UNIVERSITA' DEGLI STUDI DI NAPOLI
FEDERICO II

DOTTORATO DI RICERCA IN
BIOLOGIA
XXXI CICLO

IDENTIFICATION AND EXPLORATION OF NOVEL MOLECULAR SIGNATURES IN BIOLOGICAL SYSTEMS THROUGH GENOMICS AND BIOINFORMATICS

Tutor	Candidato
Prof. Remo Sanges	Dott. Guglielmo Roma

ANNO ACCADEMICO 2017 – 2018

To my amazing wife

& my adorable son.

Tables of Contents

Introduction

The last two decades have witnessed rapid developments in –omics technologies which enable the study of biological and disease processes in a high throughput manner. Among the -omics approaches, genomics and the related bioinformatic methods have emerged as most popular applications able to accelerate science discoveries in basic research and drug discovery and therapeutics.

Genomics is an interdisciplinary field of science focusing on the structure, function, evolution, mapping, and editing of genomes (Wikipedia, url: https://en.wikipedia.org/wiki/Genomics). Over the years, the field of genomics has undergone several revolutions. Prior to the advent of Next Generation Sequencing (NGS), genomics was limited to the characterization of single disease-associated genes (e.g. Huntington disease, cystic fibrosis, cancer) or to the study of small genomes (e.g. bacteria, viruses). As physical mapping with large-insert clones became possible, the subcloned fragments of large genomes could be sequenced as individual projects, and their finished sequences combined together to reconstruct the sequence of entire chromosomes. Using this approach and beginning from 1985, in 2003 the Human Genome Project was able to complete the sequence of the DNA in the human genome (I. H. G. S. Consortium et al., 2001; Venter et al., 2001), thus providing a basic platform for the development of new technologies. In the same period, other large genomes, including those of model organisms, were also decoded (M. G. S. Consortium et al., 2002; R. G. S. P. Consortium et al., 2004; Myers et al., 2000). Hybridization-based methods such as microarrays exploited the information gained from genome projects to develop rapid, high throughput assays to allow the measurement of genetic variation, gene expression and chromatin binding, which spread rapidly in all fields of research. Most recently, these methods were quickly replaced by NGS, which allows similar studies to be conducted with much higher sensitivity and in an unbiased whole-genome and –transcriptome fashion. As a result, sequencing has become an essential and obligatory tool and not only for biologists.

In the early days of NGS, the initial focus of every genomic scientist was on the *de-novo* assembly of novel genomes for species that were never sequenced before. These efforts led to the completion of many novel genomic sequences which include even large genomes of mammals and plants. In the case of de-novo assembly, the genomic sequence is built from scratch without the use of an existing scaffold. Advances in sequencing technology have recently led to a dramatic increase in speed and throughput capacity, and a sharp reduction in costs. These improvements enabled the shift from *de-novo* to re-sequencing of entire genomes from additional individuals of species already sequenced. In the case of re-sequencing, short reads can be aligned to reference genomes as a substrate for variation discovery or gene expression analysis. Re-sequencing applications provide the scientific community with an unprecedented opportunity to address fundamental evolutionary questions, as well as to extend the use of sequencing to population genetic studies to infer ancient population history. The availability of new data types given by an always increasing number of NGS applications continues to engage and excite the computational biology community working on software development and on the analysis of new data types generated to solve complex biomedical problems.

In this context, the main objective of my research was to explore different biological systems to identify new molecular signals through the development and implementation of genomic and bioinformatic methods. This objective was accomplished by participating to three different research projects where I applied genomic and bioinformatic solutions to different areas of biology: genome composition, organization and regulation, malaria biology, and cancer. The first chapter provides an introduction to the main technology and biology concepts explored in my research, while the following three chapters describe in details the research work conducted during my studies.

Chapter I

Next Generation Sequencing applications for research: a genomics (r)evolution.

The use of sequencers as molecule-counting devices is immensely popular. DNA sequencers are capable of sequencing large numbers of different DNA fragments in parallel in a single reaction. In general, NGS experiments consist of 4 phases: sample collection, template generation, sequencing reactions and detection, and data analysis. All the sequencing methods monitor the sequential addition of nucleotides to immobilized and spatially arrayed DNA templates in more or less similar ways, but mostly differ in how the templates are generated and interrogated to obtain the sequences (Linnarsson, 2010). Figure 1 shows a basic workflow for NGS analysis (Rizzo & Buck, 2012).

The range and the scope of DNA sequencing applications is very broad and largely depends on the biological questions to address in the study (Figure 1). Normal and diseased tissues can be used as source for the extraction of nucleic acids in a whole genome or targeted experiment. Among the most popular applications, whole genome sequencing starts from genomic DNA as input and can be applied to human genetics and evolution studies for the detection of genome-wide genetic variations like Single Nucleotide Polymorphims (SNPs), Insertion and Deletions (InDels), large genomic rearrangements (*e.g.* large deletions, duplications, insertions, inversions or translocations), and even DNA repeats like Short Tandem Repeats (STRs). Whole genome sequencing can also provide information on cancer and disease-associated mutations, which makes it a key application in the field of precision medicine. Alternatively to the whole genome, whole-exome sequencing can be used for sequencing all of the known exons of protein-coding genes in a genome (known as the "exome"). This method also starts from genomic DNA and consists of a first step to select only the subset of DNA that encodes proteins (the "exons") and a second step to sequence the selected exonic DNA using any high-throughput sequencing technology.

Transcriptome sequencing starts from either total RNA or enriched RNA fractions. This application is based on shotgun sequencing of either full-length or 3′ ends of cDNA, and is used to reveal the presence, quantity and structure of RNA in a biological sample under specific conditions (Wang, Gerstein, & Snyder, 2009). Compared to hybridization-based RNA quantification methods such as microarrays, sequencing-based transcriptome detection can quantify gene expression with low background, high accuracy and high levels of reproducibility within a large dynamic range. In addition, transcriptome sequencing does not require an existing genome sequence and can detect mutations, splice variants and fusion genes that cannot be detected by microarrays. Among the library preparation methods available, the standard poly(A)+ enrichment provides a comprehensive, detailed, and accurate view of polyadenylated RNAs. However, on samples of suboptimal quality ribosomal RNA depletion and exon capture methods have recently been reported as better alternatives (S. Schuierer et al., 2017).

ChIP-sequencing is widely used to analyze protein-DNA interactions for epigenetic studies. It combines chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing to identify binding sites of DNA-associated proteins, and can be used to precisely map global binding sites for any protein of interest. ChIP sequencing offers higher resolution and more precise and abundant information in comparison with array-based ChIP-on-chip. Likewise ChIP-seq for the genomic DNA, RIP-Sequencing identifies the binding sites of proteins to the RNA within RNA-protein complexes extracted through immunoprecipitation with antibodies targeting the protein of interest. After RNase digestion, the RNA fragments protected by protein binding are extracted, reverse-transcribed to cDNA and sequenced.

The above-listed NGS applications, and in general all the methods shown in Figure 1, generate nucleic acids (*e.g.* genomic DNA, immunoprecipitated DNA, total RNA, enriched RNA fractions) that in the case of RNA need to be converted into double-stranded DNA (dsDNA) to proceed with the library preparation. These dsDNA fragments are subsequently converted into a "library" of sequencing templates, through standard steps of fragmentation, size selection, and adapter ligation. Fragmentation allows to break the DNA templates into smaller "sequenceable" fragments, which are then "size-selected" to enrich for fragments of a size range that is compatible with the sequencing platform's specifications. The ligation of platform-specific primers (or adapters) to the ends of the library fragments is used to enable priming for downstream amplification (*e.g.* clustering generation) and the sequencing reactions. Depending on the NGS technology used, a library is either sequenced directly or is amplified then sequenced (*e.g.* cluster generation via bridge amplification). Template generation also serves to spatially separate and immobilize

DNA fragment populations for sequencing, typically by attachment to solid surfaces (*e.g.* a flow cell) or beads. This allows the downstream sequencing reaction to operate as millions of microreactions carried out in parallel on each spatially distinct template.

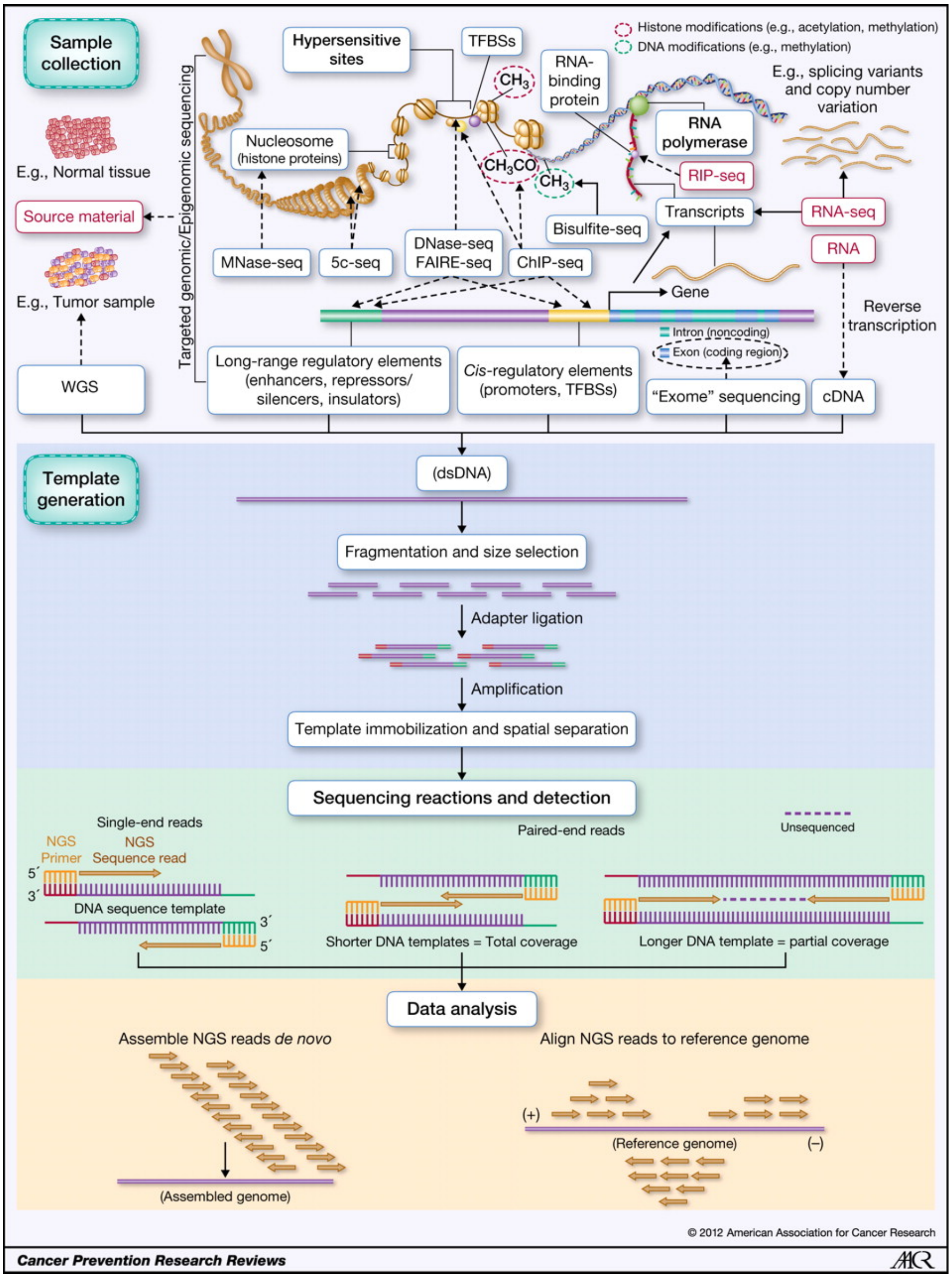

Sample collection
E.g., Normal tissue
Source material
E.g., Tumor sample
WGS
Targeted genomic/Epigenomic sequencing
Hypersensitive sites
TFBSs
Nucleosome (histone proteins)
RNA-binding protein
RNA polymerase
CH_3
CH_3CO
CH_3
Histone modifications (e.g., acetylation, methylation)
DNA modifications (e.g., methylation)
E.g., splicing variants and copy number variation
RIP-seq
RNA-seq
RNA
Transcripts
Bisulfite-seq
MNase-seq
5c-seq
DNase-seq FAIRE-seq
ChIP-seq
Gene
Intron (noncoding)
Exon (coding region)
Reverse transcription
Long-range regulatory elements (enhancers, repressors/ silencers, insulators)
Cis-regulatory elements (promoters, TFBSs)
"Exome" sequencing
cDNA
Template generation
(dsDNA)
Fragmentation and size selection
Adapter ligation
Amplification
Template immobilization and spatial separation
Sequencing reactions and detection
Single-end reads
NGS Primer
NGS Sequence read
5´
3´
DNA sequence template
3´
5´
Paired-end reads
Unsequenced
Shorter DNA templates = Total coverage
Longer DNA template = partial coverage
Data analysis
Assemble NGS reads *de novo*
(Assembled genome)
Align NGS reads to reference genome
(+)
(Reference genome)
(–)
© 2012 American Association for Cancer Research
Cancer Prevention Research Reviews
AACR

Figure 1. Basic workflow for NGS experiments.

Source: Rizzo and Buck, *Cancer Prevention Research*, 2012.

Bioinformatics: turning data into knowledge.

Driven by the rapid introduction of high-throughput sequencing in biological research, data generation has shifted to become faster and cheaper while data have been growing in complexity, diversity, and volume. This poses an important challenge to those research institutes that are not adequately prepared for the storage and for the high-performance computing analysis of "big data" and are required to access the external cloud computing to scale up to extra informatics capacity. Moreover, investing in the training of the next generation of scientists on "data science" disciplines is becoming fundamental for all research centers nowadays.

Bioinformatics is an interdisciplinary field that develops methods, databases, and software for the management, analysis and interpretation of biological data. As important as techniques to produce the NGS data are, bioinformatic approaches are equally critical for the successful analysis of those data. Many analytical approaches depend on the digital nature of NGS data, which depends on how the individual DNA fragments of the library are prepared prior to the sequencing reaction (e.g. targeted or whole genome). These fragments can be sequenced either as single-read or paired-end reads (e.g. originating from both ends of the molecule) to generate the raw data used for downstream analysis (Figure 1). The description of the standard RNA-seq workflow can serve as an example of how bioinformatics is applied to NGS data analysis.

The main goal of a standard RNA-seq experiment is to identify differentially expressed genes between two or more groups of biological samples. To this purpose, an end-to-end gene-level RNA-Seq differential expression workflow include four major steps such as quality control, alignment, quantification and identification of differentially expressed genes (Figure 2). First, raw NGS reads undergo quality assessment and filtering. Second, the quality-filtered reads in Fastq format are aligned against reference sequences (*e.g.* genome or transcriptome). The choice of the right reference genome and annotation is key for the success of the downstream analysis. Third, expression estimates are derived from the aligned reads to obtain the gene expression counts. For baseline expression, gene counts, which represent the total number of reads aligned to each gene, can be further transformed into Counts Per Million (CPM; normalization by total number of mapped reads per sample) or Fragments per Kilobase of exon per Million of fragments mapped (FPKM; further normalization by effective gene length). Fourth, for differential analysis, these count estimates are used to identify differentially expressed genes, usually by computing fold changes and *P*-values (Figure 2). The bioinformatic community has been very proficient in the development of software tools for the analysis of RNA-seq gene expression data. Tophat2 (Kim et al., 2013) and STAR (Dobin et al., 2013) are among the most popular aligners; EQP (Sven Schuierer & Roma, 2016), htseq-count (Anders, Pyl, & Huber, 2015), featureCounts (Liao, Smyth, & Shi, 2014) and Cufflinks2 (Trapnell et al., 2012) are used for the quantification step for the generation of gene counts. DESeq2 (Love, Huber, & Anders, 2014) and Cuffdiff (Trapnell et al., 2012) are used for the differential gene expression analysis.

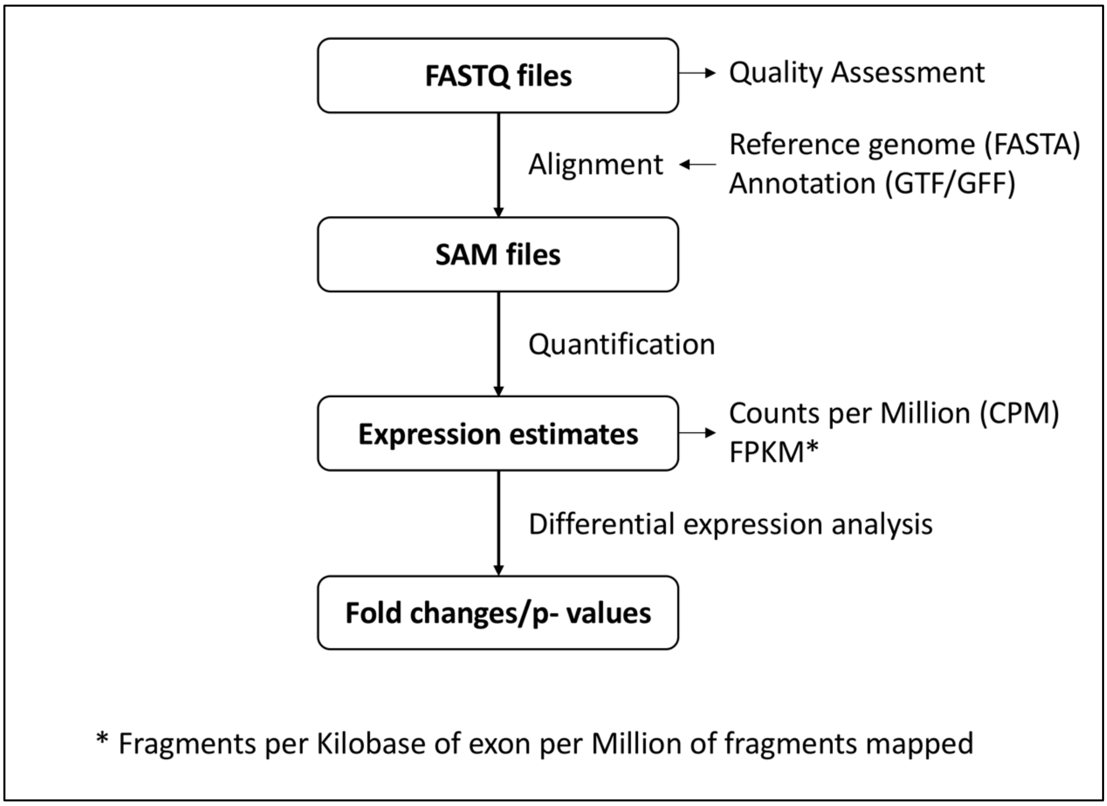


Figure 2. The major steps of a standard RNA-seq workflow.

Use of public resources for integrative data analysis.

Over the last decades, the scientific community has generated an immense amount of genomic data that is now deposited in large public repositories and available to other scientists wanting to conduct further analyses. For instance, reference genome sequences and annotation files can be accessed in genomic repositories like the "Ensembl Genome database" developed by the European Bioinformatics Institute (EBI) and the Wellcome Trust Sanger Institute in UK (Zerbino et al., 2018), the "UCSC Genome Browser" hosted at the University of Santa Cruz in California (Casper et al., 2018), and the "Genome Portal" of the Joint Genome Institute (JGI) (Nordberg et al., 2014), or even in species-specific repositories as for instance "PlasmoDB" which is a genome database for the genus Plasmodia useful to study the biology of the malaria parasites (Aurrecoechea et al., 2009).

Likewise, there has been a multitude of NGS experimental datasets deposited in the public domain which provides unprecedented opportunities for computational scientists to explore biology by data integrative approaches. For instance, the NCBI Sequence Read Archive (SRA, https://www.ncbi.nlm.nih.gov/sra) (Leinonen, Sugawara, & Shumway, 2011) and the European Nucleotide Archive (ENA, https://www.ebi.ac.uk/ena) (Leinonen, Akhtar, et al., 2011) are major public repositories hosting sequencing data generated by individual scientists or large consortia. In addition, for human relevant datasets, the database of Genotypes and Phenotypes (dbGAP, https://www.ncbi.nlm.nih.gov/gap) hosts and distributes the data and results from studies that have investigated the interaction of genotype and phenotype in humans (Tryka et al., 2014); similarly, the recent European Genome-phenome Archive (EGA, https://ega-archive.org) enables the archiving and sharing of all types of personally identifiable genetic and phenotypic data resulting from biomedical research projects (Lappalainen et al., 2015). Large scale sequencing projects that are led by public or private consortia routinely share their datasets in one or more of these repositories. Examples related to cancer studies are provided by The Cancer Genome Atlas Research Network Atlas (TCGA, https://cancergenome.nih.gov) (Gao et al., 2013; Hutter & Zenklusen, 2018), the Cancer Cell Line Encyclopedia (CCLE, https://portals.broadinstitute.org/ccle) (Barretina et al., 2012), and the Catalogue Of Somatic Mutations In Cancer (COSMIC, https://cancer.sanger.ac.uk/cosmic) (Forbes et al., 2017). The sequencing results generated by these large cancer projects, for instance, could be compared with the sequencing information from 53 non-diseased tissue sites across nearly 1000 individuals obtained by the Genotype-Tissue Expression (GTEx) project (The GTEx Consortium, 2013).

Data sharing is more and more considered good practice in computational biology to ensure research reproducibility. To enhance data reusability, the scientific community has been working on making the data FAIR, *e.g*. Findable, Accessible, Interoperable, and Reusable according to the FAIR principles (Wilkinson et al., 2016). For the same reason, scientific journals often require deposition of raw sequencing data and downstream results in public repository. This means that other scientists may access these data and materials in the future to conduct further research on these subjects.

Integrative data analysis is an emerging field of study that investigates strategies, algorithms and implementations of combining data from different sources and applies systems biology approaches in solving complex biomedical problems. Genomics, transcriptomics, proteomics and metabolomics are being combined in a systems biology approach to understand the biological system as a whole rather than focusing on individual factors. Being at the core of this new discipline, genomics approaches are now converging rapidly through the use of next-generation sequencing which enables, via a single technology, the acquisition of large datasets on genetic markers, epigenetic markers, transcriptome profiles, translational profiling, as well as relationships amongst these. The integration of such genomic datasets with proteomic and metabolomic data require the development of novel approaches for meta-analysis. Multi-omics studies are highly promising but also challenging as profound coordinated efforts in bioinformatics and biostatistics are required to connect the individual factors. One major challenge is represented by the heterogeneity of data formats that are generated by the different omic- technologies. Integration of more than two different omics data formats is still not routine and requires optimized software tools together with well-trained computational scientists to generate comprehensible workflows for the analysis of big data.

Diatoms: a biological model for regulatory genomics.

Marine diatoms are unicellular photosynthetic algae and a key phytoplankton group in the ocean (Armbrust, 2009). They play a fundamental role in global carbon cycles as they are estimated to be responsible for about 20% of global primary productivity (Armbrust, 2009; Smith et al., 2015). Diatoms have played a decisive role in the ecosystem for millions of years for the enormous amount of oxygen they generate on earth and for being the most important sources of biomass in oceans. These single-celled organisms are being studied in several commercial and industrial applications for the production of carbon-neutral fuels, pharmaceuticals, foods, biomolecules, nanomaterials, and for the bioremediation of contaminated water (Bozarth, Maier, & Zauner, 2009). Diatom cells are surrounded by a silica wall known as a "frustule" made up of two valves called "thecae", that typically overlap one another. Based on the shape of their frustule diatoms are classified into: 1) *Centrales*, centric diatoms that are radially symmetrical; 2) *Pennales,* pennate diatoms that are bilaterally symmetrical (Armbrust, 2009). The frustules of death diatoms sink to the bottom of the oceans and decomposes into diatomite, a remnant material that is used commercially as filters, mineral fillers, insulation material, insecticide, anti-caking agents, or fine abrasive. These simple eukaryotic organisms are of interest to many biologists for their extraordinary capacity to rapidly adapt to new environments. They represent an unique evolutionary model for investigating the role of genomic sequences in evolution (Russo, Annunziata, Sanges, Ferrante, & Falciatore, 2015).

The study described in Chapter II exemplifies the application of genomics and bioinformatics to the study of short tandem repeats (STRs) in diatoms. Specifically, it provides an alternative to the classical view of evolution in which changes occur via the accumulation of single point mutations by extending it to the inclusion of additional mechanisms that allow for the rapid gain, loss, and rearrangement of significant portions of the genome for instance through dynamic expansion or contraction of short repetitive sequences. These dynamic modifications of the genomes are truly fascinating as they enable simple organisms like protists to evolve rapidly in response to environmental changes, accounting for their wide dissemination in the biosphere. The study identified and characterized STR sequences in all the diatom genomes sequenced so far, including the Pennate diatoms *Phaeodactylum tricornutum* (Bowler et al., 2008), *Pseudo-nitzschia multistriata* (Basu et al., 2017) and *Fragilariopsis cylindrus* (Mock et al., 2017), and the centric diatom *Thalassiosira pseudonana* (Armbrust et al., 2004) (Figure 3). Results show, for the first time, that these genomes are enriched in triplet repeats that are mostly located in gene regulatory regions like promoters.

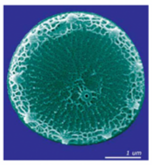
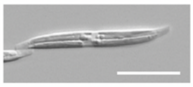
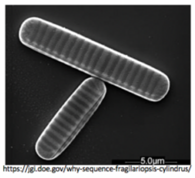
	Phaeodactylum Tricornutum	***Thalassiosira Pseudonana***	***Pseudo-nitzschia multistriata***	***Fragilariopsis cylindrus***
Genome	ASM15095v2, Feb 2010	ASM14940v2, May 2014	Psmu1.4, Jul 2015	CCMP1102 v1, Nov 2008
Length	27,568,093	32,437,365	59,304,822	80,540,407
Scaffolds	89	64	1,099	271
GC %	47.46	45.59	45.48	38.92
Genes	12,392	11,870	12,008	37,171
Source	Ensembl	Ensembl	Internal	JGI
	http://genome.jgi.doe.gov/Phatr2/Phatr2.home.html	 http://genome.jgi.doe.gov/Thaps3/Thaps3.home.html	 (D'Alelio et al, Protist, 2009)	 https://jgi.doe.gov/why-sequence-fragilariopsis-cylindrus/

Figure 3. Marine diatoms sequenced to date.

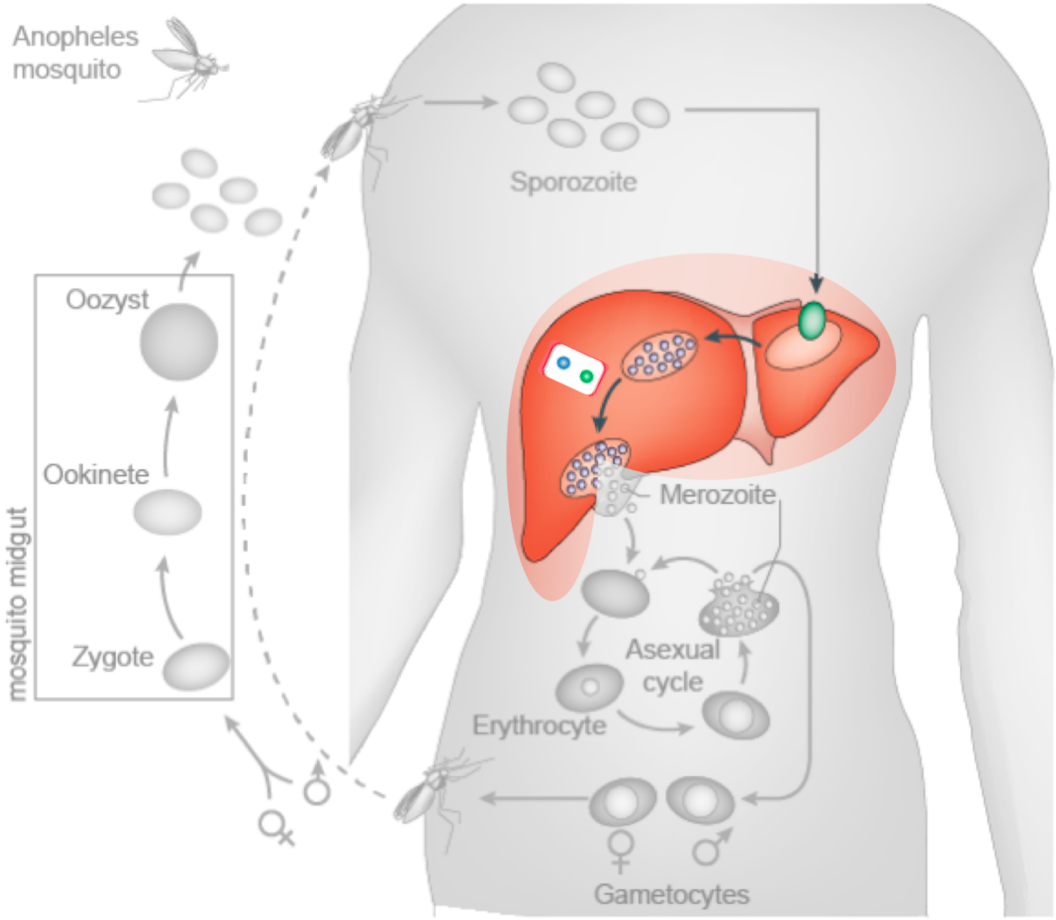
Understanding the biology of the sleeping malaria parasite.

Malaria is a life-threatening disease transmitted to humans through the bite of an infected *Anopheles* mosquito carrying the *Plasmodium* parasite. More than 100 species of *Plasmodium* have been identified so far, of which only four have long been recognized to infect humans: *P. falciparum*, *P. vivax*, *P. ovale* and *P. malariae*.

At the beginning of the parasite life cycle, female mosquitoes take blood meals to carry out egg production (Figure 4). Once injected in the human skin, the sporozoites rapidly leave the injection site and migrate through the bloodstream to the liver, where they invade hepatocytes and develop into the growing liver stage, the "schizonts". *P. vivax* and *P. Ovale*, however, can form dormant liver stages called "hypnozoites" that can re-activate months or years later giving rise to clinical malaria (relapses) without being exposed to new infectious mosquito bites. The schizonts parasites grow and multiply first in the liver cells and then in the red blood cells. When growing inside these cells the parasites destroy them, releasing daughter parasites called "merozoites" that continue the cycle by invading other cells. Malaria symptoms (like fever, headache, nausea, vomiting, abdominal pain, diarrhea, among others) are caused by the blood stage parasites. In the blood stream the parasites enter the asexual cycle. There the new forms termed "gametocytes" can be picked up by a female *Anopheles* mosquito during a new blood meal to start another cycle of growth and multiplication in the mosquito. Thus the mosquito acts as a vector carrying the disease from one human to another.

P. vivax is the major cause of malaria outside of Africa with an estimated 13.8 million malaria cases globally in 2015 (World Health Organization (WHO), 2015). Eradication of *vivax* malaria will only be feasible if effective and well-tolerated therapies kill hypnozoites and hence prevent disease relapse. Recently, the FDA approved tafenoquine as a radical cure therapy and prophylactic for *P. vivax* malaria infection (Frampton, 2018). This represents a significant advance in the field as tafenoquine is administered as a single dose regimen, which is a very important improvement for patient compliance when compared to the 14-day long drug regimen of its closely related predecessor primaquine. Like primaquine, tafenoquine cannot be administered to patients with glucose-6-phosphate dehydrogenase (G6PD) deficiency, a common genetic disorder in malaria endemic countries, due to serious adverse side-effects and life-threatening drug-induced hemolysis (Mazier, Rénia, & Snounou, 2009; Wells, Burrows, & Baird, 2010). For this reason, new drugs are urgently needed to achieve malaria elimination.

The second study presented in Chapter III describes the application of genomics and bioinformatics for the transcriptomic analysis of the malaria hypnozoites. Using a combination of genetically engineered fluorescent *P. cynomolgi* parasites (the *P. vivax* sister parasites displaying identical biology in the monkeys), *in vitro* liver stage culture, cell-sorting and RNA-seq, primary monkey hepatocytes were profiled six and seven days after infection with the *Plasmodium* parasites to investigate the hypnozoite biology. The analysis of the sequencing data revealed that hypnozoites have a reduced transcriptional rate and express a lower number of genes compared to schizonts, the hepatic forms of the developing parasite. While the schizonts express 91% of the *Plasmodium* cell pathways, the hypnozoites globally repress the gene expression to a minimum number of biological pathways that allow only the maintenance of the basic cellular functions necessary for its survival in the host hepatocyte. This data set and the analyses carried out represent a precious resource for the discovery of new vaccines and effective treatments to combat malaria.



adapted from: Nature Microbiology, 2010

Figure 4. Malaria parasite life cycle.

Hepatocellular carcinoma: the third leading cause of cancer deaths worldwide.

Representing more than 90% of all primary liver malignancies, hepatocellular carcinoma (HCC) is one of the few cancer types with rising incidence and mortality. This cancer occurs primarily in patients affected by chronic liver disease and cirrhosis. Although still under investigation, it is hypothesized that hepatic stem cells are the cells that give rise to this disease. Available treatments include only three approved systemic agents, namely sorafenib and regorafenib (both kinase inhibitors) and nivolumab (immune checkpoint inhibitor). Despite the extensive genomic and transcriptomic characterization of the features and diversity of HCC, there is still a urgent need for the identification of novel therapeutic targets in HCC and of robust biomarkers of response to therapy.

Cancer is a genetic disease caused by accumulation of DNA mutations and epigenetic alterations leading to uncontrolled cell proliferation and tumor formation. Genes involved in liver metabolism, Wnt and p53 signalling have been shown to be recurrently altered in HCC (Ahn et al., 2014; Ally et al., 2017; Fujimoto et al., 2012; Guichard et al., 2012; Hutter & Zenklusen, 2018; Schulze et al., 2015). CTNNB1 (β-catenin) and TP53 (p53) are the most frequently mutated protein-coding genes, both mutated in 20–40% of HCC patients (Ahn et al., 2014; Ally et al., 2017; Fujimoto et al., 2012; Guichard et al., 2012; Hutter & Zenklusen, 2018; Schulze et al., 2015). TP53 is also the most frequently mutated gene in human cancer (Kandoth et al., 2013)

(Figure 5). The p53 protein modulates multiple cellular functions, including transcription, DNA synthesis and repair, cell cycle arrest, senescence and apoptosis (Vogelstein, Lane, & Levine, 2000). Mutations in TP53 can abrogate these functions, leading to genetic instability and progression to cancer (Vogelstein et al., 2000).

The third study, presented in Chapter IV, exemplifies the use of genomics and bioinformatics to discover new molecular signals in the patients affected by hepatocellular carcinoma and bearing mutations in the TP53 gene. Taking advantage of the public RNA-seq data sets from The Cancer Genome Atlas (TCGA), the study defines the spectrum of the TP53 somatic mutations in HCC patients and its association with clinicopathologic features. Four distinct subsets of TP53 mutations, each characterized by specific molecular signals, were identified from 373 HCC cases. Patients with TP53 mutations had worse survival than patients with wild-type TP53. The study indicated that some genetic heterogeneity of the TP53 mutation exists in HCC cancer and that mutations in TP53 should be considered for the molecular characterization of HCC.

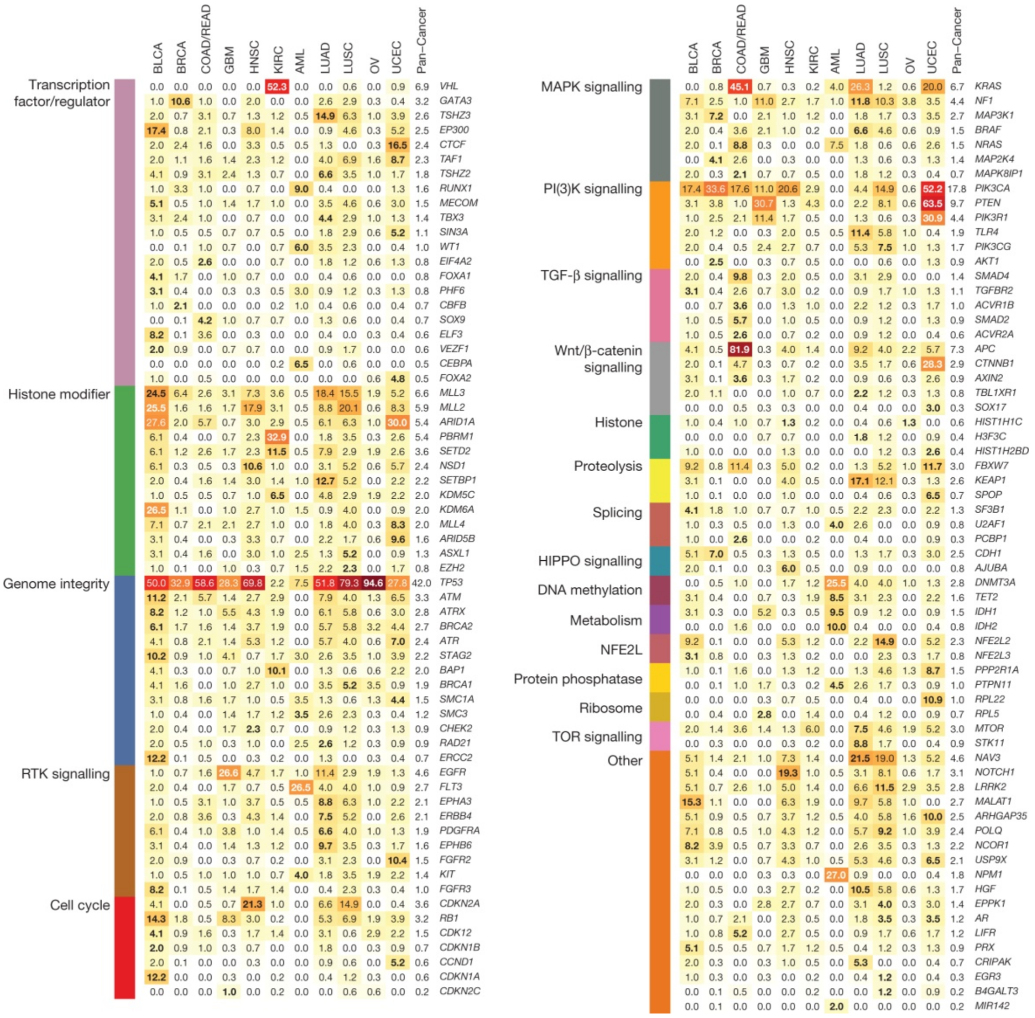


Figure 5. The 127 Significantly Mutated Genes (SMGs) from 20 cellular processes in cancer identified in 12 cancer types.

Source: Kandoth et al, *Nature*, 2013.

References

Ahn, S.-M., Jang, S. J., Shim, J. H., Kim, D., Hong, S.-M., Sung, C. O., … Kong, G. (2014). Genomic portrait of resectable hepatocellular carcinomas: Implications of RB1 and FGF19 aberrations for patient stratification. *Hepatology*, *60*(6), 1972–1982. https://doi.org/10.1002/hep.27198

Ally, A., Balasundaram, M., Carlsen, R., Chuah, E., Clarke, A., Dhalla, N., … Laird, P. W. (2017). Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. *Cell*, *169*(7), 1327–1341.e23. https://doi.org/https://doi.org/10.1016/j.cell.2017.05.046

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, *31*(2), 166–169. https://doi.org/10.1093/bioinformatics/btu638

Armbrust, E. V. (2009). The life of diatoms in the world ' s oceans, *459*(May), 185–192. https://doi.org/10.1038/nature08057

Armbrust, E. V., Berges, J. A., Bowler, C., Green, B. R., Martinez, D., Putnam, N. H., … Rokhsar, D. S. (2004). The Genome of the Diatom Thalassiosira Pseudonana: Ecology, Evolution, and Metabolism. *Science*, *306*(5693), 79–86. https://doi.org/10.1126/science.1101156

Aurrecoechea, C., Brestelli, J., Brunk, B. P., Dommer, J., Fischer, S., Gajria, B., … Wang, H. (2009). PlasmoDB : a functional genomic database for malaria parasites, *37*(October 2008), 539–543. https://doi.org/10.1093/nar/gkn814

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S.,

… Garraway, L. A. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, *483*, 603. Retrieved from http://dx.doi.org/10.1038/nature11003

Basu, S., Patil, S., Mapleson, D., Russo, M. T., Vitale, L., Fevola, C., … Ferrante, M. I. (2017). Finding a partner in the ocean: molecular and evolutionary bases of the response to sexual cues in a planktonic diatom. *New Phytologist*, *215*(1), 140–156. https://doi.org/10.1111/nph.14557

Bowler, C., Allen, A. E., Badger, J. H., Grimwood, J., Jabbari, K., Kuo, A., … Grigoriev, I. V. (2008). The Phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature*, *456*, 239. Retrieved from http://dx.doi.org/10.1038/nature07410

Bozarth, A., Maier, U., & Zauner, S. (2009). Diatoms in biotechnology : modern tools and applications, 195–201. https://doi.org/10.1007/s00253-008-1804-8

Casper, J., Zweig, A. S., Villarreal, C., Tyner, C., Speir, M. L., Rosenbloom, K. R., … Kent, W. J. (2018). The UCSC Genome Browser database: 2018 update. *Nucleic Acids Research*, *46*(D1), D762–D769. https://doi.org/10.1093/nar/gkx1020

Consortium, I. H. G. S., Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., … Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*, 860. Retrieved from http://dx.doi.org/10.1038/35057062

Consortium, M. G. S., Chinwalla, A. T., Cook, L. L., Delehaunty, K. D., Fewell, G. A., Fulton, L. A., … Zody, M. C. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, *420*, 520. Retrieved from http://dx.doi.org/10.1038/nature01262

Consortium, R. G. S. P., Gibbs, R. A., Weinstock, G. M., Metzker, M. L., Muzny, D. M., Sodergren, E. J., … Collins, F. (2004). Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, *428*, 493. Retrieved from http://dx.doi.org/10.1038/nature02426

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., … Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Forbes, S. A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., … Campbell, P. J. (2017). COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Research*, *45*(D1), D777–D783. https://doi.org/10.1093/nar/gkw1121

Frampton, J. E. (2018). Tafenoquine: First Global Approval. *Drugs*, *78*(14), 1517–1523. https://doi.org/10.1007/s40265-018-0979-2

Fujimoto, A., Totoki, Y., Abe, T., Boroevich, K. A., Hosoda, F., Nguyen, H. H., … Nakagawa, H. (2012). Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. *Nature Genetics*, *44*, 760. Retrieved from http://dx.doi.org/10.1038/ng.2291

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., … Schultz, N. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science Signaling*, *6*(269), pl1. https://doi.org/10.1126/scisignal.2004088

Guichard, C., Amaddeo, G., Imbeaud, S., Ladeiro, Y., Pelletier, L., Maad, I. Ben, … Zucman-Rossi, J. (2012). Integrated analysis of somatic mutations and focal copy-number changes identifies key genes and pathways in hepatocellular carcinoma. *Nature Genetics*, *44*, 694. Retrieved from http://dx.doi.org/10.1038/ng.2256

Hutter, C., & Zenklusen, J. C. (2018). The Cancer Genome Atlas: Creating Lasting Value beyond Its Data. *Cell*, *173*(2), 283–285. https://doi.org/https://doi.org/10.1016/j.cell.2018.03.042

Kandoth, C., McLellan, M. D., Vandin, F., Ye, K., Niu, B., Lu, C., … Ding, L. (2013). Mutational landscape and significance across 12 major cancer types. *Nature*, *502*, 333. Retrieved from http://dx.doi.org/10.1038/nature12634

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, *14*(4), R36. https://doi.org/10.1186/gb-2013-14-4-r36

Lappalainen, I., Almeida-King, J., Kumanduri, V., Senf, A., Spalding, J. D., ur-Rehman, S., … Flicek, P. (2015). The European Genome-phenome Archive of human data consented for biomedical research. *Nature Genetics*, *47*, 692. Retrieved from http://dx.doi.org/10.1038/ng.3312

Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y., … Cochrane, G. (2011). The European Nucleotide Archive. *Nucleic Acids Research*, *39*(suppl_1), D28–D31. https://doi.org/10.1093/nar/gkq967

Leinonen, R., Sugawara, H., & Shumway, M. (2011). The sequence read archive. *Nucleic Acids Research*, *39*(Database issue), D19-21. https://doi.org/10.1093/nar/gkq1019

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, *30*(7), 923–930. https://doi.org/10.1093/bioinformatics/btt656

Linnarsson, S. (2010). Recent advances in DNA sequencing methods – general principles of sample preparation. *Experimental Cell Research*, *316*(8), 1339–1343. https://doi.org/https://doi.org/10.1016/j.yexcr.2010.02.036

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Mazier, D., Rénia, L., & Snounou, G. (2009). A pre-emptive strike against malaria&#39;s stealthy hepatic forms. *Nature Reviews Drug Discovery*, *8*, 854. Retrieved from http://dx.doi.org/10.1038/nrd2960

Mock, T., Otillar, R. P., Strauss, J., McMullan, M., Paajanen, P., Schmutz, J., … Grigoriev, I. V. (2017). Evolutionary genomics of the cold-adapted diatom Fragilariopsis cylindrus. *Nature*, *541*(7638), 536–540.

Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J., … Venter, J. C. (2000). A Whole-Genome Assembly of Drosophila. *Science*, *287*(5461), 2196–2204. https://doi.org/10.1126/science.287.5461.2196

Nordberg, H., Cantor, M., Dusheyko, S., Hua, S., Poliakov, A., Shabalov, I., … Dubchak, I. (2014). The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. *Nucleic Acids Research*, *42*(D1), D26–D31.

Rizzo, J. M., & Buck, M. J. (2012). Key Principles and Clinical Applications of {\textquotedblleft}Next-Generation{\textquotedblright} DNA Sequencing. *Cancer Prevention Research*, *5*(7), 887–900. https://doi.org/10.1158/1940-6207.CAPR-11-0432

Russo, M. T., Annunziata, R., Sanges, R., Ferrante, M. I., & Falciatore, A. (2015). The upstream regulatory sequence of the light harvesting complex Lhcf2 gene of the marine diatom Phaeodactylum tricornutum enhances transcription in an orientation- and distance-independent fashion. *Marine Genomics*, *24*, 69–79. https://doi.org/https://doi.org/10.1016/j.margen.2015.06.010

Schuierer, S., Carbone, W., Knehr, J., Petitjean, V., Fernandez, A., Sultan, M., & Roma, G. (2017). A comprehensive assessment of RNA-seq protocols for degraded and low-quantity samples. *BMC Genomics*, *18*(1). https://doi.org/10.1186/s12864-017-3827-y

Schuierer, S., & Roma, G. (2016). The exon quantification pipeline (EQP): a comprehensive approach to the quantification of gene, exon and junction expression from RNA-seq data. *Nucleic Acids Research*, gkw538. https://doi.org/10.1093/nar/gkw538

Schulze, K., Imbeaud, S., Letouzé, E., Alexandrov, L. B., Calderaro, J., Rebouissou, S., … Zucman-Rossi, J. (2015). Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. *Nature Genetics*, *47*(5), 505–511. https://doi.org/10.1038/ng.3252

Smith, S. R., Glé, C., Abbriano, R. M., Traller, J. C., Davis, A., Trentacoste, E., … Hildebrand, M. (2015). Transcript level coordination of carbon pathways during silicon starvation-induced lipid accumulation in the diatom Thalassiosira pseudonana. *New Phytologist*, *210*(3), 890–904. https://doi.org/10.1111/nph.13843

The GTEx Consortium. (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, *45*(6), 580–585. https://doi.org/10.1038/ng.2653

Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., & Pachter, L. (2012). Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology*, *31*, 46. Retrieved from http://dx.doi.org/10.1038/nbt.2450

Tryka, K. A., Hao, L., Sturcke, A., Jin, Y., Wang, Z. Y., Ziyabari, L., … Feolo, M. (2014). NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Research*, *42*(D1), D975–D979. https://doi.org/10.1093/nar/gkt1211

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., … Zhu, X. (2001). The Sequence of the Human Genome. *Science*, *291*(5507), 1304–1351. https://doi.org/10.1126/science.1058040

Vogelstein, B., Lane, D., & Levine, A. J. (2000). Surfing the p53 network. *Nature*, *408*, 307. Retrieved from http://dx.doi.org/10.1038/35042675

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*, 57. Retrieved from http://dx.doi.org/10.1038/nrg2484

Wells, T. N. C., Burrows, J. N., & Baird, J. K. (2010). Targeting the hypnozoite reservoir of Plasmodium vivax: the hidden obstacle to malaria elimination. *Trends in Parasitology*, *26*(3), 145–151. https://doi.org/10.1016/j.pt.2009.12.005

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018. Retrieved from http://dx.doi.org/10.1038/sdata.2016.18

World Health Organization (WHO). (2015). World malaria report 2015. Geneva, (http://www.who.int/malaria/publications/world-malaria-report-2015/en/).

Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., … Flicek, P. (2018). Ensembl 2018. *Nucleic Acids Research*, *46*(D1), D754–D761. https://doi.org/10.1093/nar/gkx1098

Chapter II

Short tandem repeats are enriched in promoters of diatom genes

Abstract

To date, short tandem repeats (STRs) remain understudied in lower eukaryotes. Here, we present the first genome-wide survey of STRs in four marine diatom genomes, including the pennate diatoms *Phaeodactylum tricornutum*, *Pseudo-nitzschia multistriata* and *Fragilariopsis cylindrus* and the centric diatom *Thalassiosira pseudonana.* For the first time we discover that the most common STRs in diatom genomes are triplets, of which AAC is the most abundant. We found that over 75% of STRs are located in non-coding regions, particularly in promoters and in intronic regions. AAC is the most frequent repeat in the promoters of all diatom species, while AGT and ACT are copious only in *P. multistriata* promoters and TATA-box like DNA motifs (like AAT and ATT) only in *F. cylindrus* promoters. The presence of these repeats in diatom promoters might lead, in the cases of their expansion, to the gain of regulatory motifs upstream of the TSS or to their loss in cases of reduction. These sequences have therefore the capability to modulate gene expression. This dataset is a valuable resource to investigate transcriptional regulation in lower eukaryotes.

Introduction

Tandem repeats are an abundant class of genomic sequences that mutate faster than the surrounding genome (Lynch et al., 2008; Richard, Kerrest, & Dujon, 2008). At each cell division, these unstable genomic elements may change in the number of repeat units during DNA replication. These sequences, formerly-thought of as junk DNA, became of age as they are used as genomic markers and DNA fingerprints, are involved in human disease, and are linked to the evolution of coding and regulatory regions (Gemayel, Vinces, Legendre, & Verstrepen, 2010). Tandem repeats are composed of a short DNA motif, the so-called repeat unit, that is repeated several times head-to-tail. Based on the size of the repeat unit (or period size), tandem repeats are classified into microsatellites (unit size <10nt; also known as short tandem repeats, STRs) or minisatellites (unit size ≥10nt). Microsatellites are the most prevalent types of repeats (Ellegren, 2004; Gemayel et al., 2010).

Recent studies have shown that STRs are ubiquitous and abundant in higher eukaryotic genomes (Gemayel et al., 2010). They occupy 3% of the human genome and are mainly located in coding regions and in gene expression regulatory regions like promoters (Sawaya et al., 2013). STRs located in such functional regions can modulate phenotypes via expansion or contraction of their repeat units thus potentially leading to an evolutionary advantage (Iii, Hammock, Hannan, & King, 2008) or even the onset of human diseases (Campuzano et al., n.d.; Day JW, n.d.; Gijselinck et al., 2012; Maclean, Warne, & Zajac, 1996; Orr & Zoghbi, 2007; Richard et al., 2008; Sawaya et al., 2013; Tabolacci, Palumbo, Nobile, & Neri, 2016).

STRs have also been observed in promoters of other eukaryotic genomes including single-celled organisms like yeast (Vinces, Legendre, Caldara, Hagihara, & Verstrepen, 2011) and complex organisms like birds (Abe & Gemmell, 2016), dogs (Eo et al., 2016), monkeys (Ohadi M, Valipour E, Ghadimi-Haddadan S, Namdar-Aligoodarzi P & Kowsari A, Rezazadeh M, Darvish H, 2014) or humans (Sawaya et al., 2013). Vinces *et al.* found that STRs can affect gene expression by acting as nucleosome inhibitory sequences that maintain an open chromatin structure in specific regions of the promoter (Vinces et al., 2011). These authors propose a possible role of tandem repeats as nucleosome positioning elements in eukaryotes (Vinces et al., 2011). Similarly, Sawaya *et al.* described that STRs are abundant in human promoters, often highly conserved, and enriched near the transcriptional start sites (TSS) of key regulatory genes involved in growth and development. These authors suggest that promoter STRs have the potential to affect promoter function by generating mutations in regulatory elements, which may ultimately lead to variation in phenotypes (Sawaya et al., 2013). Finally, Sonay *et al.* showed that tandem repeats have taken an evolutionary role in gene expression differences in human and ape grape populations since genes with tandem repeats had higher expression divergence than genes without repeats, in the following order of decreasing divergence: repeats in 3'UTR, exons, promoters, 1st intron, and other introns (Sonay et al., 2015). While these findings highlight the potential contribution of STRs to evolution, they mostly refer to studies in higher eukaryotes. However, it remains still unknown whether STRs are also present and possibly control gene expression in simpler eukaryotic organisms like protists.

Diatoms are unicellular photosynthetic algae and a key phytoplankton group in the contemporary ocean (Vardi, Thamatrakoln, Bidle, & Falkowski, 2009). These species played a decisive role in the ecosystem for millions of years as one of the foremost set of oxygen synthesizers on earth and as one of the most important sources of biomass in oceans (Armbrust, 2009). Diatoms are used in commercial and industrial applications as the carbon neutral synthesis of fuels, pharmaceuticals, health foods, biomolecules, materials relevant to nanotechnology, and bioremediators of contaminated water (Bozarth, Maier, & Zauner, 2009). These single-celled organisms show extraordinary adaptation capacities to rapidly changing environments and therefore represent an important evolutionary model for investigating the role of tandem repeats in evolution (Russo, Annunziata, Sanges, Ferrante, & Falciatore, 2015). Here we provide the first genome-wide survey of STRs in diatoms. Using an in-house bioinformatic workflow, we identified STRs in the four diatom genomes sequenced to date.

Methods

Identification of short tandem repeats in diatom genomes

For the identification of STRs in whole genomes, we gathered genome FASTA files and annotation GTF files of the four diatoms listed in Table 1. The *P. tricornutum* and *T. pseudonana* reference files were retrieved from the Protists Ensembl database (http://protists.ensembl.org/) (Kersey et al., 2016); the *F. cylindrus* genome was obtained from the Joint Genome Institute (JGI) Genome Portal (http://genome.jgi.doe.gov) (Nordberg et al., 2014); finally, the *P. multistriata* reference genome was recently sequenced at our institute (Basu et al., 2017). Diatom genomes were interrogated for the presence of STRs using the Phobos software version 3.3.12 (Mayer, 2007) with default options except for *minUnitLen=2* to exclude homopolymer repeats and *outputFormat=3* to generate the one-per-line tabular format. The software selected STRs with a minimum length of 8 bp and a unit size of at least 2 (*e.g.* from 4 dinucleotides, 3 trinucleotides and 2 tetranucleotides). We wrote a Perl script to parse this output and to summarize STR features such as the chromosomal location, repeat size, copy number, alignment score, consensus and sequence. As a quality control, we visually inspected examples of tandem repeats from *P. tricornutum* and *T. pseudonana* using the Ensembl genome browser (Kersey et al., 2016). In addition, the presence of specific STRs (*e.g.* CCTAAC repeats known to be located at the telomere regions) was confirmed using Ensembl karyotype plots to assess the correctness of our analysis. GTF files from the four diatoms were used as the source to annotate the repeats. For each STR, we used intersectBed from the BEDTools suite (Quinlan & Hall, 2010) to verify if it overlapped a promoter region defined as 500 bp upstream of the transcriptional start site (promoter-TSS), a gene exon (exon), a gene intron (intron), or none of the above features (intergenic). The STR occurrences were then normalized by the size of the feature as annotated in the reference genomes. Finally, we used genomecov from the BEDTools suite to compute the coverage of the repeat features in the region surrounding the gene TSS. The distribution plots presented in this article were generated with scripts written in the R language (R Core Team, 2013).

Identification of short tandem repeats in diatom promoters

An in-depth analysis was carried out in promoter regions of each diatom genome. Here, we defined promoters as the sequences of the 500 bp upstream of each gene TSS, collected in the same orientation with respect to the coding strand of the related gene. Each promoter sequence was then randomized to obtain 1,000 random, shuffled sequences. In each randomization, the sequence of the promoter was shuffled so that we maintained the same base composition and length but created a non-biologically meaningful sequence. Promoter and shuffled sequences were inspected for the presence of STRs by running Phobos version 3.3.12 (Mayer, 2007) using the options *reportUnit=1* to conduct a strand-specific analysis, *minUnitLen=2* to exclude homopolymer repeats, and *outputFormat=3* to generate the one-per-line output format that is easier to parse. As in the genome-wide analysis, the software selected STRs with a minimum length of 8 bp and a unit size of at least 2 (*e.g.* from 4 dinucleotides, 3 trinucleotides and 2 tetranucleotides). To assess the statistical enrichment of STRs in promoters, the occurrences of each STR were counted in both the real and the randomized data sets, and the corresponding statistics were calculated. Calculations based on the shuffled promoters represent the 'expected' frequencies of the STRs based solely on the nucleotidic composition and were used to calculate the Z-scores, while the calculations based on real promoters represent the 'observed' occurrences. Functional enrichment analysis was conducted performing the Fisher exact test comparing for each class the proportion of class specific genes in the total set of annotated genes with the proportion of class specific genes in the set associated to the specific promoters. *P*-values were corrected using the Benjamini and Hochberg (BH) method.

A similar strand-specific STR analysis was performed for the genomic regions surrounding the gene TSS to generate the metagene plots shown in Figure 4.

Results

Diatom genomes are enriched in triplet repeats

To generate a comprehensive catalogue of STRs in marine diatoms, we surveyed the genomes of four different species: i) the centric diatom *Thalassiosira pseudonana*, ii) the pennate diatom *Phaeodactylum tricornutum*, iii) the psychrophilic diatom *Fragilariopsis cylindrus*, and iv) the neurotoxin domoic acid-producing pennate diatom *Pseudo-nitzschia multistriata*. We found that the total number of STRs varies enormously among the species, ranging from a minimum of 51,042 sequences in *P. tricornutum* to a maximum of 472,979 in *F. cylindrus* (Table 1). Not surprisingly, we found that the total number of STRs increases with the genome size (Table 1). Over 86% of the STRs are perfect, pure repeats. In order to verify the correctness of our predictions, we searched for and confirmed the presence of CCTAAC repeats that are normally located at the telomere regions of each chromosome. From the STR analysis we gathered the full list of diatom repeats along with their chromosomal location, repeat size, copy number, alignment score, consensus and sequence.

	Phaeodactylum Tricornutum		***Thalassiosira Pseudonana***		***Pseudo-nitzschia multistriata***		***Fragilariopsis cylindrus***	
Genome	ASM15095v2, Feb 2010		ASM14940v2, May 2014		Psmu1.4, Jul 2015		CCMP1102 v1, Nov 2008	
Length	27,568,093		32,437,365		59,304,822		80,540,407	
Scaffolds	89		64		1,099		271	
GC %	47.46		45.59		45.48		38.92	
Genes	12,392		11,870		12,008		37,171	
Source	Ensembl		Ensembl		Internal		JGI	
STRs	**perfect**	**imperfect**	**perfect**	**imperfect**	**perfect**	**imperfect**	**perfect**	**imperfect**
Di-nucleotide	5,101	270	7,210	430	8,526	649	26,629	4,047
Tri-	14,263	824	40,617	3,198	65,289	9,390	169,655	27,268
Tetra-	7,769	423	10,380	634	22,745	2,416	49,622	5,416
Penta-	7,203	262	11,482	624	26,711	3,998	50,793	7,420
Hexa-	11,564	466	21,260	955	40,328	6,422	85,833	14,930
Hepta-	1,635	65	3,838	202	10,030	904	13,734	1,551
Octa-	573	40	796	135	2,546	385	5,055	732
Nona-	394	40	795	80	2,958	845	6,842	1,840
Deca-	100	50	120	22	896	353	1,141	471
Total (%)	*4,602 (95.2)*	*2,440 (4.8)*	*96,498 (93.9)*	*6,280 (6.1)*	*180,029 (87.7)*	*25,362 (12.3)*	*409,304 (86.6)*	*63,675 (13.4)*
Total STRs	51,042		102,778		205,391		472,979	
Total STRs per genome length (kb)	1.85		3.17		3.46		5.87	

Table 1. Summary of short tandem repeats detected in diatom genomes.

Typically, STRs can vary in the size and in the number of copies of their repeat units. For the first time, we report that the most common STRs found in diatoms are DNA motifs with a repeat unit of 3 or 6 bases (Table 1; Fig. 1A). Triplet repeats alone account for almost one-third of the total STR sequence set (from 29.6 % in *P. tricornutum* to 42.6 % in *T. pseudonana*); while, together, triplets and hexaplets represent more than half of the repeat set. Most of the diatom STRs have up to 5 repeat copies, and only a small fraction more than 6 and up to 146 (*P. tricornutum*: 4.1%; *T. pseudonana*: 4.2%; *P. multistriata*: 7.7%; *F. cylindrus*: 13.2%) (Fig. 1B). In all species inspected, around half of the STRs has 3 repeat copies (Fig. 1B).

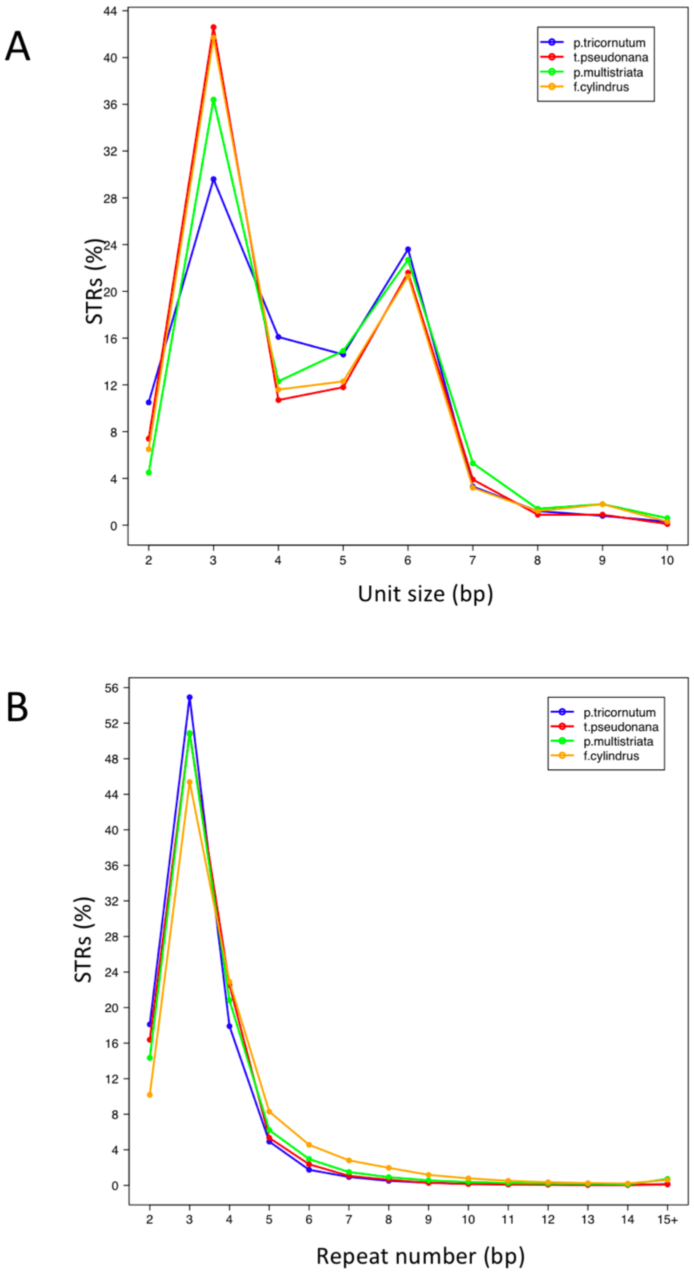


Figure 1. Characterization of STRs in diatom genomes. A) STR Distribution by Unit Size; B) STR Distribution by Repeat Number.

The sequence AAC is the most frequent repeat with a total of 61,199 occurrences across all diatoms. AAC occupies the first position in *T. pseudonana* and in *F. cylindrus,* and the second position in *P. tricornutum* and in *P. multistriata* (Table 2). This result extends our previous findings that the genome of *F. cylindrus* is enriched in CAA repeats (equivalent to AAC) (Mock et al., 2017). The second most prevalent repeat in diatoms is AAG, which is at the second position in *T. pseudonana*, at the third position in *P. tricornutum* and in *P. multistriata*, and at the fourth position in *F. cylindrus*, with a total of 43,987 occurrences collectively. Along with repeats that are highly frequent in all diatoms, we also found repeats that are at the top position in one species only. For instance, ACG is at the first position in *P. tricornutum*, but at the fifth position in *P. multistriata* and at the sixth position in *T. pseudonana* and in *F. cylindrus.* Likewise, ACT is at the first position in *P. multistriata*, but at the fifth, eighth and ninth positions in *F. cylindrus*, *T. pseudonana* and *P. tricornutum*, respectively. Finally, the AAT repeat is at the third position in *F. cylindrus* with 32,063 occurrences but at the very last position in *P. tricornutum*, *T. pseudonana* and *P. multistriata*. Taken together our results indicate for the first time that diatom genomes are highly enriched in triplet repeats.

	Unit	Number of STRs	Avg Repeat Number	StdDev Repeat Number	Phobos total score
P. tricornutum	ACG	2,851	3.51	0.86	20,176
	AAC	2,435	3.86	1.94	19,352
	AAG	2,293	3.34	0.78	15,418
	AGC	1,510	3.33	0.65	10,274
	CCG	1,395	3.28	0.55	9,236
	ACC	1,378	3.45	0.85	9,659
	ATC	1,179	3.37	0.87	8,034
	AGG	902	3.36	0.77	6,177
	ACT	781	3.69	1.93	5,750
	AAT	363	3.25	0.62	2,382
T. pseudonana	AAC	9,118	3.94	3.11	72,909
	AAG	6,817	3.43	0.85	47,302
	AGG	6,379	3.56	1	45,831
	ATC	5,925	3.55	0.97	42,873
	AGC	4,336	3.67	1.21	33,056
	ACG	4,336	3.63	0.99	32,016
	ACC	4,149	3.61	1.02	30,461
	ACT	1,337	4.03	7.38	10,229
	CCG	903	3.43	0.71	6,359
	AAT	515	3.17	0.38	3,313
P. multistriata	ACT	16,239	5.29	6.01	175,874
	AAC	12,845	4.42	3.94	116,914
	AAG	9,469	3.58	1.25	67,358
	AGC	8,276	4.07	2.27	68,513
	ACG	6,746	4	1.73	54,984
	ACC	5,838	3.65	1.55	43,735
	AGG	5,507	3.58	1.45	40,123
	ATC	4,288	3.71	3.56	32,075
	CCG	3,371	3.49	0.8	23,949
	AAT	2,100	3.51	1.42	14 847
F. cylindrus	AAC	36,801	4.56	2.24	344,162
	ATC	33,137	4.42	2.24	296,677
	AAT	32,063	4.12	1.96	265,976
	AAG	25,408	4.09	1.78	210,149
	ACT	18,782	4.28	1.88	166,747
	ACG	17,029	4.37	1.92	155,388
	AGC	14,541	4.55	2.22	139,321
	ACC	10,979	4.01	1.51	91,488
	AGG	6,609	3.96	1.54	54,975
	CCG	1,574	3.7	1.06	12,133

Table 2. Triplet STRs identified in diatom genomes.

STRs are abundant in diatom promoters and introns

To further characterize the repeat sequence set, we examined its distribution with respect to the genomic features by determining the occurrences of STRs in exons, introns, promoters (*i.e.* the 500 nucleotides upstream of gene TSS features) and intergenic regions. We identified highest occupancy in "promoter-TSS" regions in all species with the exception of *F. cylindrus* where STRs are slightly more abundant in introns (Fig. 2A). Through the visual inspection of the 20,000 nucleotides centered around the TSS of all diatom genes, we confirmed that STRs are preferentially located in promoters (Fig. 2B). We also noted that the density of STRs decreases near the gene TSS in all species (Fig. 2B). As second category after promoters-TSS, we found a high number of STRs in diatom "introns" (Fig. 2A). Further analyses indicated that STRs are significantly over-represented in the first introns of genes of *T. pseudonana* (BH-FDR adjusted *p*-value = 2.85 E-14), *P. multistriata* (BH-FDR adjusted *p*-value =1.38 E-13) and *F. cylindrus* (BH-FDR adjusted *p*-value = 1.16 E-12), and to some extent of *P. tricornutum* (BH-FDR adjusted *p*-value = 0.15). Instead there was no significant enrichment in the other introns (*e.g.* BH-FDR adjusted *p*-value = 1 in non-first introns of all diatom species). Finally, after promoters and introns, STRs occupy "exons" as third category, and "intergenic regions" as fourth and last category (Fig. 2A). Taken together these results show that STRs are very abundant in promoters and introns of diatoms and that they generally present the following decreasing order of occupancy: promoter-TSS, introns, exons, and intergenic regions.

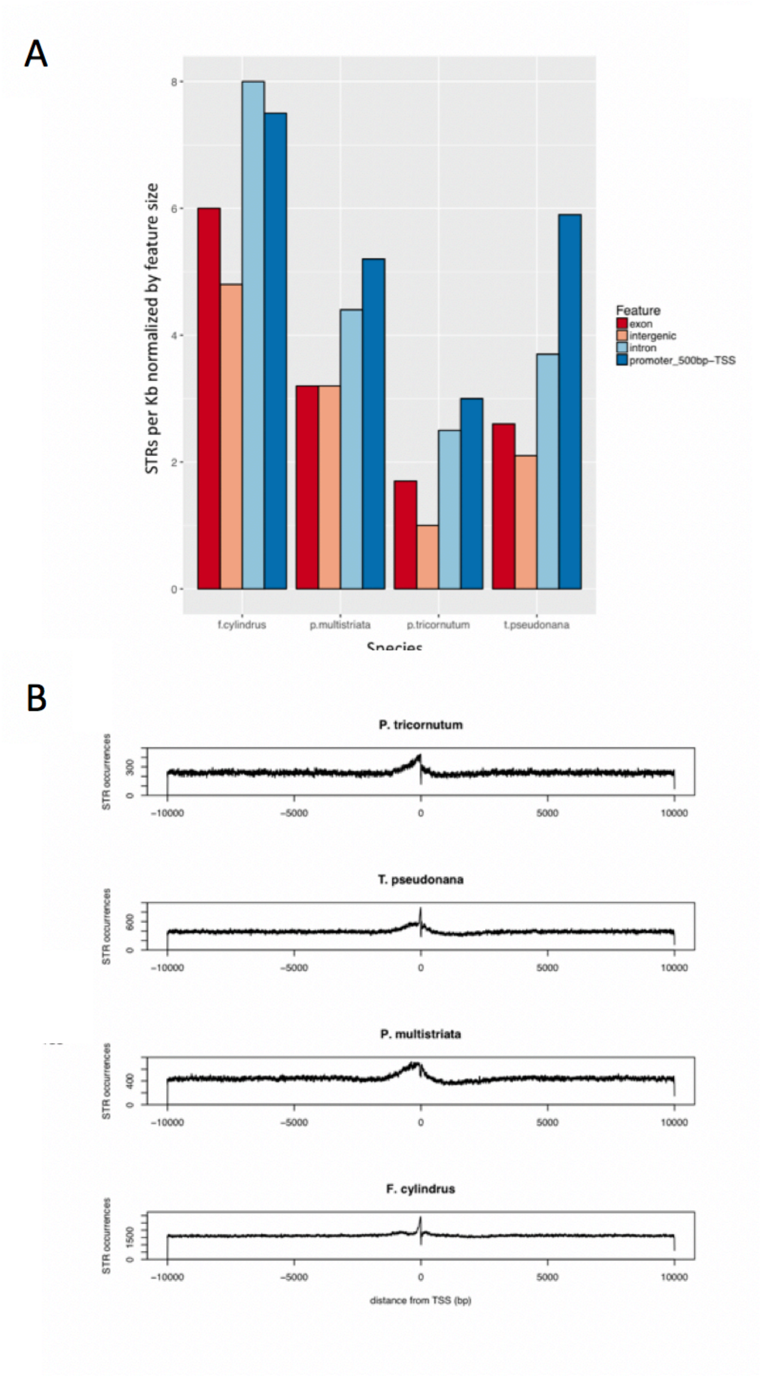


Figure 2. Distribution of STRs in annotated regions. A) Number of STRs located in exons, introns, promoter-TSS and intergenic regions. Values are normalized by feature size and reported in kilobases (kb). B) STR Distribution within the 20kb region surrounding gene TSS features.

Diatom promoters are enriched in AAC repeats

To investigate whether the strong signal in promoters was determined by one or more DNA repeat motifs, we refined our search in the 500 bp upstream of each TSS. We compared the STR occurrences found in each promoter sequence against those measured in a set of 1,000 shuffled sequences to assess the significance of the enrichment (Fig. 3A). Over two-thirds of diatom promoters harbor at least one STR, ranging from 68.3% promoters in *P. tricornutum* to 93.5% in *F. cylindrus*. Like for the genome-wide analysis, triplet repeats are very abundant also in promoters representing almost one third of the STRs (42,944 out of 147,843 total promoter STRs) (Fig. 3B); however, while triplets are the top class in *T. pseudonana* and in *F. cylindrus* promoters, tetra- and penta-nucleotide repeats are respectively more abundant than triplets in *P. tricornutum* and in *P. multistriata* promoters (Fig. 3B). Among triplets, AAC is the most frequent motif in *P. tricornutum*, in *T. pseudonana* and in *F. cylindrus* with a clear and sharp peak before the TSS (Fig. 3C), but not in *P. multistriata* where AGT is at the top (Fig. 3C). Among tetraplets, the top repeats are ACGT in *P. multistriata*, AGCT in *P. tricornutum*, AAAT in *F. cylindrus* and AAAC in *T. pseudonana*. The latter is also abundant in *F. cylindrus*. Among pentaplets, ACCGT and ACGGT are most abundant in *P. multistriata* with 1,512 counts overall (Fig. 3C), while other top frequent pentaplets are ACCCT in *P. tricornutum*, AACAC in *T. pseudonana* and AAAAC in *F. cylindrus* (Fig. 3C).

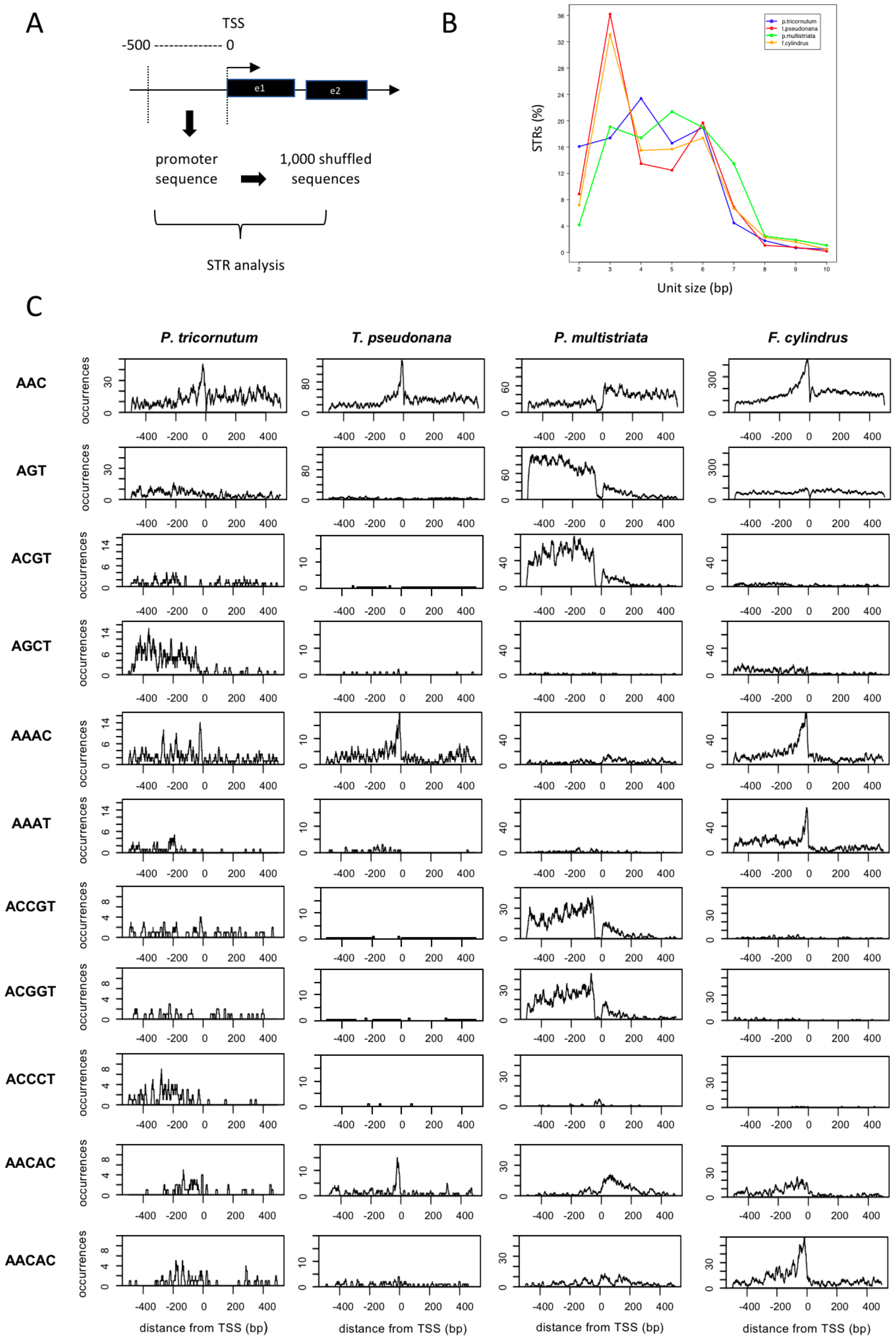
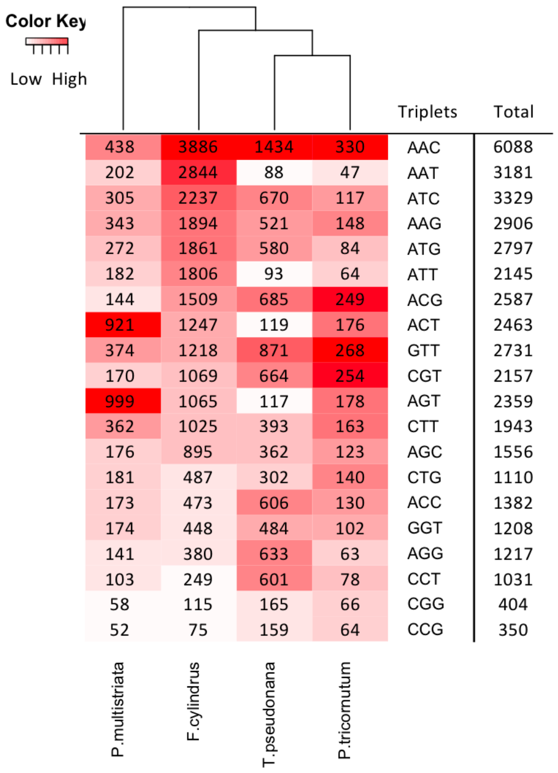


Figure 3. STR analysis in diatom promoters. A) Selection of promoters for STR analysis. B) Distribution of promoter-STRs by Unit Size. 3) Distribution around the TSS features of most frequent tri-, tetra- and penta-promoter STRs.

We further examined the promoter STRs focusing only on triplets. Hierarchical clustering analysis of STR occurrences indicated two similarity groups, with *P. tricornutum*, *T. pseudonana* and *F. cylindrus* on one side and *P. multistriata* alone on the other (Fig. 4). For comparison, we obtained similar results on tetraplets. AAC is the most statistically over-represented triplet in promoters (BH-FDR adjusted *p*-value = 0 in all diatoms) with a total of 6,088 occurrences (Fig. 4). This motif is at the top position in *P. tricornutum*, in *T. pseudonana* and in *F. cylindrus* but only third in *P. multistriata*. The relative distribution of AAC in the 4,000 nucleotides around the TSS confirms the strong enrichment in promoters of *P. tricornutum*, *T. pseudonana* and *F. cylindrus* (Fig. 3C; Fig. 4).



P.multistriata	F.cylindrus	T.pseudonana	P.tricornutum	Triplets	Total
438	3886	1434	330	AAC	6088
202	2844	88	47	AAT	3181
305	2237	670	117	ATC	3329
343	1894	521	148	AAG	2906
272	1861	580	84	ATG	2797
182	1806	93	64	ATT	2145
144	1509	685	249	ACG	2587
921	1247	119	176	ACT	2463
374	1218	871	268	GTT	2731
170	1069	664	254	CGT	2157
999	1065	117	178	AGT	2359
362	1025	393	163	CTT	1943
176	895	362	123	AGC	1556
181	487	302	140	CTG	1110
173	473	606	130	ACC	1382
174	448	484	102	GGT	1208
141	380	633	63	AGG	1217
103	249	601	78	CCT	1031
58	115	165	66	CGG	404
52	75	159	64	CCG	350

Figure 4. Heatmap representation of triplet repeat occurrences found in promoters. Rows represent all possible triplet repeats; columns represent all the diatom species under investigation. For each species (*e.g.* within each column), a white-to-red color gradient shows the number of STR occurrences (white = zero STRs; dark red = highest number of repeat occurrences found in that specific species).

Other triplets were highly frequent only in one or two species. For instance, we found a specific enrichment of A[CG]T motifs in *P. multistriata*: AGT was first in *P. multistriata* but fifth, eleventh and eighteenth in *P. tricornutum*, *F. cylindrus* and *T. pseudonana*; ACT was second in *P. multistriata* but sixth, eighth and seventeenth in *P. tricornutum*, *F. cylindrus* and *T. pseudonana*. The AGT and ACT motifs together account for one-third of all triplets located in *P. multistriata* promoters (Fig. 4). As additional example, GTT was second in *P. tricornutum* and in *T. pseudonana*, fourth in *P. multistriata*, but only ninth in *F. cylindrus* (Fig. 4). Finally, several T and A rich motifs were extremely abundant in *F. cylindrus* promoters (Fig. 4): AAT was second in *F. cylindrus* but ninth in *P. multistriata* and twentieth in both *P. tricornutum* and *T. pseudonana*, AAG was fourth in *F. cylindrus* but eleventh, thirteenth and fifteenth in *T. pseudonana, P. tricornutum* and *P. multistriata,* and ATT was sixth in *F. cylindrus* but tenth, eighteenth and nineteenth respectively in *P. multistriata*, *P. tricornutum* and *T. pseudonana*. Interestingly, we found AAT and ATT triplets significantly represented in *F. cylindrus* promoters (ATT: 1,806 observed vs 564.6 expected with BH-FDR adjusted *p*-value = 0, and AAT: 2,844 observed vs 748.9 expected with BH-FDR adjusted *p*-value = 0) and *P. multistriata* promoters (ATT: 182 observed vs 163.4 expected with BH-FDR adjusted *p*-value = 0.066, and AAT: 202 observed vs 165.8 expected with BH-FDR adjusted *p*-value = 0.0033), but not in *P. tricornutum* promoters (ATT: 64 observed vs 112.18 expected with BH-FDR adjusted *p*-value = 1, and AAT: 42 observed vs 112.91 expected with BH-FDR adjusted *p*-value = 1) and *T. pseudonana* promoters (ATT: 93 observed vs 90.4 expected with BH-FDR adjusted *p*-

value = 0.41, and AAT: 88 observed vs 110.8 expected with BH-FDR adjusted *p*-value = 0.90).

Taken together, we conclude that AAC is by far the most prevalent repeat motif in diatom promoters, although AGT and ACT are the most abundant triplets in *P. multistriata* promoters (whereas AAC is only third) and AAT and ATT (*i.e.* TATA-box like DNA motifs) are very frequent in *F. cylindrus* promoters.

Discussion

The current study presents the first genome-wide catalogue of short repetitive elements in lower eukaryotes such as marine diatoms. Because repetitive sequences are more difficult to detect using standard high-throughput sequencing technologies (Bahlo et al., 2018), the study of short tandem repeats (STRs) has been so far neglected compared to the one of single nucleotide polymorphisms (SNPs) and short insertions and deletions (InDels). As a result, our collective knowledge of variations in STRs remains scarce.

Recent studies indicated that STRs are located both in genes and in non-coding regions (Gemayel et al., 2010; Sawaya et al., 2013). In higher eukaryotic genomes, STRs are also found in promoters (Sawaya et al., 2013). The repetitive nature of these sequences might induce strand-splippage events in DNA replication resulting in mutations in the number of repeats with possible effects on the phenotype. In this light, expansion or contraction of promoter STRs might affect gene expression through several possible mechanisms: they can form transcription factor binding sites (Contente, Dittmer, Koch, Roth, & Dobbelstein, 2002), can alter spacing between regulatory elements (Rockman & Wray, 2002), or modulate epigenetics via DNA methylation (Quilez et al., 2016). These unstable repetitive elements in promoters are very important as they may facilitate evolutionary changes in phenotypes (Gemayel et al., 2010).

The advent of high-throughput sequencing technologies has enabled the development of novel genomic resources. The genome sequences of four marine diatom species have been released for public use in recent years. The small genomes of *Thalassiosira* *pseudonana* and *Phaeodactylum tricornutum* were the first ones to be sequenced and provided an unprecedented wealth of information about diatom biology and their evolution (Armbrust, 2009; Bowler et al., 2008). In addition, the genome sequences of *Fragilariopsis cylindrus* and *Pseudo-nitzschia multistriata* have recently provided insights into genome evolution of diatom species that are adapted to live in extreme conditions of the sea ice in the Southern Ocean (Basu et al., 2017; Mock et al., 2017). Our genome-wide analysis reveals for the first time that the genomes of these unicellular photosynthetic eukaryotes are rich in triplet repeats that are mostly located in non-coding regions, particularly in promoters. This finding is novel and is confirmed in all four diatoms regardless of their genome size. Promoter repeats are very abundant within the 500 nucleotides upstream of the gene TSS features and comprise over-represented DNA motifs. Overall, AAC is the most statistically over-represented triplet in promoters. This motif is at the top position in *P. tricornutum*, in *T. pseudonana* and in *F. cylindrus* but only third in *P. multistriata*. Noteworthy, AGT and ACT are the most abundant triplets in *P. multistriata* promoters, while AAT and ATT (TATA-box like DNA motifs) are also very frequent in *F. cylindrus* promoters.

We hypothesize that the significant enrichments of STRs in the promoters of diatoms might be important for the regulation of transcription. Based on this assumption, we are currently testing some of these genomic elements to understand if different number of copies of the simple repeats into STRs might modulate transcriptional levels. From what we are learning from the sex locus of *Pseudo-nitzschia multistriata* it is evident that STRs are at least linked to specific gene expression (Russo *et al*, in revision). In order to prove the molecular function of these hypothetic regulatory sequences, we selected promoters of genes of interest for a modular cloning analysis in *Phaeodactylum tricornutum* cells (Table 3).

Promoter ID	Gene ID, Description	STR Start	STR End	Unit Size	Repeat Number	Repeat Unit	Perfection
32_76837_77337_+	Phatr3_J50610, Predicted protein	342	436	3	31.667	AAC	100
bd_31x35_90668_91168_+	Phatr3_EG02359, MPDC (mevalonate diphosphate decarboxylase)	415	457	2	21.5	CT	97.674
bd_31x35_90668_91168_+	Phatr3_EG02359, MPDC (mevalonate diphosphate decarboxylase)	461	481	2	10.5	AC	95.238
bd_31x35_90633_91133_-	Phatr3_EG02362, POLA (DNA polymerase)	10	52	2	21.5	AG	97.674
bd_31x35_90633_91133_-	Phatr3_EG02362. POLA (DNA polymerase)	450	490	3	13.333	AAC	92.5

Table 3. Promoter STRs selected for modular cloning analysis.

First, we chose the promoter of the gene Phatr3_J50610 which is perhaps the most important for several reasons: 1) it contains the longest STR with the repeat AAC which also very significant in many other diatom species, therefore testing this promoter might give us useful information on STRs in diatoms in general; 2) it is long and there are variations in its length reported in Ensembl which means that there exists several alleles and therefore it could be easy to validate the regulatory activity of the reference sequence as well as of the other alleles. Second, the promoter of the gene Phatr3_EG02362 which is annotated as DNA polymerase gene and has a STR of decent length. Third, the promoter of the gene Phatr3_EG02359 which is annotated as the mevalonate diphosphate decarboxylase (MPDC), a gene involved in the isoprenoid synthesis, and has a STR of good size. It is also important to note that these two genes, Phatr3_EG02362 and Phatr3_EG02359, are located in the same genomic region and positioned head-to-head. The intergenic region between these two head-to-head genes is also of interest for the several reasons: 1) it is rather small and therefore easily to handle in the lab; 2) one of the genes should be expressed in many conditions at high levels (POLA) while the other should be more finely regulated; 3) in the case in which STRs have a regulatory function we do not know whether the strand is important in the directing the transcription and this region could be helpful in understanding; 4) it contains more than one type of STRs. In conclusion, this region is small but complex, and it would give us different information compared to the promoter region of the Phatr3_J50610 gene which is simpler.

References:

Abe, H., & Gemmell, N. J. (2016). Evolutionary Footprints of Short Tandem Repeats in Avian Promoters. *Nature Publishing Group*, (August 2015), 1–11. http://doi.org/10.1038/srep19421

Armbrust, E. V. (2009). The life of diatoms in the world ' s oceans, *459*(May), 185–192. http://doi.org/10.1038/nature08057

Basu, S., Patil, S., Mapleson, D., Russo, M. T., Vitale, L., Fevola, C., … Ferrante, M. I. (2017). Finding a partner in the ocean: molecular and evolutionary bases of the response to sexual cues in a planktonic diatom. *New Phytologist*, *215*(1), 140–156. http://doi.org/10.1111/nph.14557

Bozarth, A., Maier, U., & Zauner, S. (2009). Diatoms in biotechnology : modern tools and applications, 195–201. http://doi.org/10.1007/s00253-008-1804-8

Campuzano, V., Montermini, L., Molto, M. D., Pianese, L., Cossee, M., Cavalcanti, F., … Pandolfoll, M. (n.d.). Friedreich ' s Ataxia : Autosomal Recessive Disase Caused by an Intronic GAA Triplet Repeat Expansion, (18).

Day JW, R. L. R. pathogenesis of the myotonic dystrophies. (n.d.). RNA pathogenesis of the myotonic dystrophies. *Neuromuscul Disord*, 5–16.

Ellegren, H. (2004). MICROSATELLITES : SIMPLE SEQUENCES WITH COMPLEX EVOLUTION, *5*(June). http://doi.org/10.1038/nrg1348

Eo, J., Lee, H., Nam, G., Kwon, Y., Choi, Y., Choi, B., … Kim, H. (2016). Association of DNA methylation and monoamine oxidase A gene expression in the brains of different dog breeds. *Gene*, *580*(2), 177–182.

http://doi.org/10.1016/j.gene.2016.01.022

Gemayel, R., Vinces, M. D., Legendre, M., & Verstrepen, K. J. (2010). Variable Tandem Repeats Accelerate Evolution of Coding and Regulatory Sequences. *Annual Review of Genetics*, *44*(1), 445–477. http://doi.org/10.1146/annurev-genet-072610-155046

Gijselinck, I., Langenhove, T. Van, Zee, J. Van Der, Sleegers, K., Philtjens, S., Kleinberger, G., … Broeckhoven, C. Van. (2012). A C9orf72 promoter repeat expansion in a Flanders-Belgian cohort with disorders of the frontotemporal lobar degeneration-amyotrophic lateral sclerosis spectrum : a gene identification study. *The Lancet Neurology*, *11*(1), 54–65. http://doi.org/10.1016/S1474-4422(11)70261-7

Iii, J. W. F., Hammock, E. A. D., Hannan, A. J., & King, D. G. (2008). Simple sequence repeats : genetic modulators of brain function and behavior, (June). http://doi.org/10.1016/j.tins.2008.03.006

Kersey, P. J., Allen, J. E., Armean, I., Boddu, S., Bolt, B. J., Carvalho-Silva, D., … Staines, D. M. (2016). Ensembl Genomes 2016: more genomes, more complexity. *Nucleic Acids Research*, *44*(D1), D574–D580. Retrieved from http://dx.doi.org/10.1093/nar/gkv1209

Lynch, M., Sung, W., Morris, K., Coffey, N., Landry, C. R., Dopman, E. B., … Thomas, W. K. (2008). A genome-wide view of the spectrum of spontaneous mutations in yeast.

Maclean, H. E., Warne, G. L., & Zajac, J. D. (1996). Spinal and bulbar muscular atrophy : androgen receptor dysfunction caused by a trinucleotide repeat expansion, *135*(95), 149–157.

Mayer, C. (2007). PHOBOS – a tandem repeat search tool for complete genomes.2007. [http://www.rub.de/spezzoo/cm].

Mock, T., Otillar, R. P., Strauss, J., McMullan, M., Paajanen, P., Schmutz, J., … Grigoriev, I. V. (2017). Evolutionary genomics of the cold-adapted diatom Fragilariopsis cylindrus. *Nature*, *541*(7638), 536–540. Retrieved from http://dx.doi.org/10.1038/nature20803

Nordberg, H., Cantor, M., Dusheyko, S., Hua, S., Poliakov, A., Shabalov, I., … Dubchak, I. (2014). The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. *Nucleic Acids Research*, *42*(D1), D26–D31. Retrieved from http://dx.doi.org/10.1093/nar/gkt1069

Ohadi M, Valipour E, Ghadimi-Haddadan S, Namdar-Aligoodarzi P, B. A., & Kowsari A, Rezazadeh M, Darvish H, K. S. (2014). Core promoter short tandem repeats as evolutionary switch codes for primate speciation. *Am J Primatol.*, *77*(1), 34–43.

Orr, H. T., & Zoghbi, H. Y. (2007). Trinucleotide Repeat Disorders, 575–623. http://doi.org/10.1146/annurev.neuro.29.051605.113042

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. Retrieved from http://dx.doi.org/10.1093/bioinformatics/btq033

R Core Team. (2013). R: A Language and Environment for Statistical Computing. Vienna, Austria. Retrieved from http://www.r-project.org/

Richard, G., Kerrest, A., & Dujon, B. (2008). Comparative Genomics and Molecular Dynamics of DNA Repeats in Eukaryotes, *72*(4), 686–727. http://doi.org/10.1128/MMBR.00011-08

Russo, M. T., Annunziata, R., Sanges, R., Ferrante, M. I., & Falciatore, A. (2015).

The upstream regulatory sequence of the light harvesting complex Lhcf2 gene of the marine diatom Phaeodactylum tricornutum enhances transcription in an orientation- and distance-independent fashion. *Mar Genomics.*, *24*(Pt 1), 69–79.

Sawaya, S., Bagshaw, A., Buschiazzo, E., Kumar, P., Chowdhury, S., Black, M. A., & Gemmell, N. (2013). Microsatellite Tandem Repeats Are Abundant in Human Promoters and Are Associated with Regulatory Elements, *8*(2). http://doi.org/10.1371/journal.pone.0054710

Sonay, T. B., Carvalho, T., Robinson, M. D., Greminger, M. P., Comas, D., Highnam, G., … Wagner, A. (2015). Tandem repeat variation in human and great ape populations and its impact on gene expression divergence. *Genome Research*, *25*, 1591–1599.

Tabolacci, E., Palumbo, F., Nobile, V., & Neri, G. (2016). Transcriptional Reactivation of the FMR1 Gene . Fragile X Syndrome †, 1–16. http://doi.org/10.3390/genes7080049

Vardi, A., Thamatrakoln, K., Bidle, K. D., & Falkowski, P. G. (2009). Minireview D i a t o m g e n o m e s c o m e o f a g e. http://doi.org/10.1186/gb-2008-9-12-245

Vinces, M. D., Legendre, M., Caldara, M., Hagihara, M., & Verstrepen, K. J. (2011). NIH Public Access, *324*(5931), 1213–1216. http://doi.org/10.1126/science.1170097.Unstable

Appendix. Code Example: tandem_repeat_analysis.pl

```
#! /usr/bin/perl -w

=head1 NAME

=head1 SYNOPSIS

  perl tandem_repeat_analysis.pl [-s species_name] [-g genome_FASTA] [-a
annotation_GTF] [-c chr_sizes] [-nc_rna] [-cds] [-d source] [-force] [-h help]

=head1 DESCRIPTION

This script identifies and annotates tandem repeats from an entire genome sequence.

It requires a genomic sequence in FASTA format and related annotation in GTF format.

Typical usage is as follows:

  % perl tandem_repeat_analysis.pl -s Phaeodactylum_tricornutum -g
Phaeodactylum_tricornutum.ASM15095v2.29.dna.toplevel.fa -a
Phaeodactylum_tricornutum.ASM15095v2.29.gtf -c Phaeodactylum_tricornutum.chrsizes.txt

=head2 Options

The following options are accepted:

 --s=<species name>    Specify the species name.

 --g=<genome fasta>    Specify genome sequence in FASTA format.

 --a=<annotation gtf>  Specify annotation file name if any. GTF format is required.

 --c=<chr sizes>       Provide a file containing chromosome sizes

 --nc_rna              Consider ncRNAs, if defined.

 --cds                Build TSS, TTS and promter features using CDS coordinates
(instead of gene), if defined.

 --d=<source>          Source (ensembl, jgi or internal).

 --force               Re-load annotation.

 --help                This documentation.

=head1 AUTHOR

Guglielmo Roma

guglielmo.roma@gmail.com

=cut

use Cwd 'abs_path';
use File::Basename;
my ($name,$path,$suffix) = fileparse(abs_path($0));
require $path."../conf/conf.pl";

use strict;
use warnings;
use Getopt::Long qw( :config posix_default bundling no_ignore_case );
use Pod::Usage;
use Data::Dumper;

# Configuration variables

my %conf =  %::conf;
my $debug = $conf{'global'}{'debug'};
my $tmp_dir = $conf{'default'}{'tmp_dir'};
my $genomes_dir = $conf{'default'}{'genomes_dir'};
my $bioinfo_dir = $conf{'default'}{'bioinfo_dir'};
my $results_dir = $conf{'default'}{'results_dir'};
my %species_hash = %{$conf{'species'}};

my $PHOBOS = $conf{'PHOBOS'}{'command_genome'};
my $MaxPeriod = $conf{'PHOBOS'}{'MaxPeriod'};

# Functions declarations

=pod

      SCRIPT

=cut

my $USAGE = "perl tandem_repeat_analysis.pl [-s species_name] [-g genome_FASTA] [-a
annotation_GTF] [-c chr_sizes] [-nc_rna] [-cds] [-d source] [-force] [-h help]";

my ($species_name, $genome_fasta, $annotation_gtf, $chr_sizes, $cds, $nc_rna, $force,
$source, $show_help);

&GetOptions(
                  'species_name|s=s'      => \$species_name,
                  'genome_fasta|g=s'      => \$genome_fasta,
                  'annotation_gtf|a=s'    => \$annotation_gtf,
                  'chr_sizes|c=s'         => \$chr_sizes,
                 'source|d=s'             => \$source,
                  'nc_rna'                => \$nc_rna,
                  'cds'                  => \$cds,
                  'force'                 => \$force,
                  'help|h'                => \$show_help
                  )
  or pod2usage(-verbose=>2);
pod2usage(-verbose=>2) if $show_help;

# Considers user-specified options - if provided
if ($species_name && $genome_fasta && $annotation_gtf && $chr_sizes && $source) {
      # Overwrites default options from the configuration file
      undef(%species_hash);
      $species_hash {$species_name} {'genome_FASTA'} = $genome_fasta;
      $species_hash {$species_name} {'annotation_GTF'} = $annotation_gtf;
      $species_hash {$species_name} {'chrsizes'} = $chr_sizes;
        $species_hash {$species_name} {'source'} = $source;
}

# Dies, if users did not specify a species of interest in either command line or the
config file
die "You must specify a species name, a genome FASTA file, and an annotation file\n
Use -h for help"
        if (!%species_hash);

$cds=0 if(!$cds);

$debug && print STDOUT "Debugging species hash:\n";
$debug && print STDOUT Dumper %species_hash;
$debug && print STDOUT "\n";

#my $tmp_bed = $tmp_dir."/bed";
#mkdir($tmp_bed or die "$!");

foreach $species_name (keys %species_hash) {
      $genome_fasta = $species_hash{$species_name}{'genome_FASTA'};
      $annotation_gtf = $species_hash{$species_name}{'annotation_GTF'};
      $chr_sizes = $species_hash{$species_name}{'chrsizes'};
      $source = $species_hash{$species_name}{'source'};

      $debug && print STDOUT "Species: $species_name\n";
      $debug && print STDOUT "FASTA: $genome_fasta\n";
      $debug && print STDOUT "GTF: $annotation_gtf\n";
      $debug && print STDOUT "CHRSizes: $chr_sizes\n";
      $debug && print STDOUT "Source: $source\n";

      # Creates a result folder for each species
      my $now_string = localtime;
      $now_string =~ s/\W/_/g;
      my ($gender, $species) = split/["_","."]/, $genome_fasta;
      my $short_species_name = lc(substr($gender, 0, 1).".".$species);
      my $result_path = $results_dir.$short_species_name."_".$now_string;
      mkdir($result_path) or die "$!";
      chdir($result_path) or die "$!";

      # Executes PHOBOS
      my $TR_result_file=join('.', $genome_fasta, "dat");
      print ("$PHOBOS $genomes_dir/$genome_fasta $result_path/$TR_result_file\n");
      system ("$PHOBOS $genomes_dir/$genome_fasta $result_path/$TR_result_file");

      # Parses PHOBOS output and lists TRs in BED format
      print ("perl $path/perl/parsing_dat.pl $result_path/$TR_result_file\n");
      system ("perl $path/perl/parsing_dat.pl $result_path/$TR_result_file");

      # Generates TR distribution plots
      print ("awk '{if(\$4)print \$4}' $result_path/$TR_result_file".".txt | sort -n
>  $result_path/$TR_result_file".".txt.sort\n");
      system ("awk '{if(\$4)print \$4}' $result_path/$TR_result_file".".txt | sort -n
>  $result_path/$TR_result_file".".txt.sort");
      print "R --slave --args $result_path/$TR_result_file".".txt.sort \"Distribution
of Tandem Repeats ($gender $species)\" $MaxPeriod < $path/R/draw_TR_dist.r\n";
      system ("R --slave --args $result_path/$TR_result_file".".txt.sort
\"Distribution of Tandem Repeats ($gender $species)\" $MaxPeriod <
$path/R/draw_TR_dist.r");

      # removing the header line
      print ("sed 1d $result_path/$TR_result_file.bed >
$result_path/$TR_result_file.nh.bed\n");
      system ("sed 1d $result_path/$TR_result_file.bed >
$result_path/$TR_result_file.nh.bed");

      # Annotates non-redundant TRs
      my $annotate_cmd = "perl $path/annotate.pl -f
$result_path/$TR_result_file.nh.bed -a $genomes_dir/$annotation_gtf -c
$genomes_dir/$chr_sizes ";
      $annotate_cmd .= " --nc_rna" if ($nc_rna);
      $annotate_cmd .= " --force" if ($force);
      $annotate_cmd .= " --cds" if ($cds);
      $annotate_cmd .= " --source $source" if ($source);
      print STDOUT ($annotate_cmd."\n");
      system ($annotate_cmd);

      $debug && print "\nTandem repeat analysis for $species_name complete!\n";
}
```

Appendix. Code Example: annotate.pl

```
#! /usr/bin/perl -w

=head1 NAME

  annotate.pl

=head1 SYNOPSIS

  perl annotate.pl [-f feature_BED] [-a annotation_GTF] [-c chr_sizes] [-nc_rna] [-
cds] [-so source] [-force] [-h help]

=head1 DESCRIPTION

This script i) creates promoter, TSS and TTS annotation features, ii) annotates
features of interest provided in BED format using the annotation features.

Typical usage is as follows:

  % perl tandem_repeats/scripts/annotate.pl -f features.bed -a species.gtf -c
species.chrsizes.txt --nc_rna --force --source ensembl

=head2 Options

The following options are accepted:

 --f=<feature bed>      Specify file with genomic features. "6 fields" BED format
required.

 --a=<annotation gtf>   Specify annotation file name if any. GTF format is required.

 --c=<chr sizes>        Provide a file containing chromosome sizes.

 --nc_rna               Consider ncRNAs, if defined.

 --cds                  Build TSS, TTS and promter features using CDS coordinates
(instead of gene), if defined.

 --source               Annotation source (ensembl, jgi or internal).

 --force                Re-load annotation.

 --help                 This documentation.

=head1 AUTHOR

Guglielmo Roma

guglielmo.roma@gmail.com

=cut

use Cwd 'abs_path';
use File::Basename;
my ($name,$path,$suffix) = fileparse(abs_path($0));
require $path."../conf/conf.pl";

use strict;
use warnings;
use Getopt::Long qw( :config posix_default bundling no_ignore_case );
use Pod::Usage;
use Data::Dumper;
use File::Spec::Functions qw(catfile rootdir);
use List::Util qw[min max];
use File::Temp qw/ :POSIX /;

# Configuration variables
```

```
my %conf =  %::conf;
my $debug = $conf{'global'}{'debug'};
my $promoter_sizes = $conf{'global'}{'promoter_sizes'};
my %promoter_sizes = %{$promoter_sizes};
my $transcript_feature_types = $conf{'global'}{'transcript_feature_types'};
my %transcript_feature_types = %{$transcript_feature_types};
my @transcript_feature_types = values %transcript_feature_types;

# Functions declarations

=pod

      SCRIPT

=cut

my $USAGE = "perl annotate.pl [-f feature_BED] [-a annotation_GTF] [-c chr_sizes] [-
nc_rna] [-cds] [-so source] [-force] [-h help]";

my ($feature_bed, $annotation_gtf, $chr_sizes, $nc_rna, $cds, $source, $force,
$show_help);

&GetOptions(
                  'feature_bed|f=s'       => \$feature_bed,
                  'annotation_gtf|a=s'    => \$annotation_gtf,
                  'chr_sizes|c=s'         => \$chr_sizes,
                   'cds'                  => \$cds,
                  'nc_rna'                => \$nc_rna,
                  'source|so=s'           => \$source,
                  'force'                 => \$force,
                  'help|h'                => \$show_help
                  )
  or pod2usage(-verbose=>2);
pod2usage(-verbose=>2) if $show_help;

# Dies if files are not provided
die "You must specify an feature file in bed format\n Use -h for help"
  if !$feature_bed;

die "You must specify an annotation file in gtf format\n Use -h for help"
  if !$annotation_gtf;

die "You must specify a chr_sizes file\n Use -h for help"
  if !$chr_sizes;

die "You must specify a source (ensembl, jgi, or internal)\n Use -h for help"
  if !$source;

$cds = 0 if(!$cds);

# Checks if TSS and TTS files are already created
my ($ann_name, $ann_path, $ann_suffix) = fileparse($annotation_gtf);

# Defines annotation file names
my $promoter_file            = $annotation_gtf.".promoter";
my $tss_file                       = $annotation_gtf.".tss";
my $tts_file                       = $annotation_gtf.".tts";
my $exon_file                      = $annotation_gtf.".exon";
my $intron_file                    = $annotation_gtf.".intron";
my $gene_file                      = $annotation_gtf.".gene";

$debug && print STDOUT "Using $annotation_gtf to create promoter, TSS, TTS and exon
features.\n";

# Creates promoter features
foreach my $promoter_id (keys %promoter_sizes) {
      my $promoter_size = $promoter_sizes {$promoter_id};

      if (-e $promoter_file."_".$promoter_size && !defined($force)) {
            $debug && print STDOUT "Promoter features are available. Skipping the 
build. Use --force option to reload the annotation.\n";

      } else {
            $debug && print STDOUT "Creating promoter features with size 
$promoter_size... ";

            system("rm ".$promoter_file."_".$promoter_size);

            my $cmd;
            if ($cds==0){
                  $cmd = 'awk \'{FS="\t";OFS="\t"}{if ($3=="gene") {if ($7=="+")
{print $1,$2,"promoter_'.$promoter_size.'",$4,$4,$6,$7,$8,$9;} else if($7 == "-"){ 
print $1,$2,"promoter_'.$promoter_size.'",$5,$5,$6
,$7,$8,$9}}}\' '.$annotation_gtf.' | slopBed -g '.$chr_sizes.' -i stdin -l 
'.$promoter_size.' -r 0 -s |  awk \'{if ($4 >= "0" && $5 >= "0" && $5-
$4=='.$promoter_size.') print $_;}\' > '.$promoter_file."_".$promoter_size;
            } else{
                  $cmd = 'awk \'{FS="\t";OFS="\t"}{if ($3=="CDS") {if ($7=="+")
{print $1,$2,"promoter_'.$promoter_size.'",$4,$4,$6,$7,$8,$9;} else if($7 == "-"){ 
print $1,$2,"promoter_'.$promoter_size.'",$5,$5,$6,
$7,$8,$9}}}\' '.$annotation_gtf.' | slopBed -g '.$chr_sizes.' -i stdin -l 
'.$promoter_size.' -r 0 -s |  awk \'{if ($4 >= "0" && $5 >= "0" && $5-
$4=='.$promoter_size.') print $_;}\' > '.$promoter_file."_".$promoter_size;
            }

            system($cmd);
                $debug && print STDOUT $cmd."\n";

            $debug && print STDOUT "Done!\n";
      }
}

# Creates TSS, TTS and EXON features
if (-e $tss_file && !defined($force)) {

      $debug && print STDOUT "TSS, TTS, EXON and GENE features are available. 
Skipping the build. Use --force option to reload the annotation.\n";

} else {
      $debug && print STDOUT "Creating TSS, TTS, EXON and GENE features... ";

      my ($TSS_list, $TTS_list, $exon_list, $gene_list);

      open (FILE, $annotation_gtf) or die "Cannot open $annotation_gtf: $!";

      while (my $row = <FILE>) {
            chomp ($row);

            if ($row  !~ /^#/) {
                  my @fields =  split(/\t/, $row);

                  my $cmd;

                  if ($cds==0){
                        $gene_list .= "$row\n" if ($fields[2] eq "gene");
                  } else {
                        $gene_list .= "$row\n" if ($fields[2] eq "CDS");
                  }
                  $exon_list .= "$row\n" if ($fields[2] eq "exon");

                  my ($start_promoter, $TSS, $TTS);

                  # Retrieves only transcript feature types that are predefined in 
the config file
                  next unless ((grep {$_ eq $fields[2]} @transcript_feature_types) || 
($cds && $fields[2] eq "CDS"));

                  if ($fields[6] eq "+") {
                        $TSS = min($fields[3], $fields[4]);
                        $TTS = max($fields[3], $fields[4]);

                  } else {

                        $TSS = max($fields[3], $fields[4]);
                        $TTS = min($fields[3], $fields[4]);

                  }

                  $fields[3] = $TSS;
                  $fields[4] = $TSS;

                  $TSS_list .= join("\t", @fields)."\n";

                  $fields[3] = $TTS;
                  $fields[4] = $TTS;

                  $TTS_list .= join("\t", @fields)."\n";
            }
      }

      close FILE;

      open (TSS, ">$tss_file");
      print TSS "$TSS_list";
      close (TSS);

      open (TTS, ">$tts_file");
      print TTS "$TTS_list";
      close (TTS);

      open (EXON, ">$exon_file");
      print EXON "$exon_list";
      close (EXON);

      open (GENE, ">$gene_file");
      print GENE "$gene_list";
      close (GENE);

      # create intron features
      my $annotype = 'ensembl';
      my $exon_sort_cmd = "sort -n -k1 -k4 $exon_file > $exon_file".".sort";
      my $reformat_cmd = "perl  $path/perl/reformatGTF_forintronsize.pl 
$exon_file".".sort $source";
      my $intron_cmd = "perl $path/genomegtf2intronbed.pl -gtf $annotation_gtf > 
$intron_file";
      my $intron_sort_cmd = "sort -n -k1 -k4 $intron_file > $intron_file".".sort";
      my $merge_introns_cmd = "bedtools merge -i $intron_file".".sort > 
$intron_file".".sort.nr.bed";
      my $intron_size_cmd = 'awk \'{sum += ($3-$2+1)} END {print sum}\' 
'.$intron_file.'.sort.nr.bed > '.$intron_file.".size";

      system ($exon_sort_cmd);
      system ($exon_sort_cmd);
      system ($reformat_cmd);
      system ($intron_cmd);
      system ($intron_sort_cmd);
      system ($merge_introns_cmd);
      system ($intron_size_cmd);

      $debug && print STDOUT "Done!\n";
}

foreach my $promoter_id (keys %promoter_sizes) {
      my $promoter_size = $promoter_sizes {$promoter_id};

      # Overlapping annotation features
      my $overlapping_anno_cmd = "intersectBed -a $feature_bed -b $tss_file -c |
intersectBed -a stdin -b $promoter_file"."_"."$promoter_size -c | intersectBed -a
stdin -b $tts_file -c | intersectBed -a stdin -b $exon_
file -c | intersectBed -a stdin -b $intron_file -c | intersectBed -a stdin -b
$annotation_gtf -c | uniq > $feature_bed"."_".$promoter_size."anno.txt";
      $debug && print STDOUT ("$overlapping_anno_cmd\n");
      system ($overlapping_anno_cmd);

      my $parse_overlapping_anno_cmd = "perl $path/perl/parse_annotation.pl
$feature_bed"."_".$promoter_size."anno.txt | sort | uniq -c | sort -nr >
/$feature_bed"."_".$promoter_size."anno.stats.txt";
      $debug && print STDOUT ("$parse_overlapping_anno_cmd");
      system ($parse_overlapping_anno_cmd);

      my $pie_plot_cmd = "R --slave --args
$feature_bed"."_".$promoter_size."anno.stats.txt \"Tandem Repeat annotation\" <
$path/R/draw_annotation_pie.r ";
      $debug && print STDOUT ($pie_plot_cmd);
      system ($pie_plot_cmd);
}

# Overlapping gene features (for genic vs intergenic plot)
my $genic_anno_cmd = "intersectBed -a $feature_bed -b $gene_file -c | uniq >
$feature_bed"."_gene_anno.txt";
$debug && print STDOUT ("$genic_anno_cmd\n");
system ($genic_anno_cmd);

my $genic_anno_stat_cmd = "perl $path/perl/parse_gene_annotation.pl
$feature_bed"."_gene_anno.txt | sort | uniq -c | sort -nr >
/$feature_bed"."_gene_anno.stats.txt";
$debug && print STDOUT ("$genic_anno_stat_cmd\n");
system ($genic_anno_stat_cmd);

my $pie_gene_plot_cmd = "R --slave --args $feature_bed"."_gene_anno.stats.txt
\"Tandem Repeat annotation\" < $path/R/draw_annotation_pie.r ";
$debug && print STDOUT ("$pie_gene_plot_cmd\n");
system ($pie_gene_plot_cmd);

# Overlapping exon features (for exonic vs non-exonic plot)
my $exonic_anno_cmd = "intersectBed -a $feature_bed -b $exon_file -c | uniq >
$feature_bed"."_exonic_anno.txt";
$debug && print STDOUT ("$exonic_anno_cmd\n");
system ($exonic_anno_cmd);

my $exonic_anno_stat_cmd = "perl $path/perl/parse_exonic_annotation.pl
$feature_bed"."_exonic_anno.txt | sort | uniq -c | sort -nr >
/$feature_bed"."_exonic_anno.stats.txt";
$debug && print STDOUT ("$exonic_anno_stat_cmd\n");
system ($exonic_anno_stat_cmd);

my $pie_exonic_plot_cmd = "R --slave --args $feature_bed"."_exonic_anno.stats.txt
\"Tandem Repeat annotation\" < $path/R/draw_annotation_pie.r ";
$debug && print STDOUT ("$pie_exonic_plot_cmd\n");
system ($pie_exonic_plot_cmd);

# Computes the distance of each feature to the closest TSS.
my $closest_TSS_dist_cmd = "closestBed -D b -t first -a $feature_bed -b ".$tss_file."
| awk '{print \$NF}' > "."$feature_bed.tss.dist.txt";
$debug && print STDOUT ("$closest_TSS_dist_cmd\n");
system ($closest_TSS_dist_cmd);

# Generates a plot with the distance of each feature to the closest TSS.
my $closest_TSS_dist_plot_cmd = "R --slave --args $feature_bed.tss.dist.txt
\"Distribution of Tandem Repeats around TSS\" < $path/R/draw_TSS_linear_dist-
4kb_50b.r ";
system ($closest_TSS_dist_plot_cmd);
$debug && print STDOUT ("$closest_TSS_dist_plot_cmd\n");
```

```
my $closest_TSS_dist_plot_cmd2 = "R --slave --args $feature_bed.tss.dist.txt
\"Distribution of Tandem Repeats around TSS\" < $path/R/draw_TSS_linear_dist-
400b_10b.r ";
system ($closest_TSS_dist_plot_cmd2);
$debug && print STDOUT ("$closest_TSS_dist_plot_cmd2\n");
```

Chapter III

A comparative transcriptomic analysis of replicating and dormant liver stages of the relapsing malaria parasite *Plasmodium Cynomolgi*

Abstract

Plasmodium liver hypnozoites, which cause disease relapse, are widely considered to be the last barrier towards malaria eradication. The biology of this quiescent form of the parasite is poorly understood which hinders drug discovery. We report a comparative transcriptomic dataset of replicating liver schizonts and dormant hypnozoites of the relapsing parasite *Plasmodium cynomolgi*. Hypnozoites express only 34% of *Plasmodium* physiological pathways, while 91% are expressed in replicating schizonts. Few known malaria drug targets are expressed in quiescent parasites, but pathways involved in microbial dormancy, maintenance of genome integrity and ATP homeostasis were robustly expressed. Several transcripts encoding heavy metal transporters were expressed in hypnozoites and the copper chelator neocuproine was cidal to all liver stage parasites. This transcriptomic dataset is a valuable resource for the discovery of vaccines and effective treatments to combat vivax malaria.

Introduction

Plasmodium vivax (*P. vivax*) is the major cause of malaria outside of Africa with an estimated 13.8 million malaria cases globally in 2015 (World Health Organization (WHO), 2015). Among *P. vivax* parasites' most salient biological features are the persisting dormant liver stages (hypnozoites) that can cause relapse infections and compromise future eradication programs (Campo, Vandal, Wesche, & Burrows, 2015). Although *in vitro* hepatic cultures systems for hypnozoite-forming parasites have been developed (March et al., 2013; Zeeman et al., 2014) and rodent models of humanized liver stage infections constituted recent advances (Mikolajczak et al., 2015), the search for new drugs targeting hypnozoites is hampered by our limited knowledge of this enigmatic dormant stage.

Microbes commonly employ cellular quiescence to survive environmental stresses such as starvation, immune surveillance, or chemotherapeutic interventions and for disease causing microbes, dormancy often underlies chronic infections that considerably complicate the clinical management of infected patients (Rittershaus, Baek, & Sassetti, 2013). Cellular quiescence generally requires a physiological response underscored by a global repression of cellular metabolism but the preservation of mitochondrial respiration for ATP homeostasis and the maintenance of genome integrity (Rittershaus et al., 2013). Therapeutic interventions targeting some of these mechanisms have been proposed for a limited number of human pathogens (Andries et al., 2005; Rao, Alonso, Rand, Dick, & Pethe, 2008) but it is not clear whether *P. vivax* hypnozoites rely on similar physiological responses to survive in hepatocytes.

Some of the new drug targets that have been identified in the past decade (C. McNamara & Winzeler, 2011) have been shown to be critical in multiple stages of the parasite life cycle, such as PI4K (C. W. McNamara et al., 2013), DHODH (Phillips et al., 2015), eEF2 (Baragana et al., 2015), and pheT-RNA (Kato et al., 2016). However, none has yet been shown to be a valid target for malaria radical cure and elimination of the hypnozoite *in vivo*. Little is known about the expression pattern of these drug targets during *Plasmodium* life cycle in the liver and more specifically, it is not clear whether these genes are expressed at all in dormant parasites.

Transcriptomics approaches to assess genome-wide gene expression levels of *Plasmodium* liver stage parasites are inherently challenging given the low infection grade ratios and the higher abundance of host cell transcripts. While previous reports have emerged providing a first glance of gene expression in *Plasmodium* liver stages (Cubi et al., 2017; Vaughan et al., 2009), we provide here a comprehensive dataset derived from green fluorescent protein (GFP)-tagged *Plasmodium cynomolgi* (*P. cynomolgi*) (Voorberg-van der Wel et al., 2013) — the nonhuman primate sister taxon of *P. vivax,* known to form hypnozoites (Dembélé et al., 2014; Krotoski et al., 1982). We have collected samples from multiple independent *in vitro* hepatocyte infections, containing thousands of purified hypnozoites and liver schizonts for RNA-Seq. The sequenced reads were mapped on the new high quality, completely annotated *P. cynomolgi* genome covering 7,178 genes (Pasini et al., 2017). Using different approaches, we provide some preliminary validation of our comparative analysis of the transcriptome of replicating and quiescent liver-stages parasites, that will constitute a valuable resource for the development of *P. vivax* vaccines and therapeutics.

Material and methods

Ethics statement

Nonhuman primates were used because no other models (*in vitro* or *in vivo*) were suitable for the aims of this project. The local independent ethical committee constituted conform Dutch law (BPRC Dier Experimenten Commissie, DEC) approved the research protocol (agreement number DEC# 708) prior to the start and the experiments were all performed according to Dutch and European laws. The Council of the Association for Assessment and Accreditation of Laboratory Animal Care (AAALAC International) has awarded BPRC full accreditation. Thus, BPRC is fully compliant with the international demands on animal studies and welfare as set forth by the European Council Directive 2010/63/EU, and Convention ETS 123, including the revised Appendix A as well as the 'Standard for humane care and use of Laboratory Animals by Foreign institutions' identification number A5539-01, provided by the Department of Health and Human Services of the United States of America's National Institutes of Health (NIH) and Dutch implementing legislation. The rhesus monkeys (*Macaca mulatta*, either gender, age 4-7 years, Indian or mixed origin) used in this study were captive-bred and socially housed. Animal housing was according to international guidelines for nonhuman primate care and use. Besides their standard feeding regime, and drinking water ad libitum via an automatic watering system, the animals followed an environmental enrichment program in which, next to permanent and rotating non-food enrichment, an item of food-enrichment was offered to the macaques daily. All animals were monitored daily for health and discomfort. All intravenous injections and large blood collections were performed under ketamine sedation, and all efforts were made to minimize suffering. Liver lobes were collected from monkeys that were euthanized in the course of unrelated studies (ethically approved by the BPRC DEC) or euthanized for medical reasons, as assessed by a veterinarian. Therefore, none of the animals from which liver lobes were derived were specifically used for this work, according to the 3Rrule thereby reducing the numbers of animals used. Euthanasia was performed under ketamine sedation (10 mg/kg) and was induced by intracardiac injection of euthasol 20%, containing pentobarbital.

Transgenic *Plasmodium cynomolgi* sporozoite production

Blood stage infections were initiated in rhesus monkeys by intravenous injection of 1x10^6 *P. cynomolgi* M strain PcyC-PAC-GFP$_{hsp70}$-mCherry$_{ef1\alpha}$ (Voorberg-van der Wel et al., 2013) parasites from a cryopreserved stock. To exclude possible wild type contaminant parasites, monkeys were treated with pyrimethamine (1 mg/kg, orally on a biscuit every other day) for 3-4 times starting one day post infection. Parasitemia was monitored by Giemsa-stained smears prepared from a drop of blood obtained from thigh pricks. Animals were trained to voluntarily present for thigh pricks, and were rewarded afterwards. Around peak parasitemia, on two consecutive days, generally at days 11 and 12 post-infection, 9 ml of heparin blood was taken to feed mosquitoes and monkeys were cured from *Plasmodium* infection by intramuscular treatment with chloroquine (7.5 mg/kg) on three consecutive days. Typically ± 600 mosquitoes (two to five days old female *Anopheles stephensi* mosquitoes Sind-Kasur strain Nijmegen; Nijmegen UMC St. Radboud, Department of Medical Microbiology) were fed per blood sample using a glass feeder system. Mosquitoes were kept under standard conditions (Voorberg-van der Wel et al., 2013). Approximately one week after feeding, oocysts were counted and mosquitoes were given an uninfected blood meal to promote sporozoite invasion of the salivary glands. Mosquitoes that had received blood from the first bleeding ('feed 1') were kept separately and treated independently from mosquitoes that had received blood from the second bleeding ('feed 2').

Primary hepatocytes

Primary hepatocytes from *Macaca mulatta* or *Macaca fascicularis* were isolated freshly as described before or thawed from frozen stocks and resuspended in William's B medium (Zeeman et al., 2014): William's E with glutamax containing 10% human serum (A+), 1% MEM non-essential amino acids, 2% penicillin/streptomycin, 1% insulin/transferrin/selenium, 1% sodium pyruvate, 50 µM β-mercapto-ethanol, and 0.05 µM hydrocortisone. Hepatocytes were seeded into collagen coated (5 µg/cm^2 rat tail collagen I, Invitrogen) 6-well Costar plates at a concentration of approximately 2.25×10^6 cells/well. Following attachment, the medium was replaced by William’s B containing 2 % dimethylsulfoxide (DMSO) to prevent hepatocyte dedifferentiation.

Sporozoite isolation and inoculation

Two weeks post mosquito feeding on transgenic *P. cynomolgi* M strain infected blood, salivary gland sporozoites were isolated and used for hepatocyte inoculation. Prior to inoculation, hepatocytes were washed in William's B medium followed by sporozoite inoculation at ± 2 x 10^6 sporozoites per well. Plates were spun at RT at 500 g for 10-20 min and placed in a humidified 37 °C incubator at 5% CO_2 for 2-3 h to allow for sporozoite invasion. Medium (William's B) was replaced and incubation continued. Subsequently, infected hepatocytes were cultured with regular (every other day) medium changes until cell sorting. Sporozoite isolations and hepatocyte inoculations from 'feed 1' mosquitoes were performed separately from 'feed 2' mosquitoes.

Flow cytometry and cell sorting

At day 6 post sporozoite inoculation hepatocytes were harvested by Trypsin treatment (0.25% Trypsin-EDTA, Gibco). For logistical reasons, samples PAC22F1 and PAC22F2 were cultured for an additional day and were trypsinized at day 7 post-inoculation. Cells were washed once with PBS, followed by 3 min incubation in trypsin at 37 °C. Complete William's B medium was added to stop the trypsin digestion, cells were collected and washed two times in William's B medium that was diluted 1:5 in William's E to decrease the amount of serum in the samples. Prior to sorting, cells were passed through a 100 µM cell strainer to exclude clumps. First, a sample of uninfected hepatocytes was analysed to enable gate settings. Subsequently, infected hepatocytes were sorted with a BD FACSAria flowcytometer equipped with a 488 nm Coherent® Sapphire™ solid state 20 mW Laser. Data analyses were performed using FlowJo Version 9.4.10 (TreeStar, Inc., Ashland OR, USA). The device was equipped with a 100 µM nozzle for sorting. Gate settings were essentially the same as reported previously, except that an extra gate ('GFPdim') was included to ensure a strict separation of 'GFPlow' and 'GFPhigh' parasites (Figure 1B). Sorted samples were collected in 300 µl Trizol (*Life Technologies*). For the series of experiments relating to this paper we performed six blood stage infections. In two out of six blood stage infections, parasitemia was low (<0.2%) at the time of mosquito feeding. This resulted in poor sporozoite yields and not enough liver stage forms for FACSsort. The four other infections all resulted in successful liver stage infections with sufficient parasites for FACSsort, with one of the infections used for validation experiments. Collected 'GFPlow' samples contained 1,193-2,713 Hz (on average 1826 Hz); collected 'GFPhigh' samples contained 921-1,245 cells (on average 1,056 Sz). After sorting, tubes were vortexed ± 30 sec and transferred to a -80 °C freezer for storage until RNA extraction. During sorting, small amounts of GFPlow, GFPdim and GFPhigh samples were collected in William's B to analyze the quality of the sort: samples were transferred to a 96 well plate and analyzed using a high-throughput high-content imaging system (Operetta, Perkin-Elmer).

Neocuproine treatment

Following salivary gland dissection of infected *A. stephensi* mosquitoes 50,000 sporozoites were added per well to primary macaque hepatocyte cultures in 96-well plates as described earlier (Zeeman et al., 2014). Neocuproine (Sigma cat. 121908), dissolved in DMSO and subsequently diluted in William's B medium to 10, 1, and 0.1 µM was added in duplicate or triplicate wells to the cultures after sporozoite invasion and incubated with regular refreshments until fixation at day 6. Medium containing DMSO was used as control. Following methanol fixation immunofluorescence analysis was performed and parasites were counted using a high-content imaging system (Operetta; Perkin-Elmer) as reported previously (Zeeman et al., 2014).

Protein and antibody production

An *E. coli* codon optimized gene for PcyM_0533600 (Genscript, USA) was synthesized and protein (Q30-K145) was expressed in BL21 cells. The protein was purified using a Ni-IMAC column followed by gel-filtration/buffer exchange and used to immunize rats (Eurogentec, Belgium). In addition, monoclonal antibodies were raised against selected proteins at Genscript, USA.

Immunofluorescence analysis (IFA)

For IFA validation assays of hepatic stages, collagen coated Cell Carrier-96 well plates (Perkin Elmer) or Permanox Lab-Tek chamber slides (Nunc) were seeded with fresh primary rhesus hepatocytes and infected with wild type *P. cynomolgi* M sporozoites following procedures as described above and previously (Zeeman et al., 2014). For long-term culture, to enable IFA analysis of day 19 liver stage parasites, matrigel was placed on top of the hepatocytes as previously described (Dembélé et al., 2014). At day 6/7 (or day 19) post sporozoite inoculation, cells were briefly fixed in cold methanol followed by three washes in PBS. Infected hepatocytes were blocked in 100 mM glycine for 5 min. at room temperature. After three washes with PBS, cells were incubated for 1-2 h at room temperature with hybridoma supernatant (undiluted), polyclonal antiserum (1:100) or purified IgG (25 µg/ml) diluted in PBS. Primary antibodies were mouse mAb anti-H4K8ac (Active motif, #61525, 1:500 dilution), polyclonal rat-anti-ETRAMP (PcyM_0533600, Eurogentec), mouse mAb 1G4E7 against GAP45 (PcyM_1442700, Genscript) and mouse mAb 5B10C7 against Ferredoxin (PcyM_1419800, Genscript). Anti-*P. cynomolgi* HSP70.1 polyclonal rabbit serum (Zeeman et al., 2014) was included to detect parasites. Cells were washed three times in PBS and incubated 1-2 h at room temperature with secondary antibodies diluted in PBS with DAPI. Fluorescein isothiocyanate (FITC)-labeled goat anti-rabbit IgG (Kirkegaard and Perry Laboratories, 1:200), FITC-labeled goat anti-mouse IgG (Kirkegaard and Perry Laboratories, 1:200), Alexa-594 labeled chicken anti-mouse IgG (Invitrogen, 1:2000), or Alexa-594 labeled chicken anti-rabbit IgG (Invitrogen, 1:2000) were used as secondary antibodies. Following three washes with PBS, mounting was performed with CITIFLUOR AF1 (Agar Scientific). Images were taken using a Nikon Microphot FXA fluorescence microscope equipped with a DS-5M digital camera or with a Leica DMI6000B inverted fluorescence microscope equipped with a DFC365FX camera. For IFA staining of blood stage parasites, blood smear preparations of *P. cynomolgi* infected red blood cells were fixed with methanol. Primary and secondary antibodies were diluted in 1% FCS/PBS and each staining was for 1 h at room temperature. Primary antibodies were polyclonal rat-anti-ETRAMP (PcyM_0533600) at 25 µg/ml and rabbit anti-Band 3 monoclonal antibody (Abcam ab108414, 1:100). Secondary antibodies were Alexa-594 labeled chicken anti-rabbit IgG (Invitrogen, 1:2000) and mouse serum adsorbed FITC-labeled goat anti-rat IgG (Kirkegaard and Perry Laboratories, 1:200); DAPI was included. Slides were rinsed in PBS (4-5x) and mounted with CITIFLUOR AF1. Images were taken using a Leica DMI6000B inverted fluorescence microscope equipped with a DFC365FX camera.

RNAscope in situ hybridization

P. cynomolgi M infected primary rhesus hepatocytes cultured for 6 days in CellCarrier-96 well plates (Perkin-Elmer) were fixed for 30 min. at RT in 4% paraformaldehyde in PBS (Affymetrix), dehydrated and stored at -20°C until further processing. RNA in situ detection was performed using the RNAscope Multiplex Kit (Advanced Cell Diagnostics) according to the manufacturer's instructions. RNAscope probes used were: *gapdh* (PcyM_1250000, region 113-997) and *hsp70*

(PcyM_0515400, region 606-1837). Following the RNA-FISH protocol, IFA was performed using rabbit anti-PcyHSP70 to stain the parasites as described above. Z-Stack images were acquired on the Operetta system (Perkin-Elmer) using a 40x objective NA 0.95 and maximum projections are shown.

RNA sequencing

Total RNA was isolated from 5 different samples of FACS-sorted small parasite infected cells (GFP-low, *e.g.* hypnozoites) and 5 samples of FACS-sorted large parasite infected cells (GFP-high, *e.g.* liver schizonts). All samples were stored in TRIzol (Thermo Fisher) and total RNA extracted using the Direct-zol™ RNA MiniPrep Kit (Zymo Research) including on-column DNase digestion according to the manufacturer's instructions. RNA amplification was performed using the TargetAmp™ 2-Round aRNA Amplification Kit 2.0 (Epicentre). The quality of the RNA samples (before and after the amplification) was assessed with the RNA 6000 Pico and Nano kits using the Bioanalyzer 2100 instrument (Agilent Technologies). RNA-seq cDNA libraries were prepared from the amplified RNA samples using the TruSeq mRNA Sample Prep kit v2 (Illumina). The quality of the cDNA libraries was assessed with the Bioanalyzer 1000 DNA kit (Agilent Technologies). RNA-seq cDNA libraries were then sequenced in paired-end mode, 2 x 76 bp, using the Illumina HiSeq2500 platform. Read quality was assessed by running FastQC (version 0.10) on the FASTQ files. Sequencing reads showed high quality, with a mean Phred score higher than 30 for all base positions. Over 857 million 76-base-pair (bp) paired-end reads were used for the bioinformatics analysis. Reads from each sample were aligned to a genomic reference composed of the combination of the malaria parasite *Plasmodium cynomolgi* M strain genome and one of the following host genomes: *Macaca mulatta* (Zimin et al., 2014) (http://www.unmc.edu/rhesusgenechip/), and *Macaca fascicularis* (http://www.ncbi.nlm.nih.gov/assembly/GCF_000364345.1/). Reads mapping to the parasite genome was used to quantify gene expression by using the Exon Quantification Pipeline (EQP) (Schuierer & Roma, 2016). On average, a range of 38% (minimum) to 84% (maximum) of total reads were mapped to the parasite and host genomes, and between 17% and 65% were aligned to the parasite and host exons (expressed reads). A QC inspection of the aligned sequencing reads showed an expected coverage bias towards the 3' end of the transcripts that is due to the use of the amplification kit. Based on the alignment statistics, we decided to exclude two Sz samples and one Hz sample from further analyses. Genome and transcript alignments were used to calculate gene counts based on the *P. cynomolgi* M strain gene annotation (Pcynom M_v2, Pasini et al., 2017) provided by the BPRC and the Wellcome Trust Sanger Institute.

Gene raw counts represent the total number of reads aligned to each gene. These values were normalized using the following four-stage approach (Figure 1-figure supplement 3). First, gene raw counts were divided by the total number of mapped reads for each sample and multiplied by one million to obtain Counts Per Million (CPM) to account for varying library sizes (library size normalization). In a given sample, one CPM indicates that a specific gene was detected by one read out of one million of mapped reads. Second, a further normalization of the CPMs based on the BioConductor package DESeq2 (Love, Huber, & Anders, 2014) was performed for the samples of each stage separately to account for the variation of #parasite-cells/#host-cells fraction within one stage (group-wise normalization). Third, an adjustment of mean expression ratio between schizont and hypnozoite samples was computed by using host expression values to further account for the difference in cell size and RNA amount per cell which is expected between the schizont and the hypnozoite liver forms (host normalization). The host normalized counts were further divided by the gene length in kb to obtain the Fragments Per Kilobase per Million values (FPKM) (gene length normalization). The host normalized gene expression values were also used to identify differences in gene expression between the schizont and the hypnozoite samples using the BioConductor package DESeq2 (Love et al., 2014). We therefore calculated the list of genes that are differentially expressed between the liver schizonts and the hypnozoites along with the log2 fold changes and *p*-values after Benjamini-Hochberg false discovery rate (FDR) correction for multiple hypothesis testing.

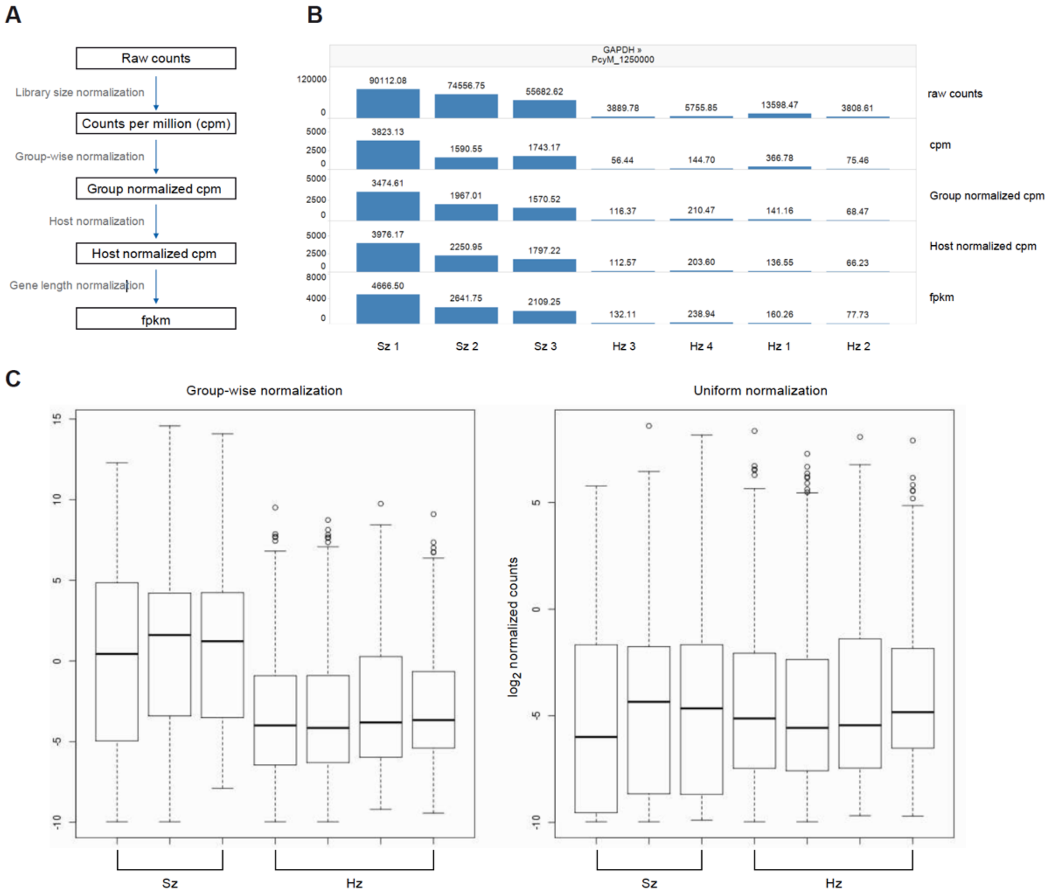


Figure 1-figure supplement 3. Normalization of gene expression values. (A) Overview of the normalization process from raw counts to FPKMs. This process comprises four steps which are library size normalization, group-wise normalization, host normalization, and gene length normalization. See Methods for a description of each step. **(B)** Effect of the normalization on gene expression levels. GAPDH is shown as an example. **(C)** Comparison of normalization strategies. Left: group-wise normalization is used to keep the expected difference in absolute level of gene expression between schizonts and hypnozoites. Right: uniform normalization (as applied by Cubi *et al.*) brings the distribution of the expression values of the hypnozoite and schizont samples onto equal levels.

Orthology and pathway analysis

In order to annotate the *Plasmodium cynomolgi* proteome, we performed an extensive orthology analysis that included the following proteomes in addition to *P. cynomolgi* M strain: *P. falciparum 3D7*, *P. berghei ANKA*, *P. knowlesi H*, *P. vivax Sal1*, *P. yoelii yoelii* 17X, *H. sapiens*, *D. melanogaster*, *M. musculus*, *R. norvegicus*, and *S. cerevisiae*. The *Plasmodia* proteomes were obtained from PlasmoDB (http://PlasmoDB.org/) version 26 (Aurrecoechea et al., 2009), the other proteomes from UniProt (release 2015_12) (Bateman et al., 2015). Our orthology analysis is based on the OrthoMCL methodology but implemented in-house to work with our local high-performance computing environment. Conceptually, this comprised the following steps: 1) alignment of all protein sequences against each other with blastp (Altschul, Gish, Miller, Myers, & Lipman, 1990); 2) calculation of the percent match length by determining all amino acids participating in any HSP between two proteins divided by the length of the shorter protein; 3) filtering out of the blast results with a percent match length below 50% or an E-value above 10^{-5}; 4) determination of potential orthologs and paralogs and their normalized E-values; 5) clustering of the resulting weighted similarity graph with MCL. See Fischer 2011 *et al* (Fischer et al., 2012) for more details, and Figure 6.12.1 contained within for an overview. The obtained groups of proteins were used to propagate protein annotations from other species to *P. cynomolgi*. Using this approach, we were able to group a total of 6,040 *P. cynomolgi* proteins (86% of the total 7,030 proteins) with at least one protein from another species, and 2295 (33%) *P. cynomolgi* proteins were linked to 257 malaria pathways mapped from PlasmoDB. For the identification of pathways that are expressed in the liver stages, we used a stringent cut-off to focus only on those genes whose expression is consistent across replicates (>1 FPKM in at least 2 replicates). This resulted into 2,748 genes and 88 pathways expressed in 2/4 Hz replicates, and 5,323 genes and 233 pathways expressed in 2/3 Sz replicates.

Pathway and Gene Ontology enrichment analyses

Gene sets were collected from two sources: PlasmoDB (Aurrecoechea et al., 2009) and Gene Ontology (Ashburner et al., 2000; The Gene Ontology Consortium, 2017). The gene sets from PlasmoDB mostly correspond to "Metabolic pathways", whereas the gene sets from the Gene Ontology correspond to general organizational principles of biology (such as "translation"). Many of the pathways from PlasmoDB are manually curated, whereas large parts of the annotations in the Gene Ontology are derived and propagated from one species to another by algorithms. The gene sets were mapped by orthology to *Plasmodium cynomolgi*. We employed two standard approaches to determine the relevance of gene sets with respect to our RNAseq data: 1) overrepresentation analysis *via* a hypergeometric test; and 2) Kolmogorov-Smirnov test, as proposed in the original GSEA publication (Subramanian et al., 2005). The main differences between the two approaches is that the first one requires a predetermined criterion to select genes of interest in which overrepresented annotations are to be determined; the second does not need any such cut-off, as the test statistic is based on a ranking of all genes in the experiment.

For the enrichment analyses, we applied several criteria of increasing stringency to select stage-specific genes of interest from our RNA-seq experiment:

- All genes within a certain stage that are expressed with at least 1 FPKM in at least 2 samples in that stage (*e.g.* Hz or Sz).
- All genes within a certain stage that are expressed with at least P_{25} FPKM in at least 2 samples in that stage, where P_{25} is the 25th percentile (1st quartile) of the expression of the pooled samples of that stage (*e.g.*_Hz_q1 or Sz_q1).
- All genes within a certain stage that are expressed with at least P_{75} FPKM in at least 2 samples in that stage, where P_{75} is the 75th percentile (3rd quartile) of the expression of the pooled samples of that stage (*e.g.* Hz_q3 or Sz_q3);
- All genes that satisfy criterion 2 in a stage but in no other stage (*e.g.* Hz_q1_specific or Sz_q1_specific).
- All genes that satisfy criterion 3 in a stage but in no other stage (*e.g.* Hz_q3_specific or Sz_q3_specific).

Genes satisfying the criteria above were determined for all stages and used as input for an overrepresentation analysis.

Targeted amplification and sequencing of the *etramp* gene

Blood stage, sporozoite, schizont and hypnozoite RNA samples were reverse transcribed using the High Capacity RNA-to-cDNA Kit (#4368814, Thermo Scientific). The *etramp* gene (PcyM_0533600) was amplified in all the samples using the Phusion DNA Polymerase kit (#F530, Thermo Scientific) with the following primers: ACTCCTTGGTGGTGCCTTAG (FWD); TGCGGGGCCCTTATCTTT (REV). The Ovation Low complexity Sequencing System kit (#9092-256, NuGEN) was used to prepare the sequencing libraries. Libraries were multiplexed and sequenced in paired-end mode, at a read length of 2 × 300 bp, using the MiSeq platform (Illumina). The resulting FASTQ files were demultiplexed and aligned against the *P. cynomolgi* M strain genome (Pcynom M_v2, unpublished) using STAR version 2.5.2a (Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S & Chaisson M, 2013) for the detection of the amplified regions. The Integrative Genomics Viewer (IGV) (James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, 2011) version 2.3.75 was used to visualize the aligned reads in the genome context. The etramp gene view presented in Figure 2D was generated using the R/Bioconductor GViz package (Hahne & Ivanek, 2016).

Comparison with published data

Published expression data of *P. cynomolgi* liver stages (Cubi et al., 2017) were downloaded from the EMBL-EBI European Nucleotide Archive [ENA: PRJEB18141; Sample group: ERS1461774] and compared to our RNA-seq data. It was not possible to compare the gene lists from Cubi *et al.* directly with the genes from this manuscript because the two studies used different *P. cynolmolgi* reference genomes and gene annotation files. The downloaded Fastq files of Cubi *et al.* were thus processed with the genome reference and annotation files from (Pasini et al., 2017), the same RNA-seq analysis pipeline, and the same normalization method as was used for our data set and which is described in the 'RNA sequencing' paragraph above. The correlation plots were generated on log10 normalized CPMs after the addition of a pseudo-count of 0.1. The Venn diagram plots were generated on genes expressed above the cut-off of 1 FPKM in the hypnozoite and schizont samples, and drawn using the on line tool available at the following website: http://bioinformatics.psb.ugent.be/webtools/Venn/.

Results

Hypnozoites express a smaller set of genes than schizonts

Six to seven days after *P. cynomolgi* sporozoite infection of primary simian hepatocytes, we FACS-purified hepatocytes containing hypnozoites and liver schizonts and prepared RNA for high-throughput sequencing (Figure 1). After quality control, we excluded 3 samples due to their low number of parasite reads, resulting in a dataset containing 3 independent schizont samples and 4 independent hypnozoite samples for analyses. To quantify parasite-specific expression for each *P. cynomolgi* gene, we determined the number of sequencing reads aligned to genes and computed gene expression values as the number of Fragments Per Kilobase per Million fragments mapped (FPKM) (Schuierer & Roma, 2016). Overall, the raw gene expression values of the schizont samples are ~14-fold higher than those of the hypnozoite samples (*p*-value 1.1e-3). This global difference in gene expression between multi-nucleated schizonts and uni-nuclear hypnozoites could be partly attributed to differences in the number of parasite transcriptionally active units per hepatocyte, however it is not possible to determine this exact number. In order to account for this difference, we normalized the gene expression values against the total number of host reads per sample which we posit to represent a constant host RNA content across all samples (see Methods). All data reported in Figures 1 to 4 show FPKM values after such normalization. A threshold of FPKM greater than 1 is deemed equivalent to one transcript copy per cell (Mortazavi, Williams, Mccue, Schaeffer, &

Wold, 2008). Using this threshold, hypnozoites generally express a lower number of genes compared to schizonts (respectively, 3,308 vs 5,702 genes at average FPKM per group ≥1). In addition, the expression level of these genes in schizonts is higher than in hypnozoites (average expression 89.14 and 9.88 FPKM, respectively) (Figure 1C). To further validate this finding we carried out RNA fluorescence *in situ* hybridization (RNA-FISH) to quantitatively evaluate the expression of abundantly expressed genes at the single-cell level in liver stage cultures. In agreement with the RNA-Seq results, the RNA-FISH staining with probes for *gapdh* and *hsp70* showed a markedly lower level in hypnozoites compared to schizonts' expression (Figure 1D). We then compared the gene expression data with those recently published by Cubi *et al.* (Cubi et al., 2017). Since the two studies used different reference genomes and annotation files, we reprocessed the raw sequencing files using the *P. cynomolgi* genome from Pasini *et al.* (Pasini et al., 2017) and the data analysis methods that we used in this manuscript. The schizonts data from Cubi *et al*. showed a high correlation of 0.95 and a large consensus between the two replicates, which compared with a slightly lower but equally high correlation (average correlation 0.88) and high overlap between the triplicates profiled in this study (Figure 1-figure supplement 1). In stark contrast, while gene expression data reported here showed a high concordance between the four biological replicates of hypnozoites (average correlation 0.68) and a large overlap of 2,804 out of 4,198 genes expressed in at least two samples (Figure 1-figure supplement 1), the data from Cubi *et al.* showed a lower correlation for the two hypnozoite samples (correlation of 0.38; Figure 1-figure supplement 1) and a scarse consensus between the two replicates (204 out of 1,147 genes; Figure 1-figure supplement 1). Of these 204 genes, 175 overlap with at least one of our hypnozoite samples (Figure 1-figure supplement 1), and can thus be considered true hypnozoite transcripts. Compared to the here reported 2,804 hypnozoite transcripts, this indicates that many genes and pathways expressed in hypnozoites were not captured in the previous study (Cubi et al., 2017).

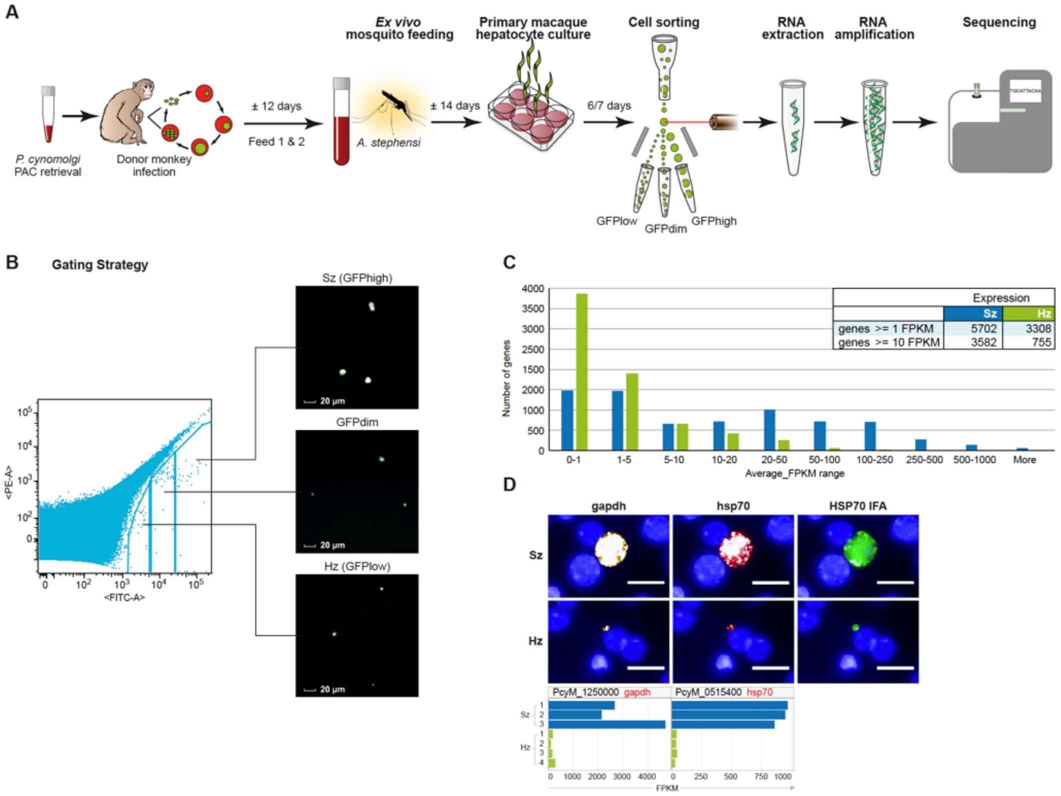


Figure 1. Transcriptomics of relapsing malaria liver stage parasites. (A) Scheme of experimental approach for purification and RNA-sequencing of cultured *P. cynomolgi* M malaria liver stage schizonts (Sz) and hypnozoites (Hz). To enable FACS purification, *P. cynomolgi* parasites that stably express GFP using a Plasmodium Artificial Chromosome (PAC) were used. For further details, see methods. (**B)** Gating strategy included an extra gate, 'GFPdim', not used in subsequent RNA-seq analysis to ensure a strict separation of 'GFPlow' and 'GFPhigh' parasites. (**C**) Distribution of average gene expression values in the hypnozoite (green; n=4) and schizont (blue; n=3) samples. FPKM, Fragments per kilobase of transcript per million mapped reads. (**D)** Top panel showing RNA fluorescence in situ hybridization (RNA-FISH) of day 6 *P. cynomolgi* Sz and Hz with probes against *gapdh* (PcyM_1250000) and *hsp70* (PcyM_0515400). Scale bars, 20 µm. Lower panel shows gene expression values (FPKM) for *gapdh* and *hsp70* of individual Hz and Sz samples as determined by RNA-sequencing

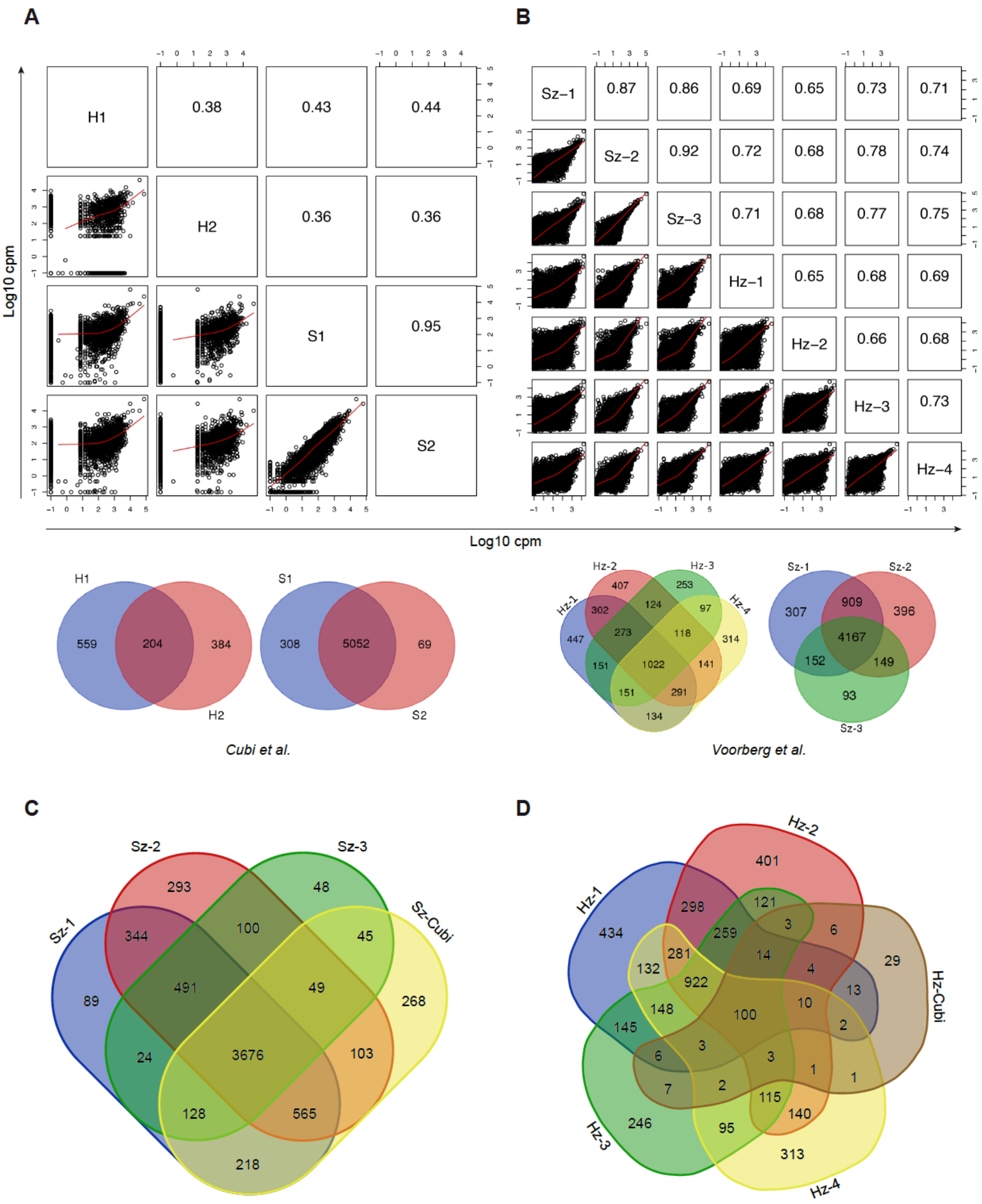


Figure 1-figure supplement 1. Comparison with published data. (A) Scatter plot showing all pairwise $\log_{10}$ normalized CPM correlations between samples from Cubi *et al*. The upper right part of the panel shows the value of the calculated Pearson correlation coefficients. The Venn diagrams show the overlap of genes expressed above 1 FPKM in the hypnozoite samples and in the schizont samples, respectively. **(B)** Same as A, but showing data from this study. **(C)** Venn diagrams showing the overlap of genes expressed above 1 FPKM in the three schizont samples from this study and the 5,502 genes shared by the two schizont samples from Cubi *et al*. (Sz-Cubi). **(D)** Venn diagrams showing the overlap of genes expressed above 1 FPKM in the four hypnozoite samples from this study and the 204 genes shared by the two hypnozoite samples from Cubi *et al*. (Hz-Cubi).

Although the transcription level in hypnozoites appears to be generally reduced, we found evidence that there is ongoing active transcription in hypnozoites, up to day 7, as demonstrated by the positive staining with antibodies recognizing the acetylated H4K8 protein, a marker of open chromatin (Gupta et al., 2013) (Figure 1-figure supplement 2). Thus, when compared to proliferating liver schizonts, "dormant" hypnozoites express only less than half of the parasite genome and the rate of transcription of individual genes also appears to be very low.

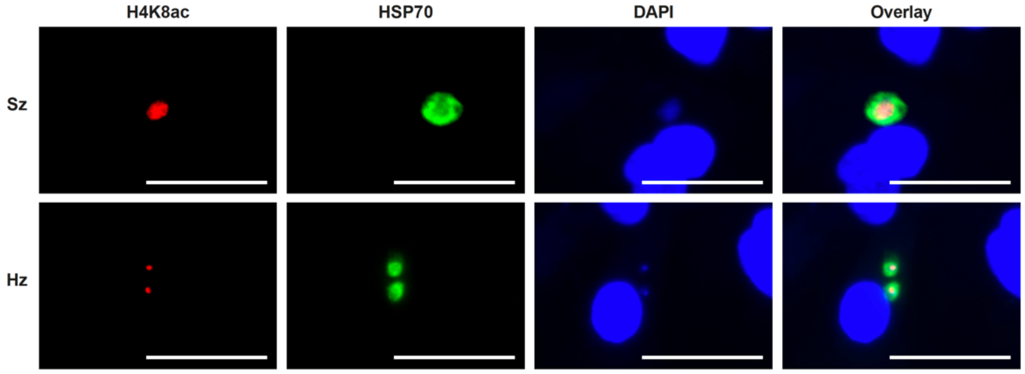


Figure 1-figure supplement 2. IFA staining of acetylated H4K8 in *P. cynomolgi* liver stages. Immunofluorescence analysis of day 7 *P. cynomolgi* liver stage schizont (upper panel) and two hypnozoites (lower panel). Scale bar 25 µm.

Comparative transcriptomic analysis allows the identification of differential markers of *P. cynomolgi* liver stages

We further explored the liver stage transcriptomes to identify those genes with significantly different expression levels between hypnozoites and schizonts (>2 fold-change absolute value, 10% false discovery rate (FDR)) (Figures 2A and 2B). Our results indicate that the expression of only a dozen genes might be enhanced in quiescent hypnozoites as compared to growing liver schizonts, while the expression of thousands of genes is significantly lower in hypnozoites than in schizonts. To determine whether protein expression follows the RNA differential expression observed, we selected a few genes that were upregulated in either stage and raised antibodies against recombinant predicted proteins. Using these antibodies, we then performed immunofluorescence analysis (IFA) on cultured liver stages. Unexpectedly, antibodies against PcyM_0533600 (ETRAMP, amino acids Q30-K145), one of the most up-regulated genes in the hypnozoite samples, failed to detect the protein in day 6 *P. cynomolgi* liver stage parasites. The same antibodies strongly reacted with *P. cynomolgi* blood parasites (Figure 2C) but failed to detect the protein in sporozoites (data not shown). Sequence analysis of RT-PCR products from different parasite stages revealed that only blood stage parasites express the predicted full-length PcyM_0533600 mRNA, while alternatively spliced transcripts (including premature stop codons) were found in sporozoite, schizont, and hypnozoite samples (Figure 2D), explaining our inability to detect the predicted protein in these stages.

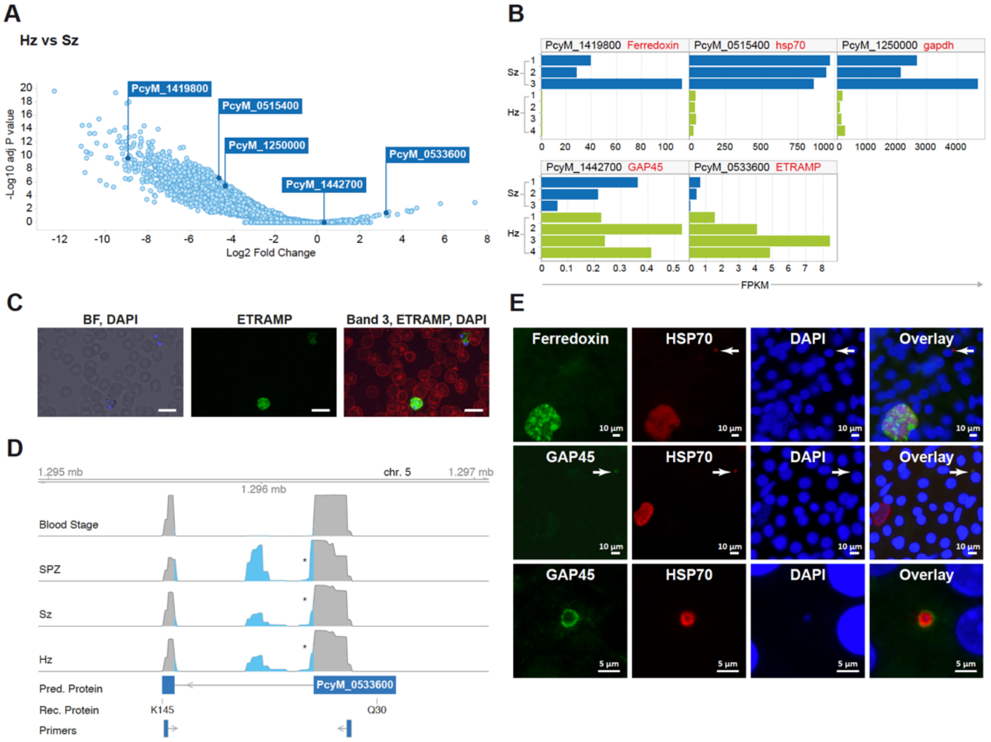


Figure 2. Relapsing malaria liver stages display low transcription levels and differ from developing stages. (A) Volcano plot showing genes differentially expressed in hypnozoites (Hz, n =4 biological replicates) versus schizonts (Sz, n=3 biological replicates). The *y*-axis represents the significance as –log10 FDR-adjusted *p*-values and the *x*-axis represents the expression changes as log2 fold-change of Hz and Sz. Genes used for validation are marked. **(B)** Gene expression values (FPKM) of individual Hz and Sz samples from genes selected for validation. **(C)** Immunofluorescent staining of ETRAMP protein (green), DAPI (blue) and red blood cell (red) in *P. cynomolgi* blood stage parasites. Scale bars 25 µm. **(D)** Genome browser view of the *etramp* gene (PcyM_0533600) showing intron splicing events detected by sequencing of RT-PCR products in Blood stages, Sporozoites (SPZ), Schizonts (Sz) and Hypnozoites (Hz). Retained intron events are highlighted in blue; asterisk shows premature termination codons (PMTs). The predicted protein (Pred. Protein), the recombinant portion of the protein (Rec. Protein) used for antibody production (amino acids Q30-K145), and the positions of the primers used to generate the RT-PCR products are shown. **(E)** Immunofluorescent staining patterns of Ferredoxin (PcyM_1419800), GAP45 (PcyM_1442700), and HSP70 (PcyM_0515400) in day 6 *P. cynomolgi* liver schizonts and hypnozoites. Arrows, hypnozoites. Lower panel shows magnified image of GAP45 stained hypnozoite.

To further validate our dataset, antibodies were raised against three other proteins, PcyM_0515400 (HSP70), PcyM_1419800 (Ferredoxin) (Miotto et al., 2015) and PcyM_1442700 (Glideosome Associated Protein GAP45). These antibodies did react with *P. cynomolgi* day 6 liver stage parasites showing staining of liver schizonts (Ferredoxin), primarily hypnozoites (GAP45) or both schizonts and hypnozoites (HSP70), mirroring precisely the RNA-seq data for these genes (Figure 2E). Antibodies against GAP45, an inner membrane complex (IMC) marker (Kono et al., 2012) and a member of the glideosome motor complex (Harding & Meissner, 2014), stained the periphery of 6-days old hypnozoites (Figure 2E middle and lower panels**)**. In contrast, the staining pattern in schizonts was weaker and sparsely distributed (early schizonts) or absent (large mature schizonts) (Figure 2E middle panel). These data concur with previous reports describing the progressive loss of the IMC during conversion of the motile sporozoite into a replication-competent metabolically active liver stage form (Jayabalasingham, Bano, & Coppens, 2010). Interestingly, we could still detect GAP45 in hypnozoites at day 19 (Figure 2-figure supplement 1). The long-term presence of GAP45 may be due to low protein turnover in hypnozoites or a functional requirement of this protein for hypnozoite maintenance. Taken together the RNA-FISH and immunofluorescence experiments confirmed the general trends we observed in the RNA-seq dataset and we anticipate that further mining of this gene list will yield differential markers of schizont development and hypnozoite maintenance.

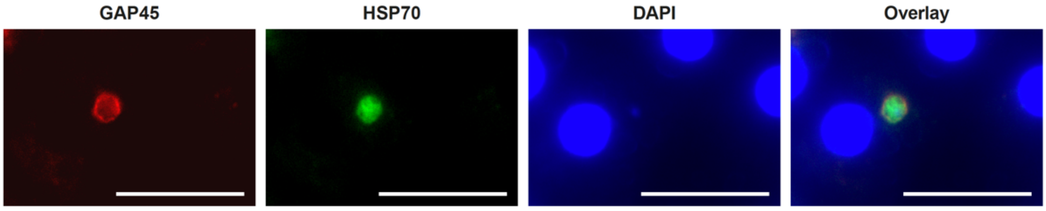


Figure 2-figure supplement 1. GAP45 protein expression in day 19 hypnozoite. Staining of GAP45 protein (PcyM_1442700) in a 19-day *P. cynomolgi* hypnozoite parasite by immunofluorescence analysis (IFA). Scale bar 25 µm.

Hypnozoites express few core pathways including the physiological hallmarks of dormancy

To investigate the physiology of *P. cynomolgi* liver stages, we performed a pathway analysis in schizonts and hypnozoites. Through orthology mapping, *P. cynomolgi* genes were assigned to 257 *Plasmodium falciparum* pathways (Aurrecoechea et al., 2009). Gene ontology and pathway enrichment analyses highlighted that hypnozoites express genes related to translation, RNA processing and epigenetic processes (*e.g.* histone acetylation and methylation). These pathways and processes were also enriched in the schizonts, which however expressed more processes related to the cell nucleus, hinting at the differences in transcriptional activity. Schizonts clearly express a much higher number of pathways than hypnozoites (Figure 3A). Of all the pathways included in this analysis, only ~34% (88 out of 257 pathways) express more than half of their constituent genes above the threshold of 1 FPKM in the hypnozoite while the equivalent is true for ~91% (233 out of 257) of the pathways in schizonts. In the schizonts, energy and glucose metabolism pathways, such as pentose phosphate cycle enzymes, CoA biosynthesis pathways and mannose/fructose metabolism are all highly expressed with nearly all genes in those pathways detected above 1 FPKM (Figure 3A). In contrast, those pathways are nearly absent in the hypnozoite, which is consistent with the quiescence and low metabolism that may be expected in dormant forms.

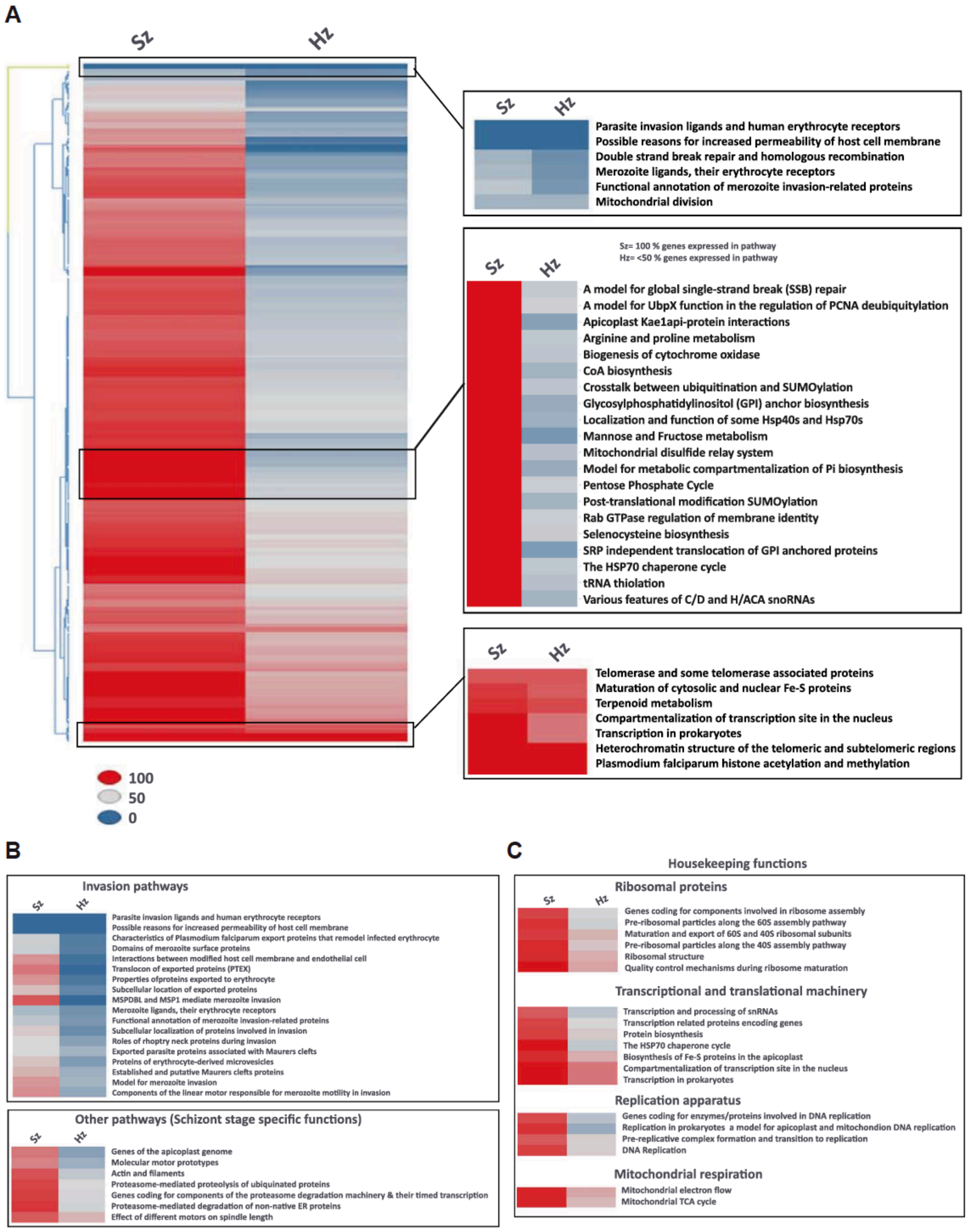


Figure 3. Pathway analysis of the malaria liver stages reveals the core biological functions required for hypnozoites maintenance. (A) Heat map showing expression of *Plasmodium* pathways in schizonts and hypnozoites. A total of 257 pathways annotated in *P. falciparum* were assigned to *P. cynomolgi* through orthology (see methods). Pathways where the fraction of genes detected above the threshold of FPKM of 1 is 100% are shown in red, between 50% and 100% in grey, between 0% and 50% in blue. **(B)** Same as a) but showing only erythrocytic invasion and schizont specific pathways. **(C)** Same as a) but showing house-keeping pathways.

Interestingly, some but not all erythrocytic invasion pathways are expressed only in schizonts, suggesting that already at day 6 the parasites express some of the genes required for merozoite function and red blood cell invasion (Figure 3B). Hypnozoites mostly express core housekeeping pathways such as those involved with nucleus and chromatin maintenance, transcription, translation and mitochondrial respiration, but no DNA replication enzymes (Figure 3C, Figure 3-figure supplement 1). Notably, genes known to be required for ATP homeostasis in non-replicating dormant *Mycobacterium tuberculosis* (Rao et al., 2008), such as various components of the F0-F1 ATPase complex, are similarly significantly expressed in hypnozoites (Figure 3-figure supplement 1). Collectively, our analyses reveal that hypnozoites express pathways previously associated with quiescence and required for the maintenance of chromosome integrity and ATP homeostasis.

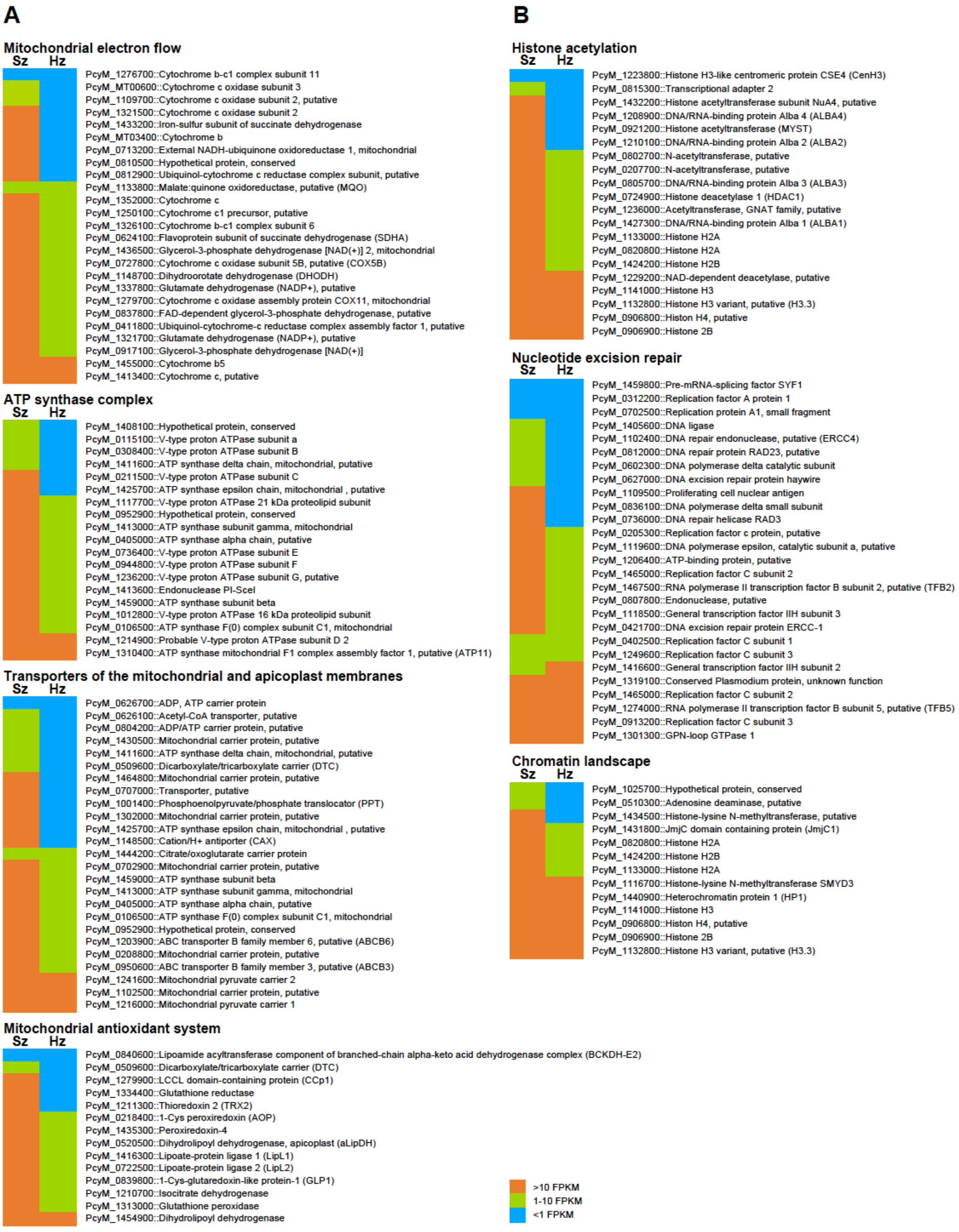


Figure 3-figure supplement 1. Liver stage schizont (Sz) and hypnozoite (Hz) gene expression values (FPKM) for pathways associated with quiescence. (**A**) Pathways involved in maintenance of membrane potential and ATP biosynthesis. (**B**) Pathways involved in preservation of genome integrity.

Expression pattern of potential drug targets in *P. cynomolgi* liver stages

In the liver schizonts and hypnozoites transcriptomic dataset, we looked at the expression FPKM values for clinically and chemically validated drug targets (reported in Figure 4A). While all drug targets are expressed in the schizonts above the threshold level of 1 FPKM, only a few of them are detectable above this level in the hypnozoites. For example, we could not detect PI4K transcripts in day 6 hypnozoites while this gene is abundantly expressed in schizonts (Figure 4A), which is consistent with previously published data on *Plasmodium* PI4K inhibitors having prophylactic but not radical curative activity in the *P. cynomolgi* model (Zeeman et al., 2016). In contrast, the antifolate drug target DHFR is detectable above 1 FPKM in hypnozoites, yet antifolates do not exhibit radical cure in the *P. cynomolgi* model (Schmidt, Fradkin, Genther, Rossan, & Squires, 1982). Likewise, DHODH, the target of the clinical candidate DSM265, is detectable in hypnozoites while this compound shows poor activity against hypnozoites *in vitro* (Phillips et al., 2015). Although the *P. cynomolgi* ATP4 ortholog, the clinically validated target of KAE609, is detectable in schizonts at low level, it is not critical as PfATP4 inhibitors are not active in liver stages (Jiménez-díaz et al., 2014; Rottmann et al., 2010; Vaidya et al., 2014). Thus, it appears that function could not be directly inferred from the liver stages expression data.

Plasmodium parasite survival and replication depends on the import of nutrients and solutes from its host cell and some transporters have been proposed to be tractable drug targets for malaria (Hapuarachchi et al., 2017; Pain et al., 2016; Slavic, Krishna, Derbyshire, & Staines, 2011; Weiner & Kooij, 2016). Consistently, we observe high FPKM values for a broad range of transporters in both liver schizonts (35 putative transporters with FPKM values >10) and hypnozoites (7 transporters with FPKM values >10 and 25 transporters with FPKM values >1. Heavy metal homeostasis has been shown to be critical to liver stage (Kenthirapalan, Waters, Matuschewski, & Kooij, 2016; Sahu et al., 2014; Stahel et al., 1988) development and consistently we found several heavy metal transporters to be expressed in all liver stages (Figure 4B). Remarkably, two putative copper transporters (PcyM_1331900 and PcyM_1277100) showed high FPKM values for both liver stage schizonts and hypnozoites (Figure 4B), suggesting a role for copper homeostasis in liver stage development and quiescence. To determine whether copper was critical to *P. cynomolgi* liver stages, we treated infected hepatocytes with a copper chelator, neocuproine (Choveaux, Przyborski, & Goldring, 2012; Kenthirapalan, Waters, Matuschewski, & Kooij, 2014). Neocuproine treatment, initiated 1-2 hours after infection with sporozoites and continued for 6 days, indeed showed pronounced cidal effects on the viability of both liver schizonts and hypnozoites (Figure 4C) at the highest concentration tested. In one of the three assays we noted a limited effect on hepatocyte viability at this concentration, as concluded from from hepatocyte nuclei counts in the analysis. These data provide some preliminary chemical validation of the hypothesis that copper homeostasis may be critical for schizonts replication and hypnozoites survival.

A

Gene ID	Target	Description	Compound	Sz (avg FPKM)	Hz (avg FPKM)
PcyM_1471000	FKBP35	FK506-binding protein (FKBP)-type peptidyl-prolyl isomerase	D44	140	19
PcyM_1337500	KRS	Lysine--tRNA ligase (apicoplast)	Cladosporin	98	17
PcyM_1245400	DXR	1-deoxy-D-xylulose 5-phosphate reductoisomerase	Fosmidomycin	211	12
PcyM_1264200	eEF2	elongation factor 2	DDD107498	124	5
PcyM_0526900	DHFR-TS	bifunctional dihydrofolate reductase-thymidylate synthase	Pyrimethamine	118	4
PcyM_1148700	DHODH	dihydroorotate dehydrogenase, mitochondrial precursor	DSM265	113	2
PcyM_1430900	DHPS	hydroxymethylpterin pyrophosphokinase-dihydropteroate synthetase	Sulfadoxine	28	n.d.
PcyM_1023600	PI4K	phosphatidylinositol 4-kinase	KDU691	25	n.d.
PcyM_0207400	PheRS	phenylalanine--tRNA ligase alpha subunit	BRD3444	20	n.d.
PcyM_1312800	ATP4	P-type ATPase4	KAE609	2	n.d.

B

Gene ID	Gene name	Description	Sz (avg FPKM)	Hz (avg FPKM)	Heavy metal chelator
PcyM_1331900	CTR2	copper transporter, putative	306	13	Neocuproine
PcyM_1277100	CTR1	copper transporter, putative	165	6	Neocuproine
PcyM_0923300	MIT1	CorA-like Mg2+ transporter protein, putative	207	4	
PcyM_1142600	ZIP1	zinc transporter protein, putative	9	3	
PcyM_0609400	ZIPCO	ZIP domain-containing protein putative (ZIPCO)	402	2	Desferrioxamine (DFO)
PcyM_1444100	VIT	iron transporter, putative	150	1	Desferrioxamine (DFO)

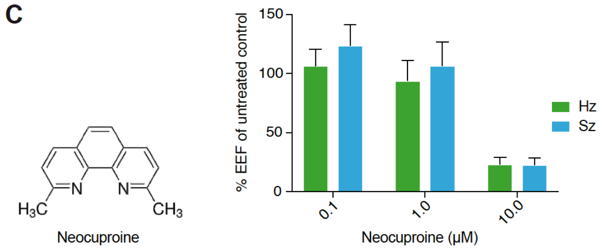


Figure 4. Expression of potential malaria drug targets in hypnozoites. (A) Table showing the list of known malaria drug targets along their expression levels in the liver stages and the targeting compound. **(B)** Table showing list of putative heavy metal transporters with chelating agents and their expression levels in the *P. cynomolgi* liver stages. **(C)** Structure formula of the copper chelator neocuproine. Dose-dependent effect of day 0-6 neocuproine treatment on *P. cynomolgi* liver stage schizonts (Sz) and hypnozoites (Hz). Bar charts show averaged results of 3 independent assays (7 wells per compound dilution in total) with standard error of the mean (sem).

Discussion

Because malaria liver stage parasites are more difficult to culture *in vitro*, the parasite hepatic life cycle has been neglected and our collective knowledge of those stages remains sparse. The need for new pre-erythrocytic vaccination strategies (Longley, Hill, & Spencer, 2015) and novel drug therapies to combat relapsing malaria parasites (Campo et al., 2015), recently fueled much interest for further investigations of the biology of liver stage parasites. As a significant contribution to these efforts, we report here a comprehensive comparative transcriptomics dataset of both developing and dormant liver stage *P. cynomolgi* malaria parasites. Using this dataset, we identified two protein markers that differentiate quiescent from actively dividing parasites and demonstrate that copper homeostasis is critically required for *P. cynomolgi* parasites replication and survival in hepatocytes. It is our hope that through multi-disciplinary collaborative efforts the research community will further mine this dataset to gain further insights in the biology of *Plasmodium* dormancy.

Recently a first *P. cynomolgi* hypnozoite transcriptomic dataset has been published (Cubi et al., 2017) which reports about 120 differentially expressed genes of which 69 are more than 3-fold upregulated, while we report here a much smaller number of upregulated genes in hypnozoites. It is important to note that Cubi *et al.* applied a uniform normalization that assumes that signals from different samples should be scaled to have the same median or average value thus not taking into account the size differences of replicating and dormant liver stages (Mikolajczak et al., 2015). This could have potentially biased their comparative analysis towards an over-estimation of the gene expression levels in hypnozoites. In contrast, we applied a group-wise normalization to the expression data in order to keep the expected difference in absolute level of gene expression between schizont and hypnozoite (Figure 1-figure supplement 3). Cubi *et al.* proposed that the gene PCYB_102390 (PcyM_1014300 in our dataset) encodes an ApiAP2 transcription factor AP2-Q (for quiescence) which could act as a master regulator of the hypnozoite fate (Cubi et al., 2017). However even after normalization, we failed to detect expression of this gene in 4 hypnozoite samples (Figure 3-figure supplement 2). Notwithstanding we detected transcripts for 9 other Api-AP2 genes in hypnozoites (Figure 3-figure supplement 2). None of the AP-AP2 genes, including PcyM_1014300, are, however, exclusive for relapsing malarias, as suggested previously (Cubi et al., 2017). Only further functional characterization, like the studies that revealed the role of the AP2-G and AP2-G2 genes in gametocyte commitment (Sinha et al., 2014), will reveal the possible role of these AP2 transcription factors in hypnozoite identity and survival.

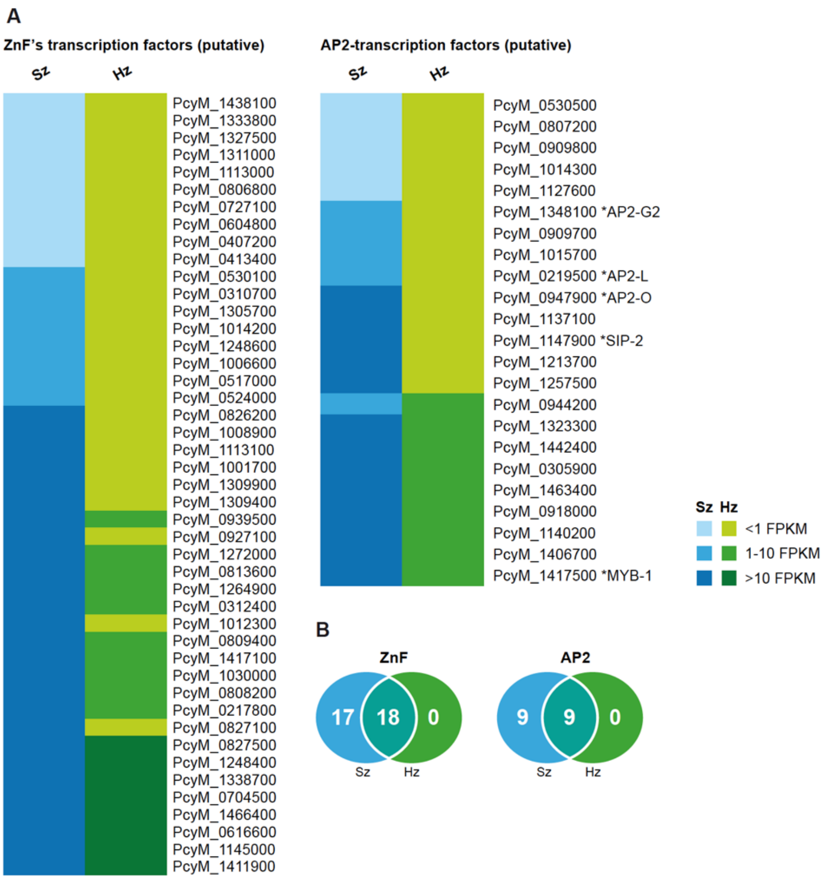


Figure 3-figure supplement 2. Expression of transcription factors in hypnozoites. (A) Heat map illustrating the expression values of genes belonging to putative target classes such as ZnF's and AP2-transcription factors. **(B)** Venn diagrams showing ZnF and AP2 transcription factors expressed in schizonts (Sz), hypnozoites (Hz), and both.

We have previously shown that hypnozoite physiology evolves over time and while PI4 kinase (PI4K) inhibitors are protective when administered shortly after the initial malaria liver infection, they fail to radically cure monkeys when administered several days after parasite inoculation (Zeeman et al., 2016). In agreement with our previous reports, we found that at least as early as day 6 post-infection, the *P. cynomolgi* PI4K gene is not expressed in hypnozoites. The current *in vitro* liver stage drug assays cannot distinguish compounds only active against developing hypnozoites from those with activity against established hypnozoites (Zeeman et al., 2016). The identification of markers specific to established hypnozoites would inform the design of parasites with transgenic reporter genes that would greatly assist the development of *in vitro* drug screening platforms (Campo et al., 2015). We found few upregulated genes in hypnozoites and unfortunately the most highly differentially expressed gene, PcyM_0533600, a member of the *etramp* family, was not translated in hypnozoites and sporozoites. The presence of such unproductive alternatively spliced transcripts in *Plasmodium* is not uncommon (Sorber, Dimon, & Derisi, 2011) and translational repression, including that of a member of the etramp family, UIS4 (Silvie, Briquet, Müller, Manzoni, & Matuschewski, 2014), has been shown to be involved in transitions between developmental stages of the life cycle (Lasonder et al., 2016). We showed nonetheless that the comparative transcriptomic dataset from this work can help select suitable proteins to produce monoclonal antibodies that differentially label specific liver stages. Further experiments are ongoing to expand the malaria liver stage research toolbox with selective and specific antibodies for replicating and quiescent liver stages.

Dormancy is a physiological response which is relevant to various chronic human infectious diseases and shared by a wide range of pathogens expressing physiological hallmarks characteristic of microbial quiescence. Our dataset suggests that several of these hallmarks are present in the *Plasmodium* hypnozoite—namely the maintenance of membrane potential, ATP biosynthesis and preservation of genome integrity (Rittershaus et al., 2013). Indeed, pathways analysis reveals that most mitochondrial electron flow genes and ATP production enzymes are robustly expressed in hypnozoites (Figure 3-figure supplement 1). Similarly, nucleus and chromatin maintenance genes are highly expressed in hypnozoites, and while canonical nonhomologous end joining (NHEJ) DNA-repair pathways are not present in *Plasmodium* (Gardner et al., 2002), we detected most of the homologous recombination repair (HR) enzymes as well as genes required for the maintenance of epigenetics marks (Figure 3-figure supplement 1). In addition, the transcriptomics data together with the FISH validation experiments, suggest that hypnozoites display a significant reduction in transcriptional rate both qualitatively and quantitatively which is another hallmark of quiescence (Figures 1C and 1D). All of these physiological hallmarks do theoretically provide proven therapeutic approaches for killing quiescent organisms with the targeting of pathogens' RNA polymerase(s) (Sala et al., 2010), proton-motive force enzymes (Andries et al., 2005) and DNA repair or epigenetic regulators (Dembélé et al., 2014; Sala et al., 2010). Establishing selective inhibition of these essential physiological processes in the parasite without significant toxicity to the host cells will be the key challenge for such approaches to be successful.

In order to survive, malaria parasites utilize membrane transport proteins that allow the uptake of nutrients, disposal of waste products and maintenance of ion homeostasis (Weiner & Kooij, 2016). While some of these transporters have been implicated in drug resistance, recent experimental work has also supported their potential as anti-malarial drug targets (Weiner & Kooij, 2016). Recent evidence has emerged for important roles of heavy metal homeostasis in sporozoite transmission and liver-stage development (Kenthirapalan et al., 2016; Sahu et al., 2014; Slavic et al., 2016; Stahel et al., 1988). Iron-deprivation inhibits liver stage growth (Goma, Renia, Miltgen, & Mazier, 1995; Stahel et al., 1988) and inactivation of a zinc-iron permease (ZIPCO) was shown to be detrimental for liver stage development (Sahu et al., 2014). We report here that *P. cynomolgi* liver stage parasites express transporters for heavy-metals, including copper that in these preliminary experiments seems to be crucially needed for liver stages. Targeting such essential import pathways will again require selective inhibition of the parasite transporters for such approaches to be viable therapeutically. Taken together the RNA-seq data indicate that drug target liver stage expression is necessary but clearly not sufficient for an inhibitor to show anti-parasitic liver stage activity. Nonetheless, it is worth noting that the 1-deoxy-D-xylulose 5-phosphate reductoisomerase (DXR), the target of Fosmidomycin (Umeda et al., 2011), and the Elongation Factor 2 (eEF2), the target of the recently discovered drug candidate DDD107498 (Baragana et al., 2015), are both expressed in the hypnozoite at day 6. This may warrant further investigations of the potential of these compounds for vivax malaria radical cure. Although we did not identify pathways or drug targets specific to hypnozoites, our data collectively show that the hypnozoite expresses a core set of genes required for its basic cellular function. Identifying those essential functions that could be safely targeted with small molecule inhibitors should reveal the Achilles' heel of the elusive hypnozoite.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.*, (215), 403–410.

Andries, K., Verhasselt, P., Guillemont, J., Neefs, J., Winkler, H., Gestel, J. Van, … Jarlier, V. (2005). A Diarylquinoline Drug Active on the ATP Synthase of Mycobacterium tuberculosis. *Science*, *307*(5707), 223–227.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., … Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nat Genet*, *25*(1), 25–29. Retrieved from http://dx.doi.org/10.1038/75556

Aurrecoechea, C., Brestelli, J., Brunk, B. P., Dommer, J., Fischer, S., Gajria, B., … Wang, H. (2009). PlasmoDB: A functional genomic database for malaria parasites. *Nucleic Acids Research*, *37*(SUPPL. 1), 539–543. http://doi.org/10.1093/nar/gkn814

Baragana, B., Hallyburton, I., Lee, M. C. S., Norcross, N. R., Grimaldi, R., Otto, T. D., … Gray, D. W. (2015). A novel multiple-stage antimalarial agent that inhibits protein synthesis. *Nature*. http://doi.org/10.1038/nature14451

Bateman, A., Martin, M. J., O'Donovan, C., Magrane, M., Apweiler, R., Alpi, E., … Zhang, J. (2015). UniProt: A hub for protein information. *Nucleic Acids Research*, *43*(D1), D204–D212. http://doi.org/10.1093/nar/gku989

Campo, B., Vandal, O., Wesche, D. L., & Burrows, J. N. (2015). Killing the hypnozoite - drug discovery approaches to prevent relapse in Plasmodium vivax. *Pathogens and Global Health*, *109*(3), 107–22.

http://doi.org/10.1179/2047773215Y.0000000013

Choveaux, D. L., Przyborski, J. M., & Goldring, J. P. D. (2012). A Plasmodium falciparum copper-binding membrane protein with copper transport motifs. *Malaria Journal*, *11*(1), 397. http://doi.org/10.1186/1475-2875-11-397

Cubi, R., Vembar, S. S., Biton, A., Franetich, J.-F., Bordessoulles, M., Sossau, D., … Mazier, D. (2017). Laser capture microdissection enables transcriptomic analysis of dividing and quiescent liver stages of Plasmodium relapsing species. *Cellular Microbiology*, *19*(8). http://doi.org/10.1111/cmi.12735

Dembélé, L., Franetich, J., Lorthiois, A., Gego, A., Zeeman, A., Kocken, C. H. M., … Mazier, D. (2014). Persistence and activation of malaria hypnozoites in long-term primary hepatocyte cultures. *Nature Medicine*, *20*(3), 307–312. http://doi.org/10.1038/nm.3461

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, B. P., & Chaisson M, G. T. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. http://doi.org/10.1093/bioinformatics/bts635

Fischer, S., Brunk, B. P., Chen, F., Gao, X., Harb, O. S., John, B., … Jr, C. J. S. (2012). Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. *Curr Protoc Bioinformatics*, 1–23. http://doi.org/10.1002/0471250953.bi0612s35.

Gardner, M. J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R. W., … Barrell, B. (2002). Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature*, *419*(6906), 498–511. Retrieved from http://dx.doi.org/10.1038/nature01097

Goma, J., Renia, L., Miltgen, F., & Mazier, D. (1995). Effects of iron deficiency on the hepatic development of Plasmodium yoelii. *Parasite (Paris, France)*, *2*(4), 351–356.

Gupta, A. P., Chin, W. H., Zhu, L., Mok, S., Luah, Y., Lim, E., & Bozdech, Z. (2013). Dynamic Epigenetic Regulation of Gene Expression during the Life Cycle of Malaria Parasite Plasmodium falciparum. *PLoS Pathog*, *9*(2). http://doi.org/10.1371/journal.ppat.1003170

Hahne, F., & Ivanek, R. (2016). Visualizing Genomic Data Using Gviz and Bioconductor. *Methods in Molecular Biology*, *1418*, 335–351. http://doi.org/10.1007/978-1-4939-3578-9

Hapuarachchi, S. V, Cobbold, S. A., Shafik, S. H., Dennis, A. S. M., McConville, M. J., Martin, R. E., … Lehane, A. M. (2017). The Malaria Parasite's Lactate Transporter PfFNT Is the Target of Antiplasmodial Compounds Identified in Whole Cell Phenotypic Screens. *PLOS Pathogens*, *13*(2), 1–24. http://doi.org/10.1371/journal.ppat.1006180

Harding, C. R., & Meissner, M. (2014). The inner membrane complex through development of Toxoplasma gondii and Plasmodium. *Cellular Microbiology*, *16*(5), 632–641. http://doi.org/10.1111/cmi.12285

James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, J. P. M. (2011). Integrative genomics viewer. *Nature Biotechnology*, *29*(1), 24–26. http://doi.org/10.1038/nbt0111-24

Jayabalasingham, B., Bano, N., & Coppens, I. (2010). Metamorphosis of the malaria parasite in the liver is associated with organelle clearance. *Cell Research*, *20*(9), 1043–59. http://doi.org/10.1038/cr.2010.88

Jiménez-díaz, M. B., Ebert, D., Salinas, Y., Pradhan, A., Lehane, A. M., Endsley, A.

N., … Horst, J. (2014). through ATP4 to induce rapid host-mediated clearance of Plasmodium, 5455–5462. http://doi.org/10.1073/pnas.1414221111

Kato, N., Comer, E., Sakata-kato, T., Sharma, A., Sharma, M., Maetani, M., … Winzeler, E. A. (2016). Diversity-oriented synthesis yields novel multistage antimalarial inhibitors. *Nature*, *538*(7625), 344–349. http://doi.org/10.1038/nature19804

Kenthirapalan, S., Waters, A. P., Matuschewski, K., & Kooij, T. W. A. (2014). Copper-transporting ATPase is important for malaria parasite fertility. *Molecular Microbiology*, *91*(2), 315–325. http://doi.org/10.1111/mmi.12461

Kenthirapalan, S., Waters, A. P., Matuschewski, K., & Kooij, T. W. A. (2016). Functional profiles of orphan membrane transporters in the life cycle of the malaria parasite. *Nature Communications*, *7*, 10519. Retrieved from http://dx.doi.org/10.1038/ncomms10519

Kono, M., Herrmann, S., Loughran, N. B., Cabrera, A., Engelberg, K., Lehmann, C., … Gilberger, T. W. (2012). Evolution and architecture of the inner membrane complex in asexual and sexual stages of the malaria parasite. *Molecular Biology and Evolution*, *29*(9), 2113–2132. http://doi.org/10.1093/molbev/mss081

Krotoski, W. A., Bray, R. S., Garnham, P. C. C., Gwadz, R. W., Killick-Kendrick, R., Draper, C. C., … Cogswell, F. B. (1982). Observations on Early and Late Post-Sporozoite Tissue Stages in Primate Malaria. *The American Journal of Tropical Medicine and Hygiene*, *31*(2).

Lasonder, E., Rijpma, S. R., Schaijk, B. C. L. Van, Hoeijmakers, W. A. M., Kensche, P. R., Gresnigt, M. S., … Sauerwein, R. W. (2016). Integrated transcriptomic and proteomic analyses of P . falciparum gametocytes : molecular insight into sex-specific processes and translational repression, *44*(13), 6087–6101. http://doi.org/10.1093/nar/gkw536

Longley, R. J., Hill, A. V. S., & Spencer, A. J. (2015). Malaria vaccines: identifying Plasmodium falciparum liver-stage targets. *Frontiers in Microbiology*, *6*, 965. http://doi.org/10.3389/fmicb.2015.00965

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. http://doi.org/10.1186/s13059-014-0550-8

March, S., Ng, S., Velmurugan, S., Galstian, A., Shan, J., Logan, D. J., … Hoffman, S. L. (2013). A Microscale Human Liver Platform that Supports the Hepatic Stages of Plasmodium falciparum and vivax. *Cell Host and Microbe*, *14*(1), 104–115. http://doi.org/10.1016/j.chom.2013.06.005

McNamara, C. W., Lee, M. C. S., Lim, C. S., Lim, S. H., Roland, J., Nagle, A., … Winzeler, E. A. (2013). Targeting Plasmodium PI(4)K to eliminate malaria. *Nature*, *504*(7479), 248–253. Retrieved from http://dx.doi.org/10.1038/nature12782

McNamara, C., & Winzeler, E. A. (2011). Target identification and validation of novel antimalarials. *Future Microbiology*, *6*(6), 693–704. http://doi.org/10.2217/fmb.11.45

Mikolajczak, S. A., Vaughan, A. M., Kangwanrangsan, N., Roobsoong, W., Fishbaugher, M., Yimamnuaychok, N., … Kappe, S. H. I. (2015). Plasmodium vivax Liver Stage Development and Hypnozoite Persistence in Human Liver-Chimeric Mice. *Cell Host & Microbe*, *17*(4), 526–535. http://doi.org/https://doi.org/10.1016/j.chom.2015.02.011

Miotto, O., Amato, R., Ashley, E. A., MacInnis, B., Almagro-Garcia, J., Amaratunga, C., … Kwiatkowski, D. P. (2015). Genetic architecture of artemisinin-resistant Plasmodium falciparum. *Nature Genetics*, *47*(3), 226–34. http://doi.org/10.1038/ng.3189

Mortazavi, A., Williams, B. A., Mccue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq, *5*(7), 1–8. http://doi.org/10.1038/NMETH.1226

Pain, M., Fuller, A. W., Basore, K., Pillai, A. D., Solomon, T., Bokhari, A. A. B., & Desai, S. A. (2016). Synergistic Malaria Parasite Killing by Two Types of Plasmodial Surface Anion Channel Inhibitors. *PLOS ONE*, *11*(2), 1–16. http://doi.org/10.1371/journal.pone.0149214

Pasini, E. M., Böhme, U., Rutledge, G. G., Voorberg-Van der Wel, A., Sanders, M., Berriman, M., … Otto, T. D. (2017). An improved Plasmodium cynomolgi genome assembly reveals an unexpected methyltransferase gene expansion. *Wellcome Open Research*, *2*, 42. http://doi.org/10.12688/wellcomeopenres.11864.1

Phillips, M. A., Lotharius, J., Marsh, K., White, J., Dayan, A., White, K. L., … Charman, S. A. (2015). A long-duration dihydroorotate dehydrogenase inhibitor (DSM265) for prevention and treatment of malaria. *Science Translational Medicine*, *7*(296), 296ra111.

Rao, S. P. S., Alonso, S., Rand, L., Dick, T., & Pethe, K. (2008). The protonmotive force is required for maintaining ATP homeostasis and viability of hypoxic, nonreplicating Mycobacterium tuberculosis. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(6), 11945–11950.

Rittershaus, E. S. C., Baek, S., & Sassetti, C. M. (2013). The Normalcy of Dormancy: Common Themes in Microbial Quiescence. *Cell Host and Microbe*, *13*(6), 643–651. http://doi.org/10.1016/j.chom.2013.05.012

Rottmann, M., McNamara, C., Yeung, B., MC, L., B, Z., B, R., … TT., D. (2010). Spiroindolones, a potent compound class for the treatment of malaria., *329*(5996), 1175–1180. http://doi.org/10.1126/science.1193225.Spiroindolones

Sahu, T., Boisson, B., Lacroix, C., Bischoff, E., Richier, Q., Formaglio, P., … Baldacci, P. (2014). ZIPCO, a putative metal ion transporter, is crucial for Plasmodium liver-stage development. *EMBO Molecular Medicine*, *6*(11), 1387–1397. http://doi.org/10.15252/emmm.201403868

Sala, C., Dhar, N., Hartkoorn, R. C., Zhang, M., Ha, Y. H., Schneider, P., & Cole, S. T. (2010). Simple Model for Testing Drugs against Nonreplicating Mycobacterium tuberculosis. *Antimicrobial Agents and Chemotherapy*, *54*(10), 4150–4158. http://doi.org/10.1128/AAC.00821-10

Schmidt, L. H., Fradkin, R., Genther, C. S., Rossan, R. N., & Squires, W. (1982). II. Responses of Sporozoite-Induced and Trophozoite-Induced Infections to Standard Antimalarial Drugs. *Am J Trop Med Hyg*, (31), 646–665.

Schuierer, S., & Roma, G. (2016). The exon quantification pipeline (EQP): a comprehensive approach to the quantification of gene, exon and junction expression from RNA-seq data. *Nucleic Acids Research*, gkw538. http://doi.org/10.1093/nar/gkw538

Silvie, O., Briquet, S., Müller, K., Manzoni, G., & Matuschewski, K. (2014). Post-transcriptional silencing of UIS4 in Plasmodium berghei sporozoites is important for host switch. *Molecular Microbiology*, *91*(6), 1200–1213.

http://doi.org/10.1111/mmi.12528

Sinha, A., Hughes, K. R., Modrzynska, K. K., Otto, T. D., Pfander, C., Dickens, N. J., … Waters, A. P. (2014). A cascade of DNA-binding proteins for sexual commitment and development in Plasmodium. *Nature*, *507*(7491), 253–257. http://doi.org/10.1038/nature12970

Slavic, K., Krishna, S., Derbyshire, E. T., & Staines, H. M. (2011). Plasmodial sugar transporters as anti-malarial drug targets and comparisons with other protozoa. *Malaria Journal*, *10*(1), 165. http://doi.org/10.1186/1475-2875-10-165

Slavic, K., Krishna, S., Lahree, A., Bouyer, G., Hanson, K. K., Vera, I., … Mota, M. M. (2016). A vacuolar iron-transporter homologue acts as a detoxifier in Plasmodium. *Nature Communications*, *7*, 10403. Retrieved from http://dx.doi.org/10.1038/ncomms10403

Sorber, K., Dimon, M. T., & Derisi, J. L. (2011). RNA-Seq analysis of splicing in Plasmodium falciparum uncovers new splice junctions, alternative splicing and splicing of antisense transcripts. *Nucleic Acids Research*, *39*(9), 3820–3835. http://doi.org/10.1093/nar/gkq1223

Stahel, E., Mazier, D., Guillouzo, A., Miltgen, F., Landau, I., Mellouk, S., … Gentilini, M. (1988). Iron Chelators: In Vitro Inhibitory Effect on the Liver Stage of Rodent and Human Malaria. *The American Journal of Tropical Medicine and Hygiene*, *39*(3), 236–240.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., … Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, *102*(43), 15545–15550.

http://doi.org/10.1073/pnas.0506580102

The Gene Ontology Consortium. (2017). Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Research*, *45*(D1), D331–D338. Retrieved from http://dx.doi.org/10.1093/nar/gkw1108

Umeda, T., Tanaka, N., Kusakabe, Y., Nakanishi, M., Kitade, Y., & Nakamura, K. T. (2011). Molecular basis of fosmidomycin's action on the human malaria parasite Plasmodium falciparum. *Scientific Reports*, 1–8. http://doi.org/10.1038/srep00009

Vaidya, A., JM, M., Z, Z., Das S, Daly TM, Otto TD, Spillman NJ, W., M, Siegl P, Marfurt J, Wirjanata G, Sebayang BF, Price RN, Chatterjee A, N. A., Stasiak M, Charman SA, Angulo-Barturen I, Ferrer S, B. J.-D. M., … Kortagere S, Burrows J, Fan E, B. L. (2014). Pyrazoleamide compounds are potent antimalarials that target Naþ homeostasis in intraerythrocytic Plasmodium falciparum. *Nat Commun.*, *5:5521*(May), 1–10. http://doi.org/10.1038/ncomms6521

Vaughan, A. M., O'neill, M. T., Tarun, A. S., Camargo, N., Phuong, T. M., Aly, A. S. I., … Kappe, S. H. I. (2009). Type II fatty acid synthesis is essential only for malaria parasite late liver stage development. *Cellular Microbiology*, *11*(3), 506–520. http://doi.org/10.1111/j.1462-5822.2008.01270.x

Voorberg-van der Wel, A., Zeeman, A. M., van Amsterdam, S. M., van den Berg, A., Klooster, E. J., Iwanaga, S., … Kocken, C. H. M. (2013). Transgenic Fluorescent Plasmodium cynomolgi Liver Stages Enable Live Imaging and Purification of Malaria Hypnozoite-Forms. *PLoS ONE*, *8*(1). http://doi.org/10.1371/journal.pone.0054888

Weiner, J., & Kooij, T. W. A. (2016). Phylogenetic profiles of all membrane transport proteins of the malaria parasite highlight new drug targets. *Microbial Cell*, *3*(10), 511–521. http://doi.org/10.15698/mic2016.10.534

World Health Organization (WHO). (2015). World malaria report 2015. Geneva, (http://www.who.int/malaria/publications/world-malaria-report-2015/en/).

Zeeman, A. M., Lakshminarayana, S. B., van der Werff, N., Klooster, E. J., Voorberg-van der Wel, A., Kondreddi, R. R., … Kocken, C. H. M. (2016). PI4K is a prophylactic, but not radical curative target in Plasmodium vivax - type malaria parasites. *Antimicrobial Agents and Chemotherapy*, *60*(5), AAC.03080-15. http://doi.org/10.1128/AAC.03080-15

Zeeman, A. M., Van Amsterdam, S. M., McNamara, C. W., Voorberg-van Der Wel, A., Klooster, E. J., Van Den Berg, A., … Kocken, C. H. M. (2014). KAI407, a potent non-8-aminoquinoline compound that kills Plasmodium cynomolgi early dormant liver stage parasites in vitro. *Antimicrobial Agents and Chemotherapy*, *58*(3), 1586–1595. http://doi.org/10.1128/AAC.01927-13

Zimin, A. V, Cornish, A. S., Maudhoo, M. D., Gibbs, R. M., Zhang, X., Pandey, S., … Norgren, R. B. (2014). A new rhesus macaque assembly and annotation for next-generation sequencing analyses. *Biology Direct*, *9*(1), 20. http://doi.org/10.1186/1745-6150-9-20

Chapter IV

Genomic analysis revealed new oncogenic signatures in *TP53*-mutant hepatocellular carcinoma

Abstract

The TP53 gene is the most commonly mutated gene in human cancers and mutations in TP53 have been shown to have either gain-of-function or loss-of-function effects. Using the data generated by The Cancer Genome Atlas, we sought to define the spectrum of TP53 mutations in hepatocellular carcinomas (HCCs) and their association with clinicopathologic features, and to determine the oncogenic and mutational signatures in TP53-mutant HCCs. Compared to other cancer types, HCCs harbored distinctive mutation hotspots at V157 and R249, whereas common mutation hotspots in other cancer types, R175 and R273, were extremely rare in HCCs. In terms of clinicopathologic features, in addition to the associations with chronic viral infection and high Edmondson grade, we found that TP53 somatic mutations were less frequent in HCCs with cholestasis or tumor infiltrating lymphocytes, but were more frequent in HCCs displaying necrotic areas. An analysis of the oncogenic signatures based on the genetic alterations found in genes recurrently altered in HCCs identified four distinct TP53-mutant subsets, three of which were defined by CTNNB1 mutations, 1q amplifications or 8q24 amplifications, respectively, that co-occurred with TP53 mutations. We also found that mutational signature 12, a liver cancer-specific signature characterized by T>C substitutions, was prevalent in HCCs

with wild-type TP53 or with missense TP53 mutations, but not in HCCs with deleterious TP53 mutations. Finally, whereas patients with HCCs harboring deleterious TP53mutations had worse overall and disease-free survival than patients with TP53-wild-type HCCs, patients with HCCs harboring missense TP53 mutations did not have worse prognosis. In conclusion, our results highlight the importance to consider the genetic heterogeneity among TP53-mutant HCCs in studies of biomarkers and molecular characterization of HCCs.

Introduction

Hepatocellular carcinomas (HCCs) display extensive histologic, transcriptomic and genetic diversity (Lee et al., 2004; Boyault et al., 2007; Chiang et al., 2008; Hoshida et al., 2009; Fujimoto et al., 2012; Guichard et al., 2012; Ahn et al., 2014; Schulze et al., 2015; The Cancer Genome Atlas Research Network, 2017). On the genetic level, genes involved in liver metabolism, Wnt and p53 signaling have been shown to be recurrently altered (Fujimoto et al., 2012; Guichard et al., 2012; Ahn et al., 2014; Schulze et al., 2015; The Cancer Genome Atlas Research Network, 2017). The most frequently mutated protein-coding genes are CTNNB1 (encoding β-catenin) and TP53 (encoding p53), both mutated in 20–40% of HCCs (Fujimoto et al., 2012; Guichard et al., 2012; Ahn et al., 2014; Schulze et al., 2015; The Cancer Genome Atlas Research Network, 2017).

TP53 is the most frequently mutated gene in human cancers (Kandoth et al., 2013). The p53 protein modulates multiple cellular functions, including transcription, DNA synthesis and repair, cell cycle arrest, senescence and apoptosis (Vogelstein et al., 2000). Mutations in TP53 can abrogate these functions, leading to genetic instability and progression to cancer (Vogelstein et al., 2000). Across 12 major cancer types (excluding HCC), 42% of cancers harbored TP53 somatic mutations, with at least 20% mutational rate in 10/12 cancer types and TP53 mutations are associated with inferior prognosis and unfavorable clinicopathologic parameters, such as tumor stage (Kandoth et al., 2013). Furthermore, TP53-mutant tumors are highly enriched among tumors driven by copy number alterations (CNAs), with most remaining TP53-mutant tumors associated with the presence of somatic mutations in the Wnt and/or the RAS-RAF-ERK signaling pathways (Ciriello et al., 2013).

The pattern of TP53 mutations is reminiscent of both an oncogene and a tumor suppressor gene (Vogelstein et al., 2013). The majority (86%) of TP53 mutations are in the DNA-binding domain (Olivier et al., 2010; Kandoth et al., 2013). Most mutations in the DNA-binding domain are missense (88%) and approximately 1/3 of missense mutations affect the hotspot residues R175, G245, R248, R249, R273, and R282 (Olivier et al., 2010). Outside the DNA-binding domain, most mutations (~60%) are nonsense or frameshift (Olivier et al., 2010). Mutant p53 proteins may lose the tumor suppressive functions and exert dominant-negative activities, but may also gain new oncogenic properties (Olivier et al., 2010; Muller and Vousden, 2014). Indeed, on the immunohistochemical level, p53 is generally detectable to various extents in samples with missense mutations but is undetectable in samples with truncating or frameshift mutations (Hall and McCluggage, 2006; Soussi et al., 2014).

In HCC, TP53 mutational frequency has been reported to range between 22 and 33% (Fujimoto et al., 2012; Guichard et al., 2012; Cleary et al., 2013; Kan et al., 2013; Ahn et al., 2014; Jhunjhunwala et al., 2014; Shiraishi et al., 2014; Totoki et al., 2014; Weinhold et al., 2014; Schulze et al., 2015; Fujimoto et al., 2016; The Cancer Genome Atlas Research Network, 2017). However, the frequency varies between geographic regions, etiological factors and carcinogen exposure, with more frequent TP53 mutations in regions where hepatitis B virus (HBV) infection is endemic (Fujimoto et al., 2012; Guichard et al., 2012; The Cancer Genome Atlas Research Network, 2017). Similar to other cancer types, TP53-mutant HCCs have been associated with features linked to poor prognosis, including high levels of alpha-fetoprotein, high Edmondson grade, expression of stem-like markers, and activation of pro-oncogenic signaling pathways (Kiani et al., 2002; Breuhahn et al., 2004; Lee et al., 2004; Peng et al., 2004; Boyault et al., 2007; Chiang et al., 2008; Hoshida et al., 2009; Goossens et al., 2015). Furthermore, patients with TP53-mutant HCCs tend to have shorter overall (OS) and disease-free survival (DFS) (Yano et al., 2007; Woo et al., 2011; Cleary et al., 2013). However, it appears that not all TP53 mutations in HCCs are equal. For instance, one of the most common mutation hotspots affecting residues R248/249 has an overall frequency of ~10% among TP53-mutant HCCs (Fujimoto et al., 2012, 2016; Ahn et al., 2014; Schulze et al., 2015; The Cancer Genome Atlas Research Network, 2017). In particular, the R249S mutation resulting from G>T transversion has specifically been linked to the combined effect of aflatoxin B1 exposure and HBV infection (Bressac et al., 1991; Hsu et al., 1991) and this mutation is detected in >75% of HCC from areas with high aflatoxin B1 exposure (Gouas et al., 2009; Kew, 2010). Further hotspot mutations affecting preferentially HCC are located at the residues V157 and H193 (both at ~2%) (Fujimoto et al., 2012, 2016; Ahn et al., 2014; Schulze et al., 2015; The Cancer Genome Atlas Research Network, 2017). Both R249S and V157F have been associated with stem cell-like traits and poor prognosis in HCC patients (Villanueva and Hoshida, 2011; Woo et al., 2011).

Finally, molecular classification studies have invariably grouped TP53-mutant HCCs under the umbrella of the aggressive subclass, but it is also clear that this subclass is molecularly, biologically and clinically heterogeneous (Boyault et al., 2007; Hoshida et al., 2009; Goossens et al., 2015).

Given the diverse pattern of TP53 mutations, taking advantage of The Cancer Genome Atlas (TCGA) dataset, in this study we sought to determine the pattern of TP53 somatic mutations in HCCs and its association with clinicopathologic features. Additionally, as TP53 mutations are associated with HCC molecular subclasses with poor prognosis, we sought to define the oncogenic and mutational signatures among TP53-mutant HCCs.

Material and methods

Sample Selection and Histologic Assessment

From TCGA liver hepatocellular carcinoma (LIHC) project (The Cancer Genome Atlas Research Network, 2017), 373 tumors with available somatic mutational data1 (accessed April 2017) (Gao et al., 2013) were included in the study. Images of diagnostic hematoxylin & eosin (H&E) slides were retrieved from the cbioportal and reviewed by three expert hepatopathologists (SA, MSM and LMT) according to the guidelines by the World Health Organization (Bosman et al., 2010) to define the presence or absence of cholestasis, Mallory bodies, tumor infiltrating lymphocytes (TILs), vessel infiltration and necrotic areas. 4-point scale Edmondson and Steiner system was adopted for tumor grading as previously described (Edmondson and Steiner, 1954; Alexandrov et al., 2013). Clinical information were obtained from the cbioportal (Gao et al., 2013).

Classification of TP53 Somatic Mutations

TP53 somatic non-synonymous and splice region mutations for the 373 HCCs were retrieved from the cbioportal (accessed April 2017) (Gao et al., 2013). TP53 mutations were stratified according to (i) the mutation type as single-nucleotide missense mutations (also encompassing synonymous mutations affecting splice region) or deleterious mutations (encompassing splice site, nonsense, in-frame, and frameshift mutations); (ii) whether the mutations were within or outside of the DNA-binding domain. For correlative analyses with clinicopathologic parameters, the sample (TCGA-DD-A1EE) with three TP53 mutations (A161S, H193R and C277∗) was classified as harboring deleterious mutation.

The spectrum of TP53 mutations in non-LIHC TCGA datasets were retrieved from the cbioportal (accessed June 2017) (Gao et al., 2013). Mutation (lolliplot) diagrams and Oncoprints were generated using cbioportal (Gao et al., 2013).

Genomic and Transcriptomic Data Analysis

Gene-level copy number ("gistic2_thresholded," 370/373 samples) and expression ("IlluminaHiSeq," 367/373 samples) data were retrieved from the UCSC Xena Functional Genomics Browser2 accessed April 2017). Gene-level copy number data were used to define genomic regions with differential frequencies of copy number alterations between HCCs with missense TP53 mutations, with deleterious TP53 mutations, or with wild-type TP53. Copy number states -2, -1, 0, 1, and 2 were considered homozygous deletion, heterozygous loss, copy number neutral, gain and high-level gain/amplification, respectively.

Transcriptomic data were in the form of gene-level, log-transformed, upper-quartile-normalized RSEM values. Molecular classification was performed according to Hoshida et al. (2009), using the Nearest Template Prediction: http://software.broadinstitute.org/cancer/software/genepattern. The R package limma was used to perform quantile normalization and for differential expression analysis. Multiple correction was performed using the Benjamini–Hochberg method. Genes with adjusted P-value < 0.05 were considered as differentially expressed.

The number of somatic mutations per sample was obtained from the cbioportal (Gao et al., 2013).

Oncogenic Signatures

Oncogenic signature ("oncosign") classification and the selection of genomic features as 'selected functional elements' (SFEs) input data were performed as described by Ciriello et al. (2013). Specifically, we selected 29 significantly mutated genes that have previously been reported as cancer genes (Futreal et al., 2004; Fujimoto et al., 2012; Kandoth et al., 2013; Lawrence et al., 2014), 27 recurrent amplifications and 34 recurrent deletions as SFEs. Robustness of the subclasses was assessed by removing 5, 10, or 20% of samples, reclassifying the reduced dataset, and calculating the Jaccard coefficients over 20 runs (Ciriello et al., 2013). Enrichment of genomic alterations was assessed using Chi-squared and Fisher's exact tests, as described (Ciriello et al., 2013).

Mutational Signatures

Decomposition of mutational signatures was performed using deconstructSigs (Rosenthal et al., 2016), based on the set of 30 mutational signatures (“signature.cosmic”) (Alexandrov et al., 2013; Nik-Zainal et al., 2016), for the 358 samples with at least 30 somatic mutations. Mutational signatures with >20% weight were considered to have substantial contribution to the overall mutational landscape. For each sample, the mutation signature with the highest weight was considered the dominant mutational signature.

Pathway Analysis

Pathways analysis was performed using Ingenuity Pathway Analysis as previously described (Piscuoglio et al., 2014; Martelotto et al., 2015). P < 0.001 was considered significant.

Statistical Analysis

Associations between TP53 mutations and clinical/histologic features were assessed using Mann–Whitney U, Chi-squared or Fisher’s exact tests as appropriate. Survival analyses were performed using the Kaplan-Meier method and the log-rank test. Univariate and multivariate analyses for OS and DFS were performed using the Cox proportional-hazards model. Mutual exclusivity and co-occurrence of somatic mutations were defined using the cbioportal (Gao et al., 2013). Statistical analyses comparing copy number profiles and defining genes up-regulated when gained or amplified and genes down-regulated when lost were performed as previously described (Piscuoglio et al., 2014). All tests were two-sided. P < 0.05 were considered statistically significant. Statistical analyses were performed with R v3.1.2 or SPSS v24 (IBM, Münchenstein, CH).

Results

Clinicopathologic Characterization and Molecular Classification of HCCs

TP53 mutation status was available for 373 HCCs subjected to whole-exome sequencing by TCGA (The Cancer Genome Atlas Research Network, 2017). Analysis of the clinical details of the patients revealed that the median age at diagnosis was 61 (range 16–90) and that 67.5% were male. Half of the patients were Caucasian (50.8%), with most remaining patients being Asian (43.9%). The most frequent primary risk factor was alcohol consumption (33.1%), followed by HBV (30.0%) and hepatitis C virus (HCV) infection (15.9%). Overall, history of at least one primary risk factor was noted in 74.2% patients.

We performed a comprehensive histopathologic review of the diagnostic H&E slides for all 373 included cases to assess Edmondson grade, the presence of cholestasis, Mallory bodies, vessel infiltration, necrotic areas, and TILs (Figure 1). Most samples were of intermediate grade, with 33.2, 60.6, and 5.4% graded as of Edmondson grades 2, 3, and 4, respectively. No sample was classified as of Edmondson grade 1. Cholestasis, Mallory bodies, vessel infiltration, necrotic areas, and TILs were present in 21.6, 22.0, 34.1, 24.8, and 47.3% of cases, respectively.

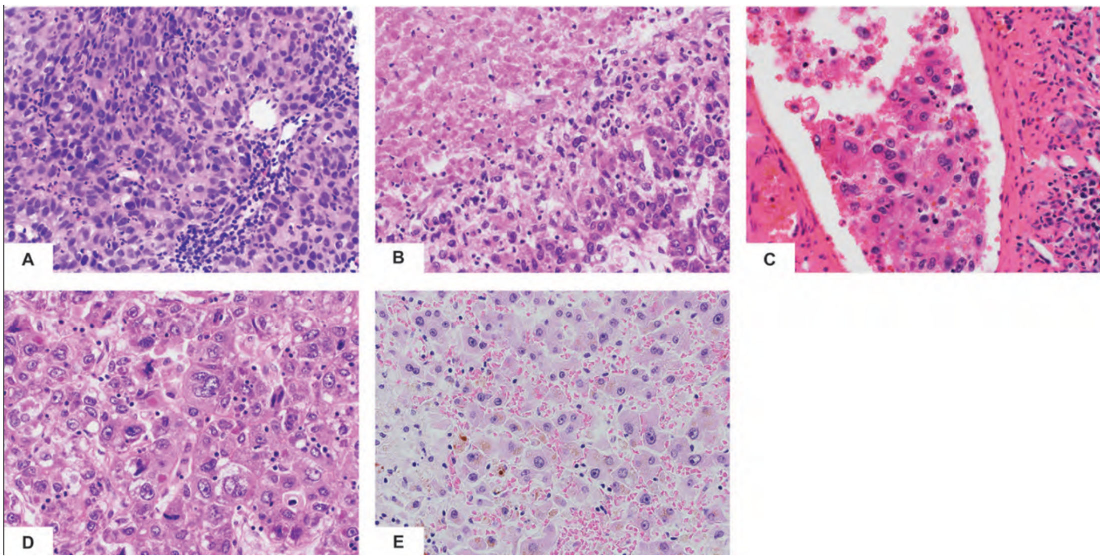


Figure 1. Histologic features of hepatocellular carcinoma. Low-power view of hepatocellular carcinomas with tumor infiltrating lymphocytes (A), necrotic areas (B), vessel infiltration (C), Mallory bodies (D), cholestasis (E) and of high Edmondson grade (F). Red arrows indicate the relevant histologic features.

Molecular classification was performed for the 367 HCCs for which expression data were available according to Hoshida et al. (2009). 31.3, 21.5, and 47.2% of HCCs were classified as S1, S2 and S3, respectively.

Spectrum of TP53 Somatic Mutations in HCCs

Given that TP53 is one of the most frequently mutated genes in HCCs and its diverse spectrum of mutations in human cancers, we sought to define the spectrum and type of TP53 mutations found in HCCs. A total of 116 somatic non-synonymous TP53 mutations and 2 synonymous TP53 mutations affecting splice regions were identified in 115 (30.8%) cases, including one case with three distinct mutations and one case with two. Missense (including missense and synonymous mutations affecting splice region) and deleterious (including nonsense, frame-shift, in-frame, splice site) mutations accounted for 73 (62%) and 45 (38%), respectively (Figure 2). Compared to other cancer types characterized by the TCGA, there was no difference between HCC and non-HCC tumor types in terms of the ratio of missense vs deleterious mutations (P = 0.197, Fisher's exact test).

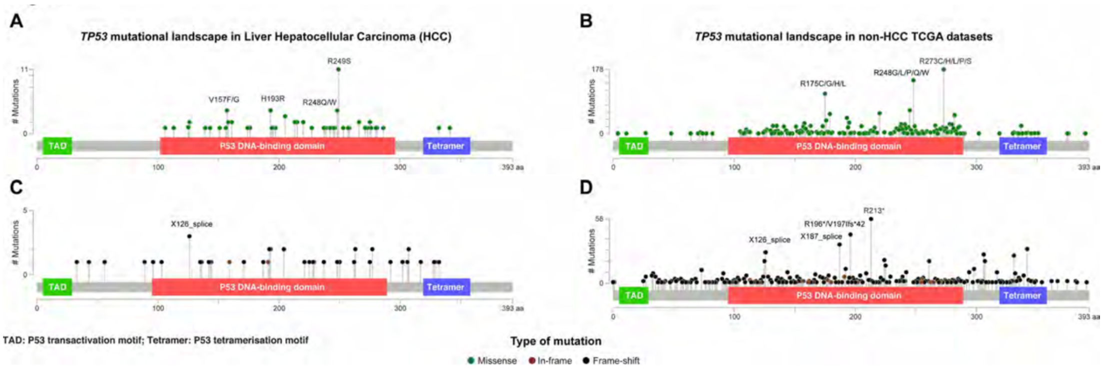


Figure 2. The distribution and the spectrum of TP53 mutations. The distribution and spectrum of TP53 missense (A,B) and deleterious (C,D) mutations in hepatocellular carcinoma (A,C) and in non-liver TCGA datasets (B,D). Diagrams represent the protein domains of p53 encoded by the TP53 gene. The presence of a mutation is shown on the x axis (lollipop), and the frequency of mutations is shown on the y axis. Missense mutations are presented as green circles, deleterious mutations (i.e., nonsense, frameshift, splice-site and in-frame) are depicted in black and brown circles. Plots were generated using cBioPortal tools (http://www.cBioPortal.org) and curated manually.

Of the 73 missense and synonymous mutations affecting splice region, 51 (70%) affected known hotspot residues (Chang et al., 2016; Gao et al., 2017) and all but one (99%) affected the DNA-binding domain (Figure 2A). All missense mutations were predicted to be pathogenic by at least 2/5 in silico mutation effect predictors, with the two synonymous mutations affecting splice region also predicted to be disease causing. The most frequent hotspot mutations were R249S (11/73, 15%), H193R (4/73, 5%), and R248Q/W (4/73, 5%). V157F, a mutation not considered to be a hotspot residue (Chang et al., 2016; Gao et al., 2017) but was reported as a mutation hotspot in HCCs (Woo et al., 2011), accounted for 4/73 (5%) of the missense mutations (Figure 2A).

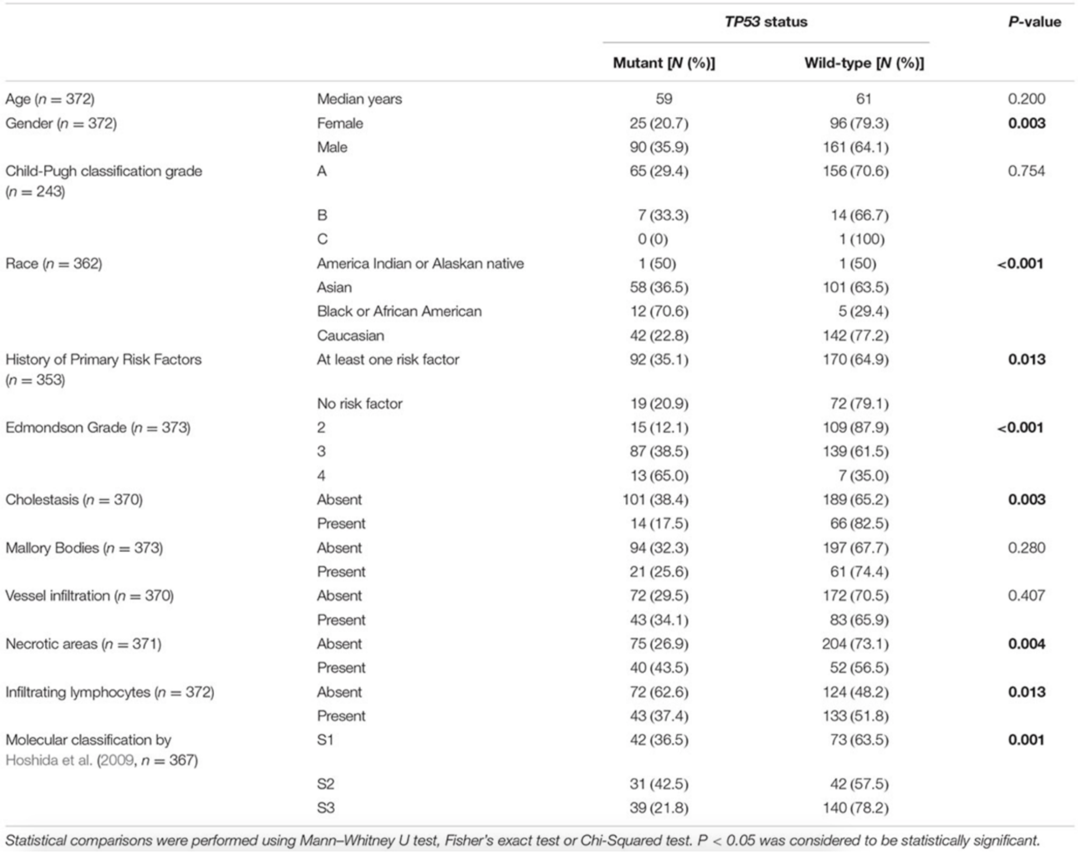
Compared to other cancer types, mutations affecting V157 and R249 accounted for greater proportions of the missense mutations in HCCs than in other cancer types (4/73, 5% vs. 22/1787, 1.2%, P = 0.017 and 11/73, 15% vs. 21/1787, 1.2%, P < 0.001, respectively, Fisher's exact tests, Figures 2A,B). In particular, R249S accounted for <0.5% of TP53 missense mutations in non-HCC TCGA samples, but accounted for 15% of the missense mutations in HCCs (P < 0.001, Fisher's exact test). In contrast, the most frequent hotspots in non-HCC tumors R273 (178/1787, 10.0% of missense mutations) and R175 (112/1787, 6.3%) were only observed once and not at all, respectively, in HCCs (P = 0.008 and P = 0.020, respectively, Fisher's exact tests).

The 45 deleterious mutations comprised 13 (29%) nonsense point mutations, 20 (44%) frameshift small insertions or deletions (indels), 3 (7%) in-frame indels and 9 (20%) mutations affecting splice sites. Unlike missense mutations, the 45 deleterious mutations were spread across the TP53 gene, with 32 (71%) in the DNA-binding domain, 3 (7%) in the tetramerization motif and 10 (22%) outside of these two domains (Figure 2C). In other cancer types, recurrent truncating mutations were observed at R196 (44/926, 4.8% of deleterious TP53 mutations) and R213 (56/926, 6.0%), both of which were not observed in HCC (Figures 2C,D).

Our results demonstrate that the spectrum of TP53 mutations in HCCs is distinct from that in non-HCC tumors, with HCC-specific recurrent hotspot mutations and a near absence of highly recurrent TP53 mutations found in other cancer types.

TP53 Status Correlates with Specific Histopathologic and Clinical Features of HCCs

Next, we sought to define whether TP53 mutation status correlated with clinicopathologic parameters. TP53 mutations were more frequently found in male patients (35.9% vs. 20.7% in female; P = 0.003, Fisher's exact test) and in patients with at least one primary risk factor (35.1% vs. 20.9%; P = 0.013, Fisher's exact test), especially in HCCs associated with HBV/HCV infection (53.1% vs. 39.7%; P = 0.021, Fisher's exact test, Table 1). Patients from different racial backgrounds were associated with different TP53 mutational frequencies (P = 0.001, Chi-squared test, Table 1). Black or African Americans had the highest frequency of TP53 mutations (70.6% vs. Asians, 36.5%, P = 0.009, and vs. Caucasians, 22.8%, P < 0.001, Fisher's exact tests), while Asians displayed more frequent TP53 mutations than Caucasians (P = 0.006, Fisher's exact test). No association with age of patients or Child-Pugh classification was observed.



		TP53 status		*P*-value
		Mutant [*N* (%)]	Wild-type [*N* (%)]	
Age (*n* = 372)	Median years	59	61	0.200
Gender (*n* = 372)	Female	25 (20.7)	96 (79.3)	**0.003**
	Male	90 (35.9)	161 (64.1)	
Child-Pugh classification grade (*n* = 243)	A	65 (29.4)	156 (70.6)	0.754
	B	7 (33.3)	14 (66.7)	
	C	0 (0)	1 (100)	
Race (*n* = 362)	America Indian or Alaskan native	1 (50)	1 (50)	**<0.001**
	Asian	58 (36.5)	101 (63.5)	
	Black or African American	12 (70.6)	5 (29.4)	
	Caucasian	42 (22.8)	142 (77.2)	
History of Primary Risk Factors (*n* = 353)	At least one risk factor	92 (35.1)	170 (64.9)	**0.013**
	No risk factor	19 (20.9)	72 (79.1)	
Edmondson Grade (*n* = 373)	2	15 (12.1)	109 (87.9)	**<0.001**
	3	87 (38.5)	139 (61.5)	
	4	13 (65.0)	7 (35.0)	
Cholestasis (*n* = 370)	Absent	101 (38.4)	189 (65.2)	**0.003**
	Present	14 (17.5)	66 (82.5)	
Mallory Bodies (*n* = 373)	Absent	94 (32.3)	197 (67.7)	0.280
	Present	21 (25.6)	61 (74.4)	
Vessel infiltration (*n* = 370)	Absent	72 (29.5)	172 (70.5)	0.407
	Present	43 (34.1)	83 (65.9)	
Necrotic areas (*n* = 371)	Absent	75 (26.9)	204 (73.1)	**0.004**
	Present	40 (43.5)	52 (56.5)	
Infiltrating lymphocytes (*n* = 372)	Absent	72 (62.6)	124 (48.2)	**0.013**
	Present	43 (37.4)	133 (51.8)	
Molecular classification by Hoshida et al. (2009, *n* = 367)	S1	42 (36.5)	73 (63.5)	**0.001**
	S2	31 (42.5)	42 (57.5)	
	S3	39 (21.8)	140 (78.2)	

Statistical comparisons were performed using Mann–Whitney U test, Fisher's exact test or Chi-Squared test. P < 0.05 was considered to be statistically significant.

Table 1. Analyses of TP53 status and clinicopathologic parameters in the 373 HCCs from The Cancer Genome Atlas cohort (The Cancer Genome Atlas Research Network, 2017).

Correlation with histologic features revealed that TP53-mutant HCCs were associated with high Edmondson grade, accounting for 12.1, 38.5, and 65.0% of cases classified as Edmondson grades 2, 3, and 4, respectively (P < 0.001, Chi-squared test, Table 1). TP53 mutations were less frequent in HCCs associated with cholestasis (17.5% vs. 38.4%; P = 0.003, Fisher's exact test) and were more frequent in HCCs with necrotic areas (43.5% vs. 26.9%; P = 0.004, Fisher's exact test, Table 1). The presence of TILs was associated with less frequent TP53 mutations (37.4% vs. 62.6%; P = 0.013, Fisher's exact test; Table 1). No association was found between TP53 mutation status and the presence of Mallory Bodies or vessel infiltration.

Further analyses comparing HCCs with missense or deleterious mutations showed that patients with HCCs with deleterious TP53 mutations were slightly older than those with missense mutations (median 64 vs. 58, P = 0.049, Mann–Whitney U test). After excluding one patient (TCGA-DD-A1EE) with both deleterious mutation (C277∗) and hotspot missense (H193R) mutations, the ages between the two groups were not different (P = 0.058, Mann–Whitney U test). Of note, TP53 recurrent hotspots V157F, R158H, H193R, Y205, and R249S were exclusively found in tumors of high Edmondson grade (grades 3/4, P = 0.038, Fisher's exact test, compared to HCCs with other TP53 mutations).

Correlating TP53 status with molecular classification, (Hoshida et al., 2009) TP53-mutant HCCs were preferentially enriched in the S1 and S2 subclasses (36.5% and 42.5% vs. 21.8% in S3, P = 0.001, Chi-squared test, Table 1). Stratifying TP53-mutant HCCs into those with missense or deleterious mutations did not reveal association between TP53 mutation types and molecular classification (P = 0.459, Chi-squared test).

These results demonstrate that, additional to the well-established associations with the male gender, HBV/HCV infection and high Edmondson grade, TP53 mutations were less frequent in HCCs with cholestasis or TILs, but were more frequent in HCCs with necrotic areas.

Genomic Instability Is Not Associated with TP53 Mutation Type

Next, we compared the number of somatic genetic alterations between TP53-wild-type and mutant cases. Mutational burden was higher in TP53-mutant HCCs, HCCs with missense TP53 mutations and HCCs with deleterious TP53 mutations than TP53-wild-type cases ($P < 0.001$, $P < 0.001$ and $P = 0.004$, respectively, Mann–Whitney U tests), but no difference was observed between cases with missense or deleterious mutations ($P = 0.799$, Mann–Whitney U test). Similarly, TP53-mutant HCCs, HCCs with missense TP53 mutations and HCCs with deleterious TP53 mutations all harbored higher number of genes affected by CNAs compared with TP53-wild-type cases ($P < 0.001$, $P < 0.001$ and $P = 0.001$, respectively, Mann–Whitney U tests), with no difference between cases with missense or deleterious TP53 mutations ($P = 0.352$, Mann–Whitney U test).

Consistent with their increased chromosomal instability, TP53-mutant HCCs displayed more frequent gains of chromosomes 1p, 3, 10p and 19p and losses of half the genome, notably of chromosomes 4, 5, 10q, 14, 17p, 18 and 19. The CNA landscapes between HCCs with TP53 missense or deleterious mutations were remarkably similar.

To identify potential CNA drivers associated with TP53 mutations, we interrogated the genes overexpressed when gained and genes downregulated when lost in the regions that showed differential CNA frequencies between TP53-mutant and TP53-wild-type cases. Pathway analysis of the copy number-regulated genes revealed that

TP53-mutant cases displayed deregulation in pathways associated with EIF2 signaling, protein ubiquitination pathway, RNA polymerase-II complex and DNA repair pathways, and in molecular and cellular functions related to cell death and survival, cell cycle, DNA replication, recombination and repair.

TP53-Mutant HCCs Displayed Heterogeneous Oncogenic Signatures

In HCCs, TP53 and CTNNB1 mutations were largely mutually exclusive (P = 0.028, Figure 3A) (Fujimoto et al., 2012; Guichard et al., 2012; Schulze et al., 2015; The Cancer Genome Atlas Research Network, 2017). Additionally, TP53 and BAP1 mutations were also mutually exclusive (P = 0.004; Figure 3A). In contrast, TP53 mutations co-occurred with RB1, JAK1 and KEAP1 mutations (P = 0.028, P = 0.034 and P = 0.044, respectively, Figure 3A). These observations suggest that TP53-mutant HCCs likely constitute a genetically heterogeneous subclass and may be subclassified into categories with distinct oncogenic signatures.

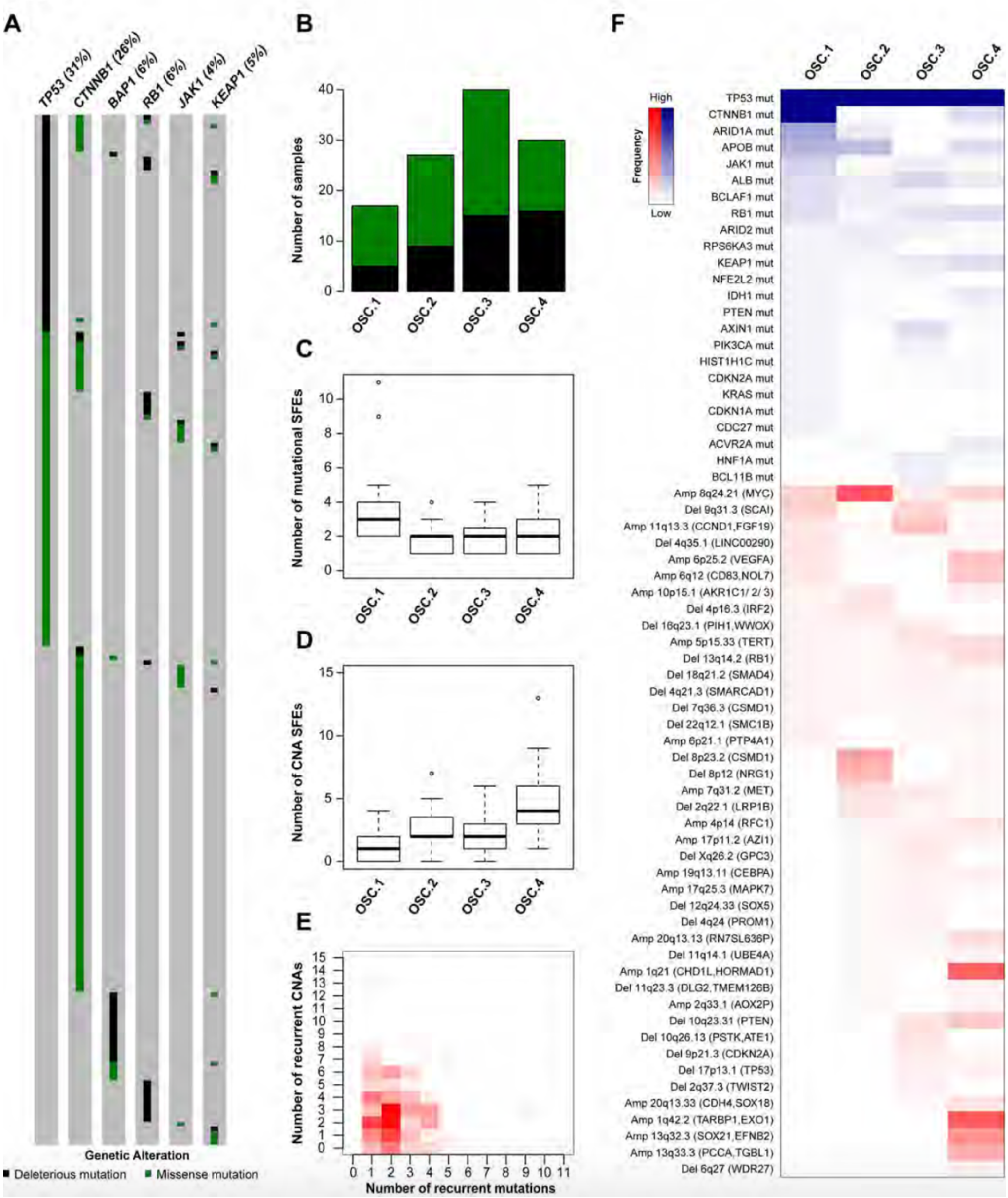


Figure 3. Oncogenic signature classes in TP53-mutant hepatocellular carcinoma. The pattern of mutations in TP53, CTNNB1, BAP1, RB1, JAK1 and KEAP1 in hepatocellular carcinoma (A). Number of TP53-mutant samples classified as OSC1, OSC2, OSC3, and OSC4, according to the color key in A (B). Number of mutational (C) and copy number (D) 'selected functional elements' (SFEs) in the different subclasses. The distribution of mutational vs copy number SFEs in TP53-mutant cases (E). The shade of red is proportional to the number of samples for a given (x,y) position. Heatmap shows the mutational and copy number SFEs altered in at least 5% of the samples in at least one oncogenic signature class (F). Shades of red and blue are proportional to the number of samples with a given genetic alteration, according to the color key. Plot in (A) was generated using cBioPortal (http://www.cBioPortal.org) and curated manually.

To define the oncogenic signatures in TP53-mutant HCCs, we performed unsupervised partitioning of the samples into classes with distinct patterns of likely 'driver' genetic alterations (or 'selected functional elements,' SFEs), (Ciriello et al., 2013) including mutations in 29 significantly mutated genes, amplifications in 27 recurrently amplified regions, and homozygous deletions in 34 recurrently deleted regions (see Materials and Methods). Among the 144 TP53-mutant HCCs with mutational and CNA data, we found median of 2 mutational (range 0–11) and 2.5 CNA (range 0–13) SFEs in each case and identified four robust oncogenic signature classes (OSCs, Figures 3B–E). HCCs with missense or deleterious TP53 mutations did not cluster separately (P = 0.305, Chi-squared test, Figure 3B), nor HCCs of distinct transcriptomic subclasses. Inspection of the SFEs that characterized each OSC revealed that OSC1 was defined by the presence of CTNNB1 mutations (100%, P < 0.001, Fisher's exact test, Figure 3F). The most frequent alteration in OSC2 was 8q24.21 amplification (encompassing MYC, 67%, P < 0.001, Fisher's exact test), while the most frequent alterations in OSC4 were 1q21.3 (encompassing CHD1L and HORMAD1, 60%) and 1q42.2 (encompassing TARBP1 and EXO1, 63%) amplifications (both P < 0.001, Fisher's exact tests, Figure 3F). OSC3 was notable for lacking highly recurrent genetic alterations, with the most frequent alteration being 11q13.3 amplification (CCND1, 23%, P = 0.011, Fisher's exact test). Additionally, ARID1A mutations were enriched in OSC1 (35%, P < 0.001, Fisher's exact test), while 10q23.21 deletion (PTEN, 20%) and 6p25.2 amplification (VEGFA, 23%) were enriched in OSC4 (P = 0.020 and P = 0.001, respectively, Fisher's exact tests). We also found that OSC1 harbored higher number of mutational SFEs and lower number of CNA SFEs (P < 0.001 and P = 0.002, respectively, Mann–Whitney U tests, Figures 3C,D) compared to other classes. By contrast, OSC4 harbored higher number of CNA SFEs than the other classes (P < 0.001, respectively, Mann–Whitney U test, Figure 3D). The TP53 R249S hotspot mutation was not associated with specific OSC classes (P = 0.591, Chi-squared test). Finally, OSC1/2 were more frequently associated with the presence of TILs than OSC3/4 (P = 0.028, Chi-squared test). No other associations between histologic or clinicopathologic parameters and OSCs were found.

These observations are concordant with the observation that tumors are primarily driven by either somatic mutations or CNAs but rarely both (Ciriello et al., 2013) (Figure 3E). Furthermore, we identified subclasses of TP53-mutant HCCs likely driven by co-occurring CTNNB1 mutations, 8q24.21 (MYC) amplification or 1q amplification in a mutually exclusive manner.

Mutational Signatures in TP53-Mutant HCCs

The somatic mutational landscapes are shaped by endogenous and/or environmental biological and chemical processes (Alexandrov et al., 2013). More than 10 mutational signatures have been identified in liver cancers, including two liver cancer-specific signatures 12 and 16 of unknown etiology, both of which are characterized by frequent T>C substitutions but with different sequence contexts (Alexandrov et al., 2013). To determine whether TP53-mutant HCCs harbored distinct mutational signatures compared to TP53-wild-type HCCs, we inferred the underlying mutational processes for the 358 HCCs with at least 30 somatic mutations (Alexandrov et al., 2013; Nik-Zainal et al., 2016). The age-associated signature 5, (Alexandrov et al., 2015) and the liver cancer-specific signatures 12 and 16 contributed substantially (≥20% weight) to the mutational landscapes in 17.0, 12.8, and 53.4% of the samples, respectively (Figure 4). Together, 72.9% of HCCs harbored signatures 5, 12 or 16 as the dominant signatures (14.0, 10.6, and 48.3%, respectively).

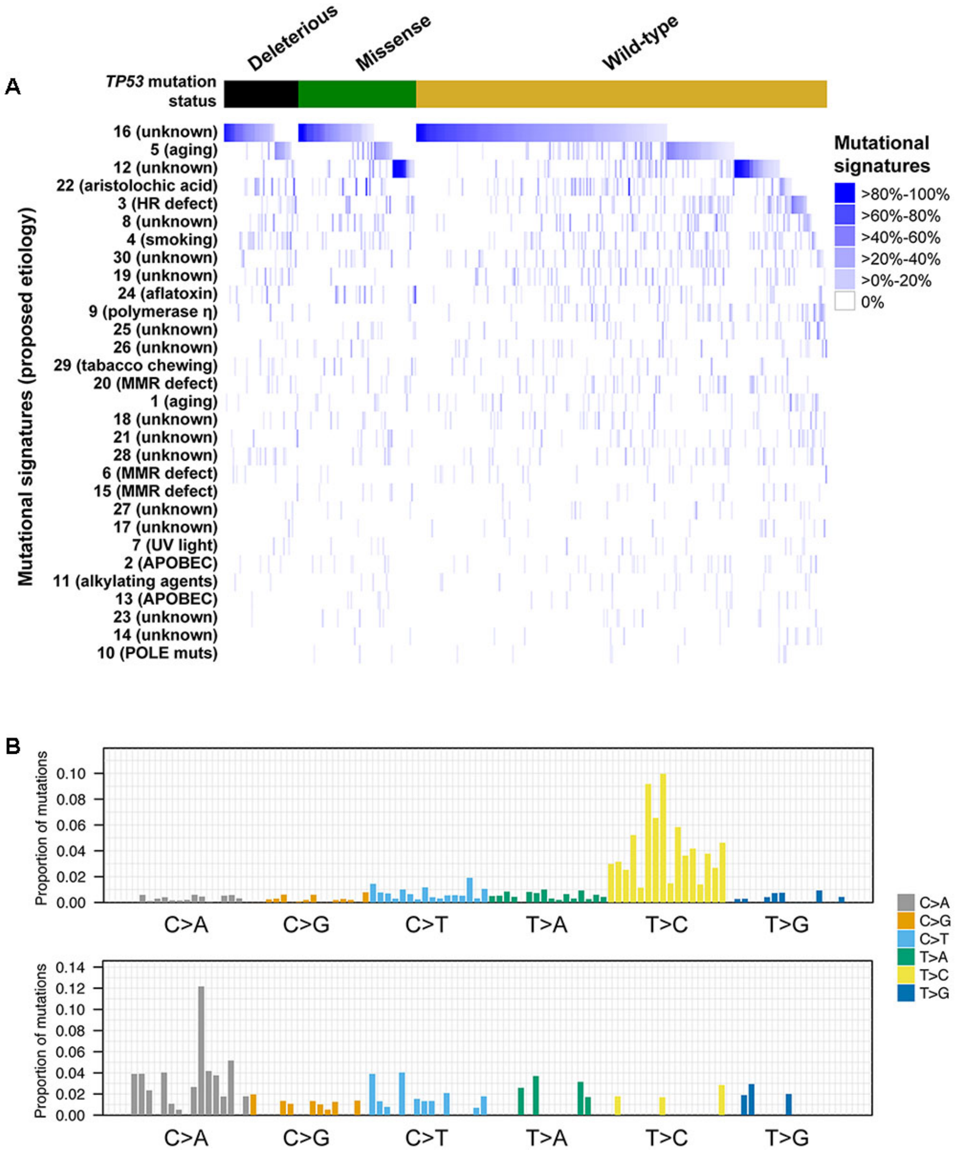


Figure 4. Mutational signatures in hepatocellular carcinoma with and without TP53 somatic mutations. Heatmap depicting the mutational signatures that shaped the genomes of the tumor samples analyzed (A) (Alexandrov et al., 2013). The similarity of the pattern of substitutions to the published mutational signatures is indicated in blue according to the color key. HCC samples were divided according to their TP53 mutational status. Mutational signatures were sorted by the number of cases classified as having a given mutational signature as the dominant signature, in decreasing order. Barplots illustrating examples of mutational signatures 12 (upper) and 24 (bottom) (B). In each panel, the colored barplot illustrates each mutational signature according to the 96 substitution classification defined by the substitution classes (C>A, C>G, C>T, T>A, T>C, and T>G bins) and the 5′ and 3′ sequence context, normalized using the observed trinucleotide frequency in the human exome to that in the human genome. The bars are ordered first by mutation classes (C>A/G>T, C>G/G>C, C>T/G>A, T>A/A>T, T>C/A>G, T>G/A>C), then by the 5′ flanking base (A, C, G, T) and then by the 3′ flanking base (A, C, G, T).

A comparison of the mutational signatures with substantial contribution (≥20%) to the mutational landscapes of TP53-mutant or TP53-wild-type HCCs revealed that only the aflatoxin-associated signature 24 was enriched among TP53-mutant HCCs (7/114, 6.1% vs. 4/244, 1.6%, P = 0.042, Fisher's exact test).

We further compared the mutational signatures between HCCs with missense or deleterious TP53 mutations. Interestingly, while 18.6% (13/70) of samples with missense TP53 mutations displayed substantial contribution from signature 12, only 4.5% (2/44) of samples harboring deleterious TP53 mutations did (P = 0.044, Fisher's exact test), with signature 12 being the dominant signature in 15.7% (11/70) and 2.3% (1/44) of samples with missense or deleterious TP53 mutations, respectively (P = 0.027, Fisher's exact test, Figure 4). No difference in other signatures was observed. The aflatoxin-associated signature 24 was enriched among R249S-mutant HCCs compared other TP53-mutant HCCs (4/11, 36% vs. 3/103, 3%, P = 0.001 for substantial contribution and 3/11, 27% vs. 2/103, 2%, P = 0.006 for dominant signature, Fisher's exact tests).

Taken together, our results suggest that the different types of TP53 mutations were associated with distinct mutational processes. Specifically, signature 12 was rarely found in HCCs with deleterious TP53 mutations.

Distinct Types of TP53 Mutations Are Associated with Different Prognoses

Previous studies found that associations between the types of TP53 mutations and prognoses in breast, and head and neck cancers (Olivier et al., 2006; Ozcelik et al., 2007; Vegran et al., 2013; Lapke et al., 2016). Here we hypothesized that patients with HCCs harboring TP53 missense or deleterious mutations may display different prognoses. Considering the patients with available data on OS (n = 372) or DFS (n = 321), we found that patients with TP53-mutant HCCs displayed a more aggressive behavior including shorter OS and DFS than TP53-wild-type patients (P = 0.018 and P = 0.005, respectively, log-rank tests, Figure 5). Patients with missense or deleterious TP53 mutations did not differ in OS or DFS (P = 0.129 and P = 0.148, respectively, log-rank tests, Figure 5). Importantly, while patients with deleterious TP53 mutations had worse OS and DFS than TP53-wild-type patients (P = 0.004 and P = 0.001, respectively, log-rank tests, Figure 5), there was no difference in OS or DFS between patients with missense TP53 mutations and those wild-type for TP53 (P = 0.192 and P = 0.084, respectively, log-rank tests, Figure 5).

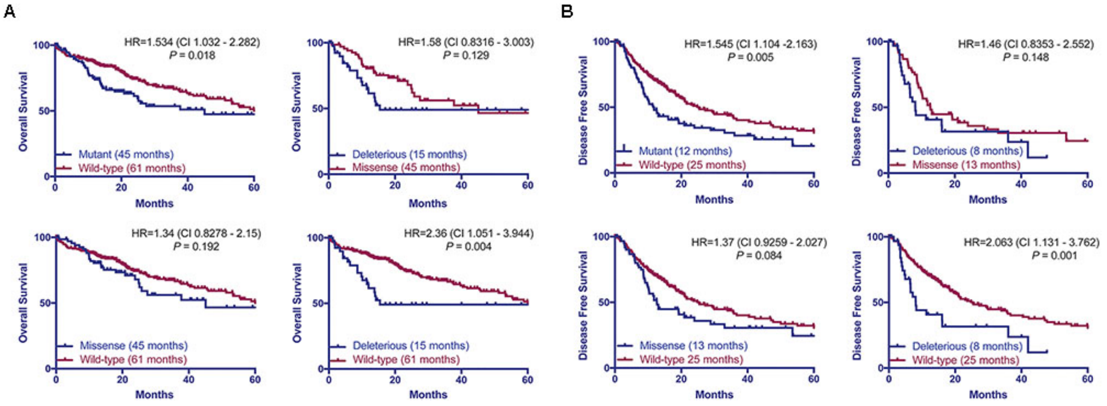


Figure 5. TP53 mutation status is associated with worse overall and disease-free survival. Overall (A) and disease-free survival (B) of HCC patients with and without TP53 somatic mutations using the Kaplan–Meier method. Median survival for each group is indicated in parentheses. Statistical comparisons were performed using log-rank tests. P < 0.05 was considered statistically significant.

As an exploratory analysis, we asked whether OSCs or mutational signatures of TP53-mutant HCCs were prognostic. Compared to OSC1 (28 months), OSC2 (26 months) and OSC3 (median not reached), OSC4 was associated with the shortest median OS of 14 months, although the difference was not statistically significant (P = 0.366, log-rank test;). Univariate Cox regression analyses revealed that the aflatoxin-associated signature 24 (HR 3.275, CI 1.279–8.384, P = 0.013), HBV infection status and the presence of necrotic areas were associated with poor prognosis. However, in a multivariate analysis, mutational signature 24 was not an independent prognostic indicator (P = 0.242).

Taken together, our results showed only patients with deleterious TP53 mutations but not missense TP53mutations were associated with significantly worse OS and DFS in this cohort.

Discussion

In this study, we performed a detailed analysis of TP53 somatic mutational spectrum in HCCs, with nearly all missense mutations (98%) and most deleterious mutations (73%) affecting the DNA-binding domain. Notably, we found that the residues mutated in HCCs differed from those in other cancer types. Hotspot mutations R249S and V157F were common in HCCs but extremely rare in other cancers, while mutations affecting R175 and R273, two of the most frequently mutated residues in other cancers, were nearly absent in HCCs. This latter observation also applies to other HCC datasets (Ahn et al., 2014; Schulze et al., 2015), suggesting that TP53 mutational spectrum in HCC is distinct from that in other cancers.

To determine the genotype–phenotype correlation between TP53 mutation status and clinicopathologic parameters, we performed a detailed assessment of histologic features using H&E slides. We confirmed the established associations with the male gender, HBV/HCV infection and high Edmondson grade. Additionally, TP53 mutations were associated with the presence of necrotic areas, and accordingly, with the absence of cholestasis, a feature more frequently observed in well-differentiated HCCs. Finally, we observed that the presence of TILs was associated with less frequent TP53 mutations, in line with the favorable prognosis associated with tumors with high TILs in other tumor types (Mahmoud et al., 2011).

Analysis of the mutational signatures revealed that signatures 16 of unknown etiology and the age-associated signature 5 (Alexandrov et al., 2015) were the most prevalent in HCCs. We also found that signature 12 of unknown etiology, characterized by frequent T>C substitutions, was prevalent in TP53-wild-type and HCCs with missense TP53 mutations but were largely absent in those with deleterious TP53 mutation. A previous study reported that the W3 signature, which was highly similar to signature 12 (Fujimoto et al., 2012), was associated with the age of patients. Here we found no difference in the age of patients when we considered tumors with strictly missense or deleterious TP53 mutations (i.e., excluding one patient with both types). The basis of signature 12 is thus unclear and further studies are required to elucidate its biological significance.

Adopting the algorithm of "oncosign" (Ciriello et al., 2013), we identified four robust subclasses of TP53-mutant HCCs with distinct oncogenic signatures. Of these classes, one subclass was likely driven by co-occurring CTNNB1 mutations, while two subclasses were likely driven by amplicon drivers on 1q and 8q. 1q21 amplification has been linked to hepatocarcinogenesis, with ALC1 (CHD1L) overexpression in HCC cells shown to promote G1/S phase transition and to inhibit apoptosis (Ma et al., 2008). The authors further suggested that the oncogenic function of ALC1 might be associated with its role in promoting cell proliferation by down-regulating p53 expression (Ma et al., 2008). The 1q21 amplicon also contains HORMAD1, a gene that has been shown to drive chromosomal instability in breast cancer (Watkins et al., 2015). As for 8q24, in addition to the well-known oncogenic role of MYC, previous studies have also shown that MYC amplification is an indicator of malignant potential and poor prognosis in HCC (Lin et al., 2010), and that the co-occurrence of MYC amplification and p53 alteration may contribute to HCC progression (Kawate et al., 1999). The remaining subclass did not have highly recurrent genetic alterations. Interestingly, this subclass was numerically, though not statistically, associated with the most favorable OS among the four classes. One may speculate that TP53-mutant HCCs lacking additional drivers may constitute a less aggressive subclass. Of note, the features that characterized the four OSCs were largely mutually exclusive, suggesting that distinct oncogenic processes are operative in non-overlapping subsets of TP53-mutant HCCs.

TP53 mutation status predicts worse OS and DFS in HCC patients (Yano et al., 2007; Woo et al., 2011; Cleary et al., 2013). However, we found that patients with deleterious mutations, but not those with missense mutations, were associated with worse OS and DFS compared to patients wild-type for TP53. This is in line with other tumor types, in which different types of TP53 mutations have been associated with different prognoses (Olivier et al., 2006; Ozcelik et al., 2007; Vegran et al., 2013; Lapke et al., 2016). In fact, the risk of death or relapse for patients harboring deleterious mutation is 2.3 times (HR = 2.36 and 2.063, respectively) higher than TP53-wild-type patients. The prognosis for patients with missense mutations appears to sit between those with wild-type TP53 or deleterious TP53 mutations, albeit not statistically different from either group. It is conceivable that the prognostic significance of the type of TP53 mutations may be confirmed in a larger cohort with extensive follow-up.

It has been suggested that TP53 missense mutations have varying capacity to transactivate p53 target genes and to alter the responsiveness to chemotherapeutic agents in breast cancer (Jordan et al., 2010). A differential expression analysis using the HCC TCGA dataset comparing HCCs with TP53 missense mutations and those with TP53 deleterious mutations identified TP53 itself as up-regulated but did not identify significantly altered genes (data not shown). Furthermore, HCCs harboring the missense mutations functionally shown to lack the ability to transactivate genes with p53 response elements (Jordan et al., 2010) did not differ from HCCs with other missense mutations on the transcriptomic level (data not shown). It is thus unclear precisely how the various TP53 mutations may differentially alter the transcriptomic landscape of HCCs. Further functional studies may be required to elucidate how the types of TP53 mutations may affect its biological functions.

In HCC molecular characterization studies to date, HCCs are typically classified as TP53-wild-type or TP53-mutant, where all TP53 mutations were treated as equal (Fujimoto et al., 2012; The Cancer Genome Atlas Research Network, 2017). However, many studies have demonstrated that TP53 can be affected by either (or both) gain-of-function or loss-of-function mutations, with missense mutations preferentially displaying gain-of-function or neomorphic properties (Muller and Vousden, 2014). Our study has demonstrated that HCCs with missense or deleterious TP53 mutations display similar clinicopathologic features, mutational/CNA burden and oncogenic signatures, but are associated with distinct mutational signatures. Clinically, while patients with tumors harboring deleterious TP53 mutations had worse prognosis compared to those wild-type for TP53, there was no statistically significant difference between those with missense mutations and those wild-type for TP53. Our study highlights the importance to consider the type of TP53 mutations in studies of biomarkers and molecular characterization of HCCs.

Our study has limitations. Despite TCGA being the largest genomic study of HCC, it is by no means the only large-scale study. However, as one of our aims was to define clinicopathologic correlates, we chose TCGA as it is the only study with publicly available H&E slides for pathology review. Secondly, the power of the OS and DFS analyses was limited due to the cohort size. Further studies may reveal whether prognosis is related to the type of TP53 mutations, as has been shown in other cancers. Thirdly, our analyses did not consider the non-coding genome due to the nature of the sequencing performed by the TCGA. Given the frequent mutations in non-coding regions such as TERT promoter, MALAT1 and NEAT1 (Fujimoto et al., 2012; Schulze et al., 2015), it is conceivable that additional oncogenic signatures within TP53-mutant HCCs may emerge.

Conclusion

Our study highlights the genetic heterogeneity among TP53-mutant HCCs and that patients with HCCs harboring different types of TP53 mutations may be associated with distinct prognoses. Future work will be required to elucidate whether the co-occurring genetic alterations act synergistically with TP53 mutations to promote carcinogenesis in HCCs.

References

Ahn, S. M., Jang, S. J., Shim, J. H., Kim, D., Hong, S. M., Sung, C. O., et al. (2014). Genomic portrait of resectable hepatocellular carcinomas: implications of RB1 and FGF19 aberrations for patient stratification. Hepatology 60, 1972–1982. doi: 10.1002/hep.27198

Alexandrov, L. B., Jones, P. H., Wedge, D. C., Sale, J. E., Campbell, P. J., Nik-Zainal, S., et al. (2015). Clock-like mutational processes in human somatic cells. Nat. Genet. 47, 1402–1407. doi: 10.1038/ng.3441

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A., Behjati, S., Biankin, A. V., et al. (2013). Signatures of mutational processes in human cancer. Nature 500, 415–421. doi: 10.1038/nature12477

Bosman, F. T., Carneiro, F., Hruban, R. H., and Theise, N. D. (2010). WHO Classification of Tumours of the Digestive System World Health Organization. Lyon: IARC. Google Scholar

Boyault, S., Rickman, D. S., De Reynies, A., Balabaud, C., Rebouissou, S., Jeannot, E., et al. (2007). Transcriptome classification of HCC is related to gene alterations and to new therapeutic targets. Hepatology 45, 42–52. doi: 10.1002/hep.21467

Bressac, B., Kew, M., Wands, J., and Ozturk, M. (1991). Selective G to T mutations of p53 gene in hepatocellular carcinoma from southern Africa. Nature 350, 429–431. doi: 10.1038/350429a0

Breuhahn, K., Vreden, S., Haddad, R., Beckebaum, S., Stippel, D., Flemming, P., et al. (2004). Molecular profiling of human hepatocellular carcinoma defines mutually exclusive interferon regulation and insulin-like growth factor II overexpression. Cancer Res. 64, 6058–6064. doi: 10.1158/0008-5472.CAN-04-0292

Chang, M. T., Asthana, S., Gao, S. P., Lee, B. H., Chapman, J. S., Kandoth, C., et al. (2016). Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. Nat. Biotechnol. 34, 155–163. doi: 10.1038/nbt.3391

Chiang, D. Y., Villanueva, A., Hoshida, Y., Peix, J., Newell, P., Minguez, B., et al. (2008). Focal gains of VEGFA and molecular classification of hepatocellular carcinoma. Cancer Res. 68, 6779–6788. doi: 10.1158/0008-5472.CAN-08-0742

Ciriello, G., Miller, M. L., Aksoy, B. A., Senbabaoglu, Y., Schultz, N., and Sander, C. (2013). Emerging landscape of oncogenic signatures across human cancers. Nat. Genet. 45, 1127–1133. doi: 10.1038/ng.2762

Cleary, S. P., Jeck, W. R., Zhao, X., Chen, K., Selitsky, S. R., Savich, G. L., et al. (2013). Identification of driver genes in hepatocellular carcinoma by exome sequencing. Hepatology 58, 1693–1702. doi: 10.1002/hep.26540

Edmondson, H. A., and Steiner, P. E. (1954). Primary carcinoma of the liver: a study of 100 cases among 48,900 necropsies. Cancer 7, 462–503. doi: 10.1002/1097-0142(195405)7:3<462::AID-CNCR2820070308>3.0.CO;2-E

Fujimoto, A., Furuta, M., Totoki, Y., Tsunoda, T., Kato, M., Shiraishi, Y., et al. (2016). Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. Nat. Genet. 48, 500–509. doi: 10.1038/ng.3547

Fujimoto, A., Totoki, Y., Abe, T., Boroevich, K. A., Hosoda, F., Nguyen, H. H., et al. (2012). Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. Nat. Genet. 44, 760–764. doi: 10.1038/ng.2291

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., et al. (2004). A census of human cancer genes. Nat. Rev. Cancer 4, 177–183. doi: 10.1038/nrc1299

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci. Signal. 6:l1. doi: 10.1126/scisignal.2004088

Gao, J., Chang, M. T., Johnsen, H. C., Gao, S. P., Sylvester, B. E., Sumer, S. O., et al. (2017). 3D clusters of somatic mutations in cancer reveal numerous rare mutations as functional targets. Genome Med. 9:4. doi: 10.1186/s13073-016-0393-x

Goossens, N., Sun, X., and Hoshida, Y. (2015). Molecular classification of hepatocellular carcinoma: potential therapeutic implications. Hepat. Oncol. 2, 371–379. doi: 10.2217/hep.15.26

Gouas, D., Shi, H., and Hainaut, P. (2009). The aflatoxin-induced TP53 mutation at codon 249 (R249S): biomarker of exposure, early detection and target for therapy. Cancer Lett. 286, 29–37. doi: 10.1016/j.canlet.2009.02.057

Guichard, C., Amaddeo, G., Imbeaud, S., Ladeiro, Y., Pelletier, L., Maad, I. B., et al. (2012). Integrated analysis of somatic mutations and focal copy-number changes identifies key genes and pathways in hepatocellular carcinoma. Nat. Genet. 44, 694–698. doi: 10.1038/ng.2256

Hall, P. A., and McCluggage, W. G. (2006). Assessing p53 in clinical contexts: unlearned lessons and new perspectives. J. Pathol. 208, 1–6. doi: 10.1002/path.1913

Hoshida, Y., Nijman, S. M., Kobayashi, M., Chan, J. A., Brunet, J. P., Chiang, D. Y., et al. (2009). Integrative transcriptome analysis reveals common molecular subclasses of human hepatocellular carcinoma. Cancer Res. 69, 7385–7392. doi: 10.1158/0008-5472.CAN-09-1089

Hsu, I. C., Metcalf, R. A., Sun, T., Welsh, J. A., Wang, N. J., and Harris, C. C. (1991). Mutational hot spot in the p53 gene in human hepatocellular carcinomas. Nature 350, 427–428. doi: 10.1038/350427a0

Jhunjhunwala, S., Jiang, Z., Stawiski, E. W., Gnad, F., Liu, J., Mayba, O., et al. (2014). Diverse modes of genomic alteration in hepatocellular carcinoma. Genome Biol. 15, 436. doi: 10.1186/s13059-014-0436-9

Jordan, J. J., Inga, A., Conway, K., Edmiston, S., Carey, L. A., Wu, L., et al. (2010). Altered-function p53 missense mutations identified in breast cancers can have subtle effects on transactivation. Mol. Cancer Res. 8, 701–716. doi: 10.1158/1541-7786.MCR-09-0442

Kan, Z., Zheng, H., Liu, X., Li, S., Barber, T. D., Gong, Z., et al. (2013). Whole-genome sequencing identifies recurrent mutations in hepatocellular carcinoma. Genome Res. 23, 1422–1433. doi: 10.1101/gr.154492.113

Kandoth, C., Mclellan, M. D., Vandin, F., Ye, K., Niu, B., Lu, C., et al. (2013). Mutational landscape and significance across 12 major cancer types. Nature 502, 333–339. doi: 10.1038/nature12634

Kawate, S., Fukusato, T., Ohwada, S., Watanuki, A., and Morishita, Y. (1999). Amplification of c-myc in hepatocellular carcinoma: correlation with clinicopathologic features, proliferative activity and p53 overexpression. Oncology 57, 157–163. doi: 10.1159/000012024

Kew, M. C. (2010). Epidemiology of chronic hepatitis B virus infection, hepatocellular carcinoma, and hepatitis B virus-induced hepatocellular carcinoma. Pathol. Biol. 58, 273–277. doi: 10.1016/j.patbio.2010.01.005

Kiani, C., Chen, L., Wu, Y. J., Yee, A. J., and Yang, B. B. (2002). Structure and function of aggrecan. Cell Res. 12, 19–32. doi: 10.1038/sj.cr.7290106

Lapke, N., Lu, Y. J., Liao, C. T., Lee, L. Y., Lin, C. Y., Wang, H. M., et al. (2016). Missense mutations in the TP53 DNA-binding domain predict outcomes in patients with advanced oral cavity squamous cell carcinoma. Oncotarget 7, 44194–44210. doi: 10.18632/oncotarget.9925

Lawrence, M. S., Stojanov, P., Mermel, C. H., Robinson, J. T., Garraway, L. A., Golub, T. R., et al. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. Nature 505, 495–501. doi: 10.1038/nature12912

Lee, J. S., Chu, I. S., Heo, J., Calvisi, D. F., Sun, Z., Roskams, T., et al. (2004). Classification and prediction of survival in hepatocellular carcinoma by gene expression profiling. Hepatology 40, 667–676. doi: 10.1002/hep.20375

Lin, C. P., Liu, C. R., Lee, C. N., Chan, T. S., and Liu, H. E. (2010). Targeting c-Myc as a novel approach for hepatocellular carcinoma. World J. Hepatol. 2, 16–20. doi: 10.4254/wjh.v2.i1.16

Ma, N. F., Hu, L., Fung, J. M., Xie, D., Zheng, B. J., Chen, L., et al. (2008). Isolation and characterization of a novel oncogene, amplified in liver cancer 1, within a commonly amplified region at 1q21 in hepatocellular carcinoma. Hepatology 47, 503–510. doi: 10.1002/hep.22072

Mahmoud, S. M., Paish, E. C., Powe, D. G., Macmillan, R. D., Grainge, M. J., Lee, A. H., et al. (2011). Tumor-infiltrating CD8+ lymphocytes predict clinical outcome in breast cancer. J. Clin. Oncol. 29, 1949–1955. doi: 10.1200/JCO.2010.30.5037

Martelotto, L. G., De Filippo, M. R., Ng, C. K., Natrajan, R., Fuhrmann, L., Cyrta, J., et al. (2015). Genomic landscape of adenoid cystic carcinoma of the breast. J. Pathol. 237, 179–189. doi: 10.1002/path.4573

Muller, P. A., and Vousden, K. H. (2014). Mutant p53 in cancer: new functions and therapeutic opportunities. Cancer Cell 25, 304–317. doi: 10.1016/j.ccr.2014.01.021

Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., et al. (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature 534, 47–54. doi: 10.1038/nature17676

Olivier, M., Hollstein, M., and Hainaut, P. (2010). TP53 mutations in human cancers: origins, consequences, and clinical use. Cold Spring Harb. Perspect. Biol. 2:a001008. doi: 10.1101/cshperspect.a001008

Olivier, M., Langerod, A., Carrieri, P., Bergh, J., Klaar, S., Eyfjord, J., et al. (2006). The clinical value of somatic TP53 gene mutations in 1,794 patients with breast cancer. Clin. Cancer Res. 12, 1157–1167. doi: 10.1158/1078-0432.CCR-05-1029

Ozcelik, H., Pinnaduwage, D., Bull, S. B., and Andrulis, I. L. (2007). Type of TP53 mutation and ERBB2 amplification affects survival in node-negative breast cancer. Breast Cancer Res. Treat. 105, 255–265. doi: 10.1007/s10549-006-9452-0

Peng, S. Y., Chen, W. J., Lai, P. L., Jeng, Y. M., Sheu, J. C., and Hsu, H. C. (2004). High alpha-fetoprotein level correlates with high stage, early recurrence and poor prognosis of hepatocellular carcinoma: significance of hepatitis virus infection, age, p53 and beta-catenin mutations. Int. J. Cancer 112, 44–50. doi: 10.1002/ijc.20279

Piscuoglio, S., Ng, C. K., Martelotto, L. G., Eberle, C. A., Cowell, C. F., Natrajan, R., et al. (2014). Integrative genomic and transcriptomic characterization of papillary carcinomas of the breast. Mol. Oncol. 8, 1588–1602. doi: 10.1016/j.molonc.2014.06.011

Rosenthal, R., Mcgranahan, N., Herrero, J., Taylor, B. S., and Swanton, C. (2016). deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biol. 17:31. doi: 10.1186/s13059-016-0893-4

Schulze, K., Imbeaud, S., Letouze, E., Alexandrov, L. B., Calderaro, J., Rebouissou, S., et al. (2015). Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. Nat. Genet. 47, 505–511. doi: 10.1038/ng.3252

Shiraishi, Y., Fujimoto, A., Furuta, M., Tanaka, H., Chiba, K., Boroevich, K. A., et al. (2014). Integrated analysis of whole genome and transcriptome sequencing reveals diverse transcriptomic aberrations driven by somatic genomic changes in liver cancers. PLOS ONE 9:e114263. doi: 10.1371/journal.pone.0114263

Soussi, T., Leroy, B., and Taschner, P. E. (2014). Recommendations for analyzing and reporting TP53 gene variants in the high-throughput sequencing era. Hum. Mutat. 35, 766–778. doi: 10.1002/humu.22561

Totoki, Y., Tatsuno, K., Covington, K. R., Ueda, H., Creighton, C. J., Kato, M., et al. (2014). Trans-ancestry mutational landscape of hepatocellular carcinoma genomes. Nat. Genet. 46, 1267–1273. doi: 10.1038/ng.3126

The Cancer Genome Atlas Research Network (2017). Comprehensive and integrative genomic characterization of hepatocellular carcinoma. Cell 169, 1327.e23–1341.e23. doi: 10.1016/j.cell.2017.05.046

Vegran, F., Rebucci, M., Chevrier, S., Cadouot, M., Boidot, R., and Lizard-Nacol, S. (2013). Only missense mutations affecting the DNA binding domain of p53 influence outcomes in patients with breast carcinoma. PLOS ONE 8:e55103. doi: 10.1371/journal.pone.0055103

Villanueva, A., and Hoshida, Y. (2011). Depicting the role of TP53 in hepatocellular carcinoma progression. J. Hepatol. 55, 724–725. doi: 10.1016/j.jhep.2011.03.018

Vogelstein, B., Lane, D., and Levine, A. J. (2000). Surfing the p53 network. Nature 408, 307–310. doi: 10.1038/35042675

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A. Jr., and Kinzler, K. W. (2013). Cancer genome landscapes. Science 339, 1546–1558. doi: 10.1126/science.1235122

Watkins, J., Weekes, D., Shah, V., Gazinska, P., Joshi, S., Sidhu, B., et al. (2015). Genomic complexity profiling reveals that hormad1 overexpression contributes to homologous recombination deficiency in triple-negative breast cancers. Cancer Discov. 5, 488–505. doi: 10.1158/2159-8290.CD-14-1092

Weinhold, N., Jacobsen, A., Schultz, N., Sander, C., and Lee, W. (2014). Genome-wide analysis of noncoding regulatory mutations in cancer. Nat. Genet. 46, 1160–1165. doi: 10.1038/ng.3101

Woo, H. G., Wang, X. W., Budhu, A., Kim, Y. H., Kwon, S. M., Tang, Z. Y., et al. (2011). Association of TP53 mutations with stem cell-like gene expression and survival of patients with hepatocellular carcinoma. Gastroenterology 140, 1063–1070. doi: 10.1053/j.gastro.2010.11.034

Yano, M., Hamatani, K., Eguchi, H., Hirai, Y., Macphee, D. G., Sugino, K., et al. (2007). Prognosis in patients with hepatocellular carcinoma correlates to mutations of p53 and/or hMSH2 genes. Eur. J. Cancer 43, 1092–1100. doi: 10.1016/j.ejca.2007.01.032

Acknowledgements

Going back to study has been at the same time challenging and extremely rewarding, bringing along positive emotions. I am very grateful to my supervisor Prof. Remo Sanges for his guidance, advice, and encouragement. Thanks for giving me this unique opportunity to work together. I gratefully acknowledge Prof. Luigi Terracciano for his friendship and for being my guidance in many occasions. I would like to thank also my many collaborators at the UniBasel, Novartis, and Biomedical Primate Research Centre (BPRC) for their scientific input. The research work on the malaria hypnozoites presented in Chapter III has been conducted with Dr. Thierry Diagana (Novartis) and Dr. Clemens Kochen (BPRC) and recently published in *eLife*. I thank Thierry and Clemens for being a great source of inspiration for my research work on malaria. The research work on TP53 described in Chapter IV, recently published in *Frontiers in Genetics*, stems from an effective collaboration with Dr. Salvatore Piscuoglio and Dr. Charlotte Ng (UniBasel): special thanks to these two bright scientists who are always ready to share their talent and knowledge with their collaborators. Last but not least, I am deeply grateful to my family for their love and their constant support throughout this period.