On analysis of protein function and variation

A PhD Thesis for the Degree of

Doctor of Philosophy

in

Computational Biology

School of Biosciences,

University of Kent

Magdalena Antczak

2020

Declaration

No part of this thesis has been submitted in support of an application for any degree or other qualification of the University of Kent, or any other University or Institution of learning.

Name: Magdalena Antczak

Date: 24.12.2020

Antczak Magdalena

Abstract

Over the last fifteen years, next-generation sequencing has overcome time and cost constraints to be widely used to generate a plethora of biological data. Around 200 million sequences are currently stored in one of the biggest protein databases, UniProt. However, less than 1% of UniProt sequences have functions supported by experimental evidence. This thesis focuses on increasing knowledge about protein functions and demonstrating why it is essential to do so.

The first aim is fulfilled by a project that used computational approaches to assign functions to the proteins of unknown function within the minimal bacterial genome. Synthesising the minimal cell was completed in 2016, and it revealed that 149 genes out of 473 had unknown functions. Our study demonstrated that using several protein function prediction methods that each explores different protein properties is an effective way to annotate unknown genes of the minimal cell. As a result, 133 out of 149 genes were assigned a function. This included 66 proteins, for which we identified more informative functions than predicted by Hutchison et al. in the initial study from 2016 (Hutchison *et al.*, 2016).

This thesis's second goal is to show how important it is to expand our knowledge about protein functions and have access to good protein function prediction methods to apply them where there is not experimental functional information. This thesis focuses specifically on applying protein function prediction methods to study the impact of DNA mutations on the proteins, which would hopefully lead to a better understanding of acquired resistance to anti-cancer therapies, which remains one of the biggest obstacles in treating cancer patients. Acquired drug resistance is developed through alterations in different molecular mechanisms such as drug efflux or binding of a drug to the target, which can sometimes be caused by a mutation in a single protein.

Here, whole-exome sequencing data of the acute myeloid leukaemia cell lines Molm13 and four Molm13 sub-lines adapted to nutlin-3 was studied to identify potential drivers of resistance. Additionally, 41 UKF-NB-3 (a neuroblastoma cell line) sub-lines adapted to tubulin-binding agents were analysed with a focus on the clonal composition of cancer cells and its impact on developing different resistance mechanisms. The analysis of *de* *novo* variants in the Molm13 sub-lines adapted to nutlin-3 revealed that three out of four sub-lines acquired loss-of-function mutations in *TP53* commonly associated with resistance to MDM2 inhibitors. Additionally, all four Molm13 sub-lines demonstrated an increased sensitivity to cytarabine that may be connected to likely deleterious *de novo* mutations identified in three out of four sub-lines in *SAMHD1*, which cleaves the triphosphorylated active form of cytarabine, causing its deactivation, while its natural function is to cleave deoxynucleoside triphosphates (dNTPs) into deoxyribonucleosides and inorganic triphosphate with the main goal of restricting viral infections by reducing dNTPs' cellular levels. These results identify *SAMHD1* mutations as a candidate biomarker for cytarabine sensitivity after the failure of MDM2 targeted therapies, which is consistent with studies demonstrating that low SAMHD1 activity, likely caused by lower SAMHD1 expression or deleterious mutations, generally tends to be associated with an increased cytarabine sensitivity in AML cells (Schneider *et al.*, 2017). Subsequently, the analysis of *de novo* variants in the UKF-NB-3 sub-lines adapted to tubulin-binding-agents demonstrated that different sub-lines adapted to the same drug can share many of the same *de novo* variants, which shows that they may have come from a similar clone. However, this is not always the case. The results revealed that different subpopulations could be selected upon the repeated adaptation of the same cancer cell line to the same drug. This emphasises the heterogeneity of processes underlying acquired resistance to anti-cancer therapies and demonstrates the need to identify these processes' biomarkers.

Acknowledgements

The ‘Acknowledgments’ section is my only chance to write in my favourite style: bullet points – so here it goes.

“No one achieves anything alone” (Leslie Knope). This PhD thesis would not have existed if it was not for the wonderful people I am surrounded by, both in person and virtually. I would like to thank:

- ***Mark Wass*** for giving me a chance to undertake a PhD in computational biology, and also for providing me with countless opportunities such as presenting at conferences and publishing my work. He has also provided me with generous support during multiple crises I have overcome during these four years.
- ***Mark Wass*** and ***Martin Michaelis*** for allowing me to analyse cancer cell lines data and for jumping through hoops to find additional funds so I could complete my projects.
- ***All the members of the Wass-Michaelis lab*** who made it a great place to work, especially ***Henry Martell***, ***Helen Grimsley***, ***Jake McGreig***, ***Stuart Masterson***, ***Stefani Dritsa*** and ***Miguel Juliá-Molina***.
- ***Jake*** and ***Stuart*** for their sense of humour with which I have a very strong love-hate relationship.
- ***Henry*** for being the greatest desk buddy that I could hope for – enduring all my questions (there were so many, especially at the beginning) and providing distractions during stressful times.
- ***Miguel,*** for introducing me to the analysis of the sequencing data and developing the first part of the pipeline that I could then expand upon.
- ***Kasia Szczepańska***, ***Ola Jakubczak***, ***Marcin Cichomski*** and ***Dorota Mikołajczak*** for all the words of encouragement that I received from them when I decided to turn my life upside-down by doing a PhD in the UK. I would like to thank them as well for the support they gave me during the PhD, especially considering the 800-mile distance between us.
- ***Helen*** and ***Kasia*** for all the conversations we had in 2020. ***Helen*** for sharing all the comedy-drama stories about her adventures and ***Kasia*** for making “a bucket for

sorrows" available all year long – this has proven to be indispensable to finishing PhD in 2020.

- My partner ***Will*** for always being there for me, for being patient and for encouraging me in all the projects that I undertook. Even from 10,000 miles away.
- ***My parents***. Chciałabym podziękować moim rodzicom za to, że od samego początku wspierali moją nieustanną potrzebę nauki. Za to, że zwalniali mnie z większości obowiązków domowych, żebym mogła rozwiązywać zadania z matematyki, czytać książki i uczyć się języków obcych. I za to, że pomimo ich własnego skromnego życia, zawsze znajdowali pieniądze, żeby sfinansować moje dodatkowe lekcje czy wysłać mnie na wymianę za granicę. Nie byłabym tu gdzie jestem, gdyby nie ich ciągłe wsparcie.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AAA protein family	ATPases Associated with diverse cellular Activities
ABC transporter	ATP-Binding Cassette transporter
ABCB1	ATP-dependent translocase ABCB1 (Multidrug resistance protein 1)
ABCC1	Multidrug resistance-associated protein 1
AI	Artificial Intelligence
AML	Acute Myeloid Leukaemia
ANOVA	ANalysis Of VAriance
API	Application Programming Interface
ATP	Adenosine Triphosphate
AUC	Area Under the Curve
BCRP	Broad substrate specificity ATP-binding cassette transporter ABCG2
BLAST	Basic Local Alignment Search Tool
BP	Biological Process
BPO	Biological Process Ontology
BRAF	Serine/threonine-protein kinase B-raf
CAFA	Critical Assessment of Function Annotation
CCO	Cellular Component Ontology
CDD	Conserved Domains Database
CNA	Copy Number Alteration
CO2	Carbon dioxide
COSMIC	Catalogue Of Somatic Mutations In Cancer
DDBJ	DNA Data Bank of Japan

DISOPRED	Disopred Prediction
DNA	Deoxyribonucleic Acid
DOOR	Database of prOkaryotic OpeRons
DSMZ	Deutsche Sammlung von Mikroorganismen und Zellkulturen
EBI	European Bioinformatics Institute
ECF transporter	Energy-coupling factor transporter
eggNOG	evolutionary genealogy of genes: Non-supervised Orthologous Groups
EMBL	European Molecular Biology Laboratory
ENA	European Nucleotide Archive
EPOB	Epothilone B
ERCC	DNA excision repair protein ERCC-1
ERI	Eribulin
FannsDB	A database for Functional ANnotations for Non Synonymous SNVs
FATHMM	Functional Analysis through Hidden Markov Models
FBS	Fetal Bovine Serum
FFPred	Feature-based Function Prediction
GATK	Genome Analysis Toolkit
GDC	Genomic Data Commons data portal
GNAQ	Guanine nucleotide-binding protein G(q) subunit alpha
gnomAD	Genome Aggregation Database
GO	Gene Ontology
GOA	Gene Ontology Annotation database
GOAT	Gene Ontology Annotation Tool

GRC	Genome Reference Consortium
HAMAP	High-quality Automated and Manual Annotation of microbial Proteomes
HEP	Inferred from High Throughput Expression Pattern
HMM	Hidden Markov Model
HPLC	High Performance Liquid Chromatography
HPO	Human Phenotype Ontology
HRAS	GTPase Hras
IC50	Half maximal inhibitory concentration
ICGC	International Cancer Genome Consortium
IDA	Inferred from Direct Assay
IDH	Isocitrate dehydrogenase
IEA	Inferred from Electronic Annotation
IGC	Inferred from Genomic Context
IMDM	Iscove's Modified Dulbecco's Medium
INDEL	Insertions and deletions
IntOGen	The Integrative OncoGenomics pipeline
ISS	Inferred from Sequence or structural Similarity
JCVI	J. Craig Venter Institute
KEGG	Kyoto Encyclopedia of Genes and Genomes
KRAS	GTPase Kras
LC–MS	Liquid Chromatography - Mass Spectrometry
LEON	multiple aLignment Evaluation Of Neighbours
MAFFT	Multiple Alignment using Fast Fourier Transform

MDM2	E3 ubiquitin-protein ligase Mdm2 (Double minute 2 protein)
MDR1	ATP-dependent translocase ABCB1 (Multidrug resistance protein 1)
ME2	2-methoxyestradiol
MF	Molecular Function
MFO	Molecular Function Ontology
MFS	Major Facilitator Superfamily
MINT	Molecular INTeraction database
MRM	Multiple Reaction Monitoring
MRP1	Multidrug resistance-associated protein 1
MSA	Multiple Sequence Alignment
MT-ND5	NADH-ubiquinone oxidoreductase chain 5
MT-ND6	NADH-ubiquinone oxidoreductase chain 6
MTT	3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyl tetrazolium bromide
MUC2	Mucin-2
MYCN	N-myc proto-oncogene protein
NAD	Nicotinamide Adenine Dinucleotide
NB	Neuroblastoma
NCBI	National Center for Biotechnology Information
NCI	National Cancer Institute
NGS	Next-Generation Sequencing
NRAS	GTPase Nras
PABPC1	Polyadenylate-binding protein 1
PBS	Phosphate Buffered Saline

PCAWG	Pan-Cancer Analysis of Whole Genomes
PCR	Polymerase Chain Reaction
PDB	Protein Data Bank
PER3	Period circadian protein homolog 3
PFP	Protein Function Prediction
PIR	Protein Information Resource
PPM1D	Protein phosphatase 1D
PSI-BLAST	Position-Specific Iterative Basic Local Alignment Search Tool
PSIC	Position-Specific Independent Counts
RCCL	Resistant Cancer Cell Line collection
RNA	Ribonucleic Acid
ROC	Receiver Operating Characteristic
SAMHD1	Deoxynucleoside triphosphate triphosphohydrolase SAMHD1
SFLD	Structure Function Linkage Database
SIFT	Sorting Intolerant From Tolerant
SLC25A5	ADP/ATP translocase 2
SMART	Simple Modular Architecture Research Tool
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
SO	Sequence Ontology
SPase	Signal peptidase
SVM	Support Vector Machine
TARGET	Therapeutically Applicable Research to Generate Effective Treatments

TBA	Tubulin-Binding Agent
TCGA	The Cancer Genome Atlas
TIGRFAM	The Institute for Genomic Research's database of protein families
TMHMM	Transmembrane Helices; Hidden Markov Model
TrEMBL	Translation of EMBL nucleotide sequence database
TrSSP	Transporter Substrate Specificity Prediction server
UniProt	Universal Protein Resource
UniProtKB	UniProt Knowledgebase
UTR	Untranslated Region
VCR	Vincristine
VEP	Variant Effect Predictor
WGS	Whole Genome Sequencing

Chapter 1 Introduction

The Human Genome Project took 13 years to complete and cost between $500 million to $1 billion (*The Cost of Sequencing a Human Genome*, 2020; Wetterstrand KA., 2020). Since its completion in 2000, the time dropped to about a day and the cost to $1000 (Figure 1.1), thus making sequencing available for high-throughput experiments (*DNA Sequencing*, 2018). The ease with which we can perform sequencing resulted in many sequences deposited in databases such as GenBank (Benson *et al.*, 2013). In April 2017, the number of sequences released by GenBank was over 200 million (200,877,884) and it has been growing steadily (see Figure 1.2). By August 2020 GenBank contained, 218,642,238 sequences, which also influences the number of available protein sequences, as 95% of the sequences in UniProt (Bateman, 2019) come from translated coding sequences deposited in ENA (European Nucleotide Archive), GenBank, and the DDBJ (DNA Data Bank of Japan) (Karsch-Mizrachi, Takagi and Cochrane, 2018). We observe exponential growth of the number of sequences in the UniProtKB database, with 195,104,019 sequence entries in the latest (7th October 2020) release of UniProtKB/TrEMBL (Bateman, 2019; *Current Release Statistics < Uniprot < EMBL-EBI*, 2020; *UniProtKB/Swiss-Prot 2020_05*, 2020) and 563,552 in the latest release of UniProtKB/SwissProt.

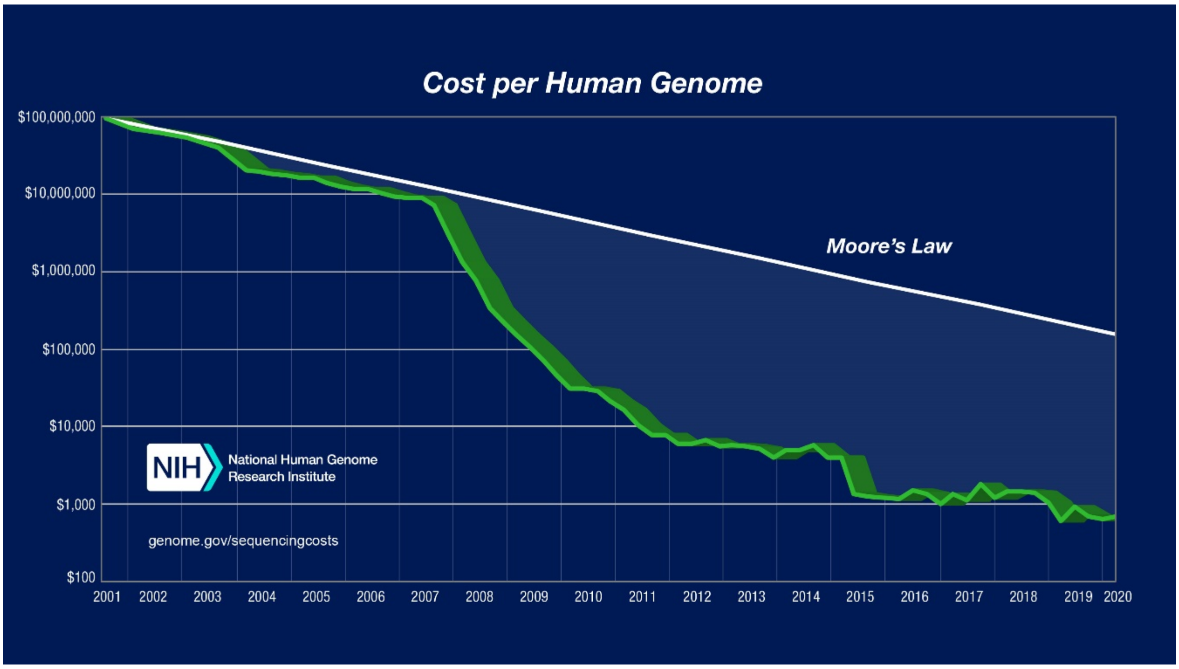


Figure 1.1 *Cost of the sequencing of a human genome from 2001 to 2020. Figure extracted from* (*The Cost of Sequencing a Human Genome*, 2020).

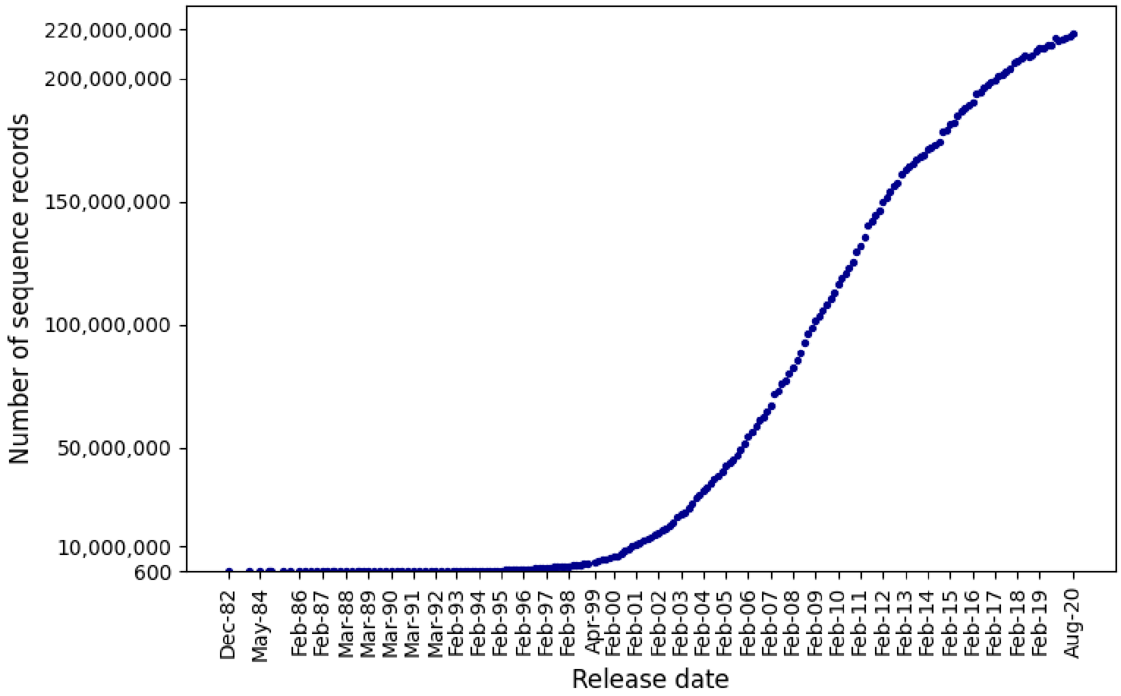


Figure 1.2 *Number of sequence records deposited in GenBank* (Benson *et al.*, 2013) *since the release in December 1982 until August 2020. The figure is based on data provided in* (*GenBank and WGS Statistics*, 2020).

Unfortunately, experimentally verified knowledge about the roles of the genes and proteins they express, lags far behind. In fact, more than 99% of functions assigned to the proteins in the UniProtKB database come from electronic annotation methods and less than 1% come from curators extracting information from papers which may incorporate functions that are experimentally confirmed (Huntley *et al.*, 2015; *About GOA | European Bioinformatics Institute*, 2020).

Many human diseases are caused by mutations of genes which in turn alter the encoded protein sequence, which may impact its function. Expanding the coverage of known protein functions is crucial for understanding the cause of disease and opening new possibilities such as identifying diagnostic tests, new targets for drugs, and also development of new medications (Bork *et al.*, 1998; Rost *et al.*, 2003; Bernardes and Pedreira, 2013).

The two research fields presented in this thesis have blossomed as a consequence of the abundance of DNA and protein sequences. The goal of protein function prediction is to fill the gap between the number of available protein sequences and those that are experimentally annotated. In turn, cancer genomics sequences and analyses cancer genomes to identify variants driving tumorigenesis and resistance to cancer therapies. These two domains of computational biology are heavily connected. It is necessary to know the function of the gene to predict the impact of mutations it harbours and how this may relate to cancer. In exchange, the information about how different variants alter the ability of the encoded proteins to perform their roles becomes an essential part of their functional annotations. Additionally, it also helps to build better function prediction tools.

1.1 Protein function prediction

The function of a protein can be determined experimentally through studying different properties of the protein. Firstly, identifying protein's subcellular location or tissue where the protein is being expressed can minimise the set of its potential functions (Punta and Ofran, 2008). Secondly, determining protein's interacting partners can shed more light on the pathway within which these proteins act. Thirdly, knocking out the gene allows comparing cell's behaviour with and without the gene present (Droit, Poirier and Hunter, 2005). This is complemented with in vitro assays which can test a diverse range of function. However, whether it is a single experiment like co-immunoprecipitation or a high-throughput experiment such as microarray analysis, wet-lab experiments are costly and time-consuming. Automated protein function prediction provides tools to build hypotheses about the functions that can be verified in the lab.

1.1.1 Definition of protein function

Protein function is a complex concept described through various aspects of proteins (Bork *et al.*, 1998; Rost *et al.*, 2003; Punta and Ofran, 2008). A function can be understood as a specific enzymatic activity, e.g. a kinase. It may also imply a pathway within which the protein and all the interacting partners perform their activity. It may represent the subcellular location, and it depends on the tissue where it is being expressed. Protein function can also be described through a malfunction caused by an alteration in its sequence and any diseases that this malfunction may induce. As summarised by Rost et al. in 2003: “Function is everything that happens to or through a protein”. The protein function definition alone is such a complex concept that it poses a challenge for prediction.

While the structure of a protein can be described by a set of coordinates for the atoms present in the protein, the concept of protein is less distinct, and function may be described in many ways. To ensure consistency in annotations and to enable computational methods to be able to utilise functional information, it is important that protein function is described using standardised and machine-readable vocabularies (Friedberg, 2006; Bernardes and Pedreira, 2013). It is also necessary that such vocabularies represent the relationships between different functions, e.g. DNA synthesis is a DNA biosynthetic process that is also a part of DNA replication. Additionally, such vocabularies need to describe functions with diverse information depth, e.g. protein binding and p53 binding (Ashburner *et al.*, 2000; Bernardes and Pedreira, 2013). Finally, protein functions need to be comparable to measure how similar they are (Punta and Ofran, 2008).

Many schemes have been developed for protein functional classification. Some of them were designed decades ago and remain popular today (Ouzounis *et al.*, 2003). They include hierarchical classification of enzymes by the Nomenclature Committee of the International Union of Biochemistry (Enzyme Commission hierarchical classification) (IUBMB, 1992; Tipton and Mcdonald, 2018), a database of genes and pathways named the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa, Sato and Kawashima, 2021), a database of molecular interactions contained within MINT and IntAct (Orchard *et al.*, 2014) or Gene Ontology terms that describe function in three categories: Molecular Function, Biological Process and Cellular Component (Carbon *et al.*, 2019).

1.1.1.1 Gene Ontology

The Gene Ontology (Ashburner *et al.*, 2000) is widely used to annotate protein function. Sequencing genomes of more and more organisms revealed that the core of biological processes is shared between them making it important that a single vocabulary was used to annotate the protein function for all species (Ashburner *et al.*, 2000). Functions such as DNA replication or transcription are conserved and performed by proteins in all prokaryotic and eukaryotic cells. The Gene Ontology was created to standardise the vocabulary describing those functions and assign them from proteins from well-studied species to their orthologues in newly sequenced organisms. Over the past twenty years, the Gene Ontology has been widely used for automated protein annotation (Carbon *et al.*, 2019).

The Gene Ontology consists of three ontologies, each representing a different aspect of protein function by answering a different question (Ashburner *et al.*, 2000). The Biological Process Ontology provides an answer to the question: "Why does the protein perform its function?". For example, Trypsin is a protease that participates in protein digestion (Kayode *et al.*, 2016), and AKT kinase is a serine/threonine kinase that phosphorylates many proteins in a plethora of pathways related to cell proliferation (Nicholson and Anderson, 2002). Meanwhile, the Molecular Function Ontology (MFO) is a response to: "What does a protein do?". It describes protein biochemical activity, e.g. a protease, kinase, or DNA binding. The Biological Process Ontology captures the higher-level processes that the molecular function of the protein is part of. Finally, the Cellular Component Ontology (CCO) answers the question: "Where in a cell does the protein performs its function?", for example, DNA polymerase acts in the nucleus (Nagasawa *et al.*, 2000), while the glycoprotein Fibronectin performs its activity in the extracellular matrix (Lee *et al.*, 2017).

Each of the ontologies is organised in a directed acyclic graph where the nodes represent terms defining the functions, and the relationships between them are denoted by the edges (Ashburner *et al.*, 2000). A Gene Ontology term (GO term) can have multiple parents and zero or more children terms. The Gene Ontology captures multiple relationships between different functions. 'Is a' and 'part of' are the most widely used. The 'Is a' relationship enables a definition of protein function from the general to specific. For example, the Gene Ontology term *DNA polymerase activity* forms the 'is a' relationship with the term *catalytic activity, acting on DNA*, which in turn forms the 'is a' relationship with the term *catalytic activity*. The 'part of' relationship captures relationships where a given function forms part of a larger process or complex. For example, *DNA polymerase activity* is a part of *DNA biosynthetic process* (Figure 1.3). Over time more relationship types have been added to the Gene Ontology, that now enable other relationships to be captured, e.g. regulation (three relationships 'regulates', 'positively regulates', 'negatively regulates').

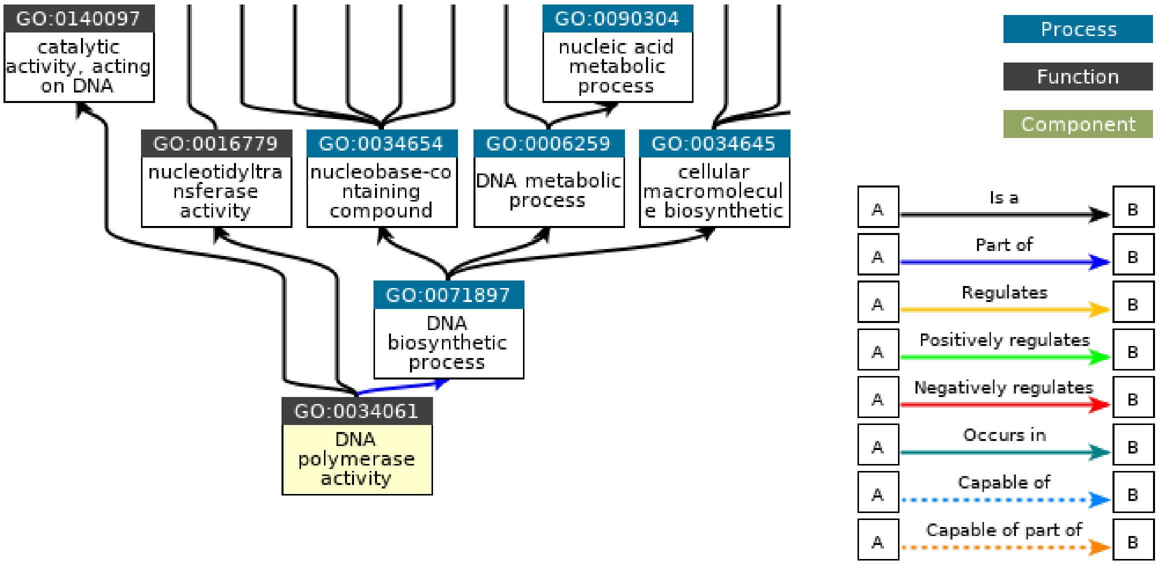


***Figure 1.3** Gene Ontology sub-graph demonstrating partial ancestry of the MFO term "DNA polymerase activity"* (Binns *et al.*, 2009; *QuickGO::Term GO:0034061*, 2020)

Each Gene Ontology annotation assigned to a gene/protein is associated with an evidence code demonstrating how the annotation was inferred (Carbon *et al.*, 2019). Evidence codes used for Gene Ontology are organised into six general classes: experimental, phylogenetic or computational evidence, author statements, curational statements and automatically generated annotations (*Guide to GO evidence codes*, 2020). Experimental evidence codes are used when the annotation is supported by an experiment such as, for example, a direct assay (evidence code: IDA) or high-throughput expression profiling (HEP), while phylogenetic evidence means that GO terms were inferred through determining evolutionary relationships between the genes.

In contrast, computational analysis evidence codes can be assigned when annotations are predicted through an *in silico* analysis of, for example, sequence similarity (ISS - Inferred from Sequence or structural Similarity) or genomic context data (IGC - Inferred from Genomic Context). However, annotations that were performed through homology-based transfer or other sequence information, but have not been manually reviewed are assigned the Inferred by Electronic Annotation (IEA) evidence code. Finally, if the GO term is assigned by either an author of the paper or by a curator, the evidence codes used in these cases belong to author statements class and the curational statements class respectively.

1.1.1.2 Enzyme Commission hierarchical classification

The Enzyme Commission (EC) hierarchical classification groups reactions catalysed by enzymes (IUBMB, 1992; Cornish-Bowden, 2014). Each EC number consists of four digits. The first digit represents the class of reaction the enzyme catalyses. The second digit represents the type of compound involved, while the third and fourth digits further specify details of the substrate (Tipton and Mcdonald, 2018). For example, Sterile alpha motif and histidine/aspartic acid domain-containing protein 1 SAMHD1 (investigated in Chapter 3 as a biomarker of sensitivity to anti-cancer drug cytarabine) is described by the following EC number in The BRENDA database of enzymes (Schomburg, Chang and Schomburg, 2002): EC 3.1.5.B1 which signifies a hydrolase (3) acting on ester bonds (3.1) further characterised as a triphosphoric-monoester hydrolase (3.1.5) and a dNTPase (3.1.5.B1). The current recommendations (2018) include seven main categories of enzymes based on the first component of the EC number (class of reaction) (Tipton and Mcdonald, 2018). These are oxidoreductases, transferases, hydrolases, lyases, isomerases, ligases and translocases.

The EC classification resembles part of the Molecular Function Ontology, which describes molecular activities performed by proteins, including catalytic activities (see above in 1.1.1.1). In fact, “catalytic activity” is a direct child of “molecular function” term among GO terms such as “binding”, “transporter activity”, “protein folding chaperone”, and many others (*QuickGO::Term GO:0003674*, 2021). Moreover, according to two publications from 2017, 70% of the catalytic activities described by EC numbers are also covered by Molecular Function Ontology (Furnham, 2017; Holliday *et al.*, 2017).

However, the EC classification, which contains only four levels of functional specification, is simpler than Gene Ontology which is represented by a directed acyclic graph where one term can have multiple parents and GO does not omit any steps when specifying a function. For example, there is one additional level of specification in GO describing “dGTPase activity” of SAMHD1 starting from “hydrolase activity” (corresponding to EC 3). “dGTPase activity” is a “triphosphoric monoester hydrolase activity” (EC 3.1.5), which in turn is a “phosphoric ester hydrolase activity” (not having a corresponding EC number), which is a “hydrolase activity, acting on ester bonds” (EC 3.1).

1.1.2 The Critical Assessment of Function Annotation

The Critical Assessment of Function Annotation (CAFA) was designed to evaluate state-of-the-art protein function prediction methods. The first edition, CAFA1, took place in 2011 (Radivojac *et al.*, 2013). Since then, the assessment has been repeated every three years resulting, with CAFA4 taking place in 2019/2020. The evaluation consists of several steps (Figure 1.4). First, proteins that do not have experimental annotations are released to the participants as targets. Secondly, the teams have a specific amount of time (usually four months) to predict the function of the proteins as GO terms together with confidence scores associated with each predicted term. Finally, after almost a year, the organisers collect the proteins that gained experimental functions throughout the period and use these to assess the predictions made by the participating teams.

In the first edition of CAFA, two ontologies were assessed: Molecular Function (MFO) and Biological Process (BPO). From CAFA2 (Jiang *et al.*, 2016) onwards, the performance in predicting Gene Ontology terms from the Cellular Component category (CCO) was also evaluated. The second edition also assessed how well the methods predicted terms from Human Phenotype Ontology (HPO) related to disease. However, this evaluation was not performed in CAFA3 (Zhou *et al.*, 2019) or CAFA4.

CAFA evaluates the methods on how well they can answer two very different questions. The first question asks: “given an amino acid sequence of a protein, what is the function of this protein?” and is protein-centric, while the second question asks: “given a function (in the form of Gene Ontology or Human Phenotype Ontology term), what are the proteins that perform this function?”. This second question is term-centric. The first problem requires predicting a subgraph of an acyclic directed graph that forms each ontology. The second task consists of deciding if a given term is associated with the protein, making it a binary classification problem.

To assess how well the methods answer these two questions, two types of metrics are calculated for each method. Protein-centric evaluation is conducted using precision (pr) and recall (rc) calculated for a given target protein and a confidence score threshold. To obtain only one measure for a method and thus to be able to compare the methods, precision and recall needed to be, first, averaged over all the assessed proteins to have one measure of each, precision and recall, per threshold. The next step is to calculate the harmonic mean of precision and recall with a goal of one metric per threshold. Finally, the maximum of all harmonic means of precision and recall over all the thresholds is taken to obtain a single score, F-max, for each method:

$$F_{max} = \max_{\tau}\left\{\frac{2 \cdot pr(\tau) \cdot \mathrm{rc}(\tau)}{pr(\tau) + \mathrm{rc}(\tau)}\right\},$$

where:

$$pr(\tau)=\frac{1}{m(\tau)}\sum_{i=1}^{m(\tau)}\frac{\sum_f \mathbb{1}\,(f \in P_i(\tau) \wedge f \in T_i)}{\sum_f \mathbb{1}\left(f \in P_i(\tau)\right)},$$

$rc(\tau)=\frac{1}{n}\sum_{i=1}^{n}\frac{\sum_f \mathbb{1}\,(f \in P_i(\tau) \wedge f \in T_i)}{\sum_f \mathbb{1}\,(f \in T_i)}$, $P_i(\tau)$ - the set of terms predicted for the protein sequence i with a score greater than or equal to τ,

T_i - the set of experimentally validated terms for the protein sequence,

f - function (Gene Ontology term),

$m(\tau)$ - the number of sequences for which terms were predicted with at least one score greater than or equal to τ,

$\mathbb{1}\,(\cdot)$ - an indicator function,

n - the number of target sequences used for the evaluation.

The second edition of CAFA introduced a further metric for assessing the performance of the methods. The new metric, S-min, is complementary to F-max, which focuses on terms predicted correctly. S-min is a minimum semantic distance between misinformation (mi) and remaining uncertainty (ru) measures over different confidence scores thresholds:

$$ru(\tau)=\frac{1}{n}\sum_{i=1}^{n}\sum_{f} ic(f) \cdot \mathbb{1}\,(f \notin P_i(\tau) \wedge f \in T_i),$$

$$mi(\tau)=\frac{1}{n}\sum_{i=1}^{n}\sum_{f} ic(f) \cdot \mathbb{1}\,(f \in P_i(\tau) \wedge f \notin T_i),$$

$$S_{min} = \min_{\tau} \left\{ \sqrt{ru(\tau)^2 + mi(\tau)^2} \right\},$$

where:

$P_i(\tau)$ - the set of terms predicted for the protein sequence i with a score greater than or equal to τ,

T_i - the set of experimentally validated terms for the protein sequence,

f - function (Gene Ontology term),

$\mathbb{1}\,(\cdot)$ - an indicator function,

n - the number of target sequences used for the evaluation,

$ic(f)$ - the information content of the Gene Ontology term.

Remaining uncertainty represents terms that were not predicted but should have been, and misinformation focuses on false-positive terms. Additionally, these two metrics included a measure of information content, which means that the deeper the terms that were incorrectly predicted or omitted, the higher misinformation and uncertainty, and consequently, S-min. Ultimately, an exemplary method should have a high F-max score and a low S-min score.

Term-centric assessment begins by calculating sensitivity and specificity for each given term and threshold. A Receiver Operating Characteristic (ROC) curve is created for each term by plotting all (1 – specificity, sensitivity) points for all the confidence scores thresholds. To obtain one measure per term, the area under the ROC curve (AUC) is computed. Averaging AUC values over all the terms for each method enables comparison of different methods. Additionally, averaging AUC values over all the methods for each term allows assessing how well the terms could be predicted by all the assessment participants.

During the first edition of CAFA (Radivojac *et al.*, 2013), 30 teams submitted predictions made using 59 methods. The number of algorithms increased to 129 in CAFA2 (Jiang *et al.*, 2016) and 144 in CAFA3 (Zhou *et al.*, 2019). Over time, participation in CAFA has grown – the preliminary results of CAFA4 presented at the Intelligent Systems in Molecular Biology conference in July 2020 reported 148 participating methods. However, not only the number of submitted methods increased from edition to edition but also their performance. The top methods in CAFA2 (Jiang *et al.*, 2016) outperformed the top methods in CAFA1 (Radivojac *et al.*, 2013), and the top methods in CAFA3 (Zhou *et al.*, 2019) outperformed the ones in CAFA2 (Jiang *et al.*, 2016). However, the improvement was not as significant as from the first to the second edition. This increase in the number of methods and the performance does not infer the increase in their diversity. The methods submitted during the first edition were far more diverse than in the second and third editions. This might be a consequence of teams designing their methods based on the top-performing methods in CAFA1 (Radivojac *et al.*, 2013). Higher performance of methods from edition to edition may be caused by better training data due to updated protein databases and improvement of the methods themselves.

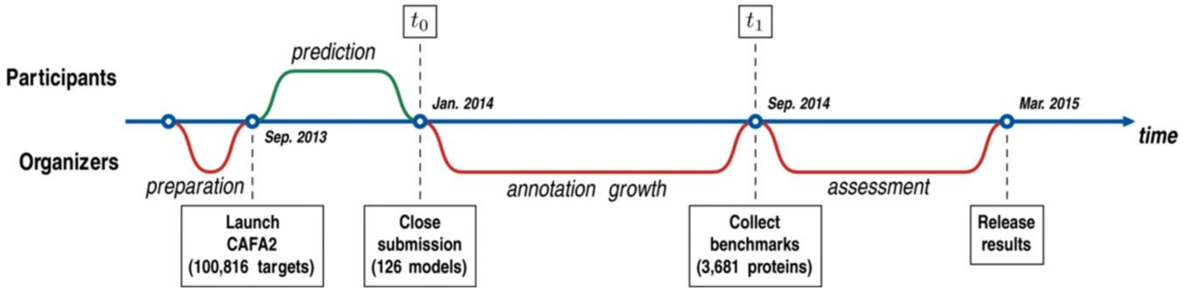


Figure 1.4 *Timeline of the second edition of CAFA. Figure extracted from* (Jiang *et al.*, 2016).

The Critical Assessment of Functional Annotation provides scientists with a review of the current state of the art of protein function methods and partially drives the development and improvement of tools. It does, however, possess some limitations.

Certain limitations were described in the paper published after the first edition of CAFA (Radivojac *et al.*, 2013). As mentioned above, in CAFA1, protein function methods in the protein-centric evaluation were assessed only using one metric - F_{max}. This measure is based on precision and recall and indicates how well a method performs. However, it does not take into consideration the specificity of the predicted terms. For example, predicting the Gene Ontology term ‘regulation of cellular process’ is rewarded the same as ‘negative regulation of cell growth’ even though the latter is more specific. The organisers addressed this issue by creating a new measure S_{min} that considers the information content of the predicted function. However, it only assesses the methods in terms of missing information and remaining uncertainty. It would be beneficial to incorporate weights based on how deep the predicted function is placed in the directed-acyclic graph representing each ontology into precision and recall. The other limitation shared by organisers after the first CAFA edition was penalising methods that annotated fewer proteins but with more accurate functions over methods that predicted less accurate functions but for all the proteins. This issue was addressed in the second edition by introducing a partial evaluation mode where the method could be assessed only on a subset of target proteins (Jiang *et al.*, 2016).

Another limitation of CAFA concerns bias towards identifying experimental functions in a small number of widely studied model organisms. Makrodimitris et al. state that ten to fifteen organisms usually accumulate experimental annotations during the CAFA challenge. Examples include *H. sapiens*, *M. musculus*, *A. thaliana*, *E. coli* or *R. norvegicus* (Makrodimitris, Van Ham and Reinders, 2020). This implies that methods are usually evaluated only on predictions for those model organisms and there is a need for initiatives that will encourage experimentalists to study more niche organisms for which we have very limited knowledge. In addition, groups that participate in the CAFA challenges are required to predict functions for many more proteins than the number that will be evaluated (due to constraints of experimental validation). For example, the number of released targets during CAFA1 and CAFA2 was 48 298 and 100 816. However, the number of proteins that gained experimental annotations and were used to evaluate protein function prediction methods at the end of the challenge were 866 and 3 681 for CAFA1 and CAFA2, respectively, representing 2-4% of the protein targets that groups were required to annotate. The CAFA challenge in the current form requires a lot of effort and resources as functional annotation of around 100 000 of proteins need considerable computational resources. It would be helpful if the target proteins could be limited to those more likely to gain experimental functions at the end of the challenge.

In addition, the fact that experiments have not validated a certain function does not necessarily mean that the protein does not possess this function. Instead, it may simply indicate that it was not a target of any experiments collecting benchmark functions. This way, a method that predicted a GO term that may be a true function of the protein but has not been experimentally validated yet will be penalised (Vesztrocy and Dessimoz, 2021). The same happens if the predicted functional term is more specific than the validated term (Radivojac *et al.*, 2013). As a solution, Vesztrocy & Dessimoz propose that methods should be evaluated on negative annotations, i.e. the knowledge is a protein does not perform or is not involved in certain roles (Vesztrocy and Dessimoz, 2021).

Finally, many of the submitted methods are not publicly available or are very difficult to set up. This contradicts the purpose of CAFA as the point of evaluating methods and presenting the state of the art of automated protein function prediction loses its value if the tools are not accessible. Thus, sharing easy to deploy code should be encouraged by the organisers.

1.1.3 Principles behind inferring protein function

Since the Basic Local Alignment Search Tool (BLAST) was first published in 1990, computational biologists have been designing new and improving already well-established methods for protein function prediction. Various approaches have been explored since then. They apply multiple hypotheses about proteins and their mechanisms of action, and they utilise the advancement of machine learning.

Protein function can be predicted either “de novo” or through the homology-based inference that transfers function to the query protein from experimentally annotated homologues. Methods can also be based solely on the sequence or structure of the protein, or they can go beyond that and explore genomic context data, protein-protein interactions or co-expression data. They can also use various machine learning algorithms, which often combine multiple different types of data to infer protein function (Aerts *et al.*, 2006; Sokolov and Ben-Hur, 2010; Kourmpetis *et al.*, 2011; Wass, Barton and Sternberg, 2012; Cozzetto *et al.*, 2016).

After every edition of the Critical Assessment of Functional Annotation (CAFA), which aims to evaluate how well the protein function prediction methods can infer Gene Ontology terms for a set of target proteins (see 1.1.1 for details), the organisers of the assessment publish a paper summarising the results. Publications from CAFA1 (Radivojac *et al.*, 2013) and CAFA2 (Jiang *et al.*, 2016) include the list of all the submitted methods and keywords, such as gene expression, protein structure, machine learning or literature, that describe best their algorithm and protein properties that are used to predict the function (summarised in Table 1.1). The paper from CAFA3 (Zhou *et al.*, 2019) does not contain the list of methods with associated keywords. It does, however, provide the insights into the frequency with which they were used (see Figure 1.5, Figure 1.6, Figure 1.7). These approaches are explored in the next section of the introduction.

Code	Keyword	Number of methods from CAFA1	Percentage of methods from CAFA1 (%)	Number of methods from CAFA2	Percentage of methods from CAFA2 (%)
cd	clinical data	0	0	0	0
cm	comparative model	3	6	5	4
dp	de novo prediction (CAFA2), derived/predicted (CAFA1)	8	15	8	6
gc	genomic context	4	7	7	6
gd	genetic data	0	0	2	2
ge	gene expression	9	17	20	16
gi	genetic interactions	1	2	7	6
gne	genome environment	0	0	0	0
hmm	hidden Markov model	0	0	24	19
ho	homolog	0	0	31	25
lt	literature	11	20	11	9
ml	machine learning	24	44	51	40
ms	mass spectrometry	0	0	0	0
nlp	natural language processing	0	0	4	3
ofi	other functional information	5	9	16	13
op	operon	0	0	0	0
or	ortholog	12	22	25	20
pa	paralog	2	4	19	15
ph	phylogeny	2	4	16	13
php	physicochemical properties	0	0	8	6
pi	protein interactions	9	17	33	26
pp	predicted properties	0	0	20	16
ppa	profile-profile alignment	13	24	25	20
pps	predicted protein structure	7	13	13	10
ps	protein structure	6	11	4	3
sa	sequence alignment	25	46	71	56
sp	sequence properties	10	19	27	21
spa	sequence-profile alignment	17	31	40	32
sta	structure alignment	0	0	2	2
sy	synteny	0	0	0	0

Table 1.1 *Number and percentage of methods associated with specific keywords. Data extracted from supplementary files of* (Radivojac *et al.*, 2013) *and* (Jiang *et al.*, 2016).

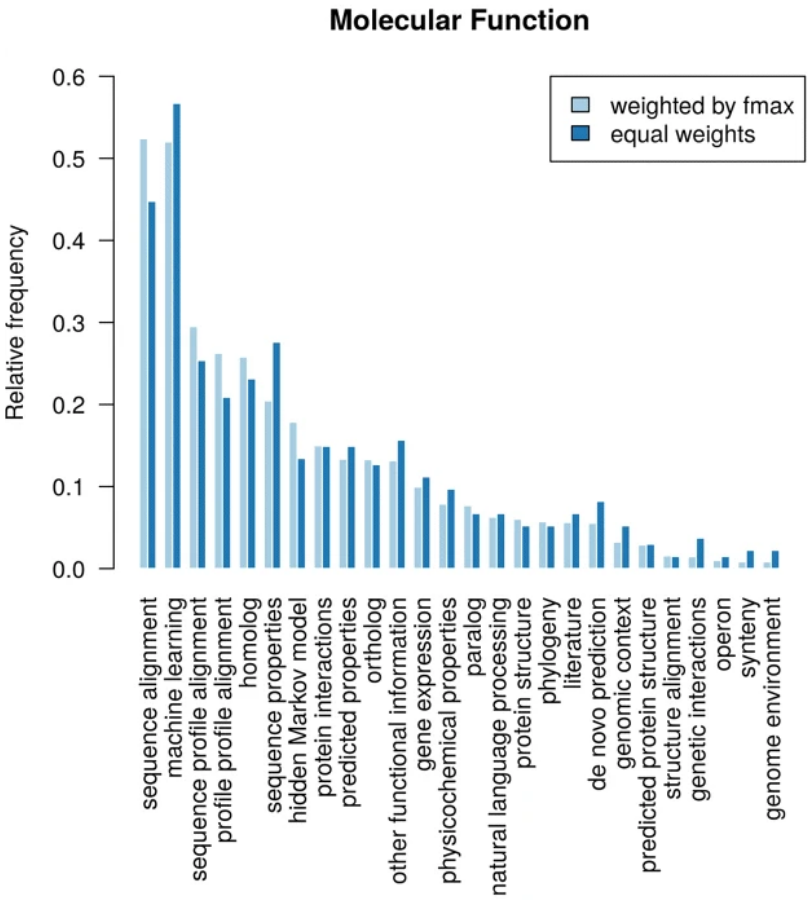


Figure 1.5 *Frequencies of the keywords across Molecular Function predictors participating in CAFA3. Figure extracted from* (Zhou *et al.*, 2019).

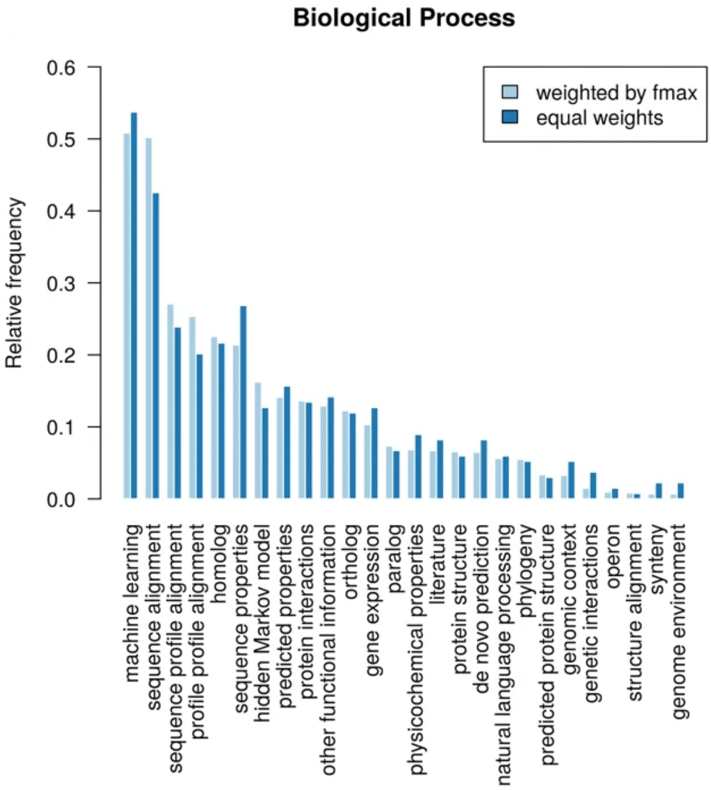


***Figure 1.6** Frequencies of the keywords across Biological Process predictors participating in CAFA3. Figure extracted from* (Zhou *et al.*, 2019).

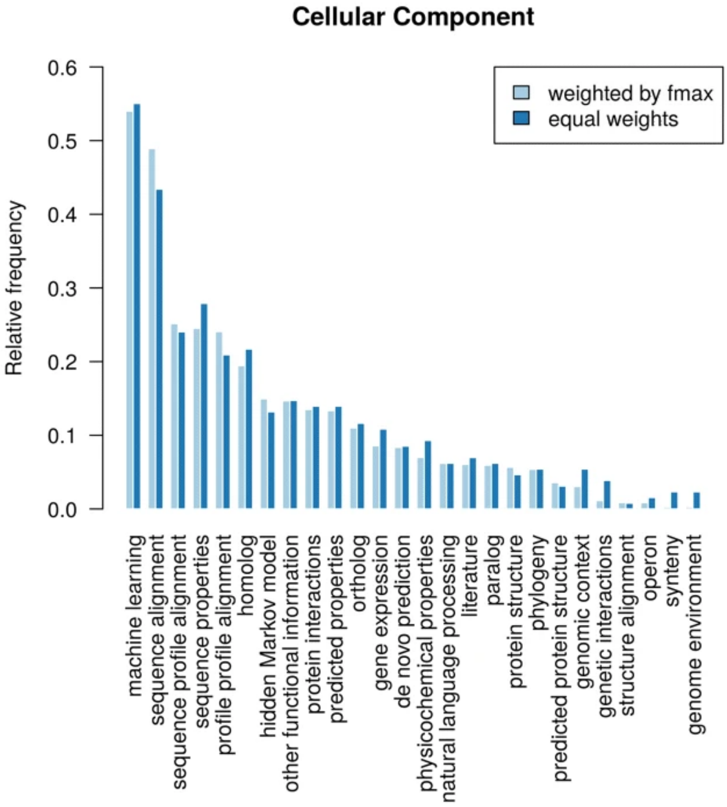


***Figure 1.7** Frequencies of the keywords across Cellular Component predictors participating in CAFA3. Figure extracted from* (Zhou *et al.*, 2019).

1.1.3.1 Homology-based function transfer

In the first two editions of CAFA, sequence alignment was the most popular technique used to predict protein function (46% of the methods in CAFA1 (Radivojac *et al.*, 2013) and 56% of the methods in CAFA2 (Jiang *et al.*, 2016) were associated with the keyword “sequence alignment”) (see Table 1.1). Machine learning was the second most popular technique, with 44% methods using it in CAFA1 and 40% in CAFA2. However, in CAFA3 (Zhou *et al.*, 2019) it was “machine learning” that was most frequently used (50-60% of the methods in all three categories, MFO, BPO, CCO) with “sequence alignment” right behind it (40-50%) (Figure 1.5, Figure 1.6, Figure 1.7).

Methods that explore sequence and profile alignments include, for example, methods that transfer functions to uncharacterised from characterised proteins with known annotations if the two sequences are highly similar (Shehu, Barbará and Molloy, 2016).

The simplest form of annotation transfer is to simply assign the function of the top BLAST hit to the query sequence. The idea behind it is that two proteins with similar sequences are likely to evolve from a common ancestor and share the same function. It is part of the commonly applied sequence-structure-function paradigm: the amino acid sequence determines protein structure, and the structure determines the function (Serçinoğlu and Ozbek, 2020).

While this concept may be effective when two proteins have very high sequence identity (e.g. >80% identical), it should be applied with caution when a protein to be annotated is divergent from the proteins with known functions (Pearson, 2013), particularly as changes to a small number of functional residues (e.g. residues in an enzyme active site) can be sufficient to alter protein function. Proteins sharing 60% or less sequence identity are considered to be "difficult" in the world of automated function prediction using sequence homology (Bernardes and Pedreira, 2013; You *et al.*, 2018). However, it has been shown that in the case of protein structure, 20% sequence similarity is enough for the structural characteristics such as the secondary structure or side-chain conformations to be preserved (Flores *et al.*, 1993), the implication being that proteins with low sequence identity can share similar structure and function (more on the structure-function relationship in 'Principles behind inferring protein function'/'Protein structure').

In addition, various studies conducted within a decade after BLAST was first published in 1990 reported that the correctness of homology-based transfer depends not only on the percentage of sequence identity itself but also on other factors (Friedberg, 2006; Punta and Ofran, 2008). These are exemplified by the aspect of function to be predicted, such as a binding substrate or a type of Gene Ontology, or by the ability to perform a catalytic activity.

It has been demonstrated that proteins with highly similar sequences may not share the same structure or play the same role in a cell (Punta and Ofran, 2008; Clark and Radivojac, 2011). Firstly, despite the same amino acid sequence, the usage of codons may slow down or speed up folding and impact the final structure (Zhou *et al.*, 2013). Secondly, two proteins with identical sequences may have a different structure due to being exposed to different solvents or ligands (Gan *et al.*, 2002). In addition, sometimes the function is determined by a small number of functional residues with specific physicochemical properties, like in the case of ATP or DNA binding residues or an enzyme active site (Punta and Ofran, 2008). Therefore, the overall sequence or structural similarity may not matter if those residues are not conserved. Finally, a protein may be simply ‘recruited’ by several mechanisms, such as point mutations, to perform a new function in the cell that will be more advantageous for the organism (Schulenburg and Miller, 2014).

Homology-based function transfer can result in functions being assigned incorrectly. If consequently used to annotate newly sequenced genes, it will propagate erroneous annotations throughout the database (Friedberg, 2006). In fact, in 2018, Mahlich et al. published a study in which they analysed the correctness of annotations in the SwissProt database (Mahlich *et al.*, 2018; Bateman, 2019) in which they revealed protein annotations might have been assigned incorrectly in a sixth of the proteins considered.

Despite all the drawbacks, homology-based inference using BLAST is still a widely used and competitive method. Methods have developed annotation transfer by considering the annotations of multiple hits from a BLAST search. For example, the method Protein Function Prediction (PFP), combines the annotations of proteins from a PSI-BLAST (Position-Specific Iterative) search with the structure of the Gene Ontology (Hawkins *et al.*, 2009). Additionally, the results from the 3rd CAFA edition revealed that the F-max metric, used to evaluate methods in CAFA (see1.1.2), for BLAST calculated for Molecular Function of proteins without any previous annotations was equal to 0.42 while F-max for the top 10 methods ranged from 0.51-0.62 (Zhou *et al.*, 2019). For Biological Process and Cellular Component, the difference between F-max of the top method and BLAST is even smaller. While F-max for Biological Process is 0.40 and 0.26 for the best method and BLAST respectively, it is equal to 0.61 and 0.46 in Cellular Component.

1.1.3.2 Sequence motifs

It is also possible to use a smaller part of the protein, called a pattern, signature or motif, which may be more likely to determine a function and thus be conserved during evolution (Bernardes and Pedreira, 2013; Shehu, Barbará and Molloy, 2016). These signatures can be a part of a domain, and they can form a functional region such as a catalytic or binding site (Bernardes and Pedreira, 2013). They can be recognised using techniques such as hidden Markov models (HMMs) or regular expressions (Friedberg, 2006).

Many methods have been developed to identify protein domains (or families), which can be thought of as evolutionary units within a protein (Kelley and Sternberg, 2015). InterPro provides access to the most widely used databases of domain and protein families and sequence motifs within one server. In 2019, it combined results from fourteen resources which allows for better coverage and more extensive exploration of signatures identified in the query sequence (Mitchell *et al.*, 2019). While protein domains are not determined by protein function, they can be used to infer function. The InterPro database is one of the methods used to identify unknown functions in the minimal bacterial genome described in Chapter 2. Sequence patterns, and protein and domain families were extracted using eleven resources out of the fourteen currently forming InterPro.

1.1.3.3 Machine learning

There are two major groups of machine learning methods – supervised and unsupervised. In supervised learning methods, also called classification algorithms, a model is built using a set of inputs (features) and known outcomes (labels) (Bernardes and Pedreira, 2013). In protein function prediction, while the functional annotations represent the labels, e.g. GO terms, the features may be created from various protein (and gene) properties. The model explains and learns correlations between the inputs derived from properties of the proteins and their known functional annotations. The resulting model can then be applied to query proteins to infer their function.

The protein properties or features used in machine learning approaches can be diverse. They can use homology-based transfer, and for example, features may then include properties such as sequence identity or e-values for matches to proteins of known function (Friedberg, 2006). Alternatively, if the function is predicted “de novo”, the inputs may consist of physicochemical properties of amino acids, such as the number of hydrophobic amino acids (Rentzsch and Orengo, 2009). They also may involve predictions of secondary structure, including the number of transmembrane helices, or prediction of glycosylation sites or post-translation modifications (Punta and Ofran, 2008). The possibility of using information about the protein that does not require identifying its homologues in other species makes algorithms of supervised machine learning very popular for the task of “de novo” function prediction (Rentzsch and Orengo, 2009).

Examples of specific algorithms of supervised machine learning include Support Vector Machines (SVMs), neural networks or logistic regression (Bernardes and Pedreira, 2013). FFPred is a “de novo” protein function prediction method that uses SVMs to build models based on inputs designed from biophysical properties of the proteins and GO terms as outputs (see 1.1.4.4 for details). It demonstrated to be indispensable for the inference of functional annotations of the proteins in the minimal bacterial genome that do not have many homologues in other species (see Chapter 2). Additionally, the Jones group methods submitted to CAFA incorporate FFPred and have frequently been among the top performing methods.

Unsupervised machine learning differs from supervised learning as it lacks provided outputs, functional annotations in the case of protein function prediction, that could help identify the correlations and build a model (Bernardes and Pedreira, 2013). Instead, it tries to identify patterns within the input data and clusters it. Unsupervised machine learning is helpful when the labels (classes) are yet to be discovered. Hierarchical clustering is a widely used clustering algorithm which was used in the work presented in Chapter 4 to identify clusters among sub-lines of the UKF-NB-3 cell line that are adapted to tubulin-binding agents (TBAs).

1.1.3.4 Orthologues

Around 20% of the submitted methods in CAFA1 (22%) and CAFA2 (20%) used orthologues to predict protein function according to the keywords assigned to methods from both CAFA papers (see Table 1.1). Fewer methods were associated with keywords such as “paralogue” or “phylogeny” in general in the first two editions of CAFA. In CAFA3, 10-20% of the methods predicting all three categories: MFO, BPO and CCO, were associated with the “orthologue” keyword, and less than 10% with each “paralogue” and “phylogeny” (Figure 1.5, Figure 1.6, Figure 1.7). Protein function tends to be more conserved within orthologues rather than paralogues (Punta and Ofran, 2008). However, this has recently been a matter of debate with some research (Nehrt *et al.*, 2011; Stamboulian *et al.*, 2020) proposing that the opposite was the case and further studies supporting the orthologue conjecture and/or suggesting that the Gene Ontology was not suitable to test this (Thomas *et al.*, 2012). Given this general concept, some methods use this to improve homology-based transfer. This can be done by building a phylogenetic tree from the homologues of the query protein and assessing the specific type of the relationships they share (Friedberg, 2006; Rentzsch and Orengo, 2009). As a result, the function is inferred from the closest orthologue and not homologue.

1.1.3.5 Protein structure

Only 13% of methods in CAFA1 (Radivojac *et al.*, 2013) were associated with the keyword “predicted protein structure” and 11% with the keyword “protein structure”. In CAFA2 (Jiang *et al.*, 2016) these numbers were 10% for “predicted protein structure” and 3% for “protein structure” respectively. That means that five times fewer methods used information derived from the structure than from the sequence even though the former provides more insights into the function of the protein (Friedberg, 2006). This is because knowing the structure permits identifying the biochemical mechanism in which the function is performed.

Molecular functions are associated with specific structural folds, and hence the homology on the structure level is more preserved than the homology on the sequence level (Hou *et al.*, 2005). This results in the possibility of identifying distant homologous relationships using structures, even though no sequence homologues were found (Punta and Ofran, 2008).

However, just as in the case of sequence motifs that can be used to infer function if the global sequence similarity to other proteins is not detected, structural motifs can also be applied. They consist of smaller spatial regions that can be associated with a specific function (Friedberg, 2006). The rationale is that functional residues, responsible for binding DNA, RNA or ligands, or performing catalytic activities, tend to cluster together in the 3D structure (Punta and Ofran, 2008).

1.1.3.6 Genomic context

The function of the protein can also be inferred using gene neighbourhood analysis, gene fusion analysis and phylogenetic profiling. These three approaches fall under the umbrella of genomic context methods. Gene neighbourhood methods explore the hypothesis that genes sharing a function may be located close to each other on the chromosome, and this location would be conserved in multiple species (Bernardes and Pedreira, 2013). Gene fusion methods consider that two genes in one species are likely to share function if they are fused as a single gene in another species. Finally, a phylogenetic profile of a gene is a vector that indicates if a gene has a homologue or not in a given genome (Friedberg, 2006). Genes that share phylogenetic profiles are expected to evolve together and carry similar roles in the cell. All three of these methods are more likely to be useful for inferring function relating to biological processes. Analysing keywords associated with methods participating in the three CAFA editions, revealed that less than 10% of the methods from each CAFA applied genomic context in their solution to the function prediction task (see Table 1.1, Figure 1.5, Figure 1.6, Figure 1.7).

1.1.3.7 Protein interactions

The rationale behind using protein-protein interactions data to predict protein function is that proteins do not perform their roles in solitude but instead in collaboration with other molecules (Bernardes and Pedreira, 2013). Hence the principle of "guilt by association" can be applied to annotate the proteins that interact with each other (Loewenstein *et al.*, 2009). The interaction data can be represented as a network where nodes are formed by proteins and the edges – the relationships between them (Rentzsch and Orengo, 2009). The hypothesis that is explored for the network-based function inference is that the closer the two proteins are in the network, the more similar role they share. 17% of the methods evaluated in CAFA1 (Radivojac *et al.*, 2013) used protein interactions, increasing to 26% in CAFA2 (Jiang *et al.*, 2016), but down to 10-20% in CAFA3 (see Table 1.1, Figure 1.5, Figure 1.6, Figure 1.7).

1.1.3.8 Gene expression

Similarly to protein interactions data, in the case of gene expression data, the "guilt by association" principle could also be applied when annotating a gene with unknown function that is co-expressed with a gene of known function (Friedberg, 2006).

1.1.3.9 Aggregate methods

As protein function is a complex concept that covers many different aspects, it has motivated scientists to develop methods for predicting function that incorporate multiple sources of data. In fact, most of the top-performing methods in all the editions of CAFA were based on machine learning algorithms and the combination of various features, e.g. protein structure, protein-protein interactions, expression data, or evolutionary relationships. The method from CAFA3 (Zhou *et al.*, 2019) that outperformed significantly all the other methods and all top methods from CAFA2 (Jiang *et al.*, 2016) and CAFA1 (Radivojac *et al.*, 2013), GOLabeler (You *et al.*, 2018), used information from the frequency of GO terms, sequence alignment, protein domains, motifs and biophysical properties, and amino acid trigrams.

1.1.3.10 Ligands

Proteins do not perform their function alone (Zhao, Cao and Zhang, 2020). Instead, they often interact with small molecules called ligands. Some of these interactions are very specific and vital for the function to be performed correctly (Gallo Cassarino, Bordoli and Schwede, 2014). They may serve either as a substrate (e.g. kinase and ADP), signalling molecules (e.g. estrogen receptors activated by estradiol) or cofactors (e.g. dehydrogenase and NAD). Therefore, identifying ligands that interact with a protein is a vital step in establishing the role of the protein in the cell.

A plethora of methods predicting ligands and ligand-binding sites have been developed over the years. For example, in Chapter 2, we applied Firestar (Lopez *et al.*, 2011) and 3DLigandSite (Wass, Kelley and Sternberg, 2010) to identify the ligands bound by the proteins of the minimal genome. Some other examples can be found in the following publications: (Roche, Brackenridge and McGuffin, 2015; Xie and Hwang, 2015; Cui *et al.*, 2019; Zhao, Cao and Zhang, 2020). In addition to the predicted protein-ligand interactions, some tools offer functional annotations such as Gene Ontology terms. For example, to identify ligands and ligand-binding sites, FunFold3 superimposes templates with biologically relevant ligands established by the BioLip database (Yang, Roy and Zhang, 2013) onto the structural model of the target protein (Roche and McGuffin, 2016). BioLip is a database created from the information on protein-ligand interactions contained in the PDB (Berman *et al.*, 2000). It also stores functional annotations such as GO terms associated with PDB structures which are then transferred to FunFold3 results. Another ligand-binding site predicting tool incorporating such functional information is Firestar (Lopez *et al.*, 2011). However, in contrast to FunFold3, Firestar uses FireDB (Maietta *et al.*, 2014) instead of BioLip.

Associating GO terms with protein-ligand interactions is possible due to GO terms assigned to PDB structures within projects such as SIFTS (Dana *et al.*, 2019) and UniProt-

GOA (Huntley *et al.*, 2015). Other ligand predicting methods, such as 3DLigandSite (Wass, Kelley and Sternberg, 2010), could also benefit from extracting GO terms associated with known PDB structures used to identify ligands and provide such functional information in addition to the clusters of residues and the ligands they bind.

1.1.4 Description of selected Gene Ontology terms predictors

The following four tools were used to predict Gene Ontology terms aiming to identify unknown functions in the minimal bacterial genome described in Chapter 2. They were selected based on their good performance in CAFA2 (Jiang *et al.*, 2016) and the ability to access their code/web-servers.

1.1.4.1 CombFunc

CombFunc integrates information from multiple sources to predict Molecular Function and Biological Process GO terms for proteins (Wass, Barton and Sternberg, 2012). It incorporates homology-based annotation transfer, domain and structural information, conserved residues, protein-protein interactions, and gene expression data. Features from these different data sources are used in a Support Vector Machine (SVM) model. For example, for homology-based transfer, e-value with which the top homologue annotated with the function is identified, the sequence identity between that homologue and the query protein, and a percentage expressing the coverage of the query by the top homologue are input into the model. On the other hand, features representing gene expression data include a fraction of proteins that are co-expressed and annotated with the function or the correlation coefficients for the co-expressed proteins. Features related to information about domains include the lowest e-value of a domain annotated with the function. CombFunc uses BLAST and PSI-BLAST (Altschul *et al.*, 1997) to collect data from the top homologues, ConFunc (Wass and Sternberg, 2008) for analysis of conserved residues, InterPro (Mitchell *et al.*, 2019) and Pfam (El-Gebali *et al.*, 2019) for the domain, and Phyre2 (Kelley *et al.*, 2015) for protein structural information. It extracts protein-protein interactions data from IntAct (Orchard *et al.*, 2014) and MINT (Licata *et al.*, 2012), and gene expression data from COXPRESdb database (Obayashi *et al.*, 2019).

1.1.4.2 Argot2

Argot2 predicts GO terms using refined homology-based inference (Falda *et al.*, 2012). The refinement step is done to select the most accurate GO terms for the protein sequence, and clusters GO terms based on their semantic similarities and a weighting scheme. Firstly, homologues of the query proteins are obtained through scanning databases with annotated sequences which results in a list of hits with scores representing how evolutionary close they are to the query. Specifically, Argot2 is based on hits from running BLAST (Altschul *et al.*, 1997) against UniProtKB (Bateman, 2019) and HMMER (Finn, Clements and Eddy, 2011) against Pfam (El-Gebali *et al.*, 2019). GO terms annotated to homologues of the target protein are taken to downstream analysis together with the corresponding confidence scores. The first step is to reconstruct various paths to the root node that these GO terms create and discard those not belonging to the reconstructed paths. The nodes are then weighted using e-values from BLAST and HMMER, and the most probable paths are selected. The remaining GO terms in these paths are clustered according to their semantic similarity, which reduces the number of similar GO terms by choosing those with the best scores, weights and the highest information content.

1.1.4.3 LocTree

LocTree3 predicts protein subcellular localisation (Goldberg *et al.*, 2014). When possible, it uses homology-based transfer, and alternatively, it applies a machine learning model. Inference from close homologues is performed using PSI-BLAST (Altschul *et al.*, 1997). First, a profile is created for a sequence by executing two iterations of PSI-BLAST against the combination of UniProt (Bateman, 2019) and the Protein Databank (PDB) (Berman *et al.*, 2000) databases. This step is followed by the profile being scanned against proteins from SwissProt (Bateman, 2019) that have a single experimental annotation of subcellular localisation (to provide non-ambiguity). If the hits are identified with an e-value less than 0.001, they are used to infer the subcellular localisation. However, where this is not the case, the protein sequence is processed through a decision tree where the decision in each step is made based on the result from a separate SVM model. This mimics the mechanism of protein targeting.

1.1.4.4 FFPred

FFPred was designed as an alternative for homology-based inference of protein function (Lobley *et al.*, 2008), and it has been successfully applied to annotate proteins with distant or no homologues in other species. FFPred3 can predict 868 GO terms from all three Gene Ontologies and includes an SVM model per GO term (Cozzetto *et al.*, 2016). Features used for the SVMs are based on various biophysical properties of protein such as secondary structure, transmembrane helices, intrinsically disordered regions, signal peptides, subcellular localisation, amino acid composition, low complexity regions, coiled coils or post-translational modification patterns. FFPred3 has been trained on annotated human proteins from UniProt-GOA (Huntley *et al.*, 2015) and UniProtKB (Bateman, 2019). However, its performance was also measured on various eukaryotic species.

1.2 Cancer genomics

1.2.1 Types of mutations

Mutations are defined as alterations in the genetic sequence. They may be caused by various environmental factors (e.g. UV light), chemicals such as free radicals (e.g. benzo[a]pyrene contained in the cigarette smoke) or errors made during DNA replication (Clancy, 2008).

Mutations can be germline or somatic. Germline mutations occur in gametes and can be passed onto the next generation. Meanwhile, somatic mutations will ever only affect the organism in which they occur (Griffiths *et al.*, 2000).

There are multiple types of mutation. The most common is a substitution of a single base (single nucleotide variants; SNVs). There are also insertions and deletions where less than 1000 bases may be inserted or deleted (INDELs). Larger insertions and deletions are referred to as structural variants (Feuk, Carson and Scherer, 2006; Clancy, 2008).

In Chapters 3 and 4, investigating potential genetic mechanisms of acquired drug resistance in cancer cell lines, we called and analysed somatic point mutations and short INDELs (insertions and deletions). Each mutation was classified according to its effect on the expressed protein using terms from Sequence Ontology (Eilbeck *et al.*, 2005) (described in more detail in 1.2.4). When SNVs occur within the protein-coding region of a gene they can have different consequences on the encoded proteins. They can cause a substitution of the amino acid, introduction of a premature STOP codon (a nonsense or stop-gain mutation), or they can be synonymous – not changing the amino acid. Where the length of indel is not divisible by three (codon length), it results in a frameshift mutation – resulting in a shifted reading frame, and many more (see Variant calling section of Chapter 3 or 4 for the full list).

Structural variants are not a focus of this thesis as currently, there are no methods to call such variants from whole-exome sequencing data with high confidence. Structural variants include copy-number alterations (CNAs), translocations and inversions (Feuk, Carson and Scherer, 2006). A copy-number alteration occurs when a DNA segment is present at a different number of copies than in the reference genome due to its duplication, deletion or insertion. A translocation is characterised by a change of position of a DNA segment without changing its sequence. Finally, an inversion happens when a DNA segment is reversed (in direction with regard to the rest of the chromosome) and reinserted.

1.2.2 Nature of cancer

Cancer primarily constitutes a genetic disease (Senft *et al.*, 2017). It is caused by mutations in DNA (see above in 1.2.1) that bypassed the cell's natural DNA repair mechanisms such as base excision repair, nucleotide excision repair, mismatch repair, homologous recombination and non-homologous end-joining (Chatterjee and Walker, 2017). However, not all mutations lead to carcinogenesis. Those that do are called "drivers". They occur in specific signalling pathways called "cancer hallmarks" (preventing apoptosis and senescence, promoting cell division without any extracellular signals, initiating metastasis and angiogenesis, deregulating energy metabolism or circumventing immune response (Gonzalez-Perez, Mustonen, et al., 2013; Martínez-Jiménez, Muiños, Sentís, et al., 2020)), and in addition, they have a damaging impact on the proteins. The latter can be checked through investigating mechanisms in which the mutations alter protein functions. This may include verifying if the mutation is present in a domain that is responsible for performing a specific function, if it will affect how the sequence is folded into a structure, or if it will interrupt interactions with ligands and other proteins.

We have a plethora of anti-cancer drugs at our disposition, and new ones are invented all the time. However, cancer is not one disease, as each tumour contains a unique set of mutations. Precision oncology relies on selecting the best treatment (drug and dose) for a specific patient and focusses on understanding how mutations present in the tumour will affect drug sensitivity. We distinguish inter-tumour heterogeneity, which signifies that cancer's genotype and phenotype differ between patients, but this is not all there is to heterogeneity in cancer. Each cancer consists of a population of cells, and as they are constantly dividing, new mutations are acquired, leading to intra-tumour heterogeneity (Senft *et al.*, 2017). We addressed the problem of intra-tumour heterogeneity and how it influences a response to treatment in Chapter 4.

Precision oncology is important for use of targeted therapies, which are usually small molecules that inhibit the function of a specific protein (Senft *et al.*, 2017). This is in contrast to cytotoxic therapies, which widely cause cellular damage without having a specific protein target. Sorafenib is an example of a BRAF inhibitor used to treat advanced melanoma with the V600E BRAF mutation (Tanda *et al.*, 2020). Another example is nutlin-3 – a drug targeting MDM2-p53 interaction in cancer patients with wild-type p53 (Khurana and Shafer, 2019). However, this type of treatment is not free of the possibility of acquiring resistance – it usually develops within 6-12 months (Senft *et al.*, 2017). Resistance acquired to nutlin-3 in acute myeloid leukaemia cell line MOLM13 was investigated in Chapter 3.

1.2.3 Next-generation sequencing in the service of precision oncology

The development of next-generation sequencing (NGS) constitutes a milestone in cancer genomics. Due to the ability to sequence millions of short reads simultaneously, NGS has provided a cost and time-efficient method to sequence even whole genomes since the mid-2000s (Behjati and Tarpey, 2013; Goodwin, McPherson and McCombie, 2016). Rapid and relatively inexpensive high-throughput sequencing opened the door to the characterisation of exomes (whole-exome sequencing), genomes (whole-genome sequencing), transcriptomes (RNA-sequencing) and epigenomes (ChIP-seq) of cancer patients (Liang and Kim, 2013; Berger and Mardis, 2018). Together with advanced computational algorithms and tools that piece the short reads together by comparing them to the reference genome, and analyse detected variation, NGS led to an increasing understanding of the biological mechanisms causing cancer (Behjati and Tarpey, 2013). This, in turn, facilitated diagnosing tumours in the clinic and selecting the most suitable treatment (Zhao, Jones and Jones, 2019). In this thesis, whole-exome sequencing data of cancer cell lines are used to analyse mechanisms of acquired drug resistance and intra-tumour heterogeneity (Chapter 3 and Chapter 4, respectively).

The advancement of our knowledge of cancer as a disease driven by genetic alterations has been primarily achieved through large-scale genomic research projects which can sequence thousands of tumours. The rationale behind these programmes is that statistical analysis of mutations in a cohort of patients can reveal the same patterns in specific genes suggesting a similar positive selection process leading to cancer development (Martínez-Jiménez, Muiños, Sentís, *et al.*, 2020).

Large-scale sequencing efforts include projects such The Cancer Genome Atlas (TCGA) (Weinstein *et al.*, 2013), Therapeutically Applicable Research to Generate Effective Treatments (TARGET) (Ma *et al.*, 2018) and those under the umbrella of the International Cancer Genome Consortium (ICGC) (International Cancer Genome Consortium *et al.*, 2010; Zhao, Jones and Jones, 2019). They generate data primarily based on whole-exome sequencing but on a smaller scale, they also analyse whole genomes, transcriptomes and epigenomes of cancer patients (Nakagawa and Fujita, 2018). Such large-effort programmes have revealed many events that could be driving cancer. They have identified differentially expressed genes, gene fusions or aberrations in gene splicing through RNA-sequencing and chemical modifications of DNA and histones through ChIP-seq (Berger and Mardis, 2018). Finally, they identified a plethora of single nucleotide variants (SNVs), small deletions and insertions, copy number and structural variants in DNA sequences. Due to the domination of whole-exome sequencing, mutations in protein-coding regions have been primarily studied. Some variants in the promoter, intronic or untranslated regions have also been recognised although the information about such mutations is still very limited (Nakagawa and Fujita, 2018; Zhao, Jones and Jones, 2019).

Despite the possibilities offered by whole-genome sequencing for precision oncology, there has so far been limited use of it in the clinical setting given the time required to generate and analyse data (Goodwin, McPherson and McCombie, 2016). Thus the clinical applications of NGS currently consist primarily of targeted sequencing that detects biomarkers using cancer gene panels (Zhao, Jones and Jones, 2019). Sequencing only a discrete number of genes that have been well characterised and whose role in cancer progression has been demonstrated has the advantage of cost, simplicity of analysing data and hence also the time over whole-exome or whole-genome sequencing (Kamps *et al.*, 2017).

1.2.4 The Cancer Genome Atlas

The Cancer Genome Atlas (TCGA) was used to identify mutations potentially relevant to carcinogenesis and drug resistance while prioritising candidates for drivers of acquired drug resistance in cancer cell lines in Chapter 3 and Chapter 4 (see Methods: Variant calling section). The project was launched in December 2005 by the National Cancer Institute (NCI) and the National Human Genome Research Institute (Weinstein *et al.*, 2013). The programme aimed to characterise changes in DNA, RNA, protein and epigenetic profiles that drive human cancer (Weinstein *et al.*, 2013). Samples from over 11,000 cancer patients were collected over a period of ten years to achieve this goal (Liu *et al.*, 2018). As of November 2020, The Cancer Genome Atlas Research Network has published 36 papers analysing molecular profiles of 32 cancer types (*The Cancer Genome Atlas - Cancers Selected for Study - National Cancer Institute*, 2020). Samples were collected from cases associated with poor prognosis, high impact on public health, and they had to fulfil requirements of high quality. Data that was integrated and analysed by the TCGA project included data such as exome sequence, gene expression, copy-number variation, transcript splice variation and DNA methylation (Weinstein *et al.*, 2013).

Aside from the studies performed separately for each type of cancer, the Pan-Cancer initiative was launched to identify commonalities and differences across multiple cancer types (Weinstein *et al.*, 2013). It started in 2012 intending to analyse 12 tumour types already characterised by the TCGA using their genomic, epigenomic, transcriptional and proteomic profiles. According to the GDC (Genomic Data Commons) website, by August

2018, 58 papers were published within the TCGA programme, including 21 within the Pan-Cancer Atlas project (*Publications | NCI Genomic Data Commons*, 2020).

The original Pan-Cancer project focused on the analysis of the coding regions. However, the ICGC and TCGA joined-up their efforts within the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium to study whole-cancer genomes to identify variation in both coding and non-coding regions across 38 tumour types (Campbell *et al.*, 2020). The analysis includes characteristics such as non-coding driver mutations, mutational signatures, tumour evolution or fusion genes. According to a special feature in Nature, there are 22 papers published up until September 2020 describing the results from PCAWG (Campbell *et al.*, 2020).

TCGA changed the way we look at cancer (Weinstein *et al.*, 2013). The analysis of collected data has demonstrated that we should assess the disease prognosis and select the appropriate therapy for a cancer patient through a molecular profile based on genomic changes instead of basing decisions on cancer histology. It has shown that cancers having the same origin tissue do not always behave in the same way, and different cancer types may be more similar if they share genomic alterations. TCGA has also driven looking at affected pathways rather than genes. When the project started, scientists expected to find cancer-driving mutations in only a few genes. They quickly realised this is not the case, and a plethora of genes can be mutated to cause carcinogenesis. However, these mutations often affect genes acting within the same pathways (*Outcomes & Impact of The Cancer Genome Atlas - National Cancer Institute*, 2019).

TCGA is one of the first projects impacting precision medicine for cancer. One of the early studies that provided a better prognostic tool and an insight into a better selection of treatment for cancer patients studied 293 lower-grade gliomas from adults (The Cancer Genome Atlas Research Network, 2015). The analysis revealed that there are three distinct molecular classes of lower-grade gliomas and patients should be stratified not based on histopathology but rather on mutations in genes such as IDH and TP53, and the status of 1p/19q chromosomal regions. The most favourable outcome was observed in patients with an IDH mutation and 1p/19q codeletion. However, most lower-grade gliomas without an IDH mutation resembled glioblastoma both molecularly and clinically.

TCGA also contributed to the development of technologies that benefit disciplines outside of cancer research. This includes, for example, reducing costs of DNA and RNA sequencing (*Outcomes & Impact of The Cancer Genome Atlas - National Cancer Institute*, 2019).

The Cancer Genome Atlas is one of the cancer genomic datasets maintained by the Genomic Data Commons (GDC) – a research programme of the National Cancer Institute. GDC is a knowledgebase and data-sharing platform. It aims to promote precision medicine in cancer treatment. TARGET is the other big GDC programme that collects, analyses and shares data related to paediatric cancers (*About the GDC | NCI Genomic Data Commons*, 2020).

1.2.5 Sequence Ontology

The Sequence Ontology (SO) is a sister project of the Gene Ontology (Ashburner *et al.*, 2000). It was established in the early 2000s to "facilitate the exchange, analysis and management of genomic data" (Eilbeck *et al.*, 2005). It was an answer to the ambiguous vocabulary used to describe "the features and properties of biological sequence" and the impossibility of comparing data from different databases and research groups. Like the Gene Ontology, the Sequence Ontology is based on a set of standardised terms and the relationships between them, and it is also organised in a directed acyclic graph. The terms are often derived from annotations that are commonly used when describing genomic data. However, they are modified to be computer-friendly.

The current release of the Sequence Ontology consists of four sub-graphs relating to sequence attributes, collection, features and variants. Examples of the SO terms include "forward" describing the feature from 5' to 3' direction (belonging to "sequence_attribute" category), "chromosomally_aberrant_genome" used for features coming from genomes containing an abnormal amount of chromosomes ("sequence_collection"), "gene", "transcript_region", "polypeptide", "amino_acid" ("sequence_feature"), "missense_variant" or "5_prime_UTR_variant" ("sequence_variant") (*The MISO Sequence Ontology Browser*, 2016). We used children terms of the "sequence_variant" sub-graph to describe mutations (e.g. missense variants, see Figure 1.8) identified in the drug-adapted sub-lines in Chapter 3 and Chapter 4.

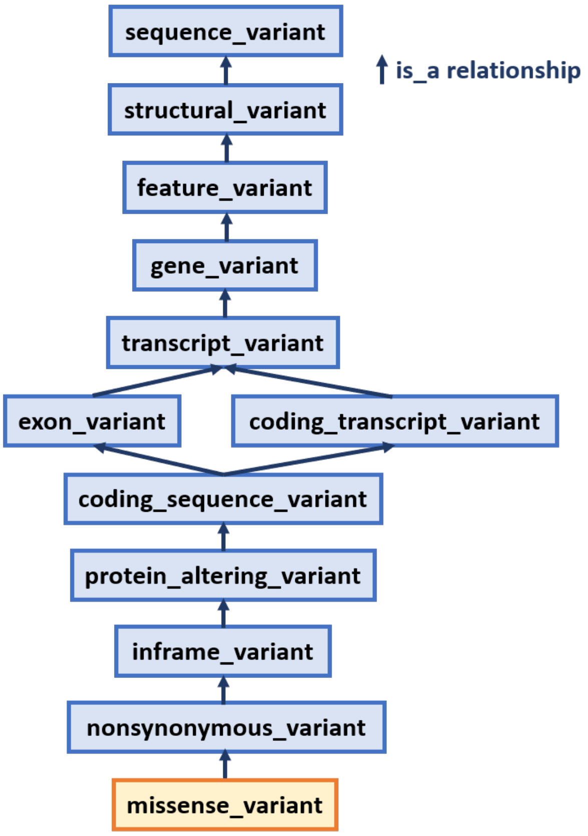


Figure 1.8 *Sequence Ontology sub-graph representing ancestry of the term "missense_variant". Figure adapted from* (Eilbeck *et al.*, 2005; *The MISO Sequence Ontology Browser - MISSENSE_VARIANT*, 2016).

1.2.6 Identification of cancer driver mutations: the holy grail of cancer genomics

Cancer is driven by genetic mutations such as single nucleotide variants (SNVs), small insertions and deletions (INDELs), copy-number alterations (CNAs) or structural variants (Cheng, Zhao and Zhao, 2016). They may have many different effects, including amino acid substitutions, change the reading frame, truncation of the encoded protein or changes to gene expression. This, in turn, may lead to loss or gain of protein function (Rajendran and Deng, 2017). However, not all mutations result in tumorigenesis. The majority of them do not provide tumour cells with a selective growth advantage, and they have minimal phenotypic effects (Liang and Kim, 2013; Cheng, Zhao and Zhao, 2016). These are called "passengers". To induce or progress carcinogenesis, the mutations need to impact certain essential functions called "cancer hallmarks". Mutations that cause carcinogenesis do it through, for example, preventing apoptosis and senescence, promoting cell division without any extracellular signals, initiating metastasis and angiogenesis, deregulating energy metabolism or circumventing immune response

(Gonzalez-Perez, Mustonen, *et al.*, 2013; Martínez-Jiménez, Muiños, Sentís, *et al.*, 2020). Mutations that contribute to cancer progression in the tumour cells are called “drivers” (Zhang *et al.*, 2014). Distinguishing between driver and passenger mutations is to this day, one of the biggest challenges of cancer genomics.

Advancement in this task is crucial for understanding how cancer is induced and how it progresses. It is also required for the development of prognostic and diagnostic biomarkers to be able to classify patients into cohorts that could benefit from a specific treatment (Jordan *et al.*, 2019). Finally, it is necessary for developing new drugs and providing patients with new treatment options tailored to their genetic profiles rather than the tissue of origin (Zhang *et al.*, 2014; Nussinov *et al.*, 2019).

Computational approaches offer a relatively inexpensive and fast way to identify potential cancer drivers and prioritise them for experimental validation and clinical application (Cheng, Zhao and Zhao, 2016), and a plethora of methods have been developed over the years. A few examples include SIFT (Sim *et al.*, 2012), PolyPhen-2 (Adzhubei *et al.*, 2010), Condel (González-Pérez and López-Bigas, 2011), FATHMM (Shihab *et al.*, 2013), dNdScv (Martincorena *et al.*, 2017), OncodriveFML (Mularoni *et al.*, 2016), OncodriveCLUSTL (Arnedo-Pac *et al.*, 2019), cBaSE (Weghorn and Sunyaev, 2017), Mutpanning (Dietlein *et al.*, 2020), HotMaps3D (Tokheim *et al.*, 2016), smRegions (Martínez-Jiménez, Muiños, López-Arribillaga, *et al.*, 2020), MutationAssessor (Reva, Antipin and Sander, 2007), MAPP (Stone and Sidow, 2005; Binkley *et al.*, 2010) and Logre (Clifford *et al.*, 2004). These tools explore many different aspects that characterise driver mutations over passengers.

Some methods assess the deleterious effect of a single nucleotide variation based on the assumption that higher functional impact implies a higher probability of being a driver (Zhang *et al.*, 2014; Nussinov *et al.*, 2019). Examples of such tools include SIFT (Sim *et al.*, 2012), PolyPhen-2 (Adzhubei *et al.*, 2010) and Condel (González-Pérez and López-Bigas, 2011) (described in details in sections 1.2.4.3, 1.2.4.4 and 1.2.4.5 respectively). They calculate amino acid conservation at the mutated position, compare amino acid properties or perform molecular dynamics simulations to assess changes in protein conformation and their impact on protein structure, stability and interactions with drugs (Cheng, Zhao and Zhao, 2016; Zhao *et al.*, 2018; Jordan *et al.*, 2019). The methods may also verify if the mutation occurs at the protein-protein or protein-nucleic acid interface

(Zhao *et al.*, 2018). It has been demonstrated that often approaches which integrate multiple sources of data perform better than individual tools (González-Pérez and López-Bigas, 2011; Cheng, Zhao and Zhao, 2016). The reason for that is that different methods explore different biological hypotheses and combining tools based on complementary premises can improve the accuracy of predictions. Some tools integrate various sources of information within machine learning algorithms and use them as features. They build a classifier based on the properties of already known passenger and driver mutations (Zhang *et al.*, 2014).

In Chapter 3 and Chapter 4, we applied SIFT and PolyPhen-2 to assess if the mutations that are candidate drivers of acquired drug resistance will have a damaging impact on the protein (see Methods: Variant calling section of both chapters).

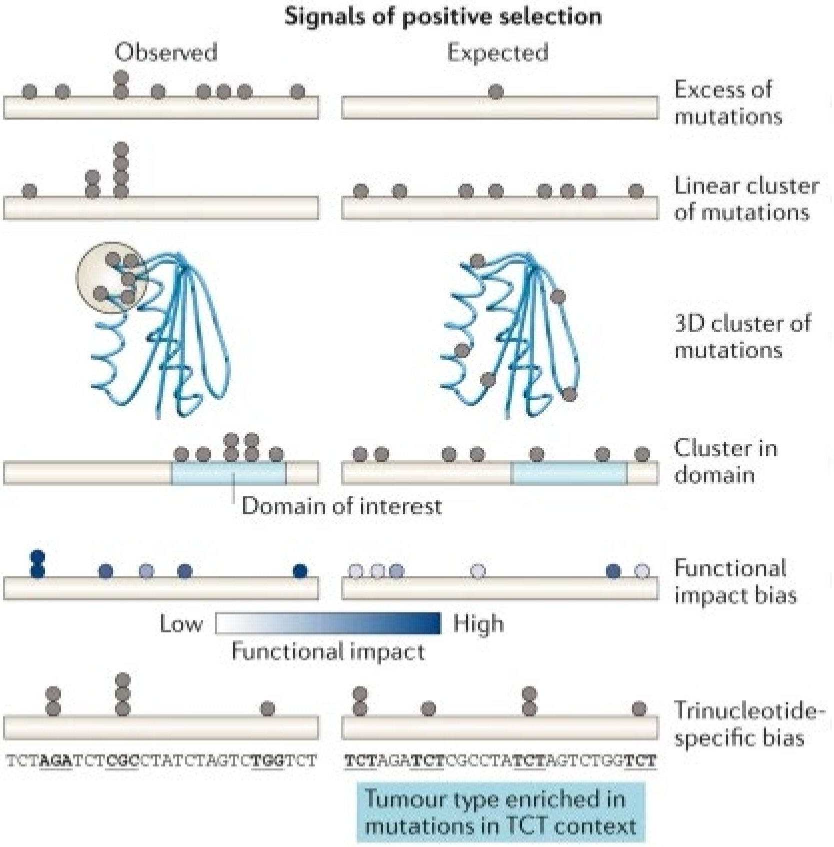

Signals of positive selection
Observed
Expected
Excess of mutations
Linear cluster of mutations
3D cluster of mutations
Cluster in domain
Domain of interest
Functional impact bias
Low
High
Functional impact
Trinucleotide-specific bias
TCTAGATCTCGCCTATCTAGTCTGGTCT
TCTAGATCTCGCCTATCTAGTCTGGTCT
Tumour type enriched in mutations in TCT context

Figure 1.9 *Signals of positive selection that are used to identify cancer driver genes across a cohort of cancer patients. Figure adapted from* (Martínez-Jiménez, Muiños, Sentís, *et al.*, 2020).

Another approach of identifying cancer gene drivers is to analyse the enrichment of mutations in specific pathways and networks (Zhang *et al.*, 2014). Many other methods identify cancer driver genes by detecting genes positively selected during tumour evolution (Martínez-Jiménez, Muiños, Sentís, *et al.*, 2020). They do that through analysing data from a cohort of cancer patients. Some tools explore accumulation in a gene of mutations with high functional impact (Figure 1.9). Other algorithms are based on the frequency of mutations across a cohort of tumour samples, and they identify genes that develop more mutations than expected. Since driver mutations provide tumour cells with a growth advantage and they are positively selected during cancer progression, given that a similar selective pressure act on different patients, the same mutation(s) should be selected and traced with the frequency rate across the patients' cohort (Gonzalez-Perez, Mustonen, *et al.*, 2013). The background mutation rate is calculated based on, for example, silent mutations in coding regions taking into account gene size and the composition of the nucleotides (Zhang *et al.*, 2014; Cheng, Zhao and Zhao, 2016). Another sign of positive selection that is commonly used to identify cancer driver genes is the clustering of mutations in a specific functional domain within the protein sequence or structure across patients' samples (Gonzalez-Perez, Mustonen, *et al.*, 2013; Hudson *et al.*, 2015; Nussinov *et al.*, 2019). Finally, some methods also verify the bias of acquiring mutations in a specific trinucleotide context (Martínez-Jiménez, Muiños, Sentís, *et al.*, 2020).

1.2.6.1 Catalogue of Somatic Mutations in Cancer

Similarly to The Cancer Genome Atlas, The Catalogue of Somatic Mutations in Cancer (COSMIC) was used to identify mutations potentially relevant to carcinogenesis and drug resistance while prioritising candidates for drivers of acquired drug resistance in cancer cell lines in Chapter 3 and Chapter 4 (see Methods: Variant calling section). It is an initiative undertaken by Wellcome Trust Sanger Institute. Somatic mutation data has been stored in the COSMIC database for over 15 years now. The project started with data collected for only four genes, HRAS, KRAS2, NRAS and BRAF, known to be somatically mutated in cancer (Bamford *et al.*, 2004). One thousand four hundred eighty-three papers were reviewed to identify a total of 10,647 mutations across tumours with mutations in these four genes. The project has expanded from this initial study, and the most recent publication (Tate *et al.*, 2019) reported COSMIC release v86 (August 2018) containing 5,977,977 coding mutations from over 26,251 publications. COSMIC also includes mutations in non-coding regions, gene fusions, copy-number alterations and mutations associated with drug resistance. There are two ways in which data is collected into COSMIC. One of them has always been manual literature curation. However, with the advent of next-generation sequencing, COSMIC started incorporating information from large-scale systematic screens and cancer data portals such as TCGA (Weinstein *et al.*, 2013).

There are now four other datasets available that are part of the COSMIC project and complement the expert-curated database somatic mutations in cancer. These are the Cancer Gene Census, Cell Lines Project, COSMIC-3D and, available from August 2020 (COSMIC v92), the Cancer Mutation Census (Tate *et al.*, 2019; *NEW PRODUCT: Discover the Cancer Mutation Census*, 2020). The Cell Lines Project incorporates data from over 1,000 cell lines (1,015 as of COSMIC v86), specifically their exome sequences and molecular profiles (Tate *et al.*, 2019). In contrast, Cancer Gene Census contains information about 719 genes that drive human cancer through somatic and germline mutations (as of COSMIC v86). It classifies genes as oncogenes, tumour suppressor genes, or both. It also assigns genes into two tiers based on the confidence of the role they have in cancer. COSMIC-3D maps COSMIC mutations to protein sequence and structure to provide a better way of understanding what impact the mutation may have on protein function and also for exploring potential drug targets (Tate *et al.*, 2019). Finally, the Cancer Mutation Census was developed to facilitate the selection of driver mutations over the passenger mutations. It comprises metrics such as ClinVar (Landrum *et al.*, 2018) significance or variant frequency in the gnomAD (Karczewski *et al.*, 2020) database (*NEW PRODUCT: Discover the Cancer Mutation Census*, 2020).

1.2.6.2 The Integrative OncoGenomics pipeline

The Integrative OncoGenomics (IntOGen) pipeline aims to identify cancer driver mutations systematically (Martínez-Jiménez, Muiños, Sentís, *et al.*, 2020). It was designed as a tool complementary to literature curation. IntOGen combines somatic SNVs and short indels from multiple studies and uses seven tools that identify drivers to create a compendium of cancer driver genes. The current version of IntOGen uses data from 28,076 tumour samples. The majority of them come from large sequencing programmes such as the ICGC (International Cancer Genome Consortium *et al.*, 2010), TCGA (Weinstein *et al.*, 2013), PCAWG (Campbell *et al.*, 2020), Hartwig Medical Foundation (Priestley *et al.*, 2019) and TARGET (Ma *et al.*, 2018).

The IntOGen pipeline consists of three steps: pre-processing of the samples, identification of cancer driver genes and lastly post-processing of the drivers. Pre-processing includes mapping all the bases to the GRCh38 human genome version to be consistent across all datasets, removing mutations that have ‘N’ as the reference or the alternative allele or in which the reference allele is the same as an alternative. Other steps undertaken within the pre-processing stage comprise of actions such as removing mutations that have a high chance of being mapped somewhere else in the genome, filtering out multiple samples from the same donor, removing hypermutated samples, or discarding datasets without synonymous variants as they are necessary to build the background mutation model.

The second step of the pipeline, detection of cancer driver genes, is performed using seven tools: dNdScv (Martincorena *et al.*, 2017), OncodriveFML (Mularoni *et al.*, 2016), OncodriveCLUSTL (Arnedo-Pac *et al.*, 2019), cBaSE (Weghorn and Sunyaev, 2017), Mutpanning (Dietlein *et al.*, 2020), HotMaps3D (Tokheim *et al.*, 2016) and smRegions (Martínez-Jiménez, Muiños, López-Arribillaga, *et al.*, 2020). Their mode of action is based on finding genes that undergo positive selection across a cohort of tumours. This is done by identifying mutational patterns different from those expected as a result of neutral mutagenesis model. These patterns cover clustering of mutations within the sequence, structure or a particular domain, excess numbers of mutations in general and the number of non-synonymous versus synonymous variants. Finally, they also involve biases towards mutations with highly damaging functional impact and in specific tri-nucleotide contexts (Figure 1.9).

The last step of the IntOGen pipeline consists of post-processing of the identified cancer driver genes. This includes genes that are not expressed based on TCGA data, highly tolerant to developing Single Nucleotide Polymorphisms (SNPs) in the human population or those that could be potentially mutated as a result of a local hypermutation activity.

Finally, in addition to predicting potential cancer driver genes, IntOGen also determines their mode of action as either activating (oncogene), loss-of-function (tumour suppressor) or ambiguous. This classification is based on the enrichment of missense versus nonsense mutations (oncogene) and vice versa (tumour suppressor).

IntOGen was used to identify drivers of acquired drug resistance in cancer cell lines in Chapter 3.

1.2.6.3 Sorting Intolerant substitutions From Tolerant (SIFT)

Sorting Intolerant substitutions From Tolerant (SIFT) predicts how amino acid substitutions may affect the protein (Ng and Henikoff, 2001). Introduced in 2001, it is one of the earliest tools developed to assess variant pathogenicity, and it has been commonly used ever since (Sim *et al.*, 2012). It is solely based on sequence homology and the assumption that functional residues are conserved within protein families. Thus it does not need any information about protein structure or function. However, due to the same reason, SIFT requires that homologous sequences are available for the query protein, and it has the best accuracy when accurate alignments of orthologues are used. SIFT first identifies homologues of the query sequence using PSI-BLAST (Altschul *et al.*, 1997) and retains only the consensus sequences for the groups of proteins with more than 90% identity. A seed profile is created from the query sequence and the consensus sequences with the similarity of over 90%. The profile is used to search similar sequences using PSI-BLAST, and they are added to the initial profile as long as the conservation of the new profile does not drop below the user-defined threshold (Ng and Henikoff, 2002). The sequences collected in the previous step are aligned, and the final step is to calculate the probability of each amino acid appearing at each position. Finally, a cut-off threshold is applied to decide if the substitution is deleterious or tolerant.

SIFT was applied in Chapter 3 and Chapter 4 to assess if the mutations that are candidate drivers of acquired drug resistance will have a damaging impact on the protein (see Methods: Variant calling section of both chapters).

1.2.6.4 PolyPhen

Similarly to SIFT (Sim *et al.*, 2012), PolyPhen-2 predicts the effect of amino acid substitutions on protein structure and function (Adzhubei *et al.*, 2010). The method consists of a Naïve Bayes classifier making predictions based on eleven features related to the properties relating to protein sequence and structure. The multiple-sequence alignment (MSA) of homologous sequences is used to extract some of these features. MSAs are generated by first using BLAST+ (Camacho *et al.*, 2009) to search UniRef100 (Bateman, 2019), then aligning the query sequence homologues using MAFFT (Katoh *et al.*, 2002) with a refinement step done with LEON (Thompson, Prigent and Poch, 2004). Some features are based on Pfam (El-Gebali *et al.*, 2019) domains, alignment depth and

CpG context. Others refer to accessible surface area and conformational mobility of the amino acid residue, change in the volume of its side chain, and identity of the query sequence to the closest homologue mutated at the queried position. Finally, the algorithm also uses Position-Specific Independent Counts (PSIC) (Sunyaev *et al.*, 1999) which are scores demonstrating the likelihood of a certain amino acid being present at a specific position and a score representing consistency of the mutated sequence with the MSA.

PolyPhen-2 was trained on two datasets. One of them combined alleles known to cause Mendelian diseases with non-damaging differences between human proteins and their homologues in mammals. The other dataset consisted of all human disease-causing mutations and human non-damaging non-synonymous SNPs. Each of them is suited better for a different task – one in diagnosing Mendelian diseases and the other for identifying complex phenotypes or cases where even mildly deleterious mutations have to be considered damaging.

PolyPhen-2 was applied in Chapter 3 and Chapter 4 to assess if the mutations that are candidate drivers of acquired drug resistance will have a damaging impact on the protein (see Methods: Variant calling section of both chapters).

1.2.6.5 Condel

Condel was developed to benefit from the complementary performance of tools predicting the impact of non-synonymous Single Nucleotide Variants on protein function (González-Pérez and López-Bigas, 2011; *Condel — FannsDB 2.0-dev documentation*, 2014). Initially, it incorporated output from five methods: SIFT (Sim *et al.*, 2012), PolyPhen-2 (Adzhubei *et al.*, 2010), MutationAssessor (Reva, Antipin and Sander, 2007), MAPP (Stone and Sidow, 2005; Binkley *et al.*, 2010) and Logre (Clifford *et al.*, 2004). Each of the tools returns a value that represents the tolerance of the specific mutation at the specific position. The final result, called CONsensus DELeteriousness score, is calculated from normalised scores multiplied by a weight. The scores are obtained from all five tools, and the weights are based on the probability that a predicted deleterious mutation is truly deleterious, and a neutral mutation is truly neutral.

The comparison of the tool combining the outputs of the five methods using such weighted average scores and the five tools separately revealed that Condel outperformed each of the individual algorithms. The current version of Condel incorporates only results from FATHMM (Shihab *et al.*, 2013) and MutationAssessor. The initial aim was to include also the outputs from SIFT and PolyPhen-2. However, adding them tended to decrease Condel's performance, and thus they were both omitted.

1.3 Scope and outline of this thesis

This thesis reports research from three manuscripts, one that investigates the essential functions of life in the minimal bacterial genome and two that explore mutations identified in drug adapted cancer cell lines to examine drivers of acquired drug resistance.

Chapter 2 – Environmental Conditions shape the nature of a minimal bacterial genome.

This chapter was published in July 2019 – Antczak M, Michaelis M, Wass MN Nature Commun 10:3100.

Chapter 3 – Acquired MDM2 inhibitor resistance is associated with collateral sensitivity to cytarabine in acute myeloid leukaemia cells.

This chapter reports the analysis of variants acquired by four sub-lines of the Molm13 cell line in the process of adaptation to nutlin-3 and their impact on resistance.

Chapter 4 – Selection of different clones upon repeated adaptation of neuroblastoma cell lines to tubulin-binding agents.

This chapter reports the analysis of variants acquired by 41 sub-lines of the UKF-NB-3 cell line in the process of adaptation to four tubulin-binding agents (eribulin, vincristine, 2-methoxyestradiol and epothilone b) and their impact on the resistance.

Chapter 5 – Discussion. This chapter discusses the research presented in chapters 2-5 and future work.

Appendix 1 – Supplementary figures and tables for Chapter 2.

Appendix 2 – Supplementary data files for Chapter 2.

Appendix 3 – Supplementary figures and tables for Chapter 3.

Appendix 4 – Supplementary tables for Chapter 3.

Appendix 5 – Supplementary figures and tables for Chapter 4.

Appendix 6 – Supplementary tables for Chapter 4.

Appendix 7 – Participation in the third edition of the Critical Assessment of Functional Annotation: the Gene Ontology Annotation Tool (GOAT).

.

Chapter 2 Environmental conditions shape the nature of a minimal bacterial genome

Antczak, M., Michaelis, M. & Wass, M.N. Environmental conditions shape the nature of a minimal bacterial genome. Nat Commun 10, 3100 (2019).
https://doi.org/10.1038/s41467-019-10837-2

2.1 My contribution

I performed all of the analysis in the study and generated all the figures. I wrote the paper's first draft, which I then worked on with Mark Wass and Martin Michaelis.

2.2 Abstract

Of the 473 genes in the genome of the bacterium with the smallest genome generated to date, 149 genes have unknown function, emphasising a universal problem; less than 1% of proteins have experimentally determined annotations. Here, we combine the results from state-of-the-art in silico methods for functional annotation and assign functions to 66 of the 149 proteins. Proteins that are still not annotated lack orthologues, lack protein domains, and/ or are membrane proteins. Twenty-four likely transporter proteins are identified indicating the importance of nutrient uptake into and waste disposal out of the minimal bacterial cell in a nutrient-rich environment after removal of metabolic enzymes. Hence, the environment shapes the nature of a minimal genome. Our findings also show that the combination of multiple different state-of-the-art in silico methods for annotating proteins is able to predict functions, even for difficult to characterise proteins and identify crucial gaps for further development.

2.3 Introduction

A long-term goal of synthetic biology has been the identification of the minimal genome, i.e., the smallest set of genes required to support a living organism. The bacterium with the smallest genome generated to date is based on *Mycoplasma mycoides* (Hutchison *et al.*, 2016). Its minimal bacterial genome consists of 473 genes including essential genes and a set of genes associated with growth, termed 'quasi-essential' (Hutchison *et al.*, 2016). The minimal genome study assigned function to proteins encoded by the minimal genome by considering matches to existing protein families in the TIGRFAM (Haft *et al.*, 2013) database, genome context and structural modelling (Hutchison *et al.*, 2016). Proteins were annotated with molecular functions and grouped into 30 biological process categories (including an unclear category, where the biological process was not known). The proteins were further assigned to five classes according to the specificity and confidence of the molecular function annotations that they had been assigned: Equivalog (confident hits to TIGRFAM families), Probable (low confidence match to TIGRFAM families supported by genome context or threading), Putative (multiple sources of evidence but lower confidence), Generic (general functional information identifiable, e.g., DNA binding or membrane protein, but specific function unknown) and Unknown (unable to infer even a general function). The final two confidence classes, Unknown (65 genes) and Generic (84 genes) form the group of genes whose function is unknown. Hence, almost a third (149) of the encoded 473 proteins are of unknown function, which emphasises our limited understanding of biological systems (Hutchison *et al.*, 2016).

This lack of functional annotation is not restricted to the minimal bacterial genome. One-third of protein-coding genes from bacterial genomes lack functional annotations (Chang *et al.*, 2016). Recent experimental approaches have begun to identify the function of 'hypothetical' proteins of unknown function (Price *et al.*, 2018). However, the continual improvement of high-throughput sequencing methods has resulted in a rapid increase in the number of organisms for which genome sequences are available and the functional annotation of the encoded gene products lags behind (Price *et al.*, 2018). Less than 1% of the 148 million protein sequences in UniProt (Bateman *et al.*, 2017) are annotated with experimentally confirmed functions in the Gene Ontology (GO) (Carbon *et al.*, 2017) (April 2019). To address this gap, computational methods for protein function prediction have been developed and significantly advanced over the past 15 years as demonstrated by the recent Critical Assessment of Functional Annotation (CAFA) challenges (Radivojac *et al.*, 2013), (Jiang *et al.*, 2016).

Here, we perform an extensive in silico analysis of the proteins of unknown function encoded by the minimal bacterial genome using an approach that combines 22 different computational methods ranging from identification of basic properties (e.g., protein domains, disorder and transmembrane helices) to state-of-the-art protein structural modelling and methods that infer GO-based protein functions, including those that have performed well in CAFA experiments.

2.4 Methods

2.4.1 Identifying basic protein properties

Protein domains were determined by running PfamScan against the library of Pfam 30.0 HMMs (Finn *et al.*, 2016). GO terms associated with Pfam domains were extracted using the pfam2go file (Finn *et al.*, 2016) (version 11 February 2017). The *e*-value of the domain matches were used to indicate the confidence of a GO term describing the function of the query protein. To test if the probability of minimal genome proteins having more domains identified increases with the increasing confidence of the annotation in the particular functional class, we performed the Mann–Whitney–Wilcoxon test. We cross-compared all the functional classes (438 proteins in total) and tested the null hypothesis that samples have the same distribution against the alternative hypothesis that there is a >0 shift in the distribution.

InterProScan was run with default settings to determine matches against InterPro databases of protein signatures (Mitchell *et al.*, 2019). Results from the following resources were included in the analysis: CDD (Marchler-Bauer *et al.*, 2017), Gene3D (Lewis *et al.*, 2018), HAMAP (Pedruzzi *et al.*, 2015), PIRSF (Wu *et al.*, 2004), PRINTS (Attwood, 2012), ProDom (Servant *et al.*, 2002), ProSitePatterns (Sigrist *et al.*, 2013), ProSiteProfiles (Sigrist *et al.*, 2013), SFLD (Akiva *et al.*, 2014), SMART (Letunic and Bork, 2018) and SUPERFAMILY (Oates *et al.*, 2015).

Orthologues were identified using eggNOG-Mapper (Huerta-Cepas *et al.*, 2017) against HMM databases for the three kingdoms of life. Additionally, precision of predictions was prioritised by restricting results to only one-to-one orthologues. The eggNOG-Mapper API was used to predict the orthologous groups in eggNOG that the minimal genome proteins belonged to. The proteins present in these orthologous groups were extracted and the species associated with the sequences were mapped to the NCBI Taxonomy to group them into phyla and used to identify the phyla where orthologues were present. Predicted features including GO terms, KEGG pathways and functional categories of Cluster of Orthologous Groups were also obtained from eggNOG-Mapper.

2.4.2 Identifying membrane transporters and lipoproteins

Proteins were classified as lipoproteins (SPaseI-cleaved proteins), SPaseI-cleaved proteins, cytoplasmic and transmembrane proteins using LipoP (Juncker *et al.*, 2003). Similarly, proteins were distinguished between membrane transporters and non-transporters using TrSSP (Mishra, Chang and Zhao, 2014). TrSSP predicted substrates of the proteins from seven groups: amino acid, anion, cation, electron, protein/mRNA, sugar and other. The functions of membrane transporters and lipoproteins were further supported by identifying transmembrane helices, signal peptides and protein topology using TMHMM (Krogh *et al.*, 2001).

2.4.3 Inferring gene ontology-based protein function

GO terms were predicted using FFPred3 (Cozzetto *et al.*, 2016), Argot2.5 (Falda *et al.*, 2012), GOAT (only Molecular Function terms; check Appendix 7 for details) and LocTree3 (Goldberg *et al.*, 2014) (only Celullar Component terms). As the FFPred3 SVMs were trained only on human proteins from UniProtKB, predicted GO terms were additionally filtered using the frequency of terms in UniProtKB-GOA (version 5 June 2017). Predicted GO terms that were not annotated to any bacterial proteins in UniProtKB-GOA were removed from the set of FFPred3 predicted functions as they were likely to be functions that are not present in prokaryotes.

Argot2.5 was run with the taxonomic constraints option. As scores returned by Argot2.5 have a minimum score of zero and no upper bound, the linear spline function recommended by the method developers (personal communication) was applied to rescale them to the range of 0 to 1.

2.4.4 Structural analysis

The CATH FunFHMMer webserver was used to identify the functional families of structural domains, CATH FunFams (Sillitoe *et al.*, 2013; Das *et al.*, 2015).

Protein disorder was predicted using DISOPRED3 (Jones and Cozzetto, 2015). For each of the proteins, the percentage of disordered regions was calculated based on the DISOPRED3 results. To verify if there is a statistically significant difference between 438 minimal genome proteins in five different functional classes, we performed a Chi-Square test for categorical data with a null hypothesis that the functional class of a protein and its disorder ratio level (0%, (0%, 10%], (10%, 20%], (20%, 30%], >30%) are independent.

Firestar (Lopez *et al.*, 2011) and 3DLigandSite (Wass, Kelley and Sternberg, 2010) were used to predict ligands binding to the proteins. For Firestar only results marked as cognate were considered. Phyre2 (Kelley *et al.*, 2015) was run using standard mode to model the structure of the minimal genome proteins. Information provided by the name and description of the best matching models was used in the process of inferring function of the proteins. To make sure that each residue was covered with the highest possible confidence, the matches were firstly sorted by *e*-value and then selected gradually if they covered residues that were not covered before by a match with lower *e*-value.

2.4.5 Identifying operons

Genes in the synthetic *M. mycoides* (JCVI-syn1.0) were grouped into operons based on the predictions made for both *M. mycoides* subsp *capri* LC str 95010 and *M. mycoides* subsp *mycoides* SC str PG1 by two methods DOOR2 (Mao *et al.*, 2014) and MicrobesOnline (Alm *et al.*, 2005). The proteins of the synthetic *M. mycoides* were first mapped to the proteins *of M. mycoides* subsp *capri* LC str 95010 and *M. mycoides* subsp *mycoides* SC str PG1 downloaded from GenBank (Benson *et al.*, 2017). This was done by using BLAST to search against databases constructed from proteomes of these two species and extracting the best hit. A protein from *M. mycoides* subsp *capri* LC str 95010 or *M. mycoides* subsp *mycoides* SC str PG1 was considered a corresponding homologue of a protein from synthetic *M. mycoides* if the coverage and identity were greater than or equal to 80%. Via the corresponding homologues, operons predicted for these two species by DOOR2 and MicrobesOnline were mapped to the proteins of the synthetic *M. mycoides*.

2.4.6 Combined protein function prediction

The results from the following methods were removed from the analysis if their *e*-value was above 0.001: TIGRFAM, Pfam, eggNOG-Mapper, CATH FunFams and domains. Models predicted by Phyre2 were kept if the probability of the match was above 80% and *e*-value was below 0.001. Only results from Firestar with a reliability score above 70% and marked as cognate were retained. Ligands predicted by 3DLigandSite were kept if they were included in at least three homologous models. The best BLAST hit from UniProt (maximum *e*-value of 0.001) was used to identify the closest homologue of the protein and the information accessible in UniProt was taken into account in the annotation. Additionally, all the predictions of Gene Ontology terms were combined together and the probability of particular terms being predicted by any of the methods were calculated using the following formula: $P(GO) = 1 – (1 – P\ (GO_{FFPred3}))^{*}\ (1 – P(GO_{Argot2.5}))^{*}\ (1 – P(GO_{GOAT})^{*}\ (1 – P(GO_{LocTree3}))$, where P(GO) is the combined probability of a given GO term and where subscripts are included this indicates the probability of that term from the named individual method. Only high probability (> 0.65) Gene Ontology terms were considered for each of the proteins. For the final prediction of protein function, results from all the methods were manually reviewed. The initial proposition of protein function was based on combining the results from TIGRFAM equivalog families, Pfam domains, InterPro families and domains, eggNOG orthologous groups, CATH functional families, the best BLAST hit from UniProt and the Phyre2 model of the structure. In considering the results from these methods, we looked for agreement between methods, particularly with highly confident results. This initial function was then verified using the predicted Gene Ontology terms and information on predicted ligands (Firestar (Lopez *et al.*, 2011), 3DLigandSite (Wass, Kelley and Sternberg, 2010)) and transmembrane helices (TMHMM). Where information was not available from the first group of methods, the second group of methods were used as a starting point to infer functions. Transporters and lipoproteins were predicted using membrane transporter (TrSSP) and lipoprotein signal sequences

(LipoP) respectively. Finally, it was considered if the predicted function was consistent within a group of genes in the same operon. Where methods made predictions that conflicted with the final predicted function, this was noted, but it did not affect the confidence as we recorded the number of methods supporting a function and the average score associated with these predictions (see below).

Confidence of predicted functions was considered for each protein by counting the results that support the final function and calculating the average score from these methods. Results used to calculate the average score come from the methods applied in the first step of function prediction (Figure 2.4), i.e., TIGRFAM, Pfam, InterPro resources (all but ProSitePatterns), eggNog-Mapper, BLAST, CATH FunFams, Phyre2, and also the overall GO term-based prediction (which already combined Argot2.5, GOAT, FFPred3 and LocTree3) resulting in 17 methods in total. Methods that concern a very specific element of a function, such as transmembrane helices or ligands were not included in the average score calculation. For all methods, scores were normalised to the range 0–100. Most of the methods use *e*-values as a measure of confidence (e.g., TIGRFAM, Pfam), for these methods –log10(*e*-value) was used capping the value at 100. Where probabilities were provided these were multiplied by 100. HAMAP and ProSiteProfiles use scores that are not probabilities or *e*-values and do not appear to have an upper bound. Considering the scores of these methods for the proteins of known function in the minimal genome indicated that scores were typically in the range 0–100 (Figure S7 and Figure S8), so scores above 100 were capped at 100.

2.5 Results

2.5.1 Orthologues for the proteins in the minimal genome

Hutchison et al. (Hutchison *et al.*, 2016) used BLAST to identify homologues of the minimal genome proteins in a set of 14 species ranging from non-mycoides mycoplasmas to archaea. They found that while many of the proteins from the Equivalog, Probable, Putative and Generic classes have homologues in all 14 species, very few of the sequences in the Unknown class had homologues outside of *M. mycoides*, with none in *M. tuberculosis*, *A. thaliana*, *S. cerevisiae* and *M. jannaschii*.

Here, eggNOG-Mapper (Huerta-Cepas *et al.*, 2017) (see methods) was used to identify orthologues for the minimal genome proteins across the three kingdoms of life. Overall the analysis showed that very few of the Unknown class of proteins (7%) have related sequences in eukaryotes or archaea (6%) while just over half (55%) have orthologues in other bacterial species, primarily in terrabacteria, the clade that *M. mycoides* belongs to (Figure 2.1a, Figure S1 and Appendix 2 Supplementary Data File 1). In contrast, many of the proteins in the other confidence classes have orthologues across the three kingdoms (Figure 2.1a and Figure S1). For example, 63%, 59% and 95% of the proteins in the Generic class have orthologues in eukaryotes, archaea and bacteria, respectively (Figure 2.1a and Figure S1), rising to 91%, 70% and 99% for the Equivalog class. Only two proteins from the Unknown class had many orthologues in both eukaryotes and archaea. These proteins MMSYN1_0298 and MMSYN1_0302 were classified by Hutchison et al. into the Unclear and Cofactor transport and salvage functional categories, respectively. Our analysis determined confident functions for both of these proteins (see below).

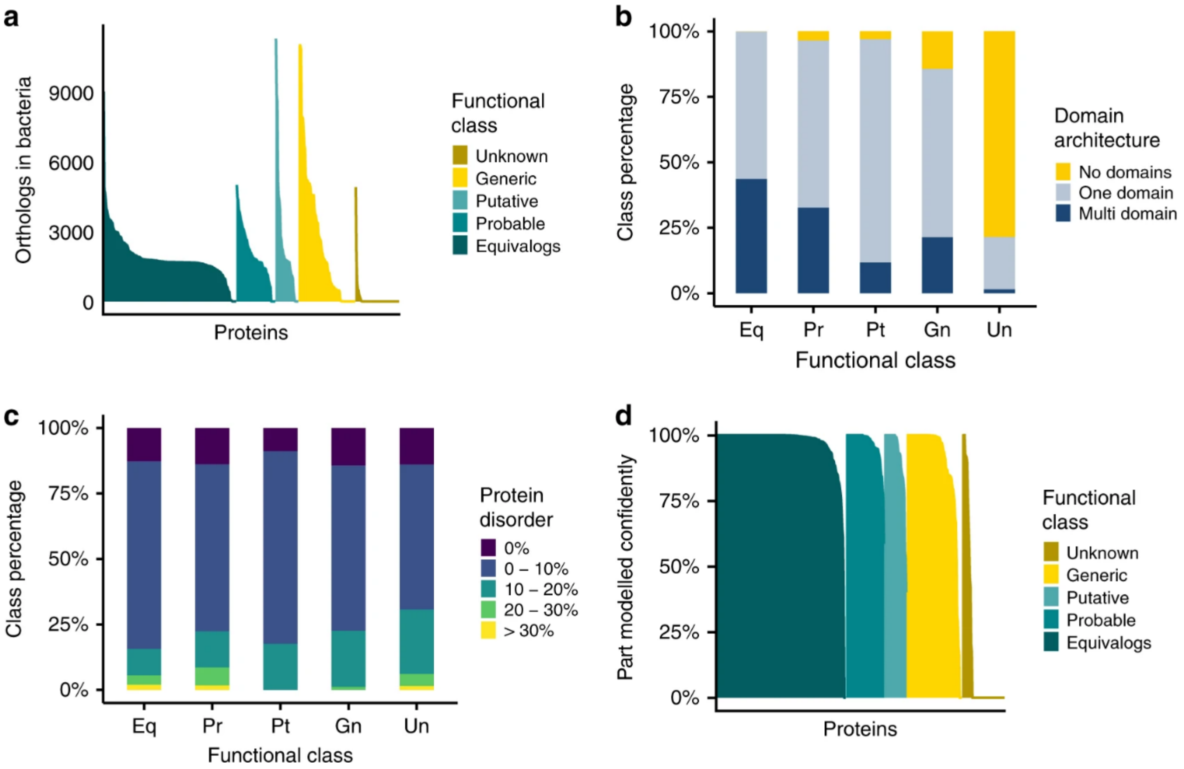


Figure 2.1 *Basic characterisation of proteins encoded by the minimal bacterial genome. a Orthologues identified in bacteria. Results for each functional class are represented by a different colour: gold for the Unknown class, yellow–Generic, light turquoise–Putative, turquoise–Probable and dark turquoise–*

Equivalog. b The domain architecture for proteins in each of the five functional confidence classes is plotted (Unknown [Un], Generic [Gn], Putative [Pt], Probable [Pr] and Equivalog [Eq]). It is represented as a number of matches to domains present in Pfam (Finn *et al.*, 2016)*. "No domains" signifies that no Pfam domains were detected in the protein. Proteins with no domains are displayed in yellow, grey represents single domain proteins and dark blue multi-domain proteins. c Predicted protein disorder in the minimal genome proteins. The results are shown for the five confidence classes from b and coloured according to the percentage of disorder present. Proteins with a percentage disorder >30% are represented by yellow, 20–30% disorder by green, 10–20%-turquoise and 0–10%-blue. Purple indicates proteins without disordered regions. d The percentage of protein structure that can be confidently modelled by Phyre2. Functional class colouring as for a.*

2.5.2 Domain architecture of minimal genome proteins

Domain analysis, using Pfam (Finn *et al.*, 2016) (Appendix 2 Supplementary Data File 2), showed that few (22%) of the proteins in the Unknown class contain known domains, significantly less than for the other four classes (Figure 2.1b; $p < 8.3e\text{-}12$; Mann–Whitney–Wilcoxon test). In contrast, all proteins in the Equivalog class contain at least one domain and nearly half of them (44%) have a multi-domain architecture (Figure 2.1b), whereas multiple domains are present in 21% of the proteins in the Generic class and only a single protein in the Unknown class (Figure 2.1b). The proteins in the Unknown class are also clearly different to those in the Generic class, where a domain is present in 86% of the proteins. Further, the proteins in the Unknown class also have more disordered regions than the other groups (Figure 2.1c), although this does not reach statistical significance (X-squared = 19.304, df = 16, p = 0.2532; Chi-Square test for categorical data).

2.5.3 Structural modelling of the minimal genome

Hutchison et al. (Hutchison *et al.*, 2016) used threading (an approach for modelling protein structure) to support functional assignment from TIGRFAM matches. Here, the Phyre2 (Kelley *et al.*, 2015) protein structure prediction server was used to model the structures of the minimal genome proteins. With the exception of the Unknown class, high confidence structural templates were identified for the vast majority of proteins for at least part of the sequence (Figure S2 and Appendix 2 Supplementary Data File 3). The proportion of proteins in each confidence class that could be accurately modelled was considered by identifying those for which at least 75% of the protein sequence could be modelled with a structural model confidence score (from Phyre2) of at least 90%. In the

Unknown class this applied to only nine proteins, whereas nearly all proteins in the four other confidence groups were successfully modelled (Figure 2.1d).

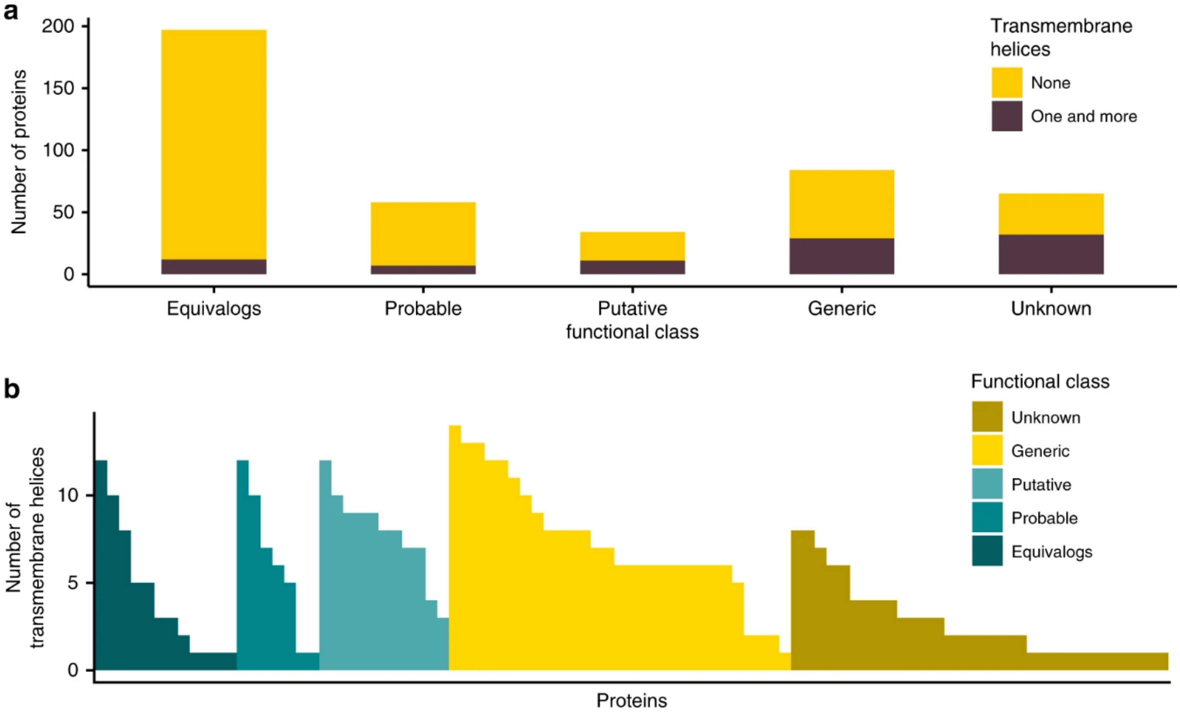


Figure 2.2 *Transmembrane proteins encoded by the minimal bacterial genome. a The number of proteins predicted by TMHMM to have transmembrane helices. Brown indicates proteins with one or more transmembrane helix. Yellow for those without transmembrane helices. b The number of transmembrane helices present in each of the proteins in the minimal genome that is predicted to have one or more transmembrane helix. Results for each functional class are represented by a different colour: gold for the Unknown class, yellow–Generic, light turquoise–Putative, turquoise–Probable and dark turquoise–Equivalog class*

2.5.4 Transmembrane proteins

Proteins in the Unknown and Generic classes are enriched with transmembrane proteins with 49% and 35%, respectively, of their proteins predicted to have transmembrane helices (Figure 2.2a and Appendix 2 Supplementary Data File 4). In contrast, very few transmembrane proteins were identified in the Equivalog and Probable classes (6% and 12% respectively), while 32% of the proteins in the Putative class are transmembrane proteins (Figure 2.2a).

These results suggest that many of the proteins that have unassigned functions may be associated with membranes. For example, 24 proteins in the Generic class are predicted to contain six or more transmembrane helices (Figure 2.2b), many of which are likely to be transporters of essential nutrients from the media (see below).

2.5.5 Prediction classification for specificity and confidence

To infer functions for the proteins of unknown function, we introduced a different way to classify our results, which separates function specificity and prediction confidence. This enabled a more nuanced interpretation of the results than the five classes (Unknown to Equivalog) used by Hutchison et al., which combined both specificity and confidence. Our specificity classes include ‘hypothetical’, where the function is completely unknown, ‘general’, where we have some basic functional information (e.g., DNA binding or transporter), ‘specific’, where we have identified a specific function (e.g., transcription factor, ABC transporter) and ‘highly specific’, where a high level of detail is known (e.g., ABC transporter with known substrate; further examples are given in Table S1).

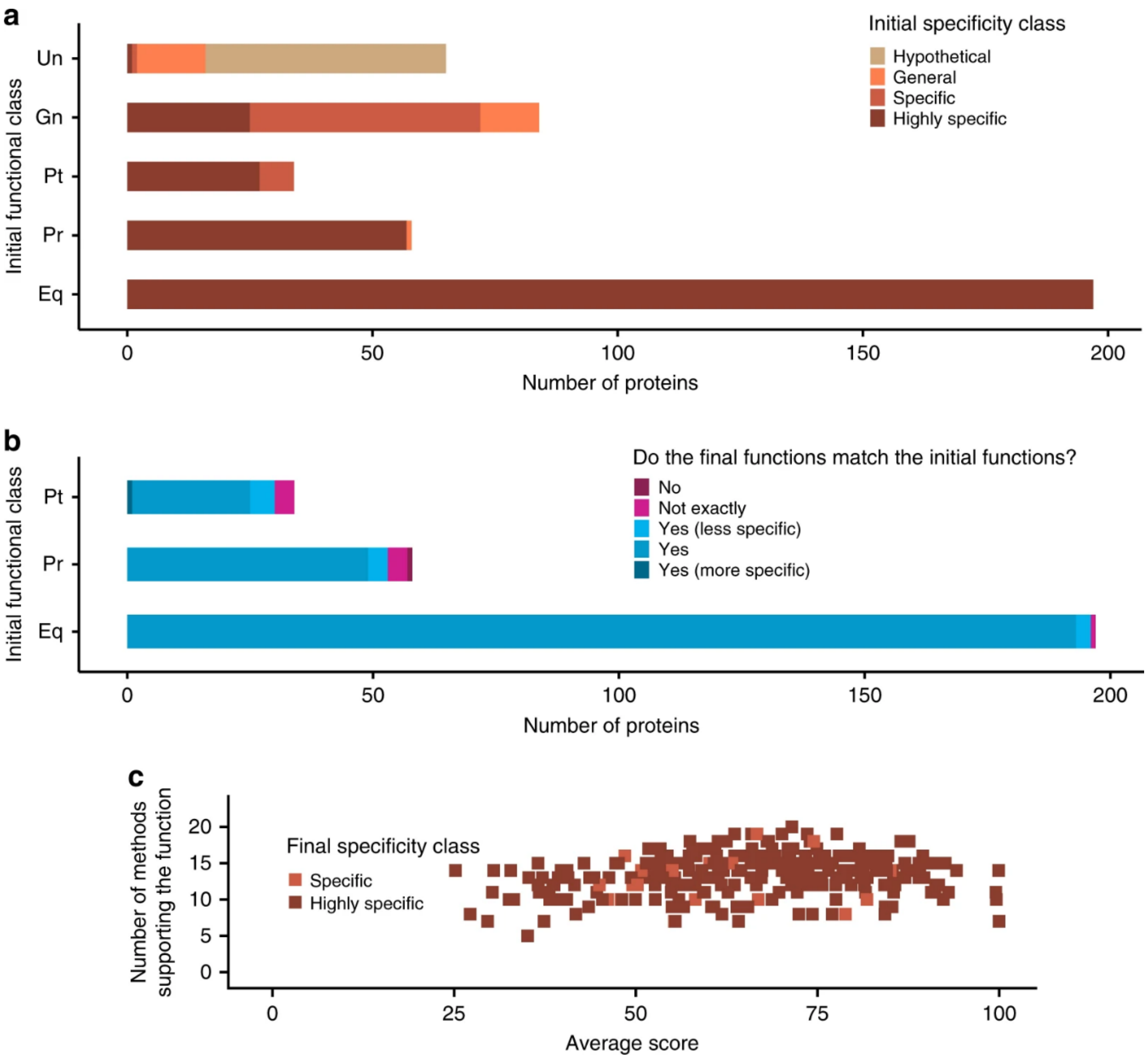


Figure 2.3 *Predictions for proteins of known function encoded by the minimal genome. a Assessment of the specificity of functions predicted by Hutchison et al. across all five initial functional classes (Unknown to Equivalog). Functions from different initial specificity classes are represented by a different colour: beige for the Hypothetical specificity class, orange–General, light brown–Specific and dark brown–Highly specific. b Comparison of our predictions with the functions predicted by Hutchison et al. for proteins of known function, i.e., from the Putative, Probable and Equivalog functional classes. Colouring indicates the level of agreement between the initial functions and the predictions made here. Dark blue where the functions exactly match, medium blue where the predictions made here were less specific than the initial ones, light blue where our predictions were more specific, dark purple where there were minor differences between the functions and light purple where the function did not agree. c Number of methods supporting the function and the average score of those methods. Each point represents a protein. Methods include those used in the first step of function prediction. Specificity class colouring as for a*

We use the number of methods that support a function and the average score associated with this function as indicators of the confidence of the annotation (see methods). The average score for each predicted function was calculated by normalising the scores from the individual methods (e.g., e-value or probability) to the range of 0–100, with 100 indicating a highly confident score (e.g., a highly significant e-value from Pfam or Gene3D; see methods). Further, each protein was assigned to a larger functional category that represents biological process using the 30 different functional categories proposed by Hutchison et al.

Before predicting protein functions, we re-analysed the annotations by Hutchison et al. and assigned the functions to our new specificity classes. Confidence levels of these initial functional annotations could not be compared, since the outputs of the individual methods from the Hutchison et al. study were not available. Our assignment to specificity classes shows that most of the proteins in the Putative, Probable and Equivalog classes had previously been assigned highly specific functions, highlighting how the classification combined both functional specificity and confidence (Figure 2.3a). Further, this analysis suggested that for some of the proteins classed as of unknown function (particularly the Generic class), there had been some suggestion of function, but with very low confidence (Figure 2.3a), i.e., these were long shots based on the results from the three methods used in the Hutchison et al. study. Most of the proteins in the Unknown class were considered to be ‘hypothetical’ according to our criteria (Figure 2.3a).

2.5.6 Benchmarking our approach using proteins of known function

In contrast to Hutchison et al. (Hutchison *et al.*, 2016), who used TIGRFAM, genome context and threading to functionally characterise the proteins encoded by the minimal genome, we applied a wider range of approaches to infer their functions. Many methods have been developed to predict protein function using properties ranging from protein sequence to interaction data and predicting features ranging from subcellular localisation to Gene Ontology (GO) terms and protein structure (Friedberg and Radivojac, 2017). Here, we applied the top performing methods from the recent CAFA (Radivojac *et al.*, 2013; Jiang *et al.*, 2016) assessments, which were available as either a webserver or for download in combination with other established methods to assign functions to the proteins encoded by the minimal bacterial genome (see methods and Figure 2.4). Overall functional inferences were made by manually investigating and combining the predictions and their consistency with genes from the same operon.

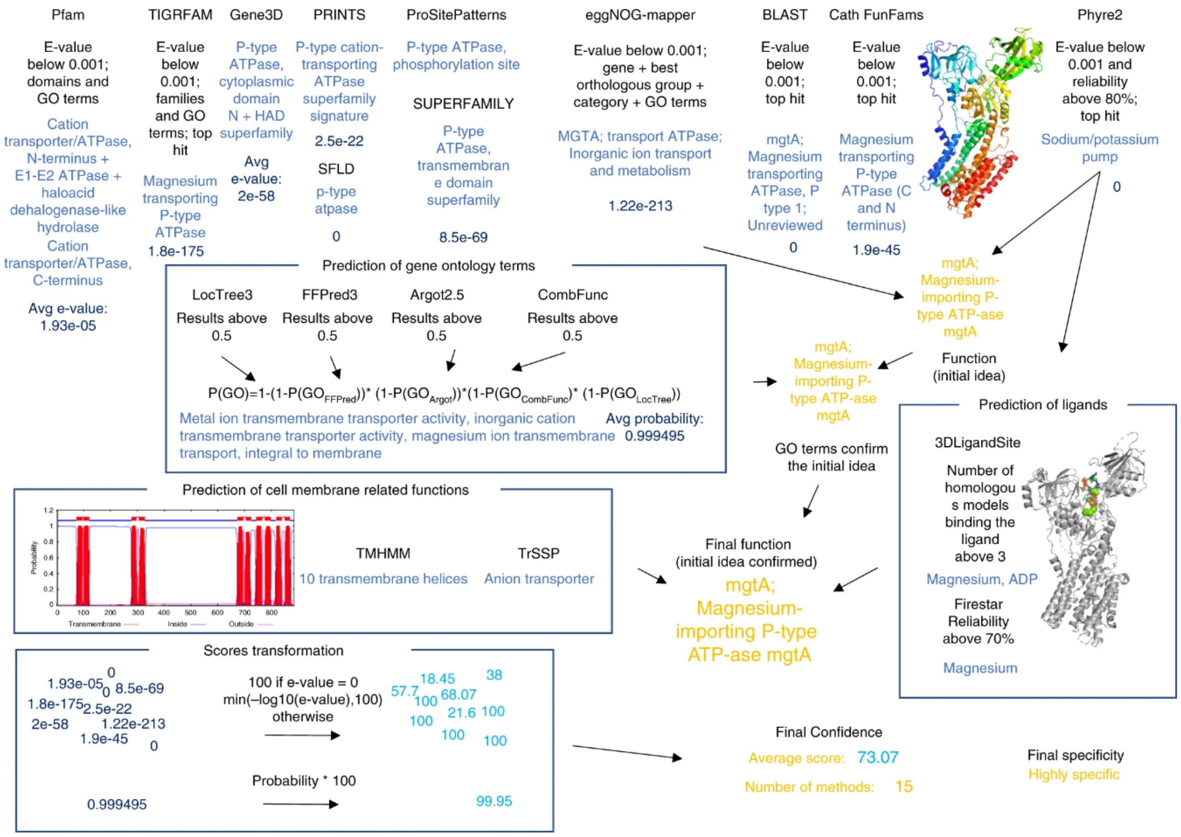

Pfam
E-value below 0.001; domains and GO terms
Cation transporter/ATPase, N-terminus + E1-E2 ATPase + haloacid dehalogenase-like hydrolase
Cation transporter/ATPase, C-terminus
Avg e-value: 1.93e-05
TIGRFAM
E-value below 0.001; families and GO terms; top hit
Magnesium transporting P-type ATPase
1.8e-175
Gene3D
P-type ATPase, cytoplasmic domain N + HAD superfamily
Avg e-value: 2e-58
PRINTS
P-type cation-transporting ATPase superfamily signature
2.5e-22
SFLD
p-type atpase
0
ProSitePatterns
P-type ATPase, phosphorylation site
SUPERFAMILY
P-type ATPase, transmembrane domain superfamily
8.5e-69
eggNOG-mapper
E-value below 0.001; gene + best orthologous group + category + GO terms
MGTA; transport ATPase; Inorganic ion transport and metabolism
1.22e-213
BLAST
E-value below 0.001; top hit
mgtA; Magnesium transporting ATPase, P type 1; Unreviewed
0
Cath FunFams
E-value below 0.001; top hit
Magnesium transporting P-type ATPase (C and N terminus)
1.9e-45
Phyre2
E-value below 0.001 and reliability above 80%; top hit
Sodium/potassium pump
0
mgtA; Magnesium-importing P-type ATP-ase mgtA
Function (initial idea)
mgtA; Magnesium-importing P-type ATP-ase mgtA
Prediction of gene ontology terms
LocTree3
Results above 0.5
FFPred3
Results above 0.5
Argot2.5
Results above 0.5
CombFunc
Results above 0.5
P(GO)=1-(1-P(GO_FFPred))* (1-P(GO_Argot))*(1-P(GO_CombFunc)* (1-P(GO_LocTree))
Metal ion transmembrane transporter activity, inorganic cation transmembrane transporter activity, magnesium ion transmembrane transport, integral to membrane
Avg probability: 0.999495
GO terms confirm the initial idea
Prediction of ligands
3DLigandSite
Number of homologous models binding the ligand above 3
Magnesium, ADP
Firestar
Reliability above 70%
Magnesium
Prediction of cell membrane related functions
Probability
Transmembrane
Inside
Outside
TMHMM
10 transmembrane helices
TrSSP
Anion transporter
Final function (initial idea confirmed)
mgtA; Magnesium-importing P-type ATP-ase mgtA
Scores transformation
1.93e-05
0
0
8.5e-69
1.8e-175
2.5e-22
2e-58
1.22e-213
1.9e-45
0
100 if e-value = 0
min(−log10(e-value),100) otherwise
18.45
38
57.7
100
68.07
100
21.6
100
100
100
0.999495
Probability * 100
99.95
Final Confidence
Average score: 73.07
Number of methods: 15
Final specificity
Highly specific

Figure 2.4 *Assigning function to proteins in the minimal genome. The flowchart outlines how functions were assigned to the proteins using MMSYN1_0879 as an example. The process is described in details in the section Methods / Combined protein function prediction. Briefly, the top row of methods are used to identify a likely function. The methods in the three groups of boxes (predicted GO terms, ligand binding predictions*

and membrane protein predictions) are then used to see if they support the function identified by the first group. Where the first group does not predict a function then this second group was used. The figure shows the results obtained for MMSYN1_0879, which was annotated as the gene mgtA, a magnesium importing P-type ATPase

To test the performance of our approach, we applied it to the proteins of known function belonging to the Hutchison classes Putative, Probable and Equivalog. For 92% (266 of 289) of the proteins, the functions predicted by our approach agreed with the annotation assigned by Hutchison et al. (Figure 2.3b). Our approach has increased the confidence of these annotations, with an average of 13 methods making predictions that supported the functional annotations, compared to a maximum of three methods used in the previous study (Figure 2.3c).

For nine proteins there were minimal differences in the annotations, for example MMSYN1_0637 was previously annotated as the gene rpsI, which encodes the 30S ribosomal protein S9, whereas our predictions suggest it to be rpsN, which encodes the 30S ribosomal protein S5 (Appendix 2 Supplementary Data File 5), which is probably due to them both belonging to the ribosomal protein S5 domain 2-like superfamily. For 12 proteins, our annotations were less specific than the original ones. These proteins were solely in the Hutchison et al. Putative class and the existing annotations were highly specific (Appendix 2 Supplementary Data File 5), such as for MMSYN1_0787, our annotation of RelA/SpoT family protein, is more general that than the original relA gene annotation. For a single protein (MMSYN1_0154) our predicted function of leucyl aminopeptidase was more specific than the initial cytosol aminopeptidase family, catalytic domain protein. Further, only for a single protein (MMSYN1_0908) was our predicted function (yidC; inner membrane protein translocase component) completely different to the existing annotation (misC-polyketide synthase). Overall, this demonstrates that for proteins with known function our approach is able to assign functions that agree with the existing annotations although in some cases, our assignment may be less specific than the existing annotations. We did not assign functions that disagreed with the known function. Further, with many methods now supporting these functions, there is greater confidence in them.

2.5.7 Annotating proteins of previously unknown function

We assigned a function to 133 of the 149 proteins of unknown function. For nearly half of them (66 of 149), new functional information was provided. This included more specific functions (25), assigning a functional category (5) or both of these (26). For the remaining ten proteins, greater functional information was added but the specificity class or functional category remained the same. For example, MMSYN1_0133 was initially annotated as a peptidase of the S8/S53 family, while we proposed a Subtilisin-like 1 serine protease function. While our annotation is more detailed, it is not highly specific and so the protein remained in the Specific class and Proteolysis functional category.

For 51 proteins, a more specific function was assigned (Figure 2.5a and Appendix 2 Supplementary Data File 5). This included 33 proteins initially classified as Hypothetical, ten classified as General and eight as Specific. An example of such protein may be an MMSYN1_0305 protein initially annotated as a metallopeptidase from the family M24 (Specific class), which we predicted to be a Xaa-Pro dipeptidase; pepQ (Highly specific class). For 33 proteins that had initially been annotated as hypothetical, a function was now assigned. Twenty-five of these annotations were classified as General, seven as Specific and one as Highly specific (Figure 2.5a and Appendix 2 Supplementary Data File 5). Eight proteins moved from a General to a Specific function (seven Specific, one Highly specific), and 10 proteins were assigned Highly specific functions having previously been assigned a Specific function (Figure 2.5a). These predictions vary in their level of confidence. Some of them are supported by many methods, while some have highly confident predictions from a smaller number of methods (Figure 2.5b, c).

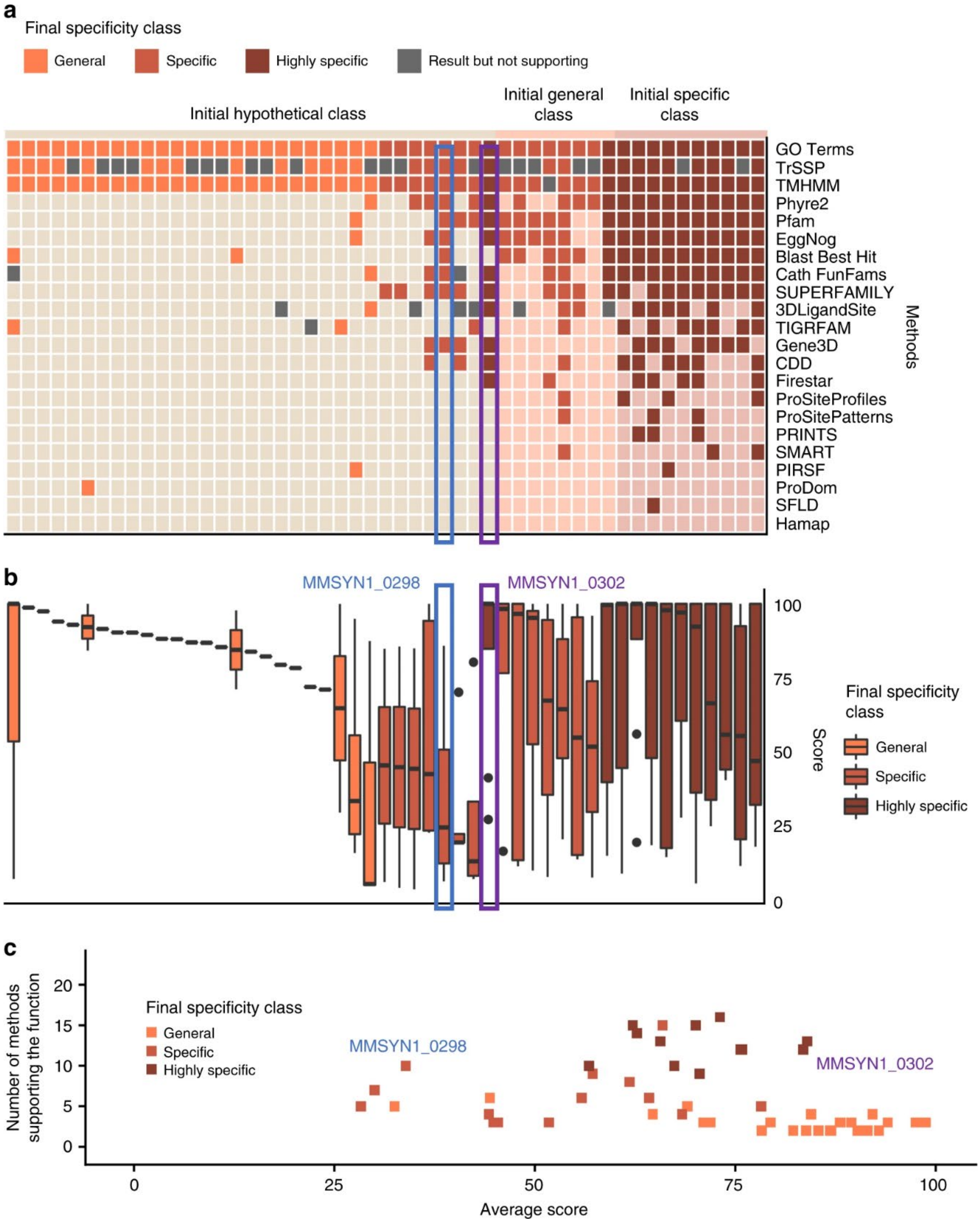


Figure 2.5 *Proteins assigned new functions. This figure shows the 51 proteins where the specificity class was increased. Results for each final specificity class are represented by a different colour: orange for the General specificity class, light brown–Specific and dark brown–Highly specific.* ***a*** *Each column represents a protein in the minimal genome and the squares show the methods that made predictions (darker colours indicate support of the final prediction), grey squares indicate predictions that did not support the function, light squares indicate that a method did not make a prediction. Proteins are grouped by their initial specificity class (Hypothetical, General, Specific and Highly specific) and then by their final specificity class.* ***b*** *Boxplot demonstrating the distribution of the scores across proteins. Proteins grouped by their initial*

specificity class and then by their final specificity class. Horizontal lines represent the median, the lower and upper hinge show respectively first quartile and third quartile, and lower and upper whisker include scores from first quartile to (distance between the first and third quartile) × 1.5 (for lower whisker) and from third quartile to (distance between the first and third quartile) × 1.5 (for upper whisker). Any scores outside of these intervals are shown as points. ***c*** *The number of methods supporting the function and the average score. Each point represents a protein*

For most proteins that were assigned a general function, we see that they were often supported by fewer methods but those methods predicted them with high confidence scores (Figure 2.5a). For example, the group of proteins in the bottom right corner of Figure 2.5c were all predicted to be transporters but only assigned a general function as further details such as substrate specificity could not be inferred. Where Specific and Highly specific functions were assigned, typically more methods supported the function but there was a greater range in the scores associated from the individual methods (Figure 2.5). For example, MMSYN1_0298 and MMSYN1_0302 were both initially classed as hypothetical and we have assigned them Specific and Highly specific functions respectively, based on data available from 10 (MMSYN1_0298) and 12 (MMSYN1_0302) methods (Figure 2.5 and Appendix 2 Supplementary Data Files 1–6, 8). Based on these data sources we propose that MMSYN1_0298 is a ribosomal protein from the family L7AE/L30e (Figure 2.6a) and that MMSYN1_0302 is an oxygen-insensitive NAD(P)H nitroreductase (Figure 2.6b), both of which are functions widespread across the kingdoms of life.

Our analysis suggests that the combination of methods improves the reliability of function annotation. For some proteins, there appeared to be evidence for a given function from multiple sources, but on closer inspection it was difficult to assign more confident annotations (Figure S3). For example, MMSYN1_0138 is homologous to the ATP-binding region of ABC transporters but the ATP-binding site is not conserved, which casts some doubt on this function (Figure S3A). For MMSYN1_0615, matches from four methods suggest a Phenylalanine-tRNA ligase function (Figure S3B). However, MMSYN1_0615 only contains 202 residues and the beta chain of bacterial Phenylalanine-tRNA ligases contain nearly 800 residues, making it unlikely that MMSYN1_0615 performs this function (Figure S3B).

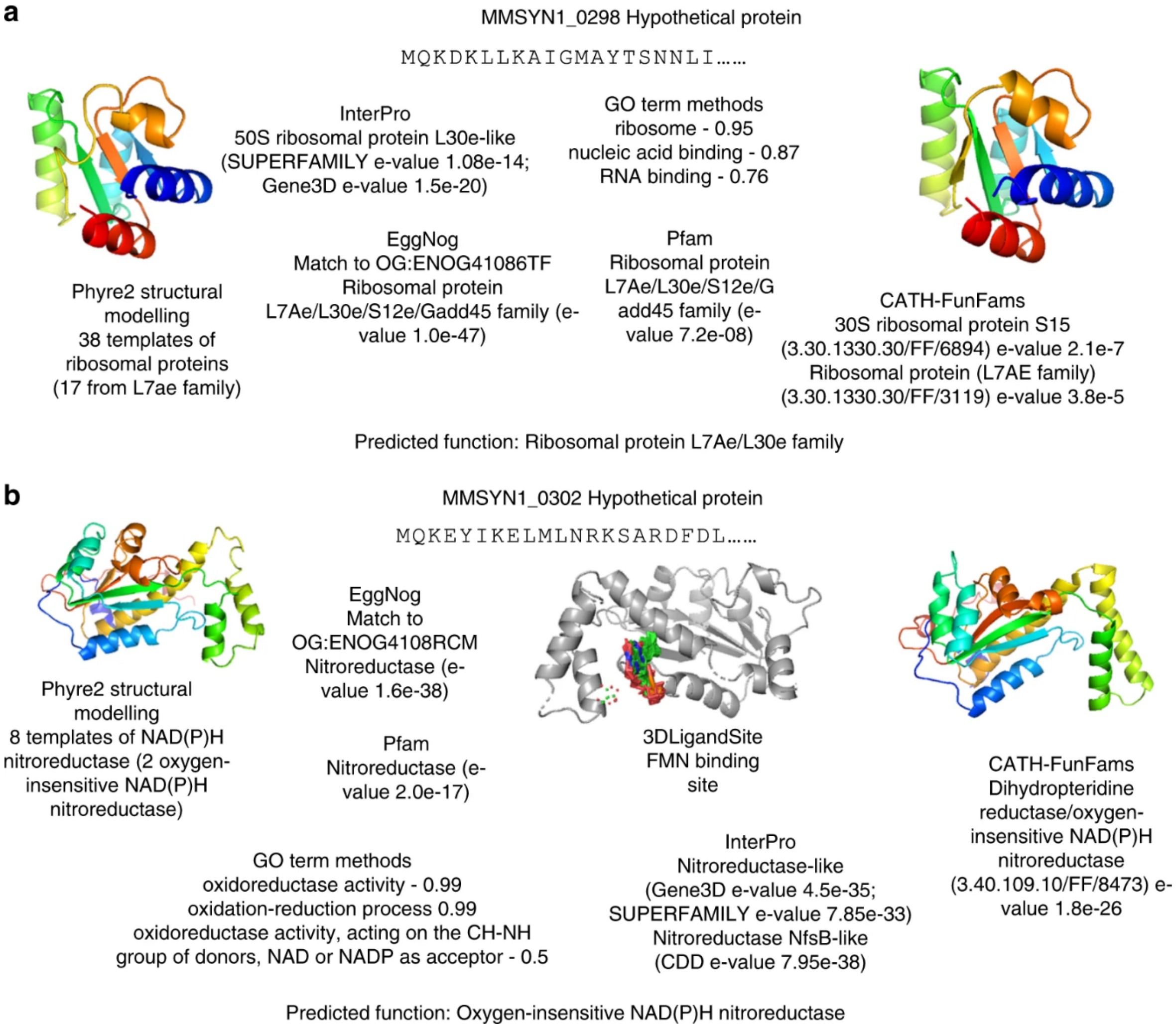

a
MMSYN1_0298 Hypothetical protein
MQKDKLLKAIGMAYTSNNLI……
InterPro
50S ribosomal protein L30e-like
(SUPERFAMILY e-value 1.08e-14;
Gene3D e-value 1.5e-20)
GO term methods
ribosome - 0.95
nucleic acid binding - 0.87
RNA binding - 0.76
EggNog
Match to OG:ENOG41086TF
Ribosomal protein
L7Ae/L30e/S12e/Gadd45 family (e-value 1.0e-47)
Pfam
Ribosomal protein
L7Ae/L30e/S12e/Gadd45 family (e-value 7.2e-08)
Phyre2 structural
modelling
38 templates of
ribosomal proteins
(17 from L7ae family)
CATH-FunFams
30S ribosomal protein S15
(3.30.1330.30/FF/6894) e-value 2.1e-7
Ribosomal protein (L7AE family)
(3.30.1330.30/FF/3119) e-value 3.8e-5
Predicted function: Ribosomal protein L7Ae/L30e family
b
MMSYN1_0302 Hypothetical protein
MQKEYIKELMLNRKSARDFDL……
EggNog
Match to
OG:ENOG4108RCM
Nitroreductase (e-value 1.6e-38)
Phyre2 structural
modelling
8 templates of NAD(P)H
nitroreductase (2 oxygen-insensitive NAD(P)H
nitroreductase)
Pfam
Nitroreductase (e-value 2.0e-17)
3DLigandSite
FMN binding
site
CATH-FunFams
Dihydropteridine
reductase/oxygen-insensitive NAD(P)H
nitroreductase
(3.40.109.10/FF/8473) e-value 1.8e-26
GO term methods
oxidoreductase activity - 0.99
oxidation-reduction process 0.99
oxidoreductase activity, acting on the CH-NH
group of donors, NAD or NADP as acceptor - 0.5
InterPro
Nitroreductase-like
(Gene3D e-value 4.5e-35;
SUPERFAMILY e-value 7.85e-33)
Nitroreductase NfsB-like
(CDD e-value 7.95e-38)
Predicted function: Oxygen-insensitive NAD(P)H nitroreductase

***Figure 2.6** Confident predictions of protein function in the minimal genome. Both **a** MMSYN1_0298 and **b** MMSYN1_0302 were previously classified as hypothetical proteins. The results from prediction methods and the function assigned are shown*

Overall, we found that the diversity of different methods used was required for inferring function, with no individual method able to predict the most detailed function assigned to more than one-third of the proteins of unknown function (Table S2). The top five methods to assign the most detailed functions each used different approaches, including a method that identifies orthologous groups (eggNOG-Mapper (Huerta-Cepas *et al.*, 2017)), the group of methods that predict GO terms (FFPred3 (Cozzetto *et al.*, 2016), Argot2.5 (Falda *et al.*, 2012), GOAT (see Appendix 7) and LocTree3 (Goldberg *et al.*, 2014)), a method that predicts protein three-dimensional structure (Phyre2 (Kelley *et al.*, 2015)), identification of protein domains from Pfam and finally the best BLAST match from UniProt. Further, any combination of the top five performing methods only obtained the final annotation for a maximum of 25% of the proteins, further highlighting the contribution of multiple different methods to assign functions (Table S3). Two methods (GO terms and TMHMM) were able to widely provide more generic functions supporting the overall assigned function (54% for GO terms and 82% for TMHMM), although TMHMM only predicts if the protein contains transmembrane helices (Table S2).

For the remaining 83 proteins, our predictions supported the existing annotation. Importantly for many of these proteins, multiple methods have now made predictions that support the annotation, thus increasing their confidence. Figure 2.7 shows that many of the proteins (28 out of 83) have predicted functions that are supported by 10 or more methods, rising to 61 supported by 5 or more methods, often with high confidence scores (or e-values) from the individual methods.

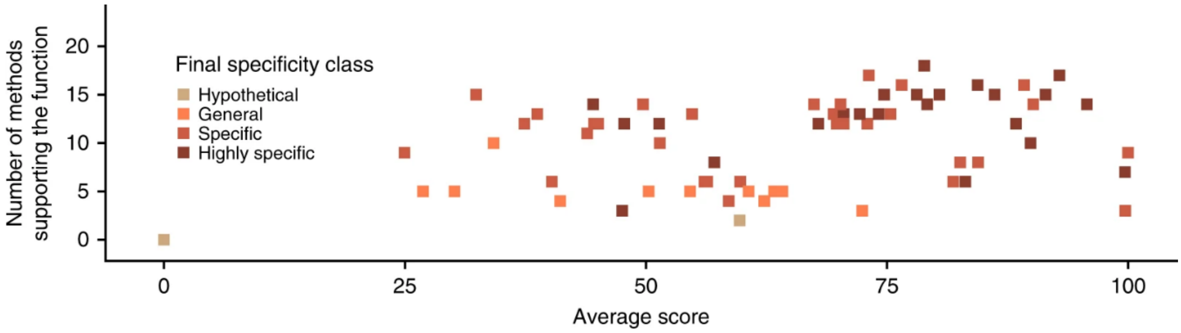


__Figure 2.7__ Multiple methods supporting existing annotations. For all proteins where the predicted function agreed with the existing annotation (i.e., the specificity class was not changed), the number of methods that predicted the function is plotted against the average score from these methods. Points for each of the final specificity classes are represented by a different colour: beige for the Hypothetical specificity class, orange–General, light brown–Specific and dark brown–Highly specific

2.5.8 Understanding biological processes in the minimal genome

Functional categories were assigned to 31 proteins that had previously been classified with Unclear biological process. The majority of the proteins with a newly assigned functional category were predicted to have transporter functions, with 24 proteins added to the 84 already assigned to this functional category (Figure 2.8a). Further, one protein (MMSYN1_0033) was assigned to the cytosolic metabolism category, three to the preservation of genetic information category (MMSYN1_0005, MMSYN1_0239, MMSYN1_0353), and three to the expression of genetic information category (MMSYN1_0615, MMSYN1_0730, MMSYN1_0873) (Figure 2.8a).

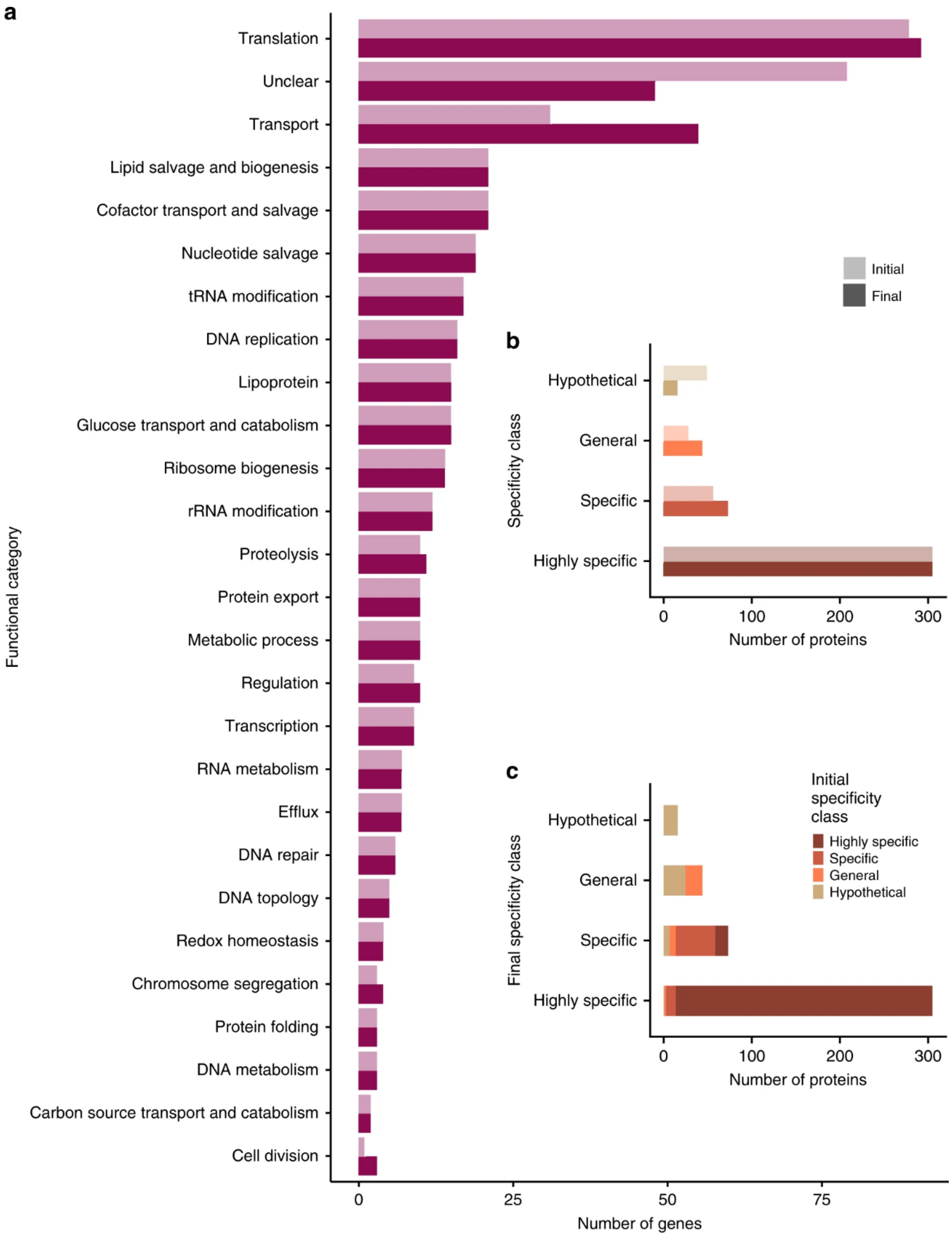


Figure 2.8 *Functional annotations of the minimal bacterial genome. The number of proteins in each of the **a** protein biological process categories (light and dark purple indicate initial and final categories, respectively). **b** Specificity classes is shown with the original minimal genome annotation and the annotations identified here. **c** Shows the change in specificity classes, coloured based on the original specificity class. Results for each initial specificity class are represented by a different colour: beige for the Hypothetical specificity class, orange–General, light brown–Specific and dark brown–Highly specific*

Overall, while functional annotations have been inferred for a considerable proportion of the proteins of unknown function, the biological process for 48 proteins remains unknown (i.e., in the Unclear category; Figure 2.8a). For 32 of these proteins, a molecular function was assigned such as Cof-like hydrolase, ATPase AAA family, or DNA-binding protein HU, but there was insufficient information to assign a functional category. The remaining sixteen proteins lack functional information and are classified as hypothetical. These proteins do not contain any known domains or transmembrane helices, none have orthologues in other kingdoms of life and only a few within bacteria. Either these are species-specific proteins that perform an important function within Mycobacteria or they have diverged significantly such that sequence relationships are not detected.

2.5.9 Newly assigned functions indicate transporters

Transmembrane helices were identified in 41% (61) of the proteins of unknown function (Figure 2.2 and Appendix 2 Supplementary Data File 4). Fifteen transmembrane proteins, which were not categorised as transporters, were annotated with functions in cell division (1), chromosome segregation (1) and proteolysis (4), while the biological process remained unknown for nine. Our analysis suggests that 46 of the 61 predicted transmembrane proteins are likely to be responsible for membrane transport (Appendix 2 Supplementary Data File 4, Figure S6). Of the 46, 23 were previously annotated by Hutchison et al. with a range of transporter functions (e.g., ABC transporters, S component of ECF transporters), all of which were further supported by our analysis. A further 15 proteins that lack transmembrane domains were also associated with transport functions, e.g., ATP-binding units of ABC transporters, 14 of them were identified by Hutchison et al. (Hutchison *et al.*, 2016).

Of the 24 newly proposed transporters (previously hypothetical or with minimal information, e.g., membrane protein), six gained specific transporter functions. All six were previously classed as membrane proteins and have now been annotated as transporters; one hexose phosphate transport protein (MMSYN1_0881), one ABC transporter (MMSYN1_0411), one S component of an ECF transporter (MMSYN1_0877), and three belonging to the Major facilitator superfamily (MMSYN1_0235, MMSYN1_0325, MMSYN1_0478) (Appendix 2 Supplementary Data Files 1–6, 8). The remaining 18 proteins annotated as transporters (general specificity level) had previously either been annotated as membrane or hypothetical proteins. Results from a few methods (with high scores–Figure 2.9) indicate that they are transporters but it was not possible to assign them to a specific family/type of transporter or to identify a substrate.

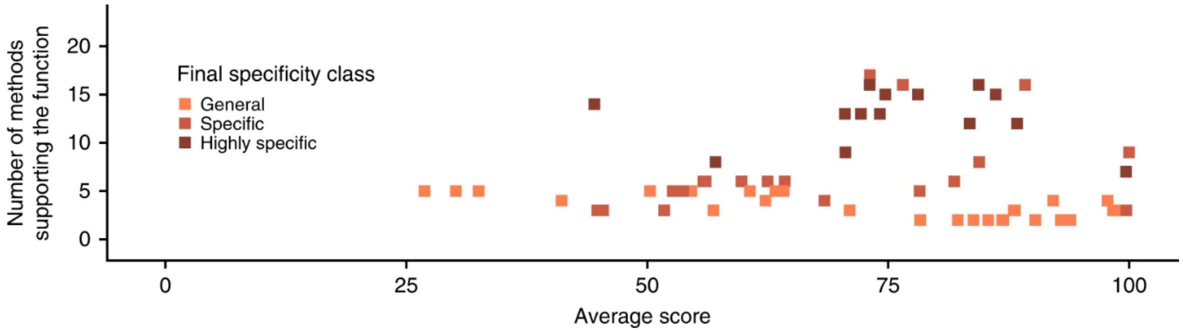


Figure 2.9 *Prediction of membrane related functions. Each point represents a protein with initially unknown function for which we assigned cell membrane related functions (e.g., transmembrane, transporter). The number of methods that supported the prediction is plotted against the average score from these methods. Points for each of the final specificity classes are represented by a different colour: orange for the General specificity class, light brown–Specific and dark brown–Highly specific*

More specific annotations could be made for proteins already annotated with transport-related functions, including four proteins (MMSYN1_0034, MMSYN1_0399, MMSYN1_0531, MMSYN1_0639) that were classed as FtsX-like permeases having previously been given generic transport-related annotations (e.g., permease). For most of these proteins, we have greater confidence in the assigned function, given that many different methods support them (Figure 2.9). This extends their initial annotations that had been assigned by only three methods. For example, one operon encodes proteins that transport oligopeptides (AmiABCDE MMSYN1_0165 - MMSYN1_0169) and another operon encodes a spermidine/putrescine transporter (PotABCD MMSYN1_0195 - MMSYN1_0197) (Appendix 2 Supplementary Data File 5, Figure S4 and Figure S5).

One of the three proteins newly proposed to be members of the Major facilitator superfamily, MMSYN1_0325, was previously classified as a membrane protein (Figure 2.10). In agreement, the transmembrane helix prediction tool TMHMM (Krogh *et al.*, 2001) predicted 13 transmembrane helices in the protein. Further, the structure was confidently modelled by Phyre2, with > 98% confidence for 26 independent structural templates, all of which had transporter functions (including members of the MFS superfamily). InterPro (Mitchell *et al.*, 2019) assigned it into the MFS transporter superfamily. Supporting this function, further methods predicted a range of transporter-related functions, including symporter activity (GO:0015293) and substrate-specific transmembrane transporter activity (GO:0022891) with probabilities >90% (Figure 2.10 and Appendix 2 Supplementary Data File 6).

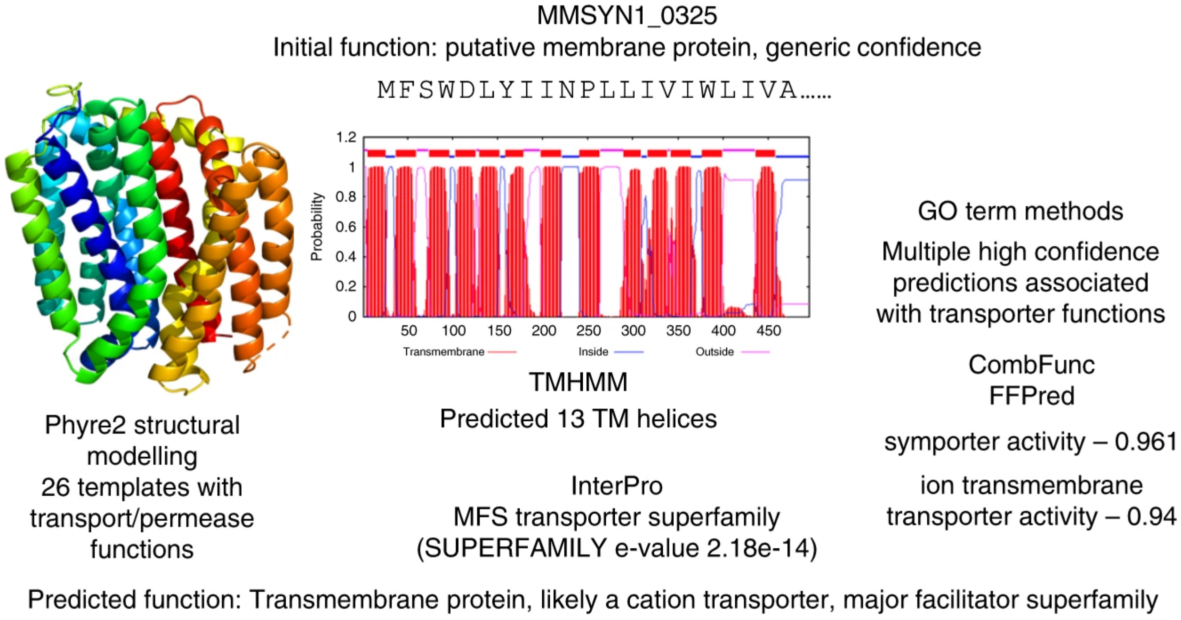

MMSYN1_0325
Initial function: putative membrane protein, generic confidence
MFSWDLYIINPLLIVIWLIVA……
1.2
1
0.8
0.6
0.4
0.2
0
Probability
50
100
150
200
250
300
350
400
450
Transmembrane
Inside
Outside
Phyre2 structural modelling
26 templates with transport/permease functions
TMHMM
Predicted 13 TM helices
InterPro
MFS transporter superfamily
(SUPERFAMILY e-value 2.18e-14)
GO term methods
Multiple high confidence predictions associated with transporter functions
CombFunc
FFPred
symporter activity – 0.961
ion transmembrane transporter activity – 0.94
Predicted function: Transmembrane protein, likely a cation transporter, major facilitator superfamily

Figure 2.10 *MMSYN1_0325 is predicted to be a transporter and member of the Major facilitator Superfamily. The results from Phyre2, TMHMM, the combination of GO term prediction methods (numbers shown are probability associated with each function) and InterPro are shown. All of these methods supported a transporter function with Phyre2 and InterPro confidently identifying association with the Major facilitator superfamily*

2.5.10 Comparison of predictions made by Danchin and Fang

Recently Danchin and Fang (Danchin and Fang, 2016) used what they referred to as an engineering-based approach to investigate the unknown functions within the minimal bacterial genome and provided annotations for 71 of the 149 proteins of unknown function. They set out to identify functions that would be expected to be in a minimal genome but were missing from the existing annotation and to then identify proteins that could perform these functions (although it is not clear how these candidates were identified as no methods were provided (Danchin and Fang, 2016)).

Comparison of the results from both studies revealed considerable overlaps (Appendix 2 Supplementary Data File 7). Using our approach, only sixteen proteins remained hypothetical without any assigned function, while Danchin and Fang did not provide any annotations for 78 of the proteins with unknown function. Thus, we leave only 10% of the previously unannotated proteins without any assigned function, while 52% remain completely uncharacterised by Danchin and Fang. This demonstrates the breadth of function that our approach is able to assign. The predictions showed complete agreement for 36 proteins and minor differences for 18 proteins (Appendix 2 Supplementary Data File 7). For a further 13 proteins the predictions were more detailed in one study than the other (Appendix 2 Supplementary Data File 7). For example, Danchin and Fang proposed that MMSYN1_0822, is an S component of an ECF transporter and is part of a folate transporter, whereas we identified three possible folate transporters (MMSYN1_0314, MMSYN1_0822, MMSYN1_0836) and could not confidently assign substrates to any of them.

Four of the predictions differed considerably (Appendix 2 Supplementary Data File 7). They are represented by proteins such as MMSYN1_0388 which here was annotated as a transmembrane protein, possibly a cation transporter, while Danchin and Fang suggested that it has a role in double-strand break repair. For three of the proteins, Danchin and Fang inferred more functional characteristics. They annotated MMSYN1_0853 MMSYN1_0530, MMSYN1_0511 with the functions energy-sensing regulator of translation, promiscuous phosphatase and double-strand break repair protein, respectively, while here they were retained as hypothetical since there was little agreement between the multiple methods used to be able to infer protein function.

2.6 Discussion

The genome size of *Mycoplasma mycoides* used to create the minimal cell is 1079 kilobase pairs (kbp). Surprisingly, it is not the smallest genome across all bacteria (Hutchison *et al.*, 2016). That title belongs to the 160 kbp-genome of *Carsonella ruddii* – an endosymbiont living in phloem sap-feeding insects (Nakabachi *et al.*, 2006). However, its genome lacks many genes essential for bacterial life, including those responsible for the biogenesis of cell envelope or metabolism of lipids and nucleotides. Because of that, it cannot fulfil the main condition for creating the minimal synthetic cell: it must be capable of autonomous growth, and as a result, mycoplasmas were selected for this task as they are the simplest organisms that can grow autonomously (Hutchison *et al.*, 2016). The smallest known genome of all mycoplasmas belongs to *Mycoplasma genitaliu*m (580 kbp). However, Hutchison et al. decided to minimise the genome of *Mycoplasma mycoides* because of its faster growth.

It is important to know that *Mycoplasma mycoides* is a parasitic bacterium - if aside from the independent growth, a "free-living" condition was to be added to the selection criteria of the smallest bacteria, a marine bacterium *Pelagibacter ubique* would be chosen as the model organism. This is because it is the smallest known self-replicating and free-living bacteria – it has all the pathways necessary to synthesise all necessary amino acids and almost all cofactors with a genome size of no more than 1309 kbp (Giovannoni *et al.*, 2005).

The synthesis of the bacterium with the smallest genome (to date), resulted in an astounding number (149 of 473) of proteins of unknown function and emphasised the gaps in our understanding of the basic principles of life. Our results demonstrate that the combined use of a range of complementary advanced methods for protein function inference is superior to the use of individual approaches. Using a combination of results from 22 different methods, we were able to assign new functional information to 66 of the 149 proteins that were originally classed as having unknown function. Further, given the use of many different methods, we have increased the confidence in existing annotations that our approach also supported. For some proteins, more detailed functions were predicted by some of the methods. However, in manually combining the predictions, there was insufficient evidence to assign them to more specific functional classes. Nevertheless, these functions should be sufficient to direct further research and experimental characterisation. Our analysis shows that the combination of many methods was essential with no single method able to identify the highest detailed function assigned to more than one-third of the proteins (Table S2).

Most of the proteins of unknown function were homologous to few other proteins with known functions and they also lacked orthologues. Thus, for many of the proteins where functions have been assigned, methods that are not dependent on homology were prevalent (e.g., FFPred3 (Cozzetto *et al.*, 2016), Figure 2.5 and Figure S6). This highlights the importance of developing further methods that do not rely on homology. Moreover, many of the difficult to characterise proteins do not contain known protein domains and are enriched for transmembrane proteins (Figure 2.1). Hence, additional approaches to predict the function of such proteins are required.

With our expanded functional assignments, 50% of the proteins encoded by the minimal genome perform functions associated with two fundamental life processes; preserving and expressing genetic information (Figure 2.8a). Most notably, many proteins were assigned transporter functions, and these proteins now represent 22% of the minimal genome. In generating the minimal genome, 32 *M. mycoide*s genes with membrane transport functions were removed (Hutchison *et al.*, 2016). Additionally, many proteins with metabolic functions were removed. Hence, the minimal genome bacterium is reliant on obtaining many nutrients from the medium and also needs to remove (toxic) metabolites from the cell. Thus, it may not be surprising that transporters are essential for the bacterium. It was not possible to assign substrates for these transporters. The reason for this may be at least in part due to the promiscuity of mycoplasmal transport systems (S Razin, 1998). Additionally, transporters may transport low-affinity substrates in a nutrient-rich environment in which nutrients are highly abundant.

The identification of many transporters also highlights the dependence of the minimal bacterial genome cells on the medium in which they grow. Hence, we postulate that there is no such thing as a generic minimal genome. Instead, the genes that shape the minimal genome partly depend on its environment. Consequently, we propose that a minimal genome consists of two sets of genes (Figure 2.11). The first set encodes functions that are an essential prerequisite for all bacteria and probably all forms of life, which on its own is not sufficient to enable life. This gene set needs to be complemented with an additional set of genes that enables life in a particular environment. In a nutrient-rich environment, these additional genes may largely have functions associated with compound uptake and efflux in agreement with our current results presented here (Figure 2.12). Under other circumstances, where nutrients are not so abundant, metabolic functions are likely to be of greater importance.

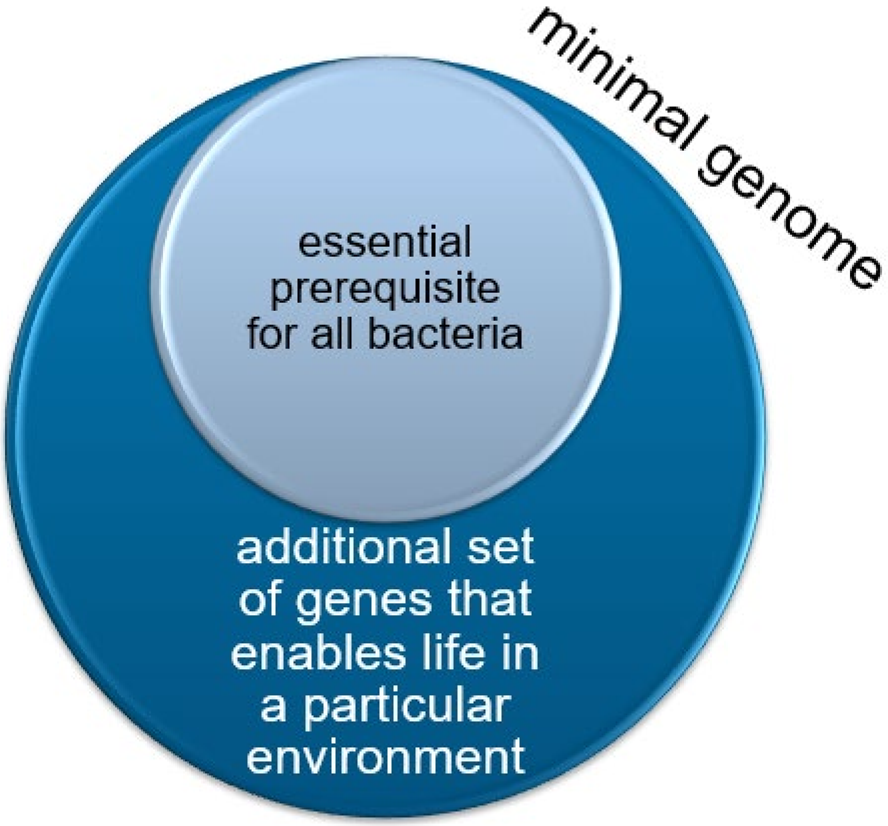

minimal genome
essential prerequisite for all bacteria
additional set of genes that enables life in a particular environment

Figure 2.11 Two sets of genes that constitute a minimal bacterial genome. *They represent genes that are essential for all bacteria and genes that enable life in a particular environment.*

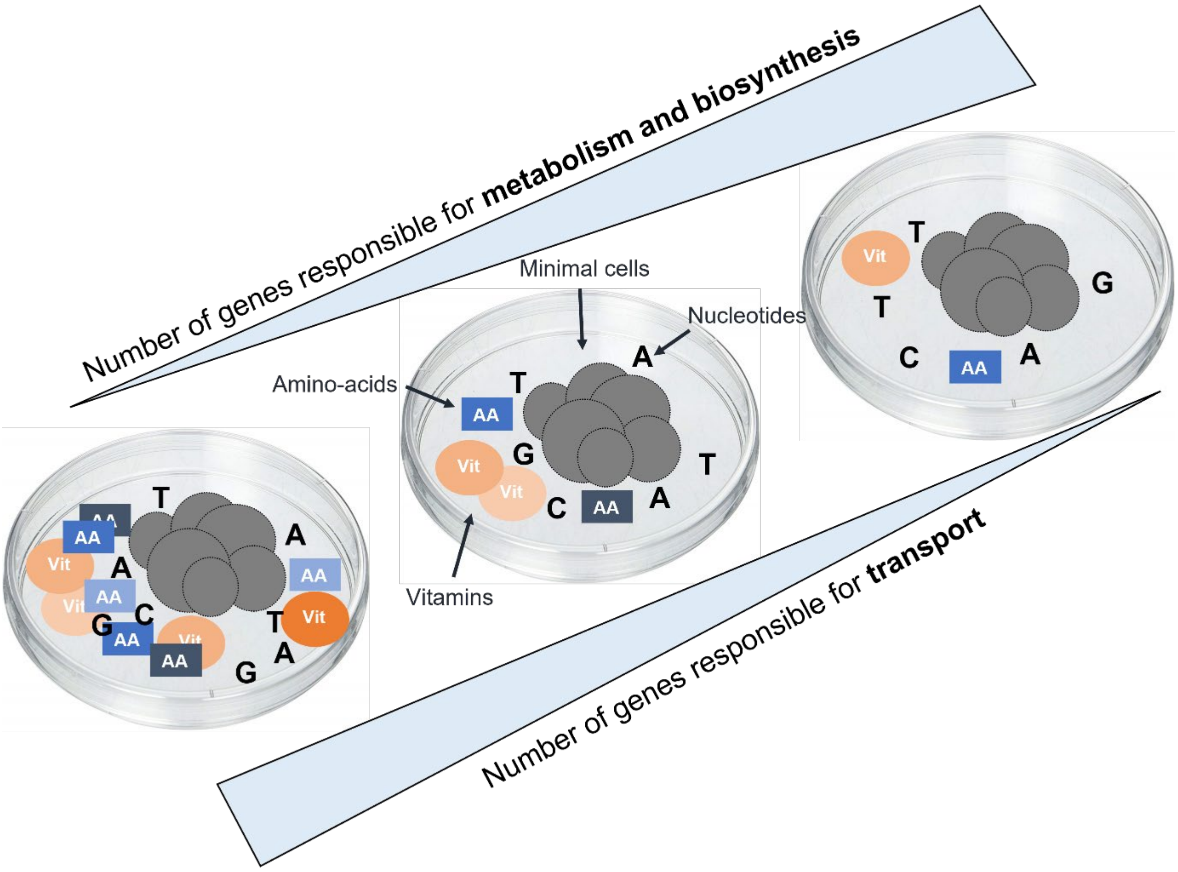

Number of genes responsible for metabolism and biosynthesis
Minimal cells
Nucleotides
Amino-acids
Vitamins
AA
Vit
T
A
G
C
Number of genes responsible for transport

Figure 2.12 Change in the number of genes responsible for metabolism and biosynthesis versus the number of genes responsible for transport depending on the richness of nutrients in the environment. *The Petri dish represents the environment in which minimal cells are grown. Grey circles show minimal cells, and the examples of nutrients include amino acids (AA), nucleotides (A, C, G, T) and vitamins (Vit).*

In summary, we have successfully applied a combined bioinformatics approach to characterise proteins with unknown function from the minimal genome that had not been annotated by previous approaches. Currently, only about 1% of all known proteins are annotated with experimentally confirmed functions. Since the experimental analysis of protein function will for the foreseeable future remain restricted to a small subset of proteins due to physical and financial limitations, optimised bioinformatics approaches will be critical for the assignment of functions to proteins and, in turn, our understanding of the essential functions of life. Proteins that are difficult to classify typically (i) do not contain known protein domains (ii) lack homology to proteins with known structure and (iii) are enriched for transmembrane proteins. Further, most of the hypothetical proteins appear to be bacteria- and clade-specific. Hence, further complementary approaches are needed that enable the assignment of functions to such proteins. Importantly, a considerable proportion of the newly annotated proteins probably have transporter functions. These transporters are likely to be involved in the uptake of nutrients and efflux of waste products in a minimal genome organism that lacks many metabolic enzymes and is cultivated in a nutrient-rich environment. Additionally, our findings indicate the existence of a core set of genes that is essential for all forms of life but not sufficient to enable life on its own. This essential gene set needs to be complemented by a second enabling gene set that facilitates life under particular environmental conditions. Thus, the concept of a minimal genome is context/environment specific.

2.6.1 The current state-of-the-art in protein structure prediction

In this work, we used Phyre2 (Kelley *et al.*, 2015) in the normal mode to predict structures of the minimal genome proteins. This method first builds a hidden Markov model (HMM) for the target protein sequence homologs detected by PSI-BLAST (Altschul *et al.*, 1997) and secondary structure is predicted by PSI-PRED (Jones, 1999). The HMM is then scanned against a library of HMMs of known structures, and a query-template alignment is generated to model the backbone of the 3D structure. Finally, the procedure ends with modelling of INDELs and side chains. In normal mode, Phyre2 creates a model of the 3D structure based on a single template.

Single-template structure prediction methods are highly reliable when a template has 50-60% identity compared to the target sequence (Buenavista, Roche and McGuffin, 2012). However, in case of a lower level of sequence–template similarity, other methods are recommended to increase the quality and coverage of predicted models.

IntFOLD-TS builds a 3D model of the protein structure based on multiple templates (Mcguffin *et al.*, 2019). The model is created iteratively using fourteen single-template and eight threading methods. The multiple target-template alignments are then scored using the ModFOLD method (McGuffin *et al.*, 2021) to minimise local errors. IntFOLD-TS has been continuously evaluated in the Critical Assessment of Structure Prediction since the 9th edition, and its performance has been competitive (for example, in CASP14 IntFOLD6 was ranked as 91 out 146 in ‘Regular Targets’ category and 99 out of 136 in ‘Inter-domain prediction’) (Kryshtafovych, Fidelis and Moult, 2011; Mcguffin *et al.*, 2019; *Home - CASP14*, 2021).

However, the first and second place in the latest CASP edition belongs to AlphaFold2 (Jumper *et al.*, 2021) and trRosetta-based methods (Yang *et al.*, 2020), respectively, among the methods that do not have to be fully automated (‘Human group’). In addition, while I-TASSER (Yang *et al.*, 2015) and C-QUARK (Mortuza *et al.*, 2021) were ranked as the best fully automated methods (‘Server’ group), it has to be noted that AlphaFold2 achieved far better results than any other method (Pearce and Zhang, 2021b).

In I-TASSER, structural templates are identified using LOMETS (Zheng *et al.*, 2019) – a multiple threading method (Yang *et al.*, 2015). The sequence is then divided into threading-aligned and threading-unaligned regions. Finally, Monte Carlo simulations (Thachuk, Shmygelska and Hoos, 2007) are applied to perform iterative template-based fragments assembly and build the full-length model.

Similarly to I-TASSER, C-QUARK and Rosetta are based on the fragment assembly, and they perform Monte Carlo simulations to construct the full-length model (Rohl *et al.*, 2004; *C-QUARK: Contact Assisted Ab Initio Protein Structure Prediction*, 2021; Mortuza *et al.*, 2021). However, in contrast to I-TASSER, C-QUARK and Rosetta are de novo protein structure prediction methods. In addition, C-QUARK uses structure fragments collected from unrelated PDB structures, and simulations are conducted under the direction of a complex force field consisting of knowledge-based energy terms, inter-residue contacts and contact-map predictions.

Finally, trRosetta and AlphaFold are deep learning-based methods (Yang *et al.*, 2020; Jumper *et al.*, 2021). trRosetta applies deep learning to predict the interresidue distances and orientations, which are then converted into smooth restraints used in Rosetta (Rohl *et al.*, 2004) to build the full-length 3D structure model under the guidance of energy minimisation. AlphaFold also incorporates novel neural networks to predict the distribution of interresidue distances and angles between the bonds that the residues form (Senior *et al.*, 2019; Jumper *et al.*, 2021). The first iteration of AlphaFold used fragment assembly to model the full protein structure. However, AlphaFold2 applies a gradient descent-based folding technique which allows for faster predictions. Created by Google's DeepMind, AlphaFold was evaluated for the first time in CASP13 in 2018, where it showed unprecedented accuracy in protein structure prediction (Pearce and Zhang, 2021a).

In addition to using novel algorithms of protein structure prediction capable of modelling structures with high accuracy and coverage, it is also recommended to use tools that assess the quality of predicted models. Examples of such tools include ProQ3 (Uziela *et al.*, 2016) and ModFold8 (McGuffin *et al.*, 2021). ProQ3 is a Support Vector Machine-based method that uses energy terms from Rosetta as features (Uziela *et al.*, 2016). Meanwhile, ModFold8 uses neural networks to combine scores from thirteen different methods (Uziela *et al.*, 2016).

Our usage of a single-template (normal) mode of Phyre2 to model 3D structures of the minimal genome proteins was justified at the time. Using this technique, we predicted confident structures with good coverage for most of the proteins from Equivalog, Probable, Putative and Generic groups. However, we found confident templates (at least 90%) that covered at least 75% of the proteins only for nine proteins from the Unknown functional class, which lacked orthologs in other species. Using newer tools such as deep learning-based methods trRosetta or AlphaFold2, complemented by methods estimating the quality of the entire model, could improve the confidence and coverage of predicted structures (Uziela *et al.*, 2016; Yang *et al.*, 2020; Jumper *et al.*, 2021; McGuffin *et al.*, 2021). For example, DMPfold (a deep learning-based method for protein structure prediction) has recently been applied to the minimal bacterial genome and has been able to obtain close to complete structural coverage (Greener *et al.*, 2020).

Our method of protein function prediction was not successful for sixteen proteins of the minimal genome. For those proteins, we could not conclude any functional information, and they remained annotated as hypothetical. Knowing the 3D structure of those proteins could help identifying their function. We could model protein's interactions with other proteins or ligands by applying molecular docking (Northey, Barešić and Martin, 2018; Salmaso and Moro, 2018) and then apply molecular dynamics simulations to examine the nature of those interactions (Hospital *et al.*, 2015).

Chapter 3 Acquired MDM2 inhibitor resistance is associated with collateral sensitivity to cytarabine in acute myeloid leukaemia cells

Magdalena Antczak[1#], Tamara Rothenburger[2#], Helen E. Grimsley[1], Miguel Julia[1], Florian Rothweiler[2], Constanze Schneider[2], Björn Rotter[3], Daniel Speidel[4,5], Andrea Nist[6], Marco Mernberger[7], Dominique Thomas[8], Gerd Geisslinger[8,9], Thorsten Stiewe[6,7], Mark N. Wass[1]*, Martin Michaelis[1]*, Jindrich Cinatl jr.[2]*

[1] School of Biosciences, University of Kent, Canterbury, UK

[2] Institute for Medical Virology, Goethe-University, Frankfurt am Main, Germany

[3] GenXPro GmbH, Altenhöferallee 3, 60438 Frankfurt am Main, Germany

[4] Children's Medical Research Institute, Westmead, Australia

[5] Sydney Medical School, The University of Sydney, Australia

[6] Genomics Core Facility, Philipps-University, Marburg, Germany

[7] Institute of Molecular Oncology, Philipps-University, Marburg, Germany

[8] pharmazentrum frankfurt/ZAFES, Institute of Clinical Pharmacology, Goethe-University, Frankfurt am Main, Germany

[9] Fraunhofer Institute for Molecular Biology and Applied Ecology (IME), Project group Translational Medicine and Pharmacology (TMP), Frankfurt am Main, Germany

#Equal contribution

***Corresponding authors**

Mark N. Wass (M.N.Wass@kent.ac.uk), Martin Michaelis (m.michaelis@kent.ac.uk), Jindrich Cinatl jr. (Cinatl@kent.ac.uk)

3.1 My contribution

I performed all analysis of the exome sequencing data for these cell lines, including setting up the computational pipeline that was used and generating the figures. This covers sections' Variant calling' and 'Mutational signatures' from 'Materials and Methods'. I wrote the paper's first draft, which I then worked on with Mark Wass and Martin Michaelis. This manuscript is being prepared to go through the peer-review process for publication in a journal.

3.2 Abstract

The standard of treatment for acute myeloid leukaemia (AML) has remained unchanged for many years despite high mortality rates. Drugs commonly used to treat AML include cytarabine (a nucleoside analogue whose active form – cytarabine triphosphate - is incorporated into a growing DNA chain preventing DNA from extending) and daunorubicin (an anthracycline that inhibits the DNA over- and under-winding activity of topoisomerase II resulting in DNA damage). In addition, in recent years the FDA has approved multiple new drugs targeting molecular drivers of AML, including a combination therapy of cytarabine and daunorubicin (Vyxeos®). However, despite this progress, the 5-year survival rate has not increased, remaining at 25%-30%; hence more efficient drugs are required.

A new promising class of drugs are small molecule inhibitors that target MDM2 – a ubiquitin ligase that is a key negative regulator of p53. MDM2 inhibitors are currently under clinical investigation for acute myeloid leukaemia (AML), and patients with wild-type p53 are good candidates for clinical trials. We here investigated four nutlin-3-adapted sublines of the AML cell line MOLM13 to anticipate acquired MDM2 inhibitor resistance mechanisms. Whole exome sequencing identified complex mutation patterns associated with nutlin-3 resistance formation. The sublines MOLM13^{r}Nutlin$^{20\mu M}$I, MOLM13^{r}Nutlin$^{20\mu M}$II, and MOLM13^{r}Nutlin$^{20\mu M}$III harboured loss-of-function *TP53* mutations. Surprisingly, nutlin-3 resistant sublines (MOLM13^{r}Nutlin$^{20\mu M}$II, MOLM13^{r}Nutlin$^{20\mu M}$III, and MOLM13^{r}Nutlin$^{20\mu M}$IV) also displayed mutations in SAMHD1, which cleaves the triphosphorylated active form of cytarabine, causing its deactivation as its natural function is to cleave deoxynucleoside triphosphates (dNTPs) into deoxyribonucleosides and inorganic triphosphate. All four nutlin-3-adapted MOLM13 sublines displayed cross-resistance to daunorubicin, probably associated with *TP53* mutations and other impairments of p53 signalling in nutlin-3-resistant cells. In contrast, all four sublines displayed increased cytarabine sensitivity, probably caused by loss-of-triphosphohydrolase-function *SAMHD1* mutations in three sublines that could prohibit cleaving the cytarabine triphosphate and thus allowing its active form to be retained in the cell. High cytarabine triphosphate levels indicated increased cytarabine activation in *SAMHD1* wild-type MOLM13rNutlin20µMI cells. In addition, *SAMHD1* point mutations cause changes in cellular nucleotide levels that promote mutations and cancer cell evolution. Hence, *SAMHD1* mutations and other mechanisms that affect deoxynucleoside phosphorylation may support evolutionary processes underlying AML cell adaptation to MDM2 inhibitors. Among 29 nutlin-3-adapted sublines of the AML cell lines MOLM-13, MV4-11, and SIG-M5, 18 (62%) displayed increased cytarabine sensitivity, 9 (31%) unchanged cytarabine response, and two (7%) cytarabine resistance. In conclusion, nutlin-3-adapted MOLM13 sublines displayed cross-resistance to daunorubicin, probably caused by nutlin-3-induced *TP53* mutations and other impairments of p53 signalling, whereas they showed increased cytarabine sensitivity, probably caused by *SAMHD1* mutations and other mechanisms affecting cytarabine phosphorylation. *SAMHD1* mutations are candidate biomarkers indicating cytarabine-sensitive AML after the failure of MDM2 therapies. AML patients treated with MDM2 inhibitors should be continuously monitored for TP53 mutations indicating an increasing chance of resistance to this therapy. Then, provided they display loss-of-function mutations in SAMHD1, cytarabine may be prescribed as a second-line treatment.

Keywords: MDM2 inhibitor, nutlin-3, acute myeloid leukaemia, AML, acquired resistance, drug resistance

3.3 Introduction

MDM2 inhibitors are a novel class of anti-cancer drugs, which are being clinically developed for *TP53* wild-type cancers from different entities including acute myeloid leukaemia (Erba *et al.*, 2019; Khurana and Shafer, 2019; Pi *et al.*, 2019; Konopleva *et al.*, 2020). They exert their anti-cancer effects by activating p53, which is arguably the most important tumour suppressor protein. In agreement with the crucial role of p53 as tumour suppressor, *TP53*, the gene that encodes p53, is the most commonly mutated gene in cancer. About 50% of cancers harbour cells without functional p53 due to loss-of-function mutations or *TP53* gene deletions. When p53 becomes activated, it induces as transcription factor the expression of a large number of target genes, which in cancer cells typically results in cell death (Huang, 2020; Levine, 2020).

MDM2 is one of the p53 target genes and encodes for an endogenous inhibitor of p53. The MDM2 protein physically interacts with p53 and mediates its ubiquitinylation and proteasomal degradation. Hence, MDM2 inhibitors are candidate drugs for the activation of p53 and induction of cell death in *TP53* wild-type cancer cells (Wade, Li and Wahl, 2013; Tisato *et al.*, 2017; Huang, 2020; Levine, 2020).

Various MDM2 inhibitors have been shown to exert anti-cancer effects in pre-clinical models of AML including patient-derived xenografts, alone or in combination with other drugs (Kojima *et al.*, 2005; Secchiero *et al.*, 2007; Long *et al.*, 2010; Samudio *et al.*, 2010; McCormack *et al.*, 2012; Weisberg *et al.*, 2015; Kojima, Ishizawa and Andreeff, 2016; Lehmann *et al.*, 2016; Cassier *et al.*, 2017; Pan *et al.*, 2017; Seipel *et al.*, 2018; Maganti *et al.*, 2018) and subsequently been introduced into clinical trials (Erba *et al.*, 2019; Khurana and Shafer, 2019; Pi *et al.*, 2019; Konopleva *et al.*, 2020).

Drug-adapted cancer cell lines can indicate clinically relevant resistance mechanisms (Engelman *et al.*, 2007; Nazarian *et al.*, 2010; Poulikakos *et al.*, 2011; Domingo-Domenech *et al.*, 2012; Joseph *et al.*, 2013; Korpal *et al.*, 2013; Crystal *et al.*, 2014; Göllner *et al.*, 2017; Schneider *et al.*, 2017; Michaelis, Wass and Cinatl, 2019). Pre-clinical studies using cancer cell lines adapted to MDM2 inhibitors indicated that the treatment of *TP53* wild-type cancer cells results in the formation of *TP53* mutations as resistance mechanism (Aziz, Shen and Maki, 2011; Michaelis *et al.*, 2011, 2012; Jones *et al.*, 2012; Cinatl *et al.*, 2014; Gianna Hoffman-Luca *et al.*, 2015; Drummond *et al.*, 2016), which was clinically confirmed in liposarcoma patients (Jung *et al.*, 2016; Marcellino *et al.*, 2020).

To further investigate acquired resistance to MDM2 inhibitors, we here used exome sequencing to characterise four sublines of the acute myeloid leukaemia cell line MOLM13 adapted to nutlin-3, an MDM2 inhibitor closely related to idasanutlin, which is currently undergoing clinical trials (Vassilev *et al.*, 2004; Cinatl *et al.*, 2014; Khurana and Shafer, 2019; Pi *et al.*, 2019; Konopleva *et al.*, 2020).

3.4 Materials and Methods

3.4.1 Cells

The AML cell lines MOLM13, MV4-11, and SIG-M5 were obtained from DSMZ (Braunschweig, Germany). The nutlin-3-resistant sub-lines were established by adaption to growth in the presence of increasing drug concentrations as previously described (Michaelis *et al.*, 2011) and derived from the resistant cancer cell line (RCCL) collection (www.kent.ac.uk/stms/cmp/RCCL/RCCLabout.html) (Michaelis, Wass and Cinatl, 2019).

All cells were propagated in IMDM supplemented with 10 % FBS, 100 IU/ml penicillin and 100 mg/ml streptomycin at 37°C. Cells were routinely tested for mycoplasma contamination and authenticated by short tandem repeat profiling.

3.4.2 Whole-exome sequencing

Whole-exome sequencing was performed with Illumina HiSeq2000 using paired-end reads of a length of 100 base pairs. Exome enrichment was conducted using Nextera Exome Enrichment Kit.

3.4.3 Variant calling

The variant calling pipeline is summarised in Figure S9. After initial quality control using FastQC (Andrews, 2010), reads were trimmed with Trimmomatic-0.38 (default parameters) (Bolger, Lohse and Usadel, 2014) and mapped onto the reference genome (version hg19) using the Burrows-Wheeler Alignment Tool with the algorithm bwa-0.7.17-mem (Li and Durbin, 2009). Duplicate PCR reads were marked and .bam files built using Picard-2.17.10 (*Picard Tools - By Broad Institute*, 2019). GenomeAnalysisTK-3.7.0 (McKenna *et al.*, 2010) was used to realign sequences around insertions/ deletions and to recalibrate base scores. The machine learning model of covariation was built using dbSNP database (Sherry *et al.*, 2001) (downloaded on 23rd of April 2018) as known sites. Variants were called with samtools-1.7 mpileup (Li, 2011) using default parameters. Phred quality score filters were set to 30 with bcftools-1.6 (Li *et al.*, 2009) and variants with base call coverage below ten and variant call coverage below three were removed. Variants that affected protein sequences were identified with VEP (release 96) (McLaren *et al.*, 2016) and categorised following the Sequence Ontology nomenclature (Eilbeck *et al.*, 2005). We considered frameshift, stop-gained, splice acceptor, splice donor, incomplete terminal codon, stop-lost, start-lost, missense, inframe insertion, and inframe deletion variants. Germline-like variants with a frequency ≥0.001% in gnomAD (Lek *et al.*, 2016) were only included if at least three sequences were recorded in TCGA and at least ten in COSMIC (Tate *et al.*, 2019). The potential impact on protein structure/ function was estimated by SIFT (Kumar, Henikoff and Ng, 2009) and PolyPhen-2 (Adzhubei *et al.*, 2010).

3.4.4 Mutational signatures

Mutational signatures were analysed with MuSiCa (Díaz-Gay *et al.*, 2018) for Whole Exome Sequencing using reference genome version hg19. Additional mutational signatures were reconstructed from thirty current COSMIC mutational signatures.

3.4.5 Viability assay

Cell viability was tested by the 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide (MTT) dye reduction assay after 120 h incubation modified after Mosmann (Mosmann, 1983) as described previously (Onafuye *et al.*, 2019). Results are expressed as mean ± S.D. of at least three experiments.

3.4.6 Quantification of cytarabine triphosphate levels by LC-MS/MS analysis

Cells were incubated with $^{13}C_3$-cytarabine (SC-217994, Santa Cruz) at 37 °C in a humidified 5% CO2 incubator (Sanyo MCO-18AIC) for six h. Subsequently, cells were washed twice in 1 ml PBS, pelleted and stored at −20 °C until measurement. The concentration of $^{13}C_3$-cytarabine triphosphate was analysed by liquid chromatography-electrospray ionisation-tandem mass spectrometry as previously described (Thomas *et al.*, 2015; Schneider *et al.*, 2017). Briefly, the analytes were extracted by protein precipitation with methanol. An anion exchange HPLC column (BioBasic AX, 150 × 2.1 mm, Thermo) was used for the chromatographic separation, and a 5500 QTrap (Sciex) instrument was used to analyse the samples, operating as triple quadrupole in positive multiple reaction monitoring (MRM) mode. The precursor-to-product ion transitions used as quantifiers was m/z 487.0 → 115.1 for $^{13}C_3$-cytarabine triphosphate. Due to the lack of a commercially available standard of $^{13}C_3$-cytarabine triphosphate, relative quantification was performed using cytidine-$^{13}C_9$,$^{15}N_3$-5′-triphosphate as internal standard.

3.5 Results

3.5.1 Number of sequence variants in MOLM13 and its nutlin-3-adapted sublines

MOLM13 cells were adapted to growth in the presence of nutlin-3 20µM by continuous exposure to stepwise increasing drug concentrations as previously described (Michaelis *et al.*, 2011). All sublines displayed substantially decreased sensitivity to nutlin-3 relative to MOLM13 as indicated by the concentrations that inhibit cell viability by 50% (IC50, MOLM13: 0.68 ± 0.22µM, MOLM13^{r}Nutlin$^{20\mu M}$I: 12.9 ± 2.5µM, MOLM13^{r}Nutlin$^{20\mu M}$II: 11.3 ± 0.7µM, MOLM13^{r}Nutlin$^{20\mu M}$III: 12.3 ± 1.8µM, MOLM13^{r}Nutlin$^{20\mu M}$IV: 10.6 ± 2.3µM).

Exome sequencing analysis of MOLM13 and its nutlin-3-adapted sublines showed that all cell lines display similar numbers of variants. In total, between 219,602 (MOLM13) and 249,532 (MOLM13^{r}Nutlin$^{20\mu M}$I) variants were called per cell line. The number of high-quality variants (Phred score ≥ 30) ranged from 122,818 (MOLM13) to 138,114 (MOLM13^{r}Nutlin$^{20\mu M}$I), the number of high-quality/ high-coverage variants (Phred score ≥ 30, ≥ 10 alleles covered, variant covered by ≥ three reads) from 58,469 to 68,084, and the number of high-quality/ high-coverage/ somatic-like variants (like high quality/ high-coverage plus either frequency in gnomAD below 0.001% or at least 10 samples with this variant in COSMIC (Tate *et al.*, 2019) and three in the TCGA (https://www.cancer.gov/tcga)) from 23,193 (Molm13^{r}Nutlin$^{20\mu M}$I) to 29,033 (MOLM13^{r}Nutlin$^{20\mu M}$III). Among the high-quality/ high-coverage/ somatic-like variants, 371 (MOLM13^{r}Nutlin$^{20\mu M}$III) to 422 (MOLM13^{r}Nutlin$^{20\mu M}$I) variants were considered to be most likely to have a functional impact because of their nature (frameshift variant, stop-gained, splice acceptor variant, splice donor variant, incomplete terminal codon variant, stop-lost, start-lost, missense variant, inframe insertion inframe deletion) (Eilbeck *et al.*, 2005) (Figure 3.1).

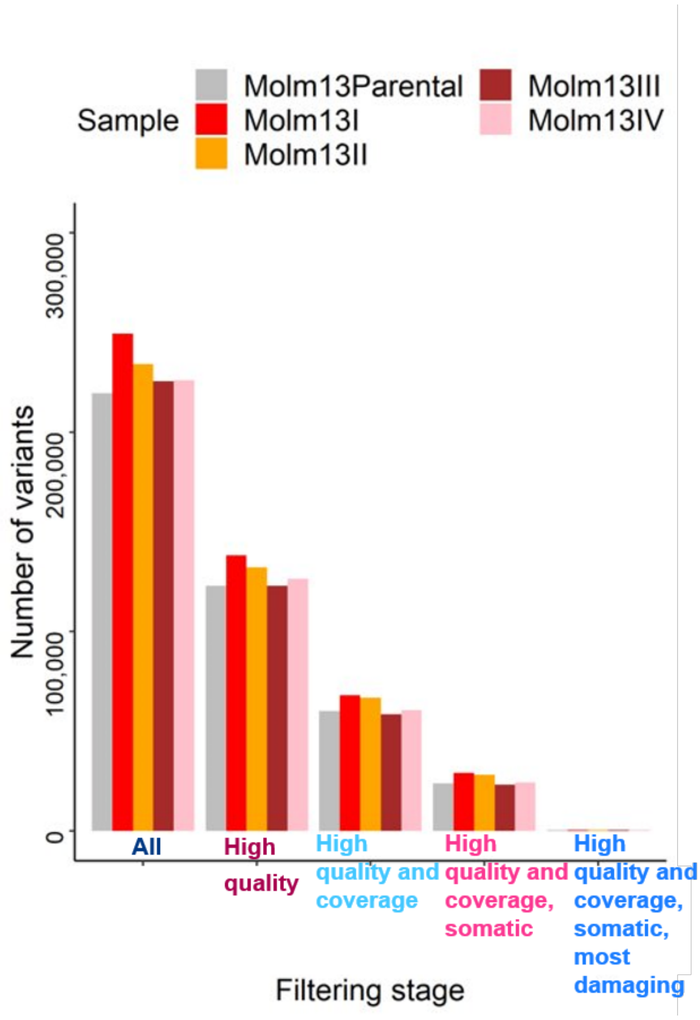


Figure 3.1 The number of variants per each filtering stage in the pipeline. *All – all the variants that were called; high quality – only variants with a base quality score (Phred score) of minimum 30; high quality and coverage variants – high-quality variants where the position is covered by at least ten reads and the variant allele itself is covered by at least three reads; high quality and coverage, somatic – high-quality and high-coverage variants that are considered to be rare with the criterium of at least ten samples carrying this variant in COSMIC* (Tate *et al.*, 2019) *and 3 in TCGA* (Weinstein *et al.*, 2013) *database; high quality and coverage, somatic, most damaging – high-quality, high-coverage, somatic-like variants with the most damaging consequences that are in the scope of sequencing (frameshift variant, stop-gained, splice acceptor variant, splice donor variant, incomplete terminal codon variant, stop-lost, start-lost, missense variant, inframe insertion and inframe deletion).*

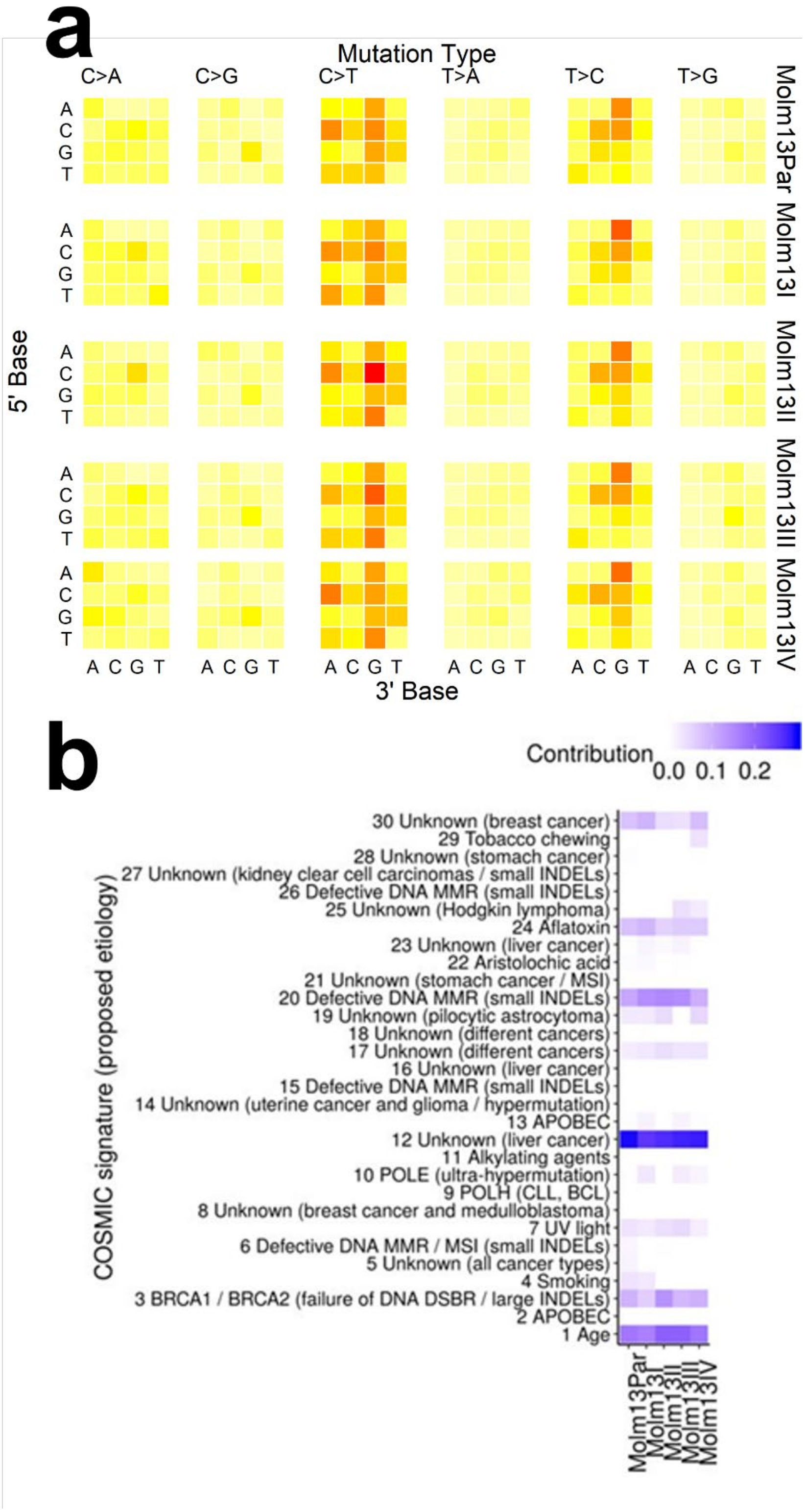


Figure 3.2 Mutational signatures. a *Mutational profile of high-quality, high-coverage, somatic-like variants in the scope of sequencing calculated by MuSiCa* (Díaz-Gay *et al.*, 2018) *for the parental cell lines and all the drug-adapted sub-lines. The X-axis represents 3' base and Y-axis – 5' base. Each column is a different substitution type, e.g. the top left corner is a mutation from ACA to AAA in the Molm13 parental cell line. The coloured heatmap represents the frequency of the mutation type in the cell line – white when the*

frequency equals to 0, yellow for 0.1 and red for 0.17. ***b*** *Contribution of the current thirty COSMIC* (Tate *et al.*, 2019) *mutational signatures into the reconstructed mutational signature of high-quality, high-coverage, somatic-like variants in the protein-coding regions created by MuSiCa for the parental cell line and all the drug-adapted sub-lines.*

3.6 Mutational signatures

The combination of mutation types can be categorised in mutational signatures (Alexandrov and Stratton, 2014; Helleday, Eshtad and Nik-Zainal, 2014). Mutational signatures of the MOLM13 parental and drug-adapted sub-lines were calculated for high-quality, high-coverage, somatic-like variants from the protein-coding regions using MuSiCa (Díaz-Gay *et al.*, 2018). The mutational signatures were very similar across MOLM13 and its nutlin-3-adapted sublines (Figure 3.2a) and resembled the mutational MOLM13 signature from the COSMIC database (Tate *et al.*, 2019) (Figure S10). All the cell lines are enriched in the X[C>T]X and X[T>C]X substitutions, especially in the context of X[C>T]G and A[T>C]G or C[T>C]G. Reconstruction of the mutational signatures using the COSMIC database (Tate *et al.*, 2019) also revealed very similar patterns (Figure 3.2b). Eleven of the 30 mutations signatures contributed to the mutational signature of the MOLM13 cell lines, with signatures 1, 3, 7, 12, 17, 20, 24, and 30 contributing to all of them and signature 12 having the highest impact in all five cell lines (Figure 3.2b). Taken together, nutlin-3 adaptation did not substantially affect the mutational signatures of the investigated MOLM13 cell lines.

3.7 Sequence changes between parental and drug-adapted cells

Next, we analysed the changes in the nutlin-3-resistant sublines relative to the parental MOLM13 cell line. We categorised changes into gained (present but not called in MOLM13, called in a subline), de novo (no read in MOLM13, called in a subline), not called (called in MOLM13, present but not called in subline), and lost (called in MOLM13, no read in a subline). Moreover, we considered significant changes in the allele frequencies of variants that were called in MOLM13 and its sublines (Table S4, Figure S11).

The majority of the high-quality/ high-coverage/ somatic-like most-likely damaging variants (from 68% in MOLM13^{r}Nutlin$^{20\mu M}$I to 72% in MOLM13^{r}Nutlin$^{20\mu M}$II and MOLM13^{r}Nutlin$^{20\mu M}$III) were shared between MOLM13 and either of its sublines (Figure 3.3, Table S5). The nutlin-3-adapted MOLM-13 sublines harboured similar proportions of de novo (13-17%), gained (13-16%), increased (0.2-1.6%), and unchanged (68-72%). In addition, 9-18% of the high-quality/ high-coverage/ somatic-like and most-likely damaging variants in MOLM13 cell line were present but not called in the sublines, and 16-18% were lost (no reads identified in the sublines) (Figure 3.3, Table S5).

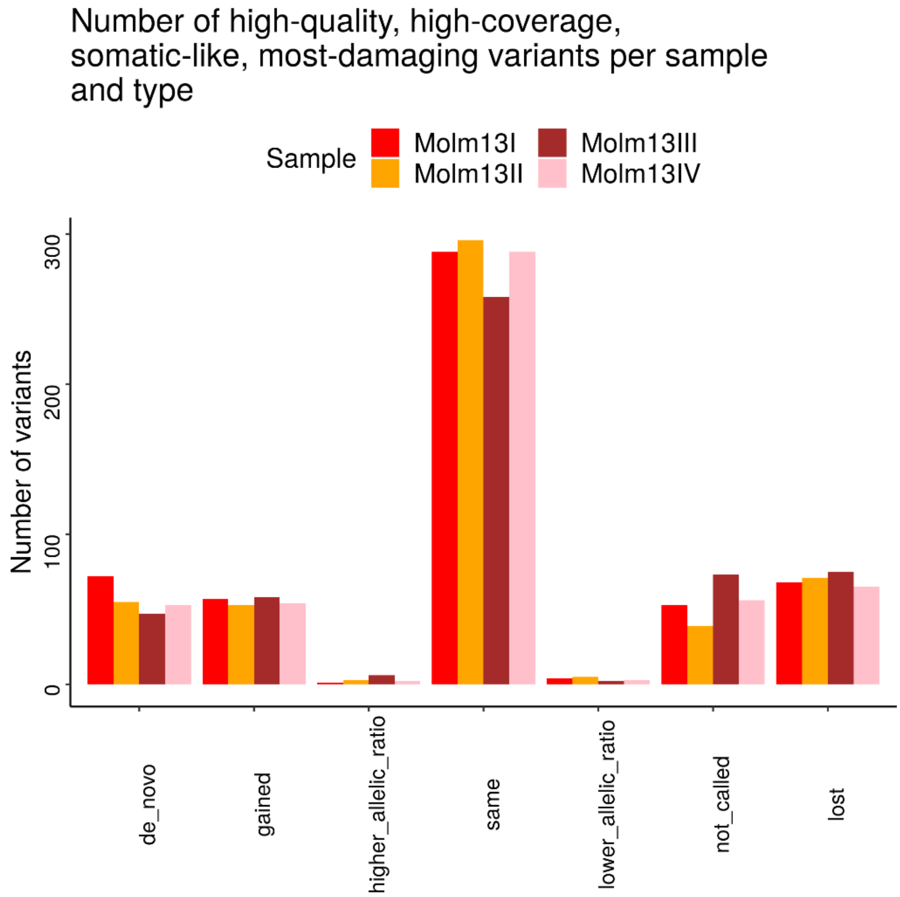


Figure 3.3 The number of variants per drug-adapted sub-line and class. *High-quality, high-coverage, somatic-like and most-damaging variants in each drug-adapted and parental cell line were classified as de novo, gained, higher allelic ratio, same, lower allelic ratio, not called and lost and then counted in each of the drug-adapted sub-line – Molm13I, Molm13II, Molm13III and Molm13IV.*

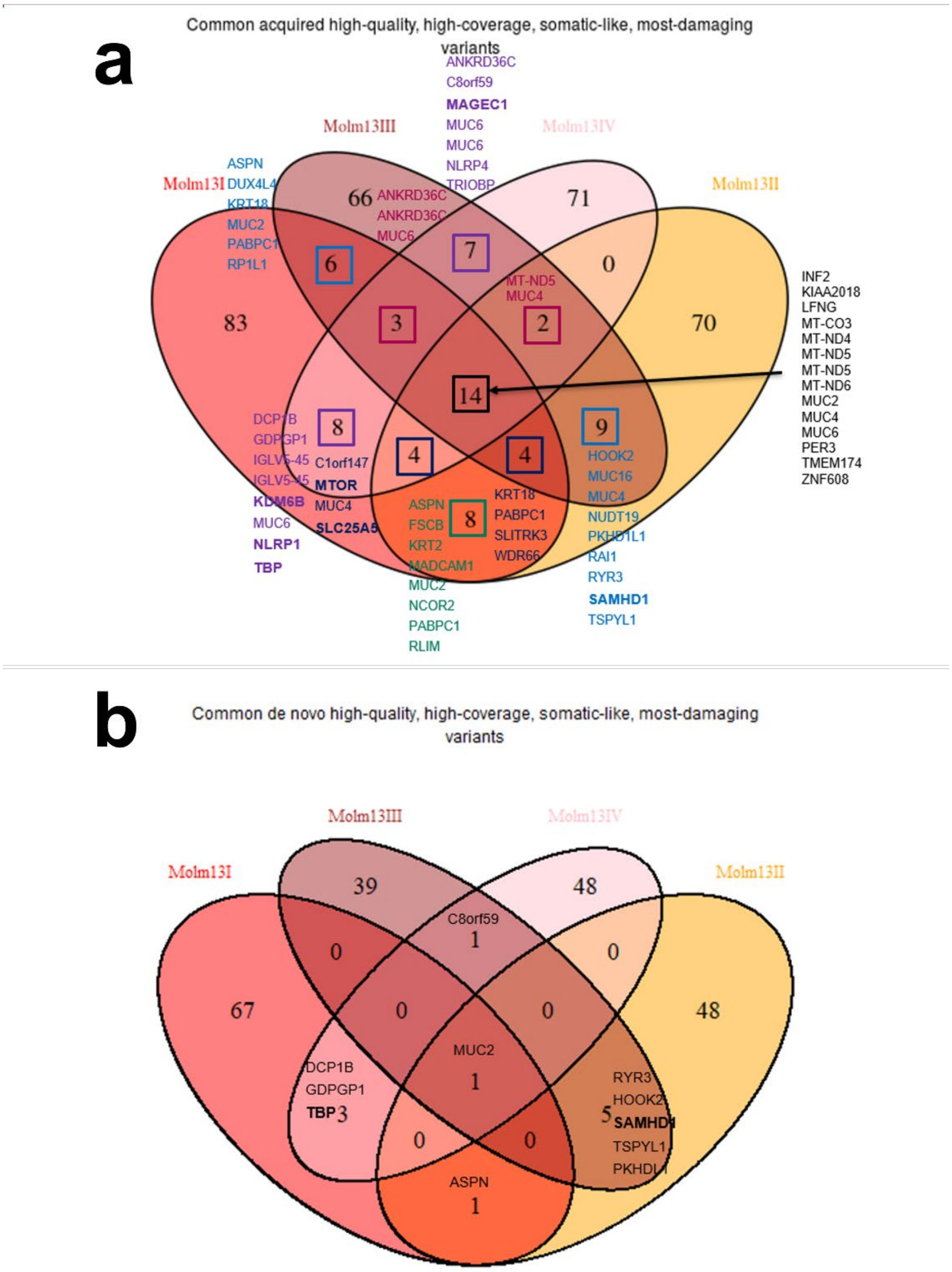


Figure 3.4 Common high-quality, high-coverage, somatic-like and most-damaging variants. *The number of high-quality, high-coverage, somatic-like and most-damaging variants that are shared between the drug-adapted sub-lines. The following variant types were considered:* ***a*** *acquired (de novo, gained, higher allelic ratio)* ***b*** *de novo.*

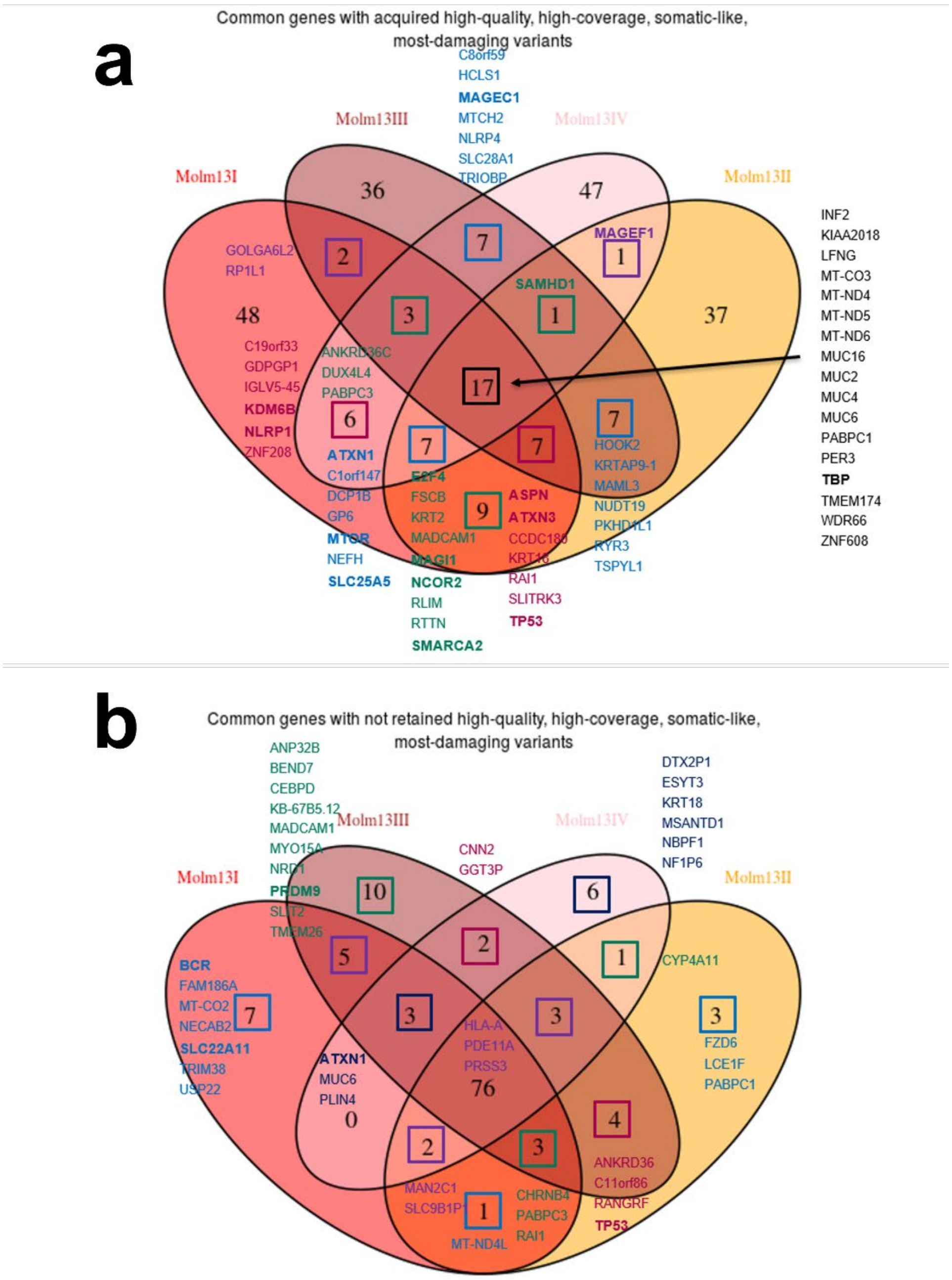


Figure 3.5 Common genes with high-quality, high-coverage, somatic-like and most-damaging variants. *The number of genes with high-quality, high-coverage, somatic-like and most-damaging variants shared between the drug-adapted sub-lines, i.e. genes do not have to carry the same variant in two different drug-adapted sub-lines to be considered shared between those cell lines. Genes with the following variant types were considered:* ***a*** *acquired (de novo, gained, higher allelic ratio)* ***b*** *not retained (lost, not called, lower allelic ratio).*

Only 14 of the high-quality, high-coverage, somatic-like, most-damaging variants that were acquired in the drug-adapted sub-lines were shared between all four drug-adapted sub-lines (Appendix 4 Supplementary Tables 1 - 8). One of the 14 variants, an inframe deletion in MUC2 Mucin-2, is a *de novo* variant (Figure 3.4a, Figure 3.4b, Appendix 4 Supplementary Tables 1 - 8). Since it is unlikely that exactly the same mutation independently occurred four times, it is more likely that it is present in MOLM13, but was not detected. This may indicate that all four sublines may be derived from a relatively rare subpopulation. Generally, shared *de novo* mutations probably rather indicate a shared clonal origin than the independent occurrence of identical mutations.

When considering genes with mutations (but not necessarily the same variant), 17 genes were found to harbour mutations in all four sublines. The numbers of genes that were only mutated in two or three sublines were lower (Figure 3.5a, Appendix 4 Supplementary Tables 1 - 8). A higher overlap was found among variants that were called in the MOLM13 but not in the sublines. Mutations disappeared in 76 genes across all four sublines (Figure 3.5b, Appendix 4 Supplementary Tables 1 - 8), while substantially lower numbers (0-5) of additional non-retained mutations were only shared between two and three cell lines.

3.8 Identification of potential driver variants

For the identification of potential driver variants, we focused on high-quality/ high-coverage/ somatic-like variants/ likely damaging variants. The nutlin-3-adapted MOLM13 cell lines harboured between one (MOLM13^{r}Nutlin$^{20\mu M}$III) and four (MOLM13^{r}Nutlin$^{20\mu M}$I) *de novo* mutations that have been reported as cancer drivers in the COSMIC (Tate *et al.*, 2019) and Intogen (Gonzalez-Perez, Perez-Llamas, *et al.*, 2013) (Appendix 4 Supplementary Tables 9 - 12). This includes *TP53* mutations in MOLM13^{r}Nutlin$^{20\mu M}$I, MOLM13^{r}Nutlin$^{20\mu M}$II, and MOLM13^{r}Nutlin$^{20\mu M}$III (Figure 3.6a, Appendix 4 Supplementary Tables 9 - 12). Moreover, MOLM13^{r}Nutlin$^{20\mu M}$I harboured *de novo* mutations in *TMED8* and *PPM1D*, MOLM13^{r}Nutlin$^{20\mu M}$II in *ATP13A4*, and MOLM13^{r}Nutlin$^{20\mu M}$IV in *GNAQ* (Appendix 4 Supplementary Tables 9 - 12). Two gained mutations were shared by at least three sublines: *PER3* (shared by all four sublines) and *PABPC1* (MOLM13^{r}Nutlin$^{20\mu M}$I, MOLM13^{r}Nutlin$^{20\mu M}$II, MOLM13^{r}Nutlin$^{20\mu M}$III) (Appendix 4 Supplementary Tables 9 - 12).

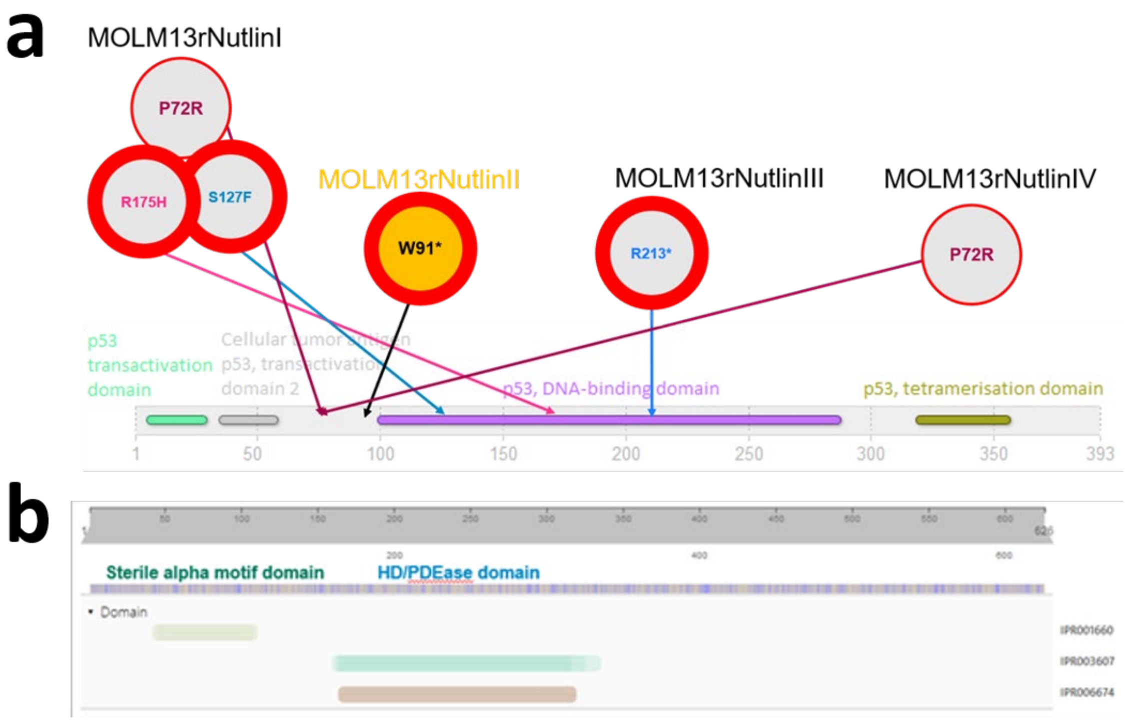


***Figure 3.6 a** Variants acquired in TP53 in the drug-adapted sub-lines, mapped to the domains of p53 protein (screenshot from InterPro* (Mitchell *et al.*, 2019)*). **b** Domains of SAMHD1 (screenshot from InterPro).*

Moreover, we used the bioinformatics tools SIFT and PolyPhen-2 to analyse the potential impact of mutations on protein structure and function. SIFT predicts the effects of amino acid substitutions on protein function based on their conservation status (Kumar, Henikoff and Ng, 2009). PolyPhen-2 predicts the effects of amino acid substitutions on protein function using a machine-learning classification based on mapping of the underlying single nucleotide variants to gene transcripts, protein sequence annotations, structural attributes, and conservation profiles (Adzhubei *et al.*, 2010).

SAMHD1 was the only gene that displayed *de novo* mutations predicted to affect protein function by SIFT and PolyPhen-2 in at least three sublines (MOLM13^{r}Nutlin$^{20\mu M}$II, MOLM13^{r}Nutlin$^{20\mu M}$III, MOLM13^{r}Nutlin$^{20\mu M}$IV). Moreover, mutations in *MT-ND5* and *MT-ND6* were gained in all four sublines, and SLC25A5 had gained mutations in three sublines (MOLM13^{r}Nutlin$^{20\mu M}$I, MOLM13^{r}Nutlin$^{20\mu M}$II, MOLM13^{r}Nutlin$^{20\mu M}$IV) (Appendix 4 Supplementary Tables 9 - 12).

We did not find an obvious potential role for TMED8, ATP13A4, and SLC25A5 in nutlin-3 resistance in AML. *GNAQ* mutations have been identified as driver mutations in uveal melanoma, and the T96S amino acid substitution identified here was detected in angiomatosis and natural killer/ T-cell lymphoma (Li *et al.*, 2019; Gaeta *et al.*, 2020; Gaffal, 2020). PABPC1 activity has been described as oncogenic event in hepatocellular carcinoma and as mediator of trastuzumab resistance in breast cancer (Dong *et al.*, 2018; Zhang *et al.*, 2020). Moreover, *PABPC1* mutations have been described in colorectal cancer (Yu *et al.*, 2014). PPMD1 is a regulator of the cellular response to DNA damage (Nahta and Castellino, 2020). Decreased PER3 levels have been associated with unfavourable outcome in breast cancer and hepatocellular carcinoma, and *PER3* deletion with recurrence of oestrogen-positive breast cancer (Chen *et al.*, 2005; Lin *et al.*, 2008; Climent *et al.*, 2010). *MT-ND5* and *MT-ND6* mutations affect mitochondrial function and are detected in different cancers (Kloss-Brandstätter *et al.*, 2010; Järviaho *et al.*, 2018; Nguyen, Kim and Jo, 2020). Hence, mutations in *GNAQ*, *PABPC1*, *PPMD1*, *PER3*, *MT-ND5*, and *MT-ND6* may contribute to acquired nutlin-3 resistance, but detail remains to be investigated.

3.8.1 Mutations in *TP53* and *SAMHD1*

The MOLM13 cell line is known to encode functional p53 harbouring a well-known (67% frequency in gnomAD database (Lek *et al.*, 2016) (Table S6) p.P72R polymorphism (Tate *et al.*, 2019), which was confirmed in our analysis (Table S7). Acquired resistance to MDM2 inhibitors is known to be associated with the formation of *TP53* mutations (Aziz, Shen and Maki, 2011; Michaelis *et al.*, 2011, 2012; Jones *et al.*, 2012; Cinatl *et al.*, 2014; Gianna Hoffman-Luca *et al.*, 2015; Drummond *et al.*, 2016; Jung *et al.*, 2016), and *TP53* mutations are associated with a poor prognosis in AML (Döhner *et al.*, 2017). As described above and in agreement with previous findings, three of the four nutlin-3-resistant MOLM13 sublines harboured TP53 mutations (Figure 3.6a, Appendix 4 Supplementary Tables 9 - 12).

MOLM13^{r}Nutlin$^{20\mu M}$I harbours two *TP53* mutations in the DNA-binding domain (p.R175H, p.S127F) (Figure 3.6a, Table S7). MOLM13^{r}Nutlin$^{20\mu M}$II and MOLM13^{r}Nutlin$^{20\mu M}$III both acquired homozygous stop mutations, MOLM13^{r}Nutlin$^{20\mu M}$II at position 91 (truncates p53 before the DNA-binding domain) and MOLM13^{r}Nutlin$^{20\mu M}$III at position 213 (truncates p53 roughly in the middle of the DNA-binding domain) (Figure 3.6a, Table S7). Only, MOLM13^{r}Nutlin$^{20\mu M}$IV retained wild-type *TP53*. This reflects previous findings showing that, while adaptation of *TP53* wild-type to MDM2 inhibitors commonly results in *TP53* mutations, some MDM2 inhibitor-adapted cell lines maintain wild-type *TP53* (Michaelis *et al.*, 2011).

Sterile alpha motif and histidine/aspartic acid domain containing protein 1 (SAMHD1) is a triphosphohydrolase that cleaves and inactivates cytarabine triphosphate, the active form of cytarabine, a nucleoside analogue commonly used for the treatment of AML (Schneider *et al.*, 2017). In AML cells, high SAMHD1 activity is associated with cytarabine resistance and low SAMHD1 activity is associated with cytarabine sensitivity (Schneider *et al.*, 2017). While MOLM13 cells harbour wild-type *SAMHD1*, potentially deleterious *SAMHD1* mutations were detected in three of the sublines (Figure 3.6b, Table S8).

MOLM13^{r}Nutlin$^{20\mu M}$II and MOLM13^{r}Nutlin$^{20\mu M}$III both display a p.Y553C amino acid substitution. As described above, SIFT classed the p.Y553C amino acid substitution in SAMHD1 as 'deleterious', PolyPhen-2 as 'probably damaging' (Table S9). MOLM13^{r}Nutlin$^{20\mu M}$IV harbours a p.S121L amino acid substitution, which is also classified as 'deleterious' by SIFT and 'probably damaging' by PolyPhen-2 (Table S9).

3.8.2 Nutlin-3 resistance is associated with collateral cytarabine sensitivity in MOLM13 cells

Since *TP53* mutations are associated with a poor outcome in AML (Döhner *et al.*, 2017) and SAMHD1 is a critical determinant of AML sensitivity to the nucleoside cytarabine (Schneider *et al.*, 2017), we investigated whether the nutlin-3-resistant MOLM13 sublines displayed altered sensitivity to cytarabine and anthracyclines that are commonly used for AML treatment (Cheung *et al.*, 2019). The drug concentrations that reduce cell viability by 50% (IC50) shown in Figure 3.7a are much higher for MOLM13 nutlin-3 resistant sublines than the parental cell line when cells are treated with daunorubicin. However, by contrast, the MOLM13 nutlin-3 adapted sublines are much more sensitive to cytarabine than the MOLM13 parental cell line. This means that all four nutlin-3-resistant MOLM13 sublines displayed increased sensitivity to cytarabine but increased resistance to the anthracycline daunorubicin (Figure 3.7a, Table S10). Thus, acquired nutlin-3 resistance seems to be associated with collateral sensitivity to cytarabine but not to other anti-cancer drugs such as daunorubicin.

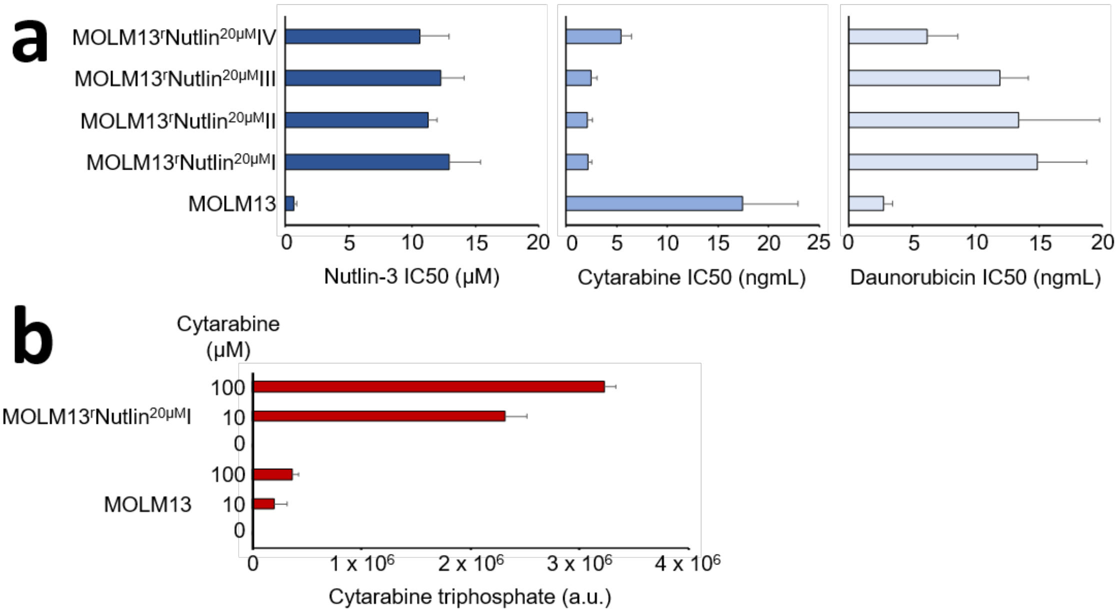


Figure 3.7 *Drug sensitivity profiles and cytarabine triphosphate formation in MOLM13 cell lines.* ***a*** *Drug concentrations that reduce cell viability by 50% (IC50) after 120h incubation as indicated by MTT assay.* ***b*** *Cytarabine triphosphate levels in MOLM13 and MOLM13^{r}Nutlin$^{20\mu M}$I cells in response to different cytarabine concentrations.*

The increased cytarabine sensitivity of MOLM13^{r}Nutlin$^{20\mu M}$II, MOLM13^{r}Nutlin$^{20\mu M}$III, and MOLM13^{r}Nutlin$^{20\mu M}$IV seems to confirm the SIFT and PolyPhen-2 predictions indicating that the observed *SAMHD1* mutations are associated with a loss of function. However, the reasons underlying the collateral cytarabine sensitivity of MOLM13^{r}Nutlin$^{20\mu M}$I are less obvious. To see whether cytarabine activation may differ between MOLM13 and MOLM13^{r}Nutlin$^{20\mu M}$I, we determined cytarabine triphosphate levels in these cell lines in response to cytarabine treatment. Results demonstrated substantially higher cytarabine triphosphate levels in MOLM13^{r}Nutlin$^{20\mu M}$I than in MOLM13 (Figure 3.7b), suggesting that resistance formation to MDM2 inhibitors may commonly be associated with changes in the processes underlying nucleoside analogue activation.

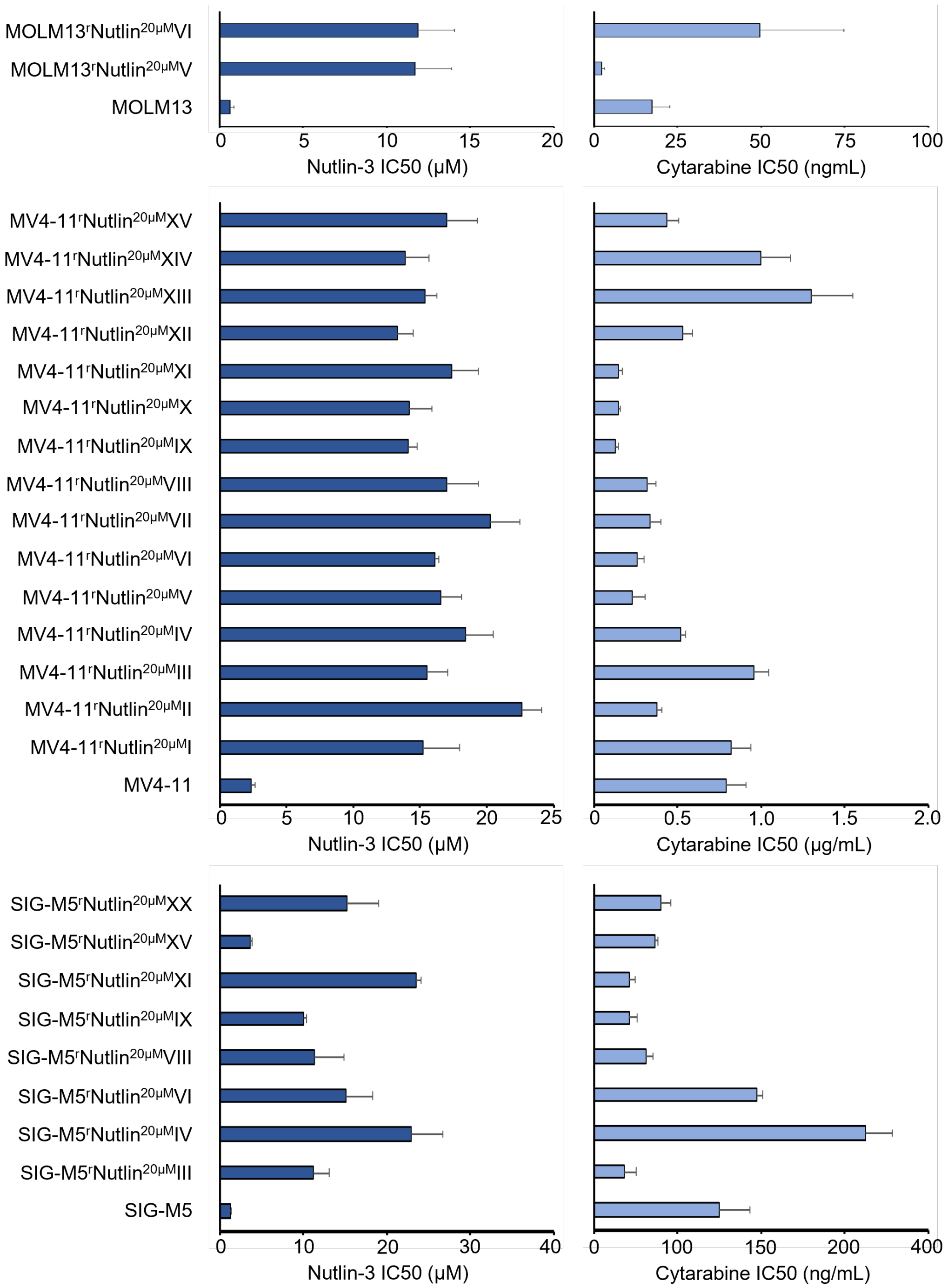


Figure 3.8 *Nutlin-3 and cytarabine concentrations that reduce the viability of parental AML cell lines and their nutlin-3-adapted sublines by 50% (IC50) after 120h incubation as indicated by MTT assay.*

Finally, we determined cytarabine sensitivity in a wider range of nutlin-3-adapted AML cell lines, including two additional MOLM13 sublines, 15 MV4-11 sublines, and eight SIG-M5 sublines (Figure 3.8, Table S11). Similarly to the four MOLM13 resistant sublines, a lower concentration of cytarabine is needed to reduce cell viability by 50% in the majority of the nutlin-3 adapted AML cell lines (compared to IC50 values for their parental cell lines). Among the in total 29 nutlin-3-adapted sublines of MOLM-13, MV4-11, and SIG-

M5, 18 (62%) displayed increased cytarabine sensitivity relative to the respective parental cell line (IC50 subline/ IC50 parental cell line ≤ 0.5), 9 (31%) displayed similar cytarabine sensitivity as the respective parental cell line (IC50 subline/ IC50 parental cell line > 0.5 and < 2), and two (7%), increased cytarabine resistance than the respective parental cell line (IC50 subline/ IC50 parental cell line ≥ 2). This demonstrates that acquired MDM2 resistance is regularly but not always associated with collateral sensitivity to cytarabine in AML cells.

3.9 Discussion

MDM2 inhibitors are under pre-clinical and clinical investigation for the treatment of AML (Kojima *et al.*, 2005; Secchiero *et al.*, 2007; Samudio *et al.*, 2010; Long *et al.*, 2010; McCormack *et al.*, 2012; Weisberg *et al.*, 2015; Kojima, Ishizawa and Andreeff, 2016; Lehmann *et al.*, 2016; Cassier *et al.*, 2017; Pan *et al.*, 2017; Seipel *et al.*, 2018; Maganti *et al.*, 2018; Erba *et al.*, 2019; Pi *et al.*, 2019; Khurana and Shafer, 2019; Konopleva *et al.*, 2020). The clinical efficacy of anti-cancer drugs is often affected by the formation of acquired resistance, at least in fraction of patients. Drug-adapted cancer cell lines can be used to investigate acquired resistance mechanisms, because they reflect clinically relevant resistance mechanisms (Engelman *et al.*, 2007; Nazarian *et al.*, 2010; Poulikakos *et al.*, 2011; Domingo-Domenech *et al.*, 2012; Joseph *et al.*, 2013; Korpal *et al.*, 2013; Crystal *et al.*, 2014; Göllner *et al.*, 2017; Schneider *et al.*, 2017; Michaelis, Wass and Cinatl, 2019). Here, we introduce the first AML models of acquired MDM2 inhibitor resistance, four sublines of the AML cell line adapted to nutlin-3.

To identify mutations with the potential to drive MDM2 inhibitor resistance, we analysed MOLM13 and its sublines by whole-exome sequencing and identified high-quality (Phred score ≥ 30)/ high-coverage (≥ 10 alleles covered, variant covered by ≥ three reads)/ somatic-like (frequency in gnomAD below 0.001% or at least ten samples with this variant in COSMIC (Tate *et al.*, 2019) and three in the TCGA (https://www.cancer.gov/tcga)) variants likely to have a functional impact due to their nature (frameshift variant, stop-gained, splice acceptor variant, splice donor variant, incomplete terminal codon variant, stop-lost, start-lost, missense variant, inframe insertion inframe deletion). This resulted in

371 (MOLM13^{r}Nutlin$^{20\mu M}$III) to 422 (MOLM13^{r}Nutlin$^{20\mu M}$I) high-quality/ high-coverage/ somatic-like/ potentially damaging variants per cell line.

Among these mutations, we focused on *de novo* mutations, which were called in at least one nutlin-3-adapted subline but not detected in MOLM13, and gained mutations (detectable but not called in MOLM13) that were detected in at least three sublines. Further criteria for candidate driver mutations included either annotation as cancer drivers in COSMIC or Intogen or being predicted to affect protein structure/ function by SIFT and PolyPhen-2.

Both SIFT and PolyPhen-2 apply similar principles behind protein function that we used in Chapter 2 to annotate proteins of the minimal bacterial genome. SIFT creates an evolutionary profile of the protein using PSI-BLAST and calculates amino acid position conservation which we applied to predict Gene Ontology terms using GOAT (see 2.4.3). PolyPhen-2, being an SVM classifier, uses features based on, among all, matching domains and sequence identity to the closest homologues. In Chapter 2, we also predicted protein domains and homologues to identify the unknown functions in the minimal genome.

Following the criteria of mutations being predicted to be damaging by SIFT and PolyPhen-2, and being present in databases of cancer drivers, we assembled a list of eleven genes of interest: *TP53*, *TMED8*, *PPM1D*, *ATP13A4*, *GNAQ* (*de novo*, annotated in COSMIC or Intogen), *PER3*, *PABPC1* (gained, annotated in COSMIC or Intogen), *SAMHD1* (*de novo*, predicted to affect protein function by SIFT and PolyPhen-2), *MT-ND5*, *MT-ND6*, and *SLC25A5* (gained, predicted to affect protein function by SIFT and PolyPhen-2).

For the mutations in *GNAQ*, *PABPC1*, *PPMD1*, *PER3*, *MT-ND5*, and *MT-ND6*, we found evidence supporting a potential role in nutlin-3 resistance (Table 3.1), but details remain to be investigated.

Mutated gene	Evidence
GNAQ	*GNAQ* driver mutations in uveal melanoma, T96S amino acid substitution in angiomatosis and natural killer/ T-cell lymphoma (Li *et al.*, 2019; Gaeta *et al.*, 2020; Gaffal, 2020).

PABPC1	PABPC1 driver in hepatocellular carcinoma, and mediator of trastuzumab resistance in breast cancer, *PABPC1* mutations in colorectal cancer (Yu *et al.*, 2014; Dong *et al.*, 2018; Zhang *et al.*, 2020).
PPMD1	DNA damage regulator (Nahta and Castellino, 2020).
PER3	Low PER3 levels associated with poor breast cancer/ hepatocellular carcinoma outcome, *PER3* deletion associated with recurrence of oestrogen-positive breast cancer (Chen *et al.*, 2005; Lin *et al.*, 2008; Climent *et al.*, 2010).
MT-ND5	*MT-ND5* mutations affect mitochondrial function and are detected in different cancers (Nguyen, Kim and Jo, 2020).
MT-ND6	*MT-ND6* mutations affect mitochondrial function and are detected in different cancers (Kloss-Brandstätter *et al.*, 2010; Järviaho *et al.*, 2018).

Table 3.1 *Evidence supporting a potential role of selected mutations in nutlin-3 resistance in MOLM13 sublines.*

The mutations with the most obvious potential relevance for AML therapies were detected in *TP53* and *SAMHD1*. To analyse the impact of these mutations on protein structure and function, we predicted Pfam (El-Gebali *et al.*, 2019) domains using the same methodology as applied in Chapter 2 to identify the functions of the minimal genome proteins. Future studies should include structural modelling using state-of-the-art protein structure prediction tools, e.g. AlphaFold2 (Jumper et al., 2021) (see 2.6.1) and assessing the effect of these mutations could have on the structure and ligand binding.

MOLM13^{r}Nutlin$^{20\mu M}$I harboured two *TP53* mutations missense in the DNA-binding domain (p.R175H, p.S127F), while MOLM13^{r}Nutlin$^{20\mu M}$II and MOLM13^{r}Nutlin$^{20\mu M}$III both acquired homozygous stop mutations. MOLM13^{r}Nutlin$^{20\mu M}$IV retained wild-type *TP53*. These findings are in agreement with previous findings showing that resistance acquisition to MDM2 inhibitors is commonly but not always associated with the formation of *TP53* mutations (Aziz, Shen and Maki, 2011; Michaelis *et al.*, 2011, 2012; Jones *et al.*, 2012;

Cinatl *et al.*, 2014; Gianna Hoffman-Luca *et al.*, 2015; Drummond *et al.*, 2016; Jung *et al.*, 2016; Marcellino *et al.*, 2020). Notably, *TP53* mutations are associated with therapy resistance and poor prognosis in AML patients (Döhner *et al.*, 2017). Thus, *TP53* mutations associated with acquired MDM2 inhibitor resistance formation may affect the prospects of potential next-line therapies.

Similarly, *SAMHD1* mutations detected in MOLM13^{r}Nutlin$^{20\mu M}$II (p.Y553C), MOLM13^{r}Nutlin$^{20\mu M}$III (p.Y553C), and MOLM13^{r}Nutlin$^{20\mu M}$IV (p.S121L) may impact follow-on treatments. The triphosphohydrolase SAMHD1 cleaves the triphosphorylated active form of cytarabine, a nucleoside analogue that is part of standard AML therapies, and is a biomarker indicating cytarabine sensitivity of AML cells (Schneider *et al.*, 2017; Cheung *et al.*, 2019).

All four nutlin-3-adapted MOLM13 sublines displayed cross-resistance to daunorubicin, which is commonly used in combination with cytarabine for AML (Cheung *et al.*, 2019). This is in agreement with the anticipated role of *TP53* mutations in treatment resistance in AML (Döhner *et al.*, 2017) and with the assumption that p53 signalling is also affected in MDM2 inhibitor-resistant cells that do not develop *TP53* mutations (Michaelis *et al.*, 2011).

In contrast to the increased daunorubicin resistance observed in nutlin-3-adapted MOLM13 cells, all four sublines displayed increased sensitivity to cytarabine. This suggests that the *SAMHD1* mutations observed in MOLM13^{r}Nutlin$^{20\mu M}$II, MOLM13^{r}Nutlin$^{20\mu M}$III, and MOLM13^{r}Nutlin$^{20\mu M}$IV are associated with a loss of its triphosphohydrolase function, which agrees with previous findings indicating that this function can be affected by various *SAMHD1* point mutations (Rentoft *et al.*, 2016). Interestingly, MOLM13^{r}Nutlin$^{20\mu M}$I also displayed increased cytarabine sensitivity, although it did not harbour a *SAMHD1* mutation. The determination of cellular cytarabine triphosphate levels in this cell line indicated reduced cytarabine activation levels in this cell line, suggesting that this process is impaired by a mechanism different from a *SAMHD1* mutation in this cell line. Notably, *SAMHD1* point mutations are associated with changes in cellular nucleotide levels that themselves can promote mutations and cancer cell evolution (Rentoft *et al.*, 2016). Hence, *SAMHD1* mutations and other mechanisms that affect deoxynucleoside phosphorylation may support evolutionary processes underlying AML cell adaptation to MDM2 inhibitors and other drugs.

The analysis of in total 29 nutlin-3-adapted sublines of the AML cell lines MOLM-13, MV4-11, and SIG-M5, indicated that AML cell resistance formation to MDM2 inhibitors is commonly but not always associated with collateral cytarabine sensitivity. 18 (62%) of the sublines displayed increased cytarabine, in 9 (31%) sublines cytarabine sensitivity was not changed, and two (7%) sublines showed increased cytarabine resistance.

In conclusion, AML resistance formation against MDM2 inhibitors is associated with complex mutation patterns. Most interestingly, nutlin-3-adapted MOLM13 sublines displayed cross-resistance to daunorubicin, which probably is associated with nutlin-3-induced *TP53* mutations and other impairments of p53 signalling, whereas they show increased sensitivity to cytarabine, which is associated with *SAMHD1* mutations resulting in a loss of triphosphohydrolase activity or other mechanisms affecting nucleoside phosphorylation. Hence, *SAMHD1* mutations are candidate biomarkers indicating cytarabine sensitivity after the failure of MDM2 therapies. Collateral cytarabine was detected in 18 (62%) out of 29 nutlin-3-adapted AML sublines and, thus, appears to be a common phenomenon.

3.10 Acknowledgements

The results <published or shown> here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

3.11 Grant support

The work was supported by the Hilfe für krebskranke Kinder Frankfurt e.V., the Frankfurter Stiftung für krebskranke Kinder, and the Kent Cancer Trust.

Chapter 4 Selection of different clones upon repeated adaptation of neuroblastoma cell lines to tubulin-binding agents

Magdalena Antczak[1], Lyto Yiangou[1], Helen E. Grimsley[1], Miguel Julia[1], Florian Rothweiler[2], Jochen Meyer[3,4], Andreas von Deimling[3,4], Frank Westermann[4], Daniel Speidel[5,6], Mark N. Wass[1#], Martin Michaelis[1#], Jindrich Cinatl jr.[1#]

[1] School of Biosciences, University of Kent, Canterbury, UK

[2] Institute for Medical Virology, Goethe-University, Frankfurt am Main, Germany

[3] Department of Neuropathology, Ruprecht-Karls-University, Heidelberg, Germany

[4] Deutsches Krebsforschungszentrum, 69120 Heidelberg, Germany

[5] Children's Medical Research Institute, Westmead, Australia

[6] Sydney Medical School, The University of Sydney, Sydney, Australia

Corresponding authors: Mark N. Wass (M.N.Wass@kent.ac.uk), Martin Michaelis (M.Michaelis@kent.ac.uk), Jindrich Cinatl jr. (Cinatl@em.uni-frankfurt.de)

4.1 My contribution

I performed all analysis of the exome sequencing data for these cell lines, including setting up the computational pipeline that was used and generating the figures. This covers section 'Variant calling' from 'Materials and Methods'. I wrote the paper's first draft, which I then worked on with Mark Wass and Martin Michaelis. This manuscript is being prepared to go through the peer-review process for publication in a journal.

4.2 Abstract

Here, we characterised a unique panel of 41 sublines of the neuroblastoma cell line UKF-NB-3 adapted to four tubulin-binding agents, vincristine (10 sublines), eribulin (12 sublines), 2-methoxyestradiol (10 sublines), or epothilone B (9 sublines), by whole-exome sequencing. These drugs bind tubulin and either prevent polymerisation (destabilising agents) or depolymerisation (stabilising agents) of microtubules. This results in failing the mitotic spindle assembly checkpoint and cell cycle arrest. We looked at sequence variants that were called in the resistant sublines but not detectable in UKF-NB-3 sublines to gain insights into the clonal relatedness of the individual sublines, as same variants are unlikely to occur repeatedly by chance and are more likely to indicate the presence of alleles that are very rare in the parental cell line and, hence, a common clonal origin. Our findings indicate that drugs do not consistently select a certain pre-existing clone during the resistance formation process. Sublines adapted to the same drug did not always cluster together with regard to the rare variants. The epothilone B-adapted UKF-NB-3 subline UKF-NB-3^{r}EPOB^{2nM}III was more closely related to three vincristine-adapted UKF-NB-3 sublines than to other epothilone B resistant sublines, although epothilone B is a stabilising agent that interacts with the taxoid domain of tubulin and vincristine a destabilising agent targeting the vinca domain. Similarly, the 2-methoxyestradiol-adapted UKF-NB-3 subline UKF-NB-3^{r}2ME$^{2\mu M}$VIII was closely related to all epothilone-adapted sublines but UKF-NB-3^{r}EPOB^{2nM}III, although 2-methoxyestradiol is a destabilising agent that binds to the colchicine domain. In conclusion, our study provides initial evidence that different subpopulations in a cancer cell line can be selected upon repeated adaptation of the same cancer cell line to the same drug, illustrating the unpredictability of resistance formation processes. Adapting the same cancer cell line to four tubulin-binding agents in around ten independent experiments generated forty-one different responses to the drugs. This analysis offers a unique insight into acquired drug resistance as it could never be performed in a clinical setting – one patient cannot be treated forty-one times. Our analysis also could not be reproduced by applying the same treatment to forty-one patients as the heterogeneity of responses, in this case, would be a sum of intra- and inter-tumour heterogeneity. A better understanding of these processes will be a prerequisite for the development of rationally designed, individualised cancer therapies.

This knowledge could be utilised in designing new biomarkers and continuous monitoring of patients through liquid biopsies.

4.3 Introduction

Cure rates remain low for most cancers that are diagnosed at an advanced stage such as metastatic disease and require systemic treatment. The main reason for this is the formation of resistance (Harbeck and Gnant, 2017; Iacobucci and Mullighan, 2017; Litwin and Tan, 2017; Fenton *et al.*, 2018; Herbst, Morgensztern and Boshoff, 2018; Michaelis, Wass and Cinatl, 2019). Resistance can be 'intrinsic', i.e. a cancer does not respond to a therapy from the outset. However, cancer diseases often respond initially well to therapy, but eventually 'acquired' resistance emerges (DeVita and Chu, 2008; Holohan *et al.*, 2013; Fenton *et al.*, 2018; Soverini *et al.*, 2018; Michaelis, Wass and Cinatl, 2019). The mechanisms underlying intrinsic and acquired resistance differ (Esposito *et al.*, 2013; Arena *et al.*, 2015; Miklos *et al.*, 2015; Carter *et al.*, 2017; Onafuye *et al.*, 2019).

Preclinical model systems are needed to study acquired resistance, because they enable the acquisition of data that cannot be derived from clinical samples. For example, the repeated adaptation of a given cancer cell population is not possible in a clinical setting, where every patient can only be treated once (Michaelis, Wass and Cinatl, 2019). Drug-adapted cancer cell lines enabled the discovery of major drug resistance mechanisms such as ABCB1 (also known as MDR1 or P-glycoprotein) and ABCC1 (also known as MRP1) (Juliano and Ling, 1976; Cole *et al.*, 1992) and have been shown to reflect clinical resistance mechanisms on many further occasions (Michaelis, Wass and Cinatl, 2019).

It is anticipated that the processes underlying acquired resistance formation are subject to a heterogeneity at a similar scale (Sequist *et al.*, 2011; Basile *et al.*, 2013; Kemper *et al.*, 2015; Soucheray *et al.*, 2015; Hata *et al.*, 2016; Michaelis, Wass and Cinatl, 2019) as that is generally observed in cancer (McGranahan and Swanton, 2017). Recent findings indicated that the repeated adaptation of the same neuroblastoma cell line to the same drugs results in phenotypically different sublines (Michaelis, Wass, *et al.*, 2020). Even long-term treatment of acute myeloid leukaemia cells with an ineffective concentration of the MDM2 inhibitor nutlin-3 resulted in modified drug sensitivity profiles in the resulting sublines, although it did not result in increased nutlin-3 resistance (Michaelis, Rothweiler, *et al.*, 2020).

To further investigate the variability of the resistance formation process, we here established 41 sublines of the neuroblastoma cell line UKF-NB-3 (Kotchetkov *et al.*, 2005) with acquired resistance to the tubulin-binding agents vincristine (10 sublines), eribulin (12 sublines), 2-methoxyestradiol (10 sublines), and epothilone B (9 sublines).

The four selected tubulin-binding agents represent the major subgroups of this drug class whose members interfere with tubulin dynamics by different mechanisms (Dumontet and Jordan, 2010; Kavallaris, 2010). Vincristine and eribulin are so-called destabilising agents that bind to the vinca domain of tubulin and inhibit at high concentrations microtubule polymerisation (Dumontet and Jordan, 2010; Kavallaris, 2010). 2-methoxyestradiol is another destabilising agent, but it binds to an alternative tubulin domain, the colchicine domain. Epothilone B belongs to the stabilising tubulin-binding agents that enhance microtubule polymerisation by binding to the taxoid domain (Dumontet and Jordan, 2010; Kavallaris, 2010).

The project cell lines were characterised by whole-exome sequencing to gain further insights into the processes underlying resistance formation.

4.4 Materials and Methods

4.4.1 Cells

The MYCN-amplified neuroblastoma cell line UKF-NB-3 and its tubulin-binding agent-adapted sublines were derived from the resistant cancer cell line (RCCL) collection (www.kent.ac.uk/stms/cmp/RCCL/RCCLabout.html) (Michaelis, Wass and Cinatl, 2019). The resistant sublines were established by adaption to growth in the presence of stepwise increasing drug concentrations as previously described (Kotchetkov *et al.*, 2005; Michaelis *et al.*, 2011).

All cells were propagated in IMDM supplemented with 10 % FBS, 100 IU/ml penicillin and 100 mg/ml streptomycin at 37°C. Cells were routinely tested for mycoplasma contamination and authenticated by short tandem repeat profiling.

4.4.2 Viability assay

Cell viability was tested by the 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide (MTT) dye reduction assay after 120 h incubation modified after Mosmann (Mosmann, 1983) as described previously (Onafuye *et al.*, 2019). Results are expressed as mean ± S.D. of at least three experiments.

4.4.3 Whole-exome sequencing

Whole-exome sequencing was performed with Illumina HiSeq2000 using paired-end reads of a length of 100 base pairs. Exome enrichment was conducted using Nextera Exome Enrichment Kit.

4.4.4 Variant calling

Our pipeline for the calling of variants from whole-exome sequencing data was based on GATK best practices for discovering short nucleotide variants (SNVs) and INDELs (Van der Auwera *et al.*, 2013; *Germline short variant discovery (SNPs + Indels) – GATK*, 2019) and is summarised in Figure S12. After initial quality control using FastQC (Andrews, 2010), reads were trimmed with Trimmomatic-0.38 (default parameters) (Bolger, Lohse and Usadel, 2014) and mapped onto the reference genome (version hg19) using the Burrows-Wheeler Alignment Tool with the algorithm bwa-0.7.17-mem (Li and Durbin, 2009). Duplicate PCR reads were marked and .bam files built using Picard-2.17.10 (*Picard Tools - By Broad Institute*, 2019). GenomeAnalysisTK-3.7.0 (McKenna *et al.*, 2010) was used to realign sequences around insertions/ deletions and to recalibrate base scores. The machine learning model of covariation was built using dbSNP database (Sherry *et al.*, 2001) (downloaded on 23rd of April 2018) as known sites. Variants were called with samtools-1.7 mpileup (Li, 2011) using default parameters. Phred quality score filters were set to 30 with bcftools-1.6 (Li and Durbin, 2009) and variants with base call coverage below 10 and variant call coverage below 3 were removed. Variants that affected protein sequences were identified with VEP (release 96) (McLaren *et al.*, 2016) and categorised following the Sequence Ontology nomenclature (Eilbeck *et al.*, 2005). We considered frameshift, stop-gained, splice acceptor, splice donor, incomplete terminal codon, stop-lost, start-lost, missense, inframe insertion, and inframe deletion variants. Germline-like variants with a frequency ≥0.001% in gnomAD (Lek *et al.*, 2016) were only included if at least three sequences were recorded in TCGA and at least ten in COSMIC (Tate *et al.*, 2019). The potential impact on protein structure/ function was estimated by SIFT (Kumar, Henikoff and Ng, 2009) and PolyPhen-2 (Adzhubei *et al.*, 2010).

4.5 Results

4.5.1 Resistance status of investigated cell line

First, we confirmed the resistance status of the investigated cell lines (Table S12). The vincristine-adapted UKF-NB-3 sublines were between 6.3 and 45 times more resistant to vincristine than UKF-NB-3. The eribulin-adapted sublines displayed resistance factors between 222 and 1900, the 2-methoxyestradiol-adapted sublines between 7.6 and 12.7, and the epothilone-adapted sublines between 4.3 and 6.8 (Table S12).

4.5.2 Number of sequence variants with potential impact per cell line

For our analysis, we considered variants that are likely to have an impact using the following criteria: high-quality (Phred score ≥ 30), high-coverage (≥ 10 alleles covered, variant covered by ≥ three reads), somatic-like (either frequency in gnomAD below 0.001% or at least 10 samples with this variant in COSMIC (Tate *et al.*, 2019) and three in the TCGA (https://www.cancer.gov/tcga)), and most-likely damaging variants (Sequence Ontology (Eilbeck *et al.*, 2005) terms frameshift variant, stop-gained, splice acceptor variant, splice donor variant, incomplete terminal codon variant, stop-lost, start-lost, missense variant, inframe insertion inframe deletion were included).

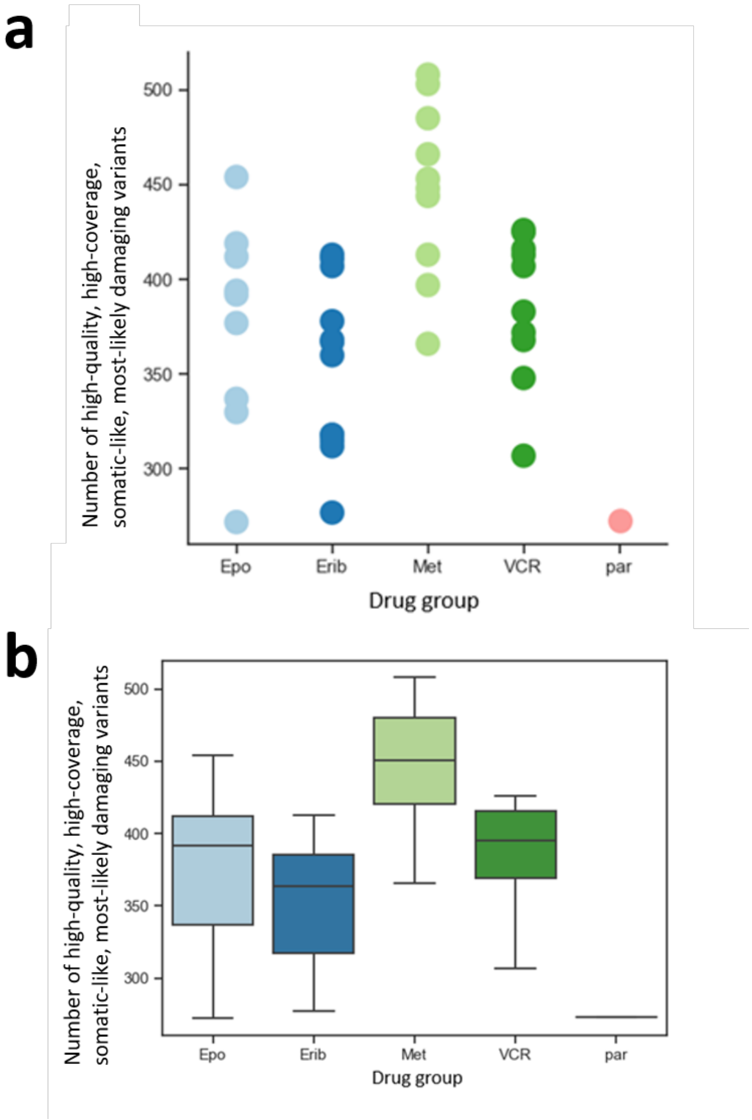


Figure 4.1 Number of high-quality, high-coverage, somatic-like variants with most-likely damaging impact on the protein structure per the same drug type. *The results plotted in the form of* ***a*** *a scatterplot* ***b*** *a boxplot. Epo = UKF-NB-3 sub-lines adapted to epothilone b; Erib = eribulin; Met = 2-methoxyestradiol; VCR = vincristine; par = UKF-NB-3 parental cell line.*

Using these criteria, we identified 2529 variants that are present in at least one of the 41 UKF-NB-3 sublines adapted to tubulin-binding agents. The cell line with the lowest number of variants was UKF-NB-3^{r}EPOB^{2nM}III (272), the one with the highest number was UKF-NB-3^{r}2ME$^{2\mu M}$V (508) (Figure 4.1).

4.5.3 Hierarchical clustering analysis

Hierarchical clustering analysis based on the shared variants indicated that clusters consist of sublines adapted to different drugs, although there is a tendency that sublines adapted to the same drug cluster together, in particular the 2-methoxyestradiol-resistant sublines (Figure 4.2a).

4.5.4 Variants not present in UKF-NB-3

We only considered variants as *de novo* variants, if the alignment of the exome sequencing data from UKF-NB-3 against the reference genome did not harbour any reads with the respective variant. According to this criterion, 1873 out of the 2529 variants were identified as *de novo* variants (Appendix 6 Supplementary Table 1). 1274 of these variants were only present in one subline, indicating that they newly occurred during resistance formation, whereas variants that are shared between different sublines probably represent variants that were already present in the UKF-NB-3 at a low frequency (Figure 4.2b).

Overlaps of unique *de novo* mutations between sublines adapted to the same drugs are shown in Figure 4.3. They revealed a particularly close relationship between UKF-NB-3^{r}ERI^{10}I and UKF-NB-3^{r}ERI^{10}IV (Figure 4.3a), but only few overlaps between UKF-NB-3^{r}EPOB^{2nM}III and the other epothilone B-resistant sublines (Figure 4.3c) and between UKF-NB-3^{r}2ME$^{2\mu M}$VIII and the other 2-methoxyestradiol-resistant sublines (Figure 4.3d).

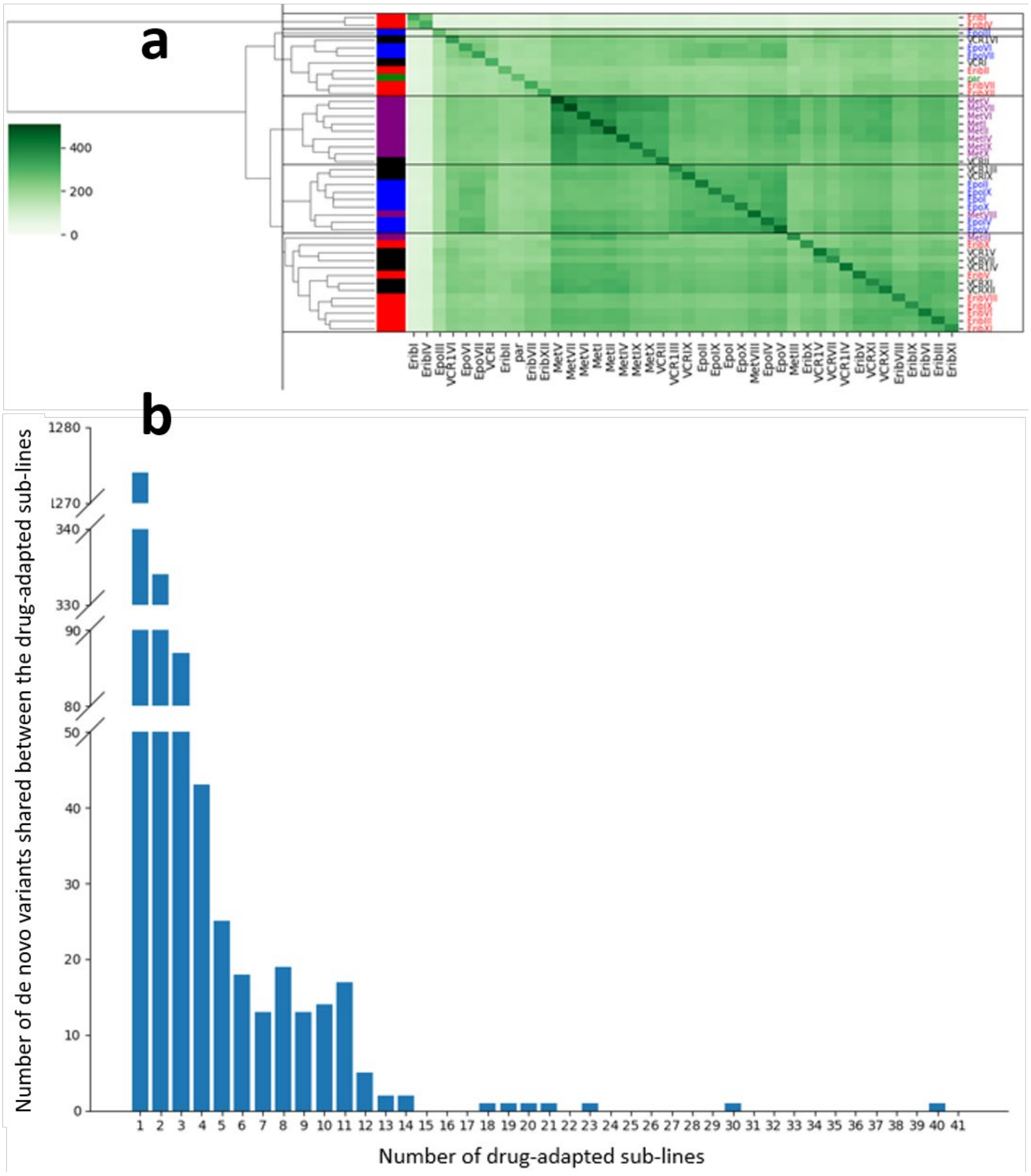


***Figure 4.2 a** Drug-adapted sub-lines hierarchically clustered by the number of shared high-quality, high-coverage, somatic-like, most-likely damaging variants. **b** Number of de novo variants shared between the drug-adapted sub-lines per number of drug-adapted sub-lines. 1274 de novo variants are present in only one unique drug-adapted sub-line, only 1 de novo variant is shared by 40 drug-adapted sub-lines, there are no de novo variants shared by all 41 drug-adapted sub-lines.*

The analysis of the overlaps of *de novo* mutations across all sublines showed that UKF-NB-3^{r}EPOB^{2nM}III is much more closely related to UKF-NB-3^{r}VCR$^{0.5}$V, UKF-NB-3^{r}VCR$^{0.5}$VI, and UKF-NB-3^{r}VCR$^{0.5}$VII suggesting that these sublines are derived from a similar clone (Figure 4.3e).

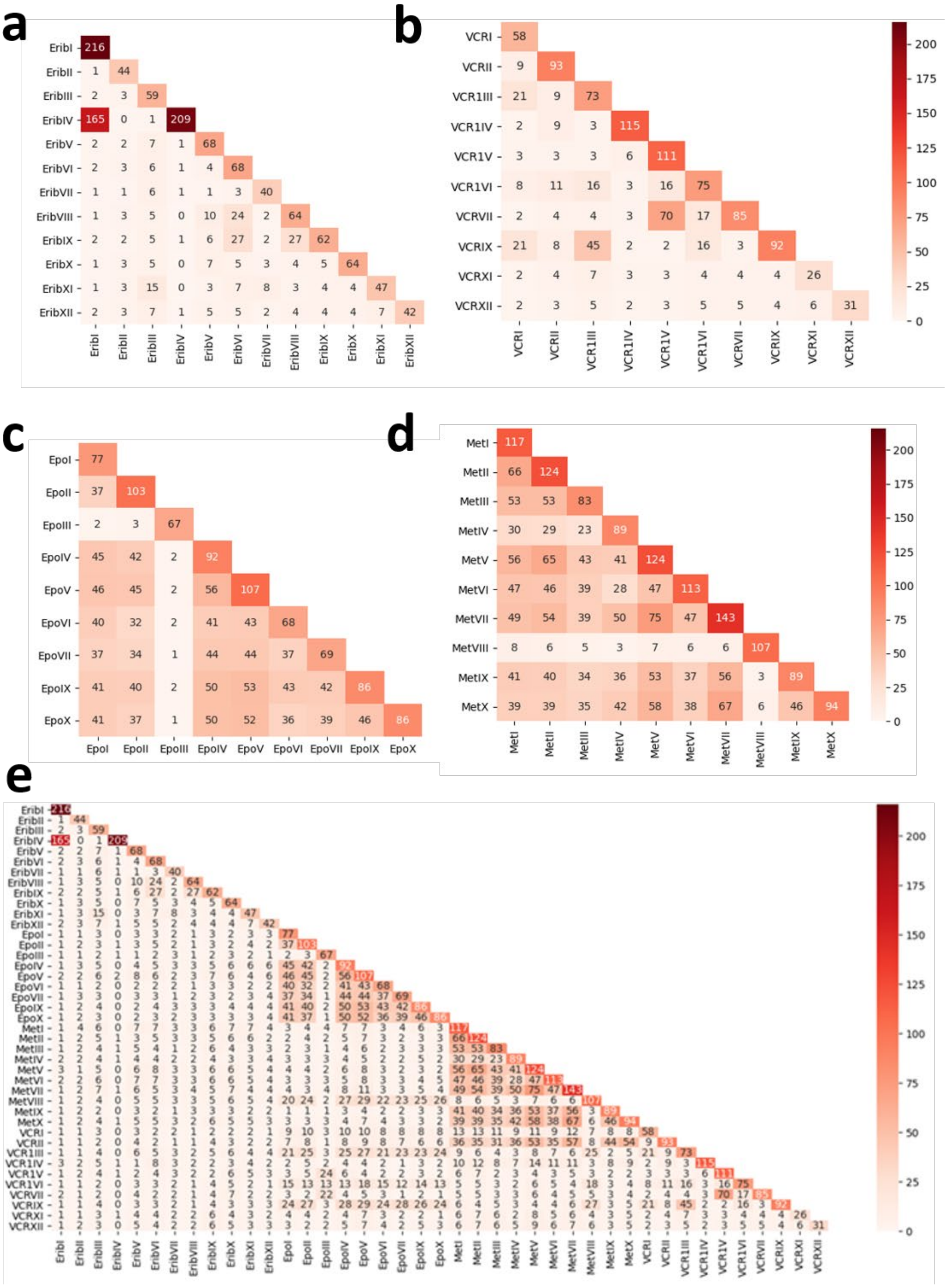


***Figure 4.3** Number of de novo variants shared by UKF-NB-3 sub-lines adapted to **a** eribulin **b** vincristine **c** epothilone b **d** 2-methoxyestradiol **e** all four tubulin-binding agents.*

UKF-NB-3^{r}2ME$^{2\mu M}$VIII is derived from a clone similar to those that all epothilone-adapted sublines are derived from, apart from UKF-NB-3^{r}EPOB^{2nM}III, the outlier in this group

(Figure 4.3e). This indicates that different subclones in the same cell line can give rise to a subpopulation resistant to a certain drug. Notably, repeated adaptation of a cell line to the same drug can result in the selection of different clones.

4.6 Discussion

Here, we performed an analysis of the changes associated with acquired resistance formation to tubulin-binding agents by investigating a unique panel of 41 sublines of the neuroblastoma cell line UKF-NB-3 adapted to vincristine (10 sublines), eribulin (12 sublines), 2-methoxyestradiol (10 sublines), or epothilone B (9 sublines).

We looked at variants that were called in the resistant sublines but were not detectable in the parental UKF-NB-3 sublines to see whether the sublines are derived from similar or distant clones. This approach is based on the assumption that exactly the same variant is unlikely to occur independently in different adaptation experiments. Hence, overlaps are more likely to indicate a common clonal origin. Since these variants are not detected in the parental cell line, they are anticipated to be rare and to provide a good indication of the clonal relatedness of sublines.

Interestingly, this analysis indicated that drugs do not consistently select a certain pre-existing clone during the resistance formation process. Although, there was a tendency that sublines adapted to the same drug share more rare variants than sublines adapted to different drugs, but this was not always the case. In particular, the vincristine-resistant sublines did not consistently cluster together, when we looked at rare variants not detectable in the parental cell line UKF-NB-3. This may be an example of convergent evolution, where different sublines acquired the same phenotype – resistance to vincristine – in a distinct way (Fortunato *et al.*, 2017; Konieczkowski, Johannessen and Garraway, 2018; Pienta *et al.*, 2020).

Sublines adapted to tubulin-binding agents with different modes of action also clustered together. The epothilone B-adapted UKF-NB-3 subline UKF-NB-3^{r}EPOB^{2nM}III was more closely related to the vincristine-adapted sublines UKF-NB-3^{r}VCR$^{0.5}$V, UKF-NB-3^{r}VCR$^{0.5}$VI, and UKF-NB-3^{r}VCR$^{0.5}$VII. This suggests that these sublines are derived from related clones although epothilone B is a stabilising agent that interacts with the taxoid domain on tubulin and vincristine a destabilising agent targeting the vinca domain (Dumontet and Jordan, 2010; Kavallaris, 2010).

Similarly, the 2-methoxyestradiol-adapted UKF-NB-3 subline UKF-NB-3^{r}2ME$^{2\mu M}$VIII is derived from a clone similar to all epothilone-adapted sublines but the previously mentioned UKF-NB-3^{r}EPOB^{2nM}III subline. Again, 2-methoxyestradiol and epothilone B target different tubulin structures, 2-methoxyestradiol as destabilising agent the colchicine domain and epothilone B as stabilising agent the taxoid domain (Dumontet and Jordan, 2010; Kavallaris, 2010).

The processes underlying resistance acquisition in cancer had previously been anticipated to be complex. It is anticipated that the processes underlying acquired resistance formation are subject to a heterogeneity at a similar scale (Sequist *et al.*, 2011; Basile *et al.*, 2013; Kemper *et al.*, 2015; Soucheray *et al.*, 2015; Hata *et al.*, 2016; Michaelis, Wass and Cinatl, 2019; Michaelis, Wass, *et al.*, 2020) as commonly observed in cancer (McGranahan and Swanton, 2017). However, the observation that the repeated adaptation of a cell line to the same drug, which is still relatively homogeneous compared to a patient tumour, results in the selection of different clones adds additional layer of complexity to this picture and stresses the need for biomarkers and effective monitoring methods for drug-induced evolution in cancer cell lines such as liquid biopsies (Heidrich *et al.*, 2020) for the development of more effective, individualised therapies.

In conclusion, our study provides initial evidence that different subpopulations in a cancer cell line can be selected upon repeated adaptation of the same cancer cell line to the same drug. A better understanding of these processes will be a prerequisite for the development of rationally designed, individualised cancer therapies.

Future studies should be focused on functional analysis of the mutations, resembling the analysis of the mutations acquired due to treatment of acute myeloid leukaemia cells presented in Chapter 3. This should include selecting candidates for TBAs-resistance drivers using criteria such as the presence of the mutations in databases of cancer drivers such as IntOGen and COSMIC and automated prediction of the variants’ impact on the protein structure and function. Literature mining should then be performed to find further evidence supporting a potential role of the selected mutations in resistance to TBAs in UKF-NB-3 sublines. Future analysis should also include a manual assessment of the effect of the mutations. This could involve tools used to identify the functions of the minimal genome proteins in Chapter 2. Pfam domains should be predicted to verify if the mutations could have a damaging impact on the function of the domains. Similarly, the analysis should include structural modelling using state-of-the-art protein structure prediction tools, e.g. AlphaFold2 (Jumper et al., 2021) (see 2.6.1) and assessing the effect that these mutations could have on the structure and ligand binding. Special focus should be put on any mutations acquired in tubulin as the four tubulin-binding agents studied here interact with beta-tubulin, and mutations in this protein have been known to be associated with resistance to this group of anti-cancer drugs (Huzil *et al.*, 2007; Kavallaris, 2010). This should include modelling changes in tubulin structures and verifying how that affects tubulin-binding drug interactions.

4.7 Acknowledgements

The results <published or shown> here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

4.8 Grant support

The work was supported by the Hilfe für krebskranke Kinder Frankfurt e.V., the Frankfurter Stiftung für krebskranke Kinder, and the Kent Cancer Trust.

Chapter 5 General discussion

5.1 Understanding the functions of proteins

The power of relatively inexpensive and fast next-generation sequencing has flooded us with more protein sequences than we can handle. At the start of the work described in Chapter 2 (June 2017), the UniProt-GOA database (Huntley *et al.*, 2015), which provides manual and electronic Gene Ontology annotations for the proteins from the UniProtKB database, contained GO annotations for over 60 million protein sequences. However, only 0.2% of them were annotated with at least one GO term assigned with an experimental evidence code (see (*Guide to GO evidence codes*, 2020) for the full list of experimental evidence codes currently available in the UniProt-GOA project). In April 2019, the number of sequences was higher than in June 2017 by 64% (over 99 million), but the percentage of sequences with experimental annotations dropped to 0.13%. Finally, in October 2020, a year and a half later, the number of sequences in the UniProt-GOA database crossed 133 million, with only 0.1% of them being experimentally annotated.

One of the aims of creating a minimal bacterial cell is to expand knowledge of protein function by identifying functions essential to sustain life. Computational annotations of the genes in the minimal bacterial cell performed by Hutchison et al. (2016) combined with our predictions presented in Chapter 2 revealed that the most fundamental life processes are preserving and expressing genetic information (nearly 50% of the genes perform these functions). We also identified 24 transporters among the proteins that were previously labelled membrane or hypothetical. This resulted in proteins involved in membrane-related processes comprising 22% of the minimal bacterial genome. The remaining proteins were either responsible for cell metabolism, or the biological process that they were involved in could not be predicted confidently.

The sixteen proteins for which we could not predict a function and which remained annotated as hypothetical lacked homologues in other species. As a result, the majority of the methods, relying largely on homology-based annotation transfer, failed to return any functional information. In such difficult cases, we relied on the results from TMHMM (a hidden Markov model representing a membrane protein topology) (Krogh *et al.*, 2001) and FFPred (a machine learning method with features based on protein biophysical properties, see Chapter 2 for details) (Cozzetto *et al.*, 2016). However, for these sixteen proteins, even these two methods were not helpful as they did not provide consistent results. Our efforts to predict functions of the proteins in the minimal genome that do not have homologues in other species and the lack of satisfying results in this matter (Chapter 2) demonstrate an urgent need to focus the protein function prediction field on bridging this gap. The current possibilities for predicting functions of proteins not having confident homologues in other species are very limited. This is represented by the methods submitted in the second edition of CAFA (Jiang *et al.*, 2016): only 6% of them were assigned a keyword “de novo” as a description. This poses quite a conundrum for the orphan genes in newly studied species. We do not know anything about them, and at the same time, we do not possess good-enough tools to guide experiments to change that.

5.2 Predicting essential functions required for life within the minimal bacterial genome

According to (Martínez-García and de Lorenzo, 2016), quite a few bacterial genomes have been subjected to the attempts to reduce them. The list starts, as expected, with the most studied bacterium, *Escherichia coli* (Kolisnychenko *et al.*, 2002; Hashimoto *et al.*, 2005; Xue *et al.*, 2015; Zhou *et al.*, 2016), followed by two Mycoplasma species: *Mycoplasma genitalium* (Gibson *et al.*, 2010) and *Mycoplasma mycoides* (Hutchison *et al.*, 2016), and then *Bacillus subtillis* (Westers *et al.*, 2003; Li *et al.*, 2016), *Corynebacterium glutamicum* (Unthan *et al.*, 2015), *Pseudomonas putida* (Leprince *et al.*, 2012; Lieder *et al.*, 2015), *Streptomyces avermitilis* (Komatsu *et al.*, 2010; Ikeda, Shin-Ya and Omura, 2014) and *Vibrio natriegens* (Weinstock *et al.*, 2016). Scientists have been working on techniques to reduce bacterial genomes to fulfil two primary goals (Martínez-García and de Lorenzo, 2016). The first one is to minimise the genes to only those that are essential for life. This way, fundamental life processes can be identified and studied. The second aim of reducing bacterial genomes comes from synthetic biology. It is to create a bacterium that will be easy to engineer and use for biotherapy, biofuels or biomaterials (Wang and Zhang, 2019). Following that aim, some of the non-essential genes, for example, those related to faster growth, may be retained.

Our efforts to identify the unknown functions of the life-essential genes in *Mycoplasma mycoides* were described in Chapter 2 of this thesis. The genome of *Mycoplasma mycoides* was minimised and synthesised by J. *Craig Venter Institute* (Hutchison *et al.*, 2016) and resulted in 473 genes. Hutchison et al. used TIGRFAM families (Haft *et al.*, 2013), threading to known structures and genomic context methods to annotate the genes, and they discovered that 149 could not be confidently annotated using these methods and they remain of unknown function (Hutchison *et al.*, 2016). In our analysis, we expanded the scope of the computational tools to 22, including methods modelling protein structure, identifying domains, transmembrane regions, predicting orthologues, ligand binding and also Gene Ontology terms (Antczak, Michaelis and Wass, 2019).

Our method has demonstrated to be effective in predicting genes of unknown functions, including those with minimal information such as a lack of orthologues in other species, matches to known domains or structural models. This is consistent with the results of CAFA where many of the best-performing methods are those integrating multiple resources and various aspects of protein function as an input for prediction (Zhou *et al.*, 2019).

By combining all the results and their manual curation, we made confident predictions for 133 out of 149 genes with previously unknown functions, including 66 that were more informative than those inferred by Hutchison et al. (2016). However, for the remaining sixteen proteins, we either did not obtain any confident or, in some cases, any results from the tools, or the results were not consistent enough to make predictions.

5.3 Development of drug resistance

Despite increasing knowledge about cancer and with many anti-cancer therapies available to treat patients, developing resistance to drugs that initially caused an improvement remains a significant obstacle on the way to recovery. There are multiple molecular mechanisms whose alterations contribute to the development of the resistance. ABC transporters such as MDR1 (coding for Multidrug resistance protein 1/P-glycoprotein 1), MRP1 (Multidrug resistance-associated protein 1) or BCRP (Broad substrate specificity ATP-binding cassette transporter ABCG2) are responsible, for example, for removing toxins from the cell. However, they have broad substrate specificity, and their high expression may result in the efflux of multiple drugs, such as vinca alkaloids or taxanes, out of the cell (Housman *et al.*, 2014).

Another mechanism causing cells to acquire tolerance to an anti-cancer drug is a decrease in the activation of the drug. For example, cytarabine is activated through multiple phosphorylation events. If the phosphorylation pathway is altered, it may result in the reduced activation of cytarabine (Housman *et al.*, 2014). Resistance may also develop through targets of the drugs acquiring mutations in them; for example, a mutation in beta-tubulin may alter the response to taxanes or vinca alkaloids (Housman *et al.*, 2014). Finally, tolerance to anti-cancer drugs may be caused by a modification of expression levels of specific genes related to DNA damage response or apoptosis. For example, while over-expression of DNA excision repair protein ERCC-1 is associated with higher tolerance to the treatment of non-small-cell lung cancer with cisplatin (Ceppi *et al.*, 2006), up-regulation of apoptosis regulator Bcl-2 may result in an increase in suppressing of apoptosis and hence also cause a higher tolerance to the anti-cancer treatment (Mansoori *et al.*, 2017).

Studies have suggested that these molecular mechanisms promoting drug resistance are not likely to be a result of specific activities of a drug but instead may be rooted in genetic instability and intra-tumour heterogeneity (Mansoori *et al.*, 2017; Nikolaou *et al.*, 2018). Anti-cancer treatments kill only the cells that are sensitive to it, leaving the resistant cells with a possibility to grow and expand into a dominating clone (Housman *et al.*, 2014). Our results from Chapter 3 and Chapter 4 show the process in which Molm13 and UKF-NB-3 were adapted to nutlin-3 (Chapter 3) and four TBAs (Chapter 4), respectively. All four Molm13 sub-lines adapted to nutlin-3 harboured the same exact mutation in MUC2 for which no supporting reads were identified in the Molm13 parental cell line. It is unlikely that this mutation developed independently in all four sub-lines, and it rather suggests that the sub-lines were derived from a relatively rare subpopulation of cells from the Molm13 parental cell line.

This result is even more apparent in the UKF-NB-3 sub-lines adapted to tubulin-binding agents (Chapter 4). Our studies of 1873 de novo variants identified in at least one of the 41 UKF-NB-3 sub-lines adapted to eribulin (12), vincristine (10), 2-methoxyestradiol (10) and epothilone b (9) revealed that 1274 of the variants developed uniquely in one sub-line while 599 were shared by at least two sub-lines. These 599 variants are more likely to be variants present in the UKF-NB-3 parental cell line with low frequency than to have occurred independently in more than one sub-line. Our results also show that sub-lines adapted to the same drug often share a lot of variants, however this is not always the case.

This was exemplified by one of the sub-lines adapted to 2-methoxyestradiol, UKF-NB-3^{r}2ME$^{2\mu M}$VIII, sharing more variants with the epothilone-adapted sub-lines (expect UKF-NB-3^{r}EPOB^{2nM}III) than the other 2-methoxyestradiol-adapted sub-lines. In addition, the fact that epothilone b and 2-methoxyestradiol have a different mechanism of action (epothilone b is a stabilising agent targeting the taxoid domain of in beta-tubulin and 2-methoxyestradiol is a destabilisng agent targeting the colchicine domain) emphasises that the variants developed during the drug adaptation process do not always depend on the drug but also the development of the same variants cannot be expected based on the drug type. Our results demonstrate that different subpopulations can be selected upon repeated adaptation of the same cancer cell line to the same drug. They also emphasise the heterogeneity and complexity of processes underlying acquired resistance and stress the need for biomarkers to monitor patients.

Additionally, we applied various principles and tools of protein function prediction explored in Chapter 2 to study the acquired variants' impact on protein structure and function. This allowed us to identify potential loss-of-function mutations that could contribute to cancer cells developing decreased sensitivity to the treatment. In Chapter 3 and Chapter 4, we used SIFT (Sim *et al.*, 2012) and PolyPhen-2 (Adzhubei *et al.*, 2010) that classify missense variants into neutral and likely damaging. They do that by applying similar principles of protein function prediction that we used in Chapter 2 to annotate proteins of the minimal bacterial genome. Further, by combining the results from SIFT and PolyPhen-2 with the information we collected from IntOGen (Martínez-Jiménez, Muiños, Sentís, *et al.*, 2020) and COSMIC (Tate *et al.*, 2019), we identified cancer drivers among genes with SNVs and established several potential drug resistance mechanisms. To further analyse the impact of these mutations, we predicted Pfam (El-Gebali *et al.*, 2019) domains using the same methodology as applied in Chapter 2.

Several mutations reported to have a role in tumorigenesis of various cancer types were identified among de novo and gained variants of nutlin-3 adapted sub-lines and may contribute to acquired resistance to nutlin-3 (Chapter 3). This includes mutations in genes such as GNAQ, PABPC1, PPMD1, PER3, MT-ND5, MT-ND6, and most importantly, TP53 and SAMHD1. While the mechanisms of resistance involving GNAQ, PABPC1, PPMD1, PER3, MT-ND5 and MT-ND6 are yet to be investigated, our studies confirm the previously suggested impact that acquired TP53 mutations have on resistance to MDM2 inhibitors (Aziz, Shen and Maki, 2011; Michaelis *et al.*, 2011; Jones *et al.*, 2012; Cinatl *et al.*, 2014; Gianna Hoffman-Luca *et al.*, 2015; Drummond *et al.*, 2016; Jung *et al.*, 2016).

In addition, likely deleterious mutations identified in SAMHD1 and increased sensitivity to cytarabine demonstrate a potential correlation between these two events. This further supports the already suggested correlation between the low activity of SAMHD1 and sensitivity to cytarabine (Schneider *et al.*, 2017). However, even in the Molm13 sub-line which did not develop SAMHD1 mutations increased sensitivity to cytarabine was observed. These results indicate the possibility of prescribing cytarabine to patients that develop resistance to nutlin-3, especially in the case of detected loss of activity of SAMHD1.

Our findings can be used as a base for further research to verify novel biomarkers of resistance. Cancer patients treated with nutlin-3 and the four TBAs that we studied in Chapter 3 and Chapter 4 could then be monitored regularly for developing mutations in the suggested genes using gene panels. The increase of understanding of the drug adaptation process and the mechanisms of resistance described in Chapter 3 and Chapter 4 could also help in designing effective drug combinations.

Our results are limited by using only whole-exome sequencing data and thus reliable variant identification primarily in protein-coding regions. While whole-exome sequencing is informative, whole-genome sequencing would provide a fuller picture of the variation that occurs as part of the adaptation process, although such non-coding variation is typically much harder to interpret. Further, this research would be strengthened by the use of multiple ‘omics methods such as transcriptomics (e.g. RNA sequencing) to measure gene expression levels and epigenomics to study DNA methylation patterns.

Our studies using whole-exome sequencing data and cancer cell lines as a cancer model allow cost-effective identification of potential drug resistance mechanisms before the drug is given to patients. In the future, this research could be expanded by monitoring the drug adaptation process by sequencing cell lines at different time points during the adaptation process. That would enable determining how the resistant sub-clones emerge and possibly pinpoint sub-clones that initiated resistance. The knowledge gained from these studies could then be applied in the clinic. The patient’s genetic profile could be assessed at first so that anti-cancer drugs associated with the smallest probability of resistance for that genetic profile could be prescribed. Then regular monitoring of any changes in the mutations (acquiring new alterations and an increase or decrease in the variant allele frequency of the already existing ones) could be monitored to predict the next cancer move and make any adjustments necessary to retain the effectiveness of the anti-cancer therapy. With the development of liquid biopsies, it could be possible to monitor mutations through analysis of circulating tumour cells and DNA.

Finally, our studies from Chapter 3 and Chapter 4 demonstrate how important it is to continue increasing the knowledge about the functions of the proteins. For example, from all the genes carrying somatic-like mutations that have possibly damaging impact on the protein function in the Molm13 parental cell line and its nutlin3-adapted sub-lines (overall 423 genes), for three of them, no functional information is present neither in the Ensembl (Yates *et al.*, 2020) nor UniProt database. Two additional genes have protein sequence entries in UniProt that are annotated with a few Gene Ontology terms. However, this information has not been reviewed by a curator. Finally, the functions of six additional proteins that have been curated and are available in SwissProt do not say more than “uncharacterised protein”. Drug-adapted sub-lines have potentially deleterious mutations in those eleven genes, but we do not know what they do. This exemplifies the urgency of developing accurate protein function prediction methods that will guide experimental work and will allow uncovering the remaining functional mysteries. This knowledge will help understand diseases, including cancer and drug resistance, better and let design more suitable treatment.

5.4 Future work

The research in this thesis could be complemented by applying current state-of-the-art methods of protein structure prediction such as AlphaFold2 (Jumper et al., 2021) (see 2.6.1) - a template-free method that explores deep learning and achieved unprecedented quality and coverage of structural modelling in the last edition of Critical Assessment of Protein Structure Prediction (*Home - CASP14*, 2021). AlphaFold2 could hopefully improve solving the structures of the minimal genome proteins which lacked orthologs in other species and thus could not be modelled using the single-template method that we applied to identify unknown functions of the minimal genome in Chapter 2. For those proteins, we could not conclude any functional information, and they remained annotated as hypothetical. Knowing their 3D structure could contribute to identifying their function. We could model protein-protein and protein-ligand interactions by applying molecular docking (Northey, Barešić and Martin, 2018; Salmaso and Moro, 2018) and then apply molecular dynamics simulations to examine the nature of those interactions (Hospital *et al.*, 2015).

Future analysis should also include applying AlphaFold2 to perform structural modelling of proteins that are candidates for drivers of resistance in cancer cell lines adapted to nutlin-3 and tubulin-binding agents that we studied in Chapter 3 and Chapter 4. Structural models could be used to assess the effect that the mutations would have on the structure and function. This would facilitate identifying loss-of-function mutations that could be prioritised for experimental validation.

Finally, the protein function prediction method that we designed and used in Chapter 2 to identify unknown functions in the minimal genome is only partially automated. Even though the information about proteins could be collected from separate tools through web servers or desktop applications, predictions were made by manual curation of all the results. Thus, the method would have to be automated to be suitable for annotating a larger number of genes. This could be done by selecting features from the 22 sources of data that we used and training a machine or deep learning algorithm to make predictions instead of relying on the expertise of a researcher. The automated method could be used to annotate proteins in other minimal genomes and identify the common set of essential functions, which would expand our knowledge of what is necessary to sustain life.

References

About GOA | European Bioinformatics Institute (2020). Available at: https://www.ebi.ac.uk/GOA/newto (Accessed: 13 October 2020).

About the GDC | NCI Genomic Data Commons (2020). Available at: https://gdc.cancer.gov/about-gdc (Accessed: 3 November 2020).

Adzhubei, I. A. *et al.* (2010) 'A method and server for predicting damaging missense mutations', *Nature Methods*. Nat Methods, pp. 248–249. doi: 10.1038/nmeth0410-248.

Aerts, S. *et al.* (2006) 'Gene prioritization through genomic data fusion', *Nature Biotechnology*. doi: 10.1038/nbt1203.

Akiva, E. *et al.* (2014) 'The Structure-Function Linkage Database', *Nucleic Acids Research*, 42(D1), pp. D521–D530. doi: 10.1093/nar/gkt1130.

Alexandrov, L. B. and Stratton, M. R. (2014) 'Mutational signatures: The patterns of somatic mutations hidden in cancer genomes', *Current Opinion in Genetics and Development*. doi: 10.1016/j.gde.2013.11.014.

Alm, E. J. *et al.* (2005) 'The MicrobesOnline Web site for comparative genomics', *Genome Research*, 15(7), pp. 1015–1022. doi: 10.1101/gr.3844805.

Altschul, S. F. *et al.* (1997) 'Gapped BLAST and PSI-BLAST: A new generation of protein database search programs', *Nucleic Acids Research*. Nucleic Acids Res, pp. 3389–3402. doi: 10.1093/nar/25.17.3389.

Andrews, S. (2010) *FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]*. Available at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (Accessed: 19 December 2020).

Antczak, M., Michaelis, M. and Wass, M. N. (2019) 'Environmental conditions shape the nature of a minimal bacterial genome', *Nature Communications*, 10(1), pp. 1–13. doi: 10.1038/s41467-019-10837-2.

Arena, S. *et al.* (2015) 'Emergence of multiple EGFR extracellular mutations during cetuximab treatment in colorectal cancer', *Clinical Cancer Research*. doi: 10.1158/1078-0432.CCR-14-2821.

Arnedo-Pac, C. *et al.* (2019) 'OncodriveCLUSTL: A sequence-based clustering method to identify cancer drivers', *Bioinformatics*. doi: 10.1093/bioinformatics/btz501.

Ashburner, M. *et al.* (2000) 'Gene ontology: Tool for the unification of biology', *Nature Genetics*, 25(1), pp. 25–29. doi: 10.1038/75556.

Attwood, T. (2012) 'The PRINTS database: a fine-grained protein sequence annotation and analysis resource–its status in 2012', *Database (Oxf.)*, 2012, p. bas019.

Van der Auwera, G. A. *et al.* (2013) 'From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline', *Current Protocols in Bioinformatics*. doi: 10.1002/0471250953.bi1110s43.

Aziz, M. H., Shen, H. and Maki, C. G. (2011) 'Acquisition of p53 mutations in response to the non-genotoxic p53 activator Nutlin-3', *Oncogene*. doi: 10.1038/onc.2011.185.

Bamford, S. *et al.* (2004) 'The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website', *British Journal of Cancer*, 91(2), pp. 355–358. doi: 10.1038/sj.bjc.6601894.

Basile, K. J. *et al.* (2013) 'In vivo MAPK reporting reveals the heterogeneity in tumoral selection of resistance to RAF inhibitors', *Cancer Research*. doi: 10.1158/0008-5472.CAN-13-1628.

Bateman, A. *et al.* (2017) 'UniProt: The universal protein knowledgebase', *Nucleic Acids Research*, 45(D1), pp. D158–D169. doi: 10.1093/nar/gkw1099.

Bateman, A. (2019) 'UniProt: A worldwide hub of protein knowledge', *Nucleic Acids Research*, 47(D1), pp. D506–D515. doi: 10.1093/nar/gky1049.

Behjati, S. and Tarpey, P. S. (2013) 'What is next generation sequencing?', *Archives of Disease in Childhood: Education and Practice Edition*, 98(6), pp. 236–238. doi: 10.1136/archdischild-2013-304340.

Benson, D. A. *et al.* (2013) 'GenBank', *Nucleic Acids Research*, 41(D1). doi: 10.1093/nar/gks1195.

Benson, D. A. *et al.* (2017) 'GenBank', *Nucleic Acids Research*, 45(D1), pp. D37–D42. doi: 10.1093/nar/gkw1070.

Berger, M. F. and Mardis, E. R. (2018) 'The emerging clinical relevance of genomics in cancer medicine', *Nature Reviews Clinical Oncology*, 15(6), pp. 353–365. doi: 10.1038/s41571-018-0002-6.

Berman, H. M. *et al.* (2000) 'The Protein Data Bank', *Nucleic Acids Research*. Oxford University Press, pp. 235–242. doi: 10.1093/nar/28.1.235.

Bernardes, J. and Pedreira, C. (2013) 'A Review of Protein Function Prediction Under Machine Learning Perspective', *Recent Patents on Biotechnology*, 7(2), pp. 122–141. doi: 10.2174/18722083113079990006.

Binkley, J. *et al.* (2010) 'ProPhylER: A curated online resource for protein function and structure based on evolutionary constraint analyses', *Genome Research*, 20(1), pp. 142–154. doi: 10.1101/gr.097121.109.

Binns, D. *et al.* (2009) 'QuickGO: A web-based tool for Gene Ontology searching', *Bioinformatics*, 25(22), pp. 3045–3046. doi: 10.1093/bioinformatics/btp536.

Bolger, A. M., Lohse, M. and Usadel, B. (2014) 'Trimmomatic: A flexible trimmer for Illumina sequence data', *Bioinformatics*. doi: 10.1093/bioinformatics/btu170.

Bork, P. *et al.* (1998) 'Predicting function: From genes to genomes and back', *Journal of Molecular Biology*, 283(4), pp. 707–725. doi: 10.1006/jmbi.1998.2144.

Bouaoun, L. *et al.* (2016) 'TP53 Variations in Human Cancers: New Lessons from the IARC TP53 Database and Genomics Data', *Human Mutation*. doi: 10.1002/humu.23035.

Buenavista, M. T., Roche, D. B. and McGuffin, L. J. (2012) 'Improvement of 3D protein models using multiple templates guided by single-template model quality assessment', *Bioinformatics*, 28(14), pp. 1851–1857. doi: 10.1093/bioinformatics/bts292.

C-QUARK: Contact Assisted Ab Initio Protein Structure Prediction (2021). Available at: https://zhanggroup.org/C-QUARK/ (Accessed: 1 September 2021).

Camacho, C. *et al.* (2009) 'BLAST+: Architecture and applications', *BMC Bioinformatics*, 10(1), pp. 1–9. doi: 10.1186/1471-2105-10-421.

Campbell, P. J. *et al.* (2020) 'Pan-cancer analysis of whole genomes', *Nature*, 578(7793), pp. 82–93. doi: 10.1038/s41586-020-1969-6.

Carbon, S. *et al.* (2017) ‘Expansion of the gene ontology knowledgebase and resources: The gene ontology consortium’, *Nucleic Acids Research*, 45(D1), pp. D331–D338. doi: 10.1093/nar/gkw1108.

Carbon, S. *et al.* (2019) ‘The Gene Ontology Resource: 20 years and still GOing strong’, *Nucleic Acids Research*, 47(D1), pp. D330–D338. doi: 10.1093/nar/gky1055.

Carter, L. *et al.* (2017) ‘Molecular analysis of circulating tumor cells identifies distinct copy-number profiles in patients with chemosensitive and chemorefractory small-cell lung cancer’, *Nature Medicine*. doi: 10.1038/nm.4239.

Cassier, P. A. *et al.* (2017) ‘Targeting apoptosis in acute myeloid leukaemia’, *British Journal of Cancer*. doi: 10.1038/bjc.2017.281.

Ceppi, P. *et al.* (2006) ‘ERCC1 and RRM1 gene expressions but not EGFR are predictive of shorter survival in advanced non-small-cell lung cancer treated with cisplatin and gemcitabine’, *Annals of Oncology*, 17(12), pp. 1818–1825. doi: 10.1093/annonc/mdl300.

Chang, Y. C. *et al.* (2016) ‘COMBREX-DB: An experiment centered database of protein function: Knowledge, predictions and knowledge gaps’, *Nucleic Acids Research*, 44(D1), pp. D330–D335. doi: 10.1093/nar/gkv1324.

Chatterjee, N. and Walker, G. C. (2017) ‘Mechanisms of DNA damage, repair, and mutagenesis’, *Environmental and Molecular Mutagenesis*. John Wiley and Sons Inc., pp. 235–263. doi: 10.1002/em.22087.

Chen, S. T. *et al.* (2005) ‘Deregulated expression of the PER1, PER2 and PER3 genes in breast cancers’, *Carcinogenesis*. doi: 10.1093/carcin/bgi075.

Cheng, F., Zhao, J. and Zhao, Z. (2016) ‘Advances in computational approaches for prioritizing driver mutations and significantly mutated genes in cancer genomes’, *Briefings in Bioinformatics*, 17(4), pp. 642–656. doi: 10.1093/bib/bbv068.

Cheung, E. *et al.* (2019) ‘The leukemia strikes back: a review of pathogenesis and treatment of secondary AML’, *Annals of Hematology*. doi: 10.1007/s00277-019-03606-0.

Cinatl, J. *et al.* (2014) ‘Resistance acquisition to MDM2 inhibitors’, in *Biochemical Society Transactions*. doi: 10.1042/BST20140035.

Clancy, S. (2008) *Genetic Mutation | Learn Science at Scitable*, *Nature Education 1(1):187*.

Available at: https://www.nature.com/scitable/topicpage/genetic-mutation-441/ (Accessed: 7 September 2021).

Clark, W. T. and Radivojac, P. (2011) ‘Analysis of protein function and its prediction from amino acid sequence’, *Proteins: Structure, Function and Bioinformatics*, 79(7), pp. 2086–2096. doi: 10.1002/prot.23029.

Clifford, R. J. *et al.* (2004) ‘Large-scale analysis of non-synonymous coding region single nucleotide polymorphisms’, *Bioinformatics*, 20(7), pp. 1006–1014. doi: 10.1093/bioinformatics/bth029.

Climent, J. *et al.* (2010) ‘Deletion of the PER3 gene on chromosome 1p36 in recurrent ER-positive breast cancer’, *Journal of Clinical Oncology*. doi: 10.1200/JCO.2009.27.0215.

Cole, S. P. C. *et al.* (1992) ‘Overexpression of a transporter gene in a multidrug-resistant human lung cancer cell line’, *Science*. doi: 10.1126/science.1360704.

Condel — FannsDB 2.0-dev documentation (2014). Available at: https://bbglab.irbbarcelona.org/fannsdb/help/condel.html (Accessed: 30 October 2020).

Cornish-Bowden, A. (2014) ‘Current IUBMB recommendations on enzyme nomenclature and kinetics’, *Perspectives in Science*, 1(1–6), pp. 74–87. doi: 10.1016/j.pisc.2014.02.006.

Cozzetto, D. *et al.* (2016) ‘FFPred 3: Feature-based function prediction for all Gene Ontology domains’, *Scientific Reports*, 6(August), pp. 1–11. doi: 10.1038/srep31865.

Crystal, A. S. *et al.* (2014) ‘Patient-derived models of acquired resistance can identify effective drug combinations for cancer’, *Science*. doi: 10.1126/science.1254721.

Cui, Y. *et al.* (2019) ‘Predicting protein-ligand binding residues with deep convolutional neural networks’, *BMC Bioinformatics*, 20(1), p. 93. doi: 10.1186/s12859-019-2672-1.

Current Release Statistics < Uniprot < EMBL-EBI (2020). Available at: https://www.ebi.ac.uk/uniprot/TrEMBLstats (Accessed: 13 October 2020).

Dana, J. M. *et al.* (2019) ‘SIFTS: Updated Structure Integration with Function, Taxonomy and Sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins’, *Nucleic Acids Research*, 47(D1), pp. D482–D489. doi: 10.1093/nar/gky1114.

Danchin, A. and Fang, G. (2016) ‘Unknown unknowns: essential genes in quest for function’, *Microbial Biotechnology*, 9(5), pp. 530–540. doi: 10.1111/1751-7915.12384.

Das, S. *et al.* (2015) ‘Functional classification of CATH superfamilies: A domain-based approach for protein function annotation’, *Bioinformatics*, 31(21), pp. 3460–3467. doi: 10.1093/bioinformatics/btv398.

DeVita, V. T. and Chu, E. (2008) ‘A history of cancer chemotherapy’, *Cancer Research*. doi: 10.1158/0008-5472.CAN-07-6611.

Díaz-Gay, M. *et al.* (2018) ‘Mutational Signatures in Cancer (MuSiCa): A web application to implement mutational signatures analysis in cancer samples’, *BMC Bioinformatics*, 19(1), p. 224. doi: 10.1186/s12859-018-2234-y.

Dietlein, F. *et al.* (2020) ‘Identification of cancer driver genes based on nucleotide context’, *Nature Genetics*. doi: 10.1038/s41588-019-0572-y.

DNA Sequencing (2018). Available at: https://www.genome.gov/dna-day/15-ways/dna-sequencing (Accessed: 7 October 2020).

Döhner, H. *et al.* (2017) ‘Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel’, *Blood*. doi: 10.1182/blood-2016-08-733196.

Domingo-Domenech, J. *et al.* (2012) ‘Suppression of Acquired Docetaxel Resistance in Prostate Cancer through Depletion of Notch- and Hedgehog-Dependent Tumor-Initiating Cells’, *Cancer Cell*. doi: 10.1016/j.ccr.2012.07.016.

Dong, H. *et al.* (2018) ‘Long non-coding RNA SNHG14 induces trastuzumab resistance of breast cancer via regulating PABPC1 expression through H3K27 acetylation’, *Journal of Cellular and Molecular Medicine*. doi: 10.1111/jcmm.13758.

Droit, A., Poirier, G. G. and Hunter, J. M. (2005) ‘Experimental and bioinformatic approaches for interrogating protein-protein interactions to determine protein function’, *Journal of Molecular Endocrinology*, 34(2), pp. 263–280. doi: 10.1677/jme.1.01693.

Drummond, C. J. *et al.* (2016) ‘TP53 mutant MDM2-amplified cell lines selected for resistance to MDM2-p53 binding antagonists retain sensitivity to ionizing radiation’, *Oncotarget*. doi: 10.18632/oncotarget.10073.

Dumontet, C. and Jordan, M. A. (2010) ‘Microtubule-binding agents: A dynamic field of cancer therapeutics’, *Nature Reviews Drug Discovery*. doi: 10.1038/nrd3253.

Eilbeck, K. *et al.* (2005) ‘The Sequence Ontology: a tool for the unification of genome annotations.’, *Genome biology*, 6(5). doi: 10.1186/gb-2005-6-5-r44.

El-Gebali, S. *et al.* (2019) ‘The Pfam protein families database in 2019’, *Nucleic Acids Research*, 47(D1), pp. D427–D432. doi: 10.1093/nar/gky995.

Engelman, J. A. *et al.* (2007) ‘MET amplification leads to gefitinib resistance in lung cancer by activating ERBB3 signaling’, *Science*. doi: 10.1126/science.1141478.

Erba, H. P. *et al.* (2019) ‘Phase 1b study of the MDM2 inhibitor AMG 232 with or without trametinib in relapsed/refractory acute myeloid leukemia’, *Blood Advances*, 3(13), pp. 1939–1949. doi: 10.1182/bloodadvances.2019030916.

Esposito, C. *et al.* (2013) ‘The S492R EGFR ectodomain mutation is never detected in KRAS wild-type colorectal carcinoma before exposure to EGFR monoclonal antibodies’, *Cancer Biology and Therapy*. doi: 10.4161/cbt.26340.

Falda, M. *et al.* (2012) ‘Argot2: A large scale function prediction tool relying on semantic similarity of weighted Gene Ontology terms’, *BMC Bioinformatics*, 13(SUPPL.4). doi: 10.1186/1471-2105-13-S4-S14.

Fenton, T. R. *et al.* (2018) ‘What really matters - response and resistance in cancer therapy’, *Cancer Drug Resistance*. doi: 10.20517/cdr.2018.19.

Feuk, L., Carson, A. R. and Scherer, S. W. (2006) ‘Structural variation in the human genome’, *Nature Reviews Genetics*. Nature Publishing Group, pp. 85–97. doi: 10.1038/nrg1767.

Finn, R. D. *et al.* (2016) ‘The Pfam protein families database: Towards a more sustainable future’, *Nucleic Acids Research*, 44(D1), pp. D279–D285. doi: 10.1093/nar/gkv1344.

Finn, R. D., Clements, J. and Eddy, S. R. (2011) ‘HMMER web server: Interactive sequence similarity searching’, *Nucleic Acids Research*, 39(SUPPL. 2), pp. W29–W37. doi: 10.1093/nar/gkr367.

Flores, T. P. *et al.* (1993) ‘Comparison of conformational characteristics in structurally similar protein pairs’, *Protein Science*, 2(11), pp. 1811–1826. doi:

10.1002/pro.5560021104.

Fortunato, A. *et al.* (2017) 'Natural selection in cancer biology: From molecular snowflakes to trait hallmarks', *Cold Spring Harbor Perspectives in Medicine*, 7(2). doi: 10.1101/cshperspect.a029652.

Friedberg, I. (2006) 'Automated protein function prediction - The genomic challenge', *Briefings in Bioinformatics*, 7(3), pp. 225–242. doi: 10.1093/bib/bbl004.

Friedberg, I. and Radivojac, P. (2017) 'Community-wide evaluation of computational function prediction', in *Methods in Molecular Biology*. Humana Press Inc., pp. 133–146. doi: 10.1007/978-1-4939-3743-1_10.

Furnham, N. (2017) 'Complementary sources of protein functional information: The far side of GO', in *Methods in Molecular Biology*. Humana Press Inc., pp. 263–274. doi: 10.1007/978-1-4939-3743-1_19.

Gaeta, R. *et al.* (2020) 'Diffuse bone and soft tissue angiomatosis with GNAQ mutation', *Pathology International*. doi: 10.1111/pin.12933.

Gaffal, E. (2020) 'Research in practice: Therapeutic targeting of oncogenic *GNAQ* mutations in uveal melanoma', *JDDG: Journal der Deutschen Dermatologischen Gesellschaft*, 18(11), pp. 1245–1248. doi: 10.1111/ddg.14288.

Gallo Cassarino, T., Bordoli, L. and Schwede, T. (2014) 'Assessment of ligand binding site predictions in CASP10', *Proteins: Structure, Function and Bioinformatics*, 82(SUPPL.2), pp. 154–163. doi: 10.1002/prot.24495.

Gan, H. H. *et al.* (2002) 'Analysis of protein sequence/structure similarity relationships', *Biophysical Journal*, 83(5), pp. 2781–2791. doi: 10.1016/s0006-3495(02)75287-9.

GenBank and WGS Statistics (2020). Available at: https://www.ncbi.nlm.nih.gov/genbank/statistics/ (Accessed: 8 October 2020).

Germline short variant discovery (SNPs + Indels) – GATK (2019). Available at: https://gatk.broadinstitute.org/hc/en-us/articles/360035535932-Germline-short-variant-discovery-SNPs-Indels (Accessed: 20 December 2020).

Gianna Hoffman-Luca, C. *et al.* (2015) 'Elucidation of acquired resistance to Bcl-2 and MDM2 inhibitors in acute Leukemia in Vitro and in Vivo', *Clinical Cancer Research*. doi:

10.1158/1078-0432.CCR-14-2506.

Gibson, D. G. *et al.* (2010) ‘Creation of a bacterial cell controlled by a chemically synthesized genome’, *Science*. doi: 10.1126/science.1190719.

Giovannoni, S. J. *et al.* (2005) ‘Genetics: Genome streamlining in a cosmopolitan oceanic bacterium’, *Science*, 309(5738), pp. 1242–1245. doi: 10.1126/science.1114057.

Goldberg, T. *et al.* (2014) ‘LocTree3 prediction of localization’, *Nucleic Acids Research*, 42(W1), pp. 350–355. doi: 10.1093/nar/gku396.

Göllner, S. *et al.* (2017) ‘Loss of the histone methyltransferase EZH2 induces resistance to multiple drugs in acute myeloid leukemia’, *Nature Medicine*. doi: 10.1038/nm.4247.

Gonzalez-Perez, A., Mustonen, V., *et al.* (2013) ‘Computational approaches to identify functional genetic variants in cancer genomes’, *Nature Methods*, 10(8), pp. 723–729. doi: 10.1038/nmeth.2562.

Gonzalez-Perez, A., Perez-Llamas, C., *et al.* (2013) ‘IntOGen-mutations identifies cancer drivers across tumor types’, *Nature Methods*. doi: 10.1038/nmeth.2642.

González-Pérez, A. and López-Bigas, N. (2011) ‘Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel’, *American Journal of Human Genetics*, 88(4), pp. 440–449. doi: 10.1016/j.ajhg.2011.03.004.

Goodwin, S., McPherson, J. D. and McCombie, W. R. (2016) ‘Coming of age: Ten years of next-generation sequencing technologies’, *Nature Reviews Genetics*, 17(6), pp. 333–351. doi: 10.1038/nrg.2016.49.

Greener, J. G. *et al.* (2020) ‘Near-complete protein structural modelling of the minimal genome’. Available at: http://arxiv.org/abs/2007.06623 (Accessed: 14 September 2021).

Griffiths, A. J. *et al.* (2000) ‘Somatic versus germinal mutation’. Available at: https://www.ncbi.nlm.nih.gov/books/NBK21894/ (Accessed: 7 September 2021).

Guide to GO evidence codes (2020). Available at: http://geneontology.org/docs/guide-go-evidence-codes/ (Accessed: 14 December 2020).

Haft, D. H. *et al.* (2013) ‘TIGRFAMs and genome properties in 2013’, *Nucleic Acids Research*, 41(D1), pp. D387–D395. doi: 10.1093/nar/gks1234.

Harbeck, N. and Gnant, M. (2017) ‘Breast cancer’, *The Lancet*. doi: 10.1016/S0140-6736(16)31891-8.

Hashimoto, M. *et al.* (2005) ‘Cell size and nucleoid organization of engineered Escherichia coli cells with a reduced genome’, *Molecular Microbiology*. doi: 10.1111/j.1365-2958.2004.04386.x.

Hata, A. N. *et al.* (2016) ‘Tumor cells can follow distinct evolutionary paths to become resistant to epidermal growth factor receptor inhibition’, *Nature Medicine*. doi: 10.1038/nm.4040.

Hawkins, T. *et al.* (2009) ‘PFP: Automated prediction of gene ontology functional annotations with confidence scores using protein sequence data’, *Proteins: Structure, Function and Bioinformatics*. doi: 10.1002/prot.22172.

Heidrich, I. *et al.* (2020) ‘Liquid biopsies: Potential and challenges’, *International Journal of Cancer*, 148(3), pp. 528–545. doi: 10.1002/ijc.33217.

Helleday, T., Eshtad, S. and Nik-Zainal, S. (2014) ‘Mechanisms underlying mutational signatures in human cancers’, *Nature Reviews Genetics*. doi: 10.1038/nrg3729.

Herbst, R. S., Morgensztern, D. and Boshoff, C. (2018) ‘The biology and management of non-small cell lung cancer’, *Nature*. doi: 10.1038/nature25183.

Holliday, G. L. *et al.* (2017) ‘Evaluating functional annotations of enzymes using the gene ontology’, in *Methods in Molecular Biology*. Humana Press Inc., pp. 111–132. doi: 10.1007/978-1-4939-3743-1_9.

Holohan, C. *et al.* (2013) ‘Cancer drug resistance: An evolving paradigm’, *Nature Reviews Cancer*. doi: 10.1038/nrc3599.

Home - CASP14 (2021). Available at: https://www.predictioncenter.org/casp14/ (Accessed: 1 September 2021).

Hospital, A. *et al.* (2015) ‘Molecular dynamics simulations: Advances and applications’, *Advances and Applications in Bioinformatics and Chemistry*. Dove Medical Press Ltd, pp. 37–47. doi: 10.2147/AABC.S70333.

Hou, J. *et al.* (2005) ‘Global mapping of the protein structure space and application in structure-based inference of protein function’, *Proceedings of the National Academy of*

Sciences of the United States of America, 102(10), pp. 3651–3656. doi: 10.1073/pnas.0409772102.

Housman, G. *et al.* (2014) ‘Drug resistance in cancer: An overview’, *Cancers*, 6(3), pp. 1769–1792. doi: 10.3390/cancers6031769.

Huang, J. (2020) ‘Current developments of targeting the p53 signaling pathway for cancer treatment’, *Pharmacology and Therapeutics*. Elsevier Inc., p. 107720. doi: 10.1016/j.pharmthera.2020.107720.

Hudson, A. M. *et al.* (2015) ‘Using large-scale genomics data to identify driver mutations in lung cancer: Methods and challenges’, *Pharmacogenomics*. doi: 10.2217/pgs.15.60.

Huerta-Cepas, J. *et al.* (2017) ‘Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper’, *Molecular Biology and Evolution*, 34(8), pp. 2115–2122. doi: 10.1093/molbev/msx148.

Huntley, R. P. *et al.* (2015) ‘The GOA database: Gene Ontology annotation updates for 2015’, *Nucleic Acids Research*, 43(D1), pp. D1057–D1063. doi: 10.1093/nar/gku1113.

Hutchison, C. A. *et al.* (2016) ‘Design and synthesis of a minimal bacterial genome’, *Science*, 351(6280), pp. aad6253–aad6253. doi: 10.1126/science.aad6253.

Huzil, J. T. *et al.* (2007) ‘The roles of β-tubulin mutations and isotype expression in acquired drug resistance’, *Cancer Informatics*. Libertas Academica Ltd., pp. 159–181. doi: 10.1177/117693510700300028.

Iacobucci, I. and Mullighan, C. G. (2017) ‘Genetic basis of acute lymphoblastic leukemia’, *Journal of Clinical Oncology*. doi: 10.1200/JCO.2016.70.7836.

Ikeda, H., Shin-Ya, K. and Omura, S. (2014) ‘Genome mining of the Streptomyces avermitilis genome and development of genome-minimized hosts for heterologous expression of biosynthetic gene clusters’, *Journal of Industrial Microbiology and Biotechnology*. doi: 10.1007/s10295-013-1327-x.

International Cancer Genome Consortium *et al.* (2010) ‘International network of cancer genome projects’, *Nature*. doi: 10.1038/nature08987.

IUBMB (1992) ‘Enzyme Nomenclature: Recommendations (1992) of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology’, *San Diego,* *CA Academic Press*.

Järviaho, T. *et al.* (2018) ‘Novel non-neutral mitochondrial DNA mutations found in childhood acute lymphoblastic leukemia’, *Clinical Genetics*. doi: 10.1111/cge.13100.

Jiang, Y. *et al.* (2016) ‘An expanded evaluation of protein function prediction methods shows an improvement in accuracy’, *Genome Biology*, 17(1), pp. 1–19. doi: 10.1186/s13059-016-1037-6.

Jones, D. T. (1999) ‘Protein secondary structure prediction based on position-specific scoring matrices’, *Journal of Molecular Biology*, 292(2), pp. 195–202. doi: 10.1006/jmbi.1999.3091.

Jones, D. T. and Cozzetto, D. (2015) ‘DISOPRED3: Precise disordered region predictions with annotated protein-binding activity’, *Bioinformatics*, 31(6), pp. 857–863. doi: 10.1093/bioinformatics/btu744.

Jones, R. J. *et al.* (2012) ‘Drug resistance to inhibitors of the human double minute-2 E3 ligase is mediated by point mutations of p53, but can be overcome with the p53 targeting agent RITA’, *Molecular Cancer Therapeutics*. doi: 10.1158/1535-7163.MCT-12-0135.

Jordan, E. J. *et al.* (2019) ‘Computational algorithms for in silico profiling of activating mutations in cancer’, *Cellular and Molecular Life Sciences*, 76(14), pp. 2663–2679. doi: 10.1007/s00018-019-03097-2.

Joseph, J. D. *et al.* (2013) ‘A clinically relevant androgen receptor mutation confers resistance to second-generation antiandrogens enzalutamide and ARN-509’, *Cancer Discovery*. doi: 10.1158/2159-8290.CD-13-0226.

Juliano, R. L. and Ling, V. (1976) ‘A surface glycoprotein modulating drug permeability in Chinese hamster ovary cell mutants’, *BBA - Biomembranes*. doi: 10.1016/0005-2736(76)90160-7.

Jumper, J. *et al.* (2021) ‘Highly accurate protein structure prediction with AlphaFold’, *Nature*, 596(7873), p. 583. doi: 10.1038/s41586-021-03819-2.

Juncker, A. S. *et al.* (2003) ‘Prediction of lipoprotein signal peptides in Gram-negative bacteria’, *Protein Science*, 12(8), pp. 1652–1662. doi: 10.1110/ps.0303703.

Jung, J. *et al.* (2016) ‘TP53 mutations emerge with HDM2 inhibitor SAR405838 treatment in de-differentiated liposarcoma', *Nature Communications*. doi: 10.1038/ncomms12609.

Kamps, R. *et al.* (2017) 'Next-generation sequencing in oncology: Genetic diagnosis, risk prediction and cancer classification', *International Journal of Molecular Sciences*, 18(2). doi: 10.3390/ijms18020308.

Kanehisa, M., Sato, Y. and Kawashima, M. (2021) 'KEGG mapping tools for uncovering hidden features in biological data', *Protein Science*, p. pro.4172. doi: 10.1002/pro.4172.

Karczewski, K. J. *et al.* (2020) 'The mutational constraint spectrum quantified from variation in 141,456 humans', *Nature*. doi: 10.1038/s41586-020-2308-7.

Karsch-Mizrachi, I., Takagi, T. and Cochrane, G. (2018) 'The international nucleotide sequence database collaboration', *Nucleic Acids Research*, 46(D1), pp. D48–D51. doi: 10.1093/nar/gkx1097.

Katoh, K. *et al.* (2002) 'MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform', *Nucleic Acids Research*, 30(14), pp. 3059–3066. doi: 10.1093/nar/gkf436.

Kavallaris, M. (2010) 'Microtubules and resistance to tubulin-binding agents', *Nature Reviews Cancer*. doi: 10.1038/nrc2803.

Kayode, O. *et al.* (2016) 'An acrobatic substrate metamorphosis reveals a requirement for substrate conformational dynamics in trypsin proteolysis', *Journal of Biological Chemistry*, 291(51), pp. 26304–26319. doi: 10.1074/jbc.M116.758417.

Kelley, L. A. *et al.* (2015) 'The Phyre2 web portal for protein modeling, prediction and analysis', *Nature Protocols*, 10(6), pp. 845–858. doi: 10.1038/nprot.2015.053.

Kelley, L. A. and Sternberg, M. J. E. (2015) 'Partial protein domains: Evolutionary insights and bioinformatics challenges', *Genome Biology*. doi: 10.1186/s13059-015-0663-8.

Kemper, K. *et al.* (2015) 'Intra- and inter-tumor heterogeneity in a vemurafenib-resistant melanoma patient and derived xenografts', *EMBO Molecular Medicine*. doi: 10.15252/emmm.201404914.

Khurana, A. and Shafer, D. A. (2019) 'MDM2 antagonists as a novel treatment option for acute myeloid leukemia: perspectives on the therapeutic potential of idasanutlin (RG7388)', *OncoTargets and Therapy*, Volume 12, pp. 2903–2910. doi:

10.2147/ott.s172315.

Kloss-Brandstätter, A. *et al.* (2010) 'Somatic mutations throughout the entire mitochondrial genome are associated with elevated PSA levels in prostate cancer patients', *American Journal of Human Genetics*. doi: 10.1016/j.ajhg.2010.11.001.

Kojima, K. *et al.* (2005) 'MDM2 antagonists induce p53-dependent apoptosis in AML: Implications for leukemia therapy', *Blood*. doi: 10.1182/blood-2005-02-0553.

Kojima, K., Ishizawa, J. and Andreeff, M. (2016) 'Pharmacological activation of wild-type p53 in the therapy of leukemia', *Experimental Hematology*. doi: 10.1016/j.exphem.2016.05.014.

Kolisnychenko, V. *et al.* (2002) 'Engineering a reduced Escherichia coli genome', *Genome Research*. doi: 10.1101/gr.217202.

Komatsu, M. *et al.* (2010) 'Genome-minimized Streptomyces host for the heterologous expression of secondary metabolism', *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.0914833107.

Konieczkowski, D. J., Johannessen, C. M. and Garraway, L. A. (2018) 'A Convergence-Based Framework for Cancer Drug Resistance', *Cancer Cell*. Cell Press, pp. 801–815. doi: 10.1016/j.ccell.2018.03.025.

Konopleva, M. *et al.* (2020) 'MDM2 inhibition: an important step forward in cancer therapy', *Leukemia*. doi: 10.1038/s41375-020-0949-z.

Korpal, M. *et al.* (2013) 'An F876l mutation in androgen receptor confers genetic and phenotypic resistance to MDV3100 (Enzalutamide)', *Cancer Discovery*. doi: 10.1158/2159-8290.CD-13-0142.

Kotchetkov, R. *et al.* (2005) 'Increased malignant behavior in neuroblastoma cells with acquired multi-drug resistance does not depend on P-gp expression', *International Journal of Oncology*. doi: 10.3892/ijo.27.4.1029.

Kourmpetis, Y. A. I. *et al.* (2011) 'Genome-wide computational function prediction of arabidopsis proteins by integration of multiple data sources', *Plant Physiology*, 155(1), pp. 271–281. doi: 10.1104/pp.110.162164.

Krogh, A. *et al.* (2001) 'Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes', *Journal of Molecular Biology*, 305(3), pp. 567–580. doi: 10.1006/jmbi.2000.4315.

Kryshtafovych, A., Fidelis, K. and Moult, J. (2011) 'CASP9 results compared to those of previous casp experiments', *Proteins: Structure, Function and Bioinformatics*, 79(SUPPL. 10), pp. 196–207. doi: 10.1002/prot.23182.

Kumar, P., Henikoff, S. and Ng, P. C. (2009) 'Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm', *Nature Protocols*. doi: 10.1038/nprot.2009.86.

Landrum, M. J. *et al.* (2018) 'ClinVar: Improving access to variant interpretations and supporting evidence', *Nucleic Acids Research*, 46(D1), pp. D1062–D1067. doi: 10.1093/nar/gkx1153.

Lee, C. S. *et al.* (2017) 'Mutations in Fibronectin Cause a Subtype of Spondylometaphyseal Dysplasia with "Corner Fractures"', *American Journal of Human Genetics*, 101(5), pp. 815–823. doi: 10.1016/j.ajhg.2017.09.019.

Lehmann, C. *et al.* (2016) 'Superior anti-tumor activity of the MDM2 antagonist idasanutlin and the Bcl-2 inhibitor venetoclax in p53 wild-type acute myeloid leukemia models', *Journal of Hematology and Oncology*. doi: 10.1186/s13045-016-0280-3.

Lek, M. *et al.* (2016) 'Analysis of protein-coding genetic variation in 60,706 humans', *Nature*. doi: 10.1038/nature19057.

Leprince, A. *et al.* (2012) 'Random and cyclical deletion of large DNA segments in the genome of Pseudomonas putida', *Environmental Microbiology*. doi: 10.1111/j.1462-2920.2012.02730.x.

Letunic, I. and Bork, P. (2018) '20 years of the SMART protein domain annotation resource', *Nucleic Acids Research*, 46(D1), pp. D493–D496. doi: 10.1093/nar/gkx922.

Levine, A. J. (2020) 'p53: 800 million years of evolution and 40 years of discovery', *Nature Reviews Cancer*. doi: 10.1038/s41568-020-0262-1.

Lewis, T. E. *et al.* (2018) 'Gene3D: Extensive prediction of globular domains in proteins', *Nucleic Acids Research*, 46(D1), pp. D435–D439. doi: 10.1093/nar/gkx1069.

Li, H. *et al.* (2009) 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics*, 25(16), pp. 2078–2079. doi: 10.1093/bioinformatics/btp352.

Li, H. (2011) ‘A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data’, *Bioinformatics*. doi: 10.1093/bioinformatics/btr509.

Li, H. and Durbin, R. (2009) ‘Fast and accurate short read alignment with Burrows-Wheeler transform’, *Bioinformatics*. doi: 10.1093/bioinformatics/btp324.

Li, Y. *et al.* (2016) ‘Characterization of genome-reduced Bacillus subtilis strains and their application for the production of guanosine and thymidine’, *Microbial Cell Factories*. doi: 10.1186/s12934-016-0494-7.

Li, Z. *et al.* (2019) ‘Recurrent GNAQ mutation encoding T96S in natural killer/T cell lymphoma’, *Nature Communications*. doi: 10.1038/s41467-019-12032-9.

Liang, H. and Kim, Y. H. (2013) ‘Identifying molecular drivers of gastric cancer through next-generation sequencing’, *Cancer Letters*. doi: 10.1016/j.canlet.2012.11.029.

Licata, L. *et al.* (2012) ‘MINT, the molecular interaction database: 2012 Update’, *Nucleic Acids Research*, 40(D1). doi: 10.1093/nar/gkr930.

Lieder, S. *et al.* (2015) ‘Genome reduction boosts heterologous gene expression in Pseudomonas putida’, *Microbial Cell Factories*. doi: 10.1186/s12934-015-0207-7.

Lin, Y. M. *et al.* (2008) ‘Disturbance of circadian gene expression in hepatocellular carcinoma’, *Molecular Carcinogenesis*. doi: 10.1002/mc.20446.

Litwin, M. S. and Tan, H. J. (2017) ‘The diagnosis and treatment of prostate cancer: A review’, *JAMA - Journal of the American Medical Association*. doi: 10.1001/jama.2017.7248.

Liu, J. *et al.* (2018) ‘An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics’, *Cell*, 173(2), pp. 400-416.e11. doi: 10.1016/j.cell.2018.02.052.

Lobley, A. E. *et al.* (2008) ‘FFPred: an integrated feature-based function prediction server for vertebrate proteomes.’, *Nucleic acids research*, 36(Web Server issue), pp. 297–302. doi: 10.1093/nar/gkn193.

Loewenstein, Y. *et al.* (2009) ‘Protein function annotation by homology-based inference.’, *Genome biology*. doi: 10.1186/gb-2009-10-2-207.

Long, J. *et al.* (2010) ‘Multiple distinct molecular mechanisms influence sensitivity and resistance to MDM2 inhibitors in adult acute myelogenous leukemia’, *Blood*. doi: 10.1182/blood-2010-01-261628.

Lopez, G. *et al.* (2011) ‘Firestar - Advances in the prediction of functionally important residues’, *Nucleic Acids Research*, 39(SUPPL. 2), pp. W235–W241. doi: 10.1093/nar/gkr437.

Ma, X. *et al.* (2018) ‘Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours’, *Nature*. doi: 10.1038/nature25795.

Maganti, H. B. *et al.* (2018) ‘Targeting the MTF2–MDM2 axis sensitizes refractory acute myeloid leukemia to chemotherapy’, *Cancer Discovery*. doi: 10.1158/2159-8290.CD-17-0841.

Mahlich, Y. *et al.* (2018) ‘HFSP: High speed homology-driven function annotation of proteins’, *Bioinformatics*, 34(13), pp. i304–i312. doi: 10.1093/bioinformatics/bty262.

Maietta, P. *et al.* (2014) ‘FireDB: A compendium of biological and pharmacologically relevant ligands’, *Nucleic Acids Research*, 42(D1), pp. D267–D272. doi: 10.1093/nar/gkt1127.

Makrodimitris, S., Van Ham, R. C. H. J. and Reinders, M. J. T. (2020) ‘Automatic gene function prediction in the 2020’s’, *Genes*. MDPI AG, pp. 1–18. doi: 10.3390/genes11111264.

Mansoori, B. *et al.* (2017) ‘The different mechanisms of cancer drug resistance: A brief review’, *Advanced Pharmaceutical Bulletin*, 7(3), pp. 339–348. doi: 10.15171/apb.2017.041.

Mao, X. *et al.* (2014) ‘DOOR 2.0: Presenting operons and their functions through dynamic and integrated views’, *Nucleic Acids Research*, 42(D1), pp. D654–D659. doi: 10.1093/nar/gkt1048.

Marcellino, B. K. *et al.* (2020) ‘Transient expansion of TP53 mutated clones in polycythemia vera patients treated with idasanutlin.’, *Blood advances*, 4(22), pp. 5735–

5744. doi: 10.1182/bloodadvances.2020002379.

Marchler-Bauer, A. *et al.* (2017) ‘CDD/SPARCLE: Functional classification of proteins via subfamily domain architectures’, *Nucleic Acids Research*, 45(D1), pp. D200–D203. doi: 10.1093/nar/gkw1129.

Martincorena, I. *et al.* (2017) ‘Universal Patterns of Selection in Cancer and Somatic Tissues’, *Cell*. doi: 10.1016/j.cell.2017.09.042.

Martínez-García, E. and de Lorenzo, V. (2016) ‘The quest for the minimal bacterial genome’, *Current Opinion in Biotechnology*. doi: 10.1016/j.copbio.2016.09.001.

Martínez-Jiménez, F., Muiños, F., Sentís, I., *et al.* (2020) ‘A compendium of mutational cancer driver genes’, *Nature Reviews Cancer*, 20(10), pp. 555–572. doi: 10.1038/s41568-020-0290-x.

Martínez-Jiménez, F., Muiños, F., López-Arribillaga, E., *et al.* (2020) ‘Systematic analysis of alterations in the ubiquitin proteolysis system reveals its contribution to driver mutations in cancer’, *Nature Cancer*. doi: 10.1038/s43018-019-0001-2.

McCormack, E. *et al.* (2012) ‘Synergistic induction of p53 mediated apoptosis by valproic acid and nutlin-3 in acute myeloid leukemia’, *Leukemia*. doi: 10.1038/leu.2011.315.

McGranahan, N. and Swanton, C. (2017) ‘Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future’, *Cell*. doi: 10.1016/j.cell.2017.01.018.

Mcguffin, L. J. *et al.* (2019) ‘IntFOLD: An integrated web resource for high performance protein structure and function prediction’, *Nucleic Acids Research*, 47(W1), pp. W408–W413. doi: 10.1093/nar/gkz322.

McGuffin, L. J. *et al.* (2021) ‘ModFOLD8: Accurate global and local quality estimates for 3D protein models’, *Nucleic Acids Research*, 49(W1), pp. W425–W430. doi: 10.1093/nar/gkab321.

McKenna, A. *et al.* (2010) ‘The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data’, *Genome Research*. doi: 10.1101/gr.107524.110.

McLaren, W. *et al.* (2016) ‘The Ensembl Variant Effect Predictor’, *Genome Biology*. doi: 10.1186/s13059-016-0974-4.

Michaelis, M. *et al.* (2011) ‘Adaptation of cancer cells from different entities to the MDM2 inhibitor nutlin-3 results in the emergence of p53-mutated multi-drug-resistant cancer cells’, *Cell Death and Disease*. doi: 10.1038/cddis.2011.129.

Michaelis, M. *et al.* (2012) ‘Human neuroblastoma cells with acquired resistance to the p53 activator RITA retain functional p53 and sensitivity to other p53 activating agents’, *Cell Death and Disease*. doi: 10.1038/cddis.2012.35.

Michaelis, M., Rothweiler, F., *et al.* (2020) ‘Long-term cultivation using ineffective MDM2 inhibitor concentrations alters the drug sensitivity profiles of PL21 leukaemia cells’, *Experimental Results*. doi: 10.1017/exp.2019.1.

Michaelis, M., Wass, M. N., *et al.* (2020) ‘YM155-adapted cancer cell lines reveal drug-induced heterogeneity and enable the identification of biomarker candidates for the acquired resistance setting’, *Cancers*. doi: 10.3390/cancers12051080.

Michaelis, M., Wass, M. N. and Cinatl, J. (2019) ‘Drug-adapted cancer cell lines as preclinical models of acquired resistance’, *Cancer Drug Resistance*, 2(3), pp. 447–456. doi: 10.20517/cdr.2019.005.

Miklos, W. *et al.* (2015) ‘Triapine-mediated ABCB1 induction via PKC induces widespread therapy unresponsiveness but is not underlying acquired triapine resistance’, *Cancer Letters*. doi: 10.1016/j.canlet.2015.02.049.

Mishra, N. K., Chang, J. and Zhao, P. X. (2014) ‘Prediction of membrane transport proteins and their substrate specificities using primary sequence information’, *PLoS ONE*, 9(6), p. e100278. doi: 10.1371/journal.pone.0100278.

Mitchell, A. L. *et al.* (2019) ‘InterPro in 2019: Improving coverage, classification and access to protein sequence annotations’, *Nucleic Acids Research*, 47(D1), pp. D351–D360. doi: 10.1093/nar/gky1100.

Mortuza, S. M. *et al.* (2021) ‘Improving fragment-based ab initio protein structure assembly using low-accuracy contact-map predictions’, *Nature Communications*, 12(1), pp. 1–12. doi: 10.1038/s41467-021-25316-w.

Mosmann, T. (1983) ‘Rapid colorimetric assay for cellular growth and survival: Application to proliferation and cytotoxicity assays’, *Journal of Immunological Methods*. doi:

10.1016/0022-1759(83)90303-4.

Mularoni, L. *et al.* (2016) ‘OncodriveFML: A general framework to identify coding and non-coding regions with cancer driver mutations’, *Genome Biology*. doi: 10.1186/s13059-016-0994-0.

Nagasawa, K. I. *et al.* (2000) ‘Identification and characterization of human DNA polymerase β2, a DNA polymerase β-related enzyme’, *Journal of Biological Chemistry*, 275(40), pp. 31233–31238. doi: 10.1074/jbc.M004263200.

Nahta, R. and Castellino, R. C. (2020) ‘Phosphatase magnesium-dependent 1 δ (PPM1D), serine/threonine protein phosphatase and novel pharmacological target in cancer’, *Biochemical Pharmacology*, 184, p. 114362. doi: 10.1016/j.bcp.2020.114362.

Nakabachi, A. *et al.* (2006) ‘The 160-kilobase genome of the bacterial endosymbiont Carsonella’, *Science*, 314(5797), p. 267. doi: 10.1126/science.1134196.

Nakagawa, H. and Fujita, M. (2018) ‘Whole genome sequencing analysis for cancer genomics and precision medicine’, *Cancer Science*, 109(3), pp. 513–522. doi: 10.1111/cas.13505.

Nazarian, R. *et al.* (2010) ‘Melanomas acquire resistance to B-RAF(V600E) inhibition by RTK or N-RAS upregulation’, *Nature*. doi: 10.1038/nature09626.

Nehrt, N. L. *et al.* (2011) ‘Testing the ortholog conjecture with comparative functional genomic data from mammals’, *PLoS Computational Biology*, 7(6). doi: 10.1371/journal.pcbi.1002073.

NEW PRODUCT: Discover the Cancer Mutation Census (2020). Available at: https://cosmic-blog.sanger.ac.uk/discover-the-cmc/ (Accessed: 4 November 2020).

Ng, P. C. and Henikoff, S. (2001) ‘Predicting deleterious amino acid substitutions’, *Genome Research*, 11(5), pp. 863–874. doi: 10.1101/gr.176601.

Ng, P. C. and Henikoff, S. (2002) ‘Accounting for human polymorphisms predicted to affect protein function’, *Genome Research*, 12(3), pp. 436–446. doi: 10.1101/gr.212802.

Nguyen, N. N. Y., Kim, S. S. and Jo, Y. H. (2020) ‘Deregulated mitochondrial DNA in diseases’, *DNA and Cell Biology*. Mary Ann Liebert Inc., pp. 1385–1400. doi: 10.1089/dna.2019.5220.

Nicholson, K. M. and Anderson, N. G. (2002) ‘The protein kinase B/Akt signalling pathway in human malignancy’, *Cellular Signalling*. Cell Signal, pp. 381–395. doi: 10.1016/S0898-6568(01)00271-6.

Nikolaou, M. *et al.* (2018) ‘The challenge of drug resistance in cancer treatment: a current overview’, *Clinical and Experimental Metastasis*, 35(4), pp. 309–318. doi: 10.1007/s10585-018-9903-0.

Northey, T. C., Barešić, A. and Martin, A. C. R. (2018) ‘IntPred: A structure-based predictor of protein-protein interaction sites’, *Bioinformatics*, 34(2), pp. 223–229. doi: 10.1093/bioinformatics/btx585.

Nussinov, R. *et al.* (2019) *Review: Precision medicine and driver mutations: Computational methods, functional assays and conformational principles for interpreting cancer drivers*, *PLoS Computational Biology*. doi: 10.1371/journal.pcbi.1006658.

Oates, M. E. *et al.* (2015) ‘The SUPERFAMILY 1.75 database in 2014: A doubling of data’, *Nucleic Acids Research*, 43(D1), pp. D227–D233. doi: 10.1093/nar/gku1041.

Obayashi, T. *et al.* (2019) ‘COXPRESdb v7: A gene coexpression database for 11 animal species supported by 23 coexpression platforms for technical evaluation and evolutionary inference’, *Nucleic Acids Research*, 47(D1), pp. D55–D62. doi: 10.1093/nar/gky1155.

Onafuye, H. *et al.* (2019) ‘Doxorubicin-loaded human serum albumin nanoparticles overcome transporter-mediated drug resistance in drug-adapted cancer cells’, *Beilstein Journal of Nanotechnology*, 10(1), pp. 1707–1715. doi: 10.3762/bjnano.10.166.

Orchard, S. *et al.* (2014) ‘The MIntAct project - IntAct as a common curation platform for 11 molecular interaction databases’, *Nucleic Acids Research*, 42(D1), pp. D358-63. doi: 10.1093/nar/gkt1115.

Outcomes & Impact of The Cancer Genome Atlas - National Cancer Institute (2019). Available at: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/history (Accessed: 2 November 2020).

Ouzounis, C. A. *et al.* (2003) ‘Classification schemes for protein structure and function’, *Nature Reviews Genetics*. Nature Publishing Group, pp. 508–519. doi: 10.1038/nrg1113.

Pan, R. *et al.* (2017) ‘Synthetic Lethality of Combined Bcl-2 Inhibition and p53 Activation in

AML: Mechanisms and Superior Antileukemic Efficacy', *Cancer Cell*. doi: 10.1016/j.ccell.2017.11.003.

Pearce, R. and Zhang, Y. (2021a) 'Deep learning techniques have significantly impacted protein structure prediction and protein design', *Current Opinion in Structural Biology*. Elsevier Ltd, pp. 194–207. doi: 10.1016/j.sbi.2021.01.007.

Pearce, R. and Zhang, Y. (2021b) 'Toward the solution of the protein structure prediction problem', *Journal of Biological Chemistry*. American Society for Biochemistry and Molecular Biology Inc., p. 100870. doi: 10.1016/j.jbc.2021.100870.

Pearson, W. R. (2013) 'An introduction to sequence similarity ("homology") searching', *Current Protocols in Bioinformatics*, 0 3(SUPPL.42). doi: 10.1002/0471250953.bi0301s42.

Pedruzzi, I. *et al.* (2015) 'HAMAP in 2015: Updates to the protein family classification and annotation system', *Nucleic Acids Research*, 43(D1), pp. D1064–D1070. doi: 10.1093/nar/gku1002.

Pi, L. *et al.* (2019) 'Evaluating dose-limiting toxicities of MDM2 inhibitors in patients with solid organ and hematologic malignancies: A systematic review of the literature', *Leukemia Research*. doi: 10.1016/j.leukres.2019.106222.

Picard Tools - By Broad Institute (2019). Available at: http://broadinstitute.github.io/picard/ (Accessed: 19 December 2020).

Pienta, K. J. *et al.* (2020) 'Convergent evolution, evolving evolvability, and the origins of lethal cancer', *Molecular Cancer Research*. American Association for Cancer Research Inc., pp. 801–810. doi: 10.1158/1541-7786.MCR-19-1158.

Poulikakos, P. I. *et al.* (2011) 'RAF inhibitor resistance is mediated by dimerization of aberrantly spliced BRAF(V600E)', *Nature*. doi: 10.1038/nature10662.

Price, M. N. *et al.* (2018) 'Mutant phenotypes for thousands of bacterial genes of unknown function', *Nature*, 557(7706), pp. 503–509. doi: 10.1038/s41586-018-0124-0.

Priestley, P. *et al.* (2019) 'Pan-cancer whole-genome analyses of metastatic solid tumours', *Nature*. doi: 10.1038/s41586-019-1689-y.

Publications | NCI Genomic Data Commons (2020). Available at: https://gdc.cancer.gov/about-data/publications (Accessed: 3 November 2020).

Punta, M. and Ofran, Y. (2008) ‘The rough guide to in silico function prediction, or how to use sequence and structure information to predict protein function’, *PLoS Computational Biology*, 4(10). doi: 10.1371/journal.pcbi.1000160.

QuickGO::Term GO:0003674 (2021). Available at: https://www.ebi.ac.uk/QuickGO/term/GO:0003674 (Accessed: 7 September 2021).

QuickGO::Term GO:0034061 (2020). Available at: https://www.ebi.ac.uk/QuickGO/GTerm?id=GO:0034061 (Accessed: 27 November 2020).

Radivojac, P. *et al.* (2013) ‘A large-scale evaluation of computational protein function prediction’, *Nature Methods*, 10(3), pp. 221–227. doi: 10.1038/nmeth.2340.

Rajendran, B. K. and Deng, C. X. (2017) ‘Characterization of potential driver mutations involved in human breast cancer by computational approaches’, *Oncotarget*, 8(30), pp. 50252–50272. doi: 10.18632/oncotarget.17225.

Rentoft, M. *et al.* (2016) ‘Heterozygous colon cancer-associated mutations of SAMHD1 have functional significance’, *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.1519128113.

Rentzsch, R. and Orengo, C. A. (2009) ‘Protein function prediction - the power of multiplicity’, *Trends in Biotechnology*, 27(4), pp. 210–219. doi: 10.1016/j.tibtech.2009.01.002.

Reva, B., Antipin, Y. and Sander, C. (2007) ‘Determinants of protein function revealed by combinatorial entropy optimization’, *Genome Biology*, 8(11), p. R232. doi: 10.1186/gb-2007-8-11-r232.

Roche, D. B., Brackenridge, D. A. and McGuffin, L. J. (2015) ‘Proteins and their interacting partners: An introduction to protein-ligand binding site prediction methods’, *International Journal of Molecular Sciences*. MDPI AG, pp. 29829–29842. doi: 10.3390/ijms161226202.

Roche, D. B. and McGuffin, L. J. (2016) ‘In silico identification and characterization of protein-ligand binding sites’, in *Methods in Molecular Biology*. Humana Press Inc., pp. 1–21. doi: 10.1007/978-1-4939-3569-7_1.

Rohl, C. A. *et al.* (2004) ‘Protein Structure Prediction Using Rosetta’, *Methods in Enzymology*, 383, pp. 66–93. doi: 10.1016/S0076-6879(04)83004-0.

Rost, B. *et al.* (2003) ‘Automatic prediction of protein function’, *Cellular and Molecular Life Sciences*, 60(12), pp. 2637–2650. doi: 10.1007/s00018-003-3114-8.

S Razin, D. Y. Y. N. (1998) ‘Molecular biology and pathogenicity of mycoplasmas’, *Microbiol. Mol. Biol. Rev.*, 62, pp. 1094–1156.

Salmaso, V. and Moro, S. (2018) ‘Bridging molecular docking to molecular dynamics in exploring ligand-protein recognition process: An overview’, *Frontiers in Pharmacology*. Frontiers Media S.A., p. 923. doi: 10.3389/fphar.2018.00923.

Sample overview for 1330947 (2020). Available at: https://cancer.sanger.ac.uk/cell_lines/sample/overview?id=1330947 (Accessed: 22 December 2020).

Samudio, I. J. *et al.* (2010) ‘Activation of p53 signaling by MI-63 induces apoptosis in acute myeloid leukemia cells’, *Leukemia and Lymphoma*, 51(5), pp. 911–919. doi: 10.3109/10428191003731325.

Schneider, C. *et al.* (2017) ‘SAMHD1 is a biomarker for cytarabine response and a therapeutic target in acute myeloid leukemia’, *Nature Medicine*. doi: 10.1038/nm.4255.

Schomburg, I., Chang, A. and Schomburg, D. (2002) ‘BRENDA, enzyme data and metabolic information’, *Nucleic Acids Research*, 30(1), pp. 47–49. doi: 10.1093/nar/30.1.47.

Schulenburg, C. and Miller, B. G. (2014) ‘Enzyme recruitment and its role in metabolic expansion’, *Biochemistry*. American Chemical Society, pp. 836–845. doi: 10.1021/bi401667f.

Secchiero, P. *et al.* (2007) ‘The MDM-2 antagonist nutlin-3 promotes the maturation of acute myeloid leukemic blasts’, *Neoplasia*. doi: 10.1593/neo.07523.

Seipel, K. *et al.* (2018) ‘The cellular p53 inhibitor MDM2 and the growth factor receptor FLT3 as biomarkers for treatment responses to the MDM2-inhibitor idasanutlin and the MEK1 inhibitor cobimetinib in acute myeloid leukemia’, *Cancers*. doi: 10.3390/cancers10060170.

Senft, D. *et al.* (2017) ‘Precision Oncology: The Road Ahead’, *Trends in Molecular Medicine*, 23(10), pp. 874–898. doi: 10.1016/j.molmed.2017.08.003.

Senior, A. W. *et al.* (2019) ‘Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13)', *Proteins: Structure, Function, and Bioinformatics*, 87(12), pp. 1141–1148. doi: 10.1002/prot.25834.

Sequist, L. V. *et al.* (2011) 'Genotypic and histological evolution of lung cancers acquiring resistance to EGFR inhibitors', *Science Translational Medicine*. doi: 10.1126/scitranslmed.3002003.

Serçinoğlu, O. and Ozbek, P. (2020) 'Sequence-structure-function relationships in class i MHC: A local frustration perspective', *PLoS ONE*. Public Library of Science, p. e0232849. doi: 10.1371/journal.pone.0232849.

Servant, F. *et al.* (2002) 'ProDom: automated clustering of homologous domains.', *Briefings in bioinformatics*, 3(3), pp. 246–251. doi: 10.1093/bib/3.3.246.

Shehu, A., Barbará, D. and Molloy, K. (2016) 'A survey of computational methods for protein function prediction', in *Big Data Analytics in Genomics*. Springer International Publishing, pp. 225–298. doi: 10.1007/978-3-319-41279-5_7.

Sherry, S. T. *et al.* (2001) 'DbSNP: The NCBI database of genetic variation', *Nucleic Acids Research*. doi: 10.1093/nar/29.1.308.

Shihab, H. A. *et al.* (2013) 'Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models', *Human Mutation*, 34(1), pp. 57–65. doi: 10.1002/humu.22225.

Sigrist, C. J. A. *et al.* (2013) 'New and continuing developments at PROSITE', *Nucleic Acids Research*, 41(D1), pp. D344–D347. doi: 10.1093/nar/gks1067.

Sillitoe, I. *et al.* (2013) 'New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures', *Nucleic Acids Research*, 41(D1), pp. D490–D498. doi: 10.1093/nar/gks1211.

Sim, N. L. *et al.* (2012) 'SIFT web server: Predicting effects of amino acid substitutions on proteins', *Nucleic Acids Research*, 40(W1), pp. 452–457. doi: 10.1093/nar/gks539.

Sokolov, A. and Ben-Hur, A. (2010) 'Hierarchical classification of gene ontology terms using the GOstruct method', *Journal of Bioinformatics and Computational Biology*, 8(2), pp. 357–376. doi: 10.1142/S0219720010004744.

Soucheray, M. *et al.* (2015) ‘Intratumoral heterogeneity in EGFR-mutant NSCLC results in divergent resistance mechanisms in response to EGFR tyrosine kinase inhibition’, *Cancer Research*. doi: 10.1158/0008-5472.CAN-15-0377.

Soverini, S. *et al.* (2018) ‘Chronic myeloid leukemia: The paradigm of targeting oncogenic tyrosine kinase signaling and counteracting resistance for successful cancer therapy’, *Molecular Cancer*. doi: 10.1186/s12943-018-0780-6.

Stamboulian, M. *et al.* (2020) ‘The ortholog conjecture revisited: the value of orthologs and paralogs in function prediction’, *Bioinformatics (Oxford, England)*, 36(1), pp. i219–i226. doi: 10.1093/bioinformatics/btaa468.

Stone, E. A. and Sidow, A. (2005) ‘Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity’, *Genome Research*, 15(7), pp. 978–986. doi: 10.1101/gr.3804205.

Sunyaev, S. R. *et al.* (1999) ‘PSIC: Profile extraction from sequence alignments with position-specific counts of independent observations’, *Protein Engineering*, 12(5), pp. 387–394. doi: 10.1093/protein/12.5.387.

Tanda, E. T. *et al.* (2020) ‘Current State of Target Treatment in BRAF Mutated Melanoma’, *Frontiers in Molecular Biosciences*. Frontiers Media S.A., p. 154. doi: 10.3389/fmolb.2020.00154.

Tate, J. G. *et al.* (2019) ‘COSMIC: The Catalogue Of Somatic Mutations In Cancer’, *Nucleic Acids Research*, 47(D1), pp. D941–D947. doi: 10.1093/nar/gky1015.

Thachuk, C., Shmygelska, A. and Hoos, H. H. (2007) ‘A replica exchange Monte Carlo algorithm for protein folding in the HP model’, *BMC Bioinformatics*, 8(1), p. 342. doi: 10.1186/1471-2105-8-342.

The Cancer Genome Atlas - Cancers Selected for Study - National Cancer Institute (2020). Available at: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/studied-cancers (Accessed: 2 November 2020).

The Cancer Genome Atlas Research Network (2015) ‘Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas’, *New England Journal of Medicine*, 372(26), pp. 2481–2498. doi: 10.1056/nejmoa1402121.

The Cost of Sequencing a Human Genome (2020). Available at: https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost (Accessed: 7 October 2020).

The MISO Sequence Ontology Browser (2016). Available at: http://www.sequenceontology.org/browser/obob.cgi (Accessed: 27 October 2020).

The MISO Sequence Ontology Browser - MISSENSE_VARIANT (2016). Available at: http://sequenceontology.org/browser/current_release/term/SO:0001583 (Accessed: 27 November 2020).

Thomas, D. *et al.* (2015) ‘Quantitation of endogenous nucleoside triphosphates and nucleosides in human cells by liquid chromatography tandem mass spectrometry’, *Analytical and Bioanalytical Chemistry*. doi: 10.1007/s00216-015-8588-3.

Thomas, P. D. *et al.* (2012) ‘On the use of gene ontology annotations to assess functional similarity among orthologs and paralogs: A short report’, *PLoS Computational Biology*, 8(2), pp. 1–7. doi: 10.1371/journal.pcbi.1002386.

Thompson, J. D., Prigent, V. and Poch, O. (2004) ‘LEON: Multiple alignment evaluation of neighbours’, *Nucleic Acids Research*, 32(4), pp. 1298–1307. doi: 10.1093/nar/gkh294.

Tipton, K. and Mcdonald, A. (2018) *A Brief Guide to Enzyme Nomenclature and Classification*.

Tisato, V. *et al.* (2017) ‘MDM2/X inhibitors under clinical evaluation: Perspectives for the management of hematological malignancies and pediatric cancer’, *Journal of Hematology and Oncology*. BioMed Central Ltd., pp. 1–17. doi: 10.1186/s13045-017-0500-5.

Tokheim, C. *et al.* (2016) ‘Exome-scale discovery of hotspot mutation regions in human cancer using 3D protein structure’, *Cancer Research*. doi: 10.1158/0008-5472.CAN-15-3190.

UniProtKB/Swiss-Prot 2020_05 (2020). Available at: https://www.uniprot.org/statistics/Swiss-Prot (Accessed: 13 October 2020).

Unthan, S. *et al.* (2015) ‘Chassis organism from Corynebacterium glutamicum - a top-down approach to identify and delete irrelevant gene clusters’, *Biotechnology Journal*. doi: 10.1002/biot.201400041.

Uziela, K. *et al.* (2016) ‘ProQ3: Improved model quality assessments using Rosetta energy terms’, *Scientific Reports*, 6(1), pp. 1–10. doi: 10.1038/srep33509.

Vassilev, L. T. *et al.* (2004) ‘In Vivo Activation of the p53 Pathway by Small-Molecule Antagonists of MDM2’, *Science*. doi: 10.1126/science.1092472.

Vesztrocy, A. W. and Dessimoz, C. (2021) ‘Benchmarking gene ontology function predictions using negative annotations’, *Bioinformatics*, 36(Suppl 1), pp. I210–I218. doi: 10.1093/BIOINFORMATICS/BTAA466.

Wade, M., Li, Y. C. and Wahl, G. M. (2013) ‘MDM2, MDMX and p53 in oncogenesis and cancer therapy’, *Nature Reviews Cancer*. doi: 10.1038/nrc3430.

Wang, F. and Zhang, W. (2019) ‘Synthetic biology: Recent progress, biosafety and biosecurity concerns, and possible solutions’, *Journal of Biosafety and Biosecurity*, 1(1), pp. 22–30. doi: 10.1016/j.jobb.2018.12.003.

Wass, M. N., Barton, G. and Sternberg, M. J. E. (2012) ‘CombFunc: Predicting protein function using heterogeneous data sources’, *Nucleic Acids Research*, 40(W1), pp. 466–470. doi: 10.1093/nar/gks489.

Wass, M. N., Kelley, L. A. and Sternberg, M. J. E. (2010) ‘3DLigandSite: Predicting ligand-binding sites using similar structures’, *Nucleic Acids Research*, 38(SUPPL. 2), pp. W469–W473. doi: 10.1093/nar/gkq406.

Wass, M. N. and Sternberg, M. J. E. (2008) ‘ConFunc - Functional annotation in the twilight zone’, *Bioinformatics*, 24(6), pp. 798–806. doi: 10.1093/bioinformatics/btn037.

Weghorn, D. and Sunyaev, S. (2017) ‘Bayesian inference of negative and positive selection in human cancers’, *Nature Genetics*. doi: 10.1038/ng.3987.

Weinstein, J. N. *et al.* (2013) ‘The cancer genome atlas pan-cancer analysis project’, *Nature Genetics*, 45(10), pp. 1113–1120. doi: 10.1038/ng.2764.

Weinstock, M. T. *et al.* (2016) ‘Vibrio natriegens as a fast-growing host for molecular biology’, *Nature Methods*. doi: 10.1038/nmeth.3970.

Weisberg, E. *et al.* (2015) ‘Inhibition of wild-type p53-expressing AML by the novel small molecule HDM2 inhibitor CGM097’, *Molecular Cancer Therapeutics*. doi: 10.1158/1535-7163.MCT-15-0429.

Westers, H. *et al.* (2003) ‘Genome Engineering Reveals Large Dispensable Regions in Bacillus subtilis’, *Molecular Biology and Evolution*. doi: 10.1093/molbev/msg219.

Wetterstrand KA. (2020) *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)*. Available at: https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data (Accessed: 7 October 2020).

Wu, C. H. *et al.* (2004) ‘PIRSF: Family classification system at the Protein Information Resource’, *Nucleic Acids Research*, 32(DATABASE ISS.), pp. D112–D114. doi: 10.1093/nar/gkh097.

Xie, Z. R. and Hwang, M. J. (2015) ‘Methods for predicting protein–ligand binding sites’, *Methods in Molecular Biology*, 1215, pp. 383–398. doi: 10.1007/978-1-4939-1465-4_17.

Xue, X. *et al.* (2015) ‘MEGA (Multiple Essential Genes Assembling) Deletion and Replacement Method for Genome Reduction in Escherichia coli’, *ACS Synthetic Biology*. doi: 10.1021/sb500324p.

Yang, J. *et al.* (2015) ‘The I-TASSER Suite: protein structure and function prediction’, *Nature Methods*, 12(1), pp. 7–8. doi: 10.1038/nmeth.3213.

Yang, J. *et al.* (2020) ‘Improved protein structure prediction using predicted interresidue orientations’, *Proceedings of the National Academy of Sciences of the United States of America*, 117(3), pp. 1496–1503. doi: 10.1073/pnas.1914677117.

Yang, J., Roy, A. and Zhang, Y. (2013) ‘BioLiP: A semi-manually curated database for biologically relevant ligand-protein interactions’, *Nucleic Acids Research*, 41(D1), pp. D1096–D1103. doi: 10.1093/nar/gks966.

Yates, A. D. *et al.* (2020) ‘Ensembl 2020’, *Nucleic Acids Research*, 48(D1), pp. D682–D688. doi: 10.1093/nar/gkz966.

You, R. *et al.* (2018) ‘GOLabeler: Improving sequence-based large-scale protein function prediction by learning to rank’, *Bioinformatics*, 34(14), pp. 2465–2473. doi: 10.1093/bioinformatics/bty130.

Yu, C. *et al.* (2014) ‘Discovery of biclonal origin and a novel oncogene SLC12A5 in colon cancer by single-cell sequencing’, *Cell Research*. doi: 10.1038/cr.2014.43.

Zhang, H. *et al.* (2020) ‘LncRNA SNHG14 promotes hepatocellular carcinoma progression via H3K27 acetylation activated PABPC1 by PTEN signaling', *Cell Death and Disease*. doi: 10.1038/s41419-020-02808-z.

Zhang, J. *et al.* (2014) 'Identifying driver mutations from sequencing data of heterogeneous tumors in the era of personalized genome sequencing', *Briefings in Bioinformatics*, 15(2), pp. 244–255. doi: 10.1093/bib/bbt042.

Zhao, E. Y., Jones, M. and Jones, S. J. M. (2019) 'Whole-genome sequencing in cancer', *Cold Spring Harbor Perspectives in Medicine*, 9(3), pp. 1–14. doi: 10.1101/cshperspect.a034579.

Zhao, F. *et al.* (2018) 'Computational approaches to prioritize cancer driver missense mutations', *International Journal of Molecular Sciences*, 19(7). doi: 10.3390/ijms19072113.

Zhao, J., Cao, Y. and Zhang, L. (2020) 'Exploring the computational methods for protein-ligand binding site prediction', *Computational and Structural Biotechnology Journal*. Elsevier B.V., pp. 417–426. doi: 10.1016/j.csbj.2020.02.008.

Zheng, W. *et al.* (2019) 'LOMETS2: improved meta-threading server for fold-recognition and structure-based function annotation for distant-homology proteins', *Nucleic Acids Research*, 47, pp. 429–436. doi: 10.1093/nar/gkz384.

Zhou, J. *et al.* (2016) 'CasHRA (Cas9-facilitated Homologous Recombination Assembly) method of constructing megabase-sized DNA', *Nucleic Acids Research*. doi: 10.1093/nar/gkw475.

Zhou, M. *et al.* (2013) 'Non-optimal codon usage affects expression, structure and function of clock protein FRQ', *Nature*, 494(7439), pp. 111–115. doi: 10.1038/nature11833.

Zhou, N. *et al.* (2019) 'The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens', *Genome Biology*, 20(1), pp. 1–23. doi: 10.1186/s13059-019-1835-8.

List of supplementary figures and tables

Appendix 1

Figure S1 Orthologs of proteins in the minimal bacterial genome. The number of orthologs for each protein identified in A) archaea and B) eukaryota. Results for each functional class are represented by a different colour: golden for the Unknown functional class, yellow – Generic, light turquoise – Putative, turquoise – Probable, dark turquoise – Equivalog. C) Summary of the total number of orthologs identified across different phyla for each of the functional confidence groups. The names of phyla from eukaryota are displayed in black, bacteria in red and archaea in grey.

Figure S2 Confidence of the top structural template identified by Phyre2. The confidence score (0-100) is shown for the top scoring template identified for each of the proteins in the minimal genome. The score indicates the confidence that the template protein sequence and the minimal genome protein sequence are homologs. Results for each functional class are represented by a different colour: golden for the Unknown functional class, yellow – Generic, light turquoise – Putative, turquoise – Probable, dark turquoise – Equivalog.

Figure S3 Examples of proteins in the minimal bacterial genome that where it was difficult to predict their function. A) Protein MMSYN_0138 was previously completely uncharacterised and listed as a hypothetical protein. Predictions for MMSYN_0138 by multiple methods identify a relationship to ATP binding domains of ABC transporters but the functional residues involved in ATP binding are not conserved making this function less likely. B) Protein MMSYN_0615 was previously classified as a tRNA binding protein in the Generic confidence class. Multiple predictions suggest that it could be a Phenylalanine- tRNA ligase β subunit, however the β subunit in other bacteria typically contains around 800 residues, whereas MMSYN_0615 is only 202 residues. It therefore seems that tRNA binding is likely but the role of this function is not known.

Figure S4 Transporter function prediction for the OppABCDF operon. Multiple sources made confident prediction for the proteins of the oligopeptide transporter system OppABCDF (AmiABCDE). These proteins form an operon in the

original *M. mycoides* subsp. *capri* and in the minimal genome. A) Permease OppB (AmiC) B) Permease OppC (AmiD) C) ATP-binding protein OppD (AmiE) D) ATP-binding protein OppF (AmiF) E) Oligopeptide binding protein OppA (AmiA).

Figure S5 Transporter function prediction for the potABCD operon. Multiple sources made confident prediction for the of the spermidine/putrescine transporter system potABCD were moved to the Putative class based on function predicted using confident results from multiple sources. These proteins form an operon in the original *M. mycoides* subsp. *capri* and in the minimal genome. A) Permease subunit potCD B) Permease subunit potB C) ATP-binding subunit potA.

Figure S6 Functional annotations where confidence was increased. This figure shows the proteins of unknown function that remained in the same specificity class. Results for each specificity class are represented by a different colour: beige for the Hypothetical specificity class, orange – General, light brown – Specific and dark brown – Highly specific. A) Each column represents a protein in the minimal genome and the squares show the methods that made predictions (darker colours indicate support of the final prediction), grey squares indicate predictions that did not support the function, light squares indicate that a method did not make a prediction. Proteins are grouped by their initial specificity class (Hypothetical, General, Specific and Highly specific) B) Boxplot showing the distribution of scores associated with the annotated functions. Proteins are grouped by their initial specificity class. Horizontal lines represent the median, the lower and upper hinge show respectively first quartile and third quartile, and lower and upper whisker include scores from first quartile to (distance between the first and third quartile)*1.5 (for lower whisker) and from third quartile to (distance between the first and third quartile)*1.5 (for upper whisker). Any scores outside of these intervals are shown as points (outliers). C) Number of methods supporting the function and the average score. Each point represents a protein. Note that the point at 0,0 represents multiple proteins classed as Hypothetical where it was not possible to assign any function.

Figure S7 Distribution of scores for matches to HAMAP. This the scores for HAMAP results for the minimal genome proteins of known function (Putative, Probable and Equivalog functional classes) are plotted. Results for each functional class are

represented by a different colour: light turquoise – Putative, turquoise – Probable, dark turquoise – Equivalog.

Figure S8 Distribution of scores from ProSiteProfiles results. This figure plots the scores for ProSiteProfiles results for the minimal genome proteins of known function (Putative, Probable and Equivalog functional classes). Results for each functional class are represented by a different colour: light turquoise – Putative, turquoise – Probable, dark turquoise – Equivalog.

Table S1 Examples of protein functions for the specificity classes.

Table S2 Comparison of the predictions made by individual methods and the final annotation assigned by the combination of methods. For each individual method we counted the predictions that agreed with the final annotation assigned to the protein (column yes – final) and if they more generally agreed with the assigned function (yes – general).

Table S3 Common predictions made by the five methods with greatest agreement with the final annotation. For each pair of methods the number of proteins where both methods make the same prediction as the final annotation is shown.

Appendix 2

Supplementary Data File 1: Orthologues of the minimal genome proteins identified using eggNOG-Mapper.

Supplementary Data File 2: Domains identified in the minimal genome proteins. Results are shown for search against the Pfam and TIGRFAM databases of protein families.

Supplementary Data File 3: Structural modelling of the minimal genome proteins. Results for modelling of the proteins using the Phyre2 server.

Supplementary Data File 4: Membrane protein predictions for the minimal genome proteins.

Supplementary Data File 5: Inferred functions of the proteins encoded by the minimal bacterial genome. The original annotation and the predicted functions from the analysis performed here are shown.

Supplementary Data File 6: Gene Ontology-based function predictions for the proteins encoded by the minimal genome.

Supplementary Data File 7: Comparison of the predicted functions of the minimal genome proteins with predictions made by Danchin and Fang.

Supplementary Data File 8: Interpro predictions for the proteins encoded by the minimal bacterial genome.

Appendix 3

Figure S1 Orthologs of proteins in the minimal bacterial genome. The number of orthologs for each protein identified in A) archaea and B) eukaryota. Results for each functional class are represented by a different colour: golden for the Unknown functional class, yellow – Generic, light turquoise – Putative, turquoise – Probable, dark turquoise – Equivalog. C) Summary of the total number of orthologs identified across different phyla for each of the functional confidence groups. The names of phyla from eukaryota are displayed in black, bacteria in red and archaea in grey.

Figure S2 Confidence of the top structural template identified by Phyre2. The confidence score (0-100) is shown for the top scoring template identified for each of the proteins in the minimal genome. The score indicates the confidence that the template protein sequence and the minimal genome protein sequence are homologs. Results for each functional class are represented by a different colour: golden for the Unknown functional class, yellow – Generic, light turquoise – Putative, turquoise – Probable, dark turquoise – Equivalog.

Figure S3 Examples of proteins in the minimal bacterial genome that where it was difficult to predict their function. A) Protein MMSYN_0138 was previously completely uncharacterised and listed as a hypothetical protein. Predictions for MMSYN_0138 by multiple methods identify a relationship to ATP binding domains of ABC transporters but the functional residues involved in ATP binding are not conserved making this function less likely. B) Protein MMSYN_0615 was previously classified as a tRNA binding protein in the Generic confidence class. Multiple predictions suggest that it could be a Phenylalanine- tRNA ligase β subunit, however the β subunit in other bacteria typically contains around 800 residues, whereas MMSYN_0615 is only 202 residues. It therefore seems that tRNA binding is likely but the role of this function is not known.

Figure S4 Transporter function prediction for the OppABCDF operon. Multiple sources made confident prediction for the proteins of the oligopeptide transporter system OppABCDF (AmiABCDE). These proteins form an operon in the original *M. mycoides* subsp. *capri* and in the minimal genome. A) Permease OppB (AmiC) B) Permease OppC (AmiD) C) ATP-binding protein OppD (AmiE) D) ATP-binding protein OppF (AmiF) E) Oligopeptide binding protein OppA (AmiA).

Figure S5 Transporter function prediction for the potABCD operon. Multiple sources made confident prediction for the of the spermidine/putrescine transporter system potABCD were moved to the Putative class based on function predicted using confident results from multiple sources. These proteins form an operon in the original *M. mycoides* subsp. *capri* and in the minimal genome. A) Permease subunit potCD B) Permease subunit potB C) ATP-binding subunit potA.

Figure S6 Functional annotations where confidence was increased. This figure shows the proteins of unknown function that remained in the same specificity class. Results for each specificity class are represented by a different colour: beige for the Hypothetical specificity class, orange – General, light brown – Specific and dark brown – Highly specific. A) Each column represents a protein in the minimal genome and the squares show the methods that made predictions (darker colours indicate support of the final prediction), grey squares indicate predictions that did not support the function, light squares indicate that a method did not make a prediction. Proteins are grouped by their initial specificity class (Hypothetical, General, Specific and Highly specific) B) Boxplot showing the distribution of scores associated with the annotated functions. Proteins are grouped by their initial specificity class. Horizontal lines represent the median, the lower and upper hinge show respectively first quartile and third quartile, and lower and upper whisker include scores from first quartile to (distance between the first and third quartile)*1.5 (for lower whisker) and from third quartile to (distance between the first and third quartile)*1.5 (for upper whisker). Any scores outside of these intervals are shown as points (outliers). C) Number of methods supporting the function and the average score. Each point represents a protein. Note that the point at 0,0 represents multiple proteins classed as Hypothetical where it was not possible to assign any function.

Figure S7 Distribution of scores for matches to HAMAP. This the scores for HAMAP results for the minimal genome proteins of known function (Putative, Probable and Equivalog functional classes) are plotted. Results for each functional class are represented by a different colour: light turquoise – Putative, turquoise – Probable, dark turquoise – Equivalog.

Figure S8 Distribution of scores from ProSiteProfiles results. This figure plots the scores for ProSiteProfiles results for the minimal genome proteins of known function (Putative, Probable and Equivalog functional classes). Results for each functional class are represented by a different colour: light turquoise – Putative, turquoise – Probable, dark turquoise – Equivalog.

Figure S9 Pipeline for calling and analysing Single Nucleotide Variants.

Figure S10 Mutational profile of the Molm13 parental cell - screenshot from COSMIC's Cell Lines Project (Sample overview for 1330947, 2020)

Figure S11 Selection of candidates for drivers. Classification of the variants into de novo, gained, higher allelic ratio, same, lower allelic ratio, not called and lost.

Figure S12 Scheme representing our in-house pipeline for calling and filtering of SNVs and INDELs.

Appendix 4

Supplementary Table 1: Acquired (de novo, gained, higher allelic ratio) high-quality, high-coverage, somatic-like, most-likely damaging variants shared between Molm13 sub-lines adapted to nutlin-3.

Supplementary Table 2: De novo high-quality, high-coverage, somatic-like, most-likely damaging variants shared between Molm13 sub-lines adapted to nutlin-3.

Supplementary Table 3: High-quality, high-coverage, somatic-like, most-likely damaging variants from Molm13 parental cell line that are not retained (lost, not called, lower allelic ratio) in nutlin-3-adapted sub-lines.

Supplementary Table 4: High-quality, high-coverage, somatic-like, most-likely damaging variants shared between Molm13 sub-lines adapted to nutlin-3.

Appendix 5

Appendix 6

Supplementary Table 1: Characterisation of de novo high-quality, high-coverage, somatic-like, most-likely damaging variants identified in UKF-NB-3 sub-lines adapted to eribulin, vincristine, epothilone b and 2-methoxyestradiol.

Appendix 1

Environmental conditions shape the nature of a minimal bacterial genome

Magdalena Antczak, Martin Michaelis*, Mark N Wass*

School of Biosciences, University of Kent, Canterbury, Kent, CT2 7NJ, UK

*to whom correspondence should be addressed: m.n.wass@kent.ac.uk
m.michaelis@kent.ac.uk

Supplementary Material

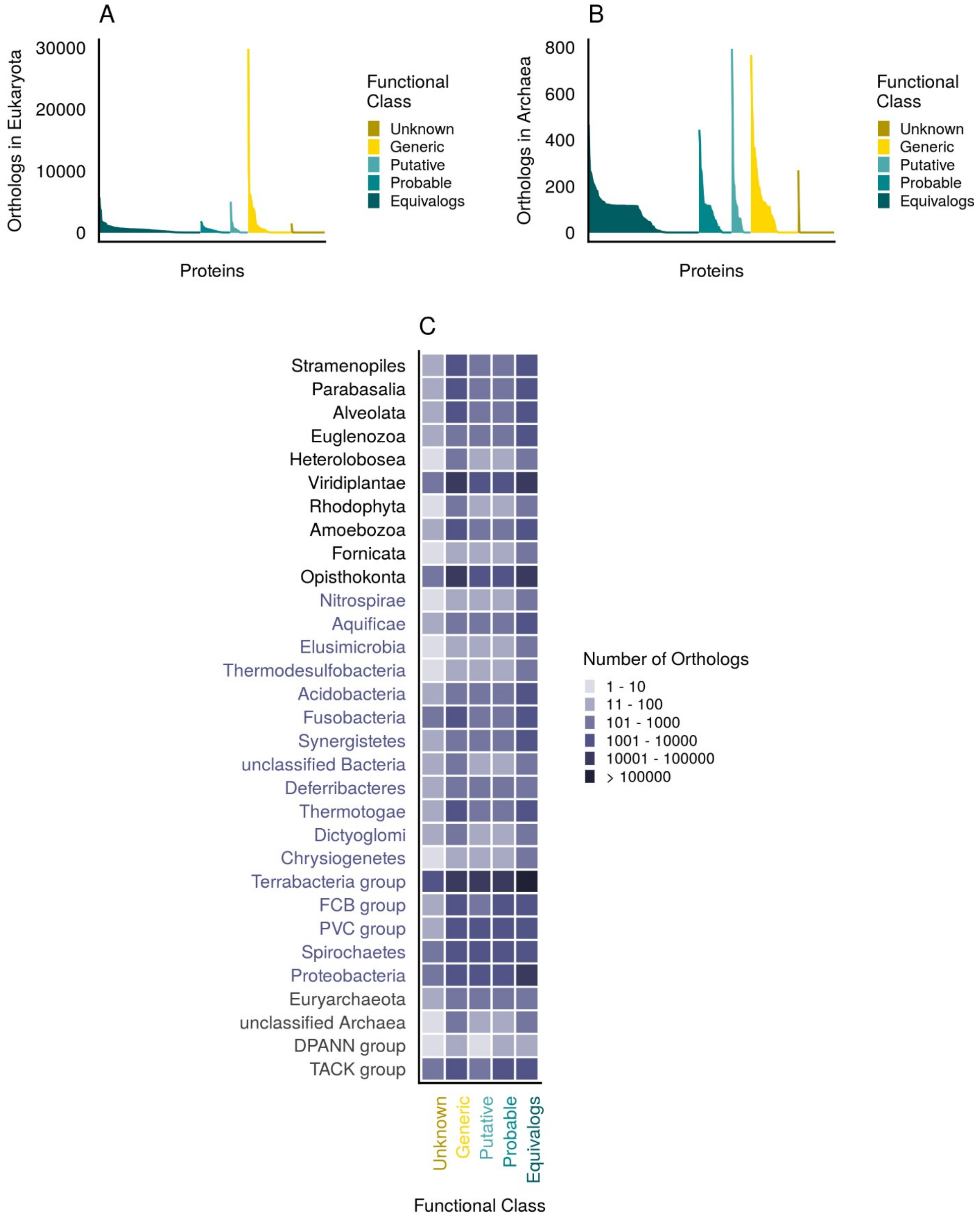


Figure S1 *Orthologs of proteins in the minimal bacterial genome. The number of orthologs for each protein identified in A) archaea and B) eukaryota. Results for each functional class are represented by a different colour: golden for the Unknown functional class, yellow – Generic, light turquoise – Putative, turquoise – Probable, dark turquoise – Equivalog. C) Summary of the total number of orthologs identified across different phyla for each of the functional confidence groups. The names of phyla from eukaryota are displayed in black, bacteria in red and archaea in grey.*

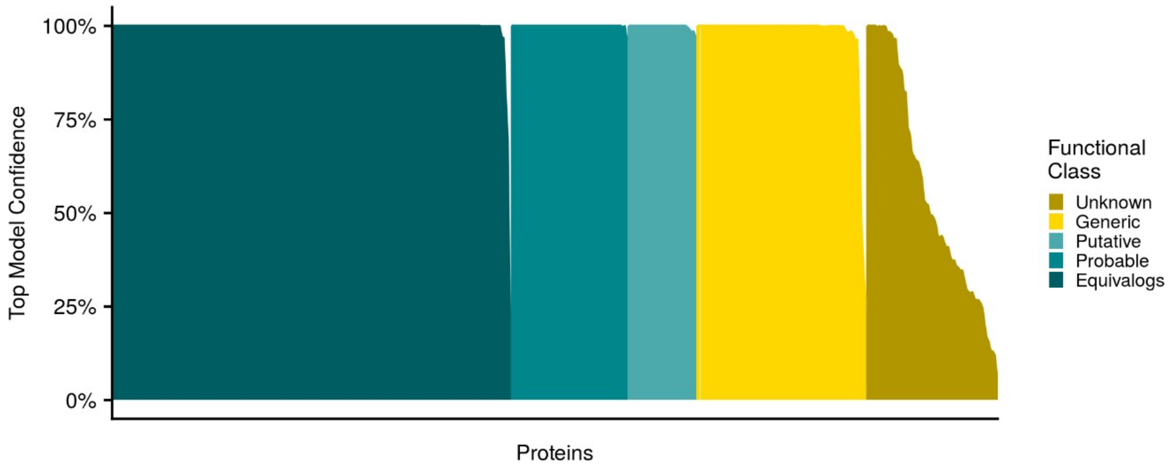
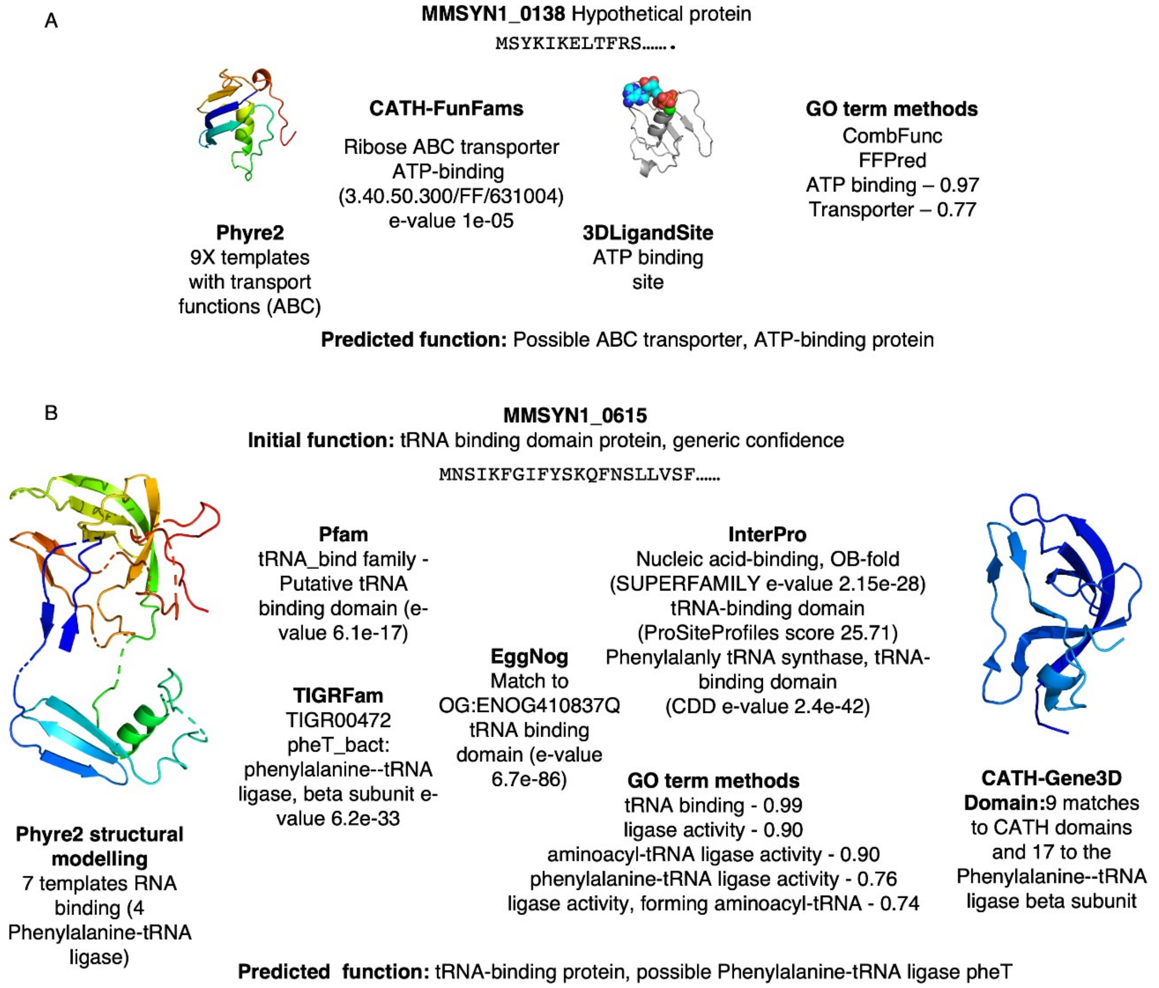


***Figure S2** Confidence of the top structural template identified by Phyre2. The confidence score (0-100) is shown for the top scoring template identified for each of the proteins in the minimal genome. The score indicates the confidence that the template protein sequence and the minimal genome protein sequence are homologs. Results for each functional class are represented by a different colour: golden for the Unknown functional class, yellow – Generic, light turquoise – Putative, turquoise – Probable, dark turquoise – Equivalog.*



MMSYN1_0138 Hypothetical protein

MSYKIKELTFRS…….

CATH-FunFams
Ribose ABC transporter
ATP-binding
(3.40.50.300/FF/631004)
e-value 1e-05

GO term methods
CombFunc
FFPred
ATP binding – 0.97
Transporter – 0.77

Phyre2
9X templates
with transport
functions (ABC)

3DLigandSite
ATP binding
site

Predicted function: Possible ABC transporter, ATP-binding protein

MMSYN1_0615
Initial function: tRNA binding domain protein, generic confidence

MNSIKFGIFYSKQFNSLLVSF……

Pfam
tRNA_bind family - Putative tRNA binding domain (e-value 6.1e-17)

InterPro
Nucleic acid-binding, OB-fold
(SUPERFAMILY e-value 2.15e-28)
tRNA-binding domain
(ProSiteProfiles score 25.71)
Phenylalanly tRNA synthase, tRNA-binding domain
(CDD e-value 2.4e-42)

EggNog
Match to
OG:ENOG410837Q
tRNA binding
domain (e-value
6.7e-86)

TIGRFam
TIGR00472
pheT_bact:
phenylalanine--tRNA
ligase, beta subunit e-value 6.2e-33

GO term methods
tRNA binding - 0.99
ligase activity - 0.90
aminoacyl-tRNA ligase activity - 0.90
phenylalanine-tRNA ligase activity - 0.76
ligase activity, forming aminoacyl-tRNA - 0.74

CATH-Gene3D Domain:9 matches to CATH domains and 17 to the Phenylalanine--tRNA ligase beta subunit

Phyre2 structural modelling
7 templates RNA binding (4 Phenylalanine-tRNA ligase)

Predicted function: tRNA-binding protein, possible Phenylalanine-tRNA ligase pheT

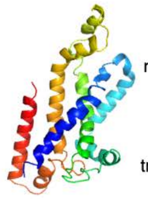
***Figure S3** Examples of proteins in the minimal bacterial genome that where it was difficult to predict their function. A) Protein MMSYN_0138 was previously completely uncharacterised and listed as a hypothetical*

protein. Predictions for MMSYN_0138 by multiple methods identify a relationship to ATP binding domains of ABC transporters but the functional residues involved in ATP binding are not conserved making this function less likely. B) Protein MMSYN_0615 was previously classified as a tRNA binding protein in the Generic confidence class. Multiple predictions suggest that it could be a Phenylalanine- tRNA ligase β subunit, however the β subunit in other bacteria typically contains around 800 residues, whereas MMSYN_0615 is only 202 residues. It therefore seems that tRNA binding is likely but the role of this function is not known.

A

MMSYN1_0165
Initial function: AmiC?
Confidence class: Generic

MKSTLKTKQEVLNLNSELLL......

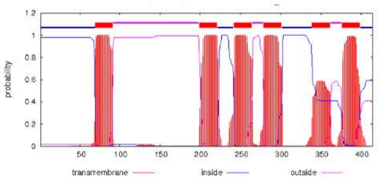


Phyre2 structural modelling
Several confident matches to templates of ABC transporter, permease

TIGRFam
nickel_nikB: nickel ABC transporter, permease subu (e-value 4e-22)

Pfam
Binding-protein-dependent transport system inner membrane component (e-value 1.1e-22)

TrSSP
Amino acid transporter



TMHMM
Predicted 6 TM helices

CATH-FunFams
Oligopeptide ABC transporter permease (1.10.3720.10/FF/58662) e-value 1.7e-21

GO term methods
Multiple high confidence predictions associated with transmembrane transporter functions

EggNog
Match to OG: ENOG41082QK ABC transporter (Permease (e-value 1.85e-104)
Predicted gene: oppB

InterPro
ABC transporter type 1, transmembrane domain MetI-like (CDD e-value 1.87e-12; ProSiteProfiles score 10.45)

Predicted function: Oligopeptide ABC transporter, permease OppB (AmiC)

B

MMSYN1_0166
Initial function: AmiD?
Confidence class: Generic

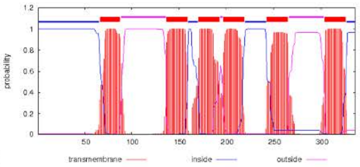
MKTKQLEQPDFSALLDSERE......

Phyre2 structural modelling
Several confident matches to templates of ABC transporter, permease

TIGRFam
nickel_nikC: nickel ABC transporter, permease subu (e-value 9.2e-25)

Pfam
Binding-protein-dependent transport system inner membrane component (e-value 6.4e-16)

TrSSP
Anion transporter



TMHMM
Predicted 6 TM helices

CATH-FunFams
Oligopeptide ABC transporter permease OppC (1.10.3720.10/FF/58605) e-value 1.2e-29

GO term methods
Multiple high confidence predictions associated with transmembrane transporter functions

EggNog
Match to OG:ENOG4107SI6 transport system permease protein (e-value 1.67e-146)
Predicted gene: oppC2

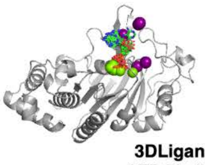
InterPro
ABC transporter type 1, transmembrane domain MetI-like (CDD e-value 7.43e-09; ProSiteProfiles score 17.83)

Predicted function: Oligopeptide ABC transporter, permease OppC (AmiD)

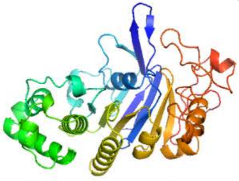
C

MMSYN1_0167
Initial function: AmiE?
Confidence class: Generic

MKNVILSIKDLVVKFRVRSK......



3DLigandSite
ATP binding site



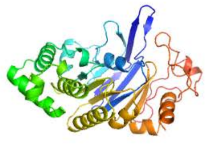
Phyre2 structural modelling
38 templates with ABC transporters, ATP-binding

Pfam
ATP-binding domain of ABC transporters (e-value 7.4e-09)
Oligopeptide/dipeptide transporter, C-terminal region (e-value 1.4e-10)

EggNog
Match to OG:ENOG411CA8C oligopeptide abc transporter atp-binding protein (4.4e-175)
Predicted gene: oppD

TIGRFam
Matches to ABC transporters, ATP-binding component

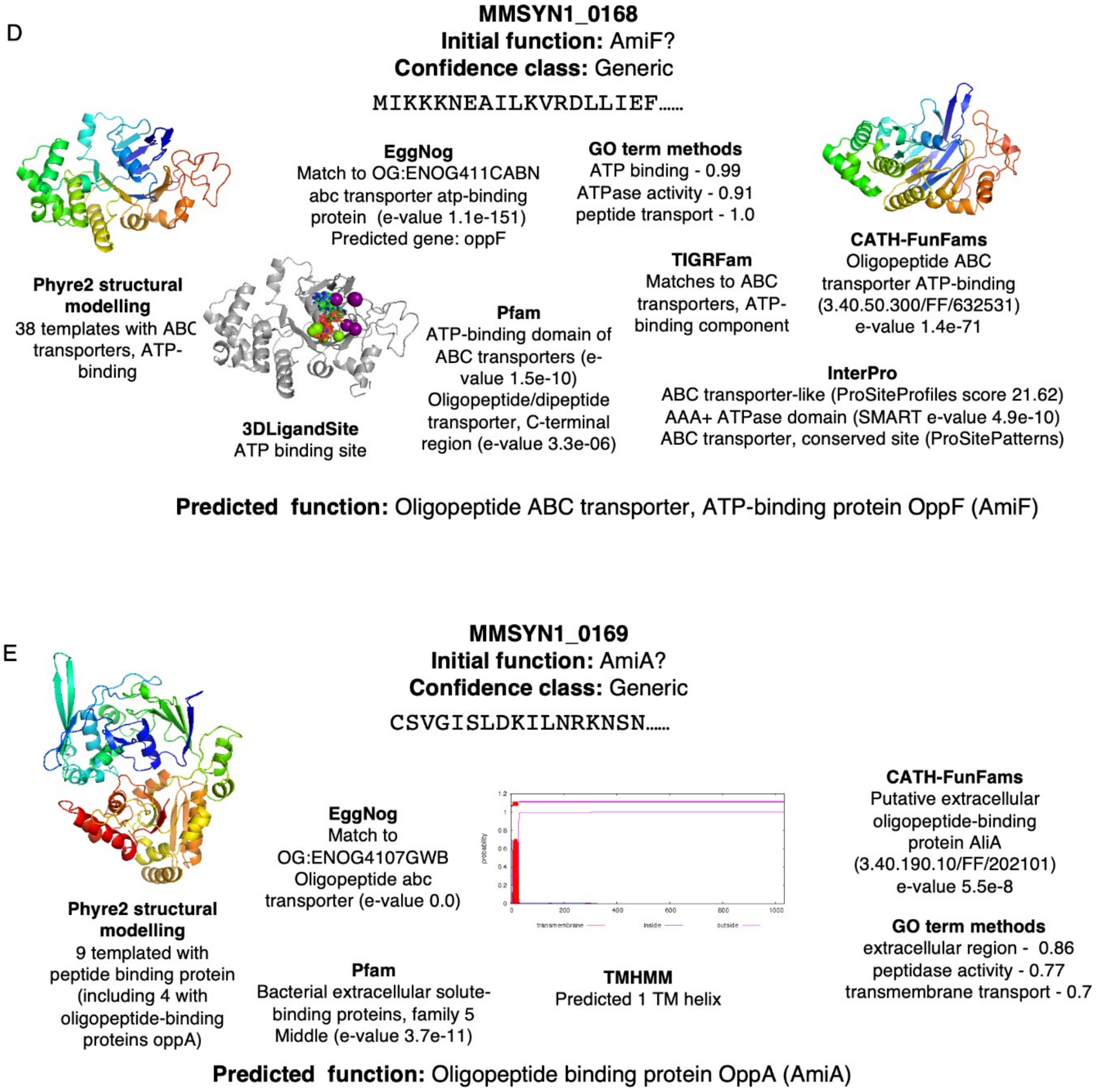
GO term methods
ATP binding - 0.99
ATPase activity - 0.92
peptide transport - 1.0



CATH-FunFams
Oligopeptide ABC transporter ATP-binding (3.40.50.300/FF/632531) e-value 3.5e-81

InterPro
ABC transporter-like (ProSiteProfiles score 19.48)
AAA+ ATPase domain (SMART e-value 3.2e-15)
ABC transporter, conserved site (ProSitePatterns)

Predicted function: Oligopeptide ABC transporter, ATP-binding protein OppD (AmiE)



D

MMSYN1_0168
Initial function: AmiF?
Confidence class: Generic

MIKKKNEAILKVRDLLIEF……

EggNog
Match to OG:ENOG411CABN abc transporter atp-binding protein (e-value 1.1e-151)
Predicted gene: oppF

GO term methods
ATP binding - 0.99
ATPase activity - 0.91
peptide transport - 1.0

CATH-FunFams
Oligopeptide ABC transporter ATP-binding (3.40.50.300/FF/632531)
e-value 1.4e-71

Phyre2 structural modelling
38 templates with ABC transporters, ATP-binding

TIGRFam
Matches to ABC transporters, ATP-binding component

Pfam
ATP-binding domain of ABC transporters (e-value 1.5e-10)
Oligopeptide/dipeptide transporter, C-terminal region (e-value 3.3e-06)

3DLigandSite
ATP binding site

InterPro
ABC transporter-like (ProSiteProfiles score 21.62)
AAA+ ATPase domain (SMART e-value 4.9e-10)
ABC transporter, conserved site (ProSitePatterns)

Predicted function: Oligopeptide ABC transporter, ATP-binding protein OppF (AmiF)

E

MMSYN1_0169
Initial function: AmiA?
Confidence class: Generic

CSVGISLDKILNRKNSN……

EggNog
Match to OG:ENOG4107GWB Oligopeptide abc transporter (e-value 0.0)

CATH-FunFams
Putative extracellular oligopeptide-binding protein AliA (3.40.190.10/FF/202101)
e-value 5.5e-8

GO term methods
extracellular region - 0.86
peptidase activity - 0.77
transmembrane transport - 0.7

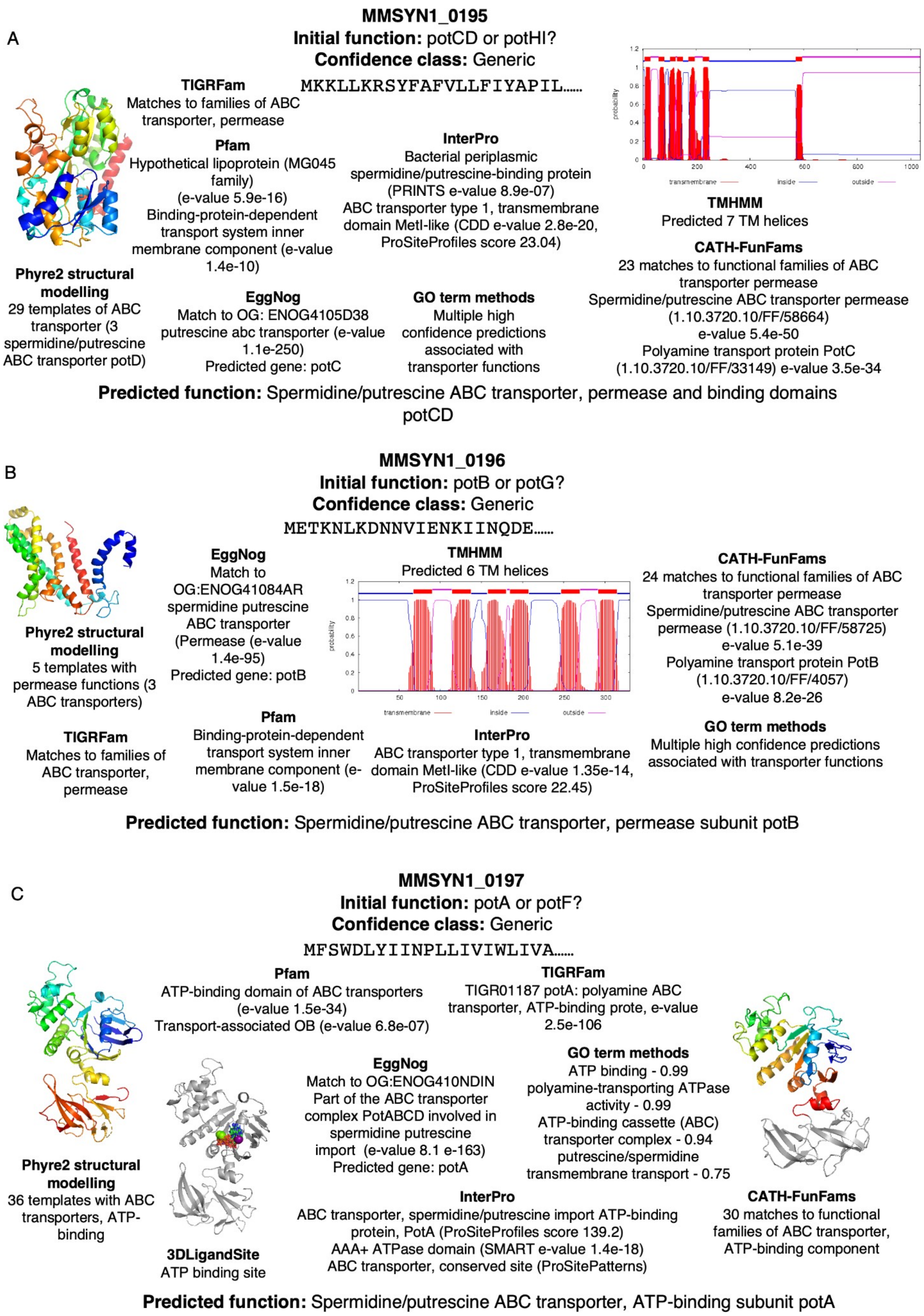
Phyre2 structural modelling
9 templated with peptide binding protein (including 4 with oligopeptide-binding proteins oppA)

Pfam
Bacterial extracellular solute-binding proteins, family 5 Middle (e-value 3.7e-11)

TMHMM
Predicted 1 TM helix

Predicted function: Oligopeptide binding protein OppA (AmiA)

Figure S4 *Transporter function prediction for the OppABCDF operon. Multiple sources made confident prediction for the proteins of the oligopeptide transporter system OppABCDF (AmiABCDE). These proteins form an operon in the original M. mycoides subsp. capri and in the minimal genome. A) Permease OppB (AmiC) B) Permease OppC (AmiD) C) ATP-binding protein OppD (AmiE) D) ATP-binding protein OppF (AmiF) E) Oligopeptide binding protein OppA (AmiA).*



MMSYN1_0195
Initial function: potCD or potHI?
Confidence class: Generic
MKKLLKRSYFAFVLLFIYAPIL......

TIGRFam
Matches to families of ABC transporter, permease

Pfam
Hypothetical lipoprotein (MG045 family) (e-value 5.9e-16)
Binding-protein-dependent transport system inner membrane component (e-value 1.4e-10)

InterPro
Bacterial periplasmic spermidine/putrescine-binding protein (PRINTS e-value 8.9e-07)
ABC transporter type 1, transmembrane domain MetI-like (CDD e-value 2.8e-20, ProSiteProfiles score 23.04)

TMHMM
Predicted 7 TM helices

Phyre2 structural modelling
29 templates of ABC transporter (3 spermidine/putrescine ABC transporter potD)

EggNog
Match to OG: ENOG4105D38 putrescine abc transporter (e-value 1.1e-250)
Predicted gene: potC

GO term methods
Multiple high confidence predictions associated with transporter functions

CATH-FunFams
23 matches to functional families of ABC transporter permease
Spermidine/putrescine ABC transporter permease (1.10.3720.10/FF/58664)
e-value 5.4e-50
Polyamine transport protein PotC (1.10.3720.10/FF/33149) e-value 3.5e-34

Predicted function: Spermidine/putrescine ABC transporter, permease and binding domains potCD

MMSYN1_0196
Initial function: potB or potG?
Confidence class: Generic
METKNLKDNNVIENKIINQDE......

EggNog
Match to OG:ENOG41084AR spermidine putrescine ABC transporter (Permease (e-value 1.4e-95)
Predicted gene: potB

TMHMM
Predicted 6 TM helices

CATH-FunFams
24 matches to functional families of ABC transporter permease
Spermidine/putrescine ABC transporter permease (1.10.3720.10/FF/58725)
e-value 5.1e-39
Polyamine transport protein PotB (1.10.3720.10/FF/4057)
e-value 8.2e-26

Phyre2 structural modelling
5 templates with permease functions (3 ABC transporters)

TIGRFam
Matches to families of ABC transporter, permease

Pfam
Binding-protein-dependent transport system inner membrane component (e-value 1.5e-18)

InterPro
ABC transporter type 1, transmembrane domain MetI-like (CDD e-value 1.35e-14, ProSiteProfiles score 22.45)

GO term methods
Multiple high confidence predictions associated with transporter functions

Predicted function: Spermidine/putrescine ABC transporter, permease subunit potB

MMSYN1_0197
Initial function: potA or potF?
Confidence class: Generic
MFSWDLYIINPLLIVIWLIVA......

Pfam
ATP-binding domain of ABC transporters (e-value 1.5e-34)
Transport-associated OB (e-value 6.8e-07)

TIGRFam
TIGR01187 potA: polyamine ABC transporter, ATP-binding prote, e-value 2.5e-106

EggNog
Match to OG:ENOG410NDIN
Part of the ABC transporter complex PotABCD involved in spermidine putrescine import (e-value 8.1 e-163)
Predicted gene: potA

GO term methods
ATP binding - 0.99
polyamine-transporting ATPase activity - 0.99
ATP-binding cassette (ABC) transporter complex - 0.94
putrescine/spermidine transmembrane transport - 0.75

Phyre2 structural modelling
36 templates with ABC transporters, ATP-binding

3DLigandSite
ATP binding site

InterPro
ABC transporter, spermidine/putrescine import ATP-binding protein, PotA (ProSiteProfiles score 139.2)
AAA+ ATPase domain (SMART e-value 1.4e-18)
ABC transporter, conserved site (ProSitePatterns)

CATH-FunFams
30 matches to functional families of ABC transporter, ATP-binding component

Predicted function: Spermidine/putrescine ABC transporter, ATP-binding subunit potA

Figure S5 *Transporter function prediction for the potABCD operon. Multiple sources made confident prediction for the of the spermidine/putrescine transporter system potABCD were moved to the Putative class based on function predicted using confident results from multiple sources. These proteins form an*

operon in the original M. mycoides subsp. capri and in the minimal genome. A) Permease subunit potCD B) Permease subunit potB C) ATP-binding subunit potA.

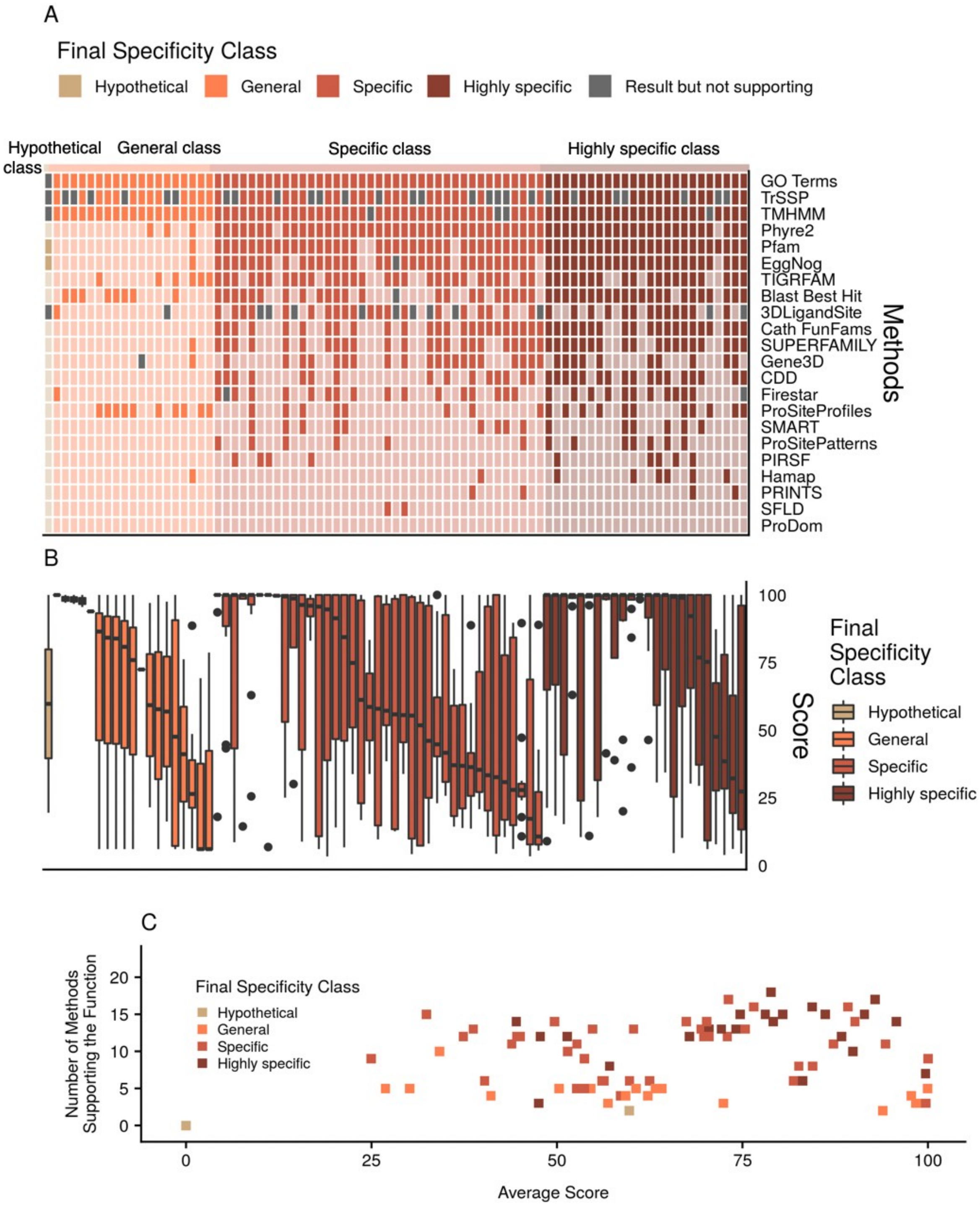


Figure S6 *Functional annotations where confidence was increased. This figure shows the proteins of unknown function that remained in the same specificity class. Results for each specificity class are represented by a different colour: beige for the Hypothetical specificity class, orange – General, light brown – Specific and dark brown – Highly specific. A) Each column represents a protein in the minimal genome and the squares show the methods that made predictions (darker colours indicate support of the final prediction), grey squares indicate predictions that did not support the function, light squares indicate that a*

*method did not make a prediction. Proteins are grouped by their initial specificity class (Hypothetical, General, Specific and Highly specific) B) Boxplot showing the distribution of scores associated with the annotated functions. Proteins are grouped by their initial specificity class. Horizontal lines represent the median, the lower and upper hinge show respectively first quartile and third quartile, and lower and upper whisker include scores from first quartile to (distance between the first and third quartile)*1.5 (for lower whisker) and from third quartile to (distance between the first and third quartile)*1.5 (for upper whisker). Any scores outside of these intervals are shown as points (outliers). C) Number of methods supporting the function and the average score. Each point represents a protein. Note that the point at 0,0 represents multiple proteins classed as Hypothetical where it was not possible to assign any function.*

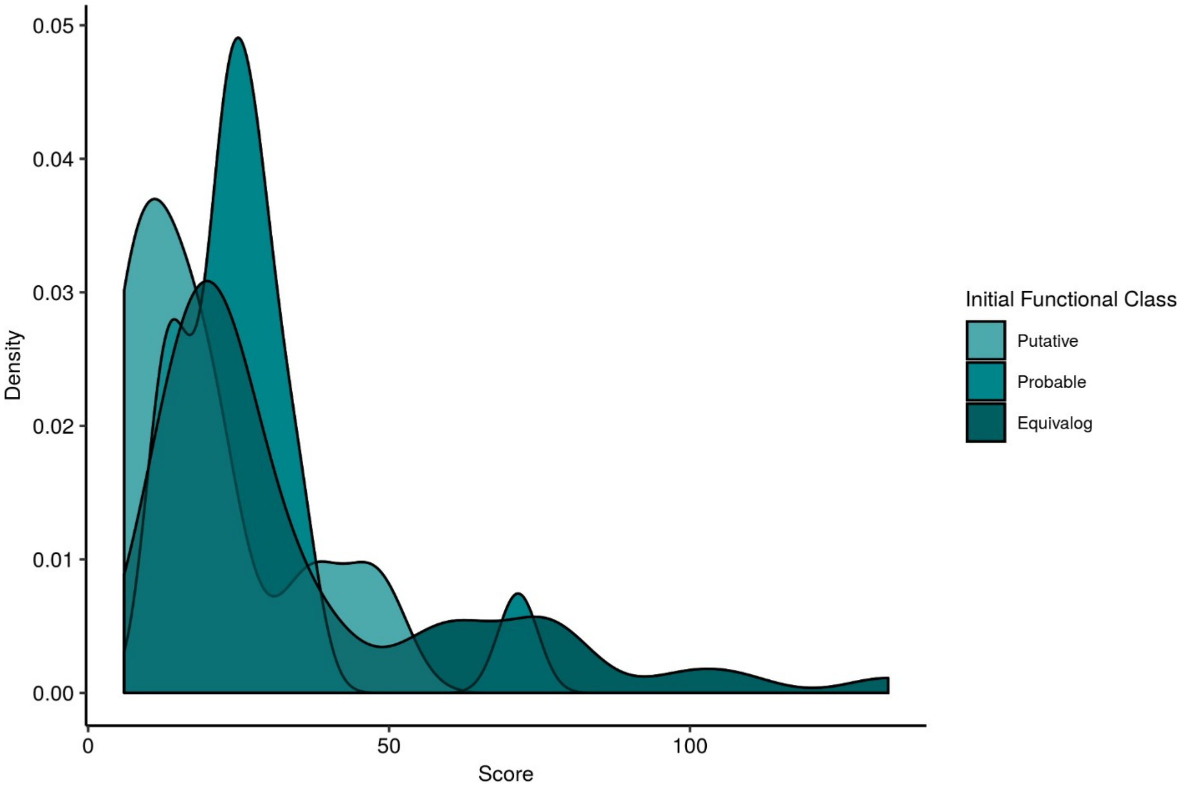


Figure S7 *Distribution of scores for matches to HAMAP. This the scores for HAMAP results for the minimal genome proteins of known function (Putative, Probable and Equivalog functional classes) are plotted. Results for each functional class are represented by a different colour: light turquoise – Putative, turquoise – Probable, dark turquoise – Equivalog.*

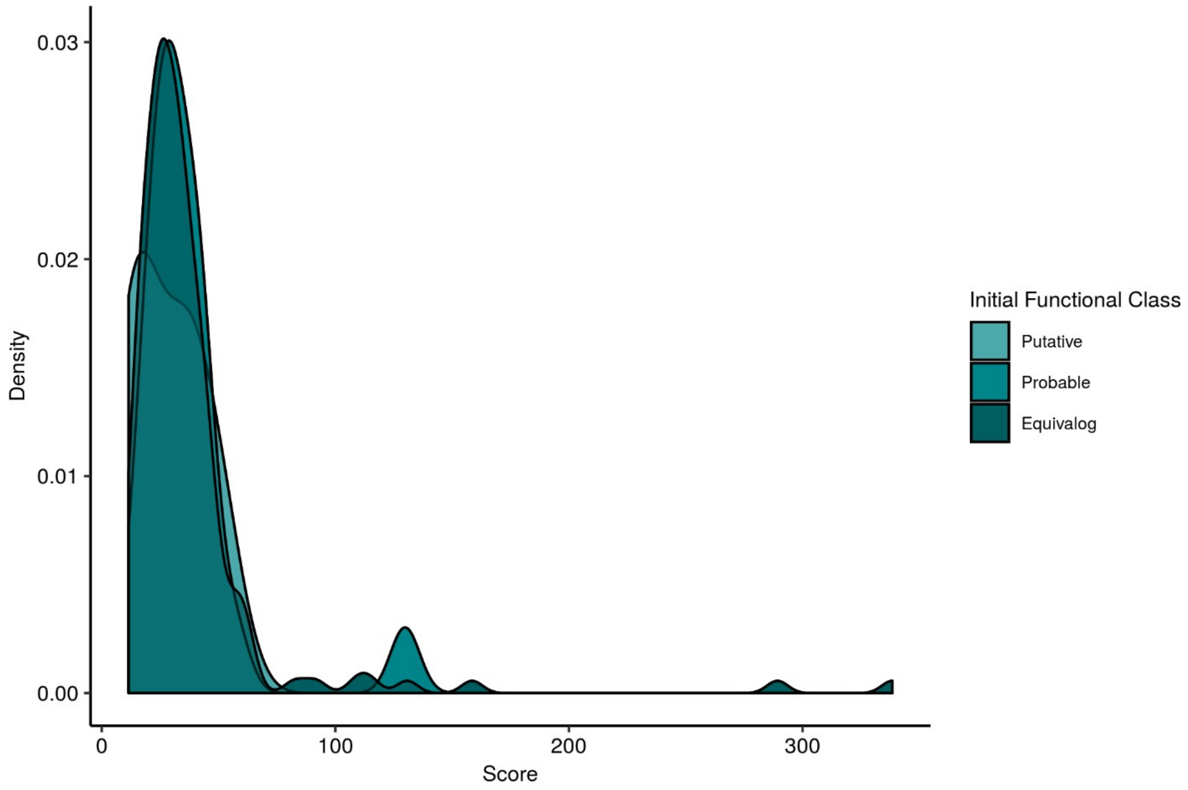


Figure S8 *Distribution of scores from ProSiteProfiles results. This figure plots the scores for ProSiteProfiles results for the minimal genome proteins of known function (Putative, Probable and Equivalog functional classes). Results for each functional class are represented by a different colour: light turquoise – Putative, turquoise – Probable, dark turquoise – Equivalog.*

General	**Specific**	**Highly specific**
Transcription factor	Transcriptional regulator, RpiR family	whiA; Sporulation transcription regulator WhiA
Ribosomal protein	Ribosomal protein L7Ae/L30e family	rpmH; 50S ribosomal protein L34
Transmembrane protein, likely a transporter	ABC transporter, ATP-binding protein	oppD; Oligopeptide ABC transporter, ATP-binding protein
Membrane metallopeptidase	Transmembrane peptidase, C39 family	pepQ; Xaa-Pro dipeptidase
DNA-binding protein	ATP-dependent DNA helicase	polA; DNA polymerase I

Table S1 *Examples of protein functions for the specificity classes.*

Methods	**Number of proteins**				**Percentage**			
	Yes (final)	**Yes (general)**	**No**	**No prediction**	**Yes (final)**	**Yes (general)**	**No**	**No prediction**

eggNOG-Mapper	55	22	1	71	37%	15%	1%	48%
GO Terms	53	80	0	16	36%	54%	0%	11%
Phyre2	53	32	0	64	36%	21%	0%	43%
BLAST against UniProt top match	51	20	1	77	34%	13%	1%	52%
Pfam	49	34	0	66	33%	23%	0%	44%
CATH FunFams	45	16	2	86	30%	11%	1%	58%
TIGRFAM	41	24	1	83	28%	16%	1%	56%
InterPro ProSiteProfile s	21	12	0	116	14%	8%	0%	78%
InterPro CDD	21	21	0	107	14%	14%	0%	72%
InterPro SUPERFAMIL Y	21	40	0	88	14%	27%	0%	59%
TrSSP	14	71	48	16	9%	48%	32%	11%
InterPro Gene3D	14	28	1	106	9%	19%	1%	71%
InterPro PIRSF	7	4	0	138	5%	3%	0%	93%
InterPro HAMAP	7	1	0	141	5%	1%	0%	95%
InterPro SMART	7	11	0	131	5%	7%	0%	88%
TMHMM	6	122	5	16	4%	82%	3%	11%
InterPro ProSitePatter ns	4	12	0	133	3%	8%	0%	89%
InterPro PRINTS	3	4	0	142	2%	3%	0%	95%

InterPro SFLD	2	1	0	146	1%	1%	0%	98%
3DLigandSite	0	44	20	85	0%	30%	13%	57%
Firestar	0	35	2	112	0%	23%	1%	75%
InterPro ProDom	0	1	0	148	0%	1%	0%	99%

***Table S2** Comparison of the predictions made by individual methods and the final annotation assigned by the combination of methods. For each individual method we counted the predictions that agreed with the final annotation assigned to the protein (column yes – final) and if they more generally agreed with the assigned function (yes – general).*

Method 1	**Method 2**	**Number of common proteins**	**Percentage**
EggNOG	BLAST - UniProt	38	25.5
EggNOG	Pfam	31	20.81
EggNOG	Phyre2	30	20.13
Phyre2	Pfam	28	18.79
Phyre2	BLAST - UniProt	26	17.45
BLAST - UniProt	Pfam	24	16.11
EggNOG	GO Terms	14	9.4
GO Terms	Phyre2	13	8.72
GO Terms	BLAST - UniProt	12	8.05
GO Terms	Pfam	9	6.04

***Table S3** Common predictions made by the five methods with greatest agreement with the final annotation. For each pair of methods the number of proteins where both methods make the same prediction as the final annotation is shown.*

Appendix 3

Acquired MDM2 inhibitor resistance is associated with collateral sensitivity to cytarabine in acute myeloid leukaemia cells

Magdalena Antczak[1#], Tamara Rothenburger[2#], Katie-May McLaughlin[1], Florian Rothweiler[2], Constanze Schneider[2], Björn Rotter[3], Daniel Speidel[4,5], Andrea Nist[6], Marco Mernberger[7], Dominique Thomas[8], Gerd Geisslinger[8,9], Thorsten Stiewe[6,7], Mark N. Wass[1]*, Martin Michaelis[1]*, Jindrich Cinatl jr.[2]*

[1] School of Biosciences, University of Kent, Canterbury, UK

[2] Institute for Medical Virology, Goethe-University, Frankfurt am Main, Germany

[3] GenXPro GmbH, Altenhöferallee 3, 60438 Frankfurt am Main, Germany

[4] Children's Medical Research Institute, Westmead, Australia

[5] Sydney Medical School, The University of Sydney, Australia

[6] Genomics Core Facility, Philipps-University, Marburg, Germany

[7] Institute of Molecular Oncology, Philipps-University, Marburg, Germany

[8] pharmazentrum frankfurt/ZAFES, Institute of Clinical Pharmacology, Goethe-University, Frankfurt am Main, Germany

[9] Fraunhofer Institute for Molecular Biology and Applied Ecology (IME), Project group Translational Medicine and Pharmacology (TMP), Frankfurt am Main, Germany

[#]Equal contribution

***Corresponding authors**

Mark N. Wass (M.N.Wass@kent.ac.uk), Martin Michaelis (m.michaelis@kent.ac.uk), Jindrich Cinatl jr. (Cinatl@kent.ac.uk)

Supplementary Material

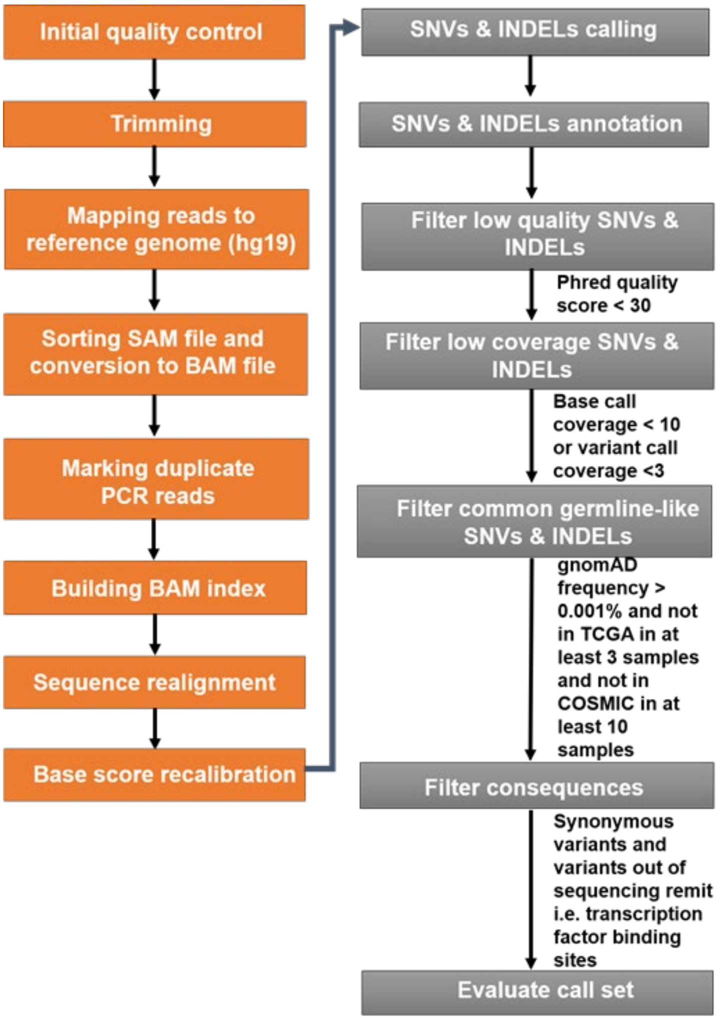

Initial quality control
Trimming
Mapping reads to reference genome (hg19)
Sorting SAM file and conversion to BAM file
Marking duplicate PCR reads
Building BAM index
Sequence realignment
Base score recalibration
SNVs & INDELs calling
SNVs & INDELs annotation
Filter low quality SNVs & INDELs
Phred quality score < 30
Filter low coverage SNVs & INDELs
Base call coverage < 10 or variant call coverage <3
Filter common germline-like SNVs & INDELs
gnomAD frequency > 0.001% and not in TCGA in at least 3 samples and not in COSMIC in at least 10 samples
Filter consequences
Synonymous variants and variants out of sequencing remit i.e. transcription factor binding sites
Evaluate call set

Figure S9 *Pipeline for calling and analysing Single Nucleotide Variants.*

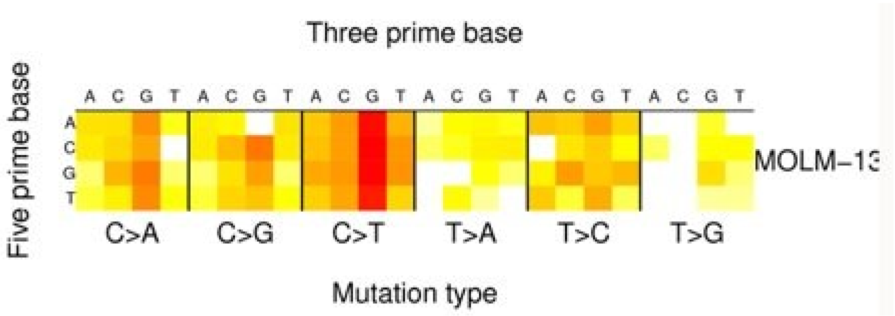


Figure S10 *Mutational profile of the Molm13 parental cell - screenshot from COSMIC's Cell Lines Project* (*Sample overview for 1330947*, 2020)

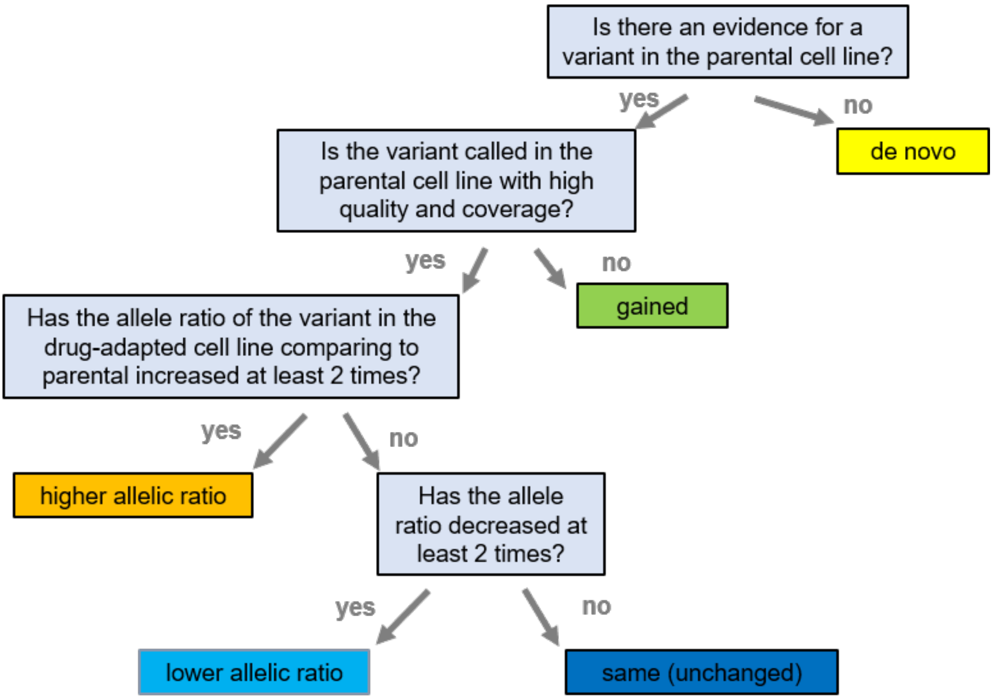

Is there an evidence for a variant in the parental cell line?
yes
no
de novo
Is the variant called in the parental cell line with high quality and coverage?
yes
no
gained
Has the allele ratio of the variant in the drug-adapted cell line comparing to parental increased at least 2 times?
yes
no
higher allelic ratio
Has the allele ratio decreased at least 2 times?
yes
no
lower allelic ratio
same (unchanged)

Figure S11 Selection of candidates for drivers. *Classification of the variants into de novo, gained, higher allelic ratio, same, lower allelic ratio, not called and lost.*

Parental \ Drug-adapted	High-quality and high-coverage	Present but either not called at all, or called with low-quality or low-coverage	Not present (no supporting reads)
High-quality and high-coverage	**same**	**not called**	**lost**
Present but either not called at all, or called with low-quality or low-coverage	**gained**		
Not present (no supporting reads)	**de novo**		

Table S4 Selection of candidates for drivers of resistance to Nutlin-3. *Classification of the variants into de novo, gained, same, not called and lost. Variants categorised as de novo, gained, not called and lost were then considered as candidates for drivers.*

Sample	Molm13I	Molm13II	Molm13III	Molm13IV
de_novo	72	55	47	53
gained	57	53	58	54
higher_allelic_ratio	1	3	6	2
same	288	296	258	288
lower_allelic_ratio	4	5	2	3
not_called	53	39	73	56
lost	68	71	75	65

Table S5 The number of variants per drug-adapted sub-line and class. *High-quality, high-coverage, somatic-like and most-damaging variants in each drug-adapted and parental cell line were classified as de novo, gained, higher allelic ratio, same, lower allelic ratio, not called and lost and then counted in each of the drug-adapted sub-line – Molm13I, Molm13II, Molm13III and Molm13IV.*

	P72R **17g.7579472 G>C**	**R175H** **17g.7578406 C>T**	**S127F** **17g.7578550 G>A**	**W91*** **17g.7579414 C>T**	**R213*** **17g.7578212 G>A**
Frequency in the population (gnomAD frequency)	6.68165e-01	3.97969e-06	-	-	0
Consequence for the protein	Change of an amino acid: from non-polar, hydrophobic to positively charged, polar, hydrophilic	Change of an amino acid: both are positively charged, polar, hydrophilic	Change of an amino acid: from no charge, polar, hydrophilic to non-polar, hydrophobic	Truncated protein	Truncated protein

SIFT/PolyPhen	tolerated/benign	tolerated/benign	deleterious/probably_damaging	-	-
Known variant?	validated polymorphism (IARC) present in dbSNP, COSMIC, gnomAD, ClinVar	validated polymorphism (IARC) present in dbSNP, COSMIC, gnomAD, ClinVar	present in dbSNP (730881999), COSMIC (44226), ClinVar (182928)	present in dbSNP (876660548), COSMIC (44492), ClinVar (233650)	present in dbSNP (397516436), COSMIC (10654), ClinVar (43590)
Associated with diseases	e.g. Li-Fraumeni syndrome, Hereditary cancer-predisposing syndrome, CODON 72 POLYMORPHISM, cisplatin response (ClinVar)	e.g. Li-Fraumeni syndrome, Hereditary cancer-predisposing syndrome, CODON 72 POLYMORPHISM, cisplatin response (ClinVar)	Li-Fraumeni syndrome (ClinVar)	Li-Fraumeni syndrome, Hereditary cancer-predisposing syndrome (ClinVar)	e.g. Li-Fraumeni syndrome, Hereditary cancer-predisposing syndrome (ClinVar)
Clinical significance	uncertain_significance&benign&drug_response; drug response (ClinVar)	uncertain_significance&benign&drug_response; drug response (ClinVar)	uncertain_significance&likely_pathogenic; Conflicting interpretations of	pathogenic; Pathogenic (ClinVar)	likely_pathogenic&pathogenic; Pathogenic (ClinVar)

			pathogenicity (ClinVar)		

Table S6 Additional information about the variants acquired in TP53 in the drug-adapted sub-lines. *The information about the gnomAD* (Karczewski *et al.*, 2020) *frequency and the assessment of SIFT* (Kumar, Henikoff and Ng, 2009) *and PolyPhen-2* (Adzhubei *et al.*, 2010) *was taken from VEP* (McLaren *et al.*, 2016) *results. The presence of the variants in the databases of cancer mutations was extracted from IARC* (Bouaoun *et al.*, 2016) *database (version from July 2019). Clinical significance and association with diseases is based on the ClinVar* (Landrum *et al.*, 2018) *entries for the variants.*

Sample	File	Value	**P72R** **17g.7579472G>C**	**R175H** **17g.7578406C>T**	**S127F** **17g.7578550G>A**	**W91*** **17g.7579414C>T**	**R213*** **17g.7578212G>A**
Molm13P	**Variants**	**Phred score**	**209**	-	-	-	-
		Coverage	**21**	-	-	-	-
		VAF	**0.381**	-	-	-	-
	Alignment	**Coverage**	30	31	39	14	141
		VAF	0.4	0	0	0	0
Molm13I	**Variants**	**Phred score**	**222**	**201**	**160**	-	-
		Coverage	**32**	**35**	**31**	-	-
		VAF	**0.563**	**0.457**	**0.323**	-	-
	Alignment	**Coverage**	39	39	41	24	135
		VAF	0.564	0.436	0.293	0	0 (1 x C)
Molm13II	**Variants**	**Phred score**	-	-	-	**225**	-
		Coverage	-	-	-	**18**	-
		VAF	-	-	-	**1**	-
	Alignment	**Coverage**	27	38	42	20	128
		VAF	0	0	0	1	0
Molm13III	**Variants**	**Phred score**	-	-	-	-	**225**
		Coverage	-	-	-	-	**48**
		VAF	-	-	-	-	**1**
	Alignment	**Coverage**	26	30	17	22	70
		VAF	0	0	0	0	1
Molm13IV	**Variants**	**Phred score**	**192**	-	-	-	-

		Coverage	**19**	-	-	-	-
		VAF	**0.47**	-	-	-	-
	Alignment	Coverage	28	50	25	19	135
		VAF	0.464	0	0	0	0

Table S7 Variant allele frequency, base coverage and Phred score for each of the TP53 variants. *Coverage of the base and variant allele frequency was calculated for each of the high-quality, high-coverage, somatic-like, most-damaging variants in TP53. Two sets of data were taken into account – the file with called variants and the alignment file where there might be additional low-quality reads that were not considered when variant was called (hence base coverage might be higher in the alignment file). Additionally, base quality score (Phred score) from the file with called variants is shown.*

Sample	File	Value	p.Y553C 20g.35526313T>C	p.P121L 20g.35563579G>A
Molm13P	Variants	Phred score	-	-
		Coverage	-	-
		VAF	-	-
	Alignment	Coverage	225	124
		VAF	0	0
Molm13I	Variants	Phred score	-	-
		Coverage	-	-
		VAF	-	-
	Alignment	Coverage	268	133
		VAF	0	0
Molm13II	Variants	Phred score	**228**	-
		Coverage	**141**	-
		VAF	**0.993**	-
	Alignment	Coverage	213	67
		VAF	0.981	0
Molm13III	Variants	Phred score	**225**	-
		Coverage	**102**	-
		VAF	**1**	-
	Alignment	Coverage	150	61
		VAF	1	0
Molm13IV	Variants	Phred score	-	**222**

		Coverage	-	**66**
		VAF	-	**0.545**
	Alignment	Coverage	269	90
		VAF	0	0.578

Table S8 Variant allele frequency, base coverage and Phred score for each of the SAMHD1 variants. *Coverage of the base and variant allele frequency was calculated for each of the high-quality, high-coverage, somatic-like, most-damaging variants in SAMHD1. Two sets of data were taken into account – the file with called variants and the alignment file where there might be additional low-quality reads that were not considered when variant was called (hence base coverage might be higher in the alignment file). Additionally, base quality score (Phred score) from the file with called variants is shown.*

	p.Y553C **20g.35526313T>C**	**p.P121L** **20g.35563579G>A**
Frequency in the population (gnomAD frequency)	-	3.97674e-06
Consequence for the protein	Change of an amino acid: both polar, Tyrosine has an aromatic ring, Cysteine forms disulfide bridges	Change of an amino acid: both hydrophobic, Proline has an aromatic ring
SIFT/PolyPhen	deleterious/probably damaging	deleterious/probably damaging
Known variant?	-	present dbSNP (rs1188635417)
Associated with diseases	Not in ClinVar	Not in ClinVar
Clinical significance	Not in ClinVar	Not in ClinVar

Table S9 Additional information about the variants acquired in SAMHD1 in the drug-adapted sub-lines. *The information about the gnomAD* (Karczewski *et al.*, 2020) *frequency and the assessment of SIFT* (Kumar, Henikoff and Ng, 2009) *and PolyPhen-2* (Adzhubei *et al.*, 2010) *was taken from VEP* (McLaren *et al.*, 2016)

results. Clinical significance and association with diseases is based on the lack of ClinVar (Landrum *et al.*, 2018) *entries for the variants.*

	Nutlin-3	Cytarabine	Daunorubicin
Cell line	IC50 (µM)	(ng/mL)	(ng/mL)
MOLM13	0.68 ± 0.22	17.4 ± 5.5	2.73 ± 0.75
MOLM13^{r}Nutlin$^{20\mu M}$I	12.9 ± 2.5 (19.0)[1]	2.15 ± 0.42 (0.12)	14.9 ± 3.9 (5.4)
MOLM13^{r}Nutlin$^{20\mu M}$II	11.3 ± 0.7 (16.6)	2.09 ± 0.56 (0.12)	13.4 ± 6.4 (4.9)
MOLM13^{r}Nutlin$^{20\mu M}$III	12.3 ± 1.8 (18.1)	2.45 ± 0.62 (0.14)	11.9 ± 2.3 (4.4)
MOLM13^{r}Nutlin$^{20\mu M}$IV	10.6 ± 2.3 (15.6)	5.41 ± 1.08 (0.31)	6.16 ± 2.44 (2.3)

[1] relative sensitivity (IC50 resistant subline/ IC50 respective parental cell line)

Table S10 *Drug concentrations that reduce cell viability by 50% (IC50) after 120h incubation as indicated by MTT assay.*

	Nutlin-3	Cytarabine
Cell line	IC50 (µM)	(ng/mL)
MOLM13	0.68 ± 0.22	17.4 ± 5.5
MOLM13^{r}Nutlin$^{20\mu M}$V	11.7 ± 2.2 (17.2)[1]	2.23 ± 1.02 (0.13)
MOLM13^{r}Nutlin$^{20\mu M}$VI	11.9 ± 2.2 (17.5)	49.6 ± 25.2 (2.9)
	Nutlin-3	Cytarabine
Cell line	IC50 (µM)	(µg/mL)
MV4-11	2.33 ± 0.35	0.79 ± 0.12
MV4-11^{r}Nutlin$^{20\mu M}$I	15.2 ± 2.8 (6.5)	0.82 ± 0.12 (1.04)
MV4-11^{r}Nutlin$^{20\mu M}$II	22.6 ± 1.5 (9.7)	0.38 ± 0.03 (0.48)
MV4-11^{r}Nutlin$^{20\mu M}$III	15.5 ± 1.6 (6.7)	0.96 ± 0.09 (1.22)
MV4-11^{r}Nutlin$^{20\mu M}$IV	18.4 ± 2.1 (7.9)	0.52 ± 0.03 (0.66)
MV4-11^{r}Nutlin$^{20\mu M}$V	16.6 ± 1.5 (7.1)	0.23 ± 0.08 (0.29)
MV4-11^{r}Nutlin$^{20\mu M}$VI	16.1 ± 0.3 (6.9)	0.26 ± 0.04 (0.33)
MV4-11^{r}Nutlin$^{20\mu M}$VII	20.3 ± 2.2 (8.7)	0.34 ± 0.06 (0.43)
MV4-11^{r}Nutlin$^{20\mu M}$VIII	17.0 ± 2.4 (7.3)	0.32 ± 0.05 (0.41)

MV4-11^{r}Nutlin$^{20\mu M}$IX	14.1 ± 0.7 (6.1)	0.13 ± 0.02 (0.16)
MV4-11^{r}Nutlin$^{20\mu M}$X	14.2 ± 1.7 (6.1)	0.15 ± 0.01 (0.19)
MV4-11^{r}Nutlin$^{20\mu M}$XI	17.4 ± 2.0 (7.5)	0.15 ± 0.02 (0.19)
MV4-11^{r}Nutlin$^{20\mu M}$XII	13.3 ± 1.2 (5.7)	0.53 ± 0.06 (0.67)
MV4-11^{r}Nutlin$^{20\mu M}$XIII	15.4 ± 0.9 (6.6)	1.30 ± 0.25 (1.65)
MV4-11^{r}Nutlin$^{20\mu M}$XIV	13.9 ± 1.8 (6.0)	1.00 ± 0.18 (1.27)
MV4-11^{r}Nutlin$^{20\mu M}$XV	17.0 ± 2.3 (7.3)	0.44 ± 0.07 (0.56)
	Nutlin-3	Cytarabine
Cell line	IC50 (µM)	(ng/mL)
SIG-M5	1.27 ± 0.16	150 ± 37
SIG-M5^{r}Nutlin$^{20\mu M}$III	11.2 ± 1.9 (8.8)	36.9 ± 13.7 (0.25)
SIG-M5^{r}Nutlin$^{20\mu M}$IV	23.0 ± 3.8 (18.1)	325 ± 32 (2.17)
SIG-M5^{r}Nutlin$^{20\mu M}$VI	15.2 ± 3.2 (12.0)	195 ± 7 (1.30)
SIG-M5^{r}Nutlin$^{20\mu M}$VIII	11.4 ± 3.5 (9.0)	63.0 ± 8.5 (0.42)
SIG-M5^{r}Nutlin$^{20\mu M}$IX	10.1 ± 0.3 (8.0)	42.5 ± 9.3 (0.28)
SIG-M5^{r}Nutlin$^{20\mu M}$XI	23.5 ± 0.7 (18.5)	43.1 ± 6.6 (0.29)
SIG-M5^{r}Nutlin$^{20\mu M}$XV	3.64 ± 0.29 (2.9)	73.7 ± 3.4 (0.49)
SIG-M5^{r}Nutlin$^{20\mu M}$XX	15.3 ± 3.8 (12.0)	80.7 ± 11.1 (0.54)

[1] relative sensitivity (IC50 resistant subline/ IC50 respective parental cell line)

Table S11 *Drug concentrations that reduce cell viability by 50% (IC50) after 120h incubation as indicated by MTT assay.*

Appendix 5

Selection of different clones upon repeated adaptation of neuroblastoma cell lines to tubulin-binding agents

Magdalena Antczak[1], Lyto Yiangou[1], Florian Rothweiler[2], Jochen Meyer[3,4], Andreas von Deimling[3,4], Frank Westermann[4], Daniel Speidel[5,6], Mark N. Wass[1#], Martin Michaelis[1#], Jindrich Cinatl jr.[1#]

[1] School of Biosciences, University of Kent, Canterbury, UK

[2] Institute for Medical Virology, Goethe-University, Frankfurt am Main, Germany

[3] Department of Neuropathology, Ruprecht-Karls-University, Heidelberg, Germany

[4] Deutsches Krebsforschungszentrum, 69120 Heidelberg, Germany

[5] Children's Medical Research Institute, Westmead, Australia

[6] Sydney Medical School, The University of Sydney, Sydney, Australia

Corresponding authors: Mark N. Wass (M.N.Wass@kent.ac.uk), Martin Michaelis (M.Michaelis@kent.ac.uk), Jindrich Cinatl jr. (Cinatl@em.uni-frankfurt.de)

Supplementary Material

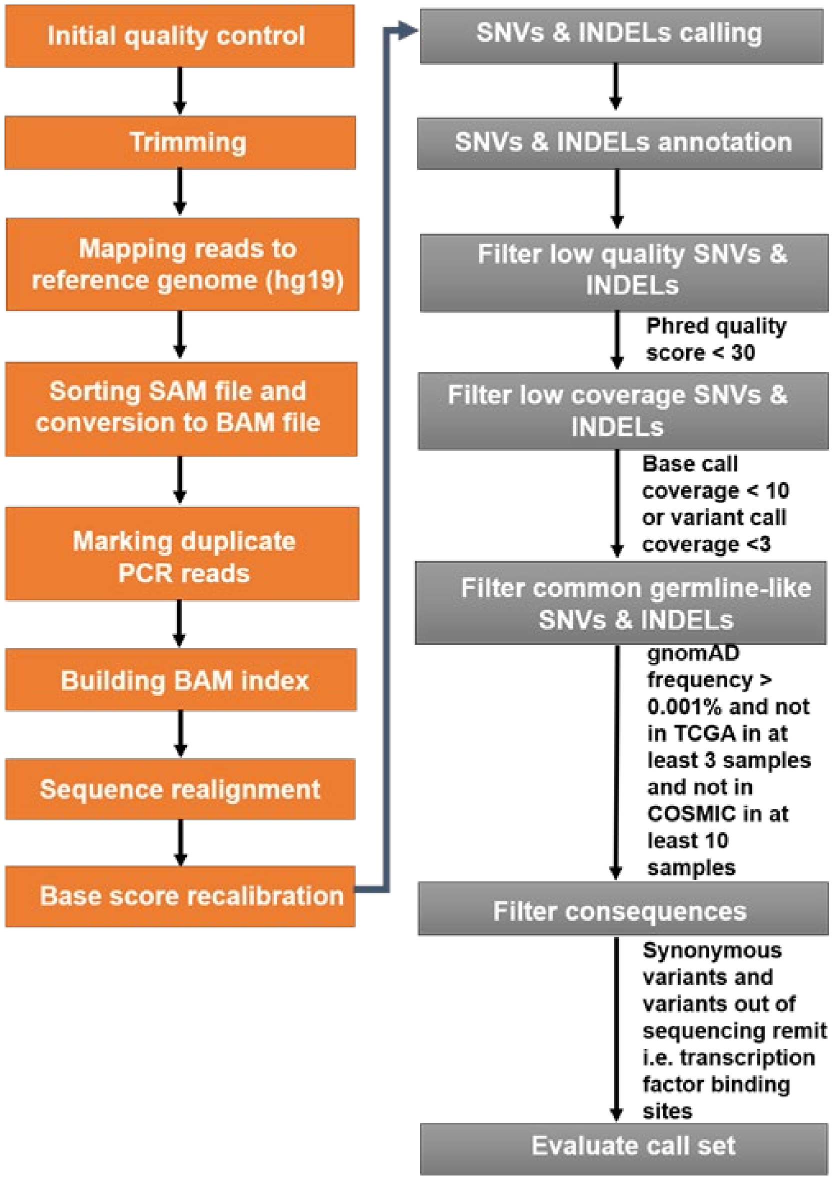

Initial quality control
Trimming
Mapping reads to reference genome (hg19)
Sorting SAM file and conversion to BAM file
Marking duplicate PCR reads
Building BAM index
Sequence realignment
Base score recalibration
SNVs & INDELs calling
SNVs & INDELs annotation
Filter low quality SNVs & INDELs
Phred quality score < 30
Filter low coverage SNVs & INDELs
Base call coverage < 10 or variant call coverage <3
Filter common germline-like SNVs & INDELs
gnomAD frequency > 0.001% and not in TCGA in at least 3 samples and not in COSMIC in at least 10 samples
Filter consequences
Synonymous variants and variants out of sequencing remit i.e. transcription factor binding sites
Evaluate call set

Figure S12 Scheme representing our in-house pipeline for calling and filtering of SNVs and INDELs.

	Vincristine
Cell line	IC50 (ng/mL)
UKF-NB-3	0.04 ± 0.02
UKF-NB-3^{r}VCR$^{0.5}$I	1.8 ± 0.4 (45.0)[1]
UKF-NB-3^{r}VCR$^{0.5}$II	0.58 ± 0.06 (14.5)
UKF-NB-3^{r}VCR$^{0.5}$III	0.60 ± 0.21 (15.0)
UKF-NB-3^{r}VCR$^{0.5}$IV	0.27 ± 0.11 (6.8)
UKF-NB-3^{r}VCR$^{0.5}$V	0.35 ± 0.11 (8.8)

UKF-NB-3^{r}VCR$^{0.5}$VI	0.84 ± 0.21 (21.0)
UKF-NB-3^{r}VCR$^{0.5}$VII	0.33 ± 0.04 (8.3)
UKF-NB-3^{r}VCR$^{0.5}$IX	1.47 ± 0.51 (36.8)
UKF-NB-3^{r}VCR$^{0.5}$XI	0.25 ± 0.01 (6.3)
UKF-NB-3^{r}VCR$^{0.5}$XII	0.35 ± 0.10 (8.8)
	Eribulin
Cell line	IC50 (ng/mL)
UKF-NB-3	0.13 ± 0.04
UKF-NB-3^{r}ERI^{10}I	247 ± 61 (1900)
UKF-NB-3^{r}ERI^{10}II	69 ± 18 (531)
UKF-NB-3^{r}ERI^{10}III	52 ± 14 (400)
UKF-NB-3^{r}ERI^{10}IV	210 ± 58 (1615)
UKF-NB-3^{r}ERI^{10}V	115 ± 42 (885)
UKF-NB-3^{r}ERI^{10}VI	159 ± 31 (1223)
UKF-NB-3^{r}ERI^{10}VII	117 ± 25 (900)
UKF-NB-3^{r}ERI^{10}VIII	112 ± 21 (862)
UKF-NB-3^{r}ERI^{10}IX	150 ± 37 (1154)
UKF-NB-3^{r}ERI^{10}X	124 ± 26 (954)
UKF-NB-3^{r}ERI^{10}XI	118 ± 22 (908)
UKF-NB-3^{r}ERI^{10}XII	28.9 ± 4.8 (222)
	2-Methoxyestradiol
Cell line	IC50 (nM)
UKF-NB-3	109 ± 16
UKF-NB-3^{r}2ME$^{2\mu M}$I	1271 ± 360 (11.7)
UKF-NB-3^{r}2ME$^{2\mu M}$II	1384 ± 26 (12.7)
UKF-NB-3^{r}2ME$^{2\mu M}$III	957 ± 495 (8.8)
UKF-NB-3^{r}2ME$^{2\mu M}$IV	1154 ± 176 (10.6)
UKF-NB-3^{r}2ME$^{2\mu M}$V	884 ± 329 (8.1)
UKF-NB-3^{r}2ME$^{2\mu M}$VI	1072 ± 294 (9.8)

UKF-NB-3^{r}2ME$^{2\mu M}$VII	1378 ± 155 (12.6)
UKF-NB-3^{r}2ME$^{2\mu M}$VIII	911 ± 183 (8.4)
UKF-NB-3^{r}2ME$^{2\mu M}$IX	1359 ± 61 (12.5)
UKF-NB-3^{r}2ME$^{2\mu M}$X	832 ± 51 (7.6)
	Epothilone B
Cell line	IC50 (nM)
UKF-NB-3	0.27 ± 0.08
UKF-NB-3^{r}EPOB^{2nM}I	1.82 ± 0.35 (6.7)
UKF-NB-3^{r}EPOB^{2nM}II	1.82 ± 0.17 (6.7)
UKF-NB-3^{r}EPOB^{2nM}III	1.16 ± 0.22 (4.3)
UKF-NB-3^{r}EPOB^{2nM}IV	1.72 ± 0.09 (6.4)
UKF-NB-3^{r}EPOB^{2nM}V	1.56 ± 0.31 (5.8)
UKF-NB-3^{r}EPOB^{2nM}VI	1.65 ± 0.24 (6.1)
UKF-NB-3^{r}EPOB^{2nM}VII	1.62 ± 0.29 (6.0)
UKF-NB-3^{r}EPOB^{2nM}IX	1.83 ± 0.36 (6.8)
UKF-NB-3^{r}EPOB^{2nM}X	1.38 ± 0.23 (5.1)

[1] relative resistance (IC50 resistant subline/ IC50 respective parental cell line)

Table S12 *Drug concentrations that reduce cell viability by 50% (IC50) after 120h incubation as indicated by MTT assay (values are presented as mean ± S.D.).*

Appendix 7

Participation in the third edition of the Critical Assessment of Functional Annotation: the Gene Ontology Annotation Tool (GOAT)

As part of my PhD research, I participated with my supervisor in the third round of the Critical Assessment of Functional Annotation (CAFA3). As a result, I am a co-author on the CAFA3 paper – (Zhou *et al.*, 2019) in Genome Biology. I developed the Gene Ontology Annotation Tool (GOAT), which was ranked in the top 10 performing methods for Molecular Function ontology (see Gene Ontology) according to F_{max} and S_{min} scores (see The Critical Assessment of Function Annotation) calculated for proteins where no prior experimental functions were available (refer to Fig S4 and Fig S5 in (Zhou *et al.*, 2019)). GOAT is a simplified version of CombFunc (Wass, Barton and Sternberg, 2012). It predicts Gene Ontology terms using the same SVM model as CombFunc (presented in detail in CombFunc); however, the features used to describe target proteins are based only on matches to Pfam domains (El-Gebali *et al.*, 2019), BLAST and PSI-BLAST hits (Altschul *et al.*, 1997) and ConFunc (Wass and Sternberg, 2008). GOAT predicts Gene Ontology terms solely from Molecular Function ontology. No other settings of CombFunc were changed.

GOAT is an automated protein function prediction method however it is currently not publicly available.