34.

[227] Toshiya Inaba, Hiroaki Honda, and Hirotaka Matsui. The enigma of monosomy 7. *Blood*, 131(26):2891–2898, 2018. Number: 26.

[228] Juliana Schwaab, Thomas Ernst, Philipp Erben, Jenny Rinke, Susanne Schnittger, Philipp Ströbel, Georgia Metzgeroth, Max Mossner, Torsten Haferlach, Nicholas C. P. Cross, Andreas Hochhaus,

Wolf-Karsten Hofmann, and Andreas Reiter. Activating CBL mutations are associated with a distinct MDS/MPN phenotype. *Annals of Hematology*, 91(11):1713–1720, November 2012. Number: 11.

[229] Hsiao-Wen Kao, Masashi Sanada, Der-Cherng Liang, Chang-Liang Lai, En-Hui Lee, Ming-Chung Kuo, Tung-Liang Lin, Yu-Shu Shih, Jin-Hou Wu, Chein-Fuang Huang, Seishi Ogawa, and Lee-Yung Shih. A high occurrence of acquisition and/or expansion of C-CBL mutant clones in the progression of high-risk myelodysplastic syndrome to acute myeloid leukemia. *Neoplasia (New York, N.Y.)*, 13(11):1035–1042, November 2011. Number: 11.

[230] Hsin-An Hou, Cheng-Hong Tsai, Chien-Chin Lin, Wen-Chien Chou, Yuan-Yeh Kuo, Chieh-Yu Liu, Mei-Hsuan Tseng, Yen-Ling Peng, Ming-Chih Liu, Chia-Wen Liu, Xiu-Wen Liao, Liang-In Lin, Ming Yao, Jih-Luh Tang, and Hwei-Fang Tien. Incorporation of mutations in five genes in the revised International Prognostic Scoring System can improve risk stratification in the patients with myelodysplastic syndrome. *Blood Cancer Journal*, 8(4):39, April 2018. Number: 4.

[231] Yusuke Shiozawa, Luca Malcovati, Anna Gallì, Andrea Pellagatti, Mohsen Karimi, Aiko Sato-Otsubo, Yusuke Sato, Hiromichi Suzuki, Tetsuichi Yoshizato, Kenichi Yoshida, Yuichi Shiraishi, Kenichi Chiba, Hideki Makishima, Jacqueline Boultwood, Eva Hellström-Lindberg, Satoru Miyano, Mario Cazzola, and Seishi Ogawa. Gene expression and risk of leukemic transformation in myelodysplasia. *Blood*, 130(24):2642–2653, 2017. Number: 24.

[232] Olivier Kosmider, Véronique Gelsi-Boyer, Meyling Cheok, Sophie Grabar, Véronique Della-Valle, Françoise Picard, Franck Viguié, Bruno Quesnel, Odile Beyne-Rauzy, Eric Solary, Norbert Vey, Mathilde Hunault-Berger, Pierre Fenaux, Véronique Mansat-De Mas, Eric Delabesse, Philippe Guardiola, Catherine Lacombe, William Vainchenker, Claude Preudhomme, François Dreyfus, Olivier A. Bernard, Daniel Birnbaum, Michaëla Fontenay, and on behalf of the

Groupe Francophone des Myélodysplasies. TET2 mutation is an independent favorable prognostic factor in myelodysplastic syndromes (MDSs). *Blood*, 114(15):3285–3291, October 2009. Number: 15.

[233] Rafael Bejar, Kristen Stevenson, Omar Abdel-Wahab, Naomi Galili, Björn Nilsson, Guillermo Garcia-Manero, Hagop Kantarjian, Azra Raza, Ross L. Levine, Donna Neuberg, and Benjamin L. Ebert. Clinical effect of point mutations in myelodysplastic syndromes. *The New England Journal of Medicine*, 364(26):2496–2506, June 2011. Number: 26.

[234] Charles G. Mullighan, Jinghui Zhang, Lawryn H. Kasper, Stephanie Lerach, Debbie Payne-Turner, Letha A. Phillips, Sue L. Heatley, Linda Holmfeldt, J. Racquel Collins-Underwood, Jing Ma, Kenneth H. Buetow, Ching-Hon Pui, Sharyn D. Baker, Paul K. Brindle, and James R. Downing. CREBBP mutations in relapsed acute lymphoblastic leukaemia. *Nature*, 471(7337):235–239, March 2011. Number: 7337.

[235] Giulio Genovese, Anna K. Kähler, Robert E. Handsaker, Johan Lindberg, Samuel A. Rose, Samuel F. Bakhoum, Kimberly Chambert, Eran Mick, Benjamin M. Neale, Menachem Fromer, Shaun M. Purcell, Oscar Svantesson, Mikael Landén, Martin Höglund, Sören Lehmann, Stacey B. Gabriel, Jennifer L. Moran, Eric S. Lander, Patrick F. Sullivan, Pamela Sklar, Henrik Grönberg, Christina M. Hultman, and Steven A. McCarroll. Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood DNA Sequence. *New England Journal of Medicine*, 371(26):2477–2487, December 2014. Number: 26.

[236] Siddhartha Jaiswal, Pierre Fontanillas, Jason Flannick, Alisa Manning, Peter V. Grauman, Brenton G. Mar, R. Coleman Lindsley, Craig H. Mermel, Noel Burtt, Alejandro Chavez, John M. Higgins, Vladislav Moltchanov, Frank C. Kuo, Michael J. Kluk, Brian Henderson, Leena Kinnunen, Heikki A. Koistinen, Claes Ladenvall, Gad Getz, Adolfo Correa, Benjamin F. Banahan, Stacey Gabriel,

Sekar Kathiresan, Heather M. Stringham, Mark I. McCarthy, Michael Boehnke, Jaakko Tuomilehto, Christopher Haiman, Leif Groop, Gil Atzmon, James G. Wilson, Donna Neuberg, David Altshuler, and Benjamin L. Ebert. Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes. *New England Journal of Medicine*, 371(26):2488–2498, December 2014. Number: 26.

[237] H. Cechova, P. Lassuthova, L. Novakova, M. Belickova, R. Stemberkova, J. Jencik, M. Stankova, P. Hrabakova, K. Pegova, H. Zizkova, and J. Cermak. Monitoring of methylation changes in 9p21 region in patients with myelodysplastic syndromes and acute myeloid leukemia. *Neoplasma*, 59(2):168–174, 2012. Number: 2.

[238] Jun Qian, Xing-Xing Chen, Wei Qian, Jing Yang, Xiang-Mei Wen, Ji-Chun Ma, Zhao-Qun Deng, Zhen Qian, Ying-Ying Zhang, and Jiang Lin. Aberrant hypermethylation of CTNNA1 gene is associated with higher IPSS risk in patients with myelodysplastic syndrome. *Clinical Chemistry and Laboratory Medicine*, 52(12):1859–1865, December 2014. Number: 12.

[239] Ying Ye, Michael A. McDevitt, Mingzhou Guo, Wei Zhang, Oliver Galm, Steven D. Gore, Judith E. Karp, Jaroslaw P. Maciejewski, Jeanne Kowalski, Hua-Ling Tsai, Lukasz P. Gondek, Hsing-Chen Tsai, Xiaofei Wang, Craig Hooker, B. Douglas Smith, Hetty E. Carraway, and James G. Herman. Progressive chromatin repression and promoter methylation of CTNNA1 associated with advanced myeloid malignancies. *Cancer Research*, 69(21):8482–8490, November 2009. Number: 21.

[240] Ting Xi Liu, Michael W. Becker, Jaroslav Jelinek, Wen-Shu Wu, Min Deng, Natallia Mikhalkevich, Karl Hsu, Clara D. Bloomfield, Richard M. Stone, Daniel J. DeAngelo, Ilene A. Galinsky, Jean-Pierre Issa, Michael F. Clarke, and A. Thomas Look. Chromosome 5q deletion and epigenetic suppression of the gene encoding alpha-

catenin (CTNNA1) in myeloid cell transformation. *Nature Medicine*, 13(1):78–83, January 2007. Number: 1.

[241] Mary Charlson, Ted P. Szatrowski, Janey Peterson, and Jeffrey Gold. Validation of a combined comorbidity index. *Journal of Clinical Epidemiology*, 47(11):1245–1251, November 1994. Number: 11.

[242] José J. Fuster, Susan MacLauchlan, María A. Zuriaga, Maya N. Polackal, Allison C. Ostriker, Raja Chakraborty, Chia-Ling Wu, Soichi Sano, Sujatha Muralidharan, Cristina Rius, Jacqueline Vuong, Sophia Jacob, Varsha Muralidhar, Avril A. B. Robertson, Matthew A. Cooper, Vicente Andrés, Karen K. Hirschi, Kathleen A. Martin, and Kenneth Walsh. Clonal hematopoiesis associated with TET2 deficiency accelerates atherosclerosis development in mice. *Science (New York, N.Y.)*, 355(6327):842–847, 2017. Number: 6327.

[243] Siddhartha Jaiswal, Pradeep Natarajan, Alexander J. Silver, Christopher J. Gibson, Alexander G. Bick, Eugenia Shvartz, Marie McConkey, Namrata Gupta, Stacey Gabriel, Diego Ardissino, Usman Baber, Roxana Mehran, Valentin Fuster, John Danesh, Philippe Frossard, Danish Saleheen, Olle Melander, Galina K. Sukhova, Donna Neuberg, Peter Libby, Sekar Kathiresan, and Benjamin L. Ebert. Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular Disease. *The New England Journal of Medicine*, 377(2):111–121, 2017. Number: 2.

[244] Adam H. Buchanan, H. Lester Kirchner, Marci L. B. Schwartz, Melissa A. Kelly, Tara Schmidlen, Laney K. Jones, Miranda L. G. Hallquist, Heather Rocha, Megan Betts, Rachel Schwiter, Loren Butry, Amanda L. Lazzeri, Lauren R. Frisbie, Alanna Kulchak Rahm, Jing Hao, Huntington F. Willard, Christa L. Martin, David H. Ledbetter, Marc S. Williams, and Amy C. Sturm. Clinical outcomes of a genomic screening program for actionable genetic conditions. *Genetics in Medicine*, pages 1–9, June 2020. Publisher: Nature Publishing Group.

[245] Rubina Manuela Trimboli, Paolo Giorgi Rossi, Nicolò Matteo Luca Battisti, Andrea Cozzi, Veronica Magni, Moreno Zanardo, and Francesco Sardanelli. Do we still need breast cancer screening in the era of targeted therapies and precision medicine? *Insights into Imaging*, 11(1):105, September 2020.

[246] Nicolas Servant, Julien Roméjon, Pierre Gestraud, Philippe La Rosa, Georges Lucotte, Séverine Lair, Virginie Bernard, Bruno Zeitouni, Fanny Coffin, Gérôme Jules-Clément, Florent Yvon, Alban Lermine, Patrick Poullet, Stéphane Liva, Stuart Pook, Tatiana Popova, Camille Barette, François Prud'homme, Jean-Gabriel Dick, Maud Kamal, Christophe Le Tourneau, Emmanuel Barillot, and Philippe Hupé. Bioinformatics for precision medicine in oncology: principles and application to the SHIVA clinical trial. *Frontiers in Genetics*, 5, 2014. Publisher: Frontiers.

[247] Anthony Letai. Functional precision cancer medicine—moving beyond pure genomics. *Nature Medicine*, 23(9):1028–1035, September 2017. Number: 9 Publisher: Nature Publishing Group.

[248] Johannes Starlinger, Steffen Pallarz, Jurica Ševa, Damian Rieke, Christine Sers, Ulrich Keilholz, and Ulf Leser. Variant information systems for precision oncology. *BMC Medical Informatics and Decision Making*, 18(1):107, November 2018.

[249] Yasi Wang, Hongxun Yao, and Sicheng Zhao. Auto-encoder based dimensionality reduction. *Neurocomputing*, 184:232–242, April 2016.

[250] Collin Tokheim and Rachel Karchin. CHASMplus Reveals the Scope of Somatic Missense Mutations Driving Human Cancers. *Cell Systems*, 9(1):9–23.e8, July 2019.

[251] Ruth Nussinov and Chung-Jung Tsai. 'Latent drivers' expand the cancer mutational landscape. *Current Opinion in Structural Biology*, 32:25–32, June 2015.

[252] Christopher D. McFarland, Kirill S. Korolev, Gregory V. Kryukov, Shamil R. Sunyaev, and Leonid A. Mirny. Impact of deleterious passenger mutations on cancer progression. *Proceedings of the National Academy of Sciences*, 110(8):2910–2915, February 2013. Publisher: National Academy of Sciences Section: Biological Sciences.

[253] Alexander Gepperth and Barbara Hammer. Incremental learning algorithms and applications. page 13.

[254] Dariusz Brzezinski and Jerzy Stefanowski. Reacting to Different Types of Concept Drift: The Accuracy Updated Ensemble Algorithm. *IEEE Transactions on Neural Networks and Learning Systems*, 25(1):81–94, January 2014. Conference Name: IEEE Transactions on Neural Networks and Learning Systems.

[255] Huixin Tian, Minwei Shuai, Kun Li, and Xiao Peng. An Incremental Learning Ensemble Strategy for Industrial Process Soft Sensors, May 2019. ISSN: 1076-2787 Pages: e5353296 Publisher: Hindawi Volume: 2019.

[256] Marko Ristin, Matthieu Guillaumin, Juergen Gall, and Luc Van Gool. Incremental Learning of Random Forests for Large-Scale Image Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):490–503, March 2016. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[257] Aiping Wang, Guowei Wan, Zhiquan Cheng, and Sikun Li. An incremental extremely random forest classifier for online learning and tracking. In *2009 16th IEEE International Conference on Image Processing (ICIP)*, pages 1449–1452, November 2009. ISSN: 2381-8549.

[258] Amir Saffari, Christian Leistner, Jakob Santner, Martin Godec, and Horst Bischof. On-line Random Forests. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pages 1393–1400, September 2009.

[259] Zardad Khan, Asma Gul, Aris Perperoglou, Miftahuddin Miftahuddin, Osama Mahmoud, Werner Adler, and Berthold Lausen. Ensemble of optimal trees, random forest and random projection ensemble classification. *Advances in Data Analysis and Classification*, 14(1):97–116, March 2020.

[260] Sharon E. Davis, Robert A. Greevy, Thomas A. Lasko, Colin G. Walsh, and Michael E. Matheny. Comparison of Prediction Model Performance Updating Protocols: Using a Data-Driven Testing Procedure to Guide Updating. *AMIA Annual Symposium Proceedings*, 2019:1002–1010, March 2020.

[261] D. B. Toll, K. J. M. Janssen, Y. Vergouwe, and K. G. M. Moons. Validation, updating and impact of clinical prediction rules: a review. *Journal of Clinical Epidemiology*, 61(11):1085–1094, November 2008.

[262] Johannes Birgmeier, Cole A. Deisseroth, Laura E. Hayward, Luisa M. T. Galhardo, Andrew P. Tierno, Karthik A. Jagadeesh, Peter D. Stenson, David N. Cooper, Jonathan A. Bernstein, Maximilian Haeussler, and Gill Bejerano. AVADA: Towards Automated Pathogenic Variant Evidence Retrieval Directly from the Full Text Literature. *Genetics in medicine : official journal of the American College of Medical Genetics*, 22(2):362–370, February 2020.

[263] Geoff Macintyre, Bauke Ylstra, and James D. Brenton. Sequencing Structural Variants in Cancer for Precision Therapeutics. *Trends in Genetics*, 32(9):530–542, September 2016.

[264] Marinka Žitnik and Blaž Zupan. Data Fusion by Matrix Factorization. *IEEE transactions on pattern analysis and machine intelligence*, 37(1):41–53, January 2015.

[265] John G. Kemeny and J. Laurie Snell. *Finite Markov Chains: With a New Appendix "Generalization of a Fundamental Matrix"*. Undergraduate Texts in Mathematics. Springer-Verlag, New York, 1976.

Publications

Journals

- **G. Nicora**, F. Vitali, A. Dagliati, N. Geifman, R. Bellazzi. Integrated Multi-Omics Analyses in Oncology: A Review of Machine Learning Methods and Tools. *Frontiers in Oncology*, https://doi.org/10.3389/fonc.2020.01030, June 2020

- **G. Nicora**, F. Moretti, E. Sauta, M. Della Porta, L. Malcovati, M. Cazzola, S. Quaglini, R. Bellazzi, "A continuous-time Markov model approach for modeling myelodysplastic syndromes progression from cross-sectional data", *Journal of Biomedical Informatics*, Volume 104, April 2020, 103398, https://doi.org/10.1016/j.jbi.2020.103398

- G. Coticchio, G. Fiorentino, **G. Nicora**,R. Sciajno,F. Cavalera R. Bellazzi,S. Garagna, A. Borini, M. Zuccotti, Harnessing cytoplasmic particles movement of the human early embryo analysed by advanced imaging and artificial intelligence to predict development to blastocyst stage, Reproductive BioMedicine Online, December 2020, https://doi.org/10.1016/j.rbmo.2020.12.008

- **G. Nicora**, Ivan Limongelli, Patrick Gambelli, Mirella Memmi, Carlo Napolitano, Alberto Malovini, Andrea Mazzanti, Silvia Priori, Riccardo Bellazzi, "CardioVAI: An automatic implementation of ACMG-AMP variant interpretation guidelines in the diagnosis of cardiovascular diseases", *Human Mutation*, October 2018, PMID:30298955, https://doi.org/10.1002/humu.

23665

Conference Proceedings

- **G. Nicora**, Riccardo Bellazzi, A Reliable Machine Learning Approach applied to Single-Cell Classification in Acute Myeloid Leukemia. AMIA (Americal Medical Informatics Association) Annual Symposium, 2020

- **G. Nicora**, Simone Marini, Ivan Limongelli, Ettore Rizzo, Stefano Montoli, Francesca Tricomi, Riccardo Bellazzi, A Semi-supervised Learning Approach for Pan-Cancer Somatic Genomic Variant Classification. In: Riaño D., Wilk S., ten Teije A. (eds) Artificial Intelligence in Medicine. AIME 2019. Lecture Notes in Computer Science, vol 11526. Springer, Cham

- **G. Nicora**, Ivan Limongelli, Riccardo Cova, Matteo Giovanni Della Porta, Luca Malcovati, Mario Cazzola, Riccardo Bellazzi, A Rule-Based Expert System for Automatic Implementation of Somatic Variant Clinical Interpretation Guidelines, In: Riaño D., Wilk S., ten Teije A. (eds) Artificial Intelligence in Medicine. AIME 2019. Lecture Notes in Computer Science, vol 11526. Springer, Cham

- G. Coticchio, R. Sciajno, G. Fiorentino, F. Cavalera **G. Nicora**, R. Bellazzi, A. Borini, S. Garagna, M. Zuccotti, Artificial neural-network analysis combined with time-lapse imaging predicts embryo ability to develop to the blastocyst stage, Fertility and Sterility, Volume 112, Issue 3, Supplement, September 2019, Pages e273-e274

Conference Abstracts

- **G. Nicora**, Ivan Limongelli, Patrick Gambelli, Mirella Memmi, Carlo Napolitano, Alberto Malovini, Andrea Mazzanti, Silvia Priori, Riccardo Bellazzi, "An automated guidelines-based approach for variants pathogenicity." podium abstract AMIA (American Medical Informatics Association), San Francisco (CA), March 2018.

- **G. Nicora**, Ivan Limongelli, Patrick Gambelli, Mirella Memmi, Carlo Napolitano, Alberto Malovini, Andrea Mazzanti, Silvia Priori, Riccardo Bellazzi, "An automatic implementation of ACMG/AMP variant interpretation guidelines.", abstract ESHG (European Society of Human Genetics) June 2018.

- **G. Nicora**, F. Moretti, E- Sauta, L. Malcovati, M. Della Porta, S. Quaglini, M. Cazzola, R- Bellazzi, "A countinuous-time Markov approach for modelling myelodysplastic syndromes progression from cross-sectional data", poster presentation at AMIA (American Medical Informatics Association) 2019

- I. Limongelli, **G. Nicora**, P. Gambelli, M. Memmi, C. Napolitano, A. Malovini, A. Mazzanti, S. Priori, R. Bellazzi, "An automated guidelines-based approach for variants pathogenicity assessment in the diagnosis of genetic cardiovascular diseases.", abstract for poster presentation SIGU (Società Italiana Genetica Umana), Naples (Italy) November 2017. (Poster presentation)

- **G. Nicora**, I. Limongelli, P. Gambelli, M. Memmi, C. Napolitano, A. Malovini, A. Mazzanti, S. Priori, R. Bellazzi, "A Rule-based Expert System for automatic genomic variant interpretation", abstract and poster presentation GNB (Gruppo Nazionale Bioingegneria), Milan (Italy) June 2018. (Poster Presentation)

- M. V. Esposito, M. Nunziato, I. Limongelli, **G. Nicora**, V. D'Argenio, "DNA variants interpretation in the next generation sequencing era: the case of eVAI tool", e-poster presentation, Italian national conference Società Italiana di Biochimica Clinica e Biologia Molecolare Clinica (SIBioC), Naples (Italy) October 2018

- R. Bartolucci, S. Grandoni, N. Melillo, **G. Nicora**, E. Sauta, E.M.Tosca, P. Magni, "Artificial Intelligence and machine learning: just a hype or a new opportunity for pharmacometrics?", abstract at PAGE (Population Approach Group in Europe) conference, Stockholm, Sweden, June 2019. (Poster Presentation)

- M. Zuccotti, G. Coticchio, G. Fiorentino, F. Cavalera, **G. Nicora**, R. Bellazzi, R. Sciajno, A. Borini, S. Garagna, Time-Lapse imaging combined with artificial neural-network analysis predicts oocytes and preimplantation embryos developmental competence. 65° Convegno Gruppo Embriologico ItalianoSocietà italiana di Biologia dello Sviluppo e della Cellula, 24-27 June 2019, Ancona (Italy).

- **G.Nicora**, I. Limongelli, S. Zucca, R. Santolisier, P. Magni, R. Bellazzi, A comparison of eVAI, CADD and VVP variant prediction results on the ICR639 hereditary cancer dataset, abstract for poster presentation, American Society of Human Genetics (ASHG) Annual Meeting, Houston (US-TX), october 2019

- F. De Paoli, I. Limongelli, E. Rizzo, **G. Nicora**, P. Magni, An automatic implementation of ACMG/ClinGen guidelines for constitutional Copy Number Variants annotation and interpretation, American Society of Human Genetics (ASHG) Annual Meeting, october 2020

Acknowledgements

This work has been funded by the Fondazione Regionale per la Ricerca Biomedica (Milan, Italy [FRRB project n. 2015-0042, Genomic profiling of rare hematologic malignancies, development of personalized medicine strategies, and their implementation into Rete Ematologica Lombarda (REL) clinical network]).

I would also like to acknowledge all the people that surrounded me during my PhD studentship.

First of all, Professor Riccardo Bellazzi, for his mentorship, inspiration and competence, for its continuous support, and for always be forthcoming.

Dr Ivan Limongelli, Dr Ettore Rizzo, Dr Susanna Zucca, Dr Simone Marini for all the opportunities, projects and discussions about bioinformatics and AI.

Professor Blaz Zupan and Dr Martin Strazar, for making my abroad period a unique experience of personal and professional growth.

Professor Luca Malcovati, for his support during our collaboration.

Dr Arianna Dagliati and Dr Francesca Vitali for our fruitful collaboration investigating multi-omics approaches.

All my friends of the University of Pavia, for all the coffee, lunches, meetings, for making my time in Pavia so enjoyable.

All the people at the University of Ljubljana, for making me feel at home during my abroad period.

Last but not least, my family and friends, for their unconditioned support and love.