

Alma Mater Studiorum - Università di Bologna

DOTTORATO DI RICERCA IN
DATA SCIENCE AND COMPUTATION

Ciclo 33

Settore Concorsuale: 05/E1 - BIOCHIMICA GENERALE

Settore Scientifico Disciplinare: BIO/10 - BIOCHIMICA

METHODS FOR BIODATA ANALYSIS IN DIFFERENT OMIC CONTEXTS

Presentata da: Davide Baldazzi

Coordinatore Dottorato

Andrea Cavalli

Supervisore

Emidio Capriotti

Co-supervisore

Rita Casadio

Esame finale anno 2022

Dedication

I dedicate this thesis to my grandmother Romana who have meant and continue to mean so much to me. She loved me unconditionally and she did everything possible to allow me to pursue my goals. Her memory will continue to shepherd me throughout my life.

Thesis Abstract

*PhD in Data Science and Computation. **Methods for Biodata Analysis in different Omic Contexts**, Candidate: Baldazzi Davide, Tutors: Prof. Rita Casadio; Dr. Roberta Maestro*

The world of Computational Biology and Bioinformatics presently integrates many different expertise, including computer science and electronic engineering. A major aim in Data Science is the development and tuning of specific computational approaches to interpret the complexity of Biology. Molecular biologists and medical doctors heavily rely on an interdisciplinary expert capable of understanding the biological background to apply algorithms for finding optimal solutions to their problems. With this problem-solving orientation, I was involved in two basic research fields: Cancer Genomics and Enzyme Proteomics.

As to Cancer Genomics, I strictly collaborate with medical doctors and researchers, within the group of Dr. Roberta Maestro at the CRO Aviano National Cancer Institute. As to Functional Proteomics, I collaborate with the Biocomputing group of the University of Bologna (UNIBO) (tutor: Prof. Rita Casadio, www.biocomp.unibo.it). My financial support was indeed due to a framework-collaboration between CRO and UNIBO.

For this reason, what I developed and implemented can be considered a general effort to help data analysis both in Cancer Genomics and in Enzyme Proteomics, focusing on enzymes which catalyse all the biochemical reactions in cells.

Specifically, as to Cancer Genomics I contributed to the characterization of intratumoral immune microenvironment in gastrointestinal stromal tumours (GISTs) correlating immune cell population levels with tumour subtypes. I was involved in the setup of strategies for the evaluation and standardization of different approaches for fusion transcript detection in sarcomas that can be applied in routine diagnostic. This was part of a coordinated effort of the Sarcoma working group of ACC (Alleanza Contro il Cancro; <https://www.alleanzacontroilcancro.it/en/>). Finally, I developed an internal application to perform standardized bioinformatics analysis with a user-friendly interface for biologist (unpublished). This application is tailored to cope with an internal database of biosamples and data from NGS technologies that I created on purpose.

As to Enzyme Proteomics, I generated a derived database collecting all the human proteins and enzymes which are known to be associated to genetic disease. I curated the data search in freely available databases such as PDB (<https://www.rcsb.org/>), UniProt (<https://www.uniprot.org/>),

Humsavar (<https://www.uniprot.org/docs/humsavar>), Clinvar (<https://www.ncbi.nlm.nih.gov/clinvar/>) and I was responsible of searching, updating, and handling the information content, and computing statistics . I also developed a web server, BENZ, which allows researchers to annotate an enzyme sequence with the corresponding Enzyme Commission number (EC number), the important feature fully describing the catalysed reaction (see 2.1). More to this, I greatly contributed to the characterization of the enzyme-genetic disease association, for a better classification of the metabolic genetic diseases. The papers I contributed and the work I presented to national and international meetings (see LAA) fully describe all the different approaches I developed and that I extensively describe in the Result section.

Index

1) DATA SCIENCE AND COMPUTATION	5
1.1) BIG DATA IN LIFE SCIENCES.....	5
1.2) ARTIFICIAL INTELLIGENCE, MACHINE LEARNING AND NEURAL NETWORKS IN BIOINFORMATICS	7
1.3) REFERENCES	9
2) SPECIFIC PROBLEMS IN DIFFERENT OMIC CONTEXTS	13
2.1) SUBTYPES CHARACTERIZATION OF SOFT TISSUE SARCOMAS	13
2.2) FUNCTIONAL ANALYSIS OF ENZYME BIOSEQUENCES	15
2.3) REFERENCES	18
3) OBJECTIVES AND MAIN RESULTS.....	20
4) MATERIALS AND METHODS.....	22
4.1) GENOMIC AND TRANSCRIPTOMIC DATA.....	22
4.2) PROTEOMIC DATABASES.....	22
4.3) TOOLS FOR GENOMIC AND TRANSCRIPTOMIC ANALYSIS	23
4.4) TOOLS FOR PROTEOMIC ANALYSIS	25
4.5) NEW TOOLS DEVELOPED FOR GENOMIC AND TRANSCRIPTOMIC ANALYSIS	25
4.6) NEW TOOLS DEVELOPED FOR PROTEOMIC ANALYSIS	26
4.7) REFERENCES	27
5) RESULTS AND DISCUSSIONS.....	32
5.1) CANCER GENOMICS	32
5.1.1) Characterization of the immune microenvironment in GISTs	32
5.1.2) Assessment of the reliability of NGS RNA-based approaches for the detection of fusion transcripts	36
5.2) ENZYME PROTEOMICS	38
5.2.1) Characterizing disease-related enzymes.....	38
A) Human Enzymes Active in Different Metabolic Pathways and Diseases	39
B) The relations between disease-related variation types, maladies, and structural conserved domains.....	43
C) Human MTHFR deficiency.	47
D) A curated Database of Disease-related Enzymes.....	48

5.2.2) BENZ WS: The Bologna ENZyme Web Server	48
B) BENZ at work	51
C) Benchmarking BENZ	53
D) The WEB Server	57
5.3) REFERENCES	57
6) CONCLUSIONS AND PERSPECTIVES.....	60
7) LIST OF ARTICLES AND ABSTRACTS (LAA)	62
7.1) ARTICLES.....	62
7.2) ABSTRACTS AND ORAL PRESENTATIONS	63
8) ACKNOWLEDGMENTS.....	65
9) APPENDIX (PUBLICATIONS)	66

1) Data Science and Computation

1.1) Big Data in Life Sciences

In Life Science, Big Data deserve a special attention. The usual question is: Do we have Big Data in Life Sciences? (See for example: <https://towardsdatascience.com/do-we-have-big-data-in-life-sciences-c6c4e9f8645c>). Indeed, the problem of data-driven sciences, as presently molecular biology and medicine, is to collect as much data as possible to cover all the different aspects of the many different phenomena under investigation. Often, the amount of data is still a bottleneck for reaching conclusions at a global level.

Over the last two decades, digitization and datafication in biology have become prominent through sequencing and bioinformatics. While many scientific disciplines such as particle physics have been long producing Big Data, only recently research data in medical and biological sciences achieved Big Data properties. Major events that led to the emergence of Big Data in life science include: 1) Technological revolution for data generation; 2) Development of dedicated tools for analysing data in Bioinformatics; 3) Conceptual changes in science practice towards open data [1].

Regarding data volume and rate of generation, Genomics is the leader field in life sciences thanks to next-generation sequencing (NGS) technologies; however, other bioanalytical platforms, such as mass spectrometry and imaging, are catching up, drastically increasing the quantity and quality of protein and metabolite characterization and quantification [2]. The emergence of high-throughput technologies promoted the development of dedicated workflows and bioinformatics tools tailored for Big Data analysis. Such tools and workflows are characterized by structured and scalable approaches, enabling researchers to study and integrate data from several heterogeneous sources. Nowadays, more and more publications use datasets from more than one life science areas such as genomics, proteomics, metabolomics, interactomics and other at organism- or cell-dimension data, overall defined multi-omics studies [3]. The ability to analyse and integrate data from multiple sources led to the need to share primary datasets and to the advent of open data and public repositories. This goal finally disclosed the true potential of Big Data which holds the promises for new model generation and hidden pattern discovery, opening new frontiers for research in Life Sciences.

Bioanalytical data management is a core element of the Big Data ecosystem. During the last years, many efforts have been spent to establish good management practices. Those efforts

stirred the concept of FAIR data as a prerequisite to perform continuous big data-driven research. FAIR stands for Findable, Accessible, Interoperable, and Reusable data [4]. In Life Sciences, multiple data-hosting platforms exist, tailored to cope with Big Data (see for example ELIXIR, the European effort for sharing Data and analytical tools in Life Science, <https://elixir-europe.org/>).

Databases provide a centralized location for specific data types such as those obtained from experiments in different omics contexts. For example, the GDC Data Portal is a data-drive platform for querying and downloading high-quality cancer-related data, mainly from The Cancer Genome Atlas (TCGA) research program (<https://www.cancer.gov/tcga>). TCGA project is a joint effort between NCI and the National Human Genome Research Institute whose goal is the molecular characterization of different cancer types, providing an exhaustive dataset to the scientific community. Similarly, freely accessible samples can be collected from the Sequence Read Archive (SRA; <https://www.ncbi.nlm.nih.gov/sra>) [5]. The SRA is a public repository for high throughput sequencing data maintained by the International Nucleotide Sequence Database Collaboration (INSDC) [6]. Records in the SRA are operated by the National Center for Biotechnology Information (NCBI) [7], the European Bioinformatics Institute (EBI) [8] and the DNA Data Bank of Japan (DDBJ) [9].

In proteomics, many wide data collections exist. The UniProt KnowledgeBase (<https://www.uniprot.org/>) is the largest collection of protein sequences and it includes two distinct sections: a manually annotated and reviewed part called Swiss-Prot, and an automatically annotated and unreviewed part called TrEMBL. In the current UniProt release (2021_03 of 02 June 2021) Swiss-Prot contains 565,254 sequences from 14,085 different species while TrEMBL collects 219,174,961 chains from 1,255,356 organisms. Only 28.9% of the protein sequences contained in Swiss-Prot have got evidence at the protein or transcript level; this percentage drops to 0.70% when considering the TrEMBL section where most of the sequences are inferred from homology and predicted. Another database collecting more than thousands of records is KEGG (<https://www.genome.jp/kegg/>) [10]. KEGG allows the inspection of high-level functions and interactions in biological systems of multiple species, from the cell perspective up to the organism level. The latest release (v.99.1) of August 2021, collects data for 36,751,502 genes from 7,454 different organisms out of which 635 are eukaryotes, 6,466 bacteria and 353 archaea, respectively. Moreover, KEGG collects 11,605 different biochemical reactions (KEGG REACTION) from 3,175 reaction classes and 825,459 pathways (KEGG

PATHWAY), summarized by 547 reference pathways. Similarly, REACTOME collects information from all the billions of biochemical reactions up to the complex protein interactions supporting physiology at a molecular level (<https://reactome.org/>).

The data ecosystem here described allows integrative analysis that were not possible in the recent past. The decentralized and standardized data-sharing procedures developed in the last years paved the foundations for increasing our knowledge in many aspects of biological processes for many different species, including *homo sapiens* and all the different pathologies affecting individuals. In Life Sciences, data science proved to be a key element to extract knowledge and insights from complicated and noisy sets of data. Multiple processes, methods and algorithms are nowadays available to cluster, filter, aggregate, manipulate and analyse the ever-growing amount of biological data.

1.2) Artificial Intelligence, Machine Learning and Neural Networks in Bio-informatics

Both inside and outside the academic environment, Big Data are quite popular in numerous distinct fields. Similarly, technologies based on Artificial Intelligence (AI) are becoming more embedded in our daily lives and consequently, nowadays companies, as well as researchers, are more heavily relying on learning algorithms than ever. Solutions to problems are “learned” from examples: the higher is the number of examples, the better the algorithm will learn possible rules of association among input and output data, and it will help in inference problems with high reliability. The underlying technologies are based on different types of machine learning procedures, including the most recent deep learning [11].

We can think of machine learning implementations like Russian nesting dolls, where each approach is a component of the previous one. Indeed, machine learning is a subfield of artificial intelligence, and deep learning algorithms are built on top of neural networks which have been the main development in the field. Machine learning includes all the algorithms designed to optimize a performance criterion, such as the value of a fitness function or the accuracy of a predictive model, using training data and/or experience. It follows that machine learning approaches are characterized by optimization problems, finding optimal solutions in the space of multiple possible ones [12].

Optimization algorithms are classified as exact and approximate. An exact method solves the problem to optimality [13], while an approximate method generates a candidate solution but

not necessarily the perfect one [14]. Approximate methods can be additionally divided in deterministic, stochastic, and heuristic approaches [15]. Well-established optimization methods include Tabu Search [16], Monte Carlo optimization [17], Genetic Algorithms [18], Greedy Algorithms [19], and simulated annealing [20]. Furthermore, computational problems tackled by machine learning approaches comprise classification problems, clustering problems and probabilistic graphical model generation.

When coping with large collections of biosequences, many problems require classification and clustering to reduce the dimensionality of the data.

In classification problems, elements belonging to a data set are classified according to a set of features and a set of classification rules. Widely used classifiers are Bayesian methods [21], logistic regression approaches [22], classification trees [23], discriminant analysis [24] and nearest neighbour approaches [25]. The most sophisticated tools to solve a classification problem are neural networks [26] and support vector machines [27].

Clustering problems consist in partitioning a set of samples into subsets with defined differences between them. Clustering problems can be solved by different approaches including partition clustering, hierarchical clustering, or mixture models. Partition clustering approaches focus on obtaining a partition of the dataset given a fixed number of expected clusters, although some methods automatically identify the most appropriate number of clusters. Vector quantization [28] and K-means algorithm [29] are two of the most popular clustering methods. Alternatively, hierarchical methods tackle the clustering problem summarizing data structures as a nested set of partitions described by a dendrogram or tree diagram. The hierarchical data structures may be defined with an agglomerative (create clusters by merging records or groups of them) or divisive approach (create clusters by splitting groups of records of a higher order) [30]. Lastly, clustering problems can be solved also by the mixture methods in which each group of records is considered as described by a different probability distribution. Overall dataset probability distribution is then considered as the result of the superimposition of the single cluster distributions, and clustering is performed based on each sample likelihood being in a specific cluster [31].

Machine learning approaches are adopted to deal with probabilistic graphical models, which support discovering structures in complex data distribution. Probabilistic graphical models describe the multivariate joint probability densities employing the product of terms involving a discrete number of variables [32]. Among probabilistic models, Hidden Markov Models are a

very famous paradigm in Bioinformatics, widely used for data analysis [33]. Probabilistic models also include Bayesian networks [34], characterized by discrete random variables, and Gaussian networks [35], determined by continuous variables, following a gaussian distribution. Machine learning methods are applied in several biological contexts to extract knowledge from the increasing amount of heterogeneous data. One of the most important domains is Genomics. Next-Generation Sequencing technologies (NGS technologies) raised exponentially the amount of data generated and deserving analysis. Machine learning tackled this urgency being able to address various problems such as the extraction of genes structure and location [36-37], the identification of regulatory elements [38] and non-coding RNA genes [39].

While genes are comparable to vectors of information, proteins can be considered the labourers that convert that information into life. In proteomics, computational methods based on machine learning are adopted to solve problems such as 3D protein structure prediction [40]. Recently a very complex architecture of deep neural networks, called AlphaFold (<https://alphafold.ebi.ac.uk/>) was able to partially solve the famous problem known as the “protein folding problem” (how a protein sequence can fold in the three-dimensional space of the solvent). Neural networks can also be adopted for predicting functional annotation [41]. Another field in which machine learning and biology cooperate is System Biology, whose aim is to model processes that take place in the cell. Such processes display an incredible level of complexity that requires computational approaches to describe all the biological networks [42]. Also, Evolution and Phylogenetics took advantage of many computational approaches to uncover knowledge hidden in the differences among genomes belonging to related or unrelated species [43].

Consequently, the application of computational methods to process the broad flow of data drastically increased, as documented by the increase in the number of publications [44]. Text mining was a side effect due to the necessity of extracting specific knowledge from a huge amount of valuable information [45].

In conclusion, new experimental technologies capable of producing large amounts of data raised the challenge of converting the volume of data into useful knowledge. In the biological context, machine learning approaches and, more specifically, neural network-based methods, tackled this problem setting themselves up as a trustworthy solution.

1.3) References

[1] Fillinger S, de la Garza L, Peltzer A, Kohlbacher O, Nahnsen S. Challenges of big data integration in the life sciences. *Anal Bioanal Chem.* 2019;411(26):6791-6800. doi:10.1007/s00216-019-02074-9

- [2] May JC, McLean JA. Advanced Multidimensional Separations in Mass Spectrometry: Navigating the Big Data Deluge. *Annu Rev Anal Chem (Palo Alto Calif)*. 2016;9(1):387-409. doi:10.1146/annurev-anchem-071015-041734
- [3] Krassowski M, Das V, Sahu SK, Misra BB. State of the Field in Multi-Omics Research: From Computational Needs to Data Mining and Sharing. *Front Genet*. 2020;11:610798. doi:10.3389/fgene.2020.610798
- [4] Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship [published correction appears in *Sci Data*. 2019 Mar 19;6(1):6]. *Sci Data*. 2016;3:160018. doi:10.1038/sdata.2016.18
- [5] Leinonen R, Sugawara H, Shumway M; International Nucleotide Sequence Database Collaboration. The sequence read archive. *Nucleic Acids Res*. 2011;39(Database issue):D19-D21. doi:10.1093/nar/gkq1019
- [6] Cochrane G, Karsch-Mizrachi I, Nakamura Y; International Nucleotide Sequence Database Collaboration. The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res*. 2011;39(Database issue):D15-D18. doi:10.1093/nar/gkq1150
- [7] Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. *Nucleic Acids Res*. 2010;38(Database issue):D46-D51. doi:10.1093/nar/gkp1024
- [8] Leinonen R, Akhtar R, Birney E, et al. Improvements to services at the European Nucleotide Archive. *Nucleic Acids Res*. 2010;38(Database issue):D39-D45. doi:10.1093/nar/gkp998
- [9] Kaminuma E, Mashima J, Kodama Y, et al. DDBJ launches a new archive database with analytical tools for next-generation sequence data. *Nucleic Acids Res*. 2010;38(Database issue):D33-D38. doi:10.1093/nar/gkp847
- [10] Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res*. 2021;49(D1):D545-D551. doi:10.1093/nar/gkaa970
- [11] Baldi P. *Deep Learning in Science*. Cambridge University Press. 2021
- [12] Larranaga P, et al. Machine Learning in Bioinformatics. *Briefings in Bioinformatics*. 2005; 7(1): 86-112. doi.org/10.1093/bib/bbk007
- [13] Fomin FV and Kaski P. Exact Exponential Algorithms. *Communications of the ACM*. 2013; 56(3): 80-88. doi.org/10.1145/2428556.2428575
- [14] Yousoff SNM, Baharin A and Abdullah A. A review on optimization algorithm for deep learning method in bioinformatics field. *IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, 2016; pp. 707-711. doi: 10.1109/IECBES.2016.7843542.
- [15] Reali F, Priami C and Marchetti L. Optimization Algorithms for Computational Systems Biology. *Frontiers in Applied Mathematics and Statistics*. 2017; 3:6. doi.org/10.3389/fams.2017.00006
- [16] Mehenni T. Multiple Guide Trees in a Tabu Search Algorithm for the Multiple Sequence Alignment Problem. *Computer Science and Its Applications*. 2015;141-152. doi: 10.1007/978-3-319-19578-0_12
- [17] Krausch N, Barz T, Sawatzki A, et al. Monte Carlo Simulations for the Analysis of Non-linear Parameter Confidence Intervals in Optimal Experimental Design. *Front Bioeng Biotechnol*. 2019;7:122. doi:10.3389/fbioe.2019.00122
- [18] Manning T, Sleator RD, Walsh P. Naturally selecting solutions: the use of genetic algorithms in bioinformatics. *Bioengineered*. 2013;4(5):266-278. doi:10.4161/bioe.23041
- [19] Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA sequences. *J Comput Biol*. 2000;7(1-2):203-214. doi:10.1089/10665270050081478

- [20] Kai Z, Yuting W, Yulin L, Jun L, Juanjuan H. An efficient simulated annealing algorithm for the RNA secondary structure prediction with Pseudoknots. *BMC Genomics*. 2019;20(Suppl 13):979. doi:10.1186/s12864-019-6300-2
- [21] Wilkinson DJ. Bayesian methods in bioinformatics and computational systems biology. *Brief Bioinform*. 2007;8(2):109-116. doi:10.1093/bib/bbm007
- [22] Bewick V, Cheek L and Ball J. Statistics review 14: Logistic regression. *Critical Care*. 2015; 9/1:112-118.
- [23] Chen X, Wang M, Zhang H. The use of classification trees for bioinformatics. *Wiley Interdiscip Rev Data Min Knowl Discov*. 2011;1(1):55-63. doi:10.1002/widm.14
- [24] Lix LM, Sajobi TT. Discriminant analysis for repeated measures data: a review. *Front Psychol*. 2010;1:146. doi:10.3389/fpsyg.2010.00146
- [25] Wang J, Yang B, An Y, et al. Systematic analysis and prediction of type IV secreted effector proteins by machine learning approaches. *Brief Bioinform*. 2019;20(3):931-951. doi:10.1093/bib/bbx164
- [26] Tang, B., Pan, Z., Yin, K. and Khateeb, A., 2019, 'Recent Advances of Deep Learning in Bioinformatics and Computational Biology', *Frontiers in Genetics* 10(214)
- [27] Huang S, Cai N, Pacheco PP, Narrandes S, Wang Y, Xu W. Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. *Cancer Genomics Proteomics*. 2018;15(1):41-51. doi:10.21873/cgp.20063
- [28] Dimopoulou M and Antonini M. Image storage in DNA using Vector Quantization. *EUSIPCO 2020*.
- [29] Oyelade J, Isewon I, Oladipupo F, et al. Clustering Algorithms: Their Application to Gene Expression Data. *Bioinform Biol Insights*. 2016;10:237-253. doi:10.4137/BBI.S38316
- [30] Vijaya , Sinha A and Bateja R. A Review on Hierarchical Clustering Algorithms. *Journal of Engineering and Applied Sciences*. 2017;12: 7501-7507. doi: 10.36478/jeasci.2017.7501.7507
- [31] Steinley D, Brusco MJ. Evaluating mixture modeling for clustering: recommendations and cautions. *Psychol Methods*. 2011;16(1):63-79. doi:10.1037/a0022673
- [32] Airolidi EM. Getting started in probabilistic graphical models. *PLoS Comput Biol*. 2007;3(12):e252. doi:10.1371/journal.pcbi.0030252
- [33] Krogh A, Brown M, Mian IS, Sjölander K, Haussler D. Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol*. 1994;235(5):1501-1531. doi:10.1006/jmbi.1994.1104
- [34] Agrahari R, Foroushani A, Docking TR, et al. Applications of Bayesian network models in predicting types of hematological malignancies. *Sci Rep*. 2018;8(1):6951. doi:10.1038/s41598-018-24758-5
- [35] Bahar I, Atilgan AR, Erman B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold Des*. 1997;2(3):173-181. doi:10.1016/S1359-0278(97)00024-2
- [36] Mathé C, Sagot MF, Schiex T, Rouzé P. Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res*. 2002;30(19):4103-4117. doi:10.1093/nar/gkf543
- [37] Won KJ, Prügél-Bennett A, Krogh A. Training HMM structure with genetic algorithm for biological sequence analysis. *Bioinformatics*. 2004;20(18):3613-3619. doi:10.1093/bioinformatics/bth454
- [38] Aerts S, Van Loo P, Moreau Y, De Moor B. A genetic algorithm for the detection of new cis-regulatory modules in sets of coregulated genes. *Bioinformatics*. 2004;20(12):1974-1976. doi:10.1093/bioinformatics/bth179
- [39] Carter RJ, Dubchak I, Holbrook SR. A computational approach to identify genes for functional RNAs in genomic sequences. *Nucleic Acids Res*. 2001;29(19):3928-3938. doi:10.1093/nar/29.19.3928

- [40] Kuhlman B, Bradley P. Advances in protein structure prediction and design. *Nat Rev Mol Cell Biol.* 2019;20(11):681-697. doi:10.1038/s41580-019-0163-x
- [41] Bernardes JS, Pedreira CE. A review of protein function prediction under machine learning perspective. *Recent Pat Biotechnol.* 2013;7(2):122-141. doi:10.2174/18722083113079990006
- [42] Bower JM, Bolouri, H. *Computational Modeling of Genetic and Biochemical Networks.* MIT Press. 2004
- [43] Azouri D, Abadi S, Mansour Y, Mayrose I, Pupko T. Harnessing machine learning to guide phylogenetic-tree search algorithms. *Nat Commun.* 2021;12(1):1983. doi:10.1038/s41467-021-22073-8
- [44] Elango, B. GROWTH OF SCIENTIFIC PUBLICATIONS: AN ANALYSIS OF TOP TEN COUNTRIES. *Library Philosophy and Practice (e-journal).* 2018 (<http://digitalcommons.unl.edu/libphilprac/1958>).
- [45] Krallinger M, Erhardt RA, Valencia A. Text-mining approaches in molecular biology and biomedicine. *Drug Discov Today.* 2005;10(6):439-445. doi:10.1016/S1359-6446(05)03376-3

2) Specific problems in different omic contexts

In the following, I will focus on two specific biological problems that had been the subjects of my thesis work, done at the CRO Aviano National Cancer Institute (<https://www.cro.it/it>), and within the Biocomputing Group of the Bologna University (www.biocomp.unibo.it), according to my study plan.

2.1) Subtypes characterization of Soft Tissue Sarcomas

Sarcomas represent a heterogeneous group of rare malignancies of mesenchymal nature. The overall incidence of sarcomas is five people per 100,000 per year and are therefore considered rare cancers (rare cancer definition: less than 6 cases per 100,000 per year). Sarcomas account for less than 1% of all adult malignancies and about 20% of paediatric cancers. In 2021, the American Cancer Society estimated more than 13,000 people in the United States diagnosed with sarcoma and more than 5,300 deaths due to this diagnosis. Incidence tends to vary accordingly to age ranging from 15-20 cases per 100,000 per year in the age range 0-20, to 1-2 cases per 100,000 per year in the elderly population [1]. Sarcomas can occur at any anatomic location and are broadly classified as bone and soft-tissue sarcomas [2]. Sarcomas are aggressive neoplasms characterized by local destructive growth, recurrence and distant metastases that usually arise in lungs, liver, bones and brain. Lymph node metastases may be observed, although the hematogenous spread is far more common. Roughly, 30% to 50% of cases metastasize, whereas 20% to 30% of the cases show local recurrence. On average, the overall five years survival rate range between 55% and 65% regardless of stage and histology of the tumours [3]. According to the fifth edition of the World Health Organization (WHO) released in 2020, there are more than 100 different histologic subtypes of sarcomas. WHO classification merges all the major advances generated in the past 20 years and classifies different sarcoma entities based on histomorphology including all the available immunophenotypic and genetic data. WHO divides soft tissue tumours into 11 different macro classes (Adipocytic tumours, Fibroblastic and myofibroblastic tumours, So-called fibrohistiocytic tumours, Vascular tumours, Pericytic tumours, smooth muscle tumours, Skeletal muscle tumours, gastrointestinal stromal tumours, Chondro-osseous tumours, Peripheral nerve sheath tumours and Tumours of uncertain differentiation) and bone tumours in 8 macro classes (Chondrogenic tumours, Osteogenic tumours,

Fibrogenic tumours, Vascular tumours of bone, Osteoclastic giant cell-rich tumours, Notochordal tumours, Other mesenchymal tumours of bone and Haematopoietic neoplasms of bone) [4].

GIST represents one of the most common types of soft tissue sarcomas [5]. GIST is a distinctive mesenchymal neoplasm with variable behaviour but characterized by an absolute tropism for the gastrointestinal tract. They are thought to originate from the interstitial cells of Cajal. Between 50% and 60% of all GISTs arise in the stomach, some 30% of the cases occur in the small intestine, 5% of GISTs arise in the large intestine and about 1% occur in the oesophagus. Rarely, GISTs can occur also in the appendix. Extragastrintestinal GISTs exists too, and they occur predominantly in retroperitoneum, omentum and mesentery, although in most cases they represent metastasis from an unrecognized primary mass [6]. Population-based studies indicate an incidence of 1.1-1.5 cases per 100,000 person-year [7]. However, sub-centimetre GISTs (also called microGISTs) are far more common and are detected in over 20% of elderly individuals [8]. From the molecular point of view, GISTs represent a relatively homogeneous class of lesions characterized by constitutive activation of KIT or PDGFRA tyrosine receptors in over 80% of cases. This activation is caused by Gain of Function (GOF) mutations that affect specific protein domains. Around 70% of GISTs cases involve oncogenic mutations affecting exon 11 of the KIT gene while exon 9 is affected in less than 10% of the cases. Mutations in exons 8, 13 and 17 were observed but in a small subset of cases. Approximately 10% of GISTs harbour PDGFRA mutations mainly in exon 18 even if exons 12 and 14 were observed being involved in cancer onset too. Between 10% and 15% of GIST cases are wild type (WT) for both KIT and PDGFRA instead. These cases are a family of tumours characterized by distinctive molecular pathogenesis and classified nowadays in 1) SDH-deficient GIST; 2) NF1-related GIST and 3) Others. Specifically, gastric WT-GISTs are mainly marked by alterations involving the succinate dehydrogenase (SDH) complex (about half of the cases). About one-third/one-half WT-GISTs carry NF1 gene mutations that lead to the activation of the RAS pathway. Other WT-GISTs were reported to have mutations of BRAF or, more rarely, HRAS, NRAS or PIK3 [9]. Finally, the use of NGS approaches has allowed us to identify rare cases driven by oncogenic fusion proteins such as ETV6-NTRK3 gene fusion [10].

WT-GISTs are currently orphans of effective therapies. Particularly for this kind of malignancies, new therapeutic vulnerabilities can be disclosed thanks to molecular profiling approaches such as NGS mutations and transcriptional analysis, and data analysis with machine learning approaches.

2.2) Functional Analysis of Enzyme Biosequences

Proteins are molecules that convert information contained in genomes to support all the reactions necessary for cell life. Proteins that act as biological catalysts (also known as biocatalysts), regulating the rate at which biochemical reactions proceed in living organisms, are defined enzymes [11]. The first usage of the term “enzyme” dates back to 1878 when the German physiologist Wilhelm Kühne described how yeast produces alcohol from sugars. The word ‘enzyme’ derives from the Greek words *en* (meaning ‘within’) and *zume* (meaning ‘yeast’) [12]. The role of proteins as direct effectors of biochemical reactions or simple carriers of those effectors remained a matter of discussion until 1926 when James. B. Sumner successfully crystallized urease, demonstrating that catalytic activity is associated with protein molecules [13-14]. Hence, enzymes can catalyse the conversion of substrates into products.

The catalytic potential of an enzyme can be expressed by a catalytic constant (k_{cat}) also called turnover rate. This constant represents the numbers of substrate molecules that are converted to product per unit time. The speed at which an enzyme catalyses the conversion of substrates to products and the factors that affect it determines the enzyme kinetics. The kinetics of monomeric enzymes is described by the Michaelis-Menten equation, modelling the speed of the reaction as a function of the substrate concentration [15]. Alternatively, the functional behaviour of multimeric enzymes, like hemoglobin, fits the Hill equation, which models a sigmoidal dependence of the reaction velocity as a function of the substrate concentration [16].

Enzymes possess common names which refers to the substrate and the reaction, they catalyse. During the years, the discovery of new enzymes led to a growing complexity and inconsistency of the enzyme naming system. Starting from 1961, The International Union of Biochemistry and Molecular Biology (IUBMB; <https://iubmb.org/>) addressed this problem providing a systematic approach to the naming of enzymes and performing a simple but still effective data-standardization. The nomenclature is based on a numerical classification of enzymes based on the reaction they catalyse. Enzymes are described by terms consisting of four digits separated by points, called Enzyme Commission (EC) numbers [17]. Each number represent a progressively more detailed description of the reaction catalysed by the enzyme. This nomenclature distributes enzymes into seven functional classes: Oxidoreductases (EC 1), Transferases (EC 2), Hydrolases (EC 3), Lyases (EC 4), Isomerases (EC 5), Ligases (EC 6) and Translocases (EC 7). The first number of the EC nomenclature defines the enzymatic class, and the fourth one describes the exact substrate of a catalysed reaction; intermediate numbers have a different meaning depending on the functional class. The oxidoreductases class includes those enzymes

that catalyse an oxidation and/or reduction reactions transferring hydrogen (H) atoms, oxygen (O) atoms or electrons (e^-) from one molecule to another. The second figure in the code number indicates the group in the donor that undergoes oxidation (-CHOH- group; -CHO group; etc.) while the third number indicates the type of involved acceptor (NAD(P)⁺; cytochrome, etc.). Transferases includes enzymes that transfer a functional group between two molecules. In the classification schema, the second level of the EC number defines the group transferred (methyl-; acyl-; amino- or phosphate group) while the third number provides further information about the group itself (i.e., EC 2.1.1 indicates methyltransferases and EC 2.1.2 indicates hydroxymethyl-transferases).

The third functional class includes enzymes that generate two products from a substrate by hydrolysis. The number in position two determines the nature of the hydrolysed bond (EC 3.1 are esterases while EC 3.2 are glycosylases) and the number in position three specify the nature of the substrate (EC 3.1.1 are carboxylic esterases while EC 3.1.2 are thioesterases). Carrying on, the lyases class includes enzymes that cleave bonds by elimination or by addition of groups to double bonds. The second figure in the EC code indicates the broken bond such as a carbon-carbon bond (EC 4.1) or a carbon-oxygen bond (EC 4.2) while the third figure provides further information about the eliminated group (i.e., CO₂ in EC 4.1.1 and H₂O in EC 4.2.1).

Isomerases (EC 5) are enzymes that cause intramolecular rearrangement like geometric or structural changes of the substrate organization. In this class, the second digit of the EC number indicates the type of isomerism (racemases, epimerases, cis-trans-isomerases, etc.) while the third one specifies the type of substrate.

The sixth enzymatic class includes enzymes that join two molecules by synthesizing a new bond (ATP is a cofactor). The second number of the classification schema indicate the type of bond formed and the third provide further specificity about the bond (it is mainly used for the sub-class EC 6.3 (forming carbon-nitrogen bonds)).

The seventh class of translocases, recently introduced in 2018 [18], includes enzymes that catalyse the movement of ions or molecules across membranes or their separation within them. Inside this class, the subclass defined by the second term designates the types of molecules or ion translocated by the enzyme while the third number provides information about the driving force for the translocation.

One major problem in protein analysis is functional annotation: how to endow with functional and structural features a protein. If the protein is an enzyme, the problem is its classification into one of the seven enzymatic classes. This allows to determine the biochemical reaction it catalyses.

Protein structure and function are related, given that structure determines function. For enzymes, this concept is even more important, due to the relevance of the active and binding sites for catalysis. Each active site is unique in determining the specificity of the reaction and its architecture has been preserved through evolution [11].

In the past, before the advent of structural bioinformatics, protein structure was routinely determined after the functional role of the molecule had already been elucidated. Therefore, structure was used as a molecular framework to explain functional properties of the protein. This approach led to the idea that predicting protein structure from sequence would almost provide automatically functional information [19]. Today, in the post genomic era, the annotation of protein sequences with functional and structural features is a fundamental task to close the gap among hundreds of millions of chains made available by NGS technologies and the much smaller number of proteins with an experimentally characterized biochemical function and 3D structure. Function predictors try to find relations between proteins that allow the transfer of functional information from one to another. Given this task, two important challenges arise: 1) defining the connection between the functional relatedness and the similarities detected and 2), set statistical thresholds for similarity levels. Several experiments demonstrated that function strictly relates to structure while, due to evolution, different sequences adopt the same structure in different organisms [20]. “Orthologous” proteins are indeed likely to have the same or a similar function in different species, while “paralogues” proteins, even though they can possibly maintain a certain level of sequence similarity, may catalyse different biochemical reactions. It is known that similarity between structures can be described as a monotonic function of their sequence similarity [21]. Moreover, from the observation of the space of proteins, we know that sequences of 250 residues or more, who share a sequence similarity of at least 30%, share structural similarities [22]. Hence, knowing that structure is more conserved than sequence in the evolutionary landscape, function predictions based on structural similarity are more reliable than those based on sequence comparison. Given this context, the scientific community developed, in recent years, different algorithms, tools and systems for protein function predictions taking advantage of advanced computer technologies. Available predictors in the field are based on detecting homologies among protein sequences [23-26] and protein structures [27-28], on recognizing similarities among protein-protein interaction (PPI) networks [29-30] and other features [31-32].

2.3) References

- [1] American Cancer Society. About Soft-Tissue Sarcoma. Overview and Types. Online. 2021 (<https://www.cancer.org/content/dam/CRC/PDF/Public/8813.00.pdf>)
- [2] Cancer Genome Atlas Research Network. Cancer Genome Atlas Research Network. Comprehensive and Integrated Genomic Characterization of Adult Soft Tissue Sarcomas. *Cell*. 2017;171(4):950-965.e28. doi:10.1016/j.cell.2017.10.014
- [3] Amankwah EK, Conley AP, Reed DR. Epidemiology and therapies for metastatic sarcoma. *Clin Epidemiol*. 2013;5:147-162. doi:10.2147/CLEP.S28390
- [4] World Health Organization. Soft Tissue and Bone Tumours. 5th Edition. WHO Classification of Tumours Editorial Board. 2020
- [5] Dei Tos AP. Soft Tissue Sarcomas. A Pattern-Based Approach to Diagnosis. Cambridge University Press. 2019
- [6] Rossi S, Miceli R, Messerini L, et al. Natural history of imatinib-naive GISTs: a retrospective analysis of 929 cases with long-term follow-up and development of a survival nomogram based on mitotic index and size as continuous variables. *Am J Surg Pathol*. 2011;35(11):1646-1656. doi:10.1097/PAS.0b013e31822d63a7
- [7] Nilsson B, Bümming P, Meis-Kindblom JM, et al. Gastrointestinal stromal tumors: the incidence, prevalence, clinical course, and prognostication in the preimatinib mesylate era--a population-based study in western Sweden. *Cancer*. 2005;103(4):821-829. doi:10.1002/cncr.20862
- [8] Burningham Z, Hashibe M, Spector L, Schiffman JD. The epidemiology of sarcoma. *Clin Sarcoma Res*. 2012;2(1):14. Published 2012 Oct 4. doi:10.1186/2045-3329-2-14
- [9] Corless CL, Barnett CM, Heinrich MC. Gastrointestinal stromal tumours: origin and molecular oncology. *Nat Rev Cancer*. 2011;11(12):865-878. doi:10.1038/nrc3143
- [10] Brenca M, Rossi S, Polano M, et al. Transcriptome sequencing identifies ETV6-NTRK3 as a gene fusion involved in GIST. *J Pathol*. 2016;238(4):543-549. doi:10.1002/path.4677
- [11] Stryer L, Berg JM and Tymoczko JL. *Biochemistry* (5th ed.). San Francisco: W.H. Freeman. 2002
- [12] Robinson PK. Enzymes: principles and biotechnological applications [published correction appears in *Essays Biochem*. 2015;59:75]. *Essays Biochem*. 2015;59:1-41. doi:10.1042/bse0590001
- [13] Sumner JB. The Isolation and Crystallization of the Enzyme Urease. Preliminary Paper. *J. Biol. Chem*. 1926;69: 435-441.
- [14] Simoni RD, Hill RH, Vaughan M. Urease, the first crystalline enzyme and the proof that enzymes are proteins: the work of James B. Sumner. *J Biol Chem*. 2002;277(35):23e.
- [15] Srinivasan B. A guide to the Michaelis-Menten equation: steady state and beyond [published online ahead of print]. *FEBS J*. 2021;10.1111/febs.16124. doi:10.1111/febs.16124.
- [16] Hill AV. The possible effect of the aggregation of the molecules of hemoglobin on its dissociation curves. *Proceeding of Physiological Society*; 22th January 1910.
- [17] Webb EC. Enzyme nomenclature 1992: recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes. Academic Press. 1992
- [18] Tripton L. Translocases (EC 7): A new EC Class', *ExplorEnz – The enzyme Database*. Online. 2018 (<https://www.enzyme-database.org/news.php>)

- [19] Loewenstein Y, Raimondo D, Redfern OC, et al. Protein function annotation by homology-based inference. *Genome Biol.* 2009;10(2):207. doi:10.1186/gb-2009-10-2-207
- [20] Whisstock JC, Lesk AM. Prediction of protein function from protein sequence and structure. *Q Rev Biophys.* 2003;36(3):307-340. doi:10.1017/s0033583503003901
- [21] Chothia C and Lesk AM. The relation between the divergence of sequence and structure in proteins. *EMBO J.* 1986;5(4):823-826.
- [22] Sander C, Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins.* 1991;9(1):56-68. doi:10.1002/prot.340090107
- [23] Yao S, You R, Wang S, Xiong Y, Huang X, Zhu S. NetGO 2.0: improving large-scale protein function prediction with massive sequence, text, domain, family and network information. *Nucleic Acids Res.* 2021;49(W1):W469-W475. doi:10.1093/nar/gkab398
- [24] Hakala K, Kaewphan S, Bjorne J, et al. Neural Network and Random Forest Models in Protein Function Prediction [published online ahead of print]. *IEEE/ACM Trans Comput Biol Bioinform.* 2020;PP:10.1109/TCBB.2020.3044230. doi:10.1109/TCBB.2020.3044230
- [25] Piovesan D, Tosatto SCE. INGA 2.0: improving protein function prediction for the dark proteome. *Nucleic Acids Res.* 2019;47(W1):W373-W378. doi:10.1093/nar/gkz375
- [26] Kulmanov M, Hoehndorf R. DeepGOPlus: improved protein function prediction from sequence [published correction appears in *Bioinformatics*]. *Bioinformatics.* 2020;36(2):422-429. doi:10.1093/bioinformatics/btz595
- [27] Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER Suite: protein structure and function prediction. *Nat Methods.* 2015;12(1):7-8. doi:10.1038/nmeth.3213
- [28] Zhang C, Zheng W, Freddolino PL, Zhang Y. MetaGO: Predicting Gene Ontology of Non-homologous Proteins Through Low-Resolution Protein Structure Prediction and Protein-Protein Network Mapping. *J Mol Biol.* 2018;430(15):2256-2265. doi:10.1016/j.jmb.2018.03.004
- [29] Kulmanov M, Khan MA, Hoehndorf R, Wren J. DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics.* 2018;34(4):660-668. doi:10.1093/bioinformatics/btx624.
- [30] You R, Zhang Z, Xiong Y, Sun F, Mamitsuka H, Zhu S. GOLabeler: improving sequence-based large-scale protein function prediction by learning to rank. *Bioinformatics.* 2018;34(14):2465-2473. doi:10.1093/bioinformatics/bty130
- [31] Kahanda I and Ben-Hur A. Gostruct 2.0: Automated protein function prediction for annotated proteins. *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (New York, NY).* 2018;60-66.
- [32] Sureyya Rifaioğlu A, Doğan T, Jesus Martin M, Cetin-Atalay R, Atalay V. DEEPred: Automated Protein Function Prediction with Multi-task Feed-forward Deep Neural Networks. *Sci Rep.* 2019;9(1):7344. doi:10.1038/s41598-019-43708-3.

3) Objectives and Main Results

The world of Computational Biology and Bioinformatics presently integrates many different expertise, including computer science and electronic engineering. A major aim in Data Science is the development and tuning of specific computational approaches to interpret the complexity of Biology. Molecular biologists and medical doctors heavily rely on an interdisciplinary expert capable of understanding the biological background to apply algorithms for finding optimal solutions to their problems. With this problem-solving orientation, I was involved in two basic research fields: Cancer Genomics and Enzyme Proteomics.

As to Cancer Genomics, I strictly collaborated with medical doctors and researchers, within the group of Dr. Roberta Maestro at the CRO Aviano National Cancer Institute. As to Functional Proteomics, I collaborated with the Biocomputing group of the University of Bologna (UNIBO) (tutor: Prof. Rita Casadio, www.biocomp.unibo.it). My financial support was indeed due to a framework-collaboration between CRO and UNIBO.

For this reason, what I developed and implemented can be considered a general effort to help data analysis both in Cancer Genomics and in Enzyme Proteomics, focusing on enzymes which catalyse all the biochemical reactions in cells.

Specifically, as to Cancer Genomics I contributed to the characterization of intratumoral immune microenvironment in gastrointestinal stromal tumours (GISTs) correlating immune cell population levels with tumour subtypes. I was involved in the setup of strategies for the evaluation and standardization of different approaches for fusion transcript detection in sarcomas that can be applied in routine diagnostic. This was part of a coordinated effort of the Sarcoma working group of ACC (Alleanza Contro il Cancro; <https://www.alleanzacontroilcancro.it/en/>). Finally, I developed an internal application to perform standardized bioinformatics analysis with a user-friendly interface for biologist (unpublished). This application is tailored to cope with an internal database of biosamples and data from NGS technologies that I created on purpose.

As to Enzyme Proteomics, I generated a derived database collecting all the human proteins and enzymes which are known to be associated to genetic disease. I curated the data search in freely available databases such as PDB (<https://www.rcsb.org/>), UniProt (<https://www.uniprot.org/>), Humsavar (<https://www.uniprot.org/docs/humsavar>), Clinvar (<https://www.ncbi.nlm.nih.gov/clinvar/>) and I was responsible of searching, updating, and handling the information content, and computing statistics. I also developed a web server, BENZ,

which allows researchers to annotate an enzyme sequence with the corresponding Enzyme Commission number (EC number), the important feature fully describing the catalysed reaction (see 2.1). More to this, I greatly contributed to the characterization of the enzyme-genetic disease association, for a better classification of the metabolic genetic diseases. The papers I contributed and the work I presented to national and international meetings (see LAA) fully describe all the different approaches I developed and that I extensively describe in the Result section.

4) Materials and Methods

4.1) Genomic and Transcriptomic data

- a) In-house sarcoma samples were molecularly profiled by using Target-seq and RNA-seq approaches as detailed in published articles (see Ref 1 and 3, LAA). Briefly, for target analysis samples were first profiled for KIT and PDGFRA mutations by Sanger sequencing. Sample scoring negative were further profiled by using a targeted NGS panel that covered the sequence of the following genes: KIT, PDGFRA, BRAF, NF1, SDH A-D, H/K/N RAS. For RNA-sequencing analysis, more than 60 million of reads per sample (paired-ends) were generated on Illumina HiSeq 1500 platform. Besides in house-sequenced samples collected from available biobanks, publicly available data from high throughput technologies were collected from different sources. Controlled access samples, namely those samples for which a special permission must be granted prior downloading, were retrieved from the National Cancer Institute (NCI) Genomic Data Commons (GDC) Data Portal (<https://portal.gdc.cancer.gov/>). Freely accessible samples were downloaded from the Sequence Read Archive (SRA; <https://www.ncbi.nlm.nih.gov/sra>) [1] and the European Nucleotide Archive (ENA; <https://www.ebi.ac.uk/ena/browser/home>). Human reference genome (build 38, patch 13; GRCh38.p13; https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.39) adopted for analysis was downloaded from the Ensembl database (<https://www.ensembl.org/index.html>) [2].
- b) Gene annotations were collected from the GENCODE consortium (v.33; <https://www.gencodegenes.org/human/>) [3] and from the GeneCards portal (<https://www.genecards.org/>) [4].

4.2) Proteomic Databases

- a) Protein sequences, functional annotations, and positions in the sequence of relevant sites (namely active sites, ligand binding sites and metal binding sites) were collected from the UniProt KnowledgeBase (<https://www.uniprot.org/>) [5], the largest collection of sequences from organisms in Life Science. BRENDA (<https://www.brenda-enzymes.org/>) and ENZYME (<https://enzyme.expasy.org/>) [6-7] were used as sources to further characterize the biochemical reactions catalysed by the enzymes collected from UniProt. Structural information of proteins, together with spatial coordinates of 3D

structures, were retrieved from the Protein Data Bank (PDB; <https://www.rcsb.org/>) [8]. Protein family models and conserved structural domains mapped along protein sequences were collected from the Protein Family (Pfam; <http://pfam.xfam.org/>) database [9], collecting HMM models of protein domains. Further, additional information for protein families were downloaded from the InterPro database (<https://www.ebi.ac.uk/interpro/>) [10].

- b) Sequence annotations of biological processes, molecular functions and metabolic pathways were collected from KEGG (<https://www.genome.jp/kegg/>) [11], REACTOME (<https://reactome.org>) [12] and Gene Ontology (GO) (<http://geneontology.org/>) [13]. Due to limitations in the download of the mapping between EC numbers and KEGG pathways, this information was collected by the means of web scraping techniques implemented in python using the dedicated library BeautifulSoup [14].
- c) Disease names associated to human enzymes were fetched from eDGAR (http://edgar.biocomp.unibo.it/gene_disease_db/) [15] and further characterized with terms from the Online Mendelian Inheritance in Man (OMIM; <https://www.omim.org>) and the Mondo Disease Ontology (<http://obofoundry.org/ontology/mondo.html>) [16], with the EMBL-EBI Ontology Lookup Service (OLS; <https://www.ebi.ac.uk/ols/index>). Genetic variations reported to associate enzymes to diseases were downloaded from Humsavar (<https://www.uniprot.org/docs/humsavar>), ClinVar (<https://www.ncbi.nlm.nih.gov/clinvar/>) [17], and curated DisGeNet (<https://www.disgenet.org/>) [18].
- d) Interatomic data, such as protein-protein interactions, genetic interactions and chemical interactions were fetched from IntAct (<https://www.ebi.ac.uk/intact/>) [19] and BioGrid (<https://thebiogrid.org/>) [20], major resources for interactomic data.

4.3) Tools for Genomic and Transcriptomic Analysis

- a) RNA-sequencing data quality was assessed with FastQC v.0.11.8 [21] and its output was further organized by means of MultiQC v.1.7 [22]. Sample sequences were pre-processed with Trimmomatic v.0.38 [23] and then aligned with the Spliced Transcripts Alignment to a Reference (STAR) software, v.2.7.0e [24]. Aligned sequences were merged and filtered using Samtools v.1.9 [25], and then visualized with the Integrative Genome Viewer (IGV) [26]. Gene expression levels were calculated by means of Cufflinks v.2.2.1 [27], HTSeq v.0.13.5 [28] and Salmon v.1.4.0 [29]. Transcript counts

(detecting transcript concentrations), generated by Salmon, were reduced to the gene level by means of the tximport package [30] in R (<https://www.r-project.org/>), based on the GENCODE annotation v.33. Differential expression analysis over the generated gene counts were performed using the DeSEQ2 package v.3.3 [31] in R.

- b) Functional annotation upon identification of differentially expressed genes were executed by means of Gene Set Enrichment Analysis [32] and Ingenuity Pathway Analysis (IPA; QIAGEN Inc., <https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis>). Abundance of tumour-infiltrating immune cells was computed by a Single Sample Gene Set Enrichment Analysis (ssGSEA) [32] implement in R package gsva [33], given the prognostic signature for the tumour immune microenvironment from Şenbabaoğlu et al. [34]. The Immune Infiltration Score (IIS) was computed as the mean of the standardized values for all the immune cell types while the T-Cells Immune Score (TIS) was defined over a limited subset of cell types including: CD8 T, T helper, T, T central and effector memory, Th1, Th2, Th17, and Treg cells. Additionally, enrichment scores were also computed for Interferon gamma (IFN- γ) [35]. Immune cell populations presence was estimated by deconvolutional approaches [36] such as CIBERSORT [37] and MCP-counter [38]. Patient specific Human Leukocyte antigens (HLA) class I alleles were predicted by PHLAT v.1.0 [39]. Then, NetMHCpan v4.0 [40] was applied to predict the binding specificity of peptides, resulting from mutated oncogenes, on the patient-matched Major Histocompatibility Complex (MHC) of class I.
- c) Correlations between gene expression levels and immune population levels were calculated by the means of the Spearman's rank method [41] while differences among the groups of Gastrointestinal Stromal Tumours (GISTs) based on their immune context were assessed via a Mann-Whitney U rank-sum test [42]. Evaluation of potential susceptibility of GISTs to immunomodulatory-based treatments were also investigated computing the Immunophenoscore (IPS) [43] and the cytolytic score (CYT) [44] per sample. Principal Component analysis (PCA) [45] and unsupervised hierarchical clustering [46] were used to evaluate differences between transcriptional profiles of samples.
- d) Fusion transcripts in samples were predicted by means of Arriba v.2.0.0 [47], Pizzly [48], which heavily relies on Kallisto v.0.46.1 [49], STAR-fusion v.1.9.1 [50] and FusionCatcher v.1.33 [51].

4.4) Tools for Proteomic Analysis

- a) Multiple sequence alignment (MSA) of clusters of proteins was performed by means of Clustal Omega v.1.2.4 [52].
- b) HMMER suite v.3.3.2 [53] was used to generate predictive Hidden Markov Models (HMMs) from the results of MSA. An intensive pairwise comparison of protein sequences was also computed with the Basic Local Alignment Searching Tool (BLAST) v.2.11.0 [54].
- c) To the aim of functionally characterize proteins starting from their sequence, multiple tools were applied such as ECPred [55], DEEPre [56] and EFICAz2.5 [57], for predicting the EC number.
- d) ISPRED4 (Interaction Site PREDictions v.4) [58] was applied to predict putative interaction sites on protein surfaces while solvent accessibility was evaluated with the DSSP program (https://swift.cmbi.umcn.nl/gv/dssp/DSSP_3.html).
- e) INPS (Impact of non-synonymous variations on Protein Stability) [59] and its variant INPS-MD [60] were used to predict the thermodynamic free energy change of single point variations over solvent accessible residues. To explore the effect of variants on protein stability, a consensus method involving the computation of the Gibbs free energy change ($\Delta\Delta G$) by means of three different tools was applied. Tools concurring in the consensus method were: INPS-MD [60], PopMuSiC2 [61] and FoldX [62]. Despite all these methods perform the same task, they adopt different approaches. In fact, INPS-MD is based on a machine learning approach while FoldX is based on statistical potentials and PopMuSiC2 on a combination of both approaches. For the consensus, we considered as destabilizing the variations for which at least two methods predicted a $\Delta\Delta G$ lower or equal to -1 kcal/mol. To assess differences of data distributions against background one, an FDR-corrected Chi-square test was applied. Furthermore, log-odds scores were calculated to easily compare and visualize these differences. The Mann-Whitney U test [42] was used to test for differences between datasets.

4.5) New tools developed for Genomic and Transcriptomic Analysis

- a) I organized the tools discussed above (4.3) in a user-friendly pipeline that wraps phase specific bash scripts (i.e., quality check phase, pre-processing phase, alignment phase, etc.) in a class data object structured in a python script. By means of the python library tkinter [63], I also developed a graphical user interface (GUI) for an easy interaction

with the pipeline. The pipeline automatically interacts with an internal database that collects in-house sequenced samples and results derived from data analysis.

- b) The internal database was developed combining scripts written in the JavaScript language [64] and records stored in json [65] format due to technical limitations. After each run, the pipeline automatically updates the internal database with the newly generated results. The pipeline is designed to perform also variants calling by implementing the Genome Analysis Toolkit (GATK) suite [66].

4.6) New tools developed for Proteomic Analysis

- a) In the proteomics context, I developed the Bologna ENZYme (BENZ) Web Server (<https://benzdb.biocomp.unibo.it/>; see Ref 4, LAA).
- b) First, enzymes were clustered starting from a graph building procedure standing out from Profiti et al. [67]. Starting from the entire UniProtKB, sequences from Swiss-Prot and TrEMBL were respectively compared by means of BLAST v.2.11.0 [68] to search for proteins pairs sharing a sequence identity (SI) greater or equal to 40% on an alignment coverage of at least 90%. A graph is built connecting pairs that fulfil both the coverage, and the identity constrains. Then clusters are defined isolating the connected component of the graph. From this background, only clusters containing sequences associated to complete EC numbers were retained.
- c) Then, for each cluster a representative model was generated after multiple sequence alignment. For technical reasons, cluster HMMs with a length higher than 5000 residues were discarded. Models of the clusters defined as described above (4.4) were then collected into a single database-like structure using the HMMER suite tool *hmmcompress*.
- d) Enzymes in clusters were further sub-clustered based on the biochemical reaction they catalyse (namely the annotated EC number/s) and of the architecture they display (namely the ordered set of Pfam domains mapped along the sequence). Thus, per cluster a reference protein was selected. Among proteins belonging to a cluster, reference sequence was selected based on the following criteria: 1) 3D structure available in PDB (mandatory for TrEMBL enzymes); 2) Highest annotation score in UniProt; 3) Complete EC number and functional annotation; 4) Available Pfam architecture. In such way, clusters were linked to EC numbers via protein architectures resulting in clusters univocally associated to a reference sequence and clusters associated to two or more

reference sequences (respectively defined as “Gold Clusters” and “Blue Clusters” in BENZ).

- e) BENZ is built on top of a PostgreSQL v.12.8 (<https://www.postgresql.org>) database hosted on an in-house server located in Bologna. The functional prediction algorithm was implemented in python and the webserver is freely accessible to users thanks to a front end developed through the use of HTML, CSS and JavaScript languages. BENZ is optimized to work with common web browsers and its functionality was fully tested in Chrome 88.0, Firefox 83.0, Edge 88.0 and Safari 14.0.

4.7) References

- [1] Leinonen R, Sugawara H, Shumway M; International Nucleotide Sequence Database Collaboration. The sequence read archive. *Nucleic Acids Res.* 2011;39(Database issue):D19-D21. doi:10.1093/nar/gkq1019
- [2] Howe KL, Achuthan P, Allen J, et al. Ensembl 2021. *Nucleic Acids Res.* 2021;49(D1):D884-D891. doi:10.1093/nar/gkaa942
- [3] Frankish A, Diekhans M, Ferreira AM, et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* 2019;47(D1):D766-D773. doi:10.1093/nar/gky955
- [4] Stelzer G, Rosen N, Plaschkes I, et al. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Curr Protoc Bioinformatics.* 2016;54:1.30.1-1.30.33. doi:10.1002/cpbi.5
- [5] UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* 2021;49(D1):D480-D489. doi:10.1093/nar/gkaa1100
- [6] Chang A, Scheer M, Grote A, Schomburg I, Schomburg D. BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009. *Nucleic Acids Res.* 2009;37(Database issue):D588-D592. doi:10.1093/nar/gkn820
- [7] Bairoch A. The ENZYME database in 2000. *Nucleic Acids Res.* 2000;28(1):304-305. doi:10.1093/nar/28.1.304
- [8] Burley SK, Bhikadiya C, Bi C, et al. RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res.* 2021;49(D1):D437-D451. doi:10.1093/nar/gkaa1038
- [9] Mistry J, Chuguransky S, Williams L, et al. Pfam: The protein families database in 2021. *Nucleic Acids Res.* 2021;49(D1):D412-D419. doi:10.1093/nar/gkaa913
- [10] Blum M, Chang HY, Chuguransky S, et al. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* 2021;49(D1):D344-D354. doi:10.1093/nar/gkaa977
- [11] Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res.* 2021;49(D1):D545-D551. doi:10.1093/nar/gkaa970
- [12] Jassal B, Matthews L, Viteri G, et al. The reactome pathway knowledgebase. *Nucleic Acids Res.* 2020;48(D1):D498-D503. doi:10.1093/nar/gkz1031

- [13] Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.* 2021;49(D1):D325-D334. doi:10.1093/nar/gkaa1113
- [14] Richardson L. Beautiful soup documentation. April. 2007;
- [15] Babbi G, Martelli PL, Profiti G, Bovo S, Savojardo C, Casadio R. eD GAR: a database of Disease-Genes Associations with annotated Relationships among genes. *BMC Genomics.* 2017;18(Suppl 5):554. doi:10.1186/s12864-017-3911-3
- [16] Mungall, Christopher J., et al. 2017 The Monarch Initiative: An Integrative Data and Analytic Platform Connecting Phenotypes to Genotypes across Species. *Nucleic Acids Research* 45 (D1): D712–22.
- [17] Landrum MJ, Lee JM, Benson M, et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* 2018;46(D1):D1062-D1067. doi:10.1093/nar/gkx1153
- [18] Piñero J, Bravo À, Queralt-Rosinach N, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* 2017;45(D1):D833-D839. doi:10.1093/nar/gkw943
- [19] Orchard S, Ammari M, Aranda B, et al. The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* 2014;42(Database issue):D358-D363. doi:10.1093/nar/gkt1115
- [20] Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* 2006;34(Database issue):D535-D539. doi:10.1093/nar/gkj109
- [21] Andrews, S. FastQC: A Quality Control Tool for High Throughput Sequence Data. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>, 2010
- [22] Ewels P, Magnusson M, Lundin S, Källér M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics.* 2016;32(19):3047-3048. doi:10.1093/bioinformatics/btw354
- [23] Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* 2014;30(15):2114-2120. doi:10.1093/bioinformatics/btu170
- [24] Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* 2013;29(1):15-21. doi:10.1093/bioinformatics/bts635
- [25] Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* 2009;25(16):2078-2079. doi:10.1093/bioinformatics/btp352
- [26] Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. *Nat Biotechnol.* 2011;29(1):24-26. doi:10.1038/nbt.1754
- [27] Trapnell C, Williams BA, Pertea G, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol.* 2010;28(5):511-515. doi:10.1038/nbt.1621
- [28] Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics.* 2015;31(2):166-169. doi:10.1093/bioinformatics/btu638
- [29] Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods.* 2017;14(4):417-419. doi:10.1038/nmeth.4197
- [30] Sonesson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res.* 2015;4:1521. doi:10.12688/f1000research.7563.2

- [31] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 2014;15(12):550. doi:10.1186/s13059-014-0550-8
- [32] Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A.* 2005;102(43):15545-15550. doi:10.1073/pnas.0506580102
- [33] Hänzelmann S, Castelo R, Guinney J. GSEA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics.* 2013;14:7. doi:10.1186/1471-2105-14-7
- [34] Şenbabaoğlu Y, Gejman RS, Winer AG, et al. Tumour immune microenvironment characterization in clear cell renal cell carcinoma identifies prognostic and immunotherapeutically relevant messenger RNA signatures [published correction appears in *Genome Biol.* 2017;18(1):46]. *Genome Biol.* 2016;17(1):231. doi:10.1186/s13059-016-1092-z
- [35] Ayers M, et al. IFN- γ -related mRNA profile predicts clinical response to PD-1 blockade. *J Clin Invest.* 2017;127(8):2930–2940.
- [36] Shen-Orr SS, Gaujoux R. Computational deconvolution: extracting cell type-specific information from heterogeneous samples. *Curr Opin Immunol.* 2013;25(5):571-578. doi:10.1016/j.coi.2013.09.015
- [37] Newman AM, Liu CL, Green MR, et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods.* 2015;12(5):453-457. doi:10.1038/nmeth.3337
- [38] Becht E, Giraldo NA, Lacroix L, et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression [published correction appears in *Genome Biol.* 2016;17(1):249]. *Genome Biol.* 2016;17(1):218. doi:10.1186/s13059-016-1070-5
- [39] Bai Y, Wang D, Fury W. PHLAT: Inference of High-Resolution HLA Types from RNA and Whole Exome Sequencing. *Methods Mol Biol.* 2018;1802:193-201. doi:10.1007/978-1-4939-8546-3_13
- [40] Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *J Immunol.* 2017;199(9):3360-3368. doi:10.4049/jimmunol.1700893
- [41] Spearman, C. (1904). The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 15(1), 72–101. <https://doi.org/10.2307/1412159>
- [42] H. B. Mann, D. R. Whitney "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other," *The Annals of Mathematical Statistics*, Ann. Math. Statist. 18(1), 50-60, (March, 1947)
- [43] Charoentong P, Finotello F, Angelova M, et al. Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. *Cell Rep.* 2017;18(1):248-262. doi:10.1016/j.celrep.2016.12.019
- [44] Rooney MS, Shukla SA, Wu CJ, Getz G, Hacohen N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell.* 2015;160(1-2):48-61. doi:10.1016/j.cell.2014.12.033
- [45] Karl Pearson F.R.S. . (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11), 559–572. <https://doi.org/10.1080/14786440109462720>
- [46] Maimon, Oded; Rokach, Lior (2006). "Clustering methods". *Data Mining and Knowledge Discovery Handbook*. Springer. pp. 321–352.

- [47] Uhrig S, Ellermann J, Walther T, et al. Accurate and efficient detection of gene fusions from RNA sequencing data. *Genome Res.* 2021;31(3):448-460. doi:10.1101/gr.257246.119
- [48] Melsted P, Hateley S, Joseph IC, Pimentel H, Bray N, Pachter L., Fusion detection and quantification by pseudoalignment. *BioRxiv.* 166322. (2017). doi: 10.1101/166322
- [49] Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification [published correction appears in *Nat Biotechnol.* 2016;34(8):888]. *Nat Biotechnol.* 2016;34(5):525-527. doi:10.1038/nbt.3519
- [50] Haas BJ, Dobin A, Li B, Stransky N, Pochet N, Regev A. Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biol.* 2019;20(1):213. doi:10.1186/s13059-019-1842-9
- [51] Nicorici D, Satalan M, Edgren H, Kangaspeska S, Murumagi A, Kallioniemi O, Virtanen S, Kilkku O, FusionCatcher – a tool for finding somatic fusion genes in paired-end RNA-sequencing data, *bioRxiv*, Nov. 2014, DOI:10.1101/011650
- [52] Sievers F, Wilm A, Dineen D, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol.* 2011;7:539. doi:10.1038/msb.2011.75
- [53] Eddy SR. Accelerated Profile HMM Searches. *PLoS Comput Biol.* 2011;7(10):e1002195. doi:10.1371/journal.pcbi.1002195
- [54] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;215(3):403-410. doi:10.1016/S0022-2836(05)80360-2
- [55] Dalkiran A, Rifaioglu AS, Martin MJ, Cetin-Atalay R, Atalay V, Doğan T. ECPred: a tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature. *BMC Bioinformatics.* 2018;19(1):334. doi:10.1186/s12859-018-2368-y
- [56] Li Y, Wang S, Umarov R, et al. DEEPre: sequence-based enzyme EC number prediction by deep learning. *Bioinformatics.* 2018;34(5):760-769. doi:10.1093/bioinformatics/btx680
- [57] Kumar N, Skolnick J. EFICAz2.5: application of a high-precision enzyme function predictor to 396 proteomes. *Bioinformatics.* 2012;28(20):2687-2688. doi:10.1093/bioinformatics/bts510
- [58] Savojardo C, Fariselli P, Martelli PL, Casadio R. ISPREd4: interaction sites PREdiction in protein structures with a refining grammar model. *Bioinformatics.* 2017;33(11):1656-1663. doi:10.1093/bioinformatics/btx044
- [59] Fariselli P, Martelli PL, Savojardo C, Casadio R. INPS: predicting the impact of non-synonymous variations on protein stability from sequence. *Bioinformatics.* 2015;31(17):2816-2821. doi:10.1093/bioinformatics/btv291
- [60] Savojardo, C.; Fariselli, P.; Martelli, P.L.; Casadio, R. INPS-MD: a web server to predict stability of protein variants from sequence and structure. *Bioinformatics*, 2016, 32, 2542-2544. doi: 10.1093/bioinformatics/btw192.
- [61] Pucci, F.; Bernaerts, K.V.; Kwasigroch, J.M.; Rooman, M. Quantification of biases in predictions of protein stability changes upon mutations. *Bioinformatics*, 2018, 34, 3659-3665. doi: 10.1093/bioinformatics/bty348.
- [62] Schymkowitz, J.; Borg, J.; Stricher, F.; Nys, R.; Rousseau, F.; Serrano, L. The FoldX web server: an online force field. *Nucleic Acids Res.*, 2004, 33, W382–W388. doi: 10.1093/nar/gki387
- [63] Lundh F. An introduction to tkinter. URL: www.pythonware.com/library/tkinter/introduction/index.htm. 1999;
- [64] Flanagan D.; JavaScript: the definitive guide. ";Reilly Media, Inc." 2006.

- [65] Pezoa, F., Reutter, J. L., Suarez, F., Ugarte, Martin, & Vrgo c, Domagoj. (2016). Foundations of JSON schema. In Proceedings of the 25th International Conference on World Wide Web (pp. 263–273).
- [66] Van der Auwera GA & O'Connor BD. (2020). Genomics in the Cloud: Using Docker, GATK, and WDL in Terra (1st Edition). O'Reilly Media.
- [67] Profiti,G., Martelli,P.L. and Casadio,R. (2017) The Bologna Annotation Resource (BAR 3.0): improving protein functional annotation, *Nucleic Acids Res.*, 45, W285–W290.
- [68] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;215(3):403-410. doi:10.1016/S0022-2836(05)80360-2

5) Results and Discussions

In the following section, the main results of my thesis work are discussed with reference to the main topics I dealt with: Cancer Genomics and Enzyme Proteomics. The section primary focus on the data that were object of scientific publications. The list of published papers is in LAA, and the papers are collected in the Appendix for references and further reading.

5.1) Cancer Genomics

5.1.1) Characterization of the immune microenvironment in GISTs

The presence of immune infiltrate in tumours has recently been gained attention due to the efficacy demonstrated by therapies targeting immune checkpoint modules and effector [1]. The entity and type of immune infiltrate in GIST, and in wild type-GIST that are currently orphan of targeted therapies, was poorly defined. To shed light on this issue, as a Bioinformatician in the unit of Oncogenetics and Functional Oncogenomics at the “CRO Aviano”, I collaborated in the exploration of transcriptomic data to decipher the intra-tumoral immune infiltrate in GISTs (see Ref 3, LAA).

To this end, 82 primary untreated gastrointestinal stromal tumours (GISTs) were retrieved from pathological files of collaboration centres (Table 5.1 from Ref 3, LAA).

Table 5.1. GITS Cohorts

	IHC cohort (38 cases)	RNA-seq cohort (77 cases)
	No. (%)	No. (%)
Sex		
Male	18 (47%)	40 (52%)
Female	20 (53%)	37 (48%)
Location		
Stomach	18 (47%)	43 (56%)
Small Intestine	19 (50%)	34 (44%)
Esophagus	1 (3%)	0
Type		
miniGIST	3 (8%)	15 (19%)
GIST	35 (92%)	62 (81%)
Mutations		
KIT	26 (69%)	47 (61%)
PDGFRA	5 (13%)	15 (20%)
BRAF	3 (8%)	3 (4%)
NF1	4 (10%)	7 (9%)
Unknown	0	5 (6%)

Detailed characterization of the analysed GIST cohorts (Ref 3, LAA).

Expert pathologists provided sample diagnosis based on tumour morphology. Samples were divided on the basis of size, mitotic index (MI: the ratio between the proportion of cells undergoing mitosis over the total number of cells) and driver mutations (KIT, PDGFRA, NF1, BRAF and unknown). 57 GISTs (diameter greater or equal than 2 cm and MI > than 5 mitoses in 5 mm²) and 25 miniGISTs were profiled.

DNA and RNA were extracted from the tumour biosamples that were selected on the basis of tumour cellularity (> 70%). Oncogenic driver mutations were identified by either Sanger sequencing or NGS targeted sequencing. Transcriptome analysis was performed on a subset of 77 tumour samples; 47 KIT mutated, 15 PDGFRA mutated and the remaining 15 wild type for KIT and PDGFRA (3 BRAF mutated, 7 NF1 mutated and 5 whose driver mutation was unknown).

Functional annotation of the expressed genes was done to identify biological processes and molecular signature enriched in the different GIST subsets known to have clinical impact (intestinal vs gastric; overt vs mini GISTs, K/P-mutated vs wild type GISTs; KIT vs PDGFRA mutated GISTs in stomach and K-mutated vs K-wild type GISTs in intestine). Gene Set Enrichment Analysis (GSEA; see materials and methods par. 4.3 sect. *b*) and Ingenuity Pathway Analysis (IPA; see materials and methods par. 4.3 sect. *b*) revealed that immune system signatures (Lymphocyte stimulation, IFN- γ -mediated signalling and T-cell selection) were significantly enriched in K/P-mutated tumours with respect to K/P-wild type ones, a phenomenon particularly evident in the intestinal subset where wild type GIST are more represented. Also, PDGFRA mutated GIST were more infiltrated compared to KIT mutated tumours of the same location (gastric tumours). Finally, overt GIST featured a greater degree of immune infiltration compared to the benign counterpart (miniGIST). To gain insights into the specific characteristics of tumour immune microenvironment, single sample Gene Set Enrichment Analysis (ssGSEA; see materials and methods par. 4.3 sect. *b*) was employed to characterize immune cell infiltrate, specifically the representation of 24 immune cell populations as well as an overall estimate of immune infiltration (IIS and TIS, computed as detailed in par. 4.3 sect. *b*) were determined by using the method developed by Şenbabaoğlu et al [2] (detail in par. 4.3 sect. *b*). Mann-Whitney *U* test (see par. 4.3 sect. *c*) revealed that the distributions of IIS, TIS as well as a measure of immune T cell activation, i.e. the cytolytic score (CYT; geometric mean of two immune effector molecules: granzyme A and perforin) [3] were significantly different (p-value < 0.05) in the above mentioned GIST categories (K/P-mutated vs K/P-wild type, intestinal K-mutated vs K-wild type and gastric K-mutated vs P-mutated). Particularly striking the difference between K/P-mutated and K/P-wild type GISTs (see Figure 5.1).

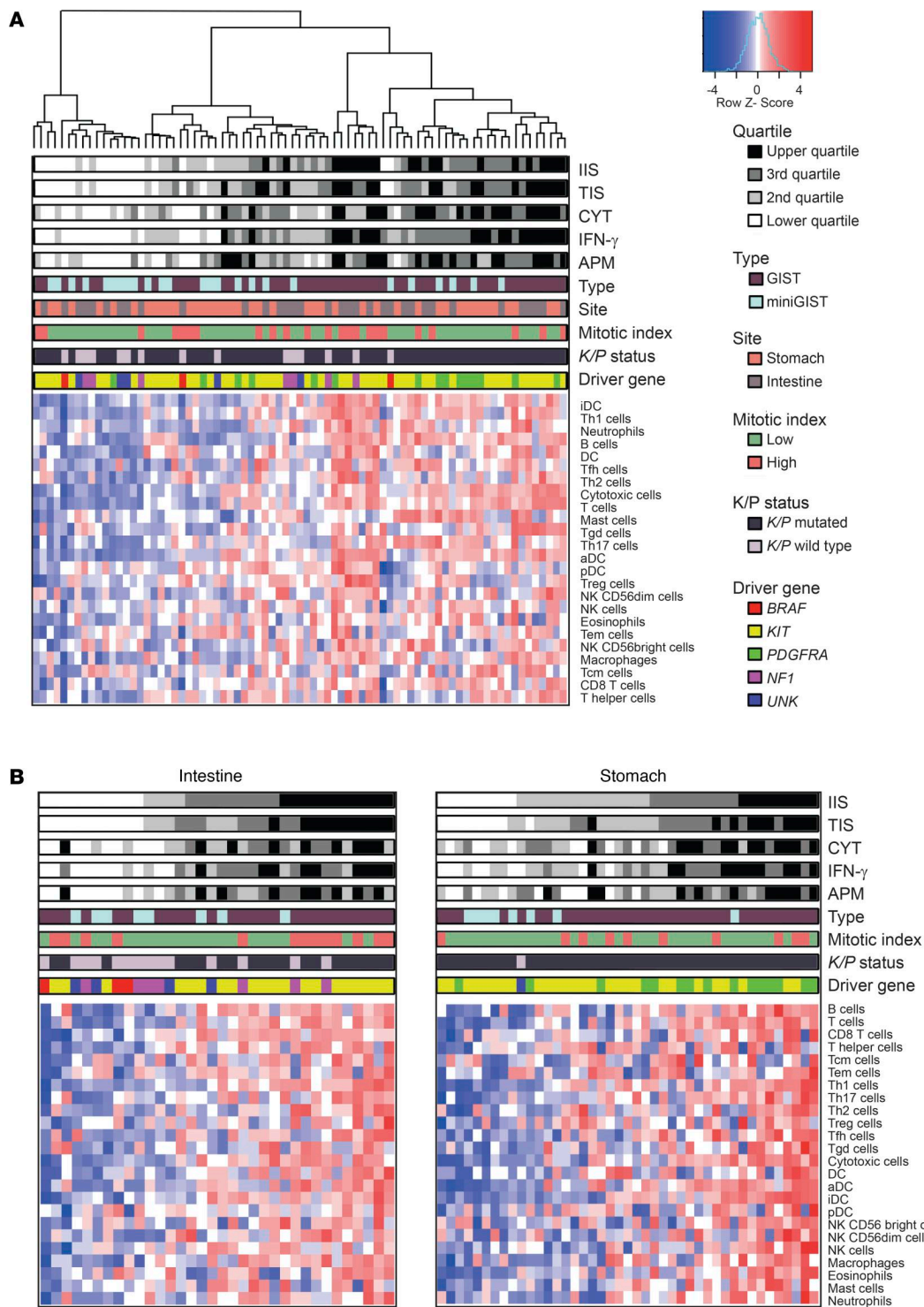


Figure 5.1. (A) Unsupervised clustering analysis of the whole GIST series ($n = 77$) based on ssGSEA scores of 24 immune cell types. Hierarchical clustering identifies 3 major groups with different extent of immune infiltration. IIS, TIS, CYT, IFN- γ , and APM scores are reported as quartiles. (B) ssGSEA in intestinal ($n = 34$) and gastric ($n = 43$) sites analysed separately highlights the impact of driver gene and malignant potential in immune infiltration. Samples are ordered according to increasing IIS. UNK, driver mutation unknown; ssGSEA, single-sample gene set enrichment analysis; IIS, immune infiltration score; TIS, T cell infiltration score; CYT, cytolytic activity score; APM, antigen-presenting machinery (Ref 3, LAA)

These results were corroborated by in situ analysis. A subset of 38 samples, representative of the different categories, were immunostained with a set of immune-specific cell markers (CD3, CD4, CD8, CD20, CD68 and FOXP3). Also, IHC indicated that wild type GIST featured an immune cold phenotype and that in immune infiltrated tumours T-lymphocytes and macrophages were the most represented immune cell populations. The overrepresentation of these two cell types was indicated also by the interrogation of the transcriptome with a deconvolution approach (CIBERSORT, see par 4.3 sect. *b*)

Overall, the consistency between IHC and in-silico analysis validates the results obtained from transcriptomics data and corroborates the hypothesis that tumour genotype, location, and malignant potential actively influence the tumour immune microenvironment.

Knowing that a substantial level of immune infiltrate exists in GISTs [4], the positive correlation of IIS with the Antigen Presenting Machinery (APM) enrichments scores and the cytolytic score (respectively, $r = 0.63$; $r = 0.62$; $p\text{-value} < 1 \times 10^{-6}$) implies that an antigen-specific immunity may exist in a relevant fraction of tumours.

In the light of the observed effect of genotype on tumor immunophenotype, we sought to address the theoretical neoantigenic capacity of epitopes generated by the mutated driver gene by using the NetMHCpan neoantigen prediction algorithm [5]. Although this analysis indicated that almost all mutations yielded at least one peptide capable of binding, with different strengths, a cognate HLA allele, no clear explanation of the immune cold phenotype of wild type GIST was obtained.

To gain further insights into the mechanisms implicated in the poor GIST immunogenicity of this subset of tumors, we interrogated transcriptome data with signatures known to be involved in immune exclusion/suppression phenomena. The Hedgehog (HH) pathway [6] and WNT/ β -catenin signaling (WNT/ β -cat) [7], known drivers of immune suppression, emerged as enriched in poorly infiltrated wild type GIST (see Figure 5.2, A and B).

Intriguingly, both pathways appear to be positively regulated by the RAS/RAF/MAPK pathway, which is activated in GIST, particularly in wild type GIST [8-10] and activation of the RAS pathway has also been associated with immune suppression [11-12]. Thus, RAS, HH, and WNT/ β -cat might cooperate to dampen the immunogenicity of K/P WT intestinal GISTs.

Thus, although our study indicates that wild type GIST currently orphan or targeted therapies are per se unlikely to respond to immune-based therapies due to their low degree of immune infiltration, the identification of Hedgehog and WNT/ β -catenin pathways as possible responsible for immune suppression discloses a potential therapeutic vulnerability, as the targeting of

these pathways might prove effective by both inhibiting pro-oncogenic signals and fostering antitumor immune responses.

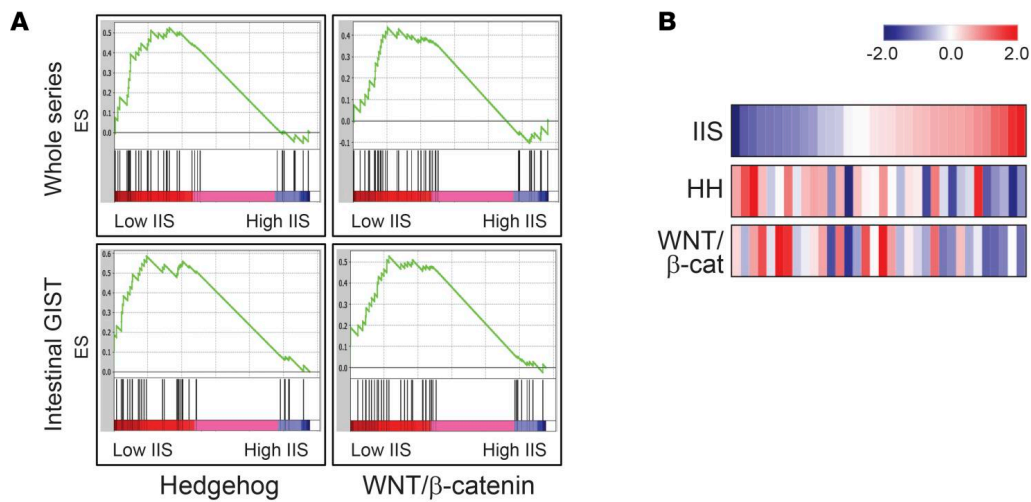


Figure 5.2. (A) GSEA analyses indicating the enrichment of HH and WNT/β-catenin signatures in immune cold GIST (low IIS), compared with immune hot GIST (high IIS) in the whole series (top) and in the intestinal subset (bottom). (B) Anticorrelation of IIS with HH and WNT/β-catenin activation scores in intestinal GIST. Color-coded score values are displayed. Site, type, mitotic index, and driver gene are as per indicated color-coded labels. UNK, driver mutation unknown; GSEA, gene set enrichment analysis; HH, Hedgehog; IIS, immune infiltration score (Ref 3, LAA)

5.1.2) Assessment of the reliability of NGS RNA-based approaches for the detection of fusion transcripts

As second project that I have contributed to as a bioinformatician, I dealt with the detection of fusion transcripts in sarcomas. About 1/3 of sarcomas carry chromosome rearrangements that result in the generation of fusion genes. The detection of these fusion genes and corresponding fusion transcripts is often pathognomonic, i.e., they represent a distinctive feature of a specific tumour subtype and therefore are helpful in differential diagnosis. Traditionally, in routine molecular diagnostics, fusion events are detected by FISH at the genomic level or by RT-PCR at the transcriptional level. These approaches are useful but are considered reflex testing, in other words they are applied when a diagnostic hypothesis has been made and needs

to be verified. The advent of omics approaches such as NGS has disclosed the possibility of interrogation of a tumour transcriptome in an agnostic manner. Nevertheless, whole transcriptome sequencing and fusion transcript detection is far too complex to be applied in a diagnostic setting. Several targeted RNA-seq assays have been developed to specifically tackle the issue of diagnostic fusion transcript. As the coordinator of the Italian Sarcoma working group of ACC (Alleanza Contro il Cancro; <https://www.alleanzacontroilcancro.it/en/>), the laboratory of Oncogenetics and Oncogenomics at the CRO in Aviano sought to compare 3 different commercially available kits and dedicated bioinformatics tools for NGS-based fusion transcript detection and assess their sensitivity, specificity and applicability to the routine diagnostics. 26 samples were tested with a hybrid capture-based panel (HC) (Illumina TS-Fusion). Nineteen of these samples plus an additional one constituted a dataset of 20 samples tested with an amplicon-based anchored multiplex PCR panel (Archer AMP-FPS). In addition, 9 samples were profiled with a more comprehensive HC panel (Illumina TS-PanCancer). TS-Fusion and TS-PanCancer panels data were analysed with BaseSpace (<https://basespace.illumina.com>) while the AMP-FPS panel with the Archer Analysis Platform (<https://analysis.archerdx.com>) (see Figure 5.3).

The result of this effort, which included the molecular profiling of over 100 sarcomas and is reported in the paper (see Ref. 1, LAA), allowed the identification of a combo kit/bioinformatics tool with feature of sensitivity, specificity and easiness of use compatible with the use in the diagnostic setting. I specifically contributed to the re-evaluation of NGS data in those cases where, although molecular diagnostics with either RT-PCR or FISH indicated the presence of a fusion event, the tested combo kit/dedicated algorithm failed in fusion detection. To this end I analysed target RNA-seq raw data with 3 different algorithms used in the identification of fusion events in whole RNA-seq, namely Arriba [13], STAR-fusion [14] and Pizzly [15], tools reported to have high fusion detection rates. Except for one single case, for which no algorithm scored positive in the detection of fusion event involving a gene apparently rearranged according to FISH, at least one fusion caller could detect a fusion transcript involving the gene suggested by molecular diagnostics (FISH or RT-PCR), thus emphasizing the importance of software sensitivity in data analysis.

AMP-FPS panel stood out as the best one, given its detection capability, easy-of-use and required time for library preparation, even if it is less comprehensive than the others. This result was also achieved by more recent assessments of commercially available assays [16]. The AMP-FPS panel was further tested on 123 additional (see Table 2, Ref 1, LAA). Out of 81

cases whose genetic abnormality was pre-detected, AMP-FPS assay correctly detected the fusion events in 71 of them.

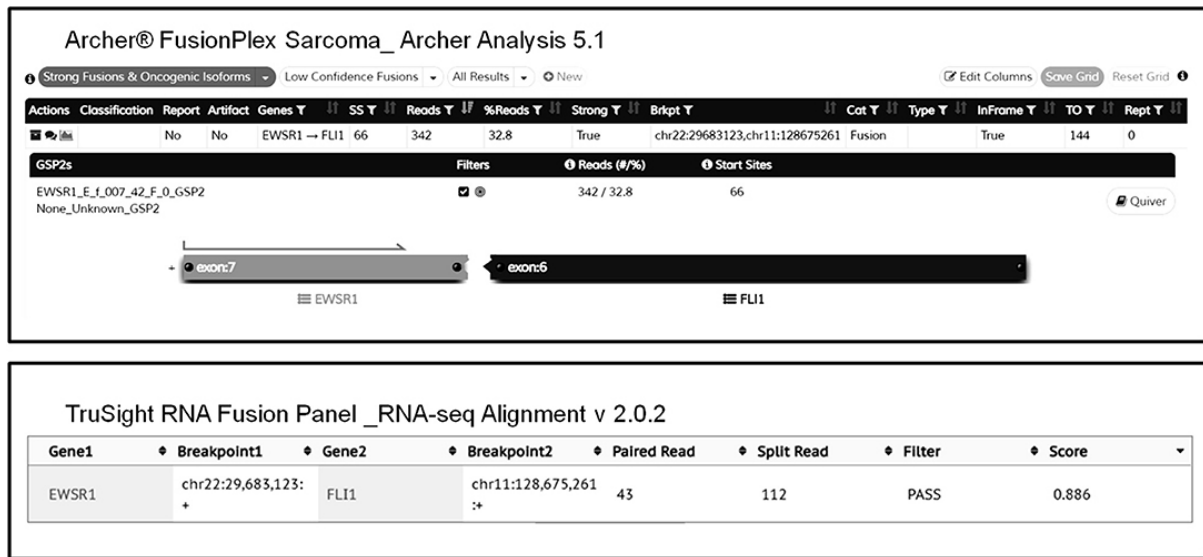


Figure 5.3. Representative graphical output of Archer Analysis (top) and Illumina BaseSpace RNA-Seq Alignment (bottom) tools. The detection of EWSR1-FLI1 fusion in an Ewing Sarcoma sample by both AMP-FPS and HC panels is shown (Ref 1, LAA)

5.2) Enzyme Proteomics

5.2.1) Characterizing disease-related enzymes

Recent advancements in cell physiology support the notion that enzymes, performing their specific molecular functions in a concerted manner, catalyse all the different biological processes occurring in a cell [17]. Consequently, the most popular model of active enzymes in the cell context is a graph, where hubs are enzyme molecules and edges are all the possible interactions, documented in specific data bases such as STRING (<https://string-db.org/>), IntAct (<https://www.ebi.ac.uk/intact/home>) and BioGRID (<https://thebiogrid.org/>). Due to experimental techniques interactomic data refer to static interpretation in the interacting network. The scientific community is debating whether enzymes transiently aggregate in the cell to generate the proper concerted actions [18]. As a member of the Bologna Biocomputing Group (<http://www.biocomp.unibo.it>), I collaborated in highlighting enzymes that are active in different metabolic pathways and that, at the same time, stand out as hubs in protein-protein interaction networks. We further linked these enzymes to diseases by characterizing their variants

with mutated residues, often associated to possible molecular mechanisms at the basis of their association with diseases (see Ref 5 and 6, LAA).

In this context, the approach is first to download from database, available data, to analyse sequences/structures with available methods (basically developed and available as web servers at the Bologna Biocomputing group, <http://www.biocomp.unibo.it/predictors.html>) and comment on what computation can infer for a possible behaviour of the molecule. I took care of data base selection, annotation, and tool applications for addressing different types of problems as described in the following. More recently I am implementing a data base of disease related enzymes which will be released soon.

A) Human Enzymes Active in Different Metabolic Pathways and Diseases

In this research we try to characterize the dimensionality of the networks of the human enzymes which are also disease related and define important features for their characterization.

Information was derived from Swiss-Prot (release 04_2019) that contains 20,365 human proteins out of which 3,428 are annotated with a complete EC number. We collect 770 human enzymes that are disease related and linked to at least one pathway in KEGG [19] for a total 90 different KEGG pathways and 930 unique biochemical reactions represented by EC numbers. By associating catalysed reactions to pathways through human enzymes, we discovered that five EC number are linked to 11 KEGG pathways: 1) EC 1.2.1.3, Aldehyde dehydrogenase (NAD⁺); 2) EC 1.14.14.1, Unspecific monooxygenase; 3) EC 2.3.1.9, Acetyl-CoA C-acetyltransferase; 4) EC 2.6.1.1, Aspartate transaminase; 5) EC 4.2.1.17, Enoyl-CoA hydratase.

In this scenario, 12 enzymes emerged as associated to 10 or more pathways in KEGG, representing four of the reactions more present in different pathways (EC 1.2.1.3; EC 2.3.1.9; EC 2.6.1.1 and EC 4.2.1.17). We explored their interactions and compared them with those of enzymes present only in one pathway. Consistently, we observed that enzymes participating in multiple pathways are endowed with a significant higher number of interactors (Mann-Whitney *U* test, p-value < 0.001) than those associated to only one pathway.

As second and more important step was to test the capability of our predictor of protein-protein interactions (ISPRED4 [20]). An interesting finding is that our predictor of protein-protein interactions can eventually predict several interaction sites on the proteins participating in interaction networks which well correlates with the number of associated pathways. Enzymatic proteins linked to at least 10 KEGG pathways turned out to have a higher number of interaction sites if compared to the enzymes acting on only one pathway (Mann-Whitney *U* test, p-value < 0.05, Figure 5.4).

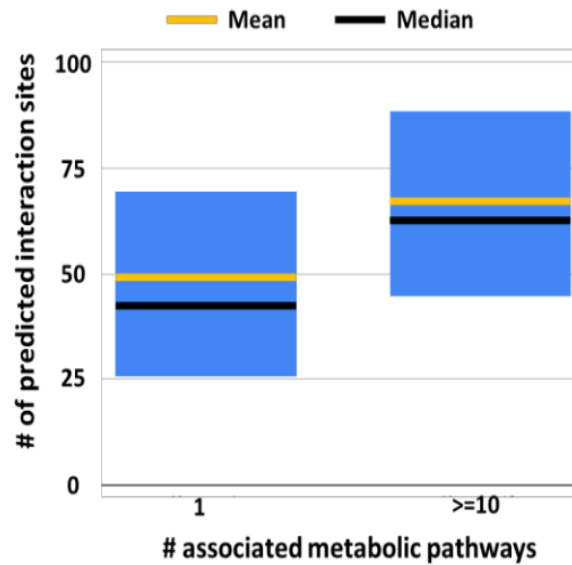


Figure 5.4. Statistical characterization of the number of interaction sites predicted with ISPRED4 in EC proteins associated with only one or at least 10 metabolic pathways. For each set, the boxes represent the first and third quartiles; yellow and black lines represent mean and median values, respectively. Significance of the reported difference on median values has been validated using the Mann–Whitney U test obtaining p-value = 0.04.

Number of. (Ref 2, LAA)

Test cases

We then focused on two highly connected enzymes to explore at the structural level protein-protein interaction sites. Furthermore, we inspected proteins that carry disease related variations to determine first the location of these residues, and then to which extent a disease related variants being in an interacting site, may hamper protein-protein interaction. A second related property to the effect of variations is the putative effect on protein stability that we can compute with a suited predictor developed in house (INPS-MD, <https://inpsmd.biocomp.unibo.it/welcome/default/index>). We selected as a test case the human alpha-amino adipic semialdehyde (AASA), also known as antiquitin (Gene: ALDH7A1, UniProt: P49419, PDB: 4ZUL). AASA is a multifunctional enzyme mediating important protective effects. The protein protects cells from oxidative stress by metabolizing lipid peroxidation-derived aldehydes (EC 1.2.1.3), and it is involved in lysine catabolism (EC 1.2.1.31). It also metabolizes betaine aldehyde to betaine (EC 1.2.1.8), an important cellular osmolyte and methyl donor. The protein is associated to 232

variations out of which 27% of them affect protein stability being in the interfaces of the biological assembly or in the protein area exposed to the solvent (see Figure 5.5). We predicted 21 residues likely to act as interaction sites in the solvent exposed surface of the protein and 11 are disease-related variations sites (4 of them associated with the AASA related disease pyridoxine-dependent epilepsy (PDE)) [21]. This suggests that the variations occurring on the protein surfaces may affect protein activity in multiple pathways where the enzyme is active by influencing interactions with other proteins but not altering the protein stability.

Similarly, we explored the putative interaction sites on the protein surface and the annotated variations of the human Acetyl-CoA C-Acetyltransferase (Gene: ACAT2/ACAT1, UniProt: Q9BWD1/P24752, PDB: 1WL4/2IBY, localization: cytosolic/mitochondrial). Both the cytosolic and the mitochondrial enzymes are linked to the same disease (alpha-methylacetoacetic aciduria) that arise because of their deficiency. We found that, in the mitochondrial protein, 8 variations match with putative interaction sites on protein surface. Moreover, 5 variations occur in a 33 residue-long mitochondrial target peptides suggesting that the disease may also be caused by an inefficient translocation of the protein to the mitochondria (see Figure 5.6). In the cytosolic protein, the only variations linked to alpha-methylacetoacetic aciduria (E176K) was observed on protein surface and we predicted it as a putative interaction site strengthening our conclusions about the relation of disease onset and restriction in the protein-protein interactions event due to specific variations (see Figure 5.7).

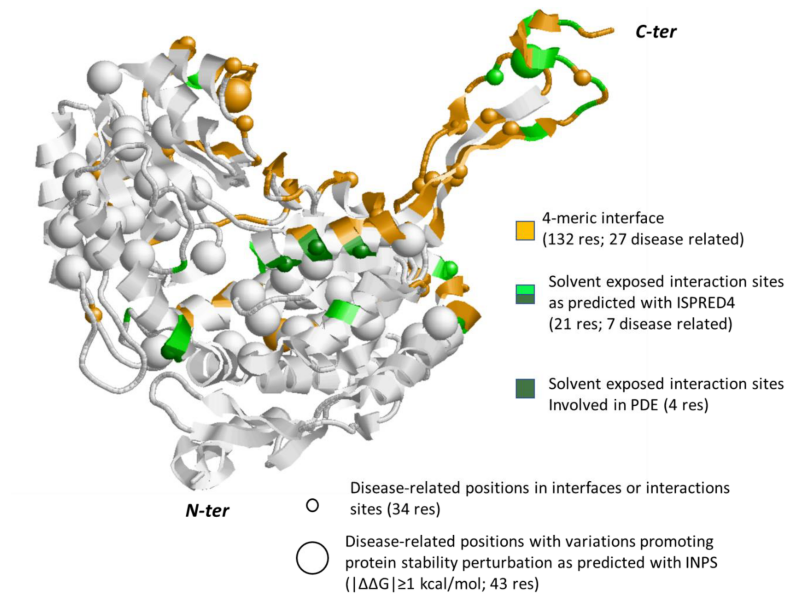


Figure 5.5. Monomeric subunit of human ALDH7A1 protein (PDB code: 4ZUL.A). Interaction surface in the tetramer as derived from the crystallographic coordinates is in orange. Interaction sites out the tetrameric interface, as predicted with ISPRED4, are in green. Positions in these regions carrying disease related variations are highlighted with small spheres. Big spheres highlight positions in the protein carrying disease related variations and promoting a large variance of folding free energy, as predicted with INPS. Grey colour: the background protein backbone (Ref 2, LAA)

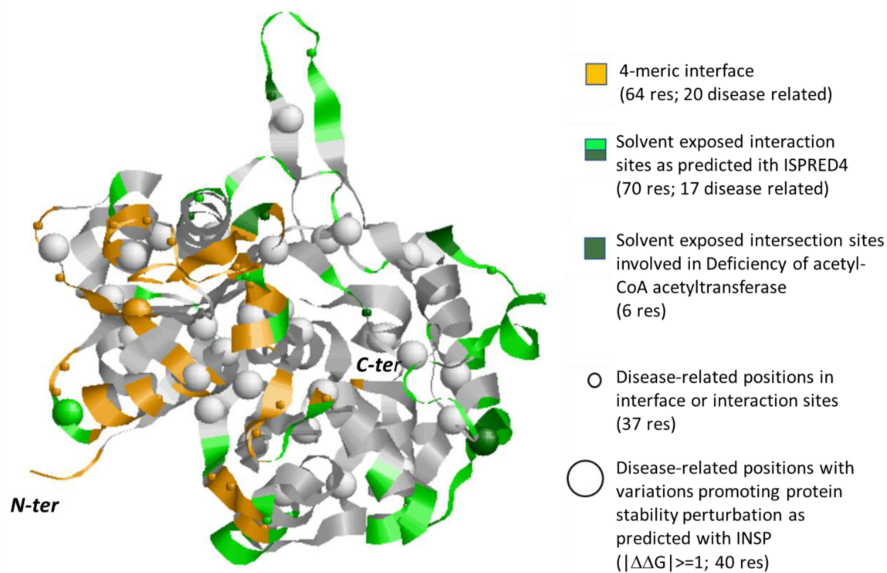


Figure 5.6. Monomeric subunit of human ACAT1 protein (PDB code: 2IBY.A). Interaction surface in the tetramer as derived from the crystallographic coordinates is in orange. Interaction sites out the tetrameric interface, as predicted with ISPRED4 are in green. Positions in these regions carrying disease related variations are highlighted with small spheres. Big spheres highlight positions in the protein carrying disease related variations and promoting a large variance of folding free energy, as predicted with INPS. Grey colour: the background protein backbone (Ref 2, LAA).

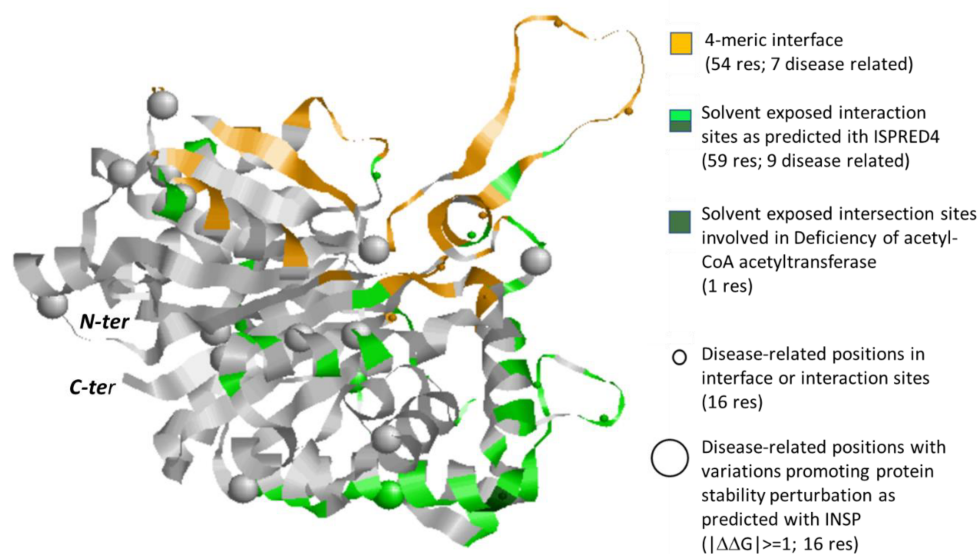


Figure 5.7. Monomeric subunit of human ACAT2 protein (PDB code: 1WL4.A). Interaction surface in the tetramer as derived from the crystallographic coordinates is in orange. Interaction sites out the tetrameric interface, as predicted with ISPRED4 are in green. Positions in these regions carrying disease related variations are highlighted with small spheres. Big spheres highlight positions in the protein carrying disease related variations and promoting a large variance of folding free energy, as predicted with INPS. Grey colour: the background protein backbone (Ref 2, LAA).

B) The relations between disease-related variation types, maladies, and structural conserved domains.

An important yet poorly explored topic is the search for specific features of disease related variations, which would allow also possible functional annotations. One major problem is the association disease-variation which is gene dependent.

To explore this context, we downloaded from Humsavar (<https://www.uniprot.org/docs/humsavar>) and ClinVar (<https://www.ncbi.nlm.nih.gov/clinvar/>), 75,145 variations, including 43,917 pathogenic ones, and carried by 3,605 unique genes. Pathogenic variations can be linked to 5,223 diseases represented by Mondo ID codes (<https://mondo.monarchinitiative.org/>).

We investigated the association gene-disease/s and we found out that Fibrillin, (Gene: FBN1, UniProt: P35555), GTPase KRas (Gene: KRAS, UniProt: P01116), the Cellular tumour antigen p53 (Gene: TP53, UniProt: P04637), Collagen alpha-1(II) chain (Gene: COL2A1, UniProt: P02458) and Prelamin-A/C (Gene: LMNA, UniProt: P02545) are associated to the highest number of diseases (at least 21). Then, we grouped variations into physicochemical classes,

corresponding to what we call the variation type. We converted the disease related variations into variation types (apolar (G, A, V, I, L, P, M); polar (S, T, C, N, Q, H); aromatic (F, W, Y); charged (D, E, K, R), giving rise to 16 possible variation types.

We then compared the distribution of pathogenic variation types with that of benign ones. In line with previous studies from our group [22], we observed differences in the distribution of variation types, discovering that nonpolar into nonpolar, polar, and charged, and charged into polar ones, are the most abundant type among pathogenic variations (see Figure 5.8).

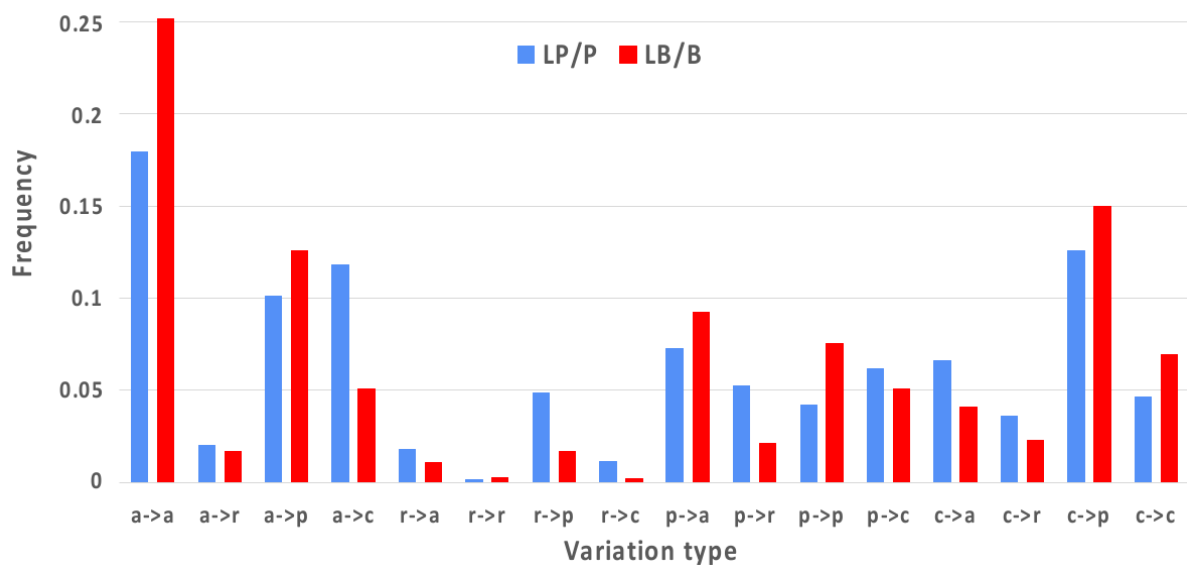


Figure 5.8. Frequency of variation types of the Union variations. Blue bars: Likely Pathogenic/Pathogenic (LP/P) variations; Red bars: Likely Benign/Benign (LB/B) variations. Labels are as follows: a, nonpolar; r, aromatic; p, polar; and c, charged. (Ref 6, LAA)

Disease related variation types are specifically and significantly linked to different Mondo categories. Making a step forward, we investigated how disease variation types map in protein structural domains as represented by Pfam and InterPro models. In Figure 5.9 we focus on the 20 most represented Pfam domains in our dataset, covering 557 genes and 6,729 pathogenic variations. It appears that Pfam domains have distinctive variational patterns (for statistical validation Ref.5, LAA).

Taking advantage of Pfam and IntePro mapping, we could establish a relation among domains containing disease related variation types and diseases, which for sake of simplicity, are grouped into anatomical system Mondo categories. Summing up, our results indicate that by mapping variation types into PFAM and InterPro gene domains, a link can be established among variations and diseases (see Figure 5.10).

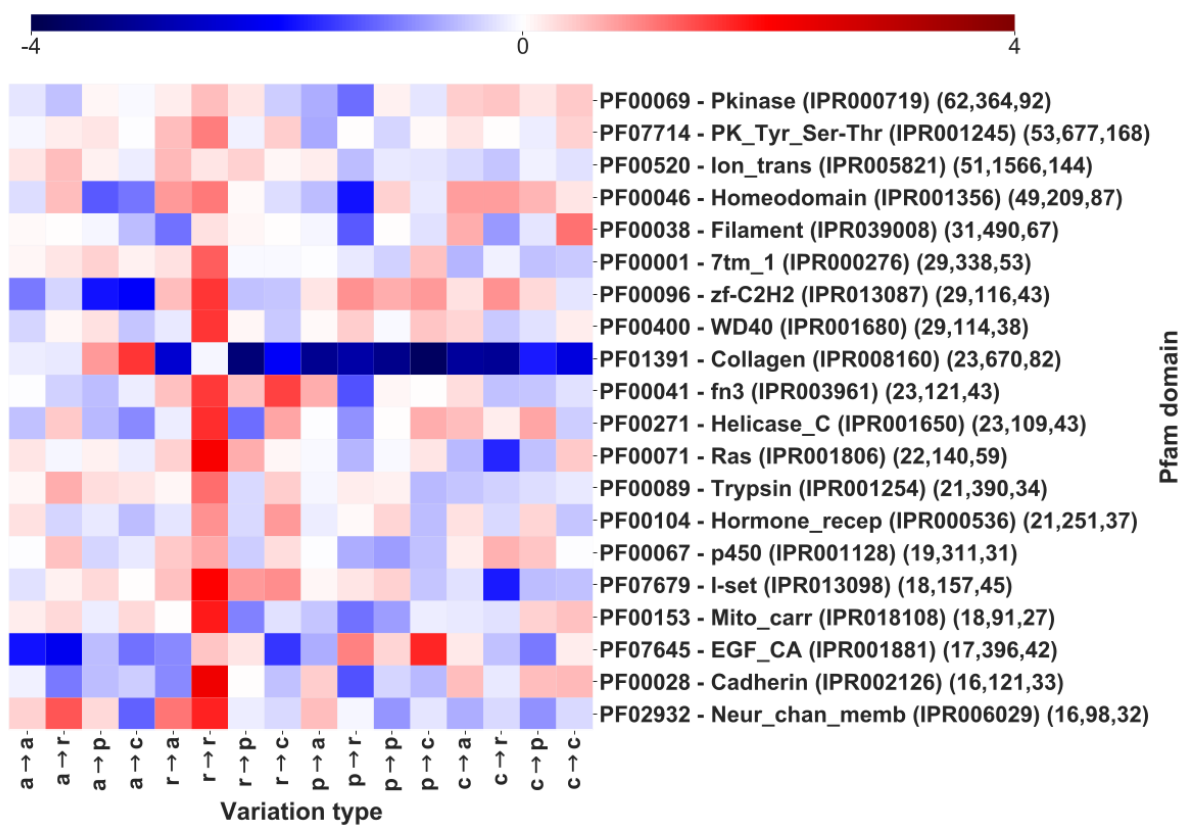


Figure 5.9. Log-odd scores of variation types for the first 20 InterPro entries (out of 5,357, Table 2), sorted by number of genes covered and not including Pfam signatures. Log-odds are computed with respect to the whole dataset background of pathogenic variations. Numbers in parentheses report, for each InterPro, the number of genes, of single residue variations and of diseases, respectively (Ref 6, LAA)

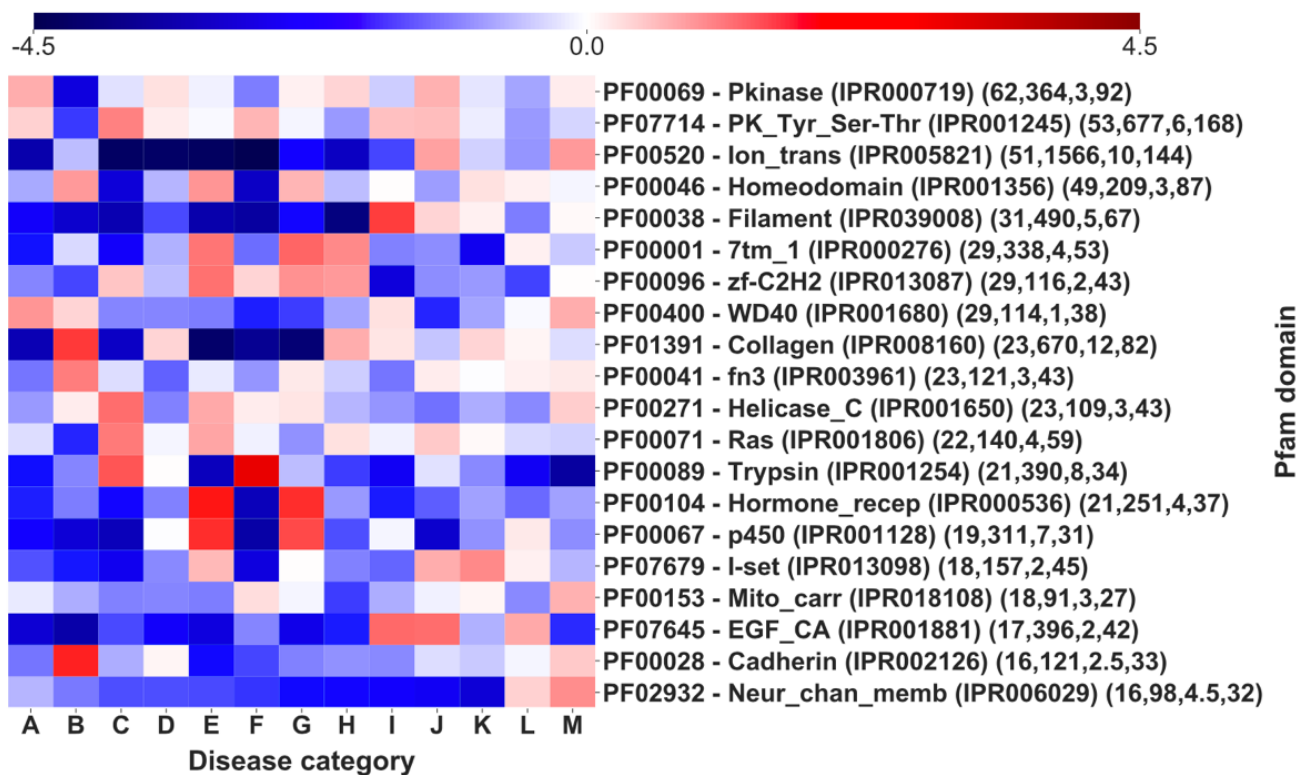


Figure 5.10. Log-odd scores for disease categories associated to different Pfam domains. Log-odds are calculated with respect to the whole-dataset background of disease categories. For each Pfam the corresponding InterPro accession is indicated. Numbers in parentheses report the number of genes, of SRVs, the median number of SRVs per gene and the number of diseases. Mondo “Disease by Anatomical System” categories as follows: A-respiratory system disease, B-auditory system disease, C-immune system disease, D-digestive system disease, E-disease of the genitourinary system, F-hematologic disease, G- endocrine system disease, H-urinary system disease, I-integumentary system disease, J- cardiovascular disease, K-musculoskeletal system disease, L-disease of the visual system, M- nervous system disorder, N-mediastinal disease. (Ref 6, LAA)

C) Human MTHFR deficiency.

As an additional effort, we investigated the relationship between structural and functional feature of the human enzyme methylenetetrahydrofolate reductase (Gene: MTHFR, UniProt: P42898) deficiency (see Ref. 6, LAA). We correctly predicted the interface region of the homodimeric structure of MTHFR (PDB: 6FCX) with ISPRED4 (<https://ispred4.biocomp.unibo.it/welcome/default/index>). Three residues belonging to the interface region are annotated as sites linked to MTHFR deficiency, suggesting that these variations may hamper protein-protein interaction and promote the disease.

The protein is endowed with two structural domains, one catalytic and one regulatory, respectively. The two domains contain most of the disease related variations, in agreement with our previous observations (Figure 5.11).

We calculated the Gibbs free energy change ($\Delta\Delta G$; as described in par 4.4. section e) associated with all the disease related variations of the enzyme with INPS (<https://inpsmd.biocomp.unibo.it/welcome/default/index>), and observed that 22 out of 42 disease related variations in the catalytic domain and 20 out of 30 variations in the regulatory domain of the protein appear to affect protein stability, suggesting that one of the major causes of disease can be protein instability promoted by the substitutions of the lateral side chains.

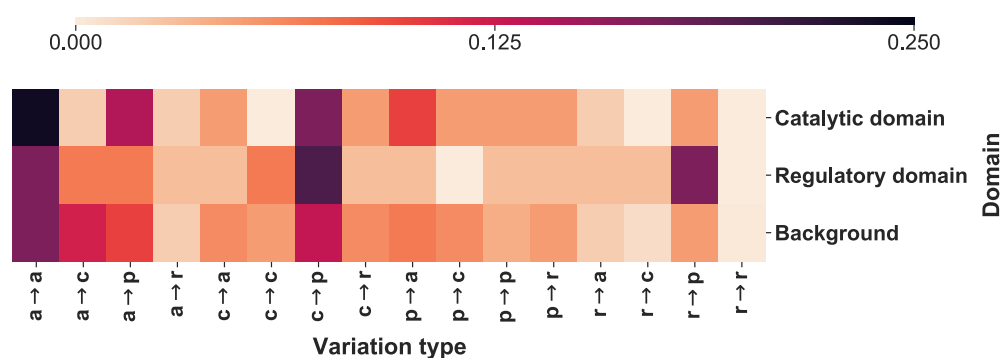


Figure 5.11. The heatmap reporting the frequency of each variation type as observed within the catalytic and the regulatory domains. The background distribution has been computed considering 22,763 pathogenic variations from Humsavar in 2,513 proteins. In variation types, labels are as follows: a, apolar; c, charged; p, polar; and r, aromatic (Ref 5, LAA)

D) A curated Database of Disease-related Enzymes

We are developing a database of human genetic disease-associated enzymes that will help in dissecting the complexity of the relationships among genes and diseases. The database includes 1,230 human enzymes. Their pathogenic variations are derived from Humsavar and ClinVar and link 1,926 different diseases represented by Mondo IDs. Enzymes participate into one or more metabolic pathways, as modelled in Reactome (<https://reactome.org/>). The project aims to establish to what extent the different Reactome pathways are affected by the different enzymes. Particularly interesting will be to explore the level of integration of the different pathways when enzymes are active in more than one biological process, in relation to the associated diseases. The aims to generate a database that will contain the results of the analysis.

5.2.2) BENZ WS: The Bologna ENZyme Web Server

I devoted my efforts in developing the Bologna ENZyme (BENZ) Web Server (<https://benzdb.biocomp.unibo.it/>; see Ref 4, LAA). BENZ closes the gap among the huge number of newly deposited protein sequences from massive application of NGS technologies, and the limited number of proteins whose biochemical function has been fully characterized. BENZ performs functional annotation taking advantage of a curated database of functional and structural protein information and of a sets of machine learning based predictive models (see Figure 5.12). The method is also described under Material and Methods (4.6), where the workflow is detailed.

The latest BENZ release contains 16,593 reference sequences, out of which 2,023 are polyfunctional enzymes (<https://benzdb.biocomp.unibo.it/statistics.html>). 6,798 reference sequences have got a structure in the PDB and 618 of them are also polyfunctional. 891 different organisms are represented in the database (Archaea: 24; Bacteria: 261; Eukaryota: 391; Viruses: 213; Unknown: 2). BENZ can predict 5,136 different EC numbers distributed in the seven functional classes thanks to 12,612 HMM models out of which 10,547 are labelled as “Gold” and 2,065 as “Blue”. Reference sequences in BENZ represents 4,158 unique conserved structural domains from Pfam (<http://pfam.xfam.org/>) [23].

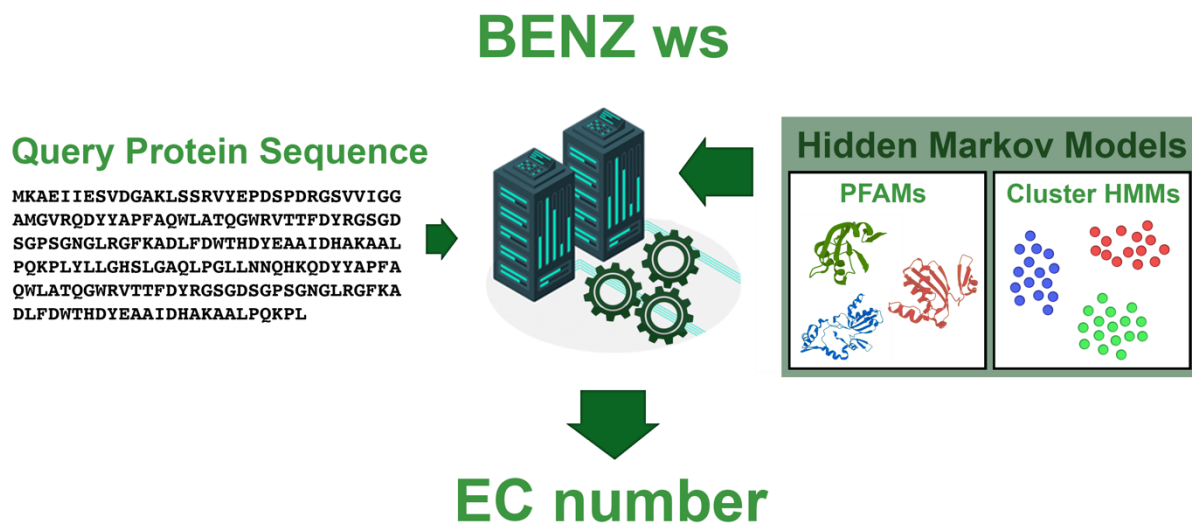


Figure 5.12. Graphical representation of BENZ WS infrastructure. When a target sequence enters the server, it is filtered by two different sets of predictive models (PFAMs and Cluster HMMs). When the target is retained with a given significance threshold and a set of conditions are met, it is endowed with four-level EC number/s (Ref 4, LAA).

A) BENZ Core Architecture

BENZ comprises two sets of HMM predictive models. One set is derived from Pfam (v.33.1) while the second set was generated in-house as described in par 4.6 (Materials and Methods). Proteins from UniProt are first clustered connecting sequence pairs that share a sequence identity greater or equal to 40% on an alignment coverage of at least 90%. Protein clusters were defined isolating the connected components and only clusters containing proteins annotated with at least one EC number in UniProt were retained.

Each cluster is represented by computing an HMM model. Enzymes in clusters were annotated with their EC number and their architecture, according to Pfam models. For each cluster reference proteins are identified provided that: 1) 3D structure is available in PDB (mandatory for TrEMBL enzymes); 2) the level of annotation score in UniProt; 3) Complete EC number and functional annotation; 4) Available Pfam architecture. Clusters were then grouped into “Gold Clusters” when univocally associated to a reference sequence, and “Blue Clusters” when associated to two or more reference sequences. Gold and Blue clusters were finally collected in a database-like structure of predictive models that constitute the core of BENZ together with the set of Pfam predictive HMM models (Figure 5.13).

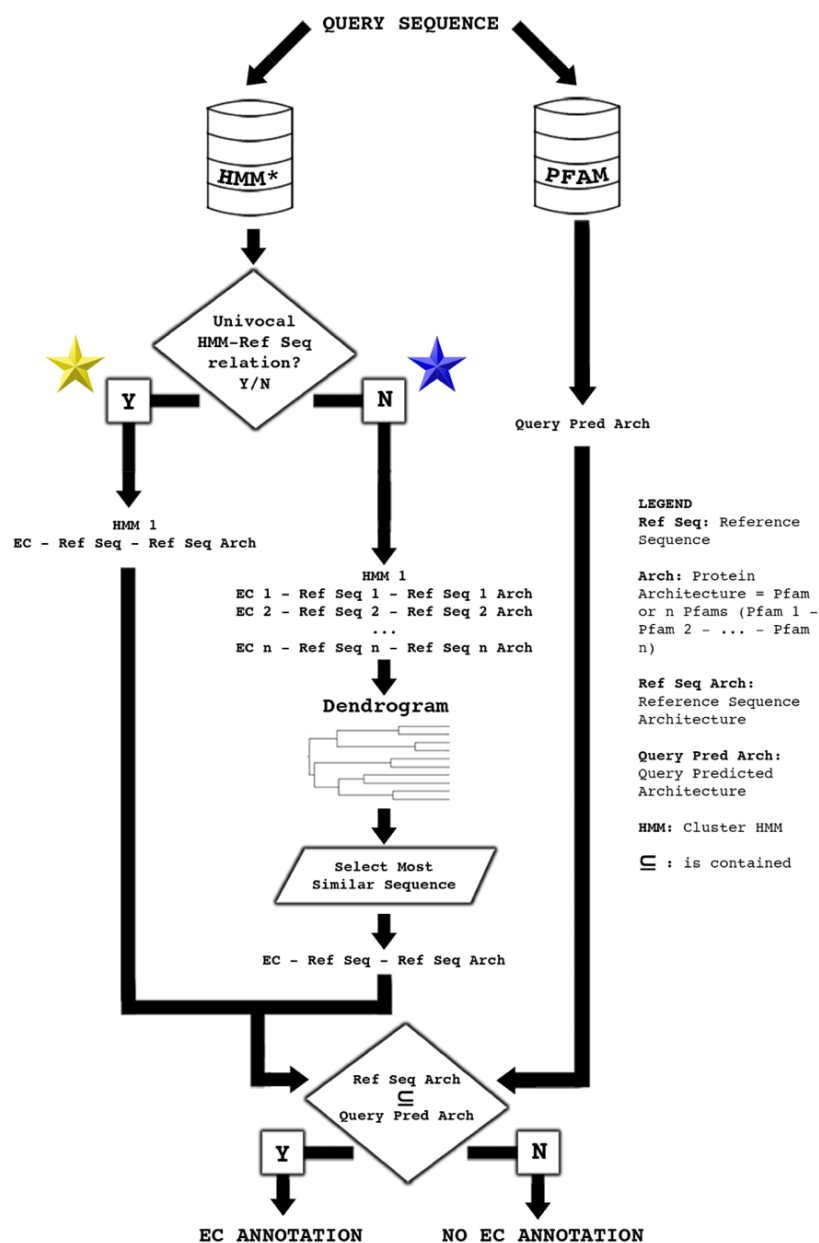


Figure 5.13. Workflow of BENZ WS. For a query sequence, in FASTA format, the annotation procedure starts with HMM filtering. If the retaining HMM is plurivocally associated to different references sequences (blue star), a dendrogram is generated to find among the reference sequences the most similar one to the target. Otherwise (yellow star), the target is associated to the only reference. The EC number-query sequence association is then made after evaluating if the reference protein architecture (Ref Seq Arch) is contained (\subseteq) in that of the predicted target Pfam architecture (Query Pred Arch), focusing on Pfams carrying relevant sites. Pfams in our system are annotated, when possible, with the positions of the active site, ligand binding site and metal binding site (relevant sites). A sequence feature viewer allows the user to verify whether the query sequence conserves the residues relevant to the protein catalysis for validating the transfer of annotation from the reference sequence. Links to the reference sequence UniProt/SwissProt file, structure PDB file and Pfam entries, together with KEGG identifiers and pathways are also present in the output (see HELP, <https://benzdb.biocomp.unibo.it/help>).

B) BENZ at work

When a sequence is pasted into the Web page, the user activates the system (Figure 5.14).

In BENZ, the collection of predictive models is organized alongside the Pfam models from Pfam v.33.1 (<http://pfam.xfam.org>). When a user submits a query sequence to BENZ, it is filtered by two different sets of models (Figure 5.13)

The target sequence finds a matching reference template when retained with a given threshold ($E\text{-value} \leq 10^{-5}$) by one of the in-house generated HMM. At the same time, within the Pfam models, when retained with a given threshold ($E\text{-value} \leq 10^{-4}$), the submitted sequence gains its architecture. Inclusion thresholds were chosen after a self-consistency test.


If the target is retained by a cluster HMM that is plurivocally associated to more than one reference sequence (Blue Cluster in Figure 5.13), BENZ runs Clustal Omega to find the most similar reference sequence, as depicted by the dendrogram generated by the multiple sequence alignment procedure (see Figure 5.13).

The predicted Pfam architecture of the target sequence is compared to that of the reference. If the target Pfam architecture matches or includes that of the template the target protein is endowed with the four-level EC number of the reference sequence. If not, the four-level EC number can be attributed based on the sharing of a common Pfam which contains relevant sites. (Ref 4, LAA).

A typical BENZ output is shown in Figure 5.15. The target sequence is endowed with the four-digit EC number when all the constraints are met. Furthermore, the user can retrieve the reference enzyme, its structure when available, the retaining HMM and the predicted target architecture.

BENZ WS

[HOME](#) [STATISTICS](#) [HELP](#) [CONTACTS](#)



**Bologna
Biocomputing
Group**

The Bologna **ENZ**yme Web Server (**BENZ WS**) annotates the Enzyme Commission numbers (EC numbers) of enzymes, defined by the International Union of Biochemistry and Molecular Biology (**IUBMB**). **BENZ WS** is based on [HMMs](#) and [PFAMs](#) and returns a four-level EC number for a target which is retained by the system.

EC number Annotation

Submit one sequence in FASTA format at the time
(max length = 5,000 residues)

Gold Example
Blue Example
Submit
Clear

```

>golden_sample.sequence
MVRLLKKSFGQHLVSEGLVKIAEELNIEEONTVVEVGGGTGNLTKVLLQHLKLYVIE
LDREMVENLKSIGDERLEVINEADSKFPFCSLGLKELVGNLPLYNVASLIENIVYNKDC
VPLAVFVFNQKEVAKELQKKKDTGWLSPVVRTFYDYNVYMTYPPRFVPPPKVQSAVKLV
KNEKFPVKDLKKNYKFLTKIFQNRKVLKKIPEELLKEAGINPDARVEQLSLEDFKLY
RLIEDSGE
          
```

Figure 5.14. **BENZ WS** homepage. Freely accessible at: <https://benzdb.biocomp.unibo.it>

EC number

Best Match

BENZ ws - Output

EC number	Predicted Pfam Target Architecture	Reference Sequence	Reference Structure	Reference Architecture	Source
2.1.1.182	PF00398 (MBS)	P37468	6IFT	PF00398 (MBS)	★

Data ↓

Predicted Target EC number

[Download Data](#)

Show entries

Search:

Retaining HMM	E-value	EC number	Reference Sequence (Length)	Reference Structure	Reference Architecture	Reference Organism	Reference KEGG ID	Reference KEGG Pathway
1	1.1e-76	2.1.1.182	P37468 (292)	6IFT	PF00398 (MBS)	Bacillus subtilis subsp. subtilis str. 168	bsu:BSU00420	
2	1.4e-32	2.1.1.184	P21236 (245)	YUB	PF00398 (MBS)	Streptococcus pneumoniae		

(Relevant Sites are present when annotated in the reference protein)

[Download Data](#) [\[SHOW\]](#)

Target Graphical Visualization

Relevant Sites: ■ AS: Active Site ■ MBS: Metal Binding Site ■ LBS: Ligand Binding Site

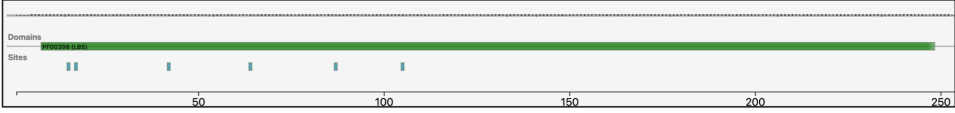


Figure 5.15. Web page screenshots from **BENZ** output page. **BENZ** returns a predicted four-level EC number and the predicted architecture of the submitted target sequence together with the reference sequence/s and their architecture/s. User may also inspect the retaining HMM together with the reference sequence. **BENZ** also allows to graphically inspect the predicted architecture and evaluate the conservation of relevant site/s annotated with the reference sequence.

C) Benchmarking BENZ

The web-server performances were assessed against four different datasets: 1) a positive dataset containing complete Swiss-Prot sequences without a PDB counterpart and annotated with only four-level EC numbers; 2) a negative dataset including complete sequences from Swiss-Prot with a PDB counterpart but without EC codes; 3) a dataset of polyfunctional enzymes comprising complete Swiss-Prot sequences annotated with at least two four-level EC numbers; 4) a dataset of complete human TrEMBL sequences annotated with four-level EC numbers. BENZ revealed outstanding performances displaying an accuracy level over 90% for all the tested dataset and a false negative rate equal to 5.6% in the worst case (see Table 5.2). BENZ performances are scored with other tools: ECPred [24], EFICAz2.5 [25] and DEEPre [26]. The benchmark dataset included 366 enzyme sequences and 1,013 non-enzymes ones as negative examples, not included in BENZ. Its performance is quite satisfactory both in terms of false positive rate, equal to 3%, and true positive rate, around 75% when four level EC number is predicted (see Table 5.3).

Table 5.2. BENZ benchmark

Dataset	Sequences (#)	Acc [^] (%)	FNR [§] (%)	FPR [°] (%)
Positive ⁺	197,880	92.4	3.9	-
Negative [£]	12,315	95.1	-	4.9
Polyfunctional [#]	10,764	93.7	5.0	-
TrEMBL-human [§]	10,024	93.4	5.6	-

⁺Positive: the positive set contains complete SwissProt sequences without any PDB counterpart and annotated with only four-level EC number.

[£]Negative: the negative set comprises complete SwissProt sequence with a PDB counterpart, without EC codes.

[#]Polyfunctional: the set includes complete SwissProt sequence that are annotated with two or more four-level EC numbers.

[§]TrEMBL-human: the set contains complete TrEMBL sequences from Homo Sapiens annotated with a four-level EC number.

[^] Acc (Accuracy) measures the number of proteins correctly assigned. For sets containing positive examples, it corresponds to the True Positive Rate as evaluated at the level of four- EC annotation. For the negative set, it corresponds to the True Negative Rate.

[§] FNR (False Negative Rate) measures the percentage of enzymes predicted as non-enzymes.

[°] FPR (False Positive Rate) measures the percentage of non-enzymes predicted as enzymes.

Table 5.3. BENZ comparison with other tools

n	Data set	TPR (%)	TPR (%)	TPR (%)	TPR (%)	FNR	FPR
		1 st level	2 nd level	3 rd level	4 th level	(%)	(%)
BENZ WS ⁺	full	87.5	87.5	87.5	85.0	12.2	3.0
BENZ WS ⁺	reduced	79.2	79.2	79.2	75.1	20.2	3.0
ECPred [§]	reduced	43.7	34.7	23.8	13.1	45.6	12.2
DEEPre [#]	reduced	38.8	35.2	27.9	20.8	51.1	2.4
EFICAz2.5.1 ^{&}	reduced	33.6	33.1	31.1	16.7	63.7	1.6

The full dataset includes 607 proteins that have gained EC annotation (7 EC classes); the reduced dataset includes a subset of 366 enzyme sequences without EC codes of the seventh. Both datasets comprise 1013 non-enzyme sequences as negative examples.

+ A BENZ WS version including only sequences and annotations available in the SwissProt release 2019_11 has been used for this test.

§ ECPred has been downloaded from <https://github.com/cansyl/ECPred> and run in-house; it does not provide multiclass predictions and the best match between the output and the list of EC numbers has been considered for multiclass enzymes. It does not include enzymes of for EC class 7.

DEEPre predictions have been run on the webserver <http://www.cbr.kaust.edu.sa/DEEPre/> in modality “I’m not sure the sequence is an enzyme”; it does not provide multiclass predictions and the best match between the output and the list of EC numbers has been considered for multiclass enzymes. It does not include enzymes of the EC class 7.

& EFICAz2.5.1 has been downloaded from <https://sites.gatech.edu/cssb/eficaz2-5/> and run in-house; it does not include enzymes of EC class 7.

To promote BENZ at the last ISMB Virtual Conference (<https://www.iscb.org/ismbeccb2021>), I tested its performance against the Human Reference Proteome (HRP; Proteome ID: UP000005640; www.uniprot.org/proteomes/UP000005640). The HRP is based on Genome assembly and annotation GCA_000001405.27 from Ensembl. The proteome presently contains 77,027 protein sequences. After discharging sequences with residue length <50, we retained 71,639 proteins out of which 20,302 belonging to Swiss-Prot and 51,337 to TrEMBL. Sequences are associated to different levels of EC number for a total of 7,446 enzymes present in the database, including 702 polyfunctional enzymes (see Table 5.4).

Table 5.4. The Benchmark Dataset.

	Sequences	w/ EC 4 th	w/ EC 3 rd	w/ EC 2 nd	w/ EC 1 st	w/o EC
Swiss-Prot	20,302	3,498 (579)*	727 (14)	123 (0)	59 (2)	15,895
TrEMBL	51,337	2,680 (106)	292 (1)	38 (0)	29 (0)	48,298
TOTAL	71,639	6,178 (685)	1,019 (15)	161 (0)	88 (2)	64,193

w/: with EC (Enzyme Commission number); w/o: without EC.

EC 4th, EC 3rd, EC 2nd, EC 1st: four, three, two and one level/s of EC

(www.qmul.ac.uk/sbcs/iubmb/enzyme/rules.html).

*Among brackets, the number of sequences endowed with more than one EC number (polyfunctional enzymes).

Of the 64,193 proteins included in the HRP without an annotated EC number, BENZ correctly predicts as non-enzyme 56,872 sequences, leading to 7,321 false positive predictions.

To mitigate the rate of false positives I developed the following strategy. For each sequence which was wrongly annotated, I asked the question whether it is or not associated to a catalytic GO term, as derived from the UniProt file. If so, I retrieved an EC number with the EC2GO (www.ebi.ac.uk/GOA/EC2GO) mapping and compared it with the prediction. Matching turned the false positive into correct prediction. When a GO catalytic term was not present, the target Pfam architecture was adopted to map to InterPro (<https://www.ebi.ac.uk/interpro/>), which allows to retrieve the associated EC numbers. Even in this case, matching validates the predicted EC (see Figure 5.16).

This external validation retrieved 5,741 proteins from the false positive predictions, increasing the number of enzymes in the data set. In the end, BENZ scores with 96.84% of accuracy with a false negative rate equal to 3.57% and a false positive rate of 2.69% (see Table 5.5).

BENZ annotates four level EC numbers, but it can also reliably annotate polyfunctional enzymes and completes annotations of partial EC numbers. BENZ can also be applied to successfully annotated sequences as enzymes, increasing the number of enzymes in the human reference proteome.

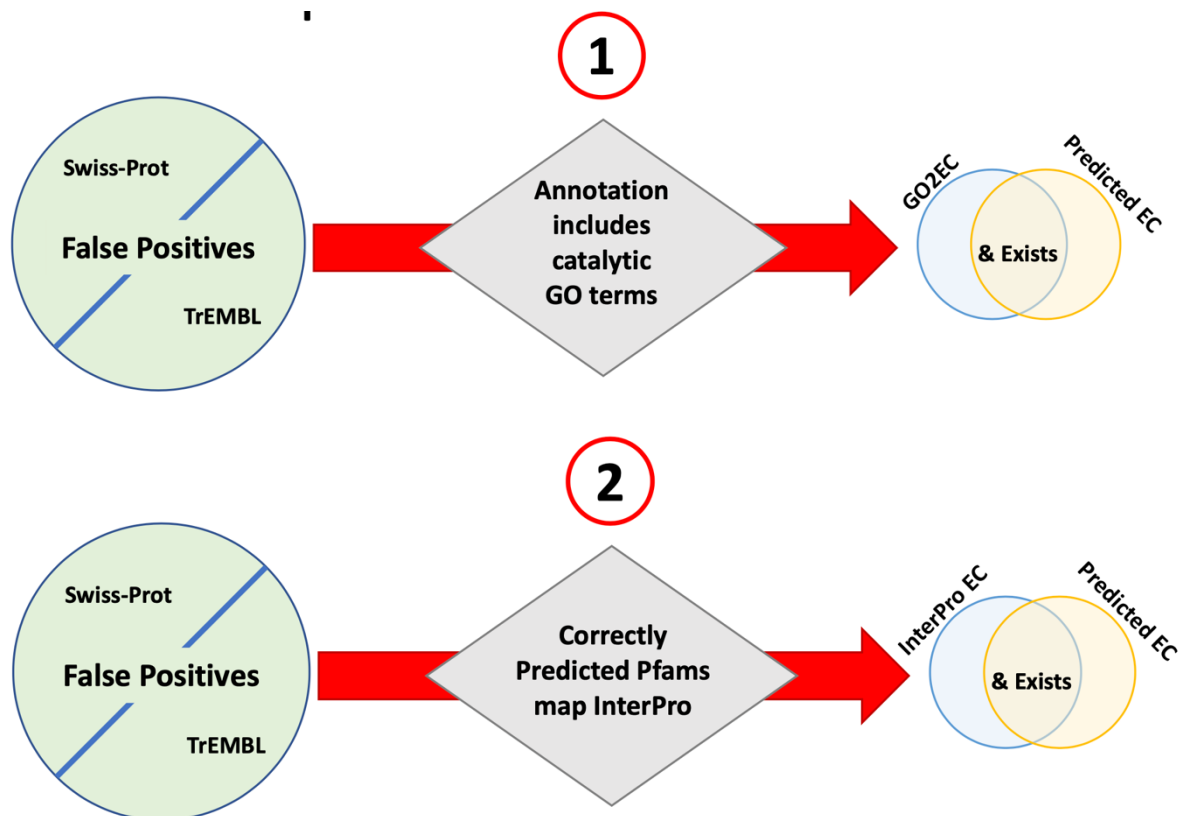


Figure 5.16. Graphical representation of the validating criteria for the re-evaluation of False Positive Predictions.

Table 5.5. BENZ overall performances

	Sequences	Rate
Correct Predictions	69,380 (Sw: 19531; Tr: 49677)	96.84 %
Wrong Predictions	205 (Sw: 150; Tr: 55)	2.06 %
Unpredicted (False Negative)	474 (Sw: 317; Tr: 157)	3.57 %
False Positive	1,580 (Sw: 304; Tr: 1276)	2.69 %
TOTAL	71,639 (Sw: 20302; Tr: 51337)	

BENZ overall performances against the Human Reference Proteome (HRP; Proteome ID: UP000005640; www.uniprot.org/proteomes/UP000005640)

D) The WEB Server

Query sequences submitted to BENZ are processed asynchronously thanks to an internal queuing service based on Sun Grid Engine [27]. The tool hmmscan from HMMER v.3.3.2 is used to align submitted sequences against predictive models and Pfam model and the user is provided with a temporary link that keeps track of the job progression and auto update every 30 seconds. On average, a job in BENZ takes within 1 minute but for sequences with a length greater than 3000 residues longer times may be required. BENZ returns to the users the predicted architecture for the submitted sequence and the EC annotation derived from the best matched reference sequence. In the output page, detailed information about the matched models and the predicted architectures can be found in the “Data” section. Tabular data are represented taking advantage of the library DataTables (<https://datatables.net>) while links are resolved with the Identifiers.org [28] services to improve interoperability. When the best retaining model is associated to more than one reference sequence, a dendrogram in Newick format is computed by the means of Clustal Omega and visualized with the Bio.Phylo module of Biopython [29]. Predicted architecture of the submitted sequences and the conservation of relevant site are displayed thanks to the Pviz.js library [30]. The web server is freely accessible without registration at <https://benzdb.biocomp.unibo.it> (see Figure 5.14).

5.3) References

- [1] Waldman AD, Fritz JM, Lenardo MJ. A guide to cancer immunotherapy: from T cell basic science to clinical practice. *Nat Rev Immunol.* 2020;20(11):651-668. doi:10.1038/s41577-020-0306-5
- [2] Şenbabaoğlu Y, Gejman RS, Winer AG, et al. Tumour immune microenvironment characterization in clear cell renal cell carcinoma identifies prognostic and immunotherapeutically relevant messenger RNA signatures [published correction appears in *Genome Biol.* 2017;18(1):46]. *Genome Biol.* 2016;17(1):231. doi:10.1186/s13059-016-1092-z
- [3] Rooney MS, Shukla SA, Wu CJ, Getz G, Hacohen N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell.* 2015;160(1-2):48-61. doi:10.1016/j.cell.2014.12.033
- [4] Ayers M, et al. IFN- γ -related mRNA profile predicts clinical response to PD-1 blockade. *J Clin Invest.* 2017;127(8):2930–2940.
- [5] Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *J Immunol.* 2017;199(9):3360-3368. doi:10.4049/jimmunol.1700893

- [6] Shou Y, Robinson DM, Amakye DD, et al. A five-gene hedgehog signature developed as a patient preselection tool for hedgehog inhibitor therapy in medulloblastoma. *Clin Cancer Res.* 2015;21(3):585-593. doi:10.1158/1078-0432.CCR-13-1711
- [7] Chang WH, Lai AG. Pan-cancer genomic amplifications underlie a WNT hyperactivation phenotype associated with stem cell-like features leading to poor prognosis. *Transl Res.* 2019;208:47-62. doi:10.1016/j.trsl.2019.02.008
- [8] Pietrobono S, Gagliardi S, Stecca B. Non-canonical Hedgehog Signaling Pathway in Cancer: Activation of GLI Transcription Factors Beyond Smoothed. *Front Genet.* 2019;10:556. doi:10.3389/fgene.2019.00556
- [9] Janssen KP, Alberici P, Fsihi H, et al. APC and oncogenic KRAS are synergistic in enhancing Wnt signaling in intestinal tumor formation and progression [published correction appears in *Gastroenterology*. 2006 Dec;131(6):2029]. *Gastroenterology.* 2006;131(4):1096-1109. doi:10.1053/j.gastro.2006.08.011
- [10] Li J, Mizukami Y, Zhang X, Jo WS, Chung DC. Oncogenic K-ras stimulates Wnt signaling in colon cancer through inhibition of GSK-3beta. *Gastroenterology.* 2005;128(7):1907-1918. doi:10.1053/j.gastro.2005.02.067
- [11] Lal N, White BS, Goussous G, et al. KRAS Mutation and Consensus Molecular Subtypes 2 and 3 Are Independently Associated with Reduced Immune Infiltration and Reactivity in Colorectal Cancer. *Clin Cancer Res.* 2018;24(1):224-233. doi:10.1158/1078-0432.CCR-17-1090
- [12] Loi S, Dushyanthen S, Beavis PA, et al. RAS/MAPK Activation Is Associated with Reduced Tumor-Infiltrating Lymphocytes in Triple-Negative Breast Cancer: Therapeutic Cooperation Between MEK and PD-1/PD-L1 Immune Checkpoint Inhibitors [published correction appears in *Clin Cancer Res.* 2019 Feb 15;25(4):1437]. *Clin Cancer Res.* 2016;22(6):1499-1509. doi:10.1158/1078-0432.CCR-15-1125
- [13] Uhrig S, Ellermann J, Walther T, et al. Accurate and efficient detection of gene fusions from RNA sequencing data. *Genome Res.* 2021;31(3):448-460. doi:10.1101/gr.257246.119
- [14] Haas BJ, Dobin A, Li B, Stransky N, Pochet N, Regev A. Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biol.* 2019;20(1):213. doi:10.1186/s13059-019-1842-9
- [15] Melsted P, Hateley S, Joseph IC, Pimentel H, Bray N, Pachter L., Fusion detection and quantification by pseudoalignment. *BioRxiv.* 166322. (2017). doi: 10.1101/166322
- [16] Heydt C, Wölwer CB, Velazquez Camacho O, et al. Detection of gene fusions using targeted next-generation sequencing: a comparative evaluation. *BMC Med Genomics.* 2021;14(1):62. doi:10.1186/s12920-021-00909-y
- [17] Bugg TDH. *Introduction to Enzyme and Coenzyme Chemistry*, 3rd ed.; Wiley: New York, NY, USA, 2012.
- [18] Savojardo C, Martelli PL, Casadio R. Protein-Protein Interaction Methods and Protein Phase Separation. *Ann. Rev. Biom. Data Sci.* 2020;89-112
- [19] Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res.* 2021;49(D1):D545-D551. doi:10.1093/nar/gkaa970
- [20] Savojardo C, Fariselli P, Martelli PL, Casadio R. ISPREd4: interaction sites PREdiction in protein structures with a refining grammar model. *Bioinformatics.* 2017;33(11):1656-1663. doi:10.1093/bioinformatics/btx044
- [21] Laciak AR, Korasick DA, Wyatt JW, Gates KS, Tanner JJ. Structural and biochemical consequences of pyridoxine-dependent epilepsy mutations that target the aldehyde binding site of aldehyde dehydrogenase ALDH7A1. *FEBS J.* 2020;287(1):173-189. doi:10.1111/febs.14997

- [22] Savojardo C, Babbi G, Martelli PL, Casadio R. Functional and Structural Features of Disease-Related Protein Variants. *Int J Mol Sci.* 2019;20(7):1530. doi:10.3390/ijms20071530
- [23] Mistry J, Chuguransky S, Williams L, et al. Pfam: The protein families database in 2021. *Nucleic Acids Res.* 2021;49(D1):D412-D419. doi:10.1093/nar/gkaa913
- [24] Dalkiran A, Rifaioglu AS, Martin MJ, Cetin-Atalay R, Atalay V, Doğan T. ECPred: a tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature. *BMC Bioinformatics.* 2018;19(1):334. doi:10.1186/s12859-018-2368-y
- [25] Kumar N, Skolnick J. EFICAz2.5: application of a high-precision enzyme function predictor to 396 proteomes. *Bioinformatics.* 2012;28(20):2687-2688. doi:10.1093/bioinformatics/bts510
- [26] Li Y, Wang S, Umarov R, et al. DEEPred: sequence-based enzyme EC number prediction by deep learning. *Bioinformatics.* 2018;34(5):760-769. doi:10.1093/bioinformatics/btx680
- [27] Gentzsch W. Sun Grid Engine: towards creating a compute power grid. *Proceedings First IEEE/ACM International Symposium on Cluster Computing and the Grid, 2001*, pp. 35-36, doi: 10.1109/CCGRID.2001.923173.
- [28] Juty N, Le Novère N, Laibe C. Identifiers.org and MIRIAM Registry: community resources to provide persistent identification. *Nucleic Acids Res.*, 2012;40, D580–D586.
- [29] Talevich E, Invergo BM, Cock PJ, Chapman BA. Bio.Phylo: a unified toolkit for processing, analyzing and visualizing phylogenetic trees in Biopython. *BMC Bioinformatics.* 2012; 13, 209.
- [30] Mukhyala K, Masselot A. Visualization of protein sequence features using JavaScript and SVG with pViz.js. *Bioinformatics.* 2014;30, 3408–3409.

6) Conclusions and Perspectives

In my thesis, I present results from different research projects. My PhD fellowship was granted by a joint effort between the University of Bologna and the Oncology Reference Center “CRO di Aviano”. During my PhD career I had the chance to greatly improve my knowledge and expertises in two research areas: Cancer Genomics and Enzyme Proteomics.

Currently, I'm the reference Bioinformatician of the department of Functional Oncogenetics and Oncogenomics at the Oncology Reference Center “CRO di Aviano” IRCCS. In the cancer genomics context, my efforts focused on understanding differences in the immune microenvironment of different types of GISTs and on assessing the ability to reliably detect fusion transcripts events of three commercially available assays.

As detailed in paragraph 5.1, we highlighted significant differences in the immune population levels of GISTs stratified accordingly to their oncogenic driver mutations. It was also demonstrated that poorly infiltrated tumours can be associated with antigen-specific immunity and oncogenes expression levels. Our results open new possibilities for patient specific immunotherapies in the field of personalized medicine. Many aspects of tumours onset and progression must still be investigated, especially for rare malignancies such as sarcomas. To this aim, we are currently expanding our in-House cohort of GISTs with some 100 new cases looking for rare and poorly characterized GIST phenotypes to increase the knowledge of this type of malignancies. Together with the ACC Sarcoma working group, I demonstrated that the Anchored Multiplex PCR panel FusionPlex Sarcoma (AMP-FPS) from ArcherDX (<https://archerdx.com/>) is the most suited for routinely diagnostic analysis requiring the detection of fusion events. I believe that the ACC consortium goal of transferring the technological innovation to clinical practice is fundamental to enable better and more sustainable care for cancer patients.

As previously stated, I also did research in the field of Enzyme proteomics. My research activities mainly focused on two connected topics: protein enzyme function prediction starting from sequences and the characterization of the enzyme-genetic disease association.

Inspecting the complex relation between enzymes and genetic diseases, we demonstrated that enzymes participating in multiple metabolic pathways are also hubs of the protein-protein interaction network. Results detailed in paragraph 5.2.1 highlight that genetic variations of hub proteins that occur in the solvent exposed surface of the protein are likely to affect enzyme

activity by influencing the ability to interact with other proteins and that this feature not necessarily affect thermodynamic stability.

One important step forward in the research area is that we established a statistically validated association between location of pathogenic variation types and structural Pfam and Interpro domains. It will be further exploited how this mapping can lead us to dissect the complexity of the relationship among disease related enzymes and Reactome pathways.

What is available online thanks to my efforts in collaboration with the Bologna Biocomputing group (<http://www.biocomp.unibo.it>) is BENZ WS The Bologna ENZYme Web Server (freely accessible at: <https://benzdb.biocomp.unibo.it/>). BENZ is first predictor of four-digit EC numbers, which for a given enzyme sequence, fully characterize its function. Noticeably, BENZ can predict polyfunctional enzymes and EC numbers belonging to the recently introduced functional class of translocases (7th class). My plans are to keep BENZ updated with respect to UniProt releases and refine the system to increase its accuracy. Additionally, I am planning to add disease-related information to BENZ reference sequences. The aim is to provide insights when the submitted target proteins are variants, not conserving important catalytic residues. Currently, I am also testing a new set of predictive criteria that will further improve the predictive power of BENZ. Overall, my work is presently described in six publications, and it has been presented in three international and national conferences (see LAA).

7) List of Articles and Abstracts (LAA)

7.1) Articles

Published

1. *Co-First Author*

Next-Generation Sequencing Approaches for the Identification of Pathognomonic Fusion Transcripts in Sarcomas: The Experience of the Italian ACC Sarcoma Working Group. Racanelli D, Brenca M, Baldazzi D, et al.; **Front Oncol.** 2020;10:489. DOI: 10.3389/fonc.2020.00489
Journal Impact Factor: 4.250 (2020)

2. **Highlighting Human Enzymes Active in Different Metabolic Pathways and Diseases: The Case Study of EC 1.2.3.1 and EC 2.3.1.9.** Babbi G, Baldazzi D, Savojardo C, Martelli PL, Casadio R. **Biomedicines** 2020, 8, 250. DOI: 10.3390/biomedicines8080250.

Journal Impact Factor: 3.600 (2020)

3. **Tumour genotype, location and malignant potential shape the immunogenicity of primary, untreated Gastrointestinal Stromal Tumors.** Gasparotto D, Sbaraglia M, Rossi S, Baldazzi D, Brenca M, Mondello A, Nardi F, Racanelli D, Cacciatore M, Dei Tos A, Maestro R. **JCI Insight** 2020. DOI: 10.1172/jci.insight.142560

Journal Impact Factor: 6.205 (2020)

4. *First Author*

BENZ WS: the Bologna ENZyme Web Server for four-level EC number annotation. Baldazzi, D; Castrense, S.; Martelli, PL, Casadio R. **Nucleic Acids Research**; 2021. DOI: 10.1093/nar/gkab328

Journal Impact Factor: 11.501 (2021)

5. **A glance into MTHFR deficiency at a molecular level.** Savojardo C, Babbi G, Baldazzi D, Martelli PL, Casadio R. **Int. J. Mol. Sci.** 2021, 23(1):167.

Journal Impact Factor: 5.900 (2021)

Under Review

6. **Disease-related variation types in human genes link to maladies with Pfam and InterPro mapping.** Babbi G, Savojardo C, Baldazzi D, Martelli PL, Casadio R. **Human Genetics**, 2021. *Under Review*
Journal Impact Factor: 5.743 (2021)

In Preparation

7. *First Author*
Baldazzi D, Savojardo C, Martelli PL, Casadio R. **The role of disease related enzymes in the Reactome pathways.** (2021/2022)

7.2) Abstracts and Oral Presentations

8. **Myoepithelial tumours of soft tissues and extraskeletal myxoid chondrosarcomas feature a distinct transcriptional pattern.** Racanelli D, Stacchiotti S, Brenca M, Sbaraglia M, Fassetta K, Baldazzi D, Piccinin S, Brich , Casali P.G., Collini P, Dagrada GP, Fiore M, Gronchi A, Astolfi A, Pantaleo MA, Righi A, Pilotti S, Dei Tos AP, Maestro R. **Annals of Oncology** 2019, Vol 30, Suppl. 5, V704 DOI: 10.1093/annonc/mdz283.054
Journal Impact Factor: 18.274 (2019)
9. **Epithelioid sarcoma: Molecular insights into proximal versus classic variant.** Frezza AM, Sigalotti L, Del Savio E, Baldazzi D, Sbaraglia M, Righi A, Gambarotti M, Barisella M, Brich S, Dei Tos AP, Stacchiotti S, Maestro R. **Journal of Clinical Oncology** 2020, 38, no. 15_suppl. DOI: 10.1200/jco.2020.38.15_suppl.e23552
Journal Impact Factor: 18.230 (2020)
10. *Presenting Author*
BENZ WS annotates sequences of the human reference proteome with four level EC numbers. Baldazzi D, Savojardo C, Martelli PL, Casadio R. **Oral presentation upon selection and poster presentation at ISMB/ECCB 2021**, the international joint conferences (July 25-30 2021) among the 29th conference on Intelligent Systems for Molecular Biology and the 20th European Conference on Computational Biology with

more than 2,100 participants from nearly 80 countries
(<https://www.iscb.org/ismbeccb2021>)

11. *Presenting Author*

BENZ WS: the Bologna ENZYme Web Server for four-level EC number annotation. Baldazzi D, Savojardo C, Martelli PL, Casadio R. **Società Italiana Di Biochimica e Biologia Molecolare (SIB)-Computational and System Biology Group. Congresso Nazionale SIB 23-24 settembre 2021. Poster and Oral presentation upon selection.**

12. **Rhabdomyoblastic Dedifferentiation in Retroperitoneal Liposarcomas is associated with reduced Immune Infiltration.** Valenti B, Pasquali S, Brenca M, Brich S, Baldazzi D, Racanelli D, De Benedictis I, Colombo C, Sbaraglia M, Vallacchi V, Castelli C, Stacchiotti S, Dei Tos AP, Collini P, Gronchi A, Maestro R. **Connective Tissue Oncology Society (CTOS) Meeting 2021** (<https://www.ctos.org/Meeting/2021AnnualMeeting.aspx>)

13. *Presenting Author*

BENZ WS: the Bologna ENZYme Web Server for four-level EC number annotation. Baldazzi D, Savojardo C, Martelli PL, Casadio R. **Elixir 3D-Bioinfo Annual Meeting** (<https://elixir-europe.org/events/3d-bioinfo-2021-annual-meeting>) **Poster and Oral presentation upon selection.**

8) Acknowledgments

First and foremost, I want to thank my supervisor, Prof. Rita Casadio, for giving me the opportunity to work with her over the past four years and before. I began with little experience, but she chose to invest endless hours into me as a student and a scientist. She was quick to scold and forgive when I made mistakes and constantly reminded me about the big picture of why we work hard every day. She has taught me that, in the academic environment, success is the result of persistent, honest science. I will do my best to carry her relentless approach to science with me in my future ventures.

I would also like to thank my co-supervisor Dr. Roberta Maestro for her guidance and support. Learning from and working alongside her was one of the greatest joys of my PhD. I undoubtedly have a much deeper respect for the complexity of Cancer Genetics due to her mentorship. Dr. Roberta Maestro generously provided for any need throughout my PhD and pushed me to apply my research beyond my expectation.

I would like to thank the entire Bologna Biocomputing Group for creating a fun, social and caring environment as well as the Unit of Oncogenetics and Functional Oncogenomics at the “CRO Aviano” National Cancer Institute. I would like to thank Pier Luigi Martelli, Castrense Savojardo and Giulia Babbi. They helped me troubleshoot my science and they endlessly encouraged me. Furthermore, I would like to thank Daniela Gasparotto, Luca Sigalotti, Valentina Damiano, Beatrice Valenti and Elisa del Savio for their relentless effort to teach me how to properly cooperate in a genomics/transcriptomics research group.

I would also like to thank Dr. Hannah Carter and Prof. Emidio Capriotti whose efforts allowed me to travel to San Diego (California, USA) and experiencing how research is done abroad.

I would like to thank my parents, Enrico, and Milena, for raising me in a loving environment without ever stopping me from following my own aspirations. I would also like to thank my sister, Lucrezia, for her continuous support. Furthermore, I would like to thank my friends starting from my childhood friends (they're too many to be cited one by one so I am mentioning them as they like it: “Associazione Bocciofila”) up to the friends I made during the last years (Fabrizio, Felix, Giulia, Idris, Leonardo, Lorenzo, and Luca).

Finally, I would like to thank my coaches and teammates from Bologna Rugby Club and Romagna Rugby Club. They were integral in instilling the value of hard work and perseverance, two qualities key to the success of this PhD.

9) Appendix (Publications)

1. **Next-Generation Sequencing Approaches for the Identification of Pathognomonic Fusion Transcripts in Sarcomas: The Experience of the Italian ACC Sarcoma Working Group.** Racanelli D, Brenca M, Baldazzi D, et al.; **Front Oncol.** 2020;10:489. DOI: 10.3389/fonc.2020.00489
Journal Impact Factor: 4.250 (2020)
2. **Highlighting Human Enzymes Active in Different Metabolic Pathways and Diseases: The Case Study of EC 1.2.3.1 and EC 2.3.1.9.** Babbi G, Baldazzi D, Savojardo C, Martelli PL, Casadio R. **Biomedicines** 2020, 8, 250. DOI: 10.3390/biomedicines8080250.
Journal Impact Factor: 3.600 (2020)
3. **Tumour genotype, location and malignant potential shape the immunogenicity of primary, untreated Gastrointestinal Stromal Tumors.** Gasparotto D, Sbaraglia M, Rossi S, Baldazzi D, Brenca M, Mondello A, Nardi F, Racanelli D, Cacciatore M, Dei Tos A, Maestro R. **JCI Insight** 2020. DOI: 10.1172/jci.insight.142560
Journal Impact Factor: 6.205 (2020)
4. **BENZ WS: the Bologna ENZYme Web Server for four-level EC number annotation.** Baldazzi, D; Castrense, S.; Martelli, PL, Casadio R. **Nucleic Acids Research**; 2021. DOI: 10.1093/nar/gkab328
Journal Impact Factor: 11.501 (2021)
5. **A glance into MTHFR deficiency at a molecular level.** Savojardo C, Babbi G, Baldazzi D, Martelli PL, Casadio R. **Int. J. Mol. Sci.** 2021, 23(1):167.
Journal Impact Factor: 5.900 (2021)
6. **Disease-related variation types in human genes link to maladies with Pfam and InterPro mapping.** Babbi G, Savojardo C, Baldazzi D, Martelli PL, Casadio R. **Human Genetics**, 2021. *Under Review*
Journal Impact Factor: 5.743 (2021)



OPEN ACCESS

Edited by:

Massimo Brogginì,
Istituto Di Ricerche Farmacologiche
Mario Negri, Italy

Reviewed by:

Don A. Baldwin,
Fox Chase Cancer Center,
United States
Tricarico Rossella,
Fox Chase Cancer Center,
United States

***Correspondence:**

Rita Falcioni
rita.falcioni@iffo.gov.it
Roberta Maestro
maestro@cro.it

†These authors have contributed
equally to this work and share first
authorship

‡These authors share last authorship

Specialty section:

This article was submitted to
Cancer Molecular Targets and
Therapeutics,
a section of the journal
Frontiers in Oncology

Received: 13 December 2019

Accepted: 18 March 2020

Published: 15 April 2020

Citation:

Racaneli D, Brenca M, Baldazzi D,
Goeman F, Casini B, De Angelis B,
Guercio M, Milano GM, Tamborini E,
Busico A, Dagrada G, Garofalo C,
Caruso C, Brunello A, Pignochino Y,
Berrino E, Grignani G, Scotlandi K,
Parra A, Hattinger CM, Ibrahim T,
Mercatali L, De Vita A, Carriero MV,
Pallocca M, Loria R, Covello R,
Sbaraglia M, Dei Tos AP, Falcioni R
and Maestro R (2020)
Next-Generation Sequencing
Approaches for the Identification of
Pathognomonic Fusion Transcripts in
Sarcomas: The Experience of the
Italian ACC Sarcoma Working Group.
Front. Oncol. 10:489.
doi: 10.3389/fonc.2020.00489

Next-Generation Sequencing Approaches for the Identification of Pathognomonic Fusion Transcripts in Sarcomas: The Experience of the Italian ACC Sarcoma Working Group

Dominga Racaneli^{1†}, Monica Brenca^{1†}, Davide Baldazzi^{1†}, Frauke Goeman^{2†}, Beatrice Casini^{2†}, Biagio De Angelis³, Marika Guercio³, Giuseppe Maria Milano³, Elena Tamborini⁴, Adele Busico⁴, Gianpaolo Dagrada⁴, Cecilia Garofalo⁵, Chiara Caruso⁵, Antonella Brunello⁶, Ymera Pignochino⁷, Enrico Berrino⁸, Giovanni Grignani⁷, Katia Scotlandi⁹, Alessandro Parra⁹, Claudia Maria Hattinger⁹, Toni Ibrahim¹⁰, Laura Mercatali¹⁰, Alessandro De Vita¹⁰, Maria Vincenza Carriero¹¹, Matteo Pallocca², Rossella Loria², Renato Covello², Marta Sbaraglia¹², Angelo Paolo Dei Tos^{12,13}, Rita Falcioni^{2*†} and Roberta Maestro^{1*†}

¹ Unit of Oncogenetics and Functional Oncogenomics, Centro di Riferimento Oncologico di Aviano (CRO Aviano) IRCCS, National Cancer Institute, Aviano, Italy, ² Department of Research, Diagnosis and Innovative Technology, IRCCS Regina Elena National Cancer Institute, Rome, Italy, ³ Department of Onco-Haematology and Cell and Gene Therapy Unit, Bambino Gesù Children's Hospital, IRCCS, Rome, Italy, ⁴ Department of Pathology, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy, ⁵ Advanced Translational Research Laboratory, Veneto Institute of Oncology IOV – IRCCS, Padua, Italy, ⁶ Medical Oncology 1, Department of Oncology, Veneto Institute of Oncology IOV – IRCCS, Padua, Italy, ⁷ Division of Medical Oncology, Candiolo Cancer Institute, FPO-IRCCS, Candiolo, Italy, ⁸ Unit of Pathology, Candiolo Cancer Institute FPO-IRCCS, Candiolo, Italy, ⁹ Laboratory of Experimental Oncology, IRCCS Istituto Ortopedico Rizzoli, Bologna, Italy, ¹⁰ Osteoncology and Rare Tumors Center, Istituto Scientifico Romagnolo per lo Studio e la Cura dei Tumori (IRST) IRCCS, Meldola, Italy, ¹¹ Tumor Progression Unit, Department of Experimental Oncology, Istituto Nazionale Tumori Fondazione “G. Pascale” IRCCS, Naples, Italy, ¹² Department of Pathology, Azienda Ospedaliera Universitaria di Padova, Padua, Italy, ¹³ Department of Medicine, University of Padua School of Medicine, Padua, Italy

This work describes the set-up of a shared platform among the laboratories of the Alleanza Contro il Cancro (ACC) Italian Research Network for the identification of fusion transcripts in sarcomas by using Next Generation Sequencing (NGS). Different NGS approaches, including anchored multiplex PCR and hybrid capture-based panels, were employed to profile a large set of sarcomas of different histotypes. The analysis confirmed the reliability of NGS RNA-based approaches in detecting sarcoma-specific rearrangements. Overall, the anchored multiplex PCR assay proved to be a fast and easy-to-analyze approach for routine diagnostics laboratories.

Keywords: sarcoma, molecular diagnosis, fusion transcripts, NGS, anchored multiplex PCR, hybrid capture-based panel

INTRODUCTION

The term “sarcoma” identifies a heterogeneous group of rare tumors comprising over 60 different histologic variants (1). Due to their rarity and heterogeneity, the accuracy of sarcoma diagnosis remains challenging. In the diagnosis of sarcomas, tumor cell morphology (shape, pattern of growth, microenvironment contexture) and the expression of differentiation markers

represent the most important factors, but molecular investigations are increasingly employed to complement these pathological assessments. Indeed, the identification of histotype-specific (pathognomonic) gene alterations is of paramount importance in the differential diagnosis among sarcoma variants, between malignant and benign mimics, as well as between sarcoma and other tumor types (1–3). In particular, about one third of all sarcomas presents pathognomonic chromosome rearrangements (translocations, deletions, insertions) that result in fusion genes and corresponding expression of fusion transcripts (4). Beside diagnostic relevance, the expression of fusion transcripts may have prognostic and/or predictive implications. For example, certain rearrangements, such as those involving *ALK* in inflammatory myofibroblastic tumors or *COL1A1-PDGFB* in dermatofibrosarcoma protuberans, are predictive of the response to tyrosine kinase inhibitors (5, 6). Moreover, the detection of NTRK fusions in a broad range of malignancies, including sarcomas, has gaining much attention due to the recent demonstration of therapeutic efficacy of a new class of tyrosine kinase inhibitors in NTRK rearranged tumors (7–9).

Commonly, FISH or RT-PCR are used to detect fusion events at the genomic or transcriptional level, respectively. However, both methods present limitations. In particular, since they are suited to investigate a specific pre-defined abnormality, they inevitably rely on a prior diagnostic hypothesis (reflex testing). The advent of technologies such as next generation sequencing (NGS), aka massive parallel sequencing, has laid down the bases to overcome this limitation. By allowing the simultaneous analysis of a large set of targets (from few genes to the whole transcriptome/genome) NGS has disclosed the possibility not only to reveal diagnostic/prognostic/predictive genetic abnormalities in the absence of a prior hypothesis but also to identify new aberrations (10–12).

Here we wanted to assess feasibility, reliability, and applicability of NGS-based methods for the detection of sarcoma-associated fusion transcripts in a routine diagnostic setting. Our multicentric analysis confirms the sensitivity of anchored-based NGS profiling approaches and corroborates the suitability of these investigations in the diagnostic setting of sarcomas.

MATERIALS AND METHODS

Case Selection

The study was conducted on a series of 150 sarcoma samples, representative of different sarcoma histotypes, retrieved from the pathological files of the participating institutions (Alleanza Contro il Cancro, ACC, Italian Research Network). Either Formalin-Fixed Paraffin-Embedded (FFPE) or frozen samples were analyzed. All sarcomas included in the study

Abbreviations: NGS, next generation sequencing; FFPE, Formalin-Fixed Paraffin-Embedded; FISH, fluorescence *in situ* hybridization; RT-PCR, reverse transcriptase-PCR; RT-qPCR, reverse transcriptase-quantitative PCR; IHC, immunohistochemistry; HC, hybrid capture-based panel; AMP-FPS, Anchored Multiplex PCR FusionPlex Sarcoma panel; TS-Fusion, TruSight RNA Fusion panel; TS-PanCancer, TruSight RNA PanCancer panel

were histopathologically re-evaluated on hematoxylin-eosin stained slides, and representative areas were selected for molecular analyses.

NGS-based Fusion Transcript Identification

RNA was extracted from 5 to 10 μm -FFPE tissue sections using the Qiagen miRNeasy FFPE kit (Qiagen, Valencia, CA, USA) or the Invitrogen RecoverAll Total Nucleic Acid Isolation kit (Thermo Fisher Scientific, Waltham, MA, USA). For frozen samples the TRIzol reagent (Life Technologies Italia, Monza, Italy) followed by the RNeasy MinElute cleanup (Qiagen, Valencia, CA, USA) was used. Total RNA was quantified by using a Qubit fluorometer (Thermo Fisher Scientific, Waltham, MA, USA). Quality was checked with the RNA 6000 Nano Kit on a 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA), or by using the Archer PreSeqTM RNA QC qPCR Assay (ArcherDX, Boulder, CO, USA) and a threshold of DV₂₀₀ >30 or PreSeq Cq <31 was used to identify high quality RNA, respectively.

FISH, RT-PCR, RT-qPCR, and IHC, used as primary detection approaches for the detection of possible fusion events, were performed during routine diagnostic procedures according to laboratory standard guidelines and validated reagents.

Three different commercially available NGS-based fusion panels were selected based on their capacity to cover most genes known to be involved in sarcoma-relevant fusions: an anchored multiplex PCR-based assay, namely the Archer FusionPlex Sarcoma kit (AMP-FPS)(ArcherDX, Boulder, CO, USA), covering 26 genes involved in sarcoma-associated fusions; two hybrid capture-based (HC) assays, namely the TruSight RNA Fusion Panel (TS-Fusion) (Illumina Inc., San Diego, CA, USA) and the TruSight RNA PanCancer Panel (TS-PanCancer) (Illumina Inc., San Diego, CA, USA) covering 507 and 1,385 genes commonly involved in cancer, respectively. Both HC assays included the 26 genes covered by the AMP-FPS kit. In a subset of samples, a customized version of the AMP-FPS panel was used to detect PAX3 fusion transcripts. Specifically, the assay was integrated with PAX3-specific primers (exons 6, 7 and 8) designed by using the Archer Assay Designer tool (ArcherDX, Boulder, CO, USA).

Libraries for all three panels were prepared and checked for quality according to the manufacturer's instructions, starting from 100 to 250 ng of RNA as input.

AMP-FPS libraries were run on either Illumina (MiSeq or NextSeq 500 Illumina Inc., San Diego, CA, USA) or Thermo (Ion S5 Thermo Fisher Scientific, Waltham, MA, USA) sequencing platforms, according to the manufacturer's instructions. HC-based libraries were sequenced on Illumina MiSeq instruments. Illumina TS-Fusion and TS-PanCancer sequencing data were analyzed by using the dedicated Illumina BaseSpace RNA-Seq Alignment tool (v.s.2.0.2), which relies on STAR and Manta algorithms (13, 14). PAR-masked/(RefSeq)hg19 was used as reference genome. A minimum of 3 million reads was obtained per sample (range 3007307–6284475). The mean percentage of reads aligned to the human genome was 98.9% (range 96.4–99.7%); the mean proportion of reads aligned to ribosomal RNA was below 2% (range 0.2–6.1%) and mean insert size was 134 bp

(range 107–155 bp), in line with literature data (15). Only high-confidence fusions that passed default thresholds of the RNA-Seq Alignment tool (PASS) were recorded.

The Archer Analysis suite (v 5.1 or v 6.0) was exploited for the analysis of AMP-FPS panel results, using default settings. Default parameters (QC PASS) that, according to the Archer user manual, allow to achieve up to 95% of sensitivity in fusion detection, were employed to assess data quality. Samples included in the study met the quality cutoffs set by the Archer Analysis platform but in a few cases that, although not fulfilling all default criteria, nevertheless yielded high confidence fusion calls (cases #9, 31, 37, 47, 57, 60, 80, 126). Fusions were recorded as “high confidence calls” (strong = true in output table) if they passed all “strong evidence” default filters as described in the Archer analysis user manual (briefly: breakpoint spanning reads that support the candidate ≥ 5 ; “fusion_percent_of_GSP2_reads”, i.e., proportion of breakpoint spanning reads that support the candidate relative to the total number of reads spanning the breakpoint $\geq 10\%$; “min_unique_start_sites_for_strong_fusion” ≥ 3 ; fusion recorded in the Quiver database or not fulfilling the “negative evidence criteria”).

Of 48 cases (12 of the first set and 36 of the second set) where a fusion was detected by NGS but the partner genes had not been previously determined by the primary detection method, material was available for orthogonal validations (RT-PCR) in 39 cases, confirming NGS results. The involvement of *SSX4* (*SS18-SSX4*), called sometime by the AMP-FPS assay in synovial sarcoma samples, was checked by nested RT-PCR (primers: Fw-SS18 GGACCACCACAGCCACCCCA, Rev-SSX ATGTTTCCCCCTTTGGGTC; Rev-SSX4 GTCTTGTTAATC TTCTCCAAGG) and Sanger sequencing on a single index case.

For second level bioinformatic analyses of HC library raw data, Arriba, STAR-Fusion and Pizzly (16–18), administered through a command line interface, were employed for fusion calling using default settings.

RESULTS

NGS-based Identification of Fusion Transcripts: Panel Comparison

As a first step toward the assessment of suitability of NGS-based approaches for the detection of pathognomonic fusions in sarcomas, performance and ease-of-use (library preparation complexity, hands-on time, user-friendly dedicated bioinformatic analysis tool) of three different NGS fusion panels were evaluated on a set of sarcoma samples previously characterized by either FISH or RT-qPCR for gene fusions (Table 1). Twenty-six samples were analyzed with a hybrid capture-based panel (HC) (Illumina TS-Fusion). Twenty samples were analyzed with an anchored multiplex PCR panel (Archer AMP-FPS), 19 of which investigated also with the Illumina TS-Fusion. In addition, 9 samples were profiled with a more comprehensive HC panel (Illumina TS-PanCancer).

All three targeted RNA-sequencing panels permit the identification of common and known fusions involved in sarcomas, but also the discovery of novel fusions. The AMP-FPS panel targets a limited set of genes (26 target genes) that are

commonly involved in sarcoma-associated fusions. This AMP-FPS panel employs unidirectional gene-specific primers to detect fusion transcripts involving target genes. In addition, molecular barcodes are included to enable single molecule counting, de-duplication and error correction, thus allowing quantitative analysis and confident mutation calling.

In HC-based panels the transcripts of interest are enriched by hybridization and capture with biotinylated probes (507 genes in TS-Fusion, 1385 genes in TS-PanCancer, in both cases including the 26 genes targeted by the AMP-FPS panel).

Raw data obtained with the different panels were then analyzed using the dedicated bioinformatic suite (BaseSpace RNA-Seq Alignment for Illumina HC panels, Archer Analysis platform for the AMP-FPS panel). The AMP-FPS assay correctly identified the pathognomonic fusion in all samples analyzed (20/20), irrespective of the sequencing platform used (Thermo and/or Illumina), demonstrating an excellent sensitivity. The pathognomonic fusion was correctly called in 22/26 samples analyzed with the TS-Fusion HC assay. Of the 9 cases analyzed with the TS-PanCancer HC panel, the dedicated bioinformatic tool identified the diagnostic fusion in 7 cases, in one of these as a reciprocal fusion. To further explore the performance of HC panels, data generated with TS-Fusion and TS-PanCancer panels were re-evaluated with additional algorithms, namely Arriba, STAR-Fusion and Pizzly (16–18). Although impractical in a routine diagnostic setting, as they rely on a command line interface, these tools are reported to have high fusion detection rates (16–18). With the exception of case #27, for which no algorithm detected, as high confidence calls, fusions involving the *CIC* gene, apparently rearranged according to FISH, at least one fusion caller was capable of detecting, among others, a fusion transcript involving the target gene in cases previously scored negative with the BaseSpace RNA-Seq Alignment tool, emphasizing the importance of software sensitivity in data analysis (Supplemental Tables 1–3).

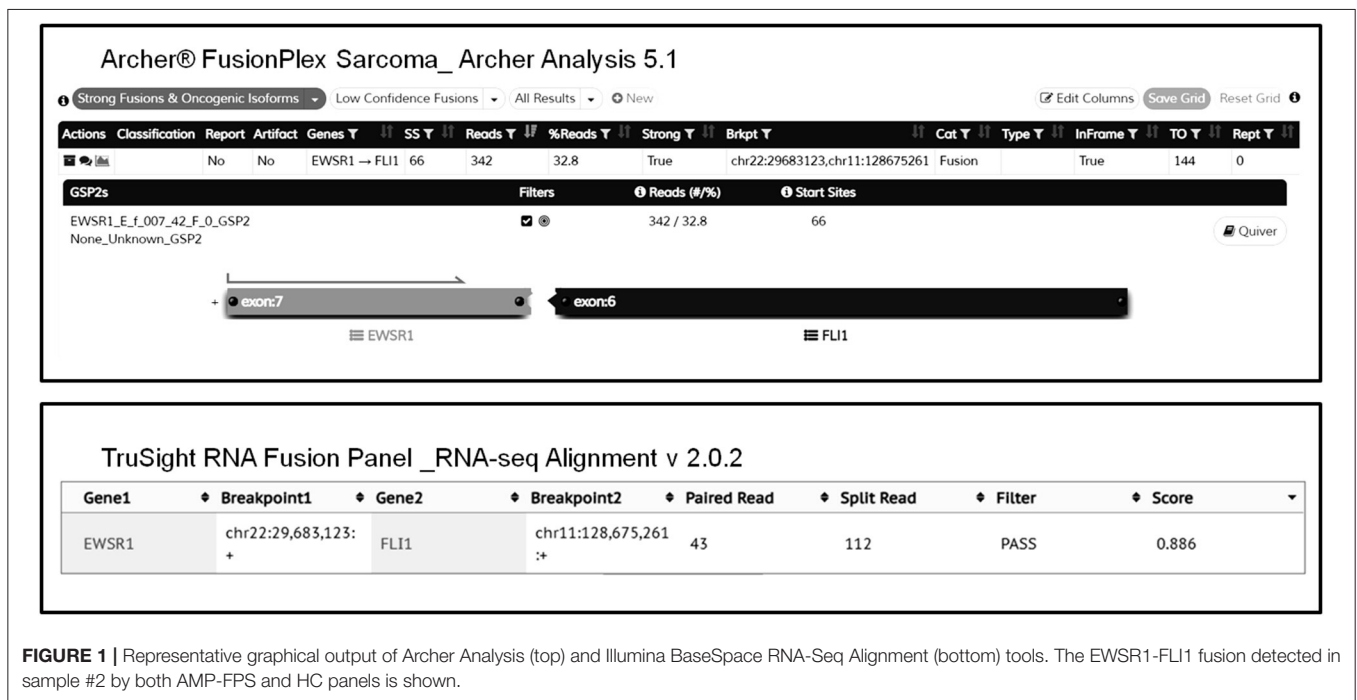
Additional passing filters fusions (in frame and out of frame) were occasionally called beside the pathognomonic one, but the actual biological significance of these alterations is unclear. For instance, beside the canonical fusion involving *SS18* and *SSX1* or *SSX2*, additional fusions involving *SSX4* were called in 5/6 synovial sarcomas analyzed with the AMP-FPS panel. It should be pointed out that the AMP-FPS approach relies on relatively small amplicons. Thus, in the presence of highly homologous genes (e.g., *SSX1*, *SSX2*, *SSX4*), this technique may fail to properly distinguish the target (19). Indeed, a deeper analysis of an index case confirmed the expression of *SS18-SSX1*, suggesting that the alleged *SS18-SSX4* fusion was likely an alignment artifact.

Overall, both AMP-FPS and HC assays demonstrated a good detection capability. The HC assays were definitively more comprehensive and suitable for a research environment. In contrast, the AMP-FPS panel was limited in breadth (only 26 target genes), and hence with reduced capacity of discovering new fusions, but definitively provided for a better ease-of-use. In particular, the hands-on-time for library preparation was reduced. Moreover, compared to the BaseSpace RNA-Seq Alignment, the AMP-FPS dedicated bioinformatic analysis tool (Archer Analysis platform) featured a more user-friendly graphical interface with detailed and straightforward information

TABLE 1 | NGS fusion profiling: panel comparison.

Nr	Diagnosis	Pre-detected genetic abnormality	Primary detection method	Histotype-specific fusion detected by the indicated NGS approach			Other passing filters fusions (assay detecting the additional fusion)
				AMP-FPS	TS-Fusion	TS-PanCancer	
1	Dermatofibrosarcoma Protuberans	<i>PDGFB</i>	FISH	<i>COL1A1-PDGFB^{IL}</i>	<i>COL1A1-PDGFB</i>	<i>COL1A1-PDGFB</i>	NFD
2	Ewing Sarcoma	<i>EWSR1</i>	FISH	<i>EWSR1-FLI1^{IL}</i>	<i>EWSR1-FLI1</i>	<i>EWSR1-FLI1</i>	NFD
3	Infantile Fibrosarcoma	<i>ETV6</i>	FISH	<i>ETV6-NTRK3^{IL}</i>	<i>ETV6-NTRK3</i>	<i>ETV6-NTRK3</i>	NFD
4	Synovial Sarcoma	<i>SS18-SSX1</i>	RT-qPCR	<i>SS18-SSX1^{IL}</i>	<i>SS18-SSX1</i>	<i>SS18-SSX1</i>	<i>SS18-SSX4</i> (AMP-FPS ^{IL})
5	Synovial Sarcoma	<i>SS18</i>	FISH	<i>SS18-SSX2^{IL}</i>	<i>SS18-SSX2</i>	<i>SS18-SSX2</i>	<i>SS18-SSX4</i> (AMP-FPS ^{IL})
6	Myoepithelioma (soft tissue)	<i>EWSR1</i>	FISH	<i>EWSR1-ATF1^{IL}</i>	<i>EWSR1-ATF1</i>	NFD	<i>ATF1-EWSR1</i> (TS-Fusion, TS-PanCancer)
7	Extraskelletal Myxoid Chondrosarcoma	<i>EWSR1-NR4A3</i>	RT-qPCR	<i>EWSR1-NR4A3^{IL}</i>	NFD	NFD	NFD
8	Clear Cell sarcoma	<i>EWSR1</i>	FISH	<i>EWSR1-ATF1^{T,IL}</i>	NFD	nd	NFD
9	Ewing Sarcoma	<i>EWSR1-FLI1</i>	RT-qPCR	<i>EWSR1-FLI1^{T,IL}</i>	<i>EWSR1-FLI1</i>	nd	NFD
10	Ewing Sarcoma	<i>EWSR1-FLI1</i>	RT-qPCR	<i>EWSR1-FLI1^{T,IL}</i>	<i>EWSR1-FLI1</i>	nd	NFD
11	Ewing Sarcoma	<i>EWSR1-ERG</i>	RT-qPCR	<i>EWSR1-ERG^{T,IL}</i>	<i>EWSR1-ERG</i>	nd	<i>EWSR1-ERG-EWSR1</i> (AMP-FPS ^{IL})
12	Extraskelletal Myxoid Chondrosarcoma	<i>EWSR1-NR4A3</i>	RT-qPCR	<i>EWSR1-NR4A3^T</i>	<i>EWSR1-NR4A3</i>	nd	NFD
13	Myxoid Liposarcoma	<i>FUS-DDIT3</i>	RT-qPCR	<i>FUS-DDIT3^{IL}</i>	<i>FUS-DDIT3</i>	nd	NFD
14	Myxoid Liposarcoma	<i>FUS-DDIT3</i>	RT-qPCR	<i>FUS-DDIT3^{T,IL}</i>	<i>FUS-DDIT3</i>	nd	<i>DDIT3-FUS</i> (TS-Fusion)
15	Myxoid Liposarcoma	<i>FUS-DDIT3</i>	RT-qPCR	<i>FUS-DDIT3^{T,IL}</i>	<i>FUS-DDIT3</i>	nd	<i>FUS-DDIT3-DLG2</i> (AMP-FPS ^{IL})
16	Synovial Sarcoma	<i>SS18-SSX1</i>	RT-qPCR	<i>SS18-SSX1^{IL}</i>	<i>SS18-SSX1</i>	nd	<i>SS18-SSX4-SS18</i> ; <i>SS18-SSX4</i> (AMP-FPS ^{IL})
17	Synovial Sarcoma	<i>SS18</i>	FISH	<i>SS18-SSX1^{IL}</i>	<i>SS18-SSX1</i>	nd	NFD
18	Synovial Sarcoma	<i>SS18-SSX1</i>	RT-qPCR	<i>SS18-SSX1^{IL}</i>	<i>SS18-SSX1</i>	nd	<i>SS18-SSX4</i> (AMP-FPS ^{IL})
19	Synovial Sarcoma	<i>SS18-SSX1</i>	RT-qPCR	<i>SS18-SSX1^{T,IL}</i>	<i>SS18-SSX1</i>	nd	<i>SS18-SSX1/4-SS18</i> ; <i>SS18-SSX4</i> (AMP-FPS ^{IL})
20	Myxoid Liposarcoma	<i>DDIT3</i>	FISH	<i>FUS-DDIT3^{IL}</i>	nd	<i>FUS-DDIT3</i>	<i>DDIT3-FUS</i> (TS-PanCancer)
21	Myxoid Liposarcoma	<i>DDIT3</i>	FISH	nd	<i>FUS-DDIT3</i>	NFD	NFD
22	Synovial Sarcoma	<i>SS18</i>	FISH	nd	<i>SS18-SSX1</i>	nd	NFD
23	Synovial Sarcoma	<i>SS18</i>	FISH	nd	<i>SS18-SSX1</i>	nd	NFD
24	Myxoid Fibrosarcoma	<i>FUS</i>	FISH	nd	<i>FUS-CREB3L2</i>	nd	NFD
25	Myxoid Liposarcoma	<i>FUS-DDIT3</i>	RT-qPCR	nd	<i>FUS-DDIT3</i>	nd	<i>DDIT3-FUS</i> (TS-Fusion)
26	Myxoid Liposarcoma	<i>DDIT3</i>	FISH	nd	NFD	nd	NFD
27	Undifferentiated Round Cell, Ewing-Like Sarcoma	<i>CIC</i>	FISH	nd	NFD	nd	NFD

NFD, no histotype-specific fusion detected; nd, not done; FISH, fluorescent in situ hybridization; RT-qPCR, reverse transcriptase- quantitative PCR; Sequencing platform used: T, Thermo platform; IL, Illumina platform.



about the fusion (exons involved, in frame/out of frame, confidence of the call) (Figure 1).

On the whole, we considered the AMP-FPS assay more suitable for routine diagnostics.

Validation on a Larger Set of Cases of the AMP-FPS Fusion Transcript Assay

Based on these results, with a view to translating NGS-based fusion identification in a routine diagnostic setting, we sought to extend the evaluation of the AMP-FPS panel (on either a Thermo or an Illumina sequencing platform) to 123 additional cases (Table 2).

Overall, the AMP-FPS panel confirmed the good performance. Of 81 cases with a pre-detected genetic abnormality suggestive of a fusion event, this NGS assay proved effective in 71, with orthogonal validations (RT-PCR) confirming the NGS result where appropriate (see Material and Methods). In the remaining 10 cases, a gene rearrangement was suggested by FISH. Nevertheless, although samples passed quality filters, the AMP-FPS assay failed to detect a fusion transcript. There are several possible explanations for this discrepancy including inadequate tumor cell fraction or low expression levels of the fusion transcript, chromosome rearrangements not yielding a fusion transcript, unusual breakpoints not covered by the assay or lack of primers covering the target gene. For instance, in two tumors (one endometrial stromal sarcoma and one sarcoma NOS) FISH indicated a rearrangement of the *BCOR* gene with an unknown partner. It is worth noting that the commercial AMP-FPS panel used in this study does not include primers for *BCOR*. Moreover, beside the common *CCNB3* partner (covered by the panel), *BCOR* has been reported to fuse with other genes which are also not targeted by the AMP-FPS assay (e.g., *ZC3H7B*, *MAML3*, *CIITA*) (20–23). Thus, in the absence of probes for

BCOR and potential partner genes, the failure of the assay in the 2 *BCOR* rearranged tumors of our series is not surprising. The same holds true for rearrangements involving *NR4A3* in extraskeletal myxoid chondrosarcomas: while the AMP-FPS assay covers the most *NR4A3* common partners (*EWSR1*, *TAF15*, *TCF12*, *TFG*) it lacks probes for both *NR4A3* and uncommon partners (24), thus scoring negative in the presence of alternative fusions.

The AMP-FPS assay failed to detect any fusion also in 3 cases of biphenotypic sinonasal sarcoma. Although in these cases no prior investigation (FISH or RT-PCR) was performed, this tumor is known to be typified by gene fusions involving the *PAX3* gene (25). Since the *PAX3* gene is not covered by the commercial AMP-FPS panel, we commissioned a customization of the assay by spiking-in primers to cover *PAX3* fusions. By using this customized AMP-FPS assay we were able to demonstrate and validate that all 3 cases expressed a *PAX3-MAML3* chimeric transcript (Figure 2).

Interestingly, a rare *EWSR1-PATZ1* fusion was detected by AMP-FPS in one *EWSR1* FISH-positive Ewing sarcoma (case #34). This fusion had been previously described in rare cases of spindle or small round cell sarcomas and it is considered to identify a distinct, Ewing-like entity (26). Moreover, the NGS profiling allowed the detection of disease-associated fusion transcripts also in a set of cases for which no prior molecular data was available or scored negative for FISH. These included one dermatofibrosarcoma protuberans (*COL1A1-PDGFB*), one endometrial stromal sarcoma (*YWHAE-NUTM2B*, aka *YWHAE-FAM22B*), one gastrointestinal neuroectodermal tumor (*EWSR1-CREB1*), one inflammatory myofibroblastic sarcoma (*TPM4-ALK*), one inflammatory myofibroblastic tumor (*TFG-ROS1*), 2 myoepitheliomas (one *FUS-NFATC2* and one *TRPS1-PLAG1*), 2 sclerosing epithelioid fibrosarcomas (one *EWSR1-CREB3L2* and one *FUS-CREB3L2*) and one solitary

TABLE 2 | Validation of the AMP-FPS fusion transcript assay.

Nr	Diagnosis	Pre-detected genetic abnormality	Primary detection method	Sequencing platform	Histotype-specific fusion detected	Other passing filters fusions
28	Askin Tumor	<i>EWSR1-ERG</i>	RT-qPCR	Illumina	<i>EWSR1-ERG</i>	<i>EWSR1-unl-ERG</i>
29	Congenital Fibrosarcoma	<i>ETV6-NTRK3</i>	RT-qPCR	Illumina	<i>ETV6-NTRK3</i>	NFD
30	Dermatofibrosarcoma Protuberans	<i>COL1A1-PDGFB</i>	FISH	Thermo	<i>COL1A1-PDGFB</i>	NFD
31	Dermatofibrosarcoma Protuberans	<i>COL1A1-PDGFB</i>	RT-qPCR	Illumina	<i>COL1A1-PDGFB</i>	NFD
32	Ewing Sarcoma	<i>EWSR1</i>	FISH	Thermo	<i>EWSR-FLI1</i>	NFD
33	Ewing Sarcoma	<i>EWSR1</i>	FISH	Thermo	<i>EWSR-FLI1</i>	NFD
34	Ewing Sarcoma	<i>EWSR1</i>	FISH	Thermo	<i>EWSR1-PATZ1</i>	NFD
35	Ewing Sarcoma	<i>EWSR1</i>	FISH	Thermo	<i>EWSR-FLI1</i>	NFD
36	Ewing Sarcoma	<i>EWSR1</i>	FISH	Thermo	<i>EWSR-FLI1</i>	NFD
37	Ewing Sarcoma	<i>EWSR1-FLI1</i>	RT-qPCR	Illumina	<i>EWSR1-FLI1</i>	<i>FXR2-CAMTA1</i>
38	Ewing Sarcoma	<i>EWSR1-FLI1</i>	RT-qPCR	Illumina	<i>EWSR1-FLI1</i>	NFD
39	Ewing Sarcoma	<i>EWSR1-FLI1</i>	RT-qPCR	Illumina	<i>EWSR1-FLI1</i>	NFD
40	Ewing Sarcoma	<i>EWSR1-ERG</i>	RT-qPCR	Illumina	<i>EWSR1-ERG</i>	<i>EWSR1-unl-EWSR1-ERG</i> ; <i>FUS-ERG</i> ; <i>EWSR1-ERG-EWSR1</i> ;
41	Ewing Sarcoma	<i>EWSR1-FLI1</i>	FISH	Illumina	<i>EWSR1-FLI1</i>	<i>EWSR1-FLI1-EWSR1</i>
42	Ewing Sarcoma	<i>EWSR1</i>	FISH	Thermo	<i>EWSR1-FLI1</i>	NFD
43	Ewing Sarcoma	<i>EWSR1-FLI1</i>	RT-qPCR	Thermo	<i>EWSR1-FLI1</i>	NFD
44	Ewing Sarcoma	<i>EWSR1-FLI1</i>	RT-qPCR	Thermo	<i>EWSR1-FLI1</i>	NFD
45	Ewing Sarcoma	<i>EWSR1-FLI1</i>	RT-qPCR	Thermo	<i>EWSR1-FLI1</i>	NFD
46	Ewing Sarcoma	<i>EWSR1-FLI1</i>	RT-qPCR	Thermo	<i>EWSR1-FLI1</i>	NFD
47	Ewing Sarcoma	<i>EWSR1-FLI1</i>	RT-qPCR	Thermo	<i>EWSR1-FLI1</i>	NFD
48	Ewing Sarcoma	<i>EWSR1-FLI1</i>	RT-qPCR	Thermo	<i>EWSR1-FLI1</i>	NFD
49	Ewing Sarcoma	<i>EWSR1-FLI1</i>	RT-qPCR	Thermo	<i>EWSR1-FLI1</i>	NFD
50	Ewing Sarcoma	<i>EWSR1-FLI1</i>	RT-qPCR	Illumina	<i>EWSR1-FLI1</i>	NFD
51	Ewing Sarcoma	<i>EWSR1</i>	FISH	Illumina	<i>EWSR1-FLI1</i>	NFD
52	Ewing Sarcoma	<i>FUS</i>	FISH	Thermo	<i>FUS-ERG</i>	NFD
53	Ewing-like Sarcoma	<i>BCOR-CCNB3</i>	RT-qPCR	Illumina	<i>BCOR-CCNB3</i>	NFD
54	Ewing-like Sarcoma	<i>CIC-DUX4</i>	RT-qPCR	Illumina	<i>CIC-DUX4</i>	NFD
55	Extraskeletal Myxoid Chondrosarcoma	<i>NR4A3</i>	FISH	Illumina	<i>EWSR1-NR4A3</i>	NFD
56	Extraskeletal Myxoid Chondrosarcoma	<i>EWSR1</i>	FISH	Illumina	<i>EWSR1-NR4A3</i>	NFD
57	Extraskeletal Myxoid Chondrosarcoma	<i>EWSR1-NR4A3</i>	RT-qPCR	Illumina	<i>EWSR1-NR4A3</i>	NFD
58	Extraskeletal Myxoid Chondrosarcoma	<i>TAF15-NR4A3</i>	RT-qPCR	Illumina	<i>TAF15-NR4A3</i>	NFD
59	Extraskeletal Myxoid Chondrosarcoma	<i>EWSR1-NR4A3</i>	RT-qPCR	Illumina	<i>EWSR1-NR4A3</i>	NFD
60	Extraskeletal Myxoid Chondrosarcoma	<i>EWSR1-NR4A3</i>	RT-qPCR	Illumina	<i>EWSR1-NR4A3</i>	NFD
61	Extraskeletal Myxoid Chondrosarcoma	<i>EWSR1-NR4A3</i>	RT-qPCR	Illumina	<i>EWSR1-NR4A3</i>	NFD
62	Extraskeletal Myxoid Chondrosarcoma	<i>EWSR1-NR4A3</i>	RT-qPCR	Illumina	<i>EWSR1-NR4A3</i>	NFD
63	Extraskeletal Myxoid Chondrosarcoma	<i>EWSR1-NR4A3</i>	RT-qPCR	Illumina	<i>EWSR1-NR4A3</i>	NFD
64	Extraskeletal Myxoid Chondrosarcoma	<i>NR4A3</i>	FISH	Illumina	<i>EWSR1-NR4A3</i>	NFD
65	Extraskeletal Myxoid Chondrosarcoma	<i>EWSR1-NR4A3</i>	RT-qPCR	Illumina	<i>EWSR1-NR4A3</i>	NFD

(Continued)

TABLE 2 | Continued

Nr	Diagnosis	Pre-detected genetic abnormality	Primary detection method	Sequencing platform	Histotype-specific fusion detected	Other passing filters fusions
66	Myoepithelial carcinoma (soft tissue)	<i>EWSR1</i>	FISH	Illumina	<i>EWSR1-ATF1</i>	NFD
67	Myoepithelioma (soft tissue)	<i>EWSR1</i>	FISH	Illumina	<i>EWSR1-ATF1</i>	NFD
68	Myxoid Liposarcoma	<i>FUS-DDIT3</i>	RT-PCR	Thermo	<i>FUS-DDIT3</i>	NFD
69	Myxoid Liposarcoma	<i>FUS-DDIT3</i>	RT-qPCR	Illumina	<i>FUS-DDIT3</i>	NFD
70	Myxoid Liposarcoma	<i>FUS-DDIT3</i>	FISH	Thermo	<i>FUS-DDIT3</i>	NFD
71	Myxoid Liposarcoma	<i>FUS-DDIT3</i>	FISH	Illumina	<i>FUS-DDIT3</i>	NFD
72	Myxoid Liposarcoma	<i>FUS-DDIT3</i>	FISH	Illumina	<i>FUS-DDIT3</i>	NFD
73	Nodular Fasciitis	<i>USP6</i>	FISH	Thermo	<i>MYH9-USP6</i>	NFD
74	Rhabdomyosarcoma, alveolar	<i>PAX3-FOXO1</i>	RT-PCR	Thermo	<i>PAX3-FOXO1</i>	NFD
75	Rhabdomyosarcoma, alveolar	<i>PAX3-FOXO1</i>	RT-PCR	Thermo	<i>PAX3-FOXO1</i>	NFD
76	Rhabdomyosarcoma, alveolar	<i>PAX3-FOXO1</i>	RT-PCR	Thermo	<i>PAX3-FOXO1</i>	NFD
77	Rhabdomyosarcoma, alveolar	<i>PAX3-FOXO1</i>	RT-qPCR	Illumina	<i>PAX3-FOXO1</i>	NFD
78	Rhabdomyosarcoma, alveolar	<i>PAX3-FOXO1</i>	RT-qPCR	Illumina	<i>PAX3-FOXO1</i>	NFD
79	Rhabdomyosarcoma, alveolar	<i>PAX3-FOXO1</i>	RT-qPCR	Illumina	<i>PAX3-FOXO1</i>	NFD
80	Rhabdomyosarcoma, alveolar	<i>PAX3-FOXO1</i>	RT-qPCR	Illumina	<i>PAX3-FOXO1</i>	NFD
81	Rhabdomyosarcoma, alveolar	<i>PAX3-FOXO1</i>	RT-qPCR	Illumina	<i>PAX3 - FOXO1</i>	<i>FOXO1-PAX3</i>
82	Rhabdomyosarcoma, alveolar	<i>PAX3-FOXO1</i>	RT-qPCR	Illumina	<i>PAX3-FOXO1</i>	NFD
83	Rhabdomyosarcoma, spindle cell	<i>SRF-NCOA2</i>	RT-qPCR	Illumina	<i>SRF- NCOA2</i>	NFD
84	Sarcoma NOS	<i>EWSR1</i>	FISH	Illumina	<i>EWSR1-FLI1</i>	NFD
85	Solitary Fibrous Tumor	<i>STAT6</i>	IHC	Thermo	<i>NAB2-STAT6</i>	NFD
86	Synovial Sarcoma	<i>SS18-SSX2</i>	RT-qPCR	Illumina	<i>SS18-SSX2</i>	<i>SS18-SSX4;SS18-SSX1; complex SS18-SSX2 fusions</i>
87	Synovial Sarcoma	<i>SS18</i>	FISH	Illumina	<i>SS18-SSX1</i>	<i>SS18-SSX4; SS18-SSX4-SS18</i>
88	Synovial Sarcoma	<i>SS18</i>	FISH	Thermo	<i>SS18-SSX1</i>	NFD
89	Synovial Sarcoma	<i>SS18-SSX1</i>	RT-qPCR	Illumina	<i>SS18-SSX1</i>	NFD
90	Synovial Sarcoma	<i>SS18-SSX1</i>	RT-qPCR	Thermo	<i>SS18-SSX1</i>	NFD
91	Synovial Sarcoma	<i>SS18-SSX1</i>	RT-qPCR	Thermo	<i>SS18-SSX1</i>	<i>SS18-SSX2</i>
92	Synovial Sarcoma	<i>SS18-SSX1</i>	RT-qPCR	Thermo	<i>SS18-SSX1</i>	<i>SS18-SSX4</i>
93	Synovial Sarcoma	<i>SS18-SSX1</i>	RT-qPCR	Thermo	<i>SS18-SSX1</i>	<i>SS18-SSX4</i>
94	Synovial Sarcoma	<i>SS18</i>	FISH	Illumina	<i>SS18-SSX1</i>	<i>SS18-SSX4-SS18</i>
95	Synovial Sarcoma	<i>SS18-SSX2</i>	RT-qPCR	Illumina	<i>SS18-SSX2</i>	NFD
96	Synovial Sarcoma	<i>SS18</i>	FISH	Illumina	<i>SS18-SSX1</i>	<i>SS18-SSX4</i>
97	Synovial Sarcoma	<i>SS18-SSX1</i>	RT-qPCR	Thermo	<i>SS18-SSX1</i>	<i>SS18-SSX4</i>
98	Clear Cell Sarcoma	<i>EWSR1</i>	FISH	Thermo	<i>EWSR1-CREB1</i>	NFD
99	Endometrial Stromal Sarcoma	<i>BCOR</i>	FISH	Thermo	NFD	NFD
100	Extraskeletal Myxoid Chondrosarcoma	<i>NR4A3</i>	FISH	Illumina	NFD	NFD
101	Myoepithelioma (soft tissue)	<i>EWSR1</i>	FISH	Illumina	NFD	NFD
102	Myxoid Fibrosarcoma	<i>FUS</i>	FISH	Illumina	NFD	NFD

(Continued)

TABLE 2 | Continued

Nr	Diagnosis	Pre-detected genetic abnormality	Primary detection method	Sequencing platform	Histotype-specific fusion detected	Other passing filters fusions
103	Myxoid Liposarcoma	<i>DDIT3</i>	FISH	Illumina	NFD	NFD
104	Nodular Fasciitis	<i>USP6</i>	FISH	Thermo	NFD	NFD
105	Rhabdomyosarcoma, alveolar	<i>FOXO1</i>	FISH	Thermo	NFD	NFD
106	Sarcoma NOS	<i>BCOR</i>	FISH	Thermo	NFD	NFD
107	Solitary Fibrous Tumor	<i>EWSR1</i>	FISH	Illumina	NFD	NFD
108	Undifferentiated round cell, Ewing-Like Sarcoma	<i>CIC</i>	FISH	Illumina	NFD	NFD
109	Lipoblastoma	<i>PLAG1 neg</i>	FISH	Illumina	NFD	NFD
110	Myxoid Fibrosarcoma	<i>EWSR1, FUS neg</i>	FISH	Thermo	NFD	NFD
111	Myxoid Fibrosarcoma	<i>EWSR1, FUS neg</i>	FISH	Thermo	NFD	NFD
112	Myxoid Fibrosarcoma	12q13-15 amp	FISH	Thermo	NFD	NFD
113	Rhabdomyosarcoma, alveolar	<i>FOXO1 neg</i>	FISH	Thermo	NFD	NFD
114	Rhabdomyosarcoma, embryonal	<i>FOXO1 neg</i>	FISH	Illumina	NFD	NFD
115	Rhabdomyosarcoma, embryonal	<i>FOXO1 neg</i>	FISH	Illumina	NFD	NFD
116	Rhabdomyosarcoma, embryonal	<i>FOXO1 neg</i>	FISH	Illumina	NFD	NFD
117	Sarcoma NOS	<i>EWSR1 neg</i>	FISH	Illumina	<i>CIC-DUX4</i>	NFD
118	Small Round Cell Tumor	<i>EWSR1, BCOR, FUS, CIC neg</i>	FISH	Thermo	NFD	NFD
119	Undifferentiated Sarcoma	<i>EWSR1 neg</i>	FISH	Illumina	<i>CIC-DUX4</i>	NFD
120	Undifferentiated Sarcoma	12q13-15 amp	FISH	Thermo	NFD	NFD
121	Undifferentiated Sarcoma	12q13-15 amp	FISH	Thermo	NFD	<i>HMGA2-LGR5</i>
122	Biphenotypic Sinonasal Sarcoma	nd	nd	Thermo	<i>PAX3-MAML3</i> [§]	NFD
123	Biphenotypic Sinonasal Sarcoma	nd	nd	Thermo	<i>PAX3-MAML3</i> [§]	NFD
124	Biphenotypic Sinonasal Sarcoma	nd	nd	Thermo	<i>PAX3-MAML3</i> [§]	NFD
125	Dermatofibrosarcoma Protuberans	nd	nd	Thermo	<i>COL1A1-PDGFB</i>	NFD
126	Endometrial Stromal Sarcoma	nd	nd	Thermo	<i>YWHAE-NUTM2B</i>	NFD
127	Gastrointestinal Neuroectodermal Tumor	nd	nd	Thermo	<i>EWSR1-CREB1</i>	<i>SS18-PTRF</i>
128	Inflammatory Myofibroblastic Sarcoma	nd	nd	Illumina	<i>TPM4-ALK</i>	NFD
129	Inflammatory Myofibroblastic Tumor	nd	nd	Thermo	<i>TFG-ROS1</i>	NFD
130	Myoepithelioma (bone)	nd	nd	Illumina	<i>FUS-NFATC2</i>	NFD
131	Myoepithelioma (soft tissue)	nd	nd	Illumina	<i>TRPS1-PLAG1</i>	NFD
132	Sclerosing Epithelioid Fibrosarcoma	nd	nd	Illumina	<i>EWSR1-CREB3L2</i>	NFD
133	Sclerosing epithelioid fibrosarcoma (soft tissue)	nd	nd	Illumina	<i>FUS-CREB3L2</i>	NFD
134	Solitary Fibrous Tumor	nd	nd	Thermo	<i>NAB2-STAT6</i>	NFD
135	Chondrosarcoma	nd	nd	Thermo	NFD	NFD
136	Endometrial Stromal Sarcoma	nd	nd	Thermo	NFD	NFD
137	Epithelioid Angiosarcoma	nd	nd	Illumina	NFD	NFD

(Continued)

TABLE 2 | Continued

Nr	Diagnosis	Pre-detected genetic abnormality	Primary detection method	Sequencing platform	Histotype-specific fusion detected	Other passing filters fusions
138	Follicular Dendritic Cell Sarcoma	nd	nd	Thermo	NFD	NFD
139	Leiomyosarcoma	nd	nd	Illumina	NFD	NFD
140	Leiomyosarcoma	nd	nd	Thermo	NFD	NFD
141	Myoepithelioma (bone)	nd	nd	Illumina	NFD	NFD
142	Myxoid Fibrosarcoma	nd	nd	Thermo	NFD	NFD
143	Myxoinflammatory Fibroblastic Sarcoma	nd	nd	Illumina	NFD	NFD
144	Osteosarcoma	nd	nd	Illumina	NFD	NFD
145	Osteosarcoma	nd	nd	Illumina	NFD	NFD
146	Pleomorphic Sarcoma	nd	nd	Thermo	NFD	NFD
147	Pleomorphic Sarcoma	nd	nd	Thermo	NFD	NFD
148	Pleomorphic Sarcoma	nd	nd	Thermo	NFD	NFD
149	Sarcoma NOS HG Myxoid	nd	FISH	Thermo	NFD	NFD
150	Undifferentiated Sarcoma	nd	nd	Illumina	NFD	NFD

NFD, no histotype-specific fusion detected; nd, not done; amp, amplification; neg, negative; RT-PCR, reverse transcriptase-PCR; FISH, fluorescent in situ hybridization; RT-qPCR, reverse transcriptase-quantitative PCR; IHC, immunohistochemistry; un1, unaligned sequence. PAX3-MAML3\$: fusion detected with a PAX3-customized AMP-FPS Panel. This sample scored negative with the standard AMP-FPS Panel.

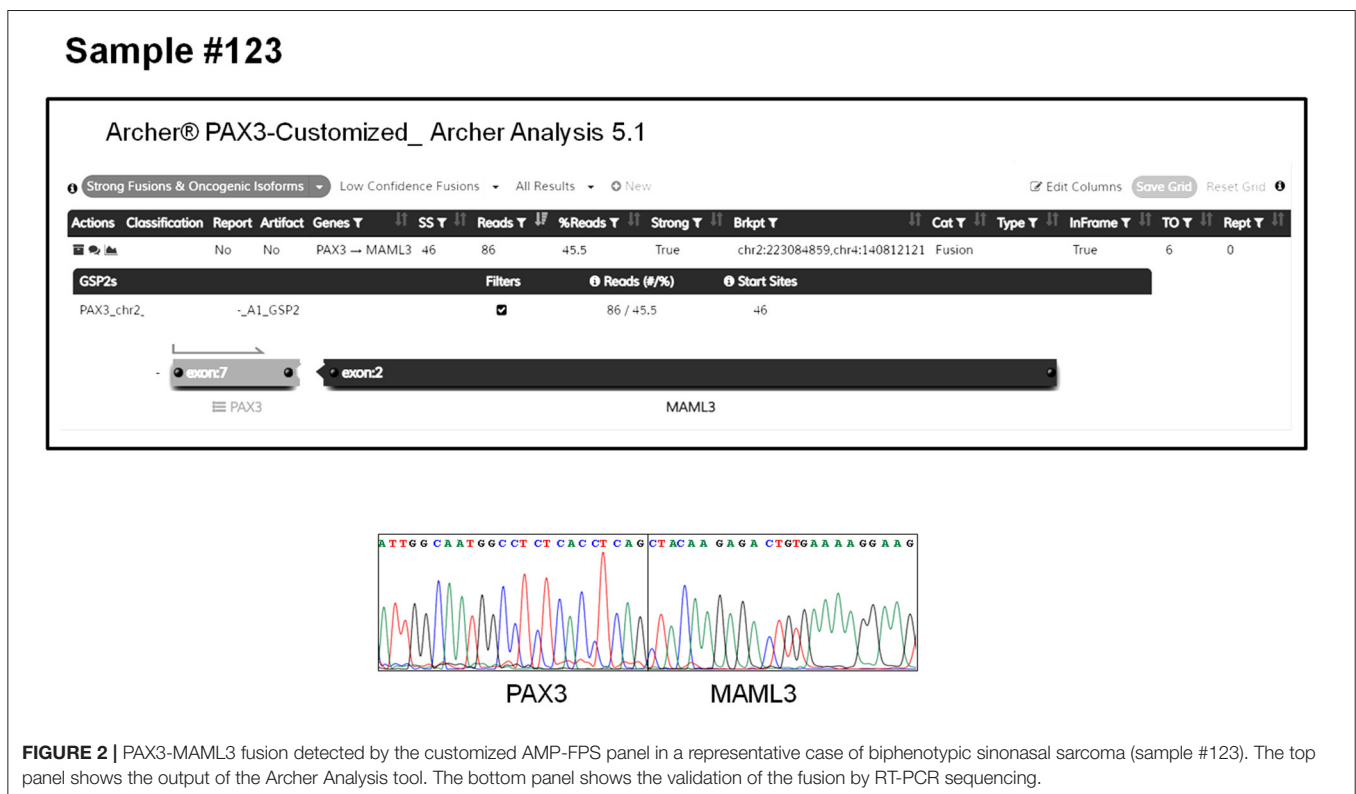


FIGURE 2 | PAX3-MAML3 fusion detected by the customized AMP-FPS panel in a representative case of biphenotypic sinonasal sarcoma (sample #123). The top panel shows the output of the Archer Analysis tool. The bottom panel shows the validation of the fusion by RT-PCR sequencing.

fibrous tumor (*NAB2-STAT6*). In addition, 2/5 tumors negative for *EWSR1* rearrangements according to FISH, turned out to express a *CIC-DUX4* fusion, leading to the diagnosis of *CIC-DUX4* fusion-positive undifferentiated round cell sarcoma (27). In all these cases the identified fusions were confirmed by RT-PCR.

Finally, the series analyzed included also sarcoma variants typically devoid of pathognomonic fusions (e.g., leiomyosarcoma, osteosarcoma). Thus, the negative result of the NGS profiling in these cases may be considered compatible with the pathological diagnosis.

DISCUSSION

The expression of fusion transcripts characterizes over a third of sarcomas where it may provide diagnostic, prognostic and predictive information. The cooperative effort described in this work was aimed at assessing feasibility, reliability, and applicability of NGS-based approaches for the detection of pathognomonic fusion transcripts in a routine diagnostic setting.

In line with recent reports (12, 19), our study corroborates the robustness of NGS, and in particular of AMP-FPS profiling, for the detection of clinically relevant fusions in sarcomas. On one hand, our analysis emphasizes the worth of implementing this type of approach in routine diagnostics. On the other hand, it underlines the importance of being aware of the actual detection capability of the panel used (genes covered by the assay) in relation to the specific tumor variant under investigation.

Our study demonstrates also the versatility of certain NGS fusion commercial panels to respond to specific diagnostic needs. In fact, the possibility of further implementing commercially available panels by spiking-in probes for genetic targets not included in the standard version of the assay allows to expand its detection capability. Indeed, beside *PAX3*, due to the recent therapeutic successes of NTRK fusions targeting drugs in solid tumors (7, 8), we are in the process of customizing the AMP-FPS panel by including primers for *NTRK1* and *NTRK2* (currently only *NTRK3* is covered by the AMP-FPS assay).

Importantly, in the presence of a negative result, a re-evaluation of RNA and library quality is mandatory as highly degraded RNA and poor quality libraries may affect the sensitivity of the assay. Nonetheless, we found that apparently low quality samples may still be effective for fusion detection. Indeed, a few cases included in this study (cases #9, 31, 37, 47, 57, 60, 80, 126), although not fulfilling all quality criteria, nevertheless yielded a correct fusion call. This indicates that this type of assay may work even in suboptimal conditions.

Finally, when reporting the result of this type of NGS analysis, especially if negative, a statement specifying the characteristics and the limits of the assay employed (type of NGS panel, number of target genes, website of the provider for the list of targeted fusions) and the actual performance of the test according to the manufacturer's standards (fulfillment of quality parameters) should always be included in the pathology report. It is worth reaffirming that the AMP-FPS assay is designed to target the most common breakpoint regions of the genes covered by the assay. Thus, unusual breakpoints may be source of "false negative" results. Moreover, when dealing with sarcoma variants expressing uncommon fusions, the presence of primers for the target genes should be verified prior to setting up the profiling because the lack of appropriate primers will yield a false negative result. The negativity in the AMP-FPS assay of the two *BCOR* rearranged tumors, included in this series, is instructive in this regard.

In the case of a positive result, beside the genes involved in the fusion, the inclusion in the pathology report of details about the fusion variant detected, including reading frame of the chimeric transcript (in frame/out of frame) and exons involved might be useful. This is of particular importance if the fusion protein is potentially actionable and the retention of specific domains in the chimeric protein is crucial for drug sensitivity, as in the case of NTRK fusions (7–9).

DATA AVAILABILITY STATEMENT

Sequencing data files are available in the NCBI-SRA (<http://www.ncbi.nlm.nih.gov/sra>) database under the accession number PRJNA608250.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethic committee Istituto Ortopedico Rizzoli IRCCS, Regina Elena National Cancer Institute IRCCS, Bambino Gesù Children's Hospital IRCCS and by the proper institutional review boards of the CRO Aviano IRCCS National Cancer Institute, Veneto Institute of Oncology (IOV) IRCCS, University of Padua, Candiolo Cancer Institute FPO-IRCCS, Istituto Scientifico Romagnolo per lo Studio e la Cura dei Tumori (IRST) Meldola IRCCS, Istituto Nazionale dei Tumori di Milano Fondazione IRCCS. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

RM conceived the work on the behalf of the ACC sarcoma working group. All authors contributed to the generation of molecular profiling data. Each center involved in panel sequencing was responsible for generation, analyses and sharing of data. RF and RM coordinated the collection and integration of data. DR, MB, DB, FG, and BC were in charge of panel comparison. DR, MB, and DB were in charge of second-level bioinformatic analyses. RM and RF wrote the first draft of the manuscript with the support of DR and MB. All authors revised and approved the final version of the manuscript.

FUNDING

This work was supported by the Ministry of Health and Alleanza Contro il Cancro (ACC).

ACKNOWLEDGMENTS

For their suggestions and support, the authors are grateful to: Valentina Laquintana (Regina Elena National Cancer Institute, Rome); Sara Piccinin, Daniela Gasparotto, Kelly Fassetta, Beatrice Valenti (Centro di Riferimento Oncologico, CRO Aviano); Franco Locatelli, Simona Caruso, Ida Russo, Rita Alaggio, Rita De Vito, Emanuele Agolini, Martina Rinelli (Bambino Gesù Children's Hospital, IRCCS, Rome); Carolina Zamuner (Veneto Institute of Oncology, Padua, Italy); Massimo Serra, Laura Pazzaglia, Marco Gambarotti, Stefania Benini, Alberto Righi (Istituto Ortopedico Rizzoli, Bologna); Federica Pieri, Michela Tebaldi, Elisa Chiadini (Istituto Scientifico Romagnolo per lo Studio e la Cura dei Tumori, Meldola). Special thanks for their work go to secretaries, preclinical, and clinical coordinators of the ACC sarcoma working group, the Italian Sarcoma Group (ISG), the Rizzoli and the CRO Aviano Institutes.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fonc.2020.00489/full#supplementary-material>

Supplemental Table 1 | Fusion transcripts called by the Arriba algorithm.

Supplemental Table 2 | Fusion transcripts called by the Pizzly algorithm.

Supplemental Table 3 | Fusion transcripts called by the STAR-Fusion algorithm.

REFERENCES

- Fletcher CDM, Bridge JA, Hogendoorn PCW, Mertens F. *WHO Classification of Tumors of Soft Tissue and Bone*. International Agency for Research on Cancer (2013).
- Schaefer I-M, Cote GM, Hornick JL. Contemporary sarcoma diagnosis, genetics, and genomics. *J Clin Oncol*. (2018) 36:101–10. doi: 10.1200/JCO.2017.74.9374
- Sbaraglia M, Dei Tos AP. The pathology of soft tissue sarcomas. *Radiol Med*. (2019) 124:266–81. doi: 10.1007/s11547-018-0882-7
- Mertens F, Antonescu CR, Mitelman F. Gene fusions in soft tissue tumors: recurrent and overlapping pathogenetic themes. *Genes Chromosomes Cancer*. (2016) 55:291–310. doi: 10.1002/gcc.22335
- Uğurel S, Mentzel T, Utikal J, Helmbold P, Mohr P, Pföhler C, et al. Neoadjuvant imatinib in advanced primary or locally recurrent dermatofibrosarcoma protuberans: a multicenter phase II DeCOG trial with long-term follow-up. *Clin Cancer Res*. (2014) 20:499–510. doi: 10.1158/1078-0432.CCR-13-1411
- Schöffski P, Sufliarsky J, Gelderblom H, Blay J-Y, Strauss SJ, Stacchiotti S, et al. Crizotinib in patients with advanced, inoperable inflammatory myofibroblastic tumours with and without anaplastic lymphoma kinase gene alterations (European Organisation for Research and Treatment of Cancer 90101 CREATE): a multicentre, single-drug, prospective, non-randomised phase 2 trial. *Lancet Respir Med*. (2018) 6:431–41. doi: 10.1016/S2213-2600(18)30116-4
- Laetsch TW, DuBois SG, Mascarenhas L, Turpin B, Federman N, Albert CM, et al. Larotrectinib for paediatric solid tumours harbouring NTRK gene fusions: phase 1 results from a multicentre, open-label, phase 1/2 study. *Lancet Oncol*. (2018) 19:705–14. doi: 10.1016/S1470-2045(18)30119-0
- Doebele RC, Drlon A, Paz-Ares L, Siena S, Shaw AT, Farago AF, et al. Entrectinib in patients with advanced or metastatic NTRK fusion-positive solid tumours: integrated analysis of three phase 1–2 trials. *Lancet Oncol*. (2020) 21:271–82. doi: 10.1016/S1470-2045(19)30691-6
- Solomon JP, Linkov I, Rosado A, Mullaney K, Rosen EY, Frosina D, et al. NTRK fusion detection across multiple assays and 33,997 cases: diagnostic implications and pitfalls. *Mod Pathol*. (2020) 33:38–46. doi: 10.1038/s41379-019-0324-7
- Brenca M, Maestro R. Massive parallel sequencing in sarcoma pathobiology: state of the art and perspectives. *Expert Rev Anticancer Ther*. (2015) 15:1473–88. doi: 10.1586/14737140.2015.1108192
- Xiao X, Garbutt CC, Hornicek F, Guo Z, Duan Z. Advances in chromosomal translocations and fusion genes in sarcomas and potential therapeutic applications. *Cancer Treat Rev*. (2018) 63:61–70. doi: 10.1016/j.ctrv.2017.12.001
- Pei J, Zhao X, Patchefsky AS, Flieder DB, Talarchek JN, Testa JR, et al. Clinical application of RNA sequencing in sarcoma diagnosis: an institutional experience. *Medicine*. (2019) 98:e16031. doi: 10.1097/MD.00000000000016031
- Dobin A, Gingeras TR. Optimizing RNA-Seq Mapping with STAR. *Methods Mol Biol*. (2016) 1415:245–62. doi: 10.1007/978-1-4939-3572-7_13
- Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*. (2016) 32:1220–2. doi: 10.1093/bioinformatics/btv710
- Kim B, Lee H, Shin S, Lee S-T, Choi JR. Clinical evaluation of massively parallel RNA sequencing for detecting recurrent gene fusions in hematologic malignancies. *J Mol Diagn*. (2019) 21:163–70. doi: 10.1016/j.jmoldx.2018.09.002
- Uhrig S. *Arriba - Fast and Accurate Gene Fusion Detection from RNA-Seq Data*. (2019). Available online at: <https://github.com/suhrig/arriba>
- Haas BJ, Dobin A, Li B, Stransky N, Pochet N, Regev A. Accuracy assessment of fusion transcript detection via read-mapping and *de novo* fusion transcript assembly-based methods. *Genome Biol*. (2019) 20:213. doi: 10.1186/s13059-019-1842-9
- Melsted P, Hateley S, Joseph IC, Pimentel H, Bray N, Pachter L. Fusion detection and quantification by pseudoalignment. *bioRxiv*. 166322. (2017). doi: 10.1101/166322
- Lam SW, Cleton-Jansen A-M, Cleven AHG, Ruano D, van Wezel T, Szuhai K, et al. Molecular analysis of gene fusions in bone and soft tissue tumors by anchored multiplex PCR-based targeted next-generation sequencing. *J Mol Diagn*. (2018) 20:653–63. doi: 10.1016/j.jmoldx.2018.05.007
- Pierron G, Tirode F, Lucchesi C, Reynaud S, Ballet S, Cohen-Gogo S, et al. A new subtype of bone sarcoma defined by BCOR-CCNB3 gene fusion. *Nat Genet*. (2012) 44:461–6. doi: 10.1038/ng.1107
- Panagopoulos I, Thorsen J, Gorunova L, Haugom L, Bjerkehagen B, Davidson B, et al. Fusion of the ZC3H7B and BCOR genes in endometrial stromal sarcomas carrying an X;22-translocation. *Genes Chromosomes Cancer*. (2013) 52:610–8. doi: 10.1002/gcc.22057
- Specht K, Zhang L, Sung Y-S, Nucci M, Dry S, Vaiyapuri S, et al. Novel BCOR-MAML3 and ZC3H7B-BCOR Gene Fusions in Undifferentiated Small Blue Round Cell Sarcomas. *Am J Surg Pathol*. (2016) 40:433–42. doi: 10.1097/PAS.0000000000000591
- Yoshida A, Arai Y, Hama N, Chikuta H, Bando Y, Nakano S, et al. Expanding the clinicopathologic and molecular spectrum of BCOR-associated sarcomas in adults. *Histopathology*. (2020) 76:509–20. doi: 10.1111/his.14023
- Urbini M, Astolfi A, Pantaleo MA, Serravalle S, Dei Tos AP, Picci P, et al. HSPA8 as a novel fusion partner of NR4A3 in extraskelatal myxoid chondrosarcoma. *Genes Chromosomes Cancer*. (2017) 56:582–6. doi: 10.1002/gcc.22462
- Carter CS, East EG, McHugh JB. Biphenotypic sinonasal sarcoma: a review and update. *Arch Pathol Laboratory Med*. (2018) 142:1196–201. doi: 10.5858/arpa.2018-0207-RA
- Bridge JA, Sumegi J, Druta M, Bui MM, Henderson-Jackson E, Linos K, et al. Clinical, pathological, and genomic features of EWSR1-PATZ1 fusion sarcoma. *Mod Pathol*. (2019) 32:1593–604. doi: 10.1038/s41379-019-0301-1
- Miettinen M, Felisiak-Golabek A, Luiña Contreras A, Glod J, Kaplan RN, Killian JK, et al. New fusion sarcomas: histopathology and clinical significance of selected entities. *Hum Pathol*. (2019) 86:57–65. doi: 10.1016/j.humpath.2018.12.006

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling editor declared a past co-authorship with two of the authors ET and ABu.

Copyright © 2020 Racanelli, Brenca, Baldazzi, Goeman, Casini, De Angelis, Guercio, Milano, Tamborini, Busico, Dagrada, Garofalo, Caruso, Brunello, Pignochino, Berrino, Grignani, Scotlandi, Parra, Hattinger, Ibrahim, Mercatali, De Vita, Carriero, Pallocca, Loria, Covello, Sbaraglia, Dei Tos, Falcioni and Maestro. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Article

Highlighting Human Enzymes Active in Different Metabolic Pathways and Diseases: The Case Study of EC 1.2.3.1 and EC 2.3.1.9

Giulia Babbi , Davide Baldazzi, Castrense Savojardo , Martelli Pier Luigi and Rita Casadio *

Biocomputing Group, University of Bologna, 40126 Bologna, Italy; giulia.babbi3@unibo.it (G.B.);
davide.baldazzi8@unibo.it (D.B.); castrense.savojardo2@unibo.it (C.S.); pierluigi.martelli@unibo.it (M.P.L.)

* Correspondence: rita.casadio@unibo.it

Received: 5 June 2020; Accepted: 24 July 2020; Published: 29 July 2020



Abstract: Enzymes are key proteins performing the basic functional activities in cells. In humans, enzymes can be also responsible for diseases, and the molecular mechanisms underlying the genotype to phenotype relationship are under investigation for diagnosis and medical care. Here, we focus on highlighting enzymes that are active in different metabolic pathways and become relevant hubs in protein interaction networks. We perform a statistics to derive our present knowledge on human metabolic pathways (the Kyoto Encyclopaedia of Genes and Genomes (KEGG)), and we found that activity aldehyde dehydrogenase (NAD(+)), described by Enzyme Commission number EC 1.2.1.3, and activity acetyl-CoA C-acetyltransferase (EC 2.3.1.9) are the ones most frequently involved. By associating functional activities (EC numbers) to enzyme proteins, we found the proteins most frequently involved in metabolic pathways. With our analysis, we found that these proteins are endowed with the highest numbers of interaction partners when compared to all the enzymes in the pathways and with the highest numbers of predicted interaction sites. As specific enzyme protein test cases, we focus on Alpha-Aminoacidic Semialdehyde Dehydrogenase (ALDH7A1, EC 2.3.1.9) and Acetyl-CoA acetyltransferase, cytosolic and mitochondrial (gene products of ACAT2 and ACAT1, respectively; EC 2.3.1.9). With computational approaches we show that it is possible, by starting from the enzyme structure, to highlight clues of their multiple roles in different pathways and of putative mechanisms promoting the association of genes to disease.

Keywords: enzymes; KEGG pathways KEGG metabolic pathways; protein-protein interaction; protein variation; protein stability

1. Introduction

It is common knowledge that enzymes are proteins characterized by specific molecular functions that, when performed in a concerted manner, give rise to the richness of biological processes at the basis of the cell complex physiology [1]. It is still a matter of debate whether different enzyme molecules tend to transiently aggregate in the cell environment, for generating the proper concerted action [2], and references therein. In the case of enzymes, any concerted biological process is modelled by a metabolic network/pathway that describes the biochemical sequential interactions and/or cycles at the basis of the cell metabolism [3]. Information on which models of metabolic pathways and reactions are known in a specific organism is also available through curated databases, such as the Kyoto Encyclopaedia of Genes and Genomes (KEGG) and REACTOME [4,5]. Each enzyme is a protein molecule endowed with a specific four-digit EC number [6], which fully describes the catalyzed biochemical reaction, and possibly with an atomic solved structure, routinely available in the Protein Data Bank (PDB), [7]. This allows for an understanding of the relationship between

sequence, structure, and function at the basis of the catalytic mechanisms at the active site/s and the role of possible effectors at the binding site/s. UniProt/SwissProt [8] is the reference database for sequences, and PDB for three dimensional (3D) structures. Many enzymes are known to be involved in genetic diseases, as reported in OMIM (Online Mendelian Inheritance in Man) [9], as well as somatic diseases, including cancers (BioMuta [10], DisGenNet [11], Clinvar [12], MalaCards [13], etc.). This makes it possible to derive information on specific molecular mechanisms when non-synonymous mutations have been associated to specific pathologies. Thanks to massive proteomic experiments, we also know partners of interactions in the cell milieu stored in databases such as IntAct [14] and BioGRID [15]. Several databases are presently available for enzyme complete functional annotation, including BRENDA [16], Enzyme Portal (EBI) [17], and M-CSA (EBI) [18]. Furthermore, among other information, available data on the extent of expression of the enzymes in the different human tissues can be found in GeneCards [19].

The more data accumulated, the more we need linking different databases in order to derive general rules of molecular functioning, which reconcile molecular mechanisms to physiological models related to specific phenotypes. A recently released version of Manet (Molecular Ancestry Network, Manet 3.0, [20,21]) groups enzymatic activities into a hierarchical system of subnetworks and mesonetworks matching KEGG classification and including structural data.

Focusing on humans, here, we ask the question of how many human enzymes are common to different metabolic pathways. The aim is highlighting the complex networks of networks where some of the proteins are involved simultaneously in different biological processes and providing evidence of possible associations to protein-protein interaction data and molecular clues.

By referring to the human KEGG metabolic maps, we provide a list of these enzymes, and their relation to maladies, when known. We find an interesting correspondence among most frequent enzymes in KEGG metabolic maps, number of interactors in the cell environment and number of predicted interaction sites.

We then investigate, at a molecular level, one of these enzymes, ALDH7A1, a member of subfamily 7 in the aldehyde dehydrogenase gene family (EC 1.2.1.3). The enzymes are described to play a major role in the detoxification of aldehydes generated by alcohol metabolism and lipid peroxidation. The protein, characterized by at least three different isoforms, is present in the cytosol, the mitochondrion, and the nucleus, and it is associated with different biological functions. By means of computational tools, we investigate which structural properties of the enzyme can be indicative of its important role and highlight possible mechanisms of its failure, associated mainly with pyridoxine-dependent epilepsy (PDE). Similarly, we describe molecular experimental and predicted details of ACAT1 and ACAT2, performing in humans Acetyl-CoA C-acetyltransferases activity, respectively in the cytosol and in mitochondria (EC 2. 3.1.9).

2. Experimental Section

2.1. Materials

For our analysis, we derived information from SwissProt/UniProt. Presently, SwissProt (release 04_2019) lists 20,365 human proteins, among which 3428 are enzymes specified with a complete enzyme commission number (EC with four digits, describing the biochemical reaction as to substrate and product) [6]. In the following, we will refer to enzyme proteins as EC proteins. We associate 7316 proteins with genetic diseases through our database eDGAR (adopting OMIM, HUMSAVAR, CLINVAR, and curated DisGeNet as primary sources of information) [22,23]. We find that 1699 proteins are EC proteins with associations to disease (Table 1).

For KEGG pathway annotation, we adopted the April 2020 KEGG release [4], with the distinction among KEGG pathways and KEGG metabolic pathways and with reference to human genes. Protein-protein interactions are retrieved from IntAct ([14], release June 2020) and BioGRID ([15], release 3.5.185).

Table 1. Disease-related human proteins with enzyme commission (EC) number.

Set	# Human Proteins
In SwissProt/UniProt	20365
Proteins with four-digit EC (EC proteins)	3428 (1411 EC) *
Proteins associated with genetic diseases	7316 (5788 diseases)
EC proteins with genetic disease associations	1669 (955 EC and 1900 diseases)

* Number of four-digit EC numbers.

2.2. Computational Methods

The likelihood of a protein lateral side chain to interact with other proteins is computed with ISPRED4 (Interaction Site PREDictions, version 4) [24,25], a machine-learning-based predictor performing at the state of the art. It predicts the interaction sites from protein structure with an accuracy as high as 85% and with a very low rate of false positive prediction (3%). When a structure is not available, an in-house version of ISPRED4 considering only sequence information is adopted. For computing the effect on protein stability of missense variations, we adopted INPS (Impact of Non synonymous variations on Protein Stability) [26,27]. Starting from information extracted from protein structure or sequence, INPS performs a non-linear regression based on machine learning approaches and reaches a Pearson's correlation coefficient as high as 0.76 (0.71 when a structure is not available). The computed $\Delta\Delta G$ values have an associated standard error of about 1 kcal/mol.

3. Results

3.1. EC Proteins and KEGG Metabolic Pathways

In order to cope with the complexity of the network of human biochemical reactions, we focused on the analysis of all the possible relationships among biological functions as described by EC numbers and KEGG pathways. The Kyoto Encyclopaedia of Genes and Genomes (KEGG), [4], includes 320 biological pathways, 90 of which are specifically termed metabolic pathways. We annotated EC human proteins with KEGG terms for pathways (Table 2). Having as a reference the human protein section of SwissProt, we find that 6904 proteins are associated with 320 KEGG pathways. When focusing on proteins associated with metabolic KEGG pathways, 1642 EC proteins participate into 90 metabolic pathways. Restricting to proteins that are enzymes and disease-related, we obtained 770 EC proteins associated with 90 metabolic pathways. The 770 proteins are associated with 930 EC numbers.

Table 2. EC human protein in Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathways.

Set	Human Proteins with KEGG Pathways		Human Proteins with KEGG Metabolic Pathways	
	#Proteins	#Pathways	#Proteins	#Pathways
In SwissProt	6904	320	1642	90
EC proteins	2258	317	1375	90
Proteins associated to genetic diseases	3391	320	895	90
EC proteins associated to genetic diseases	1255	314	770	90

Number of.

Not all the EC proteins in SwissProt are associated with KEGG metabolic pathways (883 from Table 2). The whole network model is therefore complicated [20,21], and here, we focus only on KEGG networks that describe metabolic pathways.

In Figure 1, we show the distribution of EC numbers (which we consider here the complete description of the protein molecular activity) in the KEGG metabolic pathways. We find that five EC numbers are involved in at least 11 metabolic pathways—EC 1.2.1.3, Aldehyde dehydrogenase (NAD+); EC 1.14.14.1, Unspecific monooxygenase; EC 2.3.1.9, Acetyl-CoA C-acetyltransferase; EC 2.6.1.1, Aspartate transaminase; EC 4.2.1.17, Enoyl-CoA hydratase.

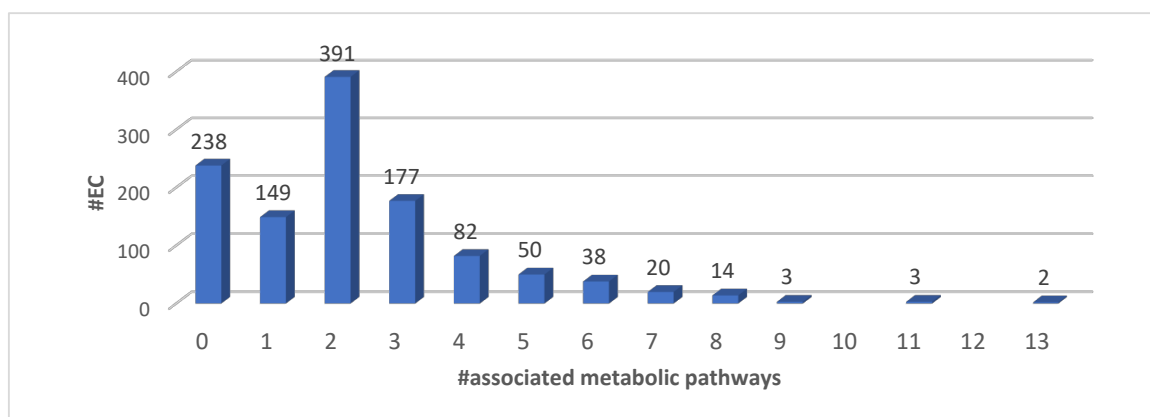


Figure 1. Distribution of functional activities (four-digit EC numbers) as a function of KEGG metabolic pathways.

The correspondence among EC numbers and proteins is plural (an EC may be associated with different proteins and a protein with different ECs). The EC proteins to KEGG metabolic pathways association is shown in Figure 2.

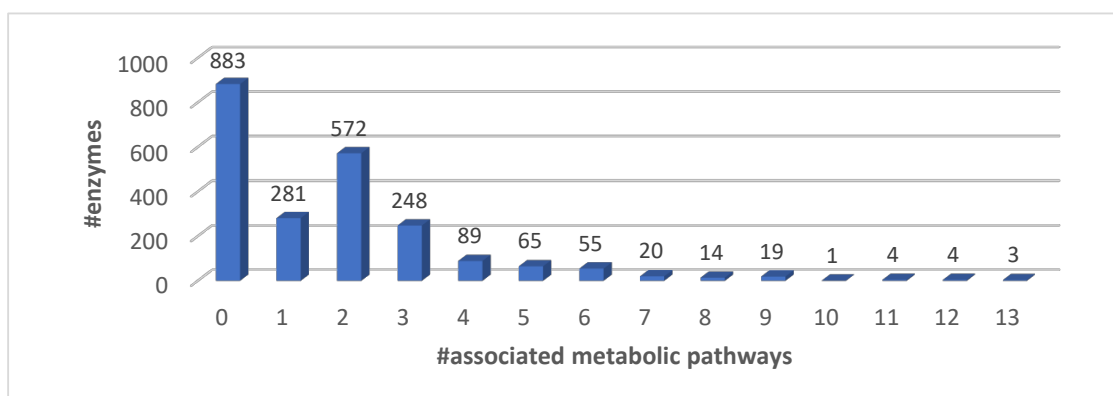


Figure 2. Distribution of EC proteins as a function of KEGG metabolic pathways.

The distribution of the EC proteins in the different KEGG metabolic pathways indicates that only 12 EC proteins are associated with 10 or more KEGG metabolic pathways (Table 3). The most frequent activities associated with the most frequent EC proteins are one oxidoreductase: EC 1.2.1.3 (aldehyde dehydrogenase (NAD⁺)); two transferases: EC 2.3.1.9 (Acetyl-CoA C-acetyltransferase), EC 2.6.1.1 (Aspartate transaminase); and one lyase: EC 4.2.1.17 (enoyl-CoA hydratase). For details on the specific biochemical reactions including the description of substrates and products, refer to the Rhea database [28].

Table 3. EC proteins involved in at least 10 KEGG metabolic pathways, and physical and predicted number of interactions.

EC Number ¹	KEGG ²	UniProt ³	PDB ⁴	IntAct ⁵	BioGRID ⁶	Int. Sites ⁷	Other EC ⁸
1.2.1.3	13: hsa00010 hsa00053 hsa00071 hsa00260 hsa00280 hsa00310 hsa00330 hsa00340 hsa00380 hsa00410 hsa00561 hsa00620 hsa01100	P49419 (13)	4ZUL (homo 4-mer)	23	62	78/235; 34/83 (21/132; 7/34)	1.2.1.8 (1) 1.2.1.31 (1)
		P49189 (12) -hsa00260	6QAP (homo 4-mer)	10	38	48/223; 0/2 (10/139; none)	1.2.1.19 (0) 1.2.1.47 (1)
		P05091 (12) -hsa00260	1O02 (homo 4-mer)	45	75	88/223; 16/41 (19/133; 4/25)	–
		P51648 (12) -hsa00260	4Q GK (homo 2-mer)	91	107	82/238; 19/54 (20/139; 4/32)	1.2.1.94 (0)
		P30837 (12)-hsa00260	–	41	93	111/517; 1/1	–
2.3.1.9	13: hsa00071 hsa00072 hsa00280 hsa00310 hsa00380 hsa00620 hsa00630 hsa00640 hsa00650 hsa00900 hsa01100 hsa01200 hsa01212	Q9BWD1 (13)	1WL4 (homo 4-mer)	20	46	94/175; 20/37 (66/113; 9/16)	–
		P24752 (13)	2IBY (homo 4-mer)	32	108	117/185; 35/59 (65/121; 17/37)	–
		P00505 (11)	5AX8 (homo 2-mer)	37	42	38/192; 5/32 (4/126; 0/21)	2.6.1.7 (0)
		P17174 (11)	6DND (1-mer)	12	73	46/200; 10/42	2.6.1.3 (0)
		P40939 (11)	6DV2 (hetero 4-mer)	116	254	24/375; 6/65 (23/337; 5/57)	1.1.1.211 (2)
		P30084 (11)	2HW5 (homo 6-mer)	65	112	45/163; 9/37 (2/68; 2/16)	–
		Q08426 (10) - hsa00062	–	123	109	234/723; 53/150	5.3.3.8 (1) 1.1.1.35 (8)
		11: hsa00062 hsa00071 hsa00280 hsa00310 hsa00380 hsa00410 hsa00640 hsa00650 hsa01100 hsa01200 hsa01212					

Hyphens in table cells refer to lack of information. ¹ The list of functional activity names corresponding to EC numbers is available in Table S1A. ² Number of metabolic KEGG associated to the EC number and list of corresponding IDs; the list of names of KEGG pathways is available in Table S1B. ³ Human protein codes included in UniProt (SwissProt section). Among brackets, number of KEGG pathways listed in the second column where the protein is active. ⁴ Representative PDB code and corresponding global stoichiometry. ⁵ Number of interacting partners in IntAct. ⁶ Number of interacting partners in BioGRID. ⁷ Number of residues predicted with ISPRED to be involved in interactions with other proteins over the total number of residues on the protein solvent accessible surface. After the semicolon, we report the number of disease related positions matching the predicted interactions sites over the number of disease related positions on the protein solvent accessible surface. Within brackets, the same numbers are restricted to the residues not involved in the PDB global stoichiometry (biological unit). When structure is not available, the number of residues in the sequence is indicated instead of the number of surface residues. ⁸ Other EC numbers associated with the protein. Within brackets, the number of metabolic KEGG pathways, where the specific activity is present.

3.2. EC Proteins and Their Interactions

A network of networks can model all the interactions that each protein can have. To exploit this possibility in the light of the available results, we focused on the EC proteins that are associated with 10 or more KEGG metabolic pathways to highlight the number of their possible interactors (Figure 3). Experimental and physical interactions are retrieved from IntAct [14] and BioGRID [15]. When restricting to interactions among human proteins, IntAct reports 337,389 interactions among 36,815 proteins (including isoforms) and BioGRID reports 471,774 interactions involving 25,420 proteins. The average number of interactors per protein is therefore equal to 18 and 37 in IntAct and BioGRID, respectively. In Figure 3, the characteristic values (average, median, first and third quartiles) of the distribution of the number of interactors reported in IntAct and BioGRID are compared among the following classes—(i) proteins involved in only one metabolic KEGG, (ii) proteins involved in at least ten metabolic KEGG pathways, and (iii) all EC proteins. EC proteins involved in a high number of KEGG metabolic pathways have also a high number of interactors, when compared to those less frequently involved.

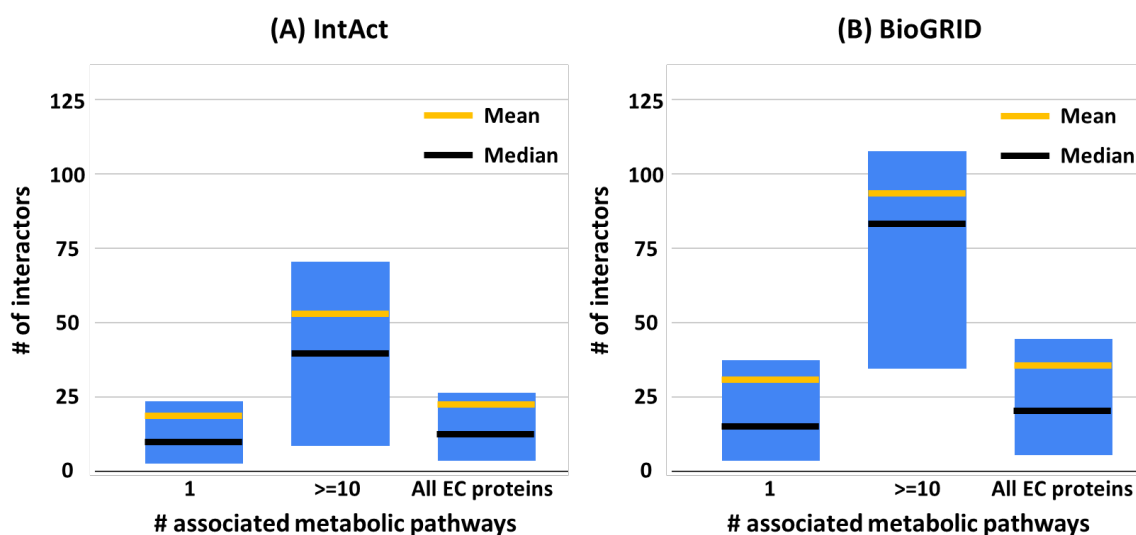


Figure 3. Statistical characterization of the number of interactors in EC proteins associated with human metabolic pathways. For each set, the boxes represent the first and third quartiles; yellow and black lines represent mean and median values, respectively. (A) and (B): from IntAct [14] and BioGRID [15], respectively. Significance of the reported differences on median values has been validated with the Mann–Whitney *U* test, obtaining *p*-value < 0.0001 when comparing the EC proteins with at least 10 interactors with the other two classes, for both IntAct and BioGRID databases. # Number of.

In Figure 4, we show that on average EC proteins that are present in at least 10 KEGG metabolic pathways, and have the highest number of interacting partners, are also endowed with the highest number of interacting sites in the solvent accessible area. This finding supports the notion that the association of experimental and theoretical data is consistent and makes it feasible to identify possible hubs in metabolic pathways.

For the human EC proteins that most frequently (≥ 10 times) participate in KEGG metabolic pathways, Table 3 lists details including the most representative PDB structure (highest coverage to the protein sequence ($\geq 70\%$) and highest atomic resolution). For each EC protein, we also indicate the putative number of predicted interaction sites (computed with ISPRED [24,25]), with the distinction among interaction sites at the protein solvent accessible surface or at the interface in the protein global stoichiometry, as reported in the PDB. We also show, for each EC protein, the total number of disease related variations and the number of disease related variations matching interactions sites.

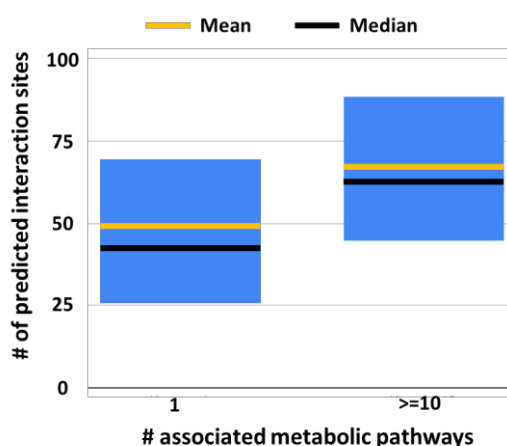


Figure 4. Statistical characterization of the number of interaction sites predicted with ISPRED4 in EC proteins associated with only one or at least 10 metabolic pathways. For each set, the boxes represent the first and third quartiles; yellow and black lines represent mean and median values, respectively. Significance of the reported difference on median values has been validated using the Mann–Whitney *U* test obtaining *p*-value = 0.04. # Number of.

3.3. The Case Study of Alpha-Aminoadipic Semialdehyde Dehydrogenase

The human protein alpha-aminoadipic semialdehyde (AASA) dehydrogenase, also known as antiquitin (P49419), coded by the gene *ALDH7A1*, is a multifunctional enzyme mediating important protective effects. The protein protects cells from oxidative stress by metabolizing lipid peroxidation-derived aldehydes (EC 1.2.1.3), and it is involved in lysine catabolism (EC 1.2.1.31). It also metabolizes betaine aldehyde to betaine (EC 1.2.1.8), an important cellular osmolyte and methyl donor. It is present with three different isoforms, one of which is only mitochondrial [19]. In human phenotype ontology [29], as reported in GeneCards, [19], the gene is associated to 59 human phenotypes and eight different REACTOME [5] and 13 KEGG [4] metabolic pathways (Table 3). In Gene Cards, expression data suggest that the protein is present in many tissues. GeneORGANizer [30] lists brain, cranial nerve, eye, head, liver, lung, peripheral nervous system, and peripheral nerve as confident expression organs. In the Human Protein Atlas [31], *ALDH7A1* is associated with 34 reactions in 17 different subsystems—cytosol, endoplasmic reticulum, lysosome, mitochondria, and peroxisome. Given its relevance for the biology of the cell, it has been the subject of more than 100 publications (they can be reached via GeneCards [19]). The protein is present in the cytoplasm, in the mitochondrion, and in the nucleus [18] and interacts with other proteins (23 interactors in IntAct [14] and 62 in BioGRID [15]). It has been crystallized 15 times [7]. Here we focus on a complete form of the biological unit (PDB code: 4ZUL), a homotetramer solved with a resolution of 0.170 nm and with the maximal coverage with the sequence P49419, without the mitochondrial target peptide [32]. Recently, important variants of the protein, associated with PDE and hampering its activity, have been also solved with atomic resolution [33]. Finally, the protein, as a major feature, according to the MobiDB database [34], does not have intrinsically disordered regions (IDPs). We are interested in highlighting at a molecular level some of the protein properties, which are related to its involvement in different metabolic pathways and diseases.

A whole list of all the variations available from different databases is listed in Table S2. The protein sequence P49419 (comprising 539 residue) is endowed with 232 variations from different data bases (Table S2); 195 variations associated to 160 positions are disease related (Table S2), and 117 disease related variations are associated to PDE.

In Figure 5, we show one of the four subunits of the homotetrameric protein (4ZUL, chain A) and highlight the interface region (in orange) in the global stoichiometric unit. This allows distinguishing between the region at the interface and the region exposed to the solvent. We map (in green) variations

predicted as possible interaction sites with ISPRED4 [24]. These sites, located in the protein-exposed region, are likely to mediate interactions with other proteins. We also map disease related residues at the interface and in the protein (small spheres).

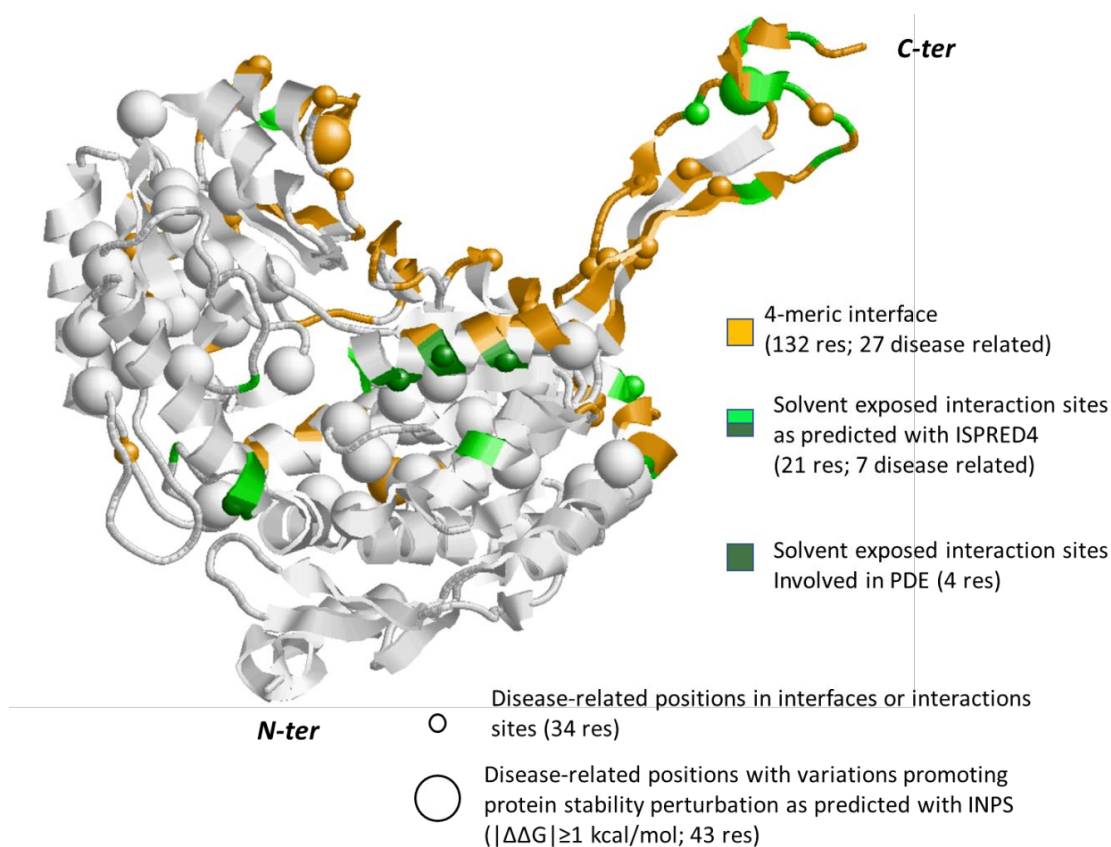


Figure 5. Monomeric subunit of human ALDH7A1 protein (PDB code: 4ZUL.A). Interaction surface in the tetramer as derived from the crystallographic coordinates is in orange. Interaction sites out the tetrameric interface, as predicted with ISPRED4, are in green. Positions in these regions carrying disease related variations (Table S2) are highlighted with small spheres. Big spheres highlight positions in the protein carrying disease related variations (Table S2, for details) and promoting a large variance of folding free energy, as predicted with INPS [26]. Grey color: the background protein backbone.

For the sake of completeness, we computed the likelihood of all the protein variations to affect protein stability (Table S2) and found as expected [2] that variations that are not always disease-related are perturbing the protein folding.

In Figure 5, big spheres highlight those variations that most affect protein stability ($|\Delta\Delta G| \geq 1$ kcal/mol). Interestingly, we found that PDE related variations V278L, Q281H, M285V, and K375R occur at the solvent accessible protein surface and match predicted interaction sites without affecting protein stability.

Table S2 provides a complete list of the properties for all the protein variations present in different databases, associated with specific diseases. For each variation, Table S2 lists its location in the protein reference sequence P49419, its location in the protein three-dimensional structure (4ZUL, chain A) and the predicted effect on the protein stability ($\Delta\Delta G$), computed with INPS, [26]. It also indicates if the disease-associated residue occurs in the target peptide, in the tetrameric interface, in the active sites, and regions annotated in the corresponding UniProt file (P49419). The ISPRED predictions are shown when present. Interestingly, many variations occur in the transit peptide (26 residue long, UniProt, P49419, [8]), a specific N-terminal peptide in the protein sequence mediating the mitochondrial import. This suggests that disease may be also due to an unpaired translocation of the protein to

the mitochondrial compartment. For the sake of comparison, in Table S2 (Supplementary Materials), we label, in red, some PDE disease-related variations, known to occur in the aldehyde substrate binding site (N195S, P197S, A199V, G202V, W203G) and recently detailed with atomic resolution on their effect on the protein structure and function [33]. INPS predicts P197S, G202V, W203G as perturbing the protein stability (Table S2).

3.4. The Case Study of Acetyl-CoA C-Acetyltransferase

In Table 3, the enzyme proteins listed for the activity EC 2.3.1.9 are Acetyl-CoA C-acetyltransferases (ACAT2, cytosolic and ACAT1 mitochondrial), which catalyze the condensation of an acetyl-CoA and an acyl-CoA (often another acetyl-CoA), leading to the synthesis of an acyl-CoA with a longer fatty acid chain [35,36]. The two enzymes are encoded by two different genes and their residue chains share 39% sequence identity. The cytosolic enzyme (UniProt Q9BWD1) is encoded by ACAT2 and the mitochondrial one by ACAT1 (UniProt P24752). The 3D structure of both proteins has been resolved at the atomic resolution. Two representative structures (2IBY:A and 1WL4:A for ACAT1 and ACAT2, respectively) structurally superimpose with a root mean square deviation as low as 0.09 nm and therefore show a high structural similarity. Moreover, they conserve the two cysteine residues that form the active site.

In humans, ACAT1 is one of the enzymes that catalyzes the last step of the mitochondrial beta-oxidation pathway, an aerobic process breaking down fatty acids into acetyl-CoA, and it plays a major role in the metabolism of ketone bodies. ACAT2 is important in the pathway of fatty acid metabolism, and in the biosynthetic pathway of cholesterol. ACAT1 and ACAT2 are both associated with the same disease—alpha-methylacetoacetic aciduria (OMIM 203,750) or deficiency of acetyl-CoA acetyltransferase, an inborn error of isoleucine catabolism. They share 13 metabolic pathways (Table 3).

In human phenotype ontology [29], as reported in GeneCards, [19], ACAT1 is associated with 118 human phenotypes, while ACAT2 is associated with 23 human phenotypes.

GeneORGANizer [30] reports that brain and head are confident expression organs for both ACAT1 and ACAT2. ACAT1 is also expressed in liver, oesophagus, and stomach. In the Human Protein Atlas [31], ACAT1 is associated with two reactions in cytosol, mitochondria, and peroxisome. Given its relevance for the biology of the cell, the two proteins are subject of many publications that can be reached via GeneCards [19]. ACAT1 protein has 32 interactors in IntAct [14] and 108 in BioGRID [15], while ACAT2 has 20 interactors in IntAct [14] and 46 in BioGRID [15] (Table 3). Particularly for ACAT1, these numbers are significantly larger than the number of interactions per protein in the whole dataset. Finally, according to the MobiDB database [34], none of the proteins have intrinsically disordered regions (IDPs).

First, we focus on a complete form of the ACAT1 biological unit (PDB code: 2IBY), a homotetramer solved with a resolution of 0.185 nm. The monomeric chain covers all the mature form of P24752, depleted of the target peptide [35]. We are interested in highlighting at a molecular level some of the protein properties, which are related to its involvement in different metabolic pathways and diseases. A whole list of all the variations available from different databases is reported in Table S3.

In Figure 6, we show the subunit A of the homotetrameric protein, and we color the interface region in the global stoichiometric unit in orange and the residues predicted with ISPRED4 as possible interaction sites in green. As in Figure 5, we represent disease-related residues at the interface and in the protein with small spheres, while big spheres highlight variations that most affect protein stability ($|\Delta\Delta G| \geq 1$ kcal/mol). Table S3 provides a complete list of the properties for all the protein variations present in different databases, associated with specific diseases and mapped on the protein reference sequence (P24752) and three-dimensional structure (2IBY, chain A) on the protein.

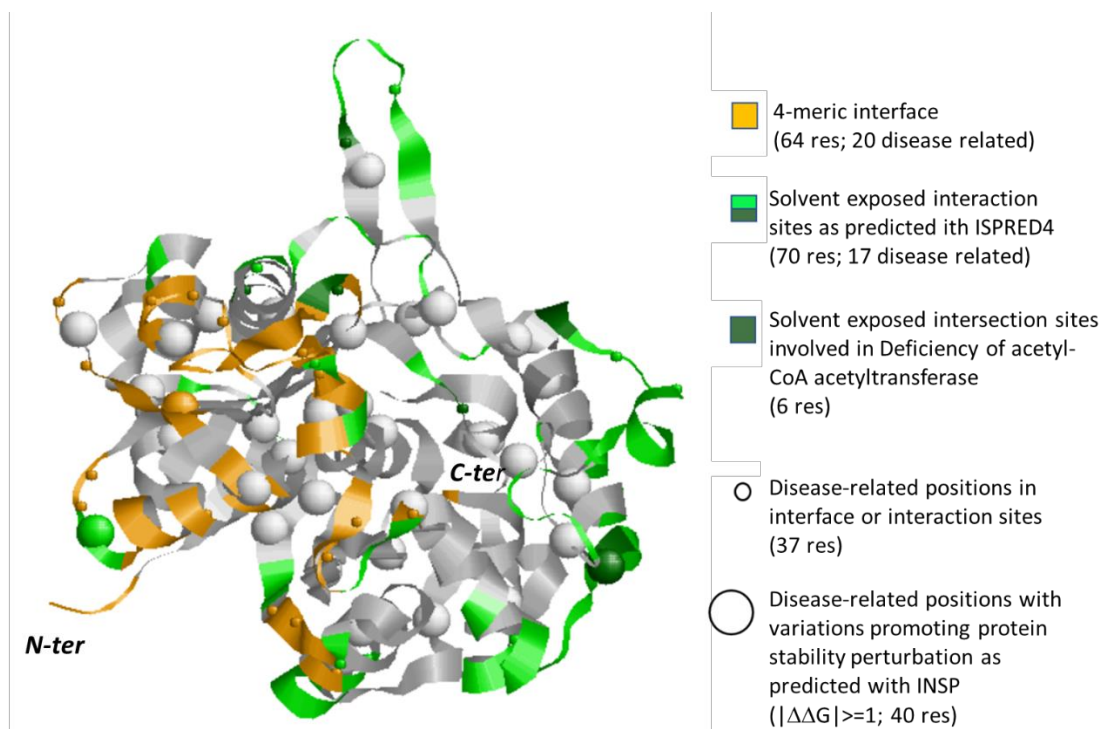


Figure 6. Monomeric subunit of human ACAT1 protein (PDB code: 2IBY.A). Interaction surface in the tetramer as derived from the crystallographic coordinates is in orange. Interaction sites out the tetrameric interface, as predicted with ISPRED4 are in green. Positions in these regions carrying disease related variations (Table S3) are highlighted with small spheres. Big spheres highlight positions in the protein carrying disease related variations (Table S3, for details) and promoting a large variance of folding free energy, as predicted with INSP [26]. Grey color: the background protein backbone.

As in the case of ALDH71, we found that variations in putative interaction sites are often conducive to the impairment of protein function. This is the case of eight variations related to Deficiency of acetyl-CoA acetyltransferase (Q73P, N158S, N158D, R208Q, R208G, T241A, R258C, T285I). Interestingly, five variations occur in the 33 residue-long mitochondrial target peptides, suggesting that disease may be also due to an unpaired translocation of the protein to the mitochondrial compartment.

For the ACAT2 protein, we adopted the PDB entry 1WL4 to represent the interaction regions and map the variations. The entry contains a homotetrameric form solved with a resolution of 0.155 nm. Each chain covers the whole sequence (Q9BWD1) [36].

In Figure 7, we show the ACAT2 subunit chain A and represent tetrameric interaction regions, predicted interaction residues and positions carrying disease related variations with the same representation as in Figures 5 and 6. Table S4 (Supplementary Materials) provides a complete list of the properties for all the protein variations present in different databases. Reinforcing the previous observations on the relevance of interaction regions, the only reported variation of ACAT2 associated with the Deficiency of acetyl-CoA acetyltransferase (E176K) occurs at the solvent accessible protein surface and it is predicted with ISPRED4 as interaction site. Moreover, this variation has a small effect on the protein stability ($\Delta\Delta G$) (-0.12 kcal/mol, see Table S4), reinforcing the concept that variations which are interaction sites can lead to disease by hampering protein-protein interactions without affecting protein stability.

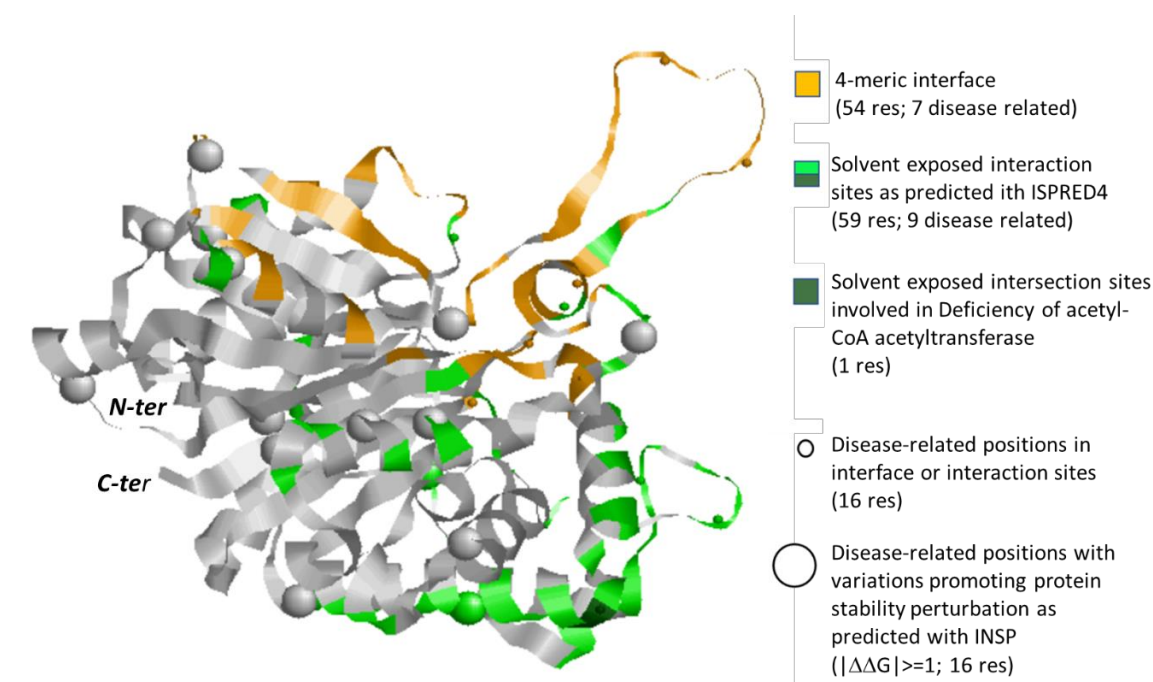


Figure 7. Monomeric subunit of human ACAT2 protein (PDB code: 1WL4.A). Interaction surface in the tetramer as derived from the crystallographic coordinates is in orange. Interaction sites out of the tetrameric interface, as predicted with ISPRED4 are in green. Positions in these regions carrying disease related variations (Table S4 (Supplementary Materials)) are highlighted with small spheres. Big spheres highlight positions in the protein carrying disease related variations (Table S4, for details) and promoting a large variance of folding free energy, as predicted with INPS [26]. Grey color: the background protein backbone.

4. Conclusions

One of the goals of system biology is to produce a three-dimensional model of the cell metabolism. As a preliminary step, nowadays, we cope with the problem of generating links among different databases that are dissecting the cell complexity into useful and important sets of data, addressing cell components from different perspectives and with different approaches. Here, we explore the problem of relating KEGG metabolic pathways to the network of protein–protein interactions (PPI) by restricting our study to human enzymes and their relation to KEGG metabolic pathways and PPI interaction maps. We found that, when enzymes are hubs in metabolic pathways, they are on average interacting with a high number of proteins as detected with different experimental methods and are also endowed with a high number of predicted interacting sites (Figures 3 and 4).

Our results suggest that enzymatic metabolic hubs are hubs in networks of protein–protein interaction. Consistently, hubs are on average endowed with the highest numbers of predicted interaction sites when compared to the other EC proteins in the networks.

Protein variants can be associated with diseases. Possible indications on the effect of disease-related variations are investigated by predicting whether the variation is located at a putative interaction site and/or whether it affects the protein stability. As a test case, we focused on the ALDH7A1 gene, which according to our data is one of the most frequent gene in KEGG metabolic pathways. The protein is associated with 232 variations in different databases (Table S2 (Supplementary Materials)). We localize the disease-related variations in the protein structure and find that 27% of them affect the protein stability, rather independently of their location in active sites, in interfaces of the biological assembly or in the protein solvent exposed area (Table S2). The protein also interacts physically with 23–62 different interactors as documented in Intact and BioGrid (Table 3). We predict that 21 residues are likely to act as interaction sites in the solvent exposed protein surface (Table S2). Among these, seven are

disease-related, and four are associated with PDE. This suggests that each disease-related variation occurring in the external surface can affect the efficiency of the protein in each of the different metabolic pathways where it is active, by affecting the interplay with all the different partners and without affecting protein stability. Similar conclusions stand also for the analysis of ACAT1 and ACAT2 gene products, representative of the second EC number of the list shown in Table 3. Again, by entering into the details of the molecular properties, we find a supportive example of the relevance of variations at the protein solvent accessible interface as conducive to disorders.

Summing up, a conclusion from our analysis is that, with the data presently available and with computational tools, it is possible to highlight enzyme proteins that are central to biochemical pathways and to identify possible molecular mechanisms at the basis of their association with specific diseases.

Supplementary Materials: The following are available online at <http://www.mdpi.com/2227-9059/8/8/250/s1>.

Author Contributions: Conceptualization, M.P.L. and R.C.; Investigation, G.B., D.B., M.P.L., C.S. and R.C.; Supervision, R.C.; Writing—original draft, R.C.; Writing—review & editing, M.P.L. and R.C. All authors have read and agreed to the published version of the manuscript.

Funding: G.B. salary is from PRIN2017 project 2017483NH8_002 delivered to C.S. of the Italian Ministry of University and Research (MIUR).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bugg, T.D.H. *Introduction to Enzyme and Coenzyme Chemistry*, 3rd ed.; Wiley: New York, NY, USA, 2012.
2. Savojardo, C.; Martelli, P.L.; Casadio, R. Protein–Protein Interaction Methods and Protein Phase Separation. *Ann. Rev. Biom. Data Sci.* **2020**, *89*–112. [[CrossRef](#)]
3. Wolfinbarger, L., Jr. *Enzyme Regulation in Metabolic Pathways*; Wiley: New York, NY, USA, 2017.
4. Kegg. Available online: <https://www.genome.jp/kegg> (accessed on 1 June 2020).
5. Reactome. Available online: <https://reactome.org> (accessed on 1 June 2020).
6. Enzyme Nomenclature. Available online: <https://www.qmul.ac.uk/sbcs/iubmb/enzyme> (accessed on 1 June 2020).
7. Protein Data Bank. Available online: <https://www.rcsb.org> (accessed on 1 June 2020).
8. UniProt. Available online: <https://www.uniprot.org> (accessed on 1 June 2020).
9. OMIM. Available online: <https://omim.org> (accessed on 1 June 2020).
10. BioMuta. Available online: <https://hive.biochemistry.gwu.edu/biomuta> (accessed on 1 June 2020).
11. DisGeNet. Available online: <https://www.disgenet.org> (accessed on 1 June 2020).
12. ClinVar. Available online: <https://www.ncbi.nlm.nih.gov/clinvar> (accessed on 1 June 2020).
13. MalaCards. Available online: <https://www.malacards.org> (accessed on 1 June 2020).
14. IntAct. Available online: <https://www.ebi.ac.uk/intact> (accessed on 1 June 2020).
15. BioGrid. Available online: <https://thebiogrid.org> (accessed on 1 June 2020).
16. Brenda. Available online: <https://www.brenda-enzymes.org> (accessed on 1 June 2020).
17. Enzyme Portal. Available online: <https://www.ebi.ac.uk/enzymeportal> (accessed on 1 June 2020).
18. Mechanism and Catalytic Site Atlas. Available online: <https://www.ebi.ac.uk/thornton-srv/m-csa> (accessed on 1 June 2020).
19. GeneCards. Available online: <https://www.genecards.org> (accessed on 1 June 2020).
20. Mughal, F.; Caetano-Anollés, G. MANET 3.0: Hierarchy and modularity in evolving metabolic networks. *PLoS ONE* **2019**, *14*, e0224201. [[CrossRef](#)] [[PubMed](#)]
21. Manet. Available online: <http://manet.illinois.edu> (accessed on 1 June 2020).
22. Babbi, G.; Martelli, P.L.; Profiti, G.; Bovo, S.; Savojardo, C.; Casadio, R. eDGAR: a database of Disease-Gene Associations with annotated Relationships among genes. *BMC Genom.* **2017**, *18*, 554. [[CrossRef](#)] [[PubMed](#)]
23. eDGAR. Available online: http://edgar.biocomp.unibo.it/gene_disease_db (accessed on 1 June 2020).
24. ISPRED4. Available online: <https://ispred4.biocomp.unibo.it/welcome/default/index> (accessed on 1 June 2020).
25. Savojardo, C.; Fariselli, P.; Martelli, P.L.; Casadio, R. ISPRED4: interaction sites PREDiction in protein structures with a refining grammar model. *Bioinformatics* **2017**, *33*, 1656. [[CrossRef](#)] [[PubMed](#)]
26. INPS-Md. Available online: <https://inpsmd.biocomp.unibo.it/inpsSuite> (accessed on 1 June 2020).

27. Savojardo, C.; Fariselli, P.; Martelli, P.L.; Casadio, R. INPS-MD: A web server to predict stability of protein variants from sequence and structure. *Bioinformatics* **2016**, *32*, 2542. [[CrossRef](#)] [[PubMed](#)]
28. Rhea-DB. Available online: <https://www.rhea-db.org> (accessed on 1 June 2020).
29. Human Phenotype Ontology. Available online: <https://hpo.jax.org/app> (accessed on 1 June 2020).
30. Gene Organizer. Available online: http://geneorganizer.huji.ac.il/browse/?GENE_IDs=ALDH7A1&FullSite=T#btn_organs-browse (accessed on 1 June 2020).
31. Protein Atlas. Available online: <https://www.proteinatlas.org/ENSG00000164904-ALDH7A1/tissue> (accessed on 1 June 2020).
32. Luo, M.; Tanner, J.J. Structural Basis of Substrate Recognition by Aldehyde Dehydrogenase 7A1. *Biochemistry* **2015**, *54*, 5513. [[CrossRef](#)] [[PubMed](#)]
33. Laciak, A.R.; Korasick, D.A.; Wyatt, J.W.; Gates, K.S.; Tanner, J.J. Structural and biochemical consequences of pyridoxine-dependent epilepsy mutations that target the aldehyde binding site of aldehyde dehydrogenase ALDH7A1. *FEBS J.* **2020**, *287*, 173. [[CrossRef](#)] [[PubMed](#)]
34. MobiDB. Available online: <https://mobidb.bio.unipd.it/P49419/db> (accessed on 1 June 2020).
35. Haapalainen, A.M.; Meriläinen, G.; Pirilä, P.L.; Kondo, N.; Fukao, T.; Wierenga, R.K. Crystallographic and Kinetic Studies of Human Mitochondrial Acetoacetyl-CoA Thiolase: The Importance of Potassium and Chloride Ions for Its Structure and Function. *Biochemistry* **2007**, *46*, 4305–4321. [[CrossRef](#)] [[PubMed](#)]
36. Kursula, P.; Sikkilä, H.; Fukao, T.; Kondo, N.; Wierenga, R.K. High Resolution Crystal Structures of Human Cytosolic Thiolase (CT): A Comparison of the Active Sites of Human CT, Bacterial Thiolase, and Bacterial KAS I. *J. Mol. Biol.* **2005**, *347*, 189–201. [[CrossRef](#)] [[PubMed](#)]



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Tumor genotype, location, and malignant potential shape the immunogenicity of primary untreated gastrointestinal stromal tumors

Daniela Gasparotto, ... , Angelo Paolo Dei Tos, Roberta Maestro

JCI Insight. 2020;5(22):e142560. <https://doi.org/10.1172/jci.insight.142560>.

Research Article

Oncology

Intratumoral immune infiltrate was recently reported in gastrointestinal stromal tumors (GISTs). However, the tumor-intrinsic factors that dictate GIST immunogenicity are still largely undefined. To shed light on this issue, a large cohort (82 samples) of primary untreated GISTs, representative of major clinicopathological variables, was investigated by an integrated immunohistochemical, transcriptomic, and computational approach. Our results indicate that tumor genotype, location, and malignant potential concur to shape the immunogenicity of primary naive GISTs. Immune infiltration was greater in overt GISTs compared with that in lesions with limited malignant potential (miniGISTs), in *KIT/PDGFRA*-mutated tumors compared with that in *KIT/PDGFRA* WT tumors, and in *PDGFRA*-mutated compared with *KIT*-mutated GISTs. Within the *KIT*-mutated subset, a higher degree of immune colonization was detected in the intestine. Immune hot tumors showed expression patterns compatible with a potentially proficient but curbed antigen-specific immunity, hinting at sensitivity to immunomodulatory treatments. Poorly infiltrated GISTs, primarily *KIT/PDGFRA* WT intestinal tumors, showed activation of Hedgehog and WNT/ β -catenin immune excluding pathways. This finding discloses a potential therapeutic vulnerability, as the targeting of these pathways might prove effective by both inhibiting pro-oncogenic signals and fostering antitumor immune responses. Finally, an intriguing anticorrelation between immune infiltration and *ANO1/DOG1* expression was observed, suggesting an immunomodulatory activity for anoctamin-1.

Find the latest version:

<https://jci.me/142560/pdf>



Tumor genotype, location, and malignant potential shape the immunogenicity of primary untreated gastrointestinal stromal tumors

Daniela Gasparotto,¹ Marta Sbaraglia,² Sabrina Rossi,³ Davide Baldazzi,¹ Monica Brenca,¹ Alessia Mondello,¹ Federica Nardi,¹ Dominga Racanelli,¹ Matilde Cacciatori,³ Angelo Paolo Dei Tos,^{2,4} and Roberta Maestro¹

¹Unit of Oncogenetics and Functional Oncogenomics, Centro di Riferimento Oncologico di Aviano (CRO Aviano) IRCCS, National Cancer Institute, Aviano, Italy. ²Department of Pathology, Azienda Ospedaliera Universitaria di Padova, Padua, Italy. ³Department of Pathology and Molecular Genetics, Treviso General Hospital, Treviso, Italy. ⁴Department of Medicine, University of Padua School of Medicine, Padua, Italy.

Intratumoral immune infiltrate was recently reported in gastrointestinal stromal tumors (GISTs). However, the tumor-intrinsic factors that dictate GIST immunogenicity are still largely undefined. To shed light on this issue, a large cohort (82 samples) of primary untreated GISTs, representative of major clinicopathological variables, was investigated by an integrated immunohistochemical, transcriptomic, and computational approach. Our results indicate that tumor genotype, location, and malignant potential concur to shape the immunogenicity of primary naive GISTs. Immune infiltration was greater in overt GISTs compared with that in lesions with limited malignant potential (miniGISTs), in *KIT/PDGFR*A-mutated tumors compared with that in *KIT/PDGFR*A WT tumors, and in *PDGFR*A-mutated compared with *KIT*-mutated GISTs. Within the *KIT*-mutated subset, a higher degree of immune colonization was detected in the intestine. Immune hot tumors showed expression patterns compatible with a potentially proficient but curbed antigen-specific immunity, hinting at sensitivity to immunomodulatory treatments. Poorly infiltrated GISTs, primarily *KIT/PDGFR*A WT intestinal tumors, showed activation of Hedgehog and WNT/ β -catenin immune excluding pathways. This finding discloses a potential therapeutic vulnerability, as the targeting of these pathways might prove effective by both inhibiting pro-oncogenic signals and fostering antitumor immune responses. Finally, an intriguing anticorrelation between immune infiltration and *ANO1/DOG1* expression was observed, suggesting an immunomodulatory activity for anoctamin-1.

Authorship note: DG and MS are co-first authors. APDT and RM are co-senior authors.

Conflict of interest: The authors have declared that no conflict of interest exists.

Copyright: © 2020, Gasparotto et al. This is an open access article published under the terms of the Creative Commons Attribution 4.0 International License.

Submitted: July 24, 2020

Accepted: October 7, 2020

Published: November 19, 2020

Reference information: *JCI Insight*. 2020;5(22):e142560.
<https://doi.org/10.1172/jci.insight.142560>.

Introduction

Gastrointestinal stromal tumors (GISTs) are the most common sarcomas of the gastrointestinal tract (1, 2). The majority of GISTs (~85%) are driven by activating mutations in the gene encoding the receptor tyrosine kinase *KIT* (65%–80%) or *PDGFR*A (15%–20%). The remaining fraction of tumors, overall referred to as *KIT/PDGFR*A WT GISTs (*K/P* WT), may rely on different oncogenic events: activation of the RAS/RAF/MAPK pathway, caused most frequently by *NF1* or *BRAF* mutations (about 10% of the cases); defects in components of the succinate dehydrogenase mitochondrial complex II (SDH) in syndromic gastric GISTs (<5%); and rare (<1%) oncogenic gene fusions (1–6).

Although localized GISTs are potentially curable by surgery alone, a significant fraction of tumors relapses after this treatment. Adjuvant therapy with imatinib targeting activated *KIT/PDGFR*A proteins proved to be significantly beneficial in the prevention of recurrence and in prolonging the survival of patients with advanced/metastatic disease (1, 7). Yet, some patients are ab initio poorly responsive to this tyrosine kinase inhibitor due to the expression of imatinib-refractory mutations (e.g., *PDGFR*A D842V) or independency of *KIT/PDGFR*A signaling (*K/P* WT tumors). Moreover, even in responsive patients, imatinib is rarely curative as secondary resistance mutations frequently occur. In these settings,

switching to other tyrosine kinase inhibitors, such as sunitinib or regorafenib, has demonstrated clinical benefit (1, 7). Recently, the portfolio of effective drugs used to treat GIST has expanded to also include avapritinib (8) and ripretinib (9).

Mounting evidence indicates that tumor immune microenvironment plays a key role in tumor inception, progression, and response to therapy. In this regard, recent works documented the presence of intratumor immune cell infiltration in GISTs and its effect on imatinib efficacy (10–14). Imatinib has been shown to amplify a preexisting cytotoxic antitumor response by inhibiting tumor cell production of the immune inhibitory enzyme indoleamine 2,3-dioxygenase. In addition, a potentiated effect of imatinib when combined with checkpoint inhibitors (anti-PD1, anti-CTLA-4, or anti-CD40) has been demonstrated in preclinical models (15–17). Based on these preliminary results and on the success of immunomodulatory treatments in other tumor types, several clinical trials aiming at assessing the efficacy of immune checkpoint inhibitors in GISTs are being conducted (NCT01643278, NCT01738139, NCT02500797, NCT02834013, NCT02880020, and NCT03291054) (18–20).

The disclosure of new therapeutic vulnerabilities in GIST is particularly relevant for that fraction of tumors, namely, *K/P* WT GISTs, that are currently orphan of effective therapies. With this in mind, we investigated the immune infiltrate by an integrated immunohistochemical, transcriptomic, and computational approach in what we believe is one of the largest cohorts of primary untreated GISTs analyzed by RNA sequencing to date. Immune contexture was examined in relation to driver gene (*KIT*, *PDGFRA*, *K/P* WT), tumor location (gastric and intestinal), and malignant potential (miniGIST and overt GIST).

Results

In situ evaluation of immune contexture. As a first step to elucidate the role of immune contexture in GISTs, an explorative cohort of 38 primary untreated GISTs was investigated by IHC. Clinicopathological characteristics of this series are reported in Table 1.

In line with previous studies (12–14), T lymphocytes and macrophages were the most abundant tumor-infiltrating immune cells, in both intestinal and gastric GISTs. CD3⁺ cells ranged between 1 and 117 (median 27.5) per HPF and were distributed as follows (median): CD4⁺ = 4.5, CD8⁺ = 15.0, Foxp3⁺ = 2.0. The number of CD68⁺ cells ranged between 17.0 and 170 (median 53.5). Few CD20⁺ B cells (range 0–19, median 0) and occasional reactivity for PD1 or PDL1 were detected (Supplemental Table 1; supplemental material available online with this article; <https://doi.org/10.1172/jci.insight.142560DS1>).

When the series was analyzed as a whole, no significant correlation among immune cell types, mitotic index, or tumor site was found. Nevertheless, differences emerged when tumors were compared according to genotype. In particular, the median number of T cells (CD3⁺, CD4⁺, and CD8⁺) was tendentially higher in *K/P*-mutated GISTs than *K/P* WT GISTs (Figure 1 and Figure 2). This difference was particularly evident for the tumors located in the intestine, where it reached statistical significance. Moreover, *KIT*-mutated gastric GISTs featured an inferior degree of infiltration both when compared with *PDGFRA*-mutated gastric tumors and when compared with the *KIT*-mutated counterpart of the intestine (Figure 1).

Transcriptional assessment of immune infiltration. To extend this initial observation, we interrogated the transcriptional profile of a cohort of 77 GISTs that were representative of different driver mutations, locations, and malignant potential (Table 1). This series included 33 of 38 cases analyzed by IHC and comprised 62 *K/P*-mutated tumors and 15 *K/P* WT tumors (3 *BRAF*, 7 *NF1*-mutated, and 5 WT for all the aforementioned genes as well as for *SDH A-D* genes, and hence defined “driver mutation unknown”) (Table 1).

After the samples were dichotomized into contrast groups according to tumor site (stomach, intestine), malignant potential (miniGIST, overt GIST), and oncogenic driver (*KIT*, *PDGFRA*, *K/P* WT), the transcriptome was interrogated for immune signatures by gene set enrichment analysis (GSEA), Ingenuity Pathway Analysis (IPA), and Reactome analyses. Pathways associated with the immune system emerged as significantly enriched in the contrast *K/P*-mutated versus *K/P* WT, particularly in the intestinal subset. A trend for enrichment of immunity-related genes also emerged when contrasting *PDGFRA* versus *KIT* gastric tumors and overt GISTs versus miniGISTs (Figure 3 and Supplemental Table 2). Finally, focusing on *KIT*-mutated GISTs, immunity-related terms were slightly more represented in intestinal tumors than gastric tumors.

These enrichments were paralleled by the differential expression of several immune cell-attracting/activating cytokines, inflammatory interleukins, and related molecules (Supplemental Table 3).

To gain better insights into the extent and nature of immune infiltration, several computational methods (single-sample GSEA [ssGSEA], CIBERSORT, and MCP-counter) were exploited to infer

Table 1. GIST cohorts

	IHC cohort (38 cases)	RNA-sequencing cohort (77 cases)
	No. (%)	No. (%)
Sex		
Male	18 (47%)	40 (52%)
Female	20 (53%)	37 (48%)
Location		
Stomach	18 (47%)	43 (56%)
Small intestine	19 (50%)	34 (44%)
Esophagus	1 (3%)	
Tumor size		
<2 cm	3 (8%)	25 (32%)
≥2≤5 cm	17 (45%)	23 (10%)
>5≤10 cm	13 (34%)	22 (29%)
>10 cm	5 (13%)	7 (9%)
MI		
≤5	19 (50%)	53 (69%)
>5	19 (50%)	24 (31%)
Type		
miniGIST (size <2cm, MI ≤5)	3 (8%)	15 (19%)
Overt GIST (size ≥2 cm, any MI)	35 (92%)	62 (81%)
Mutations		
<i>KIT</i> exon 9	2 (5%)	4 (5%)
<i>KIT</i> exon 11	23 (61%)	41 (53%)
<i>KIT</i> exon 13	1 (3%)	2 (3%)
<i>PDGFRA</i> exon 12	0	1 (1%)
<i>PDGFRA</i> exon 14	1 (3%)	3 (4%)
<i>PDGFRA</i> exon 18	4 (10%)	11 (14%)
<i>PDGFRA</i> D842V	2 cases	6 cases
<i>BRAF</i>	3 (8%)	3 (4%)
<i>NF1</i>	4 (10%)	7 (9%)
Unknown	0	5 (6%)
Risk of relapse		
Very low	5 (13%)	27 (35%)
Low	7 (18%)	13 (17%)
Intermediate	6 (16%)	9 (13%)
High	20 (53%)	28 (35%)

GIST, gastrointestinal stromal tumors; MI, mitotic index.

the presence of specific immune cell types from bulk transcriptomic data. In particular, ssGSEA was applied to estimate the distribution of 24 immune cell types as well as a cumulative immune infiltration score (IIS) and T cell-specific infiltration score (TIS) per tumor (21). Unsupervised clustering analysis, based on the scores of the 24 immune cell types, identified 3 major groups (Figure 4): the first cluster essentially comprised cold tumors (16 samples) with low IIS and TIS; the second cluster comprised tumors with an “intermediate” degree of infiltration (27 samples); and the third cluster mostly comprised highly infiltrated, “hot” tumors (34 samples).

IIS and TIS were significantly correlated (Spearman’s correlation: $r = 0.94$; $P = 2 \times 10^{-7}$). Tumors with high IIS were by and large overt GISTs, whereas miniGISTs tended to be cold. A heatmap of the samples ranked according to IIS is provided in Supplemental Figure 1A.

IIS/TIS displayed a positive correlation with the scores for IFN- γ signature (22) ($r = 0.82$, $P < 1 \times 10^{-6}$) and antigen-presenting machinery (APM) ($r = 0.63$, $P < 1 \times 10^{-6}$), a proxy for the expression of antigen-processing and presentation molecules. Moreover, IIS, IFN- γ , and APM correlated with the cytolytic activity score (CYT) (23), a surrogate estimate of cytotoxic lymphocyte activation based on the expression of granzyme A and perforin ($r = 0.62$, $r = 0.83$, $r = 0.82$, all $P < 1 \times 10^{-6}$) (Figure 4A). Overall, these data were suggestive of a potentially proficient antigen-specific immunity in a significant fraction of GISTs.

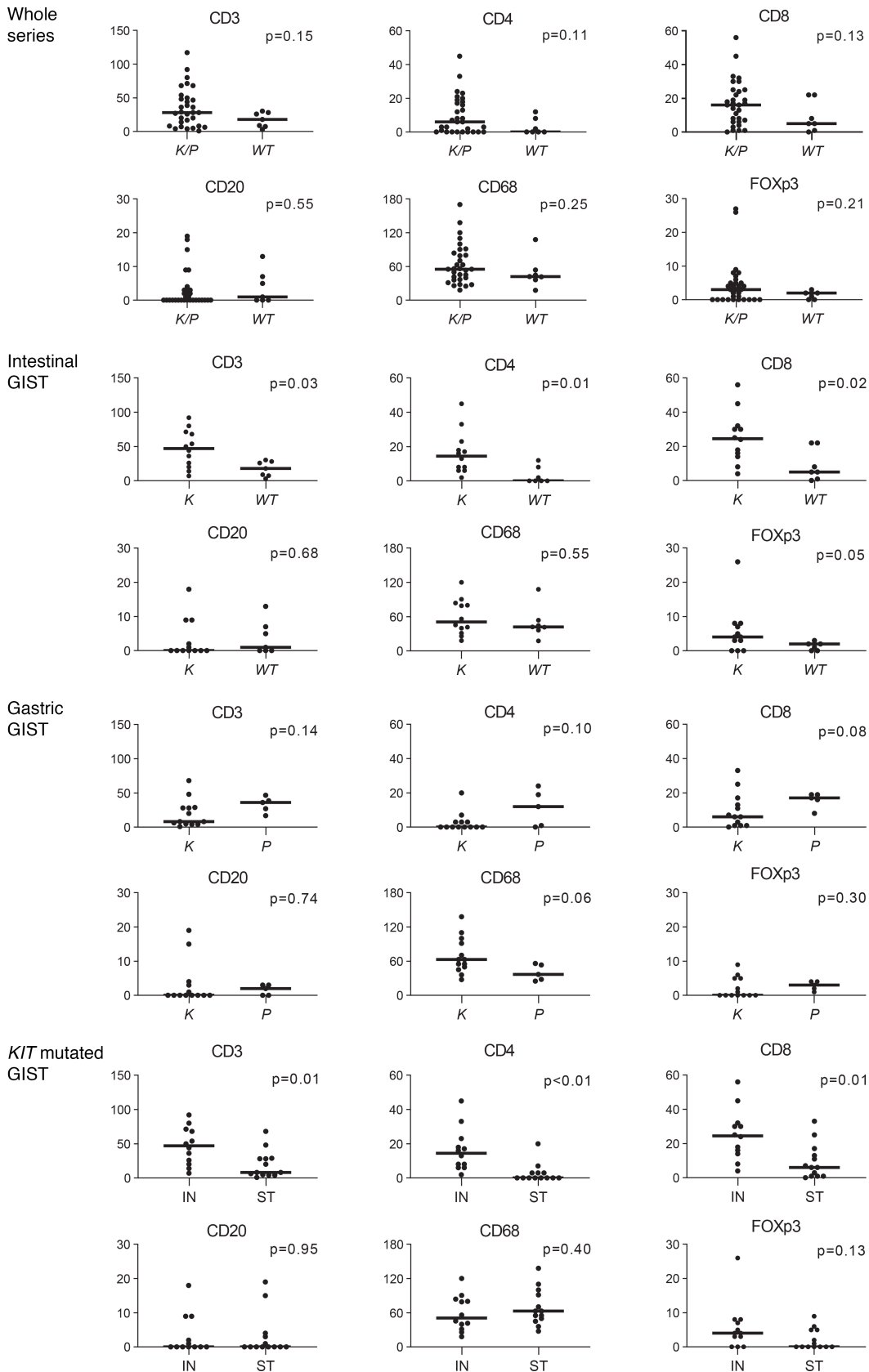


Figure 1. In situ evaluation of immune infiltration in GISTs by IHC analysis. Staining for immune cell markers in the series analyzed as a whole ($n = 38$ cases), and in intestinal ($n = 19$) and gastric ($n = 18$) GISTs analyzed separately. Cases are grouped according to the genotype (K/P, *KIT*-mutated or *PDGFRA*-mutated; WT, WT for *KIT* and *PDGFRA*; K, *KIT*-mutated; P, *PDGFRA*-mutated). The last series of plots shows the positivity for immune cell markers in the cohort of *KIT*-mutated tumors, grouped according to location (IN, intestine; ST, stomach). The ordinate indicates median number of positive cells per high-powered field. The bar indicates the median value. The Mann-Whitney *U* test was used to compare groups and the *P* value is indicated. GISTs, gastro-intestinal stromal tumors.

In line with IHC results on the explorative cohort, ssGSEA indicated that the relative degree of immune infiltration was influenced by tumor genotype. In particular, *K/P*-mutated GISTs featured higher “immunoscores” than *K/P* WT tumors (Mann-Whitney *U* test: IIS, $P = 0.021$; TIS, $P = 0.006$; CYT, $P < 0.001$) (Supplemental Table 4). Given the intrinsic difference in the biology of gastric and intestinal GISTs (24),

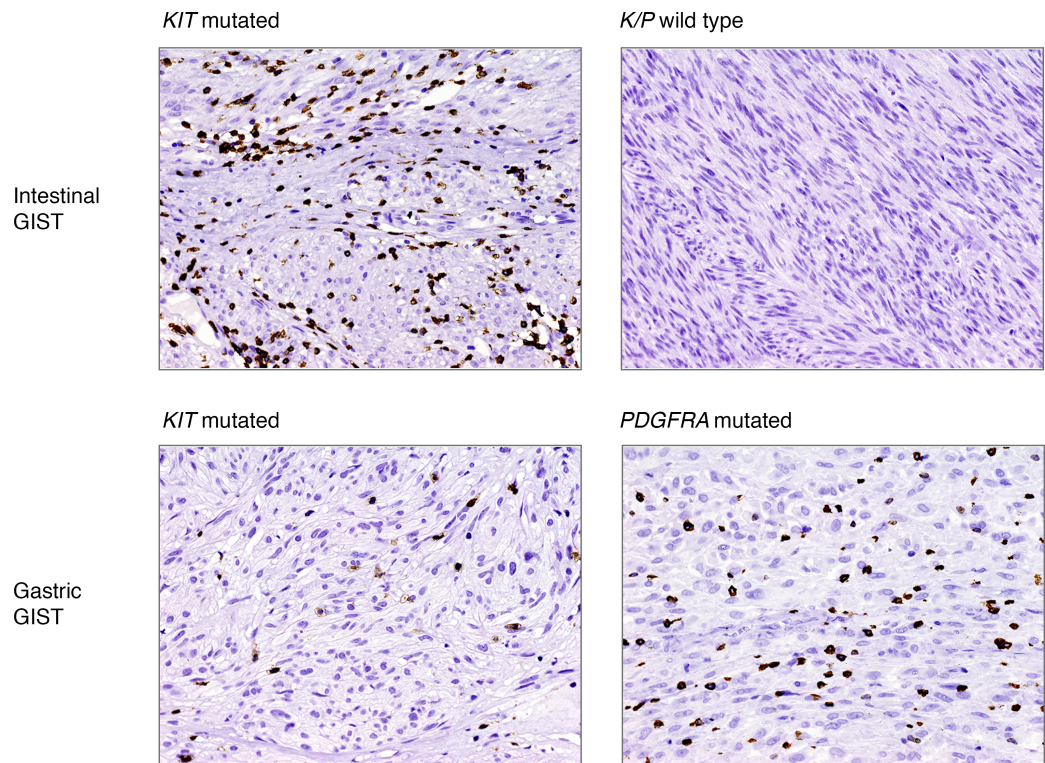


Figure 2. Representative CD3 T cells immunostainings. Location and genotypes are indicated (original magnification, $\times 20$).

we then investigated the immune infiltration correlates for the 2 locations separately (Figure 4B). In the intestine, a higher degree of infiltration was observed in *KIT*-mutated tumors versus *KIT* WT tumors (IIS, $P = 0.008$; TIS, $P = 0.006$; CYT, $P = 0.001$). The same held true for *PDGFRA*-mutated GISTs versus *KIT*-mutated GISTs in the stomach (IIS, $P = 0.005$; TIS, $P = 0.041$; CYT, $P = 0.004$). In both instances, immune infiltration was significantly associated with a greater number of cytotoxic, Th1, $T\gamma\delta$, and activated dendritic cells as well as higher APM and IFN- γ scores (Supplemental Table 4).

The interrogation of the transcriptome with deconvolution approaches (CIBERSORT and MCP-counter) yielded results coherent with ssGSEA. The cumulative scores obtained with CIBERSORT and MCP-counter showed a trend of colinearity with IIS, TIS, and CYT: all algorithms indicated that *K/P* WT tumors were colder in general compared with *K/P*-mutated GISTs as they were mini-GISTs compared with overt GISTs (Figure 5A). Moreover, in line with IHC data, an analysis focused on *KIT*-mutated genotypes highlighted the influence of tumor location on susceptibility to immune infiltration: intestinal *KIT*-mutated tumors featured higher infiltration scores compared with the gastric *KIT*-mutated counterpart (Supplemental Figure 1B and Supplemental Table 4). These patterns were observed in both the whole tumor series, including 64 archival FFPE and 13 frozen specimens, and the sole FFPE subset of samples (64 of 77), indicating that the type of processing did not bias the outcome of these analyses (compare Figure 5A and Supplemental Figure 2). Finally, CIBERSORT, which also estimates the relative proportion of the different immune types within each sample, indicated that M2 macrophages and T cells (particularly CD8 and CD4 memory resting) were the most abundant immune populations, irrespective of tumor site or mutation status (Figure 5B).

Overall, the consistency of RNA-sequencing-based analyses with IHC data supported the robustness of transcriptome-based assessment of immune infiltration and corroborated the notion that tumor genotype, malignant potential, and location impinge upon the GIST immunophenotype.

To address the potential susceptibility to immunomodulatory-based treatments of the different genotypes, we took advantage of immunophenoscore (IPS) (25), a machine learning-based classifier based on the expression of *HLA* genes, immunomodulators, and effector and suppressor cells. This scoring algorithm has proved to be effective in predicting the relative sensitivity to immune checkpoint inhibitors in diverse tumor contexts (25). After having grouped GISTs according to the driver gene, a gradient in IPS scores was observed, with *K/P* WT tumors featuring the lowest IPS values and *PDGFRA*-mutated tumors the highest

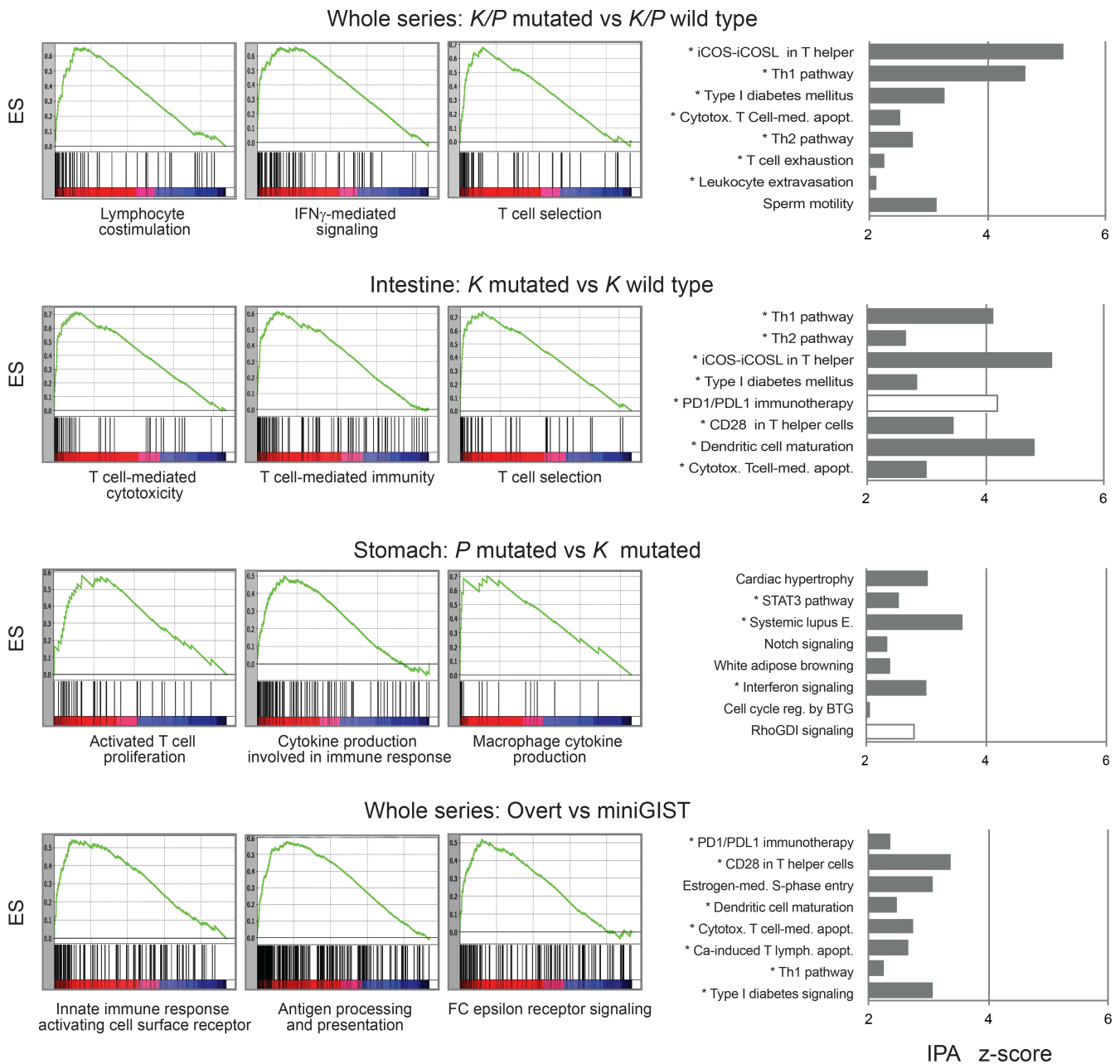


Figure 3. Transcriptional assessment of immune infiltration. The tumor series ($n = 77$ cases) was dichotomized into contrast groups as indicated and the differentially expressed genes were interrogated for immune signatures by using GSEA and IPA. The panels on the left show representative GSEA outputs (GO biological process) with associated ESs. The histograms on the right show the top 8 most significant IPA canonical pathways and associated z scores. Pathways strictly related to immunity are indicated by an asterisk. White bars indicate negative z scores. GSEA, gene set enrichment analysis; IPA, Ingenuity Pathway Analysis; ESs, enrichment scores.

IPS values (Figure 5C). These trends were supported by a differential expression in *HLA* and immune checkpoint molecules (Figure 5D). Thus, among GISTs, *K/P* WT tumors would be less likely to benefit from immune checkpoint blockade approaches.

Role of driver mutations as neoantigens. Given the observed effect of genotype on tumor immunophenotype, we sought to address the theoretical neoantigenic capacity of epitopes generated by the mutated driver gene. Neoantigen prediction algorithms, although far from being precise, may provide hints on the potential binding to patient-matched *HLA* allelotype of peptide sequences spanning the corresponding driver mutation. In this context, NetMHCpan (26) is the one of the most widely used tools. NetMHCpan predicted that almost all mutations yielded at least one peptide capable of binding, with different strengths, a cognate *HLA* allele (Supplemental Table 5).

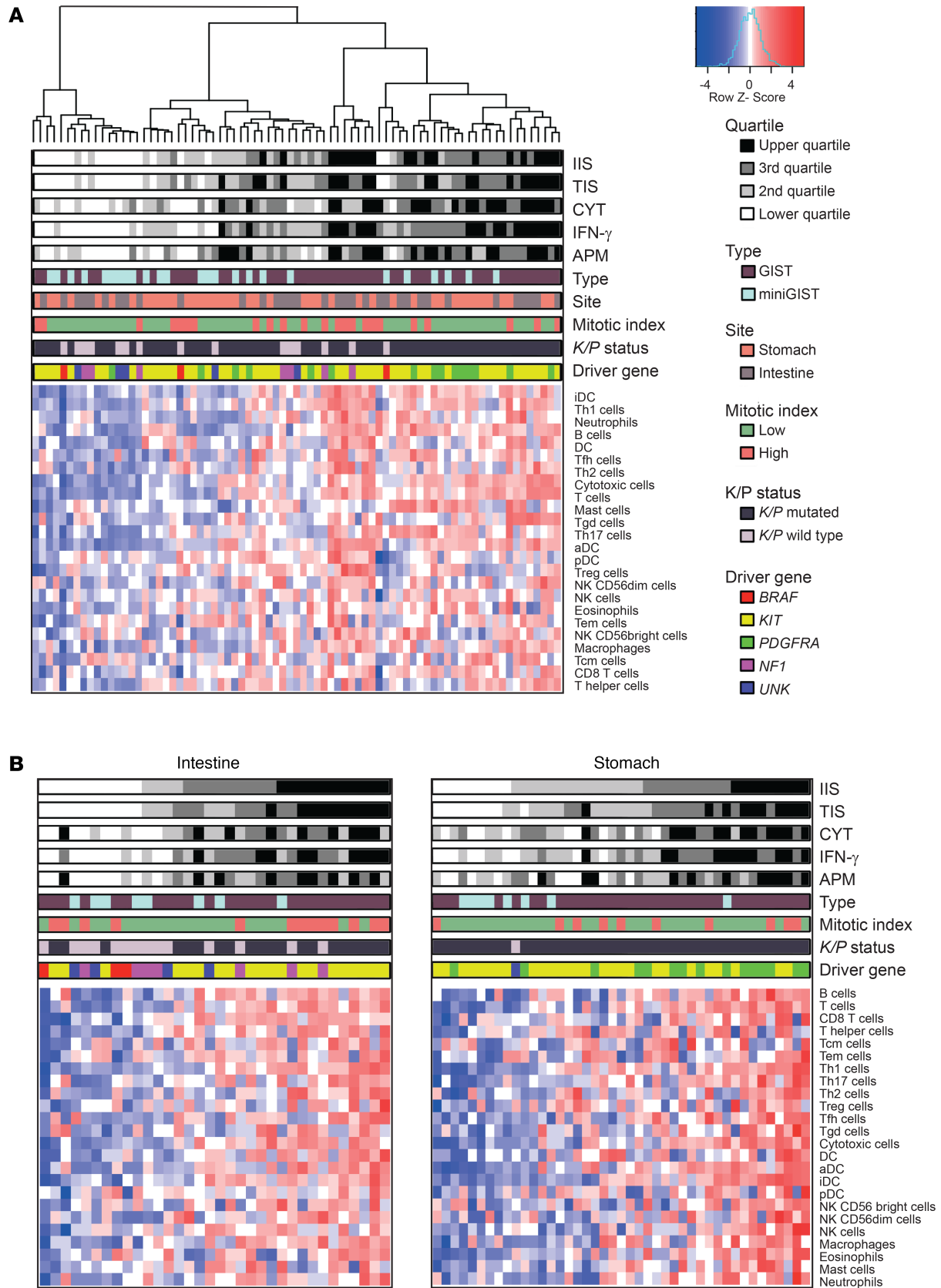


Figure 4. ssGSEA highlights a heterogeneous pattern of immune infiltration in GIST. (A) Unsupervised clustering analysis of the whole GIST series ($n = 77$) based on ssGSEA scores of 24 immune cell types. Hierarchical clustering identifies 3 major groups with different extent of immune infiltration. IIS, TIS, CYT, IFN- γ , and APM scores are reported as quartiles. **(B)** ssGSEA in intestinal ($n = 34$) and gastric ($n = 43$) sites analyzed separately highlights the impact of driver gene and malignant potential in immune infiltration. Samples are ordered according to increasing IIS. UNK, driver mutation unknown; ssGSEA, single-sample gene set enrichment analysis; IIS, immune infiltration score; TIS, T cell infiltration score; CYT, cytolytic activity score; APM, antigen-presenting machinery.

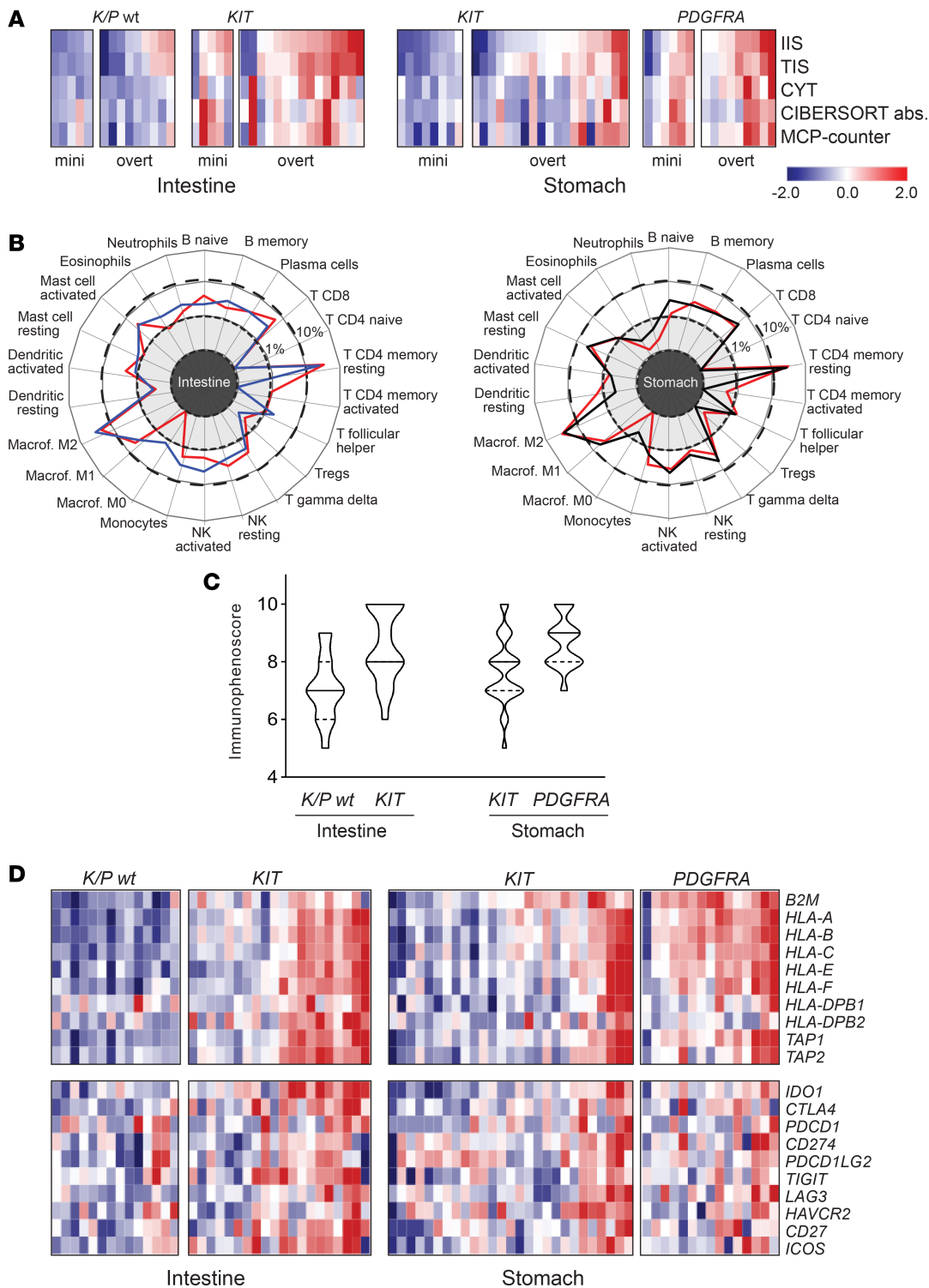


Figure 5. Dissection of genotype, location, and malignant potential in GIST immunogenicity. (A) Heatmap of the immune infiltration scores calculated with the indicated algorithms. Color-coded z scores for IIS, TIS (ssGSEA), CYT, CIBERSORT absolute (abs), and MCP-counter cumulative scores are shown. Samples are grouped according to tumor location, genotype, and malignant potential (miniGIST and overt GIST). (B) Relative proportion of the different immune cell types in intestinal (left) and gastric GIST (right) calculated by CIBERSORT. Mean proportion values (%) of the different cell types were calculated per each genotype (red line, *KIT*-mutated GIST; black line, *PDGFRA*-mutated GIST; blue line, *K/P WT* GIST) and reported in a radar plot. Dotted and dashed lines mark 1% and 10%, respectively. Macrophages M2 and T CD4 memory resting are the most represented immune cell types in both sites. (C) Violin plot showing the immunophenoscore of intestinal and gastric GISTs arranged according to genotype. The solid line indicates the median value; dashed lines indicate upper and lower quartiles. (D) Heatmap of APM genes and immune modulators in intestinal and gastric GIST. Data are presented as color-coded z scores calculated on log₂TPM of the whole series (for color coding scale, see A). IIS, immune infiltration score; TIS, T cell infiltration score; ssGSEA, single-sample gene set enrichment analysis; CYT, cytolytic activity score; APM, antigen-presenting machinery.

Because the immunogenic efficacy of a neoantigen is also affected by the extent of expression of the mutated peptide, we compared the expression levels of the driver genes. Whereas *KIT* and *PDGFRA* transcripts were robustly expressed (median TPM: *KIT*, 3239; *PDGFRA*, 2097), *BRAF* was expressed at lower levels (median TPM: 89), which probably influenced its immunogenic power (Supplemental Table 5). Regarding *NF1*, not only was this gene moderately expressed (median TPM 75) but also its alterations were typically frameshift or nonsense mutations that elicited nonsense-mediated mRNA decay. Accordingly, neurofibromin is barely detectable/absent in *NF1*-mutated GISTs (27). Thus, *NF1* mutations are unlikely to yield immunogenic peptides.

Pathways involved in differential immune colonization. To gain further insights into the mechanisms implicated in shaping GIST immunogenicity, we compared highly and poorly infiltrated tumors by performing GSEA with the MSigDB Hallmark pathway collection. As expected, signaling cascades related to immunity were markedly enriched in high IIS tumors, whereas low IIS tumors featured a trend of enrichment for the Hedgehog (HH) pathway (enrichment score [ES] = 0.52, $P = 0.05$) (Figure 6A and Supplemental Table 6). The enrichment of the HH pathway was particularly evident in poorly infiltrated GISTs located in the intestine (ES = 0.59, $P = 0.01$), where the WNT/ β -catenin signaling (WNT/ β -cat) also tended to be more represented, although not reaching statistical significance (ES = 0.53; $P = 0.24$) (Figure 6A).

As a complementary approach to address the potential involvement of these pathways in scarcely infiltrated intestinal GISTs, HH and WNT/ β -cat pathway activation scores were calculated using both the MSigDB gene sets and 2 other, nonoverlapping HH (28) and WNT/ β -cat (29) minimal activation signatures. In both instances, the degree of immune infiltration (IIS) inversely correlated with HH and WNT/ β -cat pathway activation scores (HH MSigDB, $r = -0.53$, $P = 0.02$; HH minimal signature, $r = -0.49$, $P = 0.003$; WNT/ β -cat MSigDB, $r = -0.43$, $P = 0.01$; WNT/ β -cat minimal signature, $r = -0.36$, $P = 0.038$) (Figure 6B and Supplemental Figure 3A).

Furthermore, an RNA-editing event affecting *GLII*, a major HH downstream target, was detected in a poorly infiltrated *K/P* WT GIST. This rare RNA-editing phenomenon, consisting in an RNA-only nucleotide variation that determines Arg to Gly amino acid change, is known to induce constitutive HH pathway activation (30), thus adding further support to the implication of HH in immune cold GISTs (Supplemental Figure 3B).

HH and WNT/ β -cat are highly intertwined signaling routes that have been reported to be associated with phenomena of tumor immune exclusion (31–33). Therefore, the activation of HH and WNT/ β -cat pathways might impair GIST immune cell colonization by eliciting immune evasion. Intriguingly, both pathways appear to be positively regulated by the RAS/RAF/MAPK pathway (34–36) and activation of the RAS pathway has also been associated with immune suppression (37, 38). Thus, RAS, HH, and WNT/ β -cat might cooperate to dampen the immunogenicity of *K/P* WT intestinal GISTs.

Finally, we were intrigued by the increased expression of *ANO1* (also known as *DOG1* or *TMEM16A*) in poorly infiltrated GISTs (log₂FC 0.5; FDR < 0.01). *ANO1*, commonly used as a GIST marker (39, 40), encodes anoctamin-1, an anion exchange molecule that has been recently implicated in chemokine/cytokine secretion (41, 42). We found that *ANO1* inversely correlated with the extent of immune infiltration, an anticorrelation that was particularly evident in the tumors of gastric location (*ANO1*/IIS, whole series: $r = -0.58$, $P = 3.4 \times 10^{-8}$; stomach: $r = -0.73$, $P = 2.0 \times 10^{-7}$; intestine: $r = -0.36$, $P = 0.04$) (Figure 6, C and D, and Supplemental Figure 4). The negative correlation between *ANO1* and immune infiltration in GISTs was also confirmed in an independent, publicly available gastric GIST cohort (43) (E-MTAB-373: *ANO1*/IIS, $r = -0.46$, $P = 0.003$).

To gain further insights on the role of *ANO1* in GIST immune colonization, we interrogated the list of genes described as differentially expressed following *ANO1* silencing in GIST-T1 cells (44). Although the limited size of this data set prevents definitive conclusions, overrepresentation analysis indicated enrichment for immune-related signatures (Figure 6E). Overall, these data point to a possible role for anoctamin-1 in modulating tumor immune infiltration.

Discussion

Recent evidence indicates that tumor-infiltrating immune cells populate the microenvironment of GISTs. A number of IHC studies demonstrated the presence of lymphocytes and macrophages, with some evidence of correlation with disease progression and response to tyrosine kinase inhibitors (10, 11, 13, 14). A broader approach was undertaken by Vitiello and coworkers (12), who combined transcriptional profiling, IHC, and flow cytometry to investigate in deeper detail the immune microenvironment of a large GIST cohort (75 samples). Compared with *KIT*-mutated tumors, *PDGFRA* mutant GISTs were found to feature a greater extent of immune infiltration and cytolytic activity, which were associated with increased levels of chemokines and a greater number of mutation-derived high-affinity neoepitopes. This study primarily focused on gastric tumors, and it included a limited number of small intestinal GISTs (6 of 75) as well as rare genotypes (*NF1* and *BRAF*). Moreover, the series analyzed included both primary and metastatic lesions, naive and treated tumors. Thus, the innate propensity of GISTs to immune infiltration and what tumor-specific factors affect this phenomenon, particularly in uncommon entities such as *K/P* WT tumors, remain to be fully clarified.

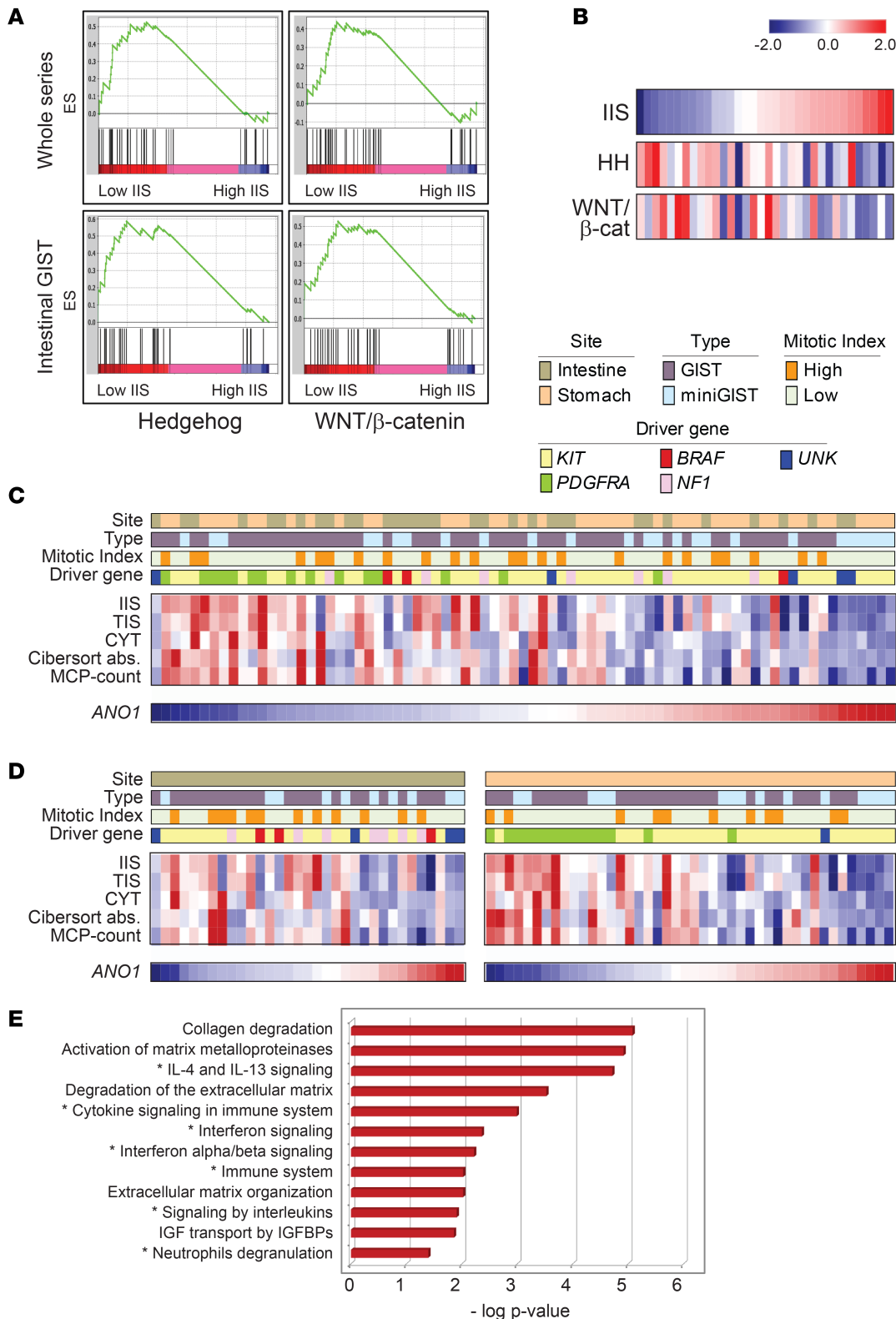


Figure 6. Pathways involved in poorly infiltrated GISTs. (A) GSEA analyses indicating the enrichment of HH and WNT/β-catenin MSigDB hallmark signatures in immune cold GIST (low IIS), compared with immune hot GIST (high IIS) in the whole series (top) and in the intestinal subset (bottom). (B) Anticorrelation of IIS with HH and WNT/β-catenin activation scores (MSigDB Hallmark) in intestinal GIST. Color-coded z score values are displayed. (C and D) Negative correlation between *ANO1* gene expression and immune infiltration scores in the whole series of 77 cases (C) and in intestinal and gastric GIST, separately (D). Site, type, mitotic index, and driver gene are as per indicated color-coded labels. z Score scale is as in B. (E) Reactome pathway analysis of the genes differentially expressed following *ANO1* silencing in GIST-T1 cells. The top most statistically significant pathways (-log P value, hypergeometric test) are shown. Immune-related pathways are indicated by an asterisk. The input gene list was from ref. 44. UNK, driver mutation unknown; GSEA, gene set enrichment analysis; HH, Hedgehog; IIS, immune infiltration score.

Bearing this in mind, we sought to specifically address GIST-intrinsic immunogenicity by focusing on primary imatinib-naïve tumors. Our cohort was assembled in a way that the major clinical-pathological and molecular variables affecting GIST biology were well represented. This allowed us to unveil that genotype, location, and malignant potential concur to shape GIST immune contexture.

The presence of intratumor immune infiltrate was demonstrated by integrating immunohistochemical, transcriptomic, and computational approaches. Macrophages and T lymphocytes appeared as the most common infiltrating elements, in line with other studies (11–14, 45).

GISTs with limited malignant potential (miniGISTs) tended to be less infiltrated compared with overt GISTs. This suggests that, in the early phases of development, despite the gain of oncogenic *KIT* or *PDGFRA* mutations, GISTs are relatively immunogenically “silent” and colonization by immune cells somehow accompanies malignant progression.

The role of tumor genotype clearly emerged in both in situ and omics analyses. Specifically, *K/P* WT GISTs turned out to be less infiltrated than *K/P*-mutated tumors.

Several factors probably contribute to the reduced infiltration observed in *K/P* WT tumors. These were primarily intestinal GISTs carrying *BRAF* and *NF1* mutations, and in silico predictions indicated that these mutations were likely had a more limited, if any (see *NF1*-inactivating mutations), neoantigenic potential compared with *KIT/PDGFR*A mutations. In addition, the major dependency of these genotypes on the RAS pathway may play a role in lowering immune colonization. In fact, RAS/RAF/MAPK is the main signaling route in *BRAF* and *NF1*-mutated tumors, whereas in *K/P*-mutated GISTs the activated kinase signals, with variable intensities, through multiple pathways (PI3K/AKT/mTOR, STAT, RAS/RAF/MAPK) (1). The activation of the RAS pathway has been shown to correlate with inhibition of *IFN γ* and *HLA* gene expression, thus lessening lymphocyte infiltration and promoting immune evasion (37, 38, 46).

More interestingly, we found that poorly infiltrated intestinal GISTs featured a peculiar activation of HH and WNT/ β -cat pathways. These are 2 highly interconnected and reciprocally regulated pathways. Both intersect the RAS/RAF/MAPK pathway (34, 47) and have been implicated in the pathogenesis of RAS-driven tumors (47–50), including GISTs (51–53). Intriguingly, HH and WNT/ β -cat pathways are known to induce immune exclusion: HH suppresses T cell recruitment by inhibiting CXCL9 and CXCL10 production (*CXCL10* was indeed significantly downregulated in *K/P* WT intestinal tumors), and WNT/ β -cat activation has been correlated with refractoriness to immune checkpoint blockers (31–33, 54–56).

Taken together, these findings suggest that RAS, HH, and WNT/ β -cat likely concur to induce an immune silent phenotype to *K/P* intestinal WT tumors.

In the gastric GIST subset, immune infiltration tended to be greater in *PDGFRA*-mutated tumors compared with *KIT*-mutated tumors, in line with previous findings (12). Higher levels of expression of a set of cytokines that can contribute to the recruitment and activation of immune cells were observed in *PDGFRA*-mutated GISTs. In particular, as reported by Vitiello and coworkers (12), these tumors featured elevated expression levels of *CXCL14*, a cytokine that promotes immune surveillance through recruitment of DC, NK, and CD8 T cells and upregulates *HLA* expression (57). In addition, we observed higher levels of immune-attractant *CCL2* and *CCL4*. *CCL2* has a major role in the recruitment of myeloid cells to tumor site and it has been recently implicated in GIST macrophage infiltration (58). Interestingly, PDGFR pathway activation has been shown to induce *CCL2* upregulation in different settings (59–61). The activation of the PDGF pathway has also been shown to induce *IL33* via *SOX7* (62). Accordingly, *SOX7* was overexpressed in *PDGFRA*-mutated versus *KIT*-mutated gastric GISTs, together with *IL33* and *IL15*. Both *IL33* and *IL15* can potentiate innate or adaptive immune responses by recruiting and stimulating T or NK cells, respectively (63). Thus, the higher level of immune colonization observed in *PDGFRA*-mutated GISTs seem to relate to the activation of the PDGF pathway.

Besides being less infiltrated than the *PDGFRA*-mutated counterpart, *KIT*-mutated gastric GISTs also featured a lower extent of immune infiltration and reduced expression of immunomodulatory cytokines when compared with intestinal GISTs with the same genetic background (*KIT* mutation). This may be due to the specific anatomic microenvironment, but it is also possible that cell-intrinsic factors may be implicated in the differential immune colonization observed in gastric versus intestinal *KIT*-mutated GISTs. In this regard, interstitial cells of Cajal, considered the bona fide cell of origin of GISTs, show distinctive features depending on location, including the expression of cytokines (64).

Finally, an unprecedented finding was the inverse correlation observed between immune infiltration and *ANO1* expression, particularly in the tumors of gastric location. *ANO1*-encoded protein, anoctamin-1, is typically expressed by GISTs, with a diffuse staining pattern generally stronger in *KIT*-mutated and *NF1*-mutated tumors (39, 40). Anoctamin-1 is a calcium-activated anion channel whose chemical inhibition affects GIST cell proliferation and viability (65). Recent evidence also implicates this molecule in chemokine signaling (41). In particular, anoctamin-1 has been shown to suppress the release of proinflammatory cytokines, thus hindering the innate immune response (66, 67). Accordingly, preliminary data suggest that *ANO1* silencing in GIST cells alters the expression of genes involved in immune-related pathways. Overall these data support the notion that anoctamin-1 may play a role in tuning GIST immunogenicity, particularly in the gastric subset.

What are the clinical implications of this study? Although imatinib and other tyrosine kinase inhibitors are active in controlling tumor recurrence and progression in patients with advanced disease, still these treatments are hardly curative. Moreover, *K/P* WT tumors currently lack targeted therapies. Our results suggest that a significant fraction of *K/P*-mutated GISTs might benefit from immune-based approaches. Specifically, the evidence of immune colonization by cytotoxic cells and a proficient APM together with the expression of molecules with immune-suppressive functions suggest that immune checkpoint-based therapies may unleash an intrinsic antitumor response in these tumors.

In contrast, *K/P* WT GISTs, in particular *BRAF* and *NF1*-mutated GISTs, were found to be essentially immune silent and hence less likely to benefit from immune checkpoint blockade approaches. The major dependency of these tumors on the RAS pathway may represent a therapeutic opportunity, even in an immunomodulatory perspective. Indeed, the combination of MEK and immune checkpoint inhibitors proved to enhance antitumor immune response in mouse models of RAS-driven cancers (38, 68), and promising results are being achieved in clinical trials with analogous combinations (38).

More interestingly, our study unveiled a therapeutic vulnerability, namely, the implication of HH and WNT/ β -cat immune-excluding pathways. In mouse models, chemical inhibition of the HH signaling has been shown to increase the recruitment of cytotoxic cells into tumor and dampen immune-suppressive innate and adaptive response (32). Moreover, combinatorial treatments of immune checkpoint inhibitors with either HH or WNT/ β -cat signaling blockade have demonstrated synergistic effects in diverse tumor settings (32, 69–72). Thus, the targeting of HH or WNT/ β -cat pathways in poorly infiltrated GISTs, in particular *K/P* WT GISTs, may represent a treatment avenue by both inhibiting intrinsic protumor oncogenic signals and alleviating immune suppression, harnessing the immune system to an antitumor attack.

Finally, the intriguing correlation between *ANO1* expression and degree of immune infiltration points to an additional possible element of vulnerability. Several compounds have demonstrated inhibitory activity toward anoctamin-1, including FDA-approved drugs (65, 73, 74), and chemical inhibition of anoctamin-1 has been shown to affect GIST cell proliferation and survival (65). It would be interesting to evaluate the effect of these compounds on cytokine secretion. Definitively, the implication of *ANO1* in tempering GIST immunogenicity is unprecedented and deserves further investigations.

Methods

Samples. Eighty-two adult cases of primary untreated GISTs were retrieved from the pathology files of the authors' institutions and reviewed by 3 sarcoma expert pathologists. GIST diagnosis was based on morphology, IHC for CD117 (KIT) and ANO1 (aka DOG1 or TMEM16A), and exclusion of other entities within the differential diagnosis. The series included 57 overt GISTs (≥ 2 cm; any mitotic index) and 25 miniGISTs, i.e., very low-risk tumors with low mitotic index (≤ 5 mitoses in 5 mm²) and small size (< 2 cm). Risk of relapse was calculated according the revised version of Joensuu risk classification (75).

Mutation analysis. DNA extraction and mutation analysis were essentially as previously described (4). Briefly, DNA was extracted from tissue sections with a tumor cellularity greater than 70%. Samples were first profiled for *KIT* (exons 9, 11, 13, and 17) and *PDGFRA* (exons 12, 14, and 18) mutations by Sanger sequencing. Samples scoring negative in this analysis were further profiled by using a targeted NGS panel that covered the whole coding sequence of *KIT*, *PDGFRA*, *BRAF*, *NF1*, *SDH A-D*, *H/K/N RAS*. The allele frequency of the mutation was greater than or equal to 30%. *SDH* deficiency was also assessed by *SDHB* immunostaining.

IHC analysis of immune infiltrate. Thirty-eight samples were evaluated for evidence of immune cell infiltration. To this end, samples were stained for CD3, CD4, CD8, FOXP3 (T cells), CD20 (B cells), CD68 (macrophages), and immune checkpoint molecules PD1/PDCD1 and PDL1/CD274. The number of positive cells was determined by counting 15 random high-power fields (HPFs) ($\times 400$) in a double-blinded fashion and expressed as median value per HPF. Further details are provided in Supplemental Methods.

Transcriptome analysis of immune infiltration. Transcriptional profiling was performed on a series of 77 GISTs, including 64 FFPE and 13 fresh-frozen samples. RNA purification, library preparation, and bioinformatic data analysis are described in detail in Supplemental Methods. Briefly, reads were first checked for quality using FastQC and MultiQC (v1.7) (76). Adapter removal and clipping was done with Trimmomatic (v0.38) (77). Samples reads were aligned against Homo sapiens genome assembly GRCh38 (hg38) with STAR (v2.7.0e) (78). SAMtools (v1.9) (79) was used for merging aligned files. Gene counts were obtained with Cufflinks (v2.2.1) (80). DESeq2 (v3.3) (81) was used for the identification of differentially expressed genes (DEGs).

Pathway analyses were performed on DEGs using IPA (QIAGEN; www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis/) and Reactome (82). GSEA (83) was run on either normalized counts or TPM using Gene Ontology (c5.bp.v7.0) and the MSigDB Hallmark collection of molecular signatures (h.all.v7.0).

The estimate of immune cell infiltration from transcriptome data was performed by using diverse computational methods including ssGSEA (21) and deconvolution approaches, namely, CIBERSORT (84) and MCP-counter (85). IPS (25) and CYT (23) were calculated as previously described.

NetMHCpan (v4.0) (26) was used to predict the binding of *KIT*, *PDGFRA*, *BRAF*, and *NF1* mutant peptides to the patient-matched *HLA* class I alleles. PHLAT (v1.0) (86) was employed for determining patient-matched *HLA* alleles. The most common *HLA* alleles in the Italian population (87) were used in 4 cases in which PHLAT typing failed.

HH and WNT/ β -cat pathway activation scores were calculated by averaging (geometric mean) \log_2 -transformed TPM values of the genes composing the corresponding MSigDB hallmark signatures (h.all.v7.0) as well as by using 5-gene HH minimal signature reported by Shou et al. (28) and the 16-gene WNT/ β -cat signature reported by Chang et al. (29). See Supplemental Methods for further details.

Data availability. Raw RNA-sequencing data are accessible at the NCBI-SRA database (<https://www.ncbi.nlm.nih.gov/sra>, accession PRJNA637476).

Statistics. Statistical analyses were performed by SigmaPlot 12.0 (SYSTAT). Correlation coefficients (r) were calculated using the Spearman's rank method. The Mann-Whitney U rank-sum test was used to compare groups. Statistical threshold was set at P values less than or equal to 0.05.

Study approval. The study was performed in compliance with relevant laws and institutional guidelines and was approved by the CRO institutional review board (IRB-04-2017) and by the Marca Ethical Committee (N. 456/CE). Written informed consent was obtained from all patients.

Author contributions

DG acquired, analyzed, and interpreted the data and drafted and edited the final draft. MS and SR acquired and analyzed the data and provided critical revision of the manuscript. DB provided bioinformatic and statistical data analyses. MB, AM, FN, DR, and MC acquired and analyzed the data. APDT conceived the study and design; analyzed and interpreted the data; acquired funding; and provided critical revision of the manuscript. RM conceived the study and design; supervised the study; analyzed and interpreted the data; acquired funding; and drafted and edited the final draft. All authors read and approved the final version of the manuscript.

Acknowledgments

This work was funded by the Italian Ministry of Health, Ricerca Finalizzata, Associazione Italiana Ricerca sul Cancro (AIRC), CRO Aviano National Cancer Institute IRCCS Intramural grant.

Address correspondence to: Roberta Maestro, Centro di Riferimento Oncologico (CRO Aviano) IRCCS, National Cancer Institute, Unit of Oncogenetics and Functional Oncogenomics, Via Gallini 2, 33081 Aviano (PN) Italy. Phone: 39.0434.659.670; Email: maestro@cro.it.

1. Corless CL, Barnett CM, Heinrich MC. Gastrointestinal stromal tumours: origin and molecular oncology. *Nat Rev Cancer*. 2011;11(12):865–878.
2. Miettinen M, Lasota J. Histopathology of gastrointestinal stromal tumor. *J Surg Oncol*. 2011;104(8):865–873.
3. Rossi S, et al. KIT, PDGFRA, and BRAF mutational spectrum impacts on the natural history of imatinib-naïve localized GIST: a population-based study. *Am J Surg Pathol*. 2015;39(7):922–930.
4. Gasparotto D, et al. Quadruple-negative GIST is a sentinel for unrecognized neurofibromatosis type 1 syndrome. *Clin Cancer Res*. 2017;23(1):273–282.
5. Brenca M, et al. Transcriptome sequencing identifies ETV6-NTRK3 as a gene fusion involved in GIST. *J Pathol*. 2016;238(4):543–549.
6. Shi E, et al. FGFR1 and NTRK3 actionable alterations in “Wild-Type” gastrointestinal stromal tumors. *J Transl Med*. 2016;14(1):339.
7. Casali PG, et al. Gastrointestinal stromal tumours: ESMO-EURACAN Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann Oncol*. 2018;29(Suppl 4):iv68–iv78.
8. Dhillon S. Avapritinib: First Approval. *Drugs*. 2020;80(4):433–439.
9. Smith BD, et al. Ripretinib (DCC-2618) is a switch control kinase inhibitor of a broad spectrum of oncogenic and drug-resistant

- KIT and PDGFRA variants. *Cancer Cell*. 2019;35(5):738–751.e9.
10. van Dongen M, et al. Anti-inflammatory M2 type macrophages characterize metastasized and tyrosine kinase inhibitor-treated gastrointestinal stromal tumors. *Int J Cancer*. 2010;127(4):899–909.
 11. Rusakiewicz S, et al. Immune infiltrates are prognostic factors in localized gastrointestinal stromal tumors. *Cancer Res*. 2013;73(12):3499–3510.
 12. Vitiello GA, et al. Differential immune profiles distinguish the mutational subtypes of gastrointestinal stromal tumor. *J Clin Invest*. 2019;129(5):1863–1877.
 13. Tan Y, Garcia-Buitrago MT, Trent JC, Rosenberg AE. The immune system and gastrointestinal stromal tumor: a wealth of opportunities. *Curr Opin Oncol*. 2015;27(4):338–342.
 14. Cameron S, Gieselmann M, Blaschke M, Ramadori G, Füzesi L. Immune cells in primary and metastatic gastrointestinal stromal tumors (GIST). *Int J Clin Exp Pathol*. 2014;7(7):3563–3579.
 15. Balachandran VP, et al. Imatinib potentiates antitumor T cell responses in gastrointestinal stromal tumor through the inhibition of Ido. *Nat Med*. 2011;17(9):1094–1100.
 16. Seifert AM, et al. PD-1/PD-L1 blockade enhances t-cell activity and antitumor efficacy of imatinib in gastrointestinal stromal tumors. *Clin Cancer Res*. 2017;23(2):454–465.
 17. Zhang JQ, et al. Macrophages and CD8⁺ T cells mediate the antitumor efficacy of combined CD40 ligation and imatinib therapy in gastrointestinal stromal tumors. *Cancer Immunol Res*. 2018;6(4):434–447.
 18. Reilly MJ, et al. Phase I clinical trial of combination imatinib and ipilimumab in patients with advanced malignancies. *J Immunother Cancer*. 2017;5:35.
 19. D'Angelo SP, et al. Combined KIT and CTLA-4 blockade in patients with refractory GIST and other advanced sarcomas: a phase Ib study of dasatinib plus ipilimumab. *Clin Cancer Res*. 2017;23(12):2972–2980.
 20. Toulmonde M, et al. Use of PD-1 targeting, macrophage infiltration, and IDO pathway activation in sarcomas: a phase 2 clinical trial. *JAMA Oncol*. 2018;4(1):93–97.
 21. Şenbabaoğlu Y, et al. Tumor immune microenvironment characterization in clear cell renal cell carcinoma identifies prognostic and immunotherapeutically relevant messenger RNA signatures. *Genome Biol*. 2016;17(1):231.
 22. Ayers M, et al. IFN- γ -related mRNA profile predicts clinical response to PD-1 blockade. *J Clin Invest*. 2017;127(8):2930–2940.
 23. Rooney MS, Shukla SA, Wu CJ, Getz G, Hacohen N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell*. 2015;160(1-2):48–61.
 24. Antonescu CR, et al. Gene expression in gastrointestinal stromal tumors is distinguished by KIT genotype and anatomic site. *Clin Cancer Res*. 2004;10(10):3282–3290.
 25. Charoentong P, et al. Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade. *Cell Rep*. 2017;18(1):248–262.
 26. Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. NetMHCpan-4.0: Improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *J Immunol*. 2017;199(9):3360–3368.
 27. Rossi S, et al. Neurofibromin C terminus-specific antibody (clone NFC) is a valuable tool for the identification of NF1-inactivated GISTs. *Mod Pathol*. 2018;31(1):160–168.
 28. Shou Y, et al. A five-gene hedgehog signature developed as a patient preselection tool for hedgehog inhibitor therapy in medulloblastoma. *Clin Cancer Res*. 2015;21(3):585–593.
 29. Chang WH, Lai AG. Pan-cancer genomic amplifications underlie a WNT hyperactivation phenotype associated with stem cell-like features leading to poor prognosis. *Transl Res*. 2019;208:47–62.
 30. Shimokawa T, et al. RNA editing of the GLI1 transcription factor modulates the output of Hedgehog signaling. *RNA Biol*. 2013;10(2):321–333.
 31. Luke JJ, Bao R, Sweis RF, Spranger S, Gajewski TF. WNT/ β -catenin pathway activation correlates with immune exclusion across human cancers. *Clin Cancer Res*. 2019;25(10):3074–3083.
 32. Hanna A, et al. Inhibition of Hedgehog signaling reprograms the dysfunctional immune microenvironment in breast cancer. *Oncimmunology*. 2019;8(3):1548241.
 33. Petty AJ, et al. Hedgehog signaling promotes tumor-associated macrophage polarization to suppress intratumoral CD8⁺ T cell recruitment. *J Clin Invest*. 2019;129(12):5151–5162.
 34. Pietrobono S, Gagliardi S, Stecca B. Non-canonical hedgehog signaling pathway in cancer: activation of GLI transcription factors beyond smoothed. *Front Genet*. 2019;10:556.
 35. Janssen KP, et al. APC and oncogenic KRAS are synergistic in enhancing Wnt signaling in intestinal tumor formation and progression. *Gastroenterology*. 2006;131(4):1096–1109.
 36. Li J, Mizukami Y, Zhang X, Jo WS, Chung DC. Oncogenic K-ras stimulates Wnt signaling in colon cancer through inhibition of GSK-3 β . *Gastroenterology*. 2005;128(7):1907–1918.
 37. Lal N, et al. KRAS mutation and consensus molecular subtypes 2 and 3 are independently associated with reduced immune infiltration and reactivity in colorectal cancer. *Clin Cancer Res*. 2018;24(1):224–233.
 38. Loi S, et al. RAS/MAPK activation is associated with reduced tumor-infiltrating lymphocytes in triple-negative breast cancer: therapeutic cooperation between MEK and PD-1/PD-L1 immune checkpoint inhibitors. *Clin Cancer Res*. 2016;22(6):1499–1509.
 39. Liegl B, Hornick JL, Corless CL, Fletcher CD. Monoclonal antibody DOG1.1 shows higher sensitivity than KIT in the diagnosis of gastrointestinal stromal tumors, including unusual subtypes. *Am J Surg Pathol*. 2009;33(3):437–446.
 40. Espinosa I, et al. A novel monoclonal antibody against DOG1 is a sensitive and specific marker for gastrointestinal stromal tumors. *Am J Surg Pathol*. 2008;32(2):210–218.
 41. Schnür A, Hegyi P, Rousseau S, Lukacs GL, Veit G. Epithelial anion transport as modulator of chemokine signaling. *Mediators Inflamm*. 2016;2016:7596531.
 42. Wanitchakool P, et al. Role of anoctamins in cancer and apoptosis. *Philos Trans R Soc Lond B Biol Sci*. 2014;369(1638):20130096.
 43. Lagarde P, et al. Mitotic checkpoints and chromosome instability are strong predictors of clinical outcome in gastrointestinal stromal tumors. *Clin Cancer Res*. 2012;18(3):826–838.
 44. Simon S, et al. DOG1 regulates growth and IGFBP5 in gastrointestinal stromal tumors. *Cancer Res*. 2013;73(12):3661–3670.

45. Pantaleo MA, et al. Immune microenvironment profiling of gastrointestinal stromal tumors (GIST) shows gene expression patterns associated to immune checkpoint inhibitors response. *Oncoimmunology*. 2019;8(9):e1617588.
46. El-Jawhari JJ, et al. Blocking oncogenic RAS enhances tumour cell surface MHC class I expression but does not alter susceptibility to cytotoxic lymphocytes. *Mol Immunol*. 2014;58(2):160–168.
47. Jeong WJ, Ro EJ, Choi KY. Interaction between Wnt/ β -catenin and RAS-ERK pathways and an anti-cancer strategy via degradations of β -catenin and RAS by targeting the Wnt/ β -catenin pathway. *NPJ Precis Oncol*. 2018;2(1):5.
48. Pelullo M, Zema S, Nardoza F, Checquolo S, Screpanti I, Bellavia D. Wnt, Notch, and TGF- β pathways impinge on hedgehog signaling complexity: an open window on cancer. *Front Genet*. 2019;10:711.
49. Luscan A, et al. The activation of the WNT signaling pathway is a hallmark in neurofibromatosis type 1 tumorigenesis. *Clin Cancer Res*. 2014;20(2):358–371.
50. Lévy P, et al. Molecular profiling of malignant peripheral nerve sheath tumors associated with neurofibromatosis type 1, based on large-scale real-time RT-PCR. *Mol Cancer*. 2004;3:20.
51. Tang CM, et al. Hedgehog pathway dysregulation contributes to the pathogenesis of human gastrointestinal stromal tumors via GLI-mediated activation of KIT expression. *Oncotarget*. 2016;7(48):78226–78241.
52. Zeng S, et al. Wnt/ β -catenin signaling contributes to tumor malignancy and is targetable in gastrointestinal stromal tumor. *Mol Cancer Ther*. 2017;16(9):1954–1966.
53. Bertucci F, Finetti P, De Nonneville A, Birnbaum D. “Wnt/ β -Catenin in GIST”-Letter. *Mol Cancer Ther*. 2018;17(1):327–328.
54. Anandappa AJ, Wu CJ, Ott PA. Directing traffic: how to effectively drive t cells into tumors. *Cancer Discov*. 2020;10(2):185–197.
55. Harding JJ, et al. Prospective genotyping of hepatocellular carcinoma: clinical implications of next-generation sequencing for matching patients to targeted and immune therapies. *Clin Cancer Res*. 2019;25(7):2116–2126.
56. Ruiz de Galarreta M, et al. β -Catenin activation promotes immune escape and resistance to anti-PD-1 therapy in hepatocellular carcinoma. *Cancer Discov*. 2019;9(8):1124–1141.
57. Westrich JA, Vermeer DW, Colbert PL, Spanos WC, Pyeon D. The multifarious roles of the chemokine CXCL14 in cancer progression and immune responses. *Mol Carcinog*. 2020;59(7):794–806.
58. Mu J, Sun P, Ma Z, Sun P. BRD4 promotes tumor progression and NF- κ B/CCL2-dependent tumor-associated macrophage recruitment in GIST. *Cell Death Dis*. 2019;10(12):935.
59. Marcelin G, et al. A PDGFR α -mediated switch toward CD9^{high} adipocyte progenitors controls obesity-induced adipose tissue fibrosis. *Cell Metab*. 2017;25(3):673–685.
60. Chen YT, et al. Platelet-derived growth factor receptor signaling activates pericyte-myofibroblast transition in obstructive and post-ischemic kidney fibrosis. *Kidney Int*. 2011;80(11):1170–1181.
61. Bethel-Brown C, Yao H, Hu G, Buch S. Platelet-derived growth factor (PDGF)-BB-mediated induction of monocyte chemoattractant protein 1 in human astrocytes: implications for HIV-associated neuroinflammation. *J Neuroinflammation*. 2012;9:262.
62. Yang Y, et al. The PDGF-BB-SOX7 axis-modulated IL-33 in pericytes and stromal cells promotes metastasis through tumour-associated macrophages. *Nat Commun*. 2016;7:11385.
63. Baker KJ, Houston A, Brint E. IL-1 Family members in cancer; two sides to every story. *Front Immunol*. 2019;10:1197.
64. Chen H, et al. Differential gene expression in functional classes of interstitial cells of Cajal in murine small intestine. *Physiol Genomics*. 2007;31(3):492–509.
65. Fröbom R, et al. Biochemical inhibition of DOG1/TMEM16A achieves antitumoral effects in human gastrointestinal stromal tumor cells in vitro. *Anticancer Res*. 2019;39(7):3433–3442.
66. Veit G, et al. Proinflammatory cytokine secretion is suppressed by TMEM16A or CFTR channel activity in human cystic fibrosis bronchial epithelia. *Mol Biol Cell*. 2012;23(21):4188–4202.
67. Dai WJ, et al. Downregulation of microRNA-9 reduces inflammatory response and fibroblast proliferation in mice with idiopathic pulmonary fibrosis through the ANO1-mediated TGF- β -Smad3 pathway. *J Cell Physiol*. 2019;234(3):2552–2565.
68. Poon E, et al. The MEK inhibitor selumetinib complements CTLA-4 blockade by reprogramming the tumor immune microenvironment. *J Immunother Cancer*. 2017;5(1):63.
69. Otsuka A, et al. Hedgehog pathway inhibitors promote adaptive immune responses in basal cell carcinoma. *Clin Cancer Res*. 2015;21(6):1289–1297.
70. Nikanjam M, Cohen PR, Kato S, Sicklick JK, Kurzrock R. Advanced basal cell cancer: concise review of molecular characteristics and novel targeted and immune therapeutics. *Ann Oncol*. 2018;29(11):2192–2199.
71. Goldsberry WN, Londoño A, Randall TD, Norian LA, Arend RC. A review of the role of Wnt in cancer immunomodulation. *Cancers (Basel)*. 2019;11(6):E771.
72. Zhao J, et al. Stromal modulation reverses primary resistance to immune checkpoint blockade in pancreatic cancer. *ACS Nano*. 2018;12(10):9881–9893.
73. Hwang SJ, Basma N, Sanders KM, Ward SM. Effects of new-generation inhibitors of the calcium-activated chloride channel anoctamin 1 on slow waves in the gastrointestinal tract. *Br J Pharmacol*. 2016;173(8):1339–1349.
74. Tradtrantip L, Namkung W, Verkman AS. Crofelemer, an antisecretory antidiarrheal proanthocyanidin oligomer extracted from *Croton lechleri*, targets two distinct intestinal chloride channels. *Mol Pharmacol*. 2010;77(1):69–78.
75. Benjamin RS, Casali PG. Adjuvant imatinib for GI stromal tumors: when and for how long? *J Clin Oncol*. 2016;34(3):215–218.
76. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*. 2016;32(19):3047–3048.
77. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114–2120.
78. Dobin A, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15–21.
79. Li H, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009;25(16):2078–2079.
80. Trapnell C, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*. 2010;28(5):511–515.
81. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550.
82. Jassal B, et al. The reactome pathway knowledgebase. *Nucleic Acids Res*. 2020;48(D1):D498–D503.

83. Subramanian A, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102(43):15545–15550.
84. Newman AM, et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods*. 2015;12(5):453–457.
85. Becht E, et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol*. 2016;17(1):218.
86. Bai Y, Wang D, Furry W. PHLAT: Inference of high-resolution HLA types from RNA and whole exome sequencing. *Methods Mol Biol*. 2018;1802:193–201.
87. Sacchi N, Castagnetta M, Miotti V, Garbarino L, Gallina A. High-resolution analysis of the HLA-A, -B, -C and -DRB1 alleles and national and regional haplotype frequencies based on 120926 volunteers from the Italian Bone Marrow Donor Registry. *HLA*. 2019;94(3):285–295.

BENZ WS: the Bologna ENZYme Web Server for four-level EC number annotation

Davide Baldazzi¹, Castrense Savojardo¹, Pier Luigi Martelli^{1,*} and Rita Casadio^{1,2}

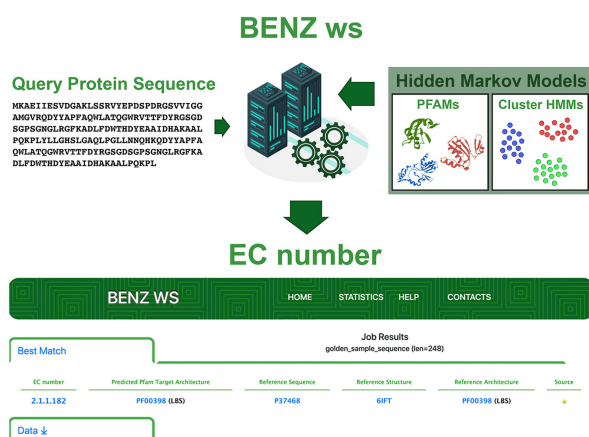
¹Biocomputing Group, Department of Pharmacy and Biotechnologies, University of Bologna, Italy and ²Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies (IBIOM), Italian National Research Council (CNR), Bari, Italy

Received March 03, 2021; Revised April 01, 2021; Editorial Decision April 15, 2021; Accepted April 20, 2021

ABSTRACT

The Bologna ENZYme Web Server (BENZ WS) annotates four-level Enzyme Commission numbers (EC numbers) as defined by the International Union of Biochemistry and Molecular Biology (IUBMB). BENZ WS filters a target sequence with a combined system of Hidden Markov Models, modelling protein sequences annotated with the same molecular function, and Pfams, carrying along conserved protein domains. BENZ returns, when successful, for any enzyme target sequence an associated four-level EC number. Our system can annotate both monofunctional and polyfunctional enzymes, and it can be a valuable resource for sequence functional annotation.

GRAPHICAL ABSTRACT



INTRODUCTION

In the post genomic era, annotating protein sequences with functional and structural features is a basic operation for bridging the gap among the hundred millions chains from different organisms, made available by deep sequencing and proteomic projects, and the much smaller number of proteins known with atomic details and with an experimentally characterised biochemical function (1, 2). The problem of functional annotation is therefore one of utmost relevance for the correct assignment of newly generated sequences to their structural and functional protein family or clan, from where they can gain some structural and functional characteristics. Indeed, the experiment Critical Assessment of Functional Annotation (CAFA) (3), since 2010, provides a large-scale assessment of computational methods developed to predict protein function as described with Gene Ontology (GO) terms, according to the three main categories, Molecular Function, Biological Process and Cellular Component (4). Yet, CAFA has no specific section on the Enzyme Commission number (EC number) prediction.

For protein enzymes, the EC number is a traditional code of the catalysed biochemical reactions, describing the relationship among the protein activity, substrates, and products. Presently, ENZYME (5) is the repository of information relative to the nomenclature of enzymes, based on the recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (https://web.expasy.org/docs/swiss-prot_guideline.html). Rhea (6), in turn, is the expert-curated knowledgebase of chemical and transport reactions of biological interest, based on the chemical dictionary ChEBI, which describes reaction participants and their transformations (<https://www.rhea-db.org/>). In Rhea, reactions are extensively curated with links to supporting literature and are mapped to other resources, including the UniProt file of each protein enzyme. Presently, the EC code includes

*To whom correspondence should be addressed: Tel: +39 0512094005; Fax: +39 0512094005; Email: pierluigi.martelli@unibo.it
Present addresses:

Davide Baldazzi, Biocomputing Group-FABIT, University of Bologna, Bologna 40126, Italy. davide.baldazzi8@unibo.it
Castrense Savojardo, Biocomputing Group-FABIT, University of Bologna, Bologna 40126, Italy. castrense.savojardo2@unibo.it
Pier Luigi Martelli, Biocomputing Group-FABIT, University of Bologna, Bologna 40126, Italy. pierluigi.martelli@unibo.it
Rita Casadio, Biocomputing Group-FABIT, University of Bologna, Bologna 40126, Italy. rita.casadio@unibo.it

seven major classes: (i) oxidoreductases, (ii) transferases, (iii) hydrolases, (iv) lyases, (v) isomerases, (vi) ligases, (vii) translocases. The EC code may range from one to four figures, when the protein catalytic activity is characterised with atomic resolution. In this case, when possible, the architecture of the catalytic site is derived from the protein structure and archived in specific databases, like M-CSA (<https://www.ebi.ac.uk/thornton-srv/m-csa>) (7), which also includes ligands.

In the UniProt reference database for protein sequences, the annotation of a protein as an enzyme is carried out whenever the automated workflow highlights specific features according to given rules (<https://www.uniprot.org/help/biocuration>). The system implements motifs derived from HAMAP (High-quality Automated and Manual Annotation of Proteins, <https://hamap.expasy.org/>), (8) and/or PROSITE, (a database of protein domains, families and functional sites, <https://prosite.expasy.org/>), (9). Feature discovering includes also the presence of motifs described in InterPro (10), which provides functional analysis of proteins by classifying them into families and by predicting domains and important sites (<https://www.ebi.ac.uk/interpro/>), and in Pfam (11), which models protein families with Hidden Markov Models (HMMs) after multiple sequence alignment (<https://pfam.xfam.org/>). Via transfer of knowledge and association rules, the enzyme gains an EC number. Eventually, manual curation allows the enzyme sequence to move from the TrEMBL to the SwissProt section of UniProt (<https://www.uniprot.org/>). EC number annotation in UniProt can include from one to four numbers, routinely depending on the annotation level of the target protein.

Other databases, by integrating different sources of information, comprising UniProt and PDB, offer a complete annotation for enzymes, such as BRENDA, (12), (<https://www.brenda-enzymes.org/index.php>) and CATH (<https://http://www.cathdb.info/>) (13). BRENDA, established in 1987, has evolved into a main collection of curated functional enzyme and metabolism data, supported by links to literature and continuously updating (12). CATH, in turn, is a free, publicly available online resource that provides information on the evolutionary relationships of protein domains. Created in the mid-1990s, it is also continuously updated. In its section FunFams, it allows the search of a target sequence and returns a functional annotation with EC number, after protein domain annotation modelled by a system of HMM hierarchical architectures. CATH is also part of InterPro and contributes therefore to the main annotation system of UniProt (<https://www.ebi.ac.uk/interpro/>).

As an alternative to transfer of knowledge, *ab-initio* computational approaches can give direct prediction/annotation of an EC number for a given input sequence or structure. This approach requires exploring the complex rules of associations among enzyme sequential and structural features and the EC codes. Methods, mainly based on different types of statistical and machine learning methods, adopt different input features, and predict EC numbers ranging from one to four levels, although with an efficiency decreasing at increasing number of levels (for an extensive review, see (14)). More recently, ECPred (15) implements an ensemble of machine learning

methods based on EC nomenclature and outperforms DEEPre, based in turn on an end-to-end feature selection and a classification model training approach (16). Both methods declare a decrease in efficiency when predicting four-level EC numbers.

A major problem in annotating EC codes remains their specificity (four-level EC codes) and the EC assignment to polyfunctional enzymes. Here, to address this problem, we develop BENZ, a system including two main sets of HMMs. One set is meant to detect sequence conservation of the target towards functional families, and the other conservation of structural architectures and family domains as described by Pfam models. The information derived from the interplay of the two different types of HMMs allows, in our case, a direct prediction of a four-level EC code for monofunctional enzymes. The system can also associate four-level EC codes to polyfunctional enzymes.

MATERIALS AND METHODS

Databases

BENZ is presently updated with UniProt/SwissProt release 2021_01. A previous version of BENZ, based on UniProt/SwissProt release 2019_11 was used in order to generate a system for CAFA-like validations. Links to Pfam (11) and KEGG (17) databases are derived from the UniProt releases. Fragments and sequences shorter than 50 residues are not considered. Annotations of active, metal, ligand-binding sites (when available) are also derived from UniProt and mapped into the Pfam architecture of the enzyme proteins.

Graph building, clustering and cluster HMM generation

The procedure stands out from a previously implemented workflow, which we adopted to generate and update our BAR 3.0 (Bologna Annotation Resource, <https://bar.biocomp.unibo.it/bar3/>), a protein functional and structural annotation resource (18). Briefly, all the UniProt sequences of a specific release (in this case, UniProtKB 2019.01) are compared with BLAST (<https://www.ncbi.nlm.nih.gov/>), and then clustered by constraining sequence identity (SI) and alignment coverage (COV, the ratio between the number of overlapping positions and the alignment length). A graph is built by connecting sequence pairs that fulfil both identity and coverage constraints. Here, we adopt $(SI) \geq 50\%$ on an alignment coverage $(COV) \geq 90\%$. Clusters are obtained by isolating the connected components of the graph. For updating, we use UniRef90 clusters (<https://www.uniprot.org/help/uniref>) which are mapped to BAR clusters, following the procedure outlined before (18). This allows the inclusion of the remaining TrEMBL sequences, and the AlignBucket algorithm (20) speeds up the alignment procedure, exploiting the constraint on COV. Each sequence in the cluster retains the annotation present in the UniProt file (PDB with the highest coverage and resolution when available, Pfam/s, KEGG links and four-level EC codes, when present). Our system allows updating (18), by adding new sequences and by reshaping clusters accordingly, with the inclusion of new annotations from UniProt.

From this background architecture, we retain only clusters containing sequences associated to four-level EC codes, particularly clusters containing SwissProt manually curated sequences, and TrEMBL sequences with an associated PDB file. For each cluster, we then trained a cluster HMM, with HMMER 3.3.2 (<http://hmmer.org/>, (20)), on the cluster specific multiple sequence alignment, as computed with Clustal Omega (21). The present version of BENZ WS, for technical reasons includes Cluster HMMs with average lengths ranging from 50 to 5000 residues, and this sets the limit of the query sequence to about 5000 residues.

Reference sequence selection and cluster HMM coloring scheme

For each cluster HMM, we select the best annotated sequence/s to be *reference sequence* for the cluster HMM-EC number/s association with the following constraints: for SwissProt sequences, chains with the highest annotation score; for TrEMBL sequences, only those with a four-level EC number and a PDB association. Each reference sequence is then associated to its specific Pfam architecture, and eventually relevant sites (including active, ligand and metal binding sites) are mapped into the corresponding Pfam/s. Cluster HMM are then grouped into two categories. GOLD cluster HMMs are univocally associated to one reference sequence, and BLUE cluster HMMs are associated to more than one reference sequence.

BENZ implementation

BENZ includes cluster HMMs and Pfam models (Pfam version 33.1). When a target sequence enters the server, it is filtered by the two different sets of models. Within the cluster HMMs, when retained (threshold for inclusion is $E\text{-value} \leq 10^{-5}$), the sequence finds a reference template; within the Pfam models, when retained (threshold for inclusion is $E\text{-value} \leq 10^{-4}$), it gains an architecture. The inclusion E-values were chosen after a self-consistency test (the prediction of the whole set of reference sequences).

This architecture is then compared to that of the reference and the target is endowed with the four-level reference EC number only when its architecture is at least equal to that of the template. If not, the four-level EC number is attributed on the basis of a common Pfam, containing relevant sites (active, metal binding, ligand binding sites). The general scheme of BENZ annotation system is depicted in Figure 1. When the retaining cluster HMM is plurivocally associated to more than one reference sequence, a dendrogram is generated after multiple sequence alignment with Clustal Omega (21), including the query sequence, which is associated with the EC code/s of the most similar among the references.

Web server

The BENZ web server interface is optimized to work with common web browsers, including Chrome 88.0, Firefox 83.0, Edge 88.0 and Safari 14.0. Upon submission, jobs are processed asynchronously adopting an internal queuing service based on Sun Grid Engine. Submitted sequences are

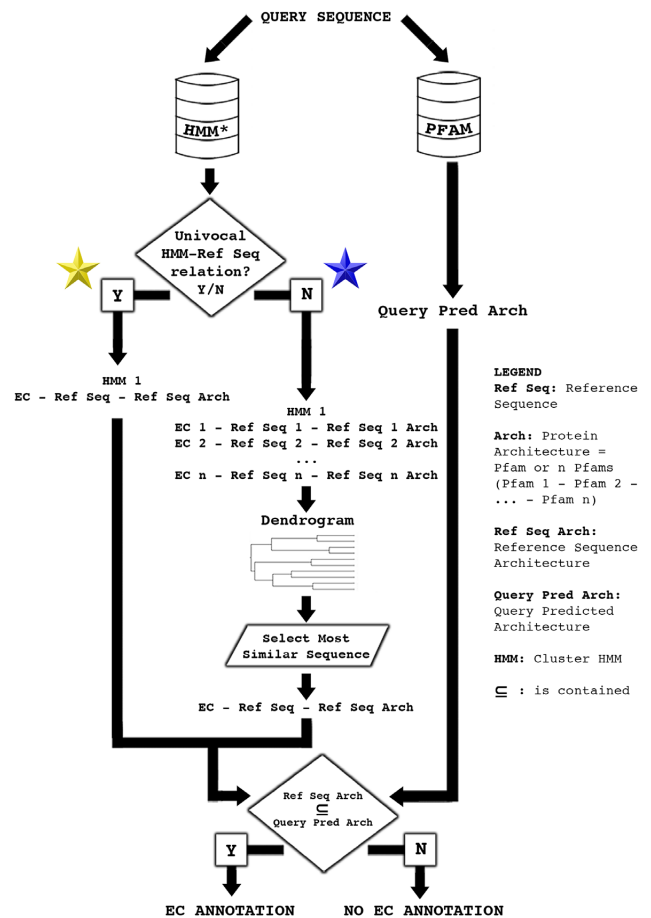


Figure 1. Workflow of BENZ WS. For a query sequence, in FASTA format, the annotation procedure starts with HMM filtering. If the retaining HMM is plurivocally associated to different reference sequences (blue star), a dendrogram is generated to find among the reference sequences the most similar one to the target. Otherwise (yellow star), the target is associated to the only reference. The EC number-query sequence association is then made after evaluating if the reference protein sequence architecture (Ref Seq Arch) is contained (\subseteq) in that of the predicted target Pfam architecture (Query Pred Arch), focusing on Pfams carrying relevant sites. Pfams in our system are annotated when possible, with the positions of the active site, ligand binding site and metal binding site (relevant sites). A sequence feature viewer allows the user to verify whether the query sequence conserves the residues relevant to the protein catalysis for validating the transfer of annotation from the reference sequence. Links to the reference sequence UniProt/SwissProt file, structure PDB file and Pfam entries, together with KEGG identifiers and pathways are also present in the output (see HELP, <https://benzdb.biocomp.unibo.it/help>).

aligned against the cluster HMMs and Pfam libraries with HMMer 3.3.2 (20). User is provided with a link to a static web page that will display results upon job completion. The page is updated every 30 s. Results are routinely returned within 1 minute since the submission. Longer times may be needed for sequences longer than 3000 residues.

When criteria described in Figure 1 are fulfilled, the result page returns the EC annotation as derived from the best matched reference sequence. The 'Best match' section also reports the PDB structure of the reference (when available), the Pfam architectures of both query and reference sequences and the type of HMM-reference association, either

univocal (GOLD star) or plurivocal (BLUE star). More details are provided in the ‘Data’ section, including the list of cluster HMMs scoring with E -value $\leq 10^{-5}$, the associated reference sequences and the links to IntEnz (<https://www.ebi.ac.uk/intenz/>), UniProt, PDB, Pfam and KEGG. Tabular data are represented with DataTables (<https://datatables.net/>), allowing to sort rows with respect to any column key and to search for text occurrences in the table. Links are resolved with the Identifiers.org service (22) to improve interoperability.

For plurivocal clusters, the dendrogram, in Newick format, representing the distances among the query and the reference sequences is computed with Clustal Omega (21) and visualized with the Bio.Phylo module of Biopython (23). The Pfam domains mapped on the query sequence are listed in the ‘Predicted Target Architecture’ table and graphically represented with tracks displayed by means of the Pviz.js library (24). Graphic view also enables to investigate the conservation of active, metal-binding, and ligand-binding sites between the query and the reference sequences. The web server is freely accessible without registration at <https://benzdb.biocomp.unibo.it>.

RESULTS

BENZ statistics

In the present version, our annotation system comprises 16 593 reference sequences (93.6% from SwissProt), from 891 organisms, included in 12 612 cluster HMMs (Table 1). Our system can annotate 5136 four-level EC numbers by means of a target-reference sequence association (Figure 1). This can be found by filtering the target with cluster HMMs and by associating the predicted target architecture to that of the reference sequence (Figure 1). When more than one reference is present in the retaining cluster HMM, a dendrogram, including the target and the cluster HMM references, allows finding the closest reference to the target. The final comparison among the predicted target architecture and the reference selected one, allows or not the association of the target with the EC number of the reference. In BENZ, 16% of the clusters HMMs (BLUE) are endowed with more than one reference sequence, including 6798 reference sequences (36% of the references, Table 1).

The reference sequence architectures comprise 4158 Pfam models, 1758 of which map relevant sites (active, ligand and metal binding) for testing the target vs reference conservation of the functional activity. 7601 reference sequences are linked to 9382 KEGG pathways.

BENZ comprises also 2023 polyfunctional reference sequences (96% from SwissProt), for a total of 1589 four-level EC numbers, included in 1485 cluster HMMs (907 GOLD and 578 BLUE). The distribution of the polyfunctional reference sequences (Supplementary Table S1S and Supplementary Figure S1S) indicates that the number of EC codes per sequence ranges from 2 to 9, following the UniProt annotation. Their associated architecture includes from one to 26 Pfam models, for a total of 1156 Pfam entries. Polyfunctional reference enzymes have relevant sites mapped into 725 Pfam entries and 1082 polyfunctional reference sequences link 2627 KEGG pathways.

BENZ at work

BENZ is tested against different protein sets (Table 2). Firstly, we run two different sets of proteins not included in our reference sequences: a positive (sequences annotated in SwissProt with a four-level EC code) and negative (sequences annotated in SwissProt without a four-level EC code). Results indicate that the system has a good efficiency in assigning four-level EC codes (92.4%) and in rejecting non-enzyme proteins (95.1%). A similar good efficiency is detected when testing two other sets, one comprising polyfunctional enzymes from SwissProt and the other including human sequences endowed with EC numbers downloaded from TrEMBL.

CAFA-like validation

The performance assessment of our method was carried out running an in-house CAFA-like benchmark. To this aim, we simulated a time-challenge experiment by computing EC annotation acquired in the time elapsed between two distant releases of SwissProt. As reference sets, we used SwissProt releases 2019_11 ($t0$: 11 December, 2019) and 2020_03 ($t1$: 17 June 2020). A BENZ test version was implemented using only sequences and annotations of the former release. Positive and negative benchmark datasets were compiled by comparing the functional annotations available in the two releases. The positive dataset consists of proteins non-annotated for EC at $t0$ but endowed with a four-level EC annotation at $t1$. Fragments were excluded. The full positive dataset therefore consists of 607 proteins not included in the ground-truth dataset of the BENZ-WS test version and endowed with a four-level EC number out of the seven main EC classes. For sake of comparison with methods not handling the EC 7 class (translocases), we considered a reduced dataset comprising 366 enzyme sequences labelled with EC codes from classes 1 to 6.

The negative dataset contains 1034 non-fragment proteins that, from $t0$ to $t1$, acquired a Gene Ontology (GO) annotation for Molecular Function (MF) different from GO:0003824 (catalytic activity) and its descendants, and that are not endowed with an EC number at any level.

We then assessed the performance of the BENZ testing version (built only on sequences and annotations available at $t0$) in discriminating enzymes from other proteins and in assigning the EC annotation. We computed different scoring measures, including the True Positive Rate (TPR), evaluating the fraction of correct predictions at different EC levels, the False Negative Rate (FNR) scoring the number of enzymes in the positive dataset predicted as non-enzymes and the False Positive Rate (FPR) scoring the number of negative proteins predicted with an EC number (Table 3).

On the full dataset (Table 3, first row), BENZ reaches FNR and FPR values of 12.2% and 3%, respectively, indicating a good ability in discriminating enzymes from other proteins. The correct EC number assignment (TPR) is equal to 85% on four-level annotations, and slightly higher when less detailed levels of annotation are considered. When the seventh Enzyme class is filtered out in the reduced data set (about 40% of the proteins), BENZ WS is still scoring with good values of FNR and FPR (second row in Table 3), highlighting the robustness of the method.

Table 1. BENZ statistics

	EC 1	EC 2	EC 3	EC 4	EC 5	EC 6	EC 7	Total
EC numbers ^a	1437	1550	1034	595	254	189	77	5136
Cluster HMM	1758	5116	3755	1006	637	636	288	12 612
Cluster HMM GOLD	1326	4315	3190	800	497	496	218	10 547
Cluster HMM BLUE	432	801	565	206	140	140	70	2065
Ref Seq ^b	2752 (390)	6455 (990)	4582 (729)	1348 (324)	842 (145)	883 (105)	405 (14)	16 593 (2023)
Ref Str ^b	1230 (152)	2252 (253)	2100 (213)	625 (110)	333 (41)	287 (22)	149 (5)	6798 (618)
Pfam ^c	682 (429)	1923 (769)	1672 (711)	463 (321)	294 (190)	276 (133)	143 (50)	4158 (1758)
KEGG ID ^d	2390	5908	3770	1185	799	894	343	14 745
KEGG Pathway ^d	2758	4812	2628	1972	952	1266	317	9382
Organisms ^e	15 158 200 13	20 187 232 1 193	15 165 208 1 135	13 125 141 4	16 109 83 6	18 119 53 12	7 57 47	24 261 391 2 213
	Arc Bac Euk Vir	Arc Bac Euk Unk Vir	Arc Bac Euk Unk Vir	Arc Bac Euk Vir	Arc Bac Euk Vir	Arc Bac Euk Vir	Arc Bac Euk	Arc Bac Euk Unk Vir

^aFour-level EC numbers are distributed according to the 7 EC classes: EC1-Oxidoreductases, EC2-Transferases, EC3-Hydrolases, EC4-Lyases, EC5-Isomerases, EC6-Ligases, EC7-Translocases.

^bRef Seq and Ref Str: number of reference sequences, and reference sequence with structure, respectively; number of polyfunctional enzymes are within brackets.

^cPfam: models from Pfam (<https://pfam.xfam.org>); within brackets Pfams, where relevant sites (active, metal, ligand binding site) are annotated.

^dKEGG ID: from UniProt annotation; KEGG pathway: from <https://www.genome.jp/kegg/>.

^enumber of organisms detailed for each kingdom. Arc: Archaea; Bac: Bacteria; Euk: Eukaryota; Oth: Others; Vir: Viruses. Unk: unknown. Annotation source: UniProt. Grand Total: 891.

Table 2. BENZ at work

Dataset	Sequences (#)	Acc ^c (%)	FNR ^f (%)	FPR ^g (%)
Positive ^a	197 880	92.4	3.9	-
Negative ^b	12 315	95.1	-	4.9
Polyfunctional ^c	10 764	93.7	5.0	-
TrEMBL-human ^d	10 024	93.4	5.6	-

^aPositive: the positive set contains complete SwissProt sequences without any PDB counterpart and annotated with only four-level EC number.

^bNegative: the negative set comprises complete SwissProt sequence with a PDB counterpart, without EC codes.

^cPolyfunctional: the set includes complete SwissProt sequence that are annotated with two or more four-level EC numbers.

^dTrEMBL-human: the set contains complete TrEMBL sequences from *Homo sapiens* annotated with a four-level EC number.

^eAcc (Accuracy) measures the number of proteins correctly assigned. For sets containing positive examples, it corresponds to the True Positive Rate as evaluated at the level of four: EC annotation. For the negative set, it corresponds to the True Negative Rate.

^fFNR (False Negative Rate) measures the percentage of enzymes predicted as non-enzymes.

^gFPR (False Positive Rate) measures the percentage of non-enzymes predicted as enzymes.

We then compared BENZ with three state-of-the-art tools: ECPred (15), DEEPre (16) and EFICAz2.5 (25). Only the reduced positive dataset was adopted since the selected methods do not consider the seventh EC class. Results indicate that BENZ outperforms the other tools in this benchmark (Table 3). TPR values of BENZ WS range from two-fold up to four-fold those obtained by the other predictors, increasing at increasing levels of EC. Concomitantly, BENZ WS FNR values overpass other predictors values by at least two or three times those of the other predictors (Table 3, FNR column).

In the reduced set, BENZ achieves a better discrimination than the other methods (FNR) and a better EC assignment sensitivity, with a TPR value ranging from 79.2% to 75% at increasing level of predicted EC (Table 3, TPR columns), and it significantly overpasses the second best-scoring method (DEEPre, 16). As to the correct recognition of non-enzymes (column FPR, Table 3), DEEPre and EFICAz2.5.1 show a better performance, which in turn

is counter-balanced by a low ability to recognise enzymes (TPR values).

DISCUSSION

A major problem in addressing EC code annotation is due to the different levels of specificity that the code carries. Only the complete four-level annotation fully characterises the protein biochemical activity. However, due to evolution, different active site architectures can catalyse the same biochemical activity and/or the same active site can bind different substrates (26). These difficulties may hamper the EC direct association to the protein sequence and rather suggest a direct prediction of GO terms, like in the CAFA experiments (3).

Here, we tackled the problem of the association of protein sequence with four-level EC code/s taking advantage of two different types of HMMs. One, the cluster HMM derives from a hierarchical clustering procedure that we adopted before for generating a system (BAR 3.0) suited to a general-purpose protein annotation and based on a rigorous and statistically validated transfer of annotation. Cluster HMMs model sequences, which have been clustered after constraining their identity ($\geq 50\%$) over 90% of the alignment length. By this, cluster HMMs retain sequences that pairwise share a high level of similarity over a large portion of the alignment length, although belonging to different organisms. Furthermore, they may conserve relevant sites in specific Pfam domains. Among the cluster-sequences, we select one reference sequence (the one with the highest score of annotation) and define its architecture by mapping Pfam domains to the chain. When present, all the relevant sites (active, ligand and metal binding) are also mapped to the corresponding Pfam domain/s. Finally, we associate each representative, its architecture and EC code/s to a more general representation, casted into the cluster HMM. Indeed, structural matching for gaining the EC code of the representative reference is checked by comparing the target predicted architecture and the reference one.

Testing BENZ on selected sets of proteins (Table 2) indicates that the system correctly rejects (97%) non enzymes

Table 3. BENZ benchmarking

Method	Data set ^a	TPR ^f (%) 1 level	TPR ^f (%) 2 level	TPR ^f (%) 3 level	TPR ^f (%) 4 level	FNR ^g (%)	FPR ^h (%)
BENZ WS ^b	Full	87.5	87.5	87.5	85.0	12.2	3.0
BENZ WS ^b	Reduced	79.2	79.2	79.2	75.1	20.2	3.0
ECPred ^c	Reduced	43.7	34.7	23.8	13.1	45.6	12.2
DEEPre ^d	Reduced	38.8	35.2	27.9	20.8	51.1	2.4
EFICAZ2.5.1 ^e	Reduced	33.6	33.1	31.1	16.7	63.7	1.6

^aBenchmark datasets are extracted by comparing SwissProt releases 2020_3 and 2019_11. The full dataset includes 607 proteins that have gained EC annotation (7 EC classes); the reduced dataset includes a subset of 366 enzyme sequences without EC codes of the seventh class for comparing with the other predictors. Both datasets comprise 1013 non-enzyme sequences as negative examples.

^bA BENZ WS version including only sequences and annotations available in the SwissProt release 2019_11 has been used for this test.

^cECPred (15) has been downloaded from <https://github.com/cansyl/ECPred> and run in-house; it does not provide multiclass predictions and the best match between the output and the list of EC numbers has been considered for multiclass enzymes. It does not include enzymes of for EC class 7.

^dDEEPre (16) predictions have been run on the webserver <http://www.cbrc.kau.edu.sa/DEEPre/> in modality 'I'm not sure the sequence is an enzyme'; it does not provide multiclass predictions and the best match between the output and the list of EC numbers has been considered for multiclass enzymes. It does not include enzymes of the EC class 7.

^eEFICAZ2.5.1 (25) has been downloaded from <https://sites.gatech.edu/cssb/eficaz2-5/> and run in-house; it does not include enzymes of EC class 7.

^fTPR (True Positive Rate) measures the number of enzymes assigned to the correct EC class. TPRs have been evaluated at the level of four-level EC annotation.

^gFNR (False Negative Rate) measures the percentage of enzymes predicted as non-enzymes.

^hFPR (False Positive Rate) measures the percentage of non-enzymes predicted as enzymes.

and that it is efficient in retaining never seen before enzyme sequences (Table 3). BENZ will eventually assign only EC codes present in the system as specific four-level EC-Cluster HMM-reference sequence association. This will be taken care of with new BENZ releases, following new UniProt releases.

When BENZ is benchmarked with other EC predictors, based on first structural principles or machine and deep learning methods, it is superior (Table 3). Predictors, which we found available, are based on different methods and not directly comparable. However, their poor performance on the specific task of EC code prediction, including poly-functional enzymes, suggests that fine-tuning of the protein functional family representation is necessary and that machine learning, including end-to-end models, poorly captures it.

We introduce BENZ as a reliable method for transfer of knowledge after generalisation over subsets of proteins belonging to specific functional and structural families.

DATA AVAILABILITY

BENZ WS is freely available as a web server at the following URL: <https://benzdb.biocomp.unibo.it/>.

SUPPLEMENTARY DATA

[Supplementary Data](#) are available at NAR Online.

ACKNOWLEDGEMENTS

D. Baldazzi is the recipient of a PhD fellowship for the Data Science and Computation PhD program of the University of Bologna, Italy, supported by Centro di Riferimento Oncologico (CRO), Aviano, Italy.

FUNDING

Italian Ministry of Education, University and Research [PRIN2017 grant, project no. 2017483NH8.002 to C.S.];

European Commission H2020 programme [CIRCLES project, grant no. 818290 to P.L.M.]. Funding for open access charge: Italian Ministry of Education, University and Research.

Conflict of interest statement. Not declared.

REFERENCES

1. The UniProt Consortium. (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.*, **49**, D480–D489.
2. Burley, S.K., Bhikadiya, C., Bi, C., Bittrich, S., Chen, L., Crichlow, G.V., Christie, C.H., Kenneth Dalenberg, K., Di Costanzo, L. *et al.* (2021) RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res.*, **49**, D437–D451.
3. Zhou, N., Jiang, Y., Bergquist, T.R., Lee, A.J., Kacsóh, B.Z., Crocker, A.W., Lewis, K.A., Georghiou, G., Nguye, H.N. and Friedberg, I. (2019) The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome biology*, **20**, 244.
4. Gene Ontology Consortium. (2021) The Gene Ontology resource: enriching a GOLD mine. *Nucleic Acids Res.*, **49**, D325–D334.
5. Pundir, S., Onwubiko, J., Zaru, R., Rosanoff, S., Antunes, R., Bingley, M., Watkins, X., O'Donovan, C. and Martin, M.J. (2017) An update on the Enzyme Portal: an integrative approach for exploring enzyme knowledge. *Protein Eng. Des. Sel.*, **30**, 245–251.
6. Lombardot, T., Morgat, A., Axelsen, K.B., Aimo, L., Hyka-Nouspikel, N., Niknejad, A., Ignatchenko, A., Xenarios, I., Coudert, E. and Bridge, A. (2019) Updates in Rhea: SPARQLing biochemical reaction data. *Nucleic Acids Res.*, **47**, D596–D600.
7. Ribeiro, A.J.M., Holliday, G.L., Furnham, N., Tyzack, J.D., Ferris, K. and Thornton, J.M. (2018) Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. *Nucleic Acids Res.*, **46**, D618–D623.
8. Pedruzzi, I., Rivoire, C., Auchincloss, A.H., Coudert, E., Keller, G., de Castro, E., Baratin, D., Cuhe, B.A., Bougueleret, L. and Bridge, A. (2015) HAMAP in 2015: updates to the protein family classification and annotation system. *Nucleic Acids Res.*, **43**, D1064–D1070.
9. Sigrist, C.J.A., de Castro, E., Cerutti, L., Cuhe, B.A., Hulo, N., Bridge, A., Bougueleret, L. and Xenarios, I. (2013) New and continuing developments at PROSITE. *Nucleic Acids Res.*, **41**, D344–D347.
10. Blum, M., Chang, H.-Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., Nuka, G., Paysan-Lafosse, T., Qureshi, M. and Finn, R.D.

- (2021) The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.*, **49D1**, D344–D354.
11. Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Silio Tosatto, S., Paladin, L., Raj, S. and Bateman, A. (2021) Pfam: the protein families database in 2021. *Nucleic Acids Res.*, **49D1**, D412–D419.
 12. Chang, A., Jeske, L., Ulbrich, S., Hofmann, J., Koblitz, J., Schomburg, I., Neumann-Schaal, M., Jahn, D. and Schomburg, D. (2021) BRENDA, the ELIXIR core data resource in 2021: new developments and updates. *Nucleic Acids Res.*, **49**, D498–D508.
 13. Sillitoe, I., Dawson, N., Lewis, T.E., Das, S., Lees, J.G., Ashford, P., Tolulope, A., Scholes, H.M., Senatorov, I. and Orengo, C.A. (2019) CATH: expanding the horizons of structure-based functional annotations for genome sequences. *Nucleic Acids Res.*, **47D1**, D280–D284.
 14. Tan, J.X., Lv, H., Wang, F., Dao, F.Y., Chen, W. and Ding, H. (2019) A survey for predicting enzyme family classes using machine learning methods. *Curr. Drug Targets*, **20**, 540–550.
 15. Dalkiran, A., Rifaioglu, A.S., Martin, M.J., Cetin-Atalay, R., Atalay, V. and Doğan, T. (2018) ECPred: a tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature. *BMC Bioinformatics*, **19**, 334.
 16. From, Li Y., Wang, S., Umarov, R., Xie, B., Fan, M., Li, L. and Gao, X. (2018) DEEPred: sequence-based enzyme EC number prediction by deep learning. *Bioinformatics*, **34**, 760–769.
 17. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. and Tanabe, M. (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.*, **44**, D457–D462.
 18. Profiti, G., Martelli, P.L. and Casadio, R. (2017) The Bologna Annotation Resource (BAR 3.0): improving protein functional annotation. *Nucleic Acids Res.*, **45**, W285–W290.
 19. Profiti, G., Fariselli, P. and Casadio, R. (2015) AlignBucket: a tool to speed up ‘all-against-all’ protein sequence alignments optimizing length constraints. *Bioinformatics*, **31**, 3841–3843.
 20. Eddy, S.R. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol.*, **7**, e1002195.
 21. Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W. and Higgins, D. G. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.
 22. Juty, N., Le Novère, N. and Laibe, C. (2012) Identifiers.org and MIRIAM Registry: community resources to provide persistent identification. *Nucleic Acids Res.*, **40**, D580–D586.
 23. Talevich, E., Invergo, B.M., Cock, P.J. and Chapman, B.A. (2012) Bio.Phylo: a unified toolkit for processing, analyzing and visualizing phylogenetic trees in Biopython. *BMC Bioinformatics*, **13**, 209.
 24. Mukhyala, K. and Masselot, A. (2014) Visualization of protein sequence features using JavaScript and SVG with pViz.js. *Bioinformatics*, **30**, 3408–3409.
 25. Kumar, N. and Skolnick, J. (2012) EFICAz2.5: application of a high-precision enzyme function predictor to 396 proteomes. *Bioinformatics*, **28**, 2687–2688.
 26. Tyzack, D.J., Furnham, N., Sillitoe, I., Orengo, C.M. and Thornton, J.M. (2017) Understanding enzyme function evolution from a computational perspective. *Curr. Opin. Struct. Biol.*, **47**, 131–139.



Article

A Glance into MTHFR Deficiency at a Molecular Level

Castrense Savojardo ¹, Giulia Babbi ¹, Davide Baldazzi ¹, Pier Luigi Martelli ^{1,*} and Rita Casadio ^{1,2}

¹ Biocomputing Group, Department of Pharmacy and Biotechnology, University of Bologna, 40126 Bologna, Italy; castrense.savojardo2@unibo.it (C.S.); giulia.babbi3@unibo.it (G.B.); davide.baldazzi8@unibo.it (D.B.); rita.casadio@unibo.it (R.C.)

² Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies (IBIOM), Italian National Research Council (CNR), 70126 Bari, Italy

* Correspondence: pierluigi.martelli@unibo.it

Abstract: MTHFR deficiency still deserves an investigation to associate the phenotype to protein structure variations. To this aim, considering the MTHFR wild type protein structure, with a catalytic and a regulatory domain and taking advantage of state-of-the-art computational tools, we explore the properties of 72 missense variations known to be disease associated. By computing the thermodynamic $\Delta\Delta G$ change according to a consensus method that we recently introduced, we find that 61% of the disease-related variations destabilize the protein, are present both in the catalytic and regulatory domain and correspond to known biochemical deficiencies. The propensity of solvent accessible residues to be involved in protein-protein interaction sites indicates that most of the interacting residues are located in the regulatory domain, and that only three of them, located at the interface of the functional protein homodimer, are both disease-related and destabilizing. Finally, we compute the protein architecture with Hidden Markov Models, one from Pfam for the catalytic domain and the second computed in house for the regulatory domain. We show that patterns of disease-associated, physicochemical variation types, both in the catalytic and regulatory domains, are unique for the MTHFR deficiency when mapped into the protein architecture.

Keywords: MTHFR deficiency; MTHFR variants; functional annotation; structural annotation; disease related variations; solvent accessibility; $\Delta\Delta G$ predictions; consensus method; protein-protein interactions; disease HMM models



Citation: Savojardo, C.; Babbi, G.; Baldazzi, D.; Martelli, P.L.; Casadio, R. A Glance into MTHFR Deficiency at a Molecular Level. *Int. J. Mol. Sci.* **2022**, *23*, 167. <https://doi.org/10.3390/ijms23010167>

Academic Editor: Paola Ghiorzo

Received: 11 October 2021

Accepted: 21 December 2021

Published: 23 December 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The one-carbon metabolism cycle, including the folate and methionine cycles, is a critical pathway for cell survival. The human enzyme methylenetetrahydrofolate reductase (encoded by the gene MTHFR, UniProt code: P42898) exchanges one-carbon unit from the folate to methionine cycle. This is exclusively used for methionine and S-adenosylmethionine (SAM) synthesis, and MTHFR is the rate-limiting enzyme in the methyl cycle, undergoing allosteric inhibition by its end product SAM (S-Adenosil-Methionine) [1–4]. The protein is functional in its homodimeric form [5].

MTHFR catalyzes the conversion of 5,10-methylenetetrahydrofolate to 5-methyltetrahydrofolate, a co-substrate for homocysteine re-methylation to methionine (EC number: 1.5.1.20). The recent release of its structure (PDB code: 6FCX, 0.25 nm resolution) highlights the organization of the protein into two flexible domains, one catalytic and one regulatory, with a connecting linker allowing domain-domain interactions, possibly due to a phosphorylation cascade. The structure clarifies the molecular mechanism of the reaction, which requires FAD as a cofactor, NAD(P)H to provide reducing equivalents and homodimerisation for allosteric regulation upon SAM binding at the regulatory domain [6]. The 36-residue N-terminal portion is not resolved in the available PDB file and MobiDB predicts only here a flexible region (<https://mobidb.bio.unipd.it/P42898> accessed on 10 October 2021). The PDB contains the homodimeric protein organization.

MTHFR deficiency and upregulation result in various disease states, which have been extensively described in relation to a number of variants characterized in many studies. 109 MTHFR mutations have been reported in 171 families, including 70 missense mutations, 17 that primarily affect splicing, 11 nonsense mutations, seven small deletions, two no-stop mutations, one small duplication, and one large duplication [7]. Two other variants, A222V and E429A, distributed worldwide in the population, are characterized by a reduced enzymatic activity, and are associated to different risk factors [8,9]. Variations, reducing the MTHFR activity to different extents, result in hyperhomocysteinemia and varying severities of disease, including ischemic stroke, folate sensitive neural tube defects and schizophrenia [1]. Evidently, the protein is also an attractive drug target [10]. All known missense variations are distributed in the three-dimensionally resolved catalytic and regulatory domains.

In this study, we are interested in exploiting with computational tools the structural properties of the protein missense variations associated to the disease to highlight possible mechanisms of protein destabilization due to residue change. To this aim we first map disease variations on the protein structure in relation to their solvent accessibility and compute for the accessible variations their likelihood of being involved in protein-protein interactions. We also compute the Gibbs free energy change ($\Delta\Delta G$) for each variation with a consensus method and find that positions 387 (G387D), 506 (Y506D) and 628 (L628T) of the protein homodimeric interface at the level of the two regulatory domains, besides being correctly predicted as interaction sites, are also destabilizing the protein homodimer. This corroborates the relevance of the interaction of the two regulatory domains for the stability of the functional protein. We then grouped all the disease-related variations according to their physicochemical types and mapped them into the computed HMM modelled architecture of the protein. By this we establish a link among protein domains and variation types, which is a unique marker of the MTHFR deficiency.

2. Results

Application of state-of-the art tools for functional annotation of a protein is common routine in the field of computational biology. Here, having the solved structure of the MTHFR gene, we aim at highlighting possible structural properties of the missense variations associated to the deficiency, and most cases associated to a decreased biochemical efficiency. Our goal is to relate computational properties, such as solvent exposure, being an interaction site and promote protein instability, to their annotation of being disease-associated. This highlights some interesting properties of the disease related variations and in turn benchmarks tools in the difficult task of their prediction.

2.1. MTHFR and Protein-Protein Interactions

Large scale experiments of interactomics indicate that human MTHFR interacts with many specific partners. BioGRID (<https://thebiogrid.org/> accessed on 10 October 2021), the Database of Protein, Genetic and Chemical Interactions, lists 33 physical interactors, 22 of which are also present in IntAct (<https://www.ebi.ac.uk/intact/> accessed on 10 October 2021), the other molecular data base collecting data from large scale experiments. It is worth noticing that none of the enzymes involved in the folate and methionine cycles are present among the physical interactors, and that many membrane and nuclear proteins are in the interacting protein pool. Why this is so perhaps deserves more experiments, and it can be interpreted considering the presence of MTHFR in different cell compartments, including its putative interaction with mitochondria, endoplasmic reticulum, and the nucleus [1]. For the time being, we can compute the likelihood of solvent-exposed residues to be in contact with a putative partner. We adopt our ISPREd4 predictor [11], which is based on machine learning, and it is specifically suited to compute the likelihood of an exposed residue to be involved in a protein contact. We compute 44 interacting sites in the protein structure (see Supplementary Material Table S1 for details), 17 of which are at the interface between the two regulatory domains of the homodimer. Many of the

interactions reported in the databases are likely to be non-obligate and therefore different interactions can involve the same sites, in different compartments and phases of the cell lifespan. This can be considered the reason why the number of interaction sites as derived from a structure rarely coincides with the number of interactors as derived from large scale experiments. In the following, our interest is on disease-related variations which are in interaction sites and whose functional annotations are already documented (Table 1).

In Figure 1, predicted contacts are represented with hard spheres centered on the C-alpha atom of the specific residue. The color code follows the organization of the protein in the catalytic (yellow) and regulatory (pale blue) domain, inclusive of the linker region [6]. The bound FAD and SAH molecules, present in the protein crystal (6FCX), are also shown for clarity, and their binding sites are relevant for protein catalytic activity.

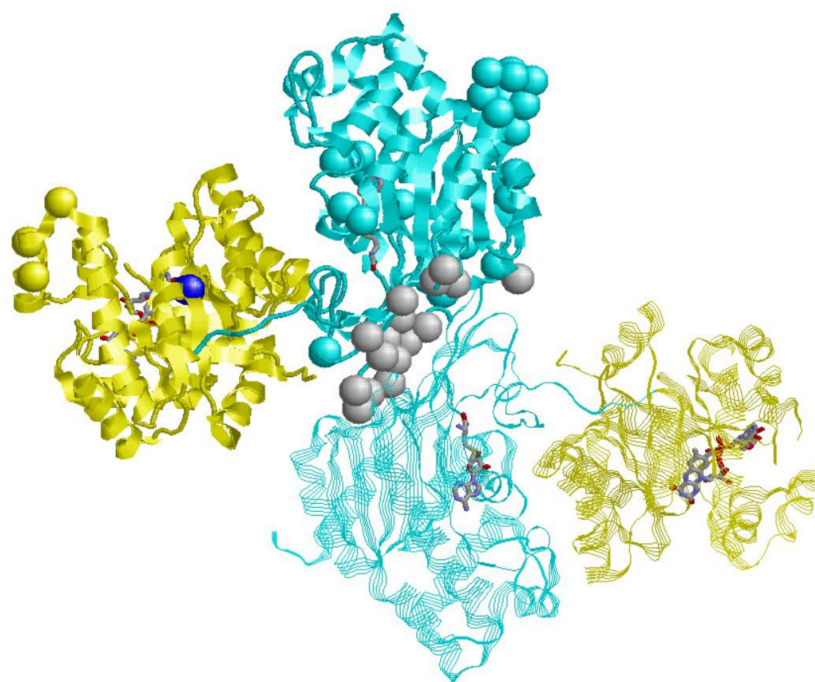


Figure 1. Protein-protein interaction sites predicted with ISPRED4 [10] on the MTHFR PDB 6FCX. The catalytic and regulatory domains are depicted in yellow and pale blue, respectively. Interaction sites are represented with hard spheres centered on the C-alpha atom of the specific residue. Grey spheres are residues in the homodimeric interface, correctly predicted as interaction sites.

Interestingly enough, we correctly predict the interface region of the homodimer (hard spheres in grey). Other predicted PPI sites are distributed in different regions of the protein surface. These residues are candidates for taking part in the interaction with the 33 proteins reported in the IntAct and BioGRID databases. Only in position 387 (G387D), 506 (Y506D) and 628 (L628T) of the homodimeric interface at the level of the regulatory domain do the predicted interaction sites coincide with missense variations associated with the MTHFR disease. These variations are also predicted as destabilizing (see below). This observation finally highlights the role of the regulatory domain interactions not only in being part of the protein functional stability, but also in playing a role in the disease [5].

2.2. MTHFR and Protein Stability

We can investigate whether disease-related missense variations are related to protein instability. To this aim, we adopt a consensus method, computing (with three state-of-the-art methods) the Gibbs free energy change ($\Delta\Delta G$) associated with a specific variation in the protein. We select a consensus method, given the variability of the different methods in predicting the $\Delta\Delta G$ values [12], and adopt three of the art methods: INPS-MD [13] is based

on machine learning, FoldX [14] on statistical potentials, and PoPMuSiC2 [15] on statistical potentials and machine learning.

Table 1. MDHFR deficiency-related variations.

Variation	Effects	$\Delta\Delta G$ (kcal/mol)					RSA (%)
		INPS3D	FoldX	PoPMuSiC2	ISPRED4		
Catalytic Domain							
R46Q	No effect on NAD(P) affinity	−0.76	−0.26	−1.05	N	29	
R46W	No effect on NAD(P) affinity	−0.5	−1.04	−0.35	N	29	
R51P		−1.24	−1.13	−1.47	N	49	
R52Q	Reduced affinity for NAD(P)	−1.06	0.08	−0.77	N	23	
W59C		−1.59	−3.58	−2.52	N	2	
W59S		−2.67	−3.92	−3.3	N	2	
P66L	NAD(P) binding site	−0.46	−4.09	−0.01	N	20	
R68G	Reduced affinity for NAD(P) NAD(P) binding site	−0.92	−0.40	−0.59	N	96	
R82W	No effect on NAD(P) affinity	−0.66	0.2	−0.81	N	44	
A113T	No effect NADPH	−1.17	−1.44	−1.71	N	0	
A116T		−0.65	−2.29	−1.95	N	0	
H127Y	FAD binding site	−0.18	1.37	−0.33	N	5	
T129N	Reduced affinity for NAD(P) FAD binding site	−1.17	−1.37	−0.81	N	7	
C130R	No effect on NAD(P) affinity	−1.99	−16.08	−1.34	N	1	
T139M		−0.34	0.83	0.29	N	18	
Q147P		−0.46	−2.95	−0.91	N	73	
G149V		−1.02	−13.0	−3.26	N	2	
I153M	No effect on NAD(P) affinity	−1.56	0.18	−1.71	N	1	
R157Q	No effect on NAD(P) affinity FAD binding site	−1.31	−0.72	−0.57	N	25	
A175T	Reduced affinity for NAD(P) FAD binding site	−1.13	−0.73	−0.54	N	8	
H181D		−1.79	−2.23	−1.5	N	10	
R183Q	No effect on NAD(P) affinity	−1.48	−3.34	−0.82	N	16	
C193Y		−1.19	−10.91	−0.03	N	17	
A195V	Reduced affinity for NAD(P) FAD binding site	−0.43	0.39	0	N	11	
G196D	Reduced affinity for NAD(P)	−1.08	−3.26	−1.18	N	2	
P202T	FAD binding site	−0.73	−1.57	−0.16	N	68	
V218L	Decreased affinity for FAD	−1.00	−0.42	−0.42	N	12	
A222V *	Decreased affinity for FAD	−0.71	−1.08	−0.09	N	11	
I225L	No effect on NAD(P) affinity	−1.32	−0.57	−1.17	N	0	
T227M		−1.58	−2.9	−0.14	N	1	
P251L		−0.56	0.62	−0.68	N	38	
V253F	Reduced affinity for NAD(P)	−0.82	−1.26	−1	N	0	
P254S	No effect on NAD(P) affinity	−1.22	−3.7	−0.86	N	0	
G255V		−0.55	−2.81	0.33	N	1	
I256N		−3.24	−3.27	−2.47	N	1	
F257V		−1.34	−1.61	−1.83	N	11	
L323P	Substrate binding site NAD(P) binding site	−2.19	−4.95	−1.94	N	32	
N324S		−0.77	−3.52	−1.92	N	8	
R325C	Substrate binding site	−0.78	0.41	−0.34	N	43	
L333P		−3.39	−6.05	−3.62	N	0	
R335C		−0.67	−1.14	−0.86	N	60	

Table 1. Cont.

Variation	Effects	$\Delta\Delta G$ (kcal/mol)			ISPRED4	RSA (%)
		INPS3D	FoldX	PoPMuSiC2		
Regulatory Domain						
M338T		−1.58	−3.74	−1.21	N	18
W339G		−2.78	−4.46	−2.55	N	20
R345C		−0.67	−1.31	−0.23	N	43
P348S	Reduced affinity for NAD(P) SAH binding site	−1.19	−3.53	−1.16	N	26
H354Y	Reduced affinity for NAD(P)	−0.24	−0.2	−0.67	N	18
R357C		−1.32	−2.3	−1.54	N	5
R357H		−1.28	−1.09	−0.29	N	5
R363H	Reduced affinity for NAD(P)	−1.39	−1.34	−0.83	N	6
K372E	Reduced affinity for NAD(P)	−0.46	0.99	−0.31	N	52
R377C	Reduced affinity for NAD(P)	−1.17	−3.99	−1.4	N	0
R377H	Reduced affinity for NAD(P)	−1.2	−4.59	−0.68	N	0
W381R		−1.83	−2.29	−1.95	N	14
G387D	Reduced affinity for NAD(P)	−0.82	−3.35	−1.31	I	33
G390D		−0.88	−2.23	0.13	N	64
W421S	Reduced affinity for NAD(P)	−3.07	−6.97	−4	N	1
E429A *		−0.13	−0.79	0.2	N	50
F435S		−3.45	−5.56	−2.94	N	1
S440L		0.03	2.15	−0.55	N	25
Y506D	Reduced affinity for NAD(P)	−1.77	−5.1	−3.16	I	61
Y512C		−1.94	−4.31	−2.18	N	2
R535Q		−0.79	−1.61	−0.77	N	25
R535W		0.07	−1.39	−0.26	N	25
V536F	Reduced affinity for NAD(P)	−1.38	−3.21	−0.6	N	1
P572L	Reduced affinity for NAD(P)	−0.43	−7.32	−0.09	N	0
V574G	Reduced affinity for NAD(P)	−3.32	−4.13	−3.47	N	1
V575G	Reduced affinity for NAD(P)	−3.64	−4.06	−3.07	N	8
E586K		−0.8	−5.23	−0.99	N	1
L598P	Reduced affinity for NAD(P)	−2.49	−7.34	−2.68	N	22
S603C		−1.03	−1.81	−0.7	N	15
L628P	Reduced affinity for NAD(P)	−0.83	−4.47	−2.25	I	57
M338T		−1.58	−3.74	−1.21	N	18

The table lists 72 variations associated to the MTHFR deficiency, as reported in the MTHFR UniProt file MTHFR, UniProt code: P42898 and in [7]. * Variations are described in [8] (A222V) and [9] (E229), respectively. Effects of the variations on the MTHFR enzymatic activity are listed when reported. Bold style indicates variations for which at least two of the three methods adopted for computing $\Delta\Delta G$ (INPS3D, [13]; FoldX [14]; and PoPMuSiC2 [15], compute negative results, lower than -1 kcal/mol, indicating protein destabilization (for details see text). For completeness, we include results (I, Interaction; N, No Interaction) of the Interaction site prediction method (ISPRED4) [11] and values of the relative solvent accessibility (RSA%) (see Materials and Methods for details) (second to last column and right-most column, respectively).

We select as a threshold value $|1 \text{ kcal/mol}|$, which takes into account the variability of the experimental thermodynamic data on protein stability adopted for training the predictors. In Table 1, we list the 42 disease-related variations in the catalytic domain and the 30 disease related variations in the regulatory domain. Alongside this, we indicate the corresponding effects on the protein function, the computed $\Delta\Delta G$ according to the three predictors, the prediction of the wild-type residue to be in contact or not and the computed relative solvent accessibility [16]. It appears that 22 variations in the catalytic domain and 20 variations in the regulatory domain decrease protein stability, according to at least two of the three predictors. Among the remaining ones, seven are in the NAD(P)H binding site, and the other seven are in the FAD binding site, respectively. These variations, as reported in Table 1, decrease the binding affinity without perturbing the protein stability, including A222V and E429A. In any case, the available structure 6FCX contains an A in position 429

instead of E, and in this case, we compute $\Delta\Delta G$ of the reverse variation [12]. R325C, in the substrate-binding site, decreases substrate affinity without affecting protein stability.

In the protein catalytic domain, we map 41 disease related variations, 14 of which are exposed and not in interaction sites, 20 are destabilizing and mostly (90%) buried. In the protein regulatory domain, we map the remaining 31 disease related variations, 12 of which are exposed, 21 are destabilizing and 15 of these are buried. Interestingly, positions 387 (G387D), 506 (Y506D) and 628 (L628T) at the protein homodimeric interface, correctly predicted as interaction sites (see above), promote also protein destabilization, supporting a role of the regulatory domain interactions in the stability of the functional protein homodimeric complex. Overall, 61% of the disease related variations are affecting the protein stability and most of them have been experimentally found to promote instability of cofactor binding.

Out of the pool of the MDRFH deficiency variations listed in Table 1, UniProt in the protein file P42898 lists eight other variations with a dbSNP code (<https://www.ncbi.nlm.nih.gov/snp/> accessed on 10 October 2021) not yet associated to disease (likely benign?). Six of these maps into the protein structure and two of them, (G422R, exposed, and G566E, at the homodimer interface) destabilize the protein structure according to our criterion. Results are in line with previous observations highlighting how protein instability is not a necessary condition for being disease-related [17,18], although in this specific case many of the variations are indeed destabilizing the protein organization (Table 1).

2.3. MTHFR Deficiency and Its Structural Model

Recently we introduced the concept of mapping disease variation types into associated Pfam structural protein models (<https://pfam.xfam.org> accessed on 10 October 2021), finding that by this it is possible to establish a relation among genes and maladies [18,19]. Indeed, mapping of variation types into Pfam is unique for a given disease. Here we exploit our strategy with MTHFR, considering Pfam 02219 for the catalytic domain. This Pfam is shared by similar proteins in Eukarya, Bacteria and Archaea. The regulatory domain does not have a Pfam model and is present only in Eukarya. We build a model for the regulatory domain aiming for a structural representation of the complete protein. The model is an HMM of the profile of a multiple alignment of some 50 sequences from Eukarya with a length similar to that of MTHFR. The model includes the linker, and it spans from residue 336 to residue 566 (see Supplementary Material, where the Pfam-like model of the regulatory domain is reported). We then converted the disease related variations of Table 1 into variation types (apolar (G, A, V, I, L, P, M); polar (S, T, C, N, Q, H); aromatic (F, W, Y); charged (D, E, K, R) giving rise to 16 possible variation types. We associated MTHFR related variation types to the protein architecture, as represented by P02219 and our Pfam-like model of the regulatory domain. The frequency of the variation types in each domain is represented in Figure 2.

It appears that the variational pattern is different in the two domains and different from the background variational pattern, obtained considering the pathogenic variations from Humsavar (<https://www.uniprot.org/docs/humsavar> accessed on 10 October 2021), in 2513 human proteins (22,763 disease related variations).

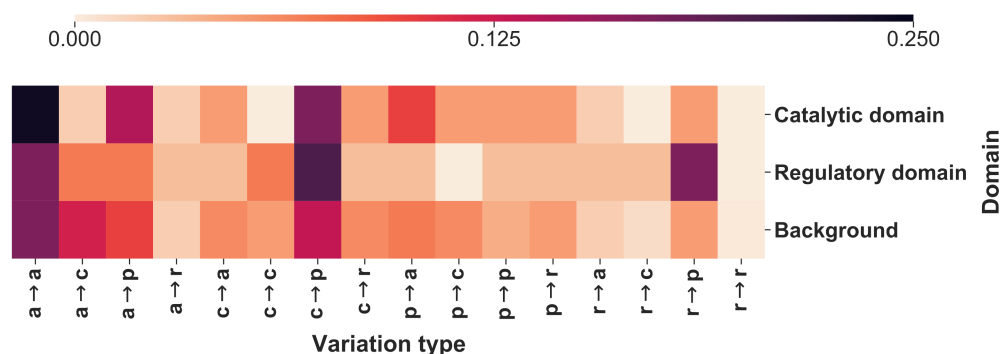


Figure 2. The heatmap reporting the frequency of each variation type as observed within the catalytic and the regulatory domains. The background distribution has been computed considering 22,763 pathogenic variations from Humsavar in 2513 proteins. In variation types, labels are as follows: a, apolar; c, charged; p, polar; and r, aromatic (for details see text). Differences between Catalytic and Regulatory sites are significant at 10% when a Chi-square test is applied after adding pseudocounts (with value 0.5) for regularization.

3. Materials and Methods

3.1. Characterization of Protein Surface and Annotation of Protein-Protein Interaction Sites

The solvent accessibility of residues of PDB entry 6FCX, chain A, [6] was computed with the DSSP program (https://swift.cmbi.umcn.nl/gv/dssp/DSSP_3.html accessed on 10 October 2021) and normalized with respect to the residue-specific maximal accessibility values as previously described [16]. Residues interacting in the homodimer interface are those undergoing a decrease of the absolute solvent accessibility (ASA) $\geq 1 \text{ \AA}^2$ in the complex with respect to the isolated monomer.

Protein-protein interaction sites were predicted with ISPRED4 [11], a tool based on support vector machines and grammatical restrained hidden conditional random fields that integrate 46 different features extracted from the monomer sequence, its multiple sequence alignment against the UniProt database and its 3D structure. ISPRED4 has been trained and cross-validated on 151 protein complexes and reaches a per-residue Matthews correlation coefficient of 0.48 and an overall accuracy of 0.85. Similar values are obtained on blind test sets, and therefore ISPRED4 is one of the top-performing tools for the computational annotation of protein-protein interaction sites.

3.2. Prediction of $\Delta\Delta G$ Changes upon Single Residue Variation

The possible effect on protein stability induced by single residue variation starting from protein structure has been predicted with three state-of-the-art methods: (i) INPS3D [13], a tool based on a machine-learning approach; (ii) FoldX [14], that estimated energy changes on the basis of a knowledge-based potential; and (iii) PoPMuSic2 [15], a method implementing a combination of statistical potentials optimized with a neural network. The following convention has been adopted for the definition of the $\Delta\Delta G$ sign:

$$\Delta\Delta G = (\Delta G_{\text{wt}} - \Delta G_{\text{mut}}) \quad (1)$$

where ΔG_{wt} and ΔG_{mut} are the folding free energy of the wild-type and mutated proteins, respectively. Negative values of $\Delta\Delta G$ mean that the mutated form is less stable than the wild-type. We considered as destabilizing the variations for which at least two methods predict $\Delta\Delta G \leq -1 \text{ kcal/mol}$.

Since the structure 6FCX carries the mutant allele (A) in position 429, the thermodynamic effect of variation E429A was estimated by computing the $\Delta\Delta G$ of variation A429E on the crystal and applying the antisymmetric principle.

3.3. Pfam-Like Model of the Regulatory Domain

From the UniRef50 cluster UniRef50_P42898 (https://www.uniprot.org/uniref/UniRef50_P42898 accessed on 10 October 2021), we collected 150 complete protein sequences from Eukarya and with length ranging between 640 and 670 residues. These sequences cover both domains of human MTHFR protein and share more than 50% sequence identity with it. We aligned the sequences with ClustalOmega (<https://www.ebi.ac.uk/Tools/msa/clustalo/> accessed on 10 October 2021) and extracted the multiple sequence alignment of the regulatory domain, spanning from position 336 to 566 in the human sequence. We then trained a HMM, with HMMER 3.3.2 (<http://hmmmer.org/> accessed on 10 October 2021). The trained model is available in the Supplementary Materials.

4. Conclusions

In this paper we exploit different computational methods for refining the annotation of the disease related variants of MTHFR, promoting MTHFR deficiency. Due to the numerous biological processes in which the protein is directly and/or indirectly involved, MTHFR is of particular interest, since its partial or total dysfunction may have a range of effects on human health, spanning from mild to lethal ones. Among the known 72 disease related variations, we characterize those that are at the protein surface, participate into protein-protein contacts and are at the homodimer interface which involves the protein regulatory domain. We also highlight other properties of the protein, like the exposed residues that eventually participate in the protein-protein interaction (Table S1). Furthermore, we show that 61% of the disease related variants are destabilizing the protein, highlighting a possible source of structural destabilization causing the decreased binding affinity of the protein cofactors when documented. Noteworthy is that positions 387 (G387D), 506 (Y506D), and 628 (L628T) in the interface of the two regulatory domains of the homodimeric protein, besides being disease associated, are correctly predicted as interaction sites, and predicted also as destabilizing. This confirms the role of the regulatory domains interaction in supporting the homodimeric functional unit [5].

Finally, we propose a structural variational model for MTHFR deficiency by associating variation types to the protein architecture, as modelled with HMMs representing the catalytic and regulatory domain, respectively.

Supplementary Materials: The following supplementary files are available online at <https://www.mdpi.com/article/10.3390/ijms23010167/s1>.

Author Contributions: Conceptualization, R.C., C.S., P.L.M.; data curation: G.B., C.S., D.B.; data analysis: C.S., G.B., D.B.; writing, R.C., C.S., P.L.M.; supervision, R.C. All authors have read and agreed to the published version of the manuscript.

Funding: The work was supported by the PRIN2017 grant (project 2017483NH8_002), delivered to CS from the Italian Ministry of University and Research.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available in the tables, figures and Supplementary Materials of this article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Froese, D.S.; Huemer, M.; Suormala, T.; Burda, P.; Coelho, D.; Guéant, J.L.; Landolt, M.A.; Kožich, V.; Fowler, B.; Baumgartner, M.R. Mutation Update and Review of Severe Methylenetetrahydrofolate Reductase Deficiency. *Hum. Mutat.* **2016**, *37*, 427–438. [[CrossRef](#)] [[PubMed](#)]
2. Ducker, G.S.; Rabinowitz, J.D. One-Carbon Metabolism in Health and Disease. *Cell Metab.* **2017**, *25*, 27–42. [[CrossRef](#)] [[PubMed](#)]
3. Froese, D.S.; Fowler, B.; Baumgartner, M.R. Vitamin B12, folate, and the methionine remethylation cycle-biochemistry, pathways, and regulation. *J. Inherit. Metab. Dis.* **2019**, *42*, 673–685. [[CrossRef](#)] [[PubMed](#)]

4. Yamada, K.; Chen, Z.; Rozen, R.; Matthews, R.G. Effects of common polymorphisms on the properties of recombinant human methylenetetrahydrofolate reductase. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 14853–14858. [[CrossRef](#)] [[PubMed](#)]
5. Bhatia, M.; Thakur, J.; Suyal, S.; Oniel, R.; Chakraborty, R.; Pradhan, S.; Sharma, M.; Sengupta, S.; Laxman, S.; Masakapalli, S.K.; et al. Allosteric inhibition of MTHFR prevents futile SAM cycling and maintains nucleotide pools in one-carbon metabolism. *J. Biol. Chem.* **2020**, *295*, 16037–16057. [[CrossRef](#)] [[PubMed](#)]
6. Froese, D.S.; Kopec, J.; Rembeza, E.; Bezerra, G.A.; Oberholzer, A.E.; Suormala, T.; Lutz, S.; Chalk, R.; Borkowska, O.; Baumgartner, M.R.; et al. Structural basis for the regulation of human 5,10-methylenetetrahydrofolate reductase by phosphorylation and S-adenosylmethionine inhibition. *Nat. Commun.* **2018**, *11*, 2261–2273. [[CrossRef](#)] [[PubMed](#)]
7. Burda, P.; Schäfer, A.; Suormala, T.; Rummel, T.; Bürer, C.; Heuberger, D.; Frapolli, M.; Giunta, C.; Sokolová, J.; Vlášková, H.; et al. Insights into severe 5,10-methylenetetrahydrofolate reductase deficiency: Molecular genetic and enzymatic characterization of 76 patients. *Hum. Mutat.* **2015**, *36*, 611–621. [[CrossRef](#)] [[PubMed](#)]
8. Frosst, P.; Blom, H.J.; Milos, R.; Goyette, P.; Sheppard, C.A.; Matthews, R.G.; Boers, G.J.H.; den Heijer, M.; Kluijtmans, L.A.J.; van den Heuvel, L.P.; et al. Worldwide distribution of a common methylenetetrahydrofolate reductase mutation. *Nat. Genet.* **1995**, *10*, 111–113. [[CrossRef](#)] [[PubMed](#)]
9. Skibola, C.F.; Smith, M.T.; Kane, E.; Roman, E.; Rollinson, S.; Cartwright, R.A.; Morgan, G. Polymorphisms in the methylenetetrahydrofolate reductase gene are associated with susceptibility to acute leukemia in adults. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 12810–12815. [[CrossRef](#)] [[PubMed](#)]
10. Bezerra, G.A.; Hostenstein, A.; Foster, W.R.; Xie, B.; Hicks, K.G.; Bürer, C.; Lutz, S.; Mukherjee, A.; Sarkar, D.; Bhattacharya, D.; et al. Identification of small molecule allosteric modulators of 5,10-methylenetetrahydrofolate reductase (MTHFR) by targeting its unique regulatory domain. *Biochimie* **2021**, *183*, 100–107. [[CrossRef](#)] [[PubMed](#)]
11. Savojardo, C.; Fariselli, P.; Martelli, P.L.; Casadio, R. ISPPRED4: Interaction sites prediction in protein structures with a refining grammar model. *Bioinformatics* **2017**, *33*, 1656–1663. [[CrossRef](#)] [[PubMed](#)]
12. Casadio, R.; Savojardo, C.; Fariselli, P.; Capriotti, E.; Martelli, P.L. Turning failures into applications: The problem of protein $\Delta\Delta G$ prediction. *Methods Mol. Biol.* **2021**, *in press*.
13. Savojardo, C.; Fariselli, P.; Martelli, P.L.; Casadio, R. INPS-MD: A web server to predict stability of protein variants from sequence and structure. *Bioinformatics* **2016**, *32*, 2542–2544. [[CrossRef](#)] [[PubMed](#)]
14. Schymkowitz, J.; Borg, J.; Stricher, F.; Nys, R.; Rousseau, F.; Serrano, L. The FoldX web server: An online force field. *Nucleic Acids Res.* **2004**, *33*, W382–W388. [[CrossRef](#)] [[PubMed](#)]
15. Pucci, F.; Bernaerts, K.V.; Kwasigroch, J.M.; Rooman, M. Quantification of biases in predictions of protein stability changes upon mutations. *Bioinformatics* **2018**, *34*, 3659–3665. [[CrossRef](#)] [[PubMed](#)]
16. Savojardo, C.; Manfredi, M.; Martelli, P.L.; Casadio, R. Solvent accessibility of residues undergoing pathogenic variations in Humans: From protein structures to protein Sequences. *Front. Mol. Biosci.* **2021**, *7*, 626363. [[CrossRef](#)]
17. Casadio, R.; Vassura, M.; Tiwari, S.; Fariselli, P.; Martelli, P.L. Correlating disease-related mutations to their effect on protein stability: A large-scale analysis of the human proteome. *Hum. Mutat.* **2011**, *32*, 1161–1170. [[CrossRef](#)] [[PubMed](#)]
18. Savojardo, C.; Babbi, G.; Martelli, P.L.; Casadio, R. Functional and structural features of disease-related protein variants. *Int. J. Mol. Sci.* **2019**, *20*, 1530. [[CrossRef](#)] [[PubMed](#)]
19. Savojardo, C.; Babbi, G.; Martelli, P.L.; Casadio, R. Mapping OMIM disease-related variations on protein domains reveals an association among variation type, Pfam models, and disease classes. *Front. Mol. Biosci.* **2021**, *8*, 617016. [[CrossRef](#)] [[PubMed](#)]

Disease-related variation types in human genes link to maladies with Pfam and InterPro mapping

Giulia Babbi^{1,*}, Castrense Savojardo^{1,*}, Davide Baldazzi¹, Pier Luigi Martelli¹, Rita Casadio^{1,2}

¹ Biocomputing Group, Department of Pharmacy and Biotechnology, University of Bologna, Italy

² Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies (IBIOM), Italian National Research Council (CNR), Bari, Italy

* Authors equally contributed to the work

Correspondence to: pierluigi.martelli@unibo.it

Authors' ORCID:

GB 0000-0002-9816-4737

CS 0000-0002-7359-0633

DB 0000-0001-5731-3663

PLM 0000-0002-0274-5669

RC 0000-0002-7462-7039

Abstract

We previously introduced the concept of disease-related variation type by grouping variations according to their physicochemical properties. Here, by using a large data set of proteins with disease related and benign variations, as derived by merging Humsavar and ClinVar data, we investigate to which extent our physicochemical grouping procedure can help in determining whether patterns of variation types are related to specific groups of diseases and whether they occur in Pfam and/or InterPro gene domains.

We download 75,145 germline disease-related and benign variations of 3,605 genes, group them according to physicochemical categories and map them into Pfam and InterPro gene domains.

Statistically validated analysis indicates that each cluster of genes associated to Mondo anatomical system categorizations is characterized by a specific variation pattern. Patterns identify specific Pfam and InterPro domain–Mondo category associations. Our data suggest that the association of variation patterns to Mondo categories is unique and may help in associating gene variants to genetic diseases.

Keywords: Disease Associated Variant; Variation physicochemical type; Pfam domain; Interpro Domain; Mondo anatomical system categories.

Background

Modern sequencing technologies and intensive research on the molecular origins of humans are increasing exponentially the number of missense single-nucleotide mutations leading to observable changes in protein sequences, and evidently, in their function. For many of these single residue variations (SRVs), links to disease are reported in public databases such as Humsavar (The UniProt Consortium 2021, <https://www.uniprot.org/docs/humsavar>), the UniProt dataset of human missense variants, and ClinVar (Landrum MJ et al 2018, <https://www.ncbi.nlm.nih.gov/clinvar>), the NCBI resource of relationships among human variations and disease phenotypes.

In this scenario, harmonisation of disease definition is an issue for a better association of molecular events to phenotypes (McInnes et al 2021). Recently the Mondo Disease Ontology, in its semi-automatic version that includes also manual curation (Mungall CJ et al 2017, <https://mondo.monarchinitiative.org/>), integrates multiple disease resources to yield a coherent merged ontology. Furthermore, thanks to the interoperability provided by the Ontology Lookup Service (part of the ELIXIR infrastructure, <https://elixir-europe.org/>), it is now available for browsing (<https://www.ebi.ac.uk/ols/ontologies/mondo>), making it feasible to merge data from different databases for a larger inclusion of variations when characterising variant disease-association. Indeed, the relationship between sequence variation and disease predisposition can identify processes that are responsible of pathogenesis and can help in highlighting new treatments (McCarthy et al 2017; Clausnitzer et al 2020, Sheils et al 2021).

More to this, genome-wide association studies (GWAS) have identified thousands of noncoding loci that are associated with human diseases and complex traits, each of which could reveal insights into the mechanisms of disease. Particularly interesting is the network of genome-wide enhancers, which links variations to target disease genes recently described (Nasser et al 2021, and references therein). This stands from the estimation of which enhancers regulate which genes in the genome and the enhancer-promoter contact frequency from epigenomic datasets, supporting the general notion that variations and gene-mediated disease associations are a very complex phenomenon, which occurs at the cell level (Nasser et al 2021, <https://www.engreitzlab.org/resources/>).

On the other hand, computational methods try to establish rules of associations between variations and diseases with the purpose of helping the annotation process of the newly sequenced variants, exomes, and genomes (for recent implementations see Pei and Grishin 2021, Woodard J et al 2021, and references therein). Methods rely on inference processes standing upon the knowledge present in databases and require validated sets of variation-disease associations (Glusman et al, 2017, Peng et al 2019, Sakar et al 2020, Vihinen 2021).

With the increasing amount of available data, we are now interested in understanding to which extent gene structural and functional features may help in relating variations to diseases. For this, we decided to focus on structural and functional mapping genes and their variants with Pfam and InterPro domains (Mistry et al, 2021, <https://pfam.xfam.org/> and

<https://www.ebi.ac.uk/interpro/about/interpro>, respectively). We found that in human proteins pathogenic variations group into variational patterns that differ depending on the Pfam domain and the group of diseases they link (Savojardo et al 2019, Savojardo et al 2021 a and b). Here, we extend the analysis to a much larger data set of germline variations generated by the union of Humsavar and ClinVar. Besides Pfam, we include functional features as described by InterPro domains and find that Pfam and InterPro regions, covering most of the union data set, specifically link variations to associated diseases. Furthermore we show that different Mondo categories link different Pfam and InterPro regions in a significant manner, supporting the notion that a specific disease may link the gene variant knowing the location of the corresponding variations in specific structural or functional domains.

Materials and methods

Data collection

Variations were collected from Humsavar (UniProt Consortium 2021, <https://www.uniprot.org/docs/humsavar>) and ClinVar (Landrum MJ et al 201, <https://www.ncbi.nlm.nih.gov/clinvar/>) along with the annotation of their effect on human health following the classification scheme of the American College of Medical Genetics and Genomics/Association for Molecular Pathology terminology (Richards et al. 2015). In this work, we focus on germline variations, and we identify genes with the corresponding UniProt reference protein. ClinVar adopts a more detailed labelling than Humsavar. For sake of simplicity, ClinVar variations labelled as Likely Pathogenic or Pathogenic (LP/P), Pathogenic (P) and Likely Pathogenic (LP) were merged into a unique LP/P class, like in Humsavar. Similarly Likely Benign or Benign (LB/B), Likely Benign (LB) and Benign (B) were grouped in the class LB/B, following Humsavar. Furthermore, LB/B variations were collected only when associated to genes with disease-related variations. Variations of Uncertain Significance were discarded from both databases.

We collected our data set, adopting the following procedure.

- i) From Humsavar (release: 8/04/2021) we collected 30,415 unique single residue variations annotated as LP/P in 3,043 genes and their included LB/B variations; from ClinVar (release: 29/03/2021) we extracted 38,415 missense variations annotated as pathogenic, likely pathogenic or pathogenic/likely pathogenic in 3,842 genes and their included LB, B and LB/B variations. With this, we consider only LB/B variations in disease associated genes.
- ii) Gene variations were mapped on the corresponding UniProt canonical protein sequences by means of the RefSeq transcript (NM) and protein (NP and WP) accessions. We found that 93% of the whole variation set mapped to the UniProt

canonical sequence. We checked the consistency between the protein sequence and the wild type residue of the reported missense variation.

- iii) Somatic variations and variations with contrasting effect in the two databases were discarded.
- iv) Associations of gene variations to specific diseases were retrieved by means of the OMIM disease codes (Amberger et al 2019) in Humsavar and of the OMIM, Orphanet, HPO, MeSH, and Mondo codes in ClinVar.
- v) Associated diseases were annotated with the “disease or disorder” branch in the Mondo ontology (Mungall CJ et al 2017, <https://www.ebi.ac.uk/ols/ontologies/mondo>), apart from 71 OMIM diseases without any IDs in Mondo. All the variations associated to diseases without an OMIM and/or a Mondo ID were discharged.

Disease classification

We classify diseases following the Mondo “Disease by Anatomical System” categorization, as reported by EMBL-EBI Ontology Lookup Service (<https://www.ebi.ac.uk/ols/index>). According to this Mondo categorisation (<http://obofoundry.org/ontology/mondo.html>), diseases group in relation to their effects on the functioning of an organ system. For sake of brevity, when necessary, we arbitrary label the 14 Mondo “Disease by Anatomical System” categories as follows: **A**-respiratory system disease, **B**-auditory system disease, **C**-immune system disease, **D**-digestive system disease, **E**-disease of the genitourinary system, **F**-hematologic disease, **G**-endocrine system disease, **H**-urinary system disease, **I**-integumentary system disease, **J**-cardiovascular disease, **K**-musculoskeletal system disease, **L**-disease of the visual system, **M**-nervous system disorder, **N**-mediastinal disease.

5,223 Mondo IDs are classified in 13 of the 14 Mondo anatomical system categories, except for the “mediastinal disease” anatomical category, which includes only one variation, and it has been therefore excluded from the analysis.

Pfam and InterPro annotation

Pfam annotations (version 33.1) were downloaded for the human proteome from the Pfam FTP server (<ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam33.1/proteomes/9606.tsv.gz>). Annotations were filtered to retain only those occurring in genes included in our dataset and covering at least one pathogenic SRV.

Analogously, InterPro annotations including all signatures for human genes were extracted from the complete UniProt protein annotation file available in the InterPro website (<ftp.ebi.ac.uk/pub/databases/interpro/protein2ipr.dat.gz>). We retained only InterPro signatures mapping on genes in our set and covering pathogenic SRVs.

Statistical validation

The significance of the observed difference between Pfam/InterPro-specific distributions of variation types and Mondo anatomical system categories against respective background distributions has been assessed using an FDR-corrected Chi-squared test. Given a domain-specific observed counts $c_o = (c_o^1, \dots, c_o^K)$ for K possible events (either counting SRV types or Mondo categories) and a corresponding background distribution $f_b = (f_b^1, \dots, f_b^K)$, we compute the Chi-squared test statistics as:

$$\chi^2 = \sum_{i=1}^K \frac{(c_o^i - f_b^i N_o)^2}{f_b^i N_o} \quad (1).$$

Where $N_o = \sum_{i=1}^K c_o^i$ is the total number of observations.

P-values are then computed using a χ^2 distribution with $K - 1$ degrees of freedom, where K is the number of events. False-discovery rate (FDR) correction is also applied to correct p-values for multiple testing. We computed statistical validation for classes with at least 20 observations.

Computation of log-odds

Given a domain-specific (either Pfam or InterPro) observed frequencies f_o (either the frequency of SRV types or Mondo categories) and a corresponding background distribution f_b , we compute log-odd scores as follows:

$$LOGD = \log \frac{f_o}{f_b} \quad (2).$$

For avoiding numerical errors in the computation of the logarithm, we introduced pseudocounts when computing f_o .

When appropriate, we report the median value of variations per protein, grouped according to the Pfam/InterPro domains, to highlight the central value of the distribution, independently of outliers.

Results

The Union data set

Our dataset is described in Table 1. When the union between Humsavar and ClinVar is considered (Union), it includes 75,145 variations (43,917 of which are pathogenic) in 3,605 genes. Pathogenic variations (LP/P) are linked to 5,223 diseases. Humsavar and ClinVar differently contribute to the Union data set; interestingly ClinVar contributes with a larger LB/B number of variations and a larger number of diseases to Union (Table 1, between brackets). When LP/P variations are annotated with OMIM or Mondo codes in both datasets, the overlap between the lists of associated diseases is 82.4%. Considering the 2,576 shared genes, the overlap of the associated diseases between ClinVar and Humsavar is 74.2%.

Table 1. General description of the Union dataset

	Humsavar #	ClinVar #	°Intersection #	°Union #
Disease-associated genes	2,984 (408)*	3,197 (621)*	2,576	3,605
Variations in disease-associated genes	41,693 (25,035)*	50,110 (33,452)*	16,658	75,145
- Pathogenic	29,579 (17,371)*	26,546 (14,338)*	12,208	43,917
- Benign	12,114 (7,664)*	23,564 (19,114)*	4,450	31,228
Associated diseases[^]	3,898 (593)*	4,629 (1,324)*	3,305	5,223

° Intersection, ° Union: Intersection and Union of Humsavar and ClinVar, respectively.

[^]Mondo IDs (5152) and OMIM (71)

* Between brackets: Exclusive items for each database, included in Union.

Union genes and their disease association

The molecular function of the 3,605 genes in the Union dataset has been derived from the UniProt entries of their encoded proteins. We considered the annotation in terms of 30 high-level terms of the Molecular Function branch of the Gene Ontology (GO-MF; Gene Ontology Consortium 2021, <http://geneontology.org/>) and of the Enzyme Commission numbers (EC; Pundir et al., 2017). Some 38% of the dataset consist of enzymes: 1230 proteins are endowed with one or more EC number (Supplementary Table 1). Some 136 are annotated with a catalytic activity (GO:000382) and 15 are annotated as ATPases (GO:0016887) without EC number.

The other high-level GO-MF terms significantly over-represented in our dataset are GO:0140110 (transcription regulator activity, 277 genes), GO:0005198 (transporter activity, 239 genes), GO:0005198 (structural molecule activity, 159 genes), GO:0098772 (molecular function regulator activity, 135 genes), GO:0060089 (molecular transducer activity, 119 genes). GO:0005488 (binding) annotates 598 genes and the remaining high-level GO classes

for MF account for a total of 76 proteins. Multiple high-level GO-MF terms are annotated for 308 genes and 313 genes lack GO-MF annotation.

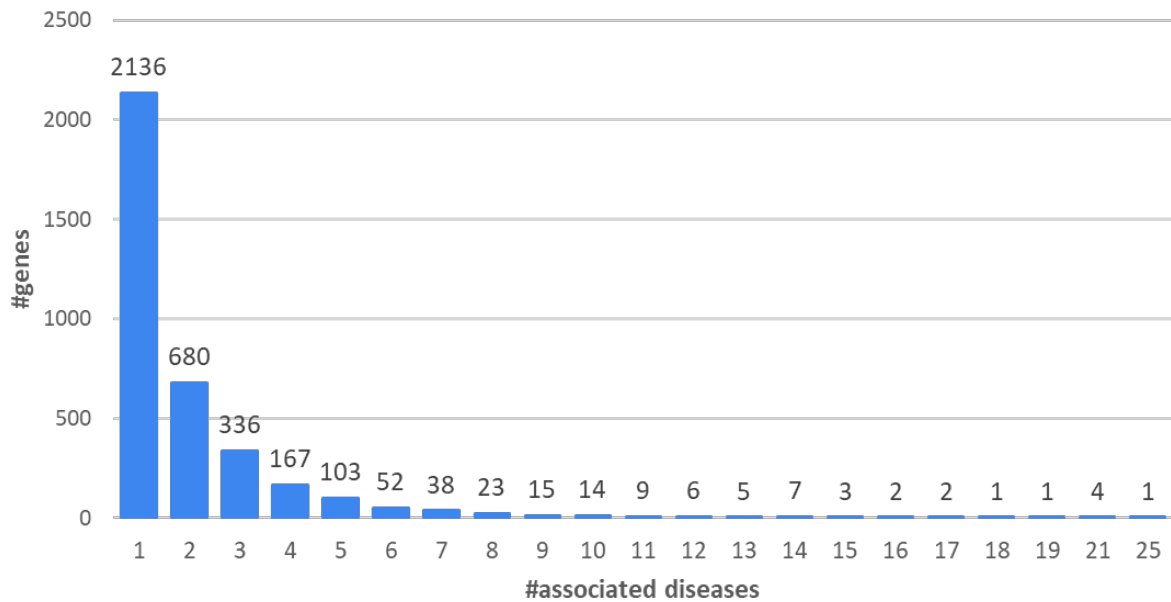


Fig. 1 Distribution of Union genes as a function of the number of associated diseases (5,223 diseases) (Table 1).

Union genes are associated to diseases (Fig.1) and 59% of the genes are associated to one disease. 41% of the Union genes are associated to more than one disease. Genes associated with the highest numbers of diseases are Fibrillin, (FBN1, UniProt code: P35555), the GTPase KRas (KRAS, UniProt code: P01116), the Cellular tumor antigen p53 (TP53, UniProt code: P04637) and the Collagen alpha-1(II) chain (COL2A1, UniProt code: P02458), with 21 disease-associations. Prelamin-A/C (LMNA, UniProt code: P02545) is associated with 25 diseases.

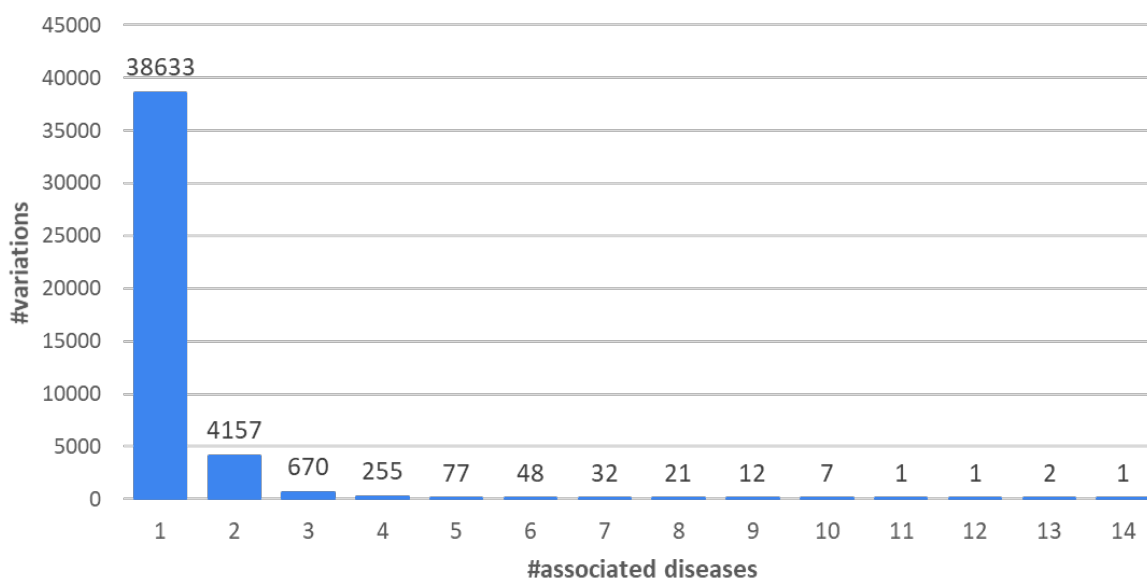


Fig. 2 Distribution of the 43,917 LP/P variations in the Union data set as a function of the number of associated diseases (5,223).

Union variations are listed as a function of the number of associated diseases, as represented by Mondo IDs and 71 OMIM codes (Fig.2). 88% of the variations have only one disease-association. The variation associated with more diseases (14 in Fig.2) is P250R on FGFR3, the Fibroblast growth factor receptor 3 (UniProt code: P22607). Its variation is associated to the Muenke syndrome (MNKS), a condition characterized by coronal craniosynostosis, which affects the shape of the head and face, often with a decrease in the depth of the orbits and hypoplasia of the maxillae. Therefore, the variation, linked to 14 Mondo IDs, maps to 5 Mondo anatomical system categories (E, H, I, K, L; see Disease classification in Materials and Methods).

For finding distinguished features among genes, variations, and diseases, we first grouped the disease related variations by variation types. To this aim, we firstly grouped residues according to their physicochemical properties, obtaining four major groups: nonpolar (GAVPLIM), aromatic (FWY), polar (STCNQH) and charged (DEKR) residues. We define a variation type in relation to the conservation or substitution of nonpolar (a), polar (p), aromatic (r) and charged (c) residues (Savojardo et al 2019). Variations are then grouped into the 16 possible variation types, which allows to distinguish between residue substitutions which may affect protein stability and function based on the notion of being conservative or not, respectively. Results are in Fig.3, show the different distribution of pathogenic versus benign variations in the different types.

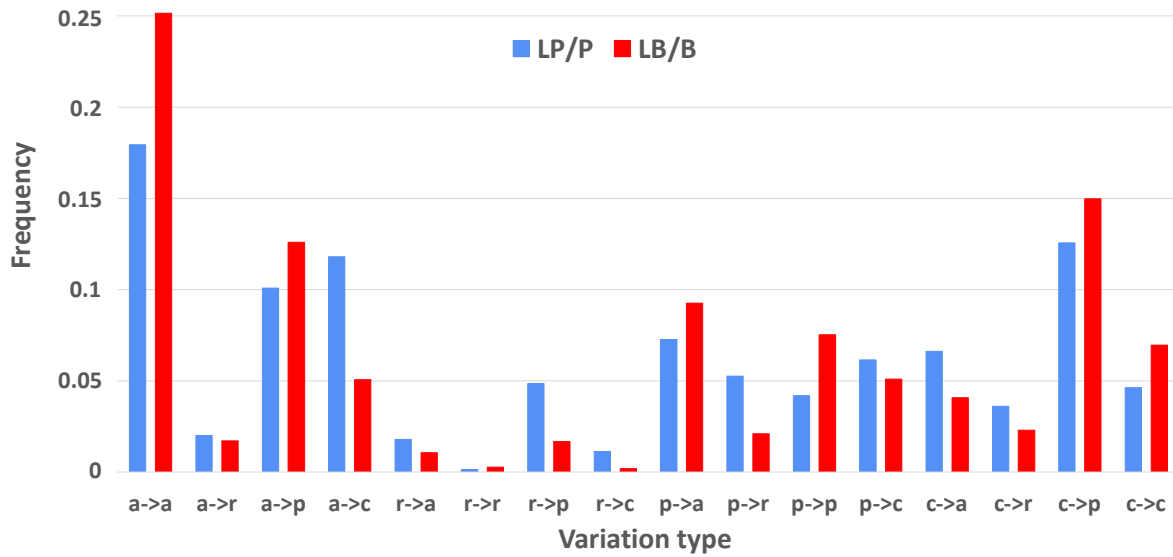


Fig. 3 Frequency of variation types of the Union variations. Blue bars: LP/P variations; Red bars: LB/B variations. Labels are as follows: a, nonpolar; r, aromatic; p, polar; and c, charged.

Disease related and benign variations have a different distribution and from now on we will focus on disease related variations, being our goal to explore gene-disease association. The most abundant types of disease related variations are nonpolar into nonpolar, polar, and charged, respectively, and charged into polar. These results agree with the more frequent variation types that we described as disease associated in a much smaller data set (Savojardo et al 2019).

The relationship among pathogenic variations associated to Mondo IDs and Mondo anatomical system categories is shown in the heat map of Fig. 4. Here we list as a function of the variation type, all the variations which are associated to the different Mondo anatomical system categories. For sake of clarity, we include the number of diseases in the set, the genes (*italic*) and the number of disease-related variations. The color-coded heat map indicates that for each category, the pattern of disease related variation types is different. A statistical validation of our findings is in Supplementary Table 2. To better highlight over/under-representation, we show log-odds between each disease-type distribution and the background frequency of LP/P variations in the whole dataset (Supplementary Fig. 1).

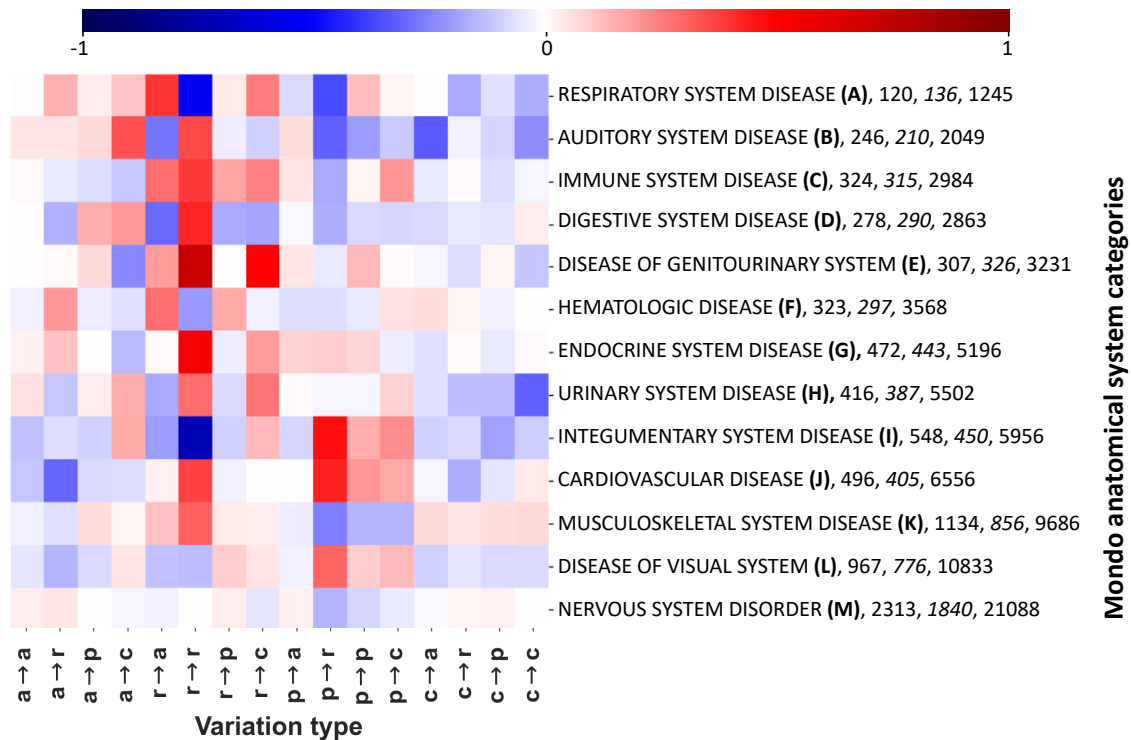


Fig. 4 Log-odd scores of variation types associated to the different Mondo anatomical system categories. The heatmap shows the log-odd score of each variation type with respect to the corresponding LP/P background (shown in Supplementary Fig.1). For each Mondo category, we show the number of diseases, genes (*italic*) and disease related variations. In variation types, labels are as follows: a, nonpolar; r, aromatic; p, polar; and c, charged. Statistical validation is reported in Supplementary Table2.

Pfam and InterPro coverage

In the following we take advantage of Pfam and InterPro coverage of each single gene to locate disease related variation types into structural and functional regions (Table 2). Pfam entries cover at least one pathogenic variant in 2,987 genes (83% of the 3,605 Union disease related genes, Table 1). Overall, 1,949 Pfam entries are identified in Union genes, including 32,575 pathogenic variations (74%). 1685 Pfams are endowed with an associated PDB structural domain. This analysis complements and confirm previous observation in a smaller data set (Savojarado et al. 2019, Savojarado et al. 2021 a and b).

InterPro (<https://www.ebi.ac.uk/interpro/>), which integrates Pfam annotations with signatures taken from other member databases such as PROSITE, PRINTS and PANTHER provide a larger number of functional regions. Indeed, with InterPro mapping we further enlarge the coverage at both gene and variation levels and can include some more 8,515 pathogenic variations in 459 genes (Table 2).

156 disease genes (4% of the total) do not have Pfam and/or InterPro domains including their pathogenic SRV positions. Finally, three SwissProt disease genes (Dentin sialophosphoprotein (UniProt: Q9NZW4), Uncharacterized protein FAM120AOS (UniProt: Q5T036) and Ribitol-5-

phosphate xylosyltransferase 1 (UniProt: Q9Y2B1) do not have Pfam and/or InterPro signatures.

Table 2. Pfam and InterPro coverage statistics

	Pfam #	InterPro #
Union genes with at least one pathogenic variant in a Pfam and/or InterPro region	2,987 (83%) ^a	3,446 (96%) ^a
Domains covering pathogenic variants	1,949	5,357
Pathogenic variants in Pfam and/or InterPro regions	32,575 (74%) ^b	41,090 (94%) ^b
Benign variants in Pfam and/or InterPro regions	13,195 (42%) ^c	24,461 (78%) ^c

^a Percentages computed with respect to the total number of diseases associated Union genes (3,605, Table 1).

^b Percentages computed with respect to the total number of pathogenic variants (43,917, Table 1)

^c Percentages computed with respect to the total number of benign variants (31,228, Table 1)

A complete list of the Pfam and InterPro regions, detailed for each gene, is reported in Supplementary Table 1. For each gene, we report the accession, the name, the functional annotation (EC, GO MF), the list of Pfam and InterPro domains, the numbers of pathogenic variations and associated disease, the disease name and the associated Mondo disease anatomical system categories. Results highlight that the Pfam domain covering the highest number of disease related genes (62) is Pkinase (PF00069) while the domain mostly enriched in pathogenic variations (1,566) is Ion_trans (PF00520). Supplementary Table 1 lists also the results obtained with the InterPro coverage. Among the most abundant InterPro entries we found many conserved, binding, and active sites (as expected, being these important sites driving the gene/protein function). Some of them are within Pfam domains: e.g., the Homeobox_CS (IPR017970), included in the Homeodomain (PF00046) domain. This finding provides an additional specification of the most critical regions containing pathogenic variations.

Distinctive patterns of pathogenic variation types within Pfam and InterPro regions

After structural and functional Pfam and IterPro gene mapping, we can analyze the relationship among variation type and diseases (grouped by Mondo anatomical system categories). With the concept of variation types (Fig.3), the 16 different SRV types can be associated to individual Pfam and InterPro (complete results are provided in Supplementary

Table 3, which for Pfam and InterPro entry, include the number of genes, the number of LP/P variations, the frequencies of the variation type, the statistical validation and log-odds scores between domain-specific distributions and LP/P background frequency).

In Fig.5 we show the log-odd scores of pathogenic variation types for the 20 most populated Pfam domains (Fig. 5). Pfams are sorted by the number of genes covered. For each domain, we report its Pfam accession and name with the number of genes and pathogenic SRVs covered, respectively (within parentheses). Overall, the 20 Pfams shown in Fig. 5 cover 557 genes and 6,729 pathogenic SRVs, corresponding to 19% and 21% of the total number of Pfam-covered genes and SRVs, respectively (Table 2). In particular, genes covered by 6 out of 20 Pfams (p450, Pkinase, Ras, Trypsin, Helicase_C and PK_Tyr_Ser-Thr) are mainly associated with enzymatic activities, 2 (Homeodomain and zf-C2H2) occur in proteins performing transcription regulation activities (GO:0140110), 2 (Filament and Collagen) cover structural proteins (GO:0005488), 2 (Mito_carr and Ion_trans) are in transporters (GO:0005215), 1 (7tm_1) cover transducers (GO:0060089), 1 (Hormone_recep) is associated with proteins performing either transduction or transcription regulation activities, 1 (Neur_Chann_memb) is found in proteins associated to transport or transduction. The remaining 4 domains (fn3, EGF_CA, I-set, and Cadherin) have multiple associated functions and mainly act as mediators of interactions in proteins associated with a diverse range of functional activities.

Noticeably, the different Pfam domains show a distinctive variational pattern with significant deviations from the background distribution. Overall, our results confirm over a larger dataset, previous observations (Savojardo et al., 2021 b). Statistical validation and resulting FDR-corrected p-values for each Pfam entry are also reported in Supplementary Table 3.

A similar analysis is performed for those InterPro regions that do not include Pfam domains (Fig. 6 and Supplementary Table 3). The 20 InterPro entries in Fig. 6 cover 836 genes and 9,208 pathogenic SRVs, corresponding to 24% and 22% of total number of InterPro-covered genes and SRVs, respectively (Table 2). Among the 20 InterPros, 9 cover proteins that are clearly associated to specific functions: 6 InterPros (Kinase-like_dom_sf, Znf_RING/FYVE/PHD, Protein_kinase_ATP_BS, Tyr_kinase_cat_dom, P-loop_NTPase and NAD(P)-bd_dom_sf) cover enzymes while 3 entries (Homeobox-like_sf, Homeobox_CS and Znf_C2H2_sf) are associated to transcription factors. The other 11 InterPros are predominantly (not univocally) associated with proteins having different functions, including binding activities (Growth_fact_rcpt_cys_sf, WD40/YVTN_repeat-like_dom_sf, WD40_repeat_dom, LRR_dom_sf, Ig-like_dom_sf and WD40_repeat_dom_sf), molecular transducer activities (Ig-like_fold, FN3_sf) and 2 to enzymes (Ig_sub, TPR-like_helical_dom_sf).

Also in this case, different variational patterns can be observed for different InterPro entries.

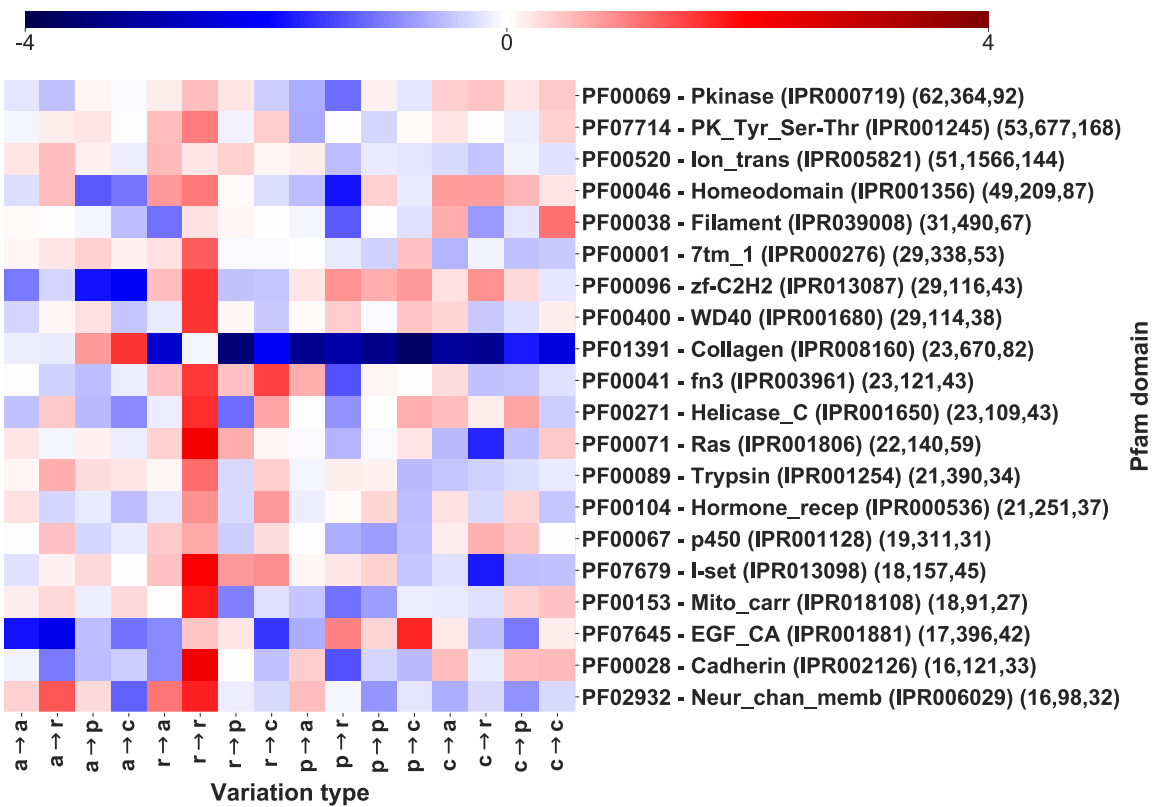


Fig. 5 Log-odd scores of variation types in Pfam entries sorted by number of genes covered (the first 20, out of 1,940 Pfams, Supplementary Table 3). Log-odds are computed with respect to the whole dataset LP/P background (Fig. 3). For each Pfam, the corresponding InterPro accession is also included. Numbers within parentheses report the number of genes, variations, and diseases, respectively. Statistical validation and resulting FDR-corrected p-values for each Pfam entry are reported in Supplementary Table 3.

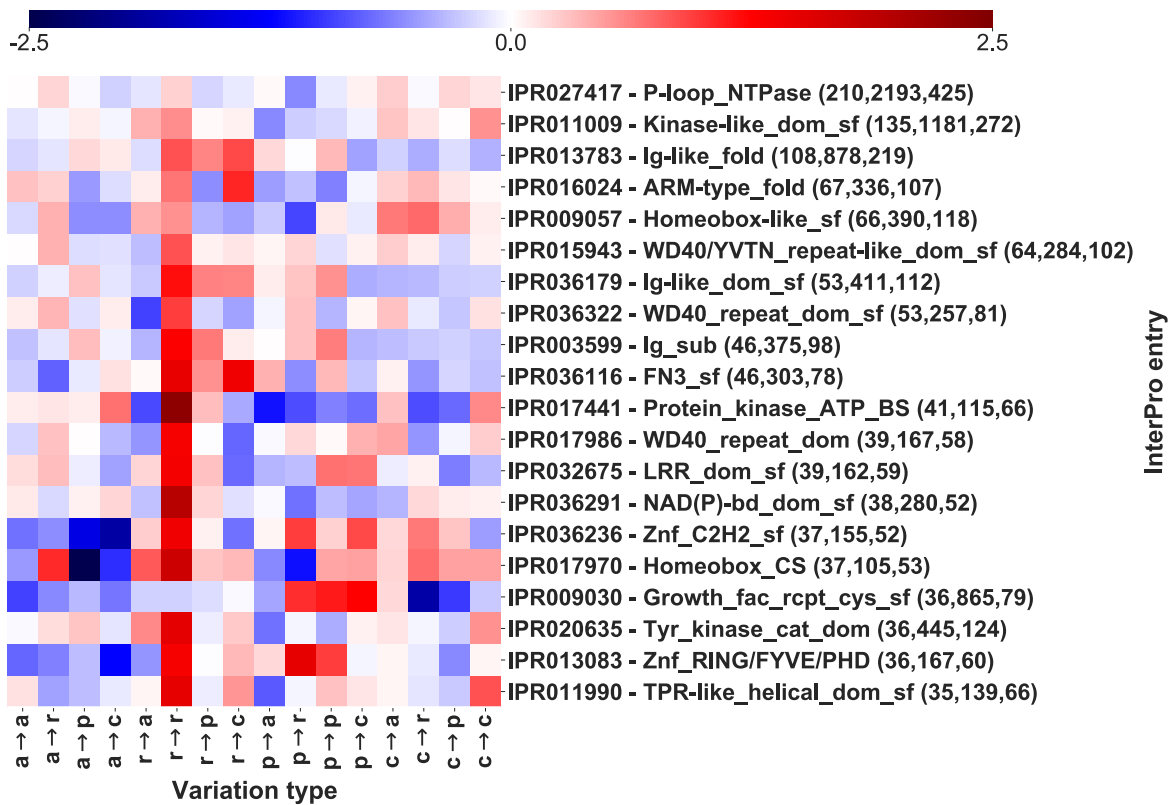


Fig.6 Log-odd scores of variation types for the first 20 InterPro entries (out of 5,357, Table 2), sorted by number of genes covered and not including Pfam signatures. Log-odds are computed with respect to the whole dataset LP/P background (Fig. 3). Numbers in parentheses report, for each InterPro, the number of genes, of SRVs and of diseases, respectively. Statistical validation and resulting FDR-corrected p-values for each InterPro entry are reported in Supplementary Table 3.

Linking Pfam/InterPro to Mondo anatomical system categories

In Fig 4, we established a relation between Mondo anatomical system categories and pathogenic variation types. In Fig. 5-6, we detailed the association among variation types and Pfam/InterPro regions in the different genes. For sake of generalization, an important question to answer is then to which extent Pfam and/or InterPro domains can be directly related to diseases grouped according to Mondo categories.

Fig. 7 shows log-odd scores for the disease Mondo categories associated to the 20 most populated Pfams (the full association with the 1949 Pfam domains covering our Union set are listed in Supplementary Table 4, also including the background distribution frequency of disease categories in the entire set and the statistical validation results).

Pfams are associated to multiple disease categories, as visible by comparing with the background signal. However, it is evident (Fig. 7) that there is often one or more prevalent category/ies with an evident and significantly high log-odd score. For instance, in the case of Trypsin domain (PF00089), about 63% of the pathogenic variations link to Hematologic

diseases (F), a percentage significantly higher than the background frequency of this type of disease in the whole set (4%). Remarkably, these SRVs come from different genes (the median number of SRVs per gene for the Trypsin domain is 8). Similar conclusions can be drawn for other domains, like Ion_trans (PF00520) particularly enriched in neurological diseases (M). Finally, similar conclusions are obtained, when a similar heat map is generated considering the relationship among Mondo anatomical system categories and InterPro regions not included in Pfam (Fig. 8, reporting log-odd scores).

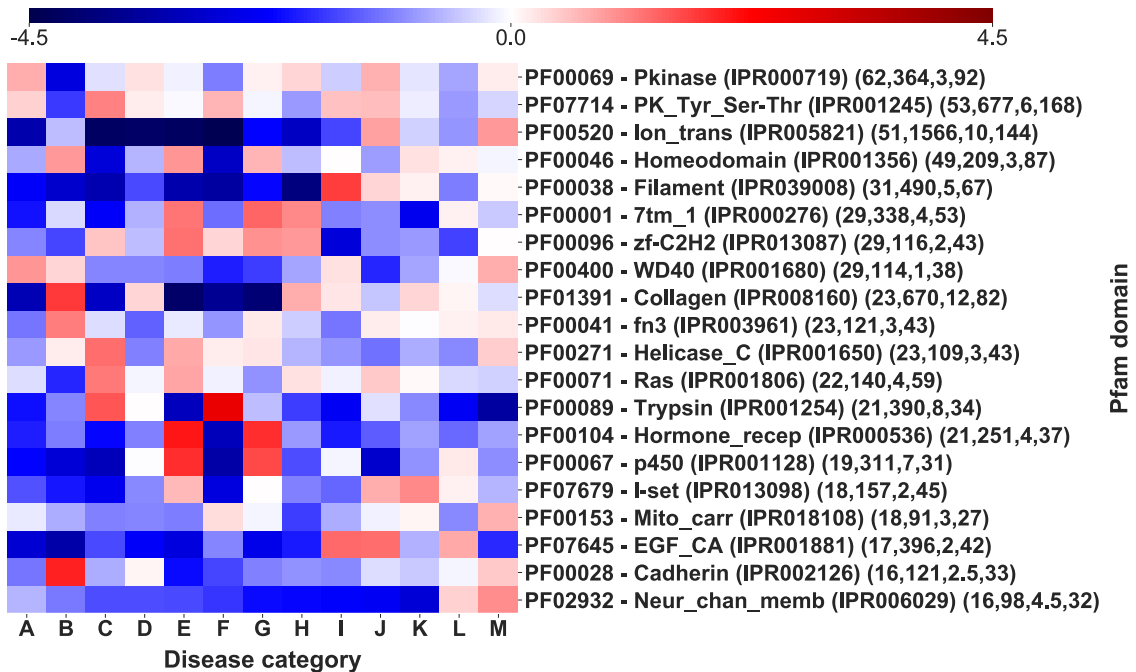


Fig. 7 Log-odd scores for disease categories associated to different Pfam domains. Log-odds are calculated with respect to the whole-dataset background of disease categories (Supplementary Table 4). For each Pfam the corresponding InterPro accession is indicated. Numbers in parentheses report the number of genes, of SRVs, **the median number of SRVs per gene** and the number of diseases (for statistical validation see Supplementary Table 4).

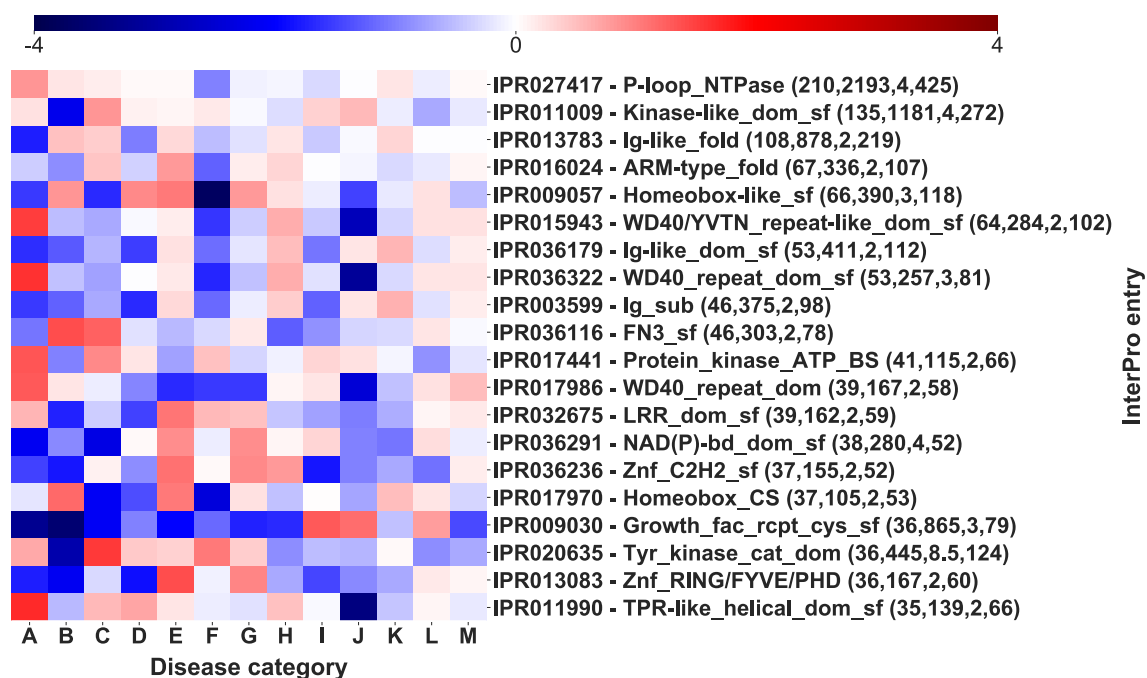


Fig. 8 Log-odd scores for disease categories associated to different InterPro domains. Log-odds are calculated with respect to the whole-dataset background of disease categories (Supplementary Table 4). Numbers in parentheses report the number of genes, of SRVs, **the median number of SRVs per gene** and the number of diseases (for statistical validation see Supplementary Table 4).

Conclusions and perspectives

We investigate the association between variants and disease with the aim of finding possible descriptors for the association of genes carrying pathological variations and the corresponding diseases. To this aim we generated a data set of variants with pathological and benign variations, union of the last releases of Humsavar and Clinvar (Table 1). Our focus are germline variations excluding somatic ones, whose associations to different types of cancers may require different ontologies.

We represent variations with variation types, which refer to their physicochemical properties. The distribution of disease related and benign variation types of the union set is different (Fig.3). We therefore focused on the pathological variations, the carrying genes and the associated diseases, grouped into the corresponding Mondo anatomical system categories. We recognise that disease related variation types are specifically and significantly linked to different Mondo categories (Fig.4) and detailed the specificity by mapping variations into Pfam and InterPro regions. We find that these regions include most of the pathological variants (Table 2) and that the Pfam and InterPro mapping (Fig. 7 and 8) significantly correlates to Mondo disease categories.

To our knowledge, this type of analysis is new and provide insights into the complex

relationship among genes, variants, and associated diseases. Our final goal is to provide a mapping of the complex space relating variations, genes, and disease by means of gene structural and functional features. This can be useful for future algorithmic developments focusing on variant annotation. Possibly, new incoming data will be framed into our basic representation and will allow a better understanding of the mechanisms eliciting specific phenotypes linked to germline variations.

Declarations

Funding

This work was supported by the PRIN2017 grant (project 2017483NH8_002), delivered to CS from the Italian Ministry of University and Research.

Conflicts of interest/Competing interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Availability of data and material

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Authors' contributions

RC supervised the study and wrote the final manuscript. GB and CS collected data and performed the analyses; DB contributed to the functional annotation of the dataset; PLM reviewed the methods and results and revised the manuscript. All the authors read and approved the final manuscript.

References

Amberger JS, Bocchini CA, Scott AF, Hamosh A (2019) OMIM.org: leveraging knowledge across phenotype–gene relationships. *Nucleic Acids Res* 47: D1038–D1043. doi.org/10.1093/nar/gky1151

Clausnitzer M, Cho JH, Collins R et al (2020). A brief history of human disease genetics. *Nature* 577:179–189. doi.org/10.1038/s41586-019-1879-7

Gene Ontology Consortium (2021) The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res* 49 (D1):D325–D334. doi: 10.1093/nar/gkaa1113

Glusman G, Rose PW, Prlić A et al (2017) Mapping genetic variations to three-dimensional protein structures to enhance variant interpretation: a proposed framework. *Genome Med* 9:113. doi: 10.1186/s13073-017-0509-y

Landrum MJ, Chitipiralla S, Brown GR et al (2020) ClinVar: improvements to accessing data. *Nucleic Acids Res.* 48, D835–D844. doi: 10.1093/nar/gkz972

McCarthy MI, MacArthur DG (2017) Human disease genomics: from variants to biology. *Genome Biol* 18:20–22. DOI 10.1186/s13059-017-1160-z

McInnes G, Sharo AG, Koleske ML et al (2021) Opportunities and challenges for the computational interpretation of rare variation in clinically important genes. *Am J Hum Genet* 108:535–548. doi: 10.1016/j.ajhg.2021.03.003

Mistry J, Chuguransky S, Williams L et al (2021) Pfam: The protein families database in 2021. *Nucleic Acids Res* 49(D1):D412–D419. doi: 10.1093/nar/gkaa913. doi: 10.1093/nar/gkaa913

Mungall JC, McMurry JA, Köhler S et al (2017) The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res* 45: D712–D722. doi.org/10.1093/nar/gkw1128

Nasser J, Bergman DT, Fulco CP et al (2021) Genome-wide enhancer maps link risk variants to disease genes. *Nature* 593: 238–243. doi.org/10.1038/s41586-021-03446-x

Pei J, Grishin NV (2021) The DBSAV Database: Predicting Deleteriousness of Single Amino Acid Variations in the Human Proteome. *Journal of Molecular Biology* 433: 166915, ISSN 0022-2836. doi.org/10.1016/j.jmb.2021.166915

Peng Y, Alexov E, Basu S (2019) Structural Perspective on Revealing and Altering Molecular Functions of Genetic Variants Linked with Diseases. *Int J Mol Sci* 20:548. doi: 10.3390/ijms20030548

Pundir S, Onwubiko J, Zaru R et al (2017) An update on the Enzyme Portal: an integrative approach for exploring enzyme knowledge. *Protein Engineering, Design and Selection* 30:247–254. doi.org/10.1093/protein/gzx008

Richards S, Aziz N, Bale S et al (2015) 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)* 17:405–424. doi.org/10.1038/gim.2015.30

Sarkar A, Yang Y, Vihinen M (2020). Variation benchmark datasets: update, criteria, quality and applications. *Database* 2020, baz117. doi.org/10.1093/database/baz117

Savojardo C, Babbi G, Martelli PL, Casadio R (2019) Functional and Structural Features of Disease-Related Protein Variants. *Int J Mol Sci* 20:1530. doi: 10.3390/ijms20071530.

Savojardo C, Manfredi M, Martelli PL, Casadio R (2021a) Solvent Accessibility of Residues Undergoing Pathogenic Variations in Humans: From Protein Structures to Protein Sequences. *Front Mol Biosci* 7:626363. Published 2021 Jan 7. doi:10.3389/fmolb.2020.626363

Savojardo C, Babbi G, Martelli PL, Casadio R (2021b) Mapping OMIM Disease-Related Variations on Protein Domains Reveals an Association Among Variation Type, Pfam Models, and Disease Classes. *Front Mol Biosci* 8:617016. DOI: 10.3389/fmolb.2021.617016.

Sheils T, Mathias S, Kelleher KJ et al (2021) TCRD and Pharos 2021: mining the human proteome for disease biology. *Nucleic Acids Res* 49(D1):D1334-D1346. doi: 10.1093/nar/gkaa993.

The UniProt Consortium (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* 49: D480–D489. doi.org/10.1093/nar/gkaa1100

Vihinen M (2021) Functional effects of protein variants. *Biochimie* 80:104-120. doi.org/10.1016/j.biochi.2020.10.009.

Woodard J, Zhang C, Zhang Y (2021) ADDRESS: A Database of Disease-associated Human Variants Incorporating Protein Structure and Folding Stabilities. *J Mol Biol* 433: 166840. doi.org/10.1016/j.jmb.2021.166840.