DATA NOTE

# REVISED PhenoPlasm: a database of disruption phenotypes for malaria parasite genes [version 2; referees: 2 approved]

Theo Sanderson, Julian C. Rayner

Malaria Programme, Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK

**v2** **First published:** 21 Jun 2017, **2**:45 (doi: 10.12688/wellcomeopenres.11896.1)
**Latest published:** 24 Jul 2017, **2**:45 (doi: 10.12688/wellcomeopenres.11896.2)

## Abstract

Two decades after the first *Plasmodium* transfection, attempts have been made to disrupt more than 3,151 genes in malaria parasites, across five *Plasmodium* species. While results from rodent malaria transfections have been curated and systematised, empowering large-scale analysis, phenotypic data from human malaria parasite transfections currently exists as individual reports scattered across a the literature. To facilitate systematic analysis of published experimental genetic data across *Plasmodium* species, we have built PhenoPlasm (http://www.phenoplasm.org), a database of phenotypes generated by transfection experiments in all *Plasmodium* parasites. The site provides a simple interface linking citation-backed *Plasmodium* reverse-genetic phenotypes to gene IDs. The database has been populated with phenotypic data on 367 *P. falciparum* genes, curated from 176 individual publications, as well as existing data on rodent *Plasmodium* species from RMgmDB and PlasmoGEM. This is the first time that all available data on *P. falciparum* transfection experiments has been brought together in a single place. These data are presented using ortholog mapping to allow a researcher interested in a gene in one species to see results across other *Plasmodium* species. The collaborative nature of the database enables any researcher to add new phenotypes as they are discovered. As an example of database utility, we use the currently available datasets to identify RAP (RNA-binding domain abundant in Apicomplexa)-domain containing proteins as crucial to parasite survival.

**Open Peer Review**

**Referee Status:** ✓ ✓

| | Invited Referees 1 | 2 |
|---|---|---|
| REVISED **version 2** published 24 Jul 2017 | | |
| **version 1** published 21 Jun 2017 | ✓ report | ✓ report |

1 **Tania F. de Koning-Ward**, Deakin University, Australia
**Natalie Counihan**, Deakin University, Australia

2 **Omar S. Harb**, University of Pennsylvania, USA

**Discuss this article**

Comments (0)

**Corresponding author:** Theo Sanderson (ts10@sanger.ac.uk)

**Author roles: Sanderson T**: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – Original Draft Preparation; **Rayner JC**: Funding Acquisition, Writing – Review & Editing

**Competing interests:** No competing interests were disclosed.

**How to cite this article:** Sanderson T and Rayner JC. **PhenoPlasm: a database of disruption phenotypes for malaria parasite genes [version 2; referees: 2 approved]** Wellcome Open Research 2017, **2**:45 (doi: 10.12688/wellcomeopenres.11896.2)



**Grant information:** This work was supported by the Wellcome Trust [098051].
*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

**First published:** 21 Jun 2017, **2**:45 (doi: 10.12688/wellcomeopenres.11896.1)

**REVISED** **Amendments from Version 1**

We have made a number of small improvements to the online database as a result of the reviewers' comments. In addition, we have made minor changes to the text to clarify meaning. The most significant change is that the domain-phenotype enrichment analysis has been repeated, now associating phenotypes with InterPro families as opposed to individual signatures. This does not significantly impact the results, but makes them easier to interpret. Figure 3 and Figure 4 have been updated to reflect additional phenotype data added since submission.

**See referee reports**

## Introduction

The increasing use of experimental genetics in *Plasmodium spp.* has provided numerous insights into the biology of the malaria parasite (de Koning-Ward *et al.*, 2015). Nevertheless, to date such results for *P. falciparum* transfection experiments are scattered across a range of journals, with no unified or queryable interface. This means that a researcher whose experiment or analysis identifies a set of genes of interest must devote considerable time to reviewing all available literature if they are to understand what is already known about these genes from previous knockout or other genetic manipulation experiments. To facilitate rapid functional profiling using already established phenotypes, we set out to build a database to contain this information.

There were three key functional requirements for such a database:

### Systematic, and synergistic with existing resources

To allow for automated bioinformatic analyses, it is crucial that the database have a defined, machine-comprehensible, schema for recording phenotypes. It is also important that this schema is compatible with existing resources. The rodent malaria genetically modified parasite database (RMgmDB, http://pberghei.eu; Khan *et al.*, 2013) provides a powerful curated resource for the rodent *Plasmodium* species, and contains curated data on disruption attempts for over 500 genes from individual studies, making it the largest manually curated database for *Plasmodium* experimental genetic data. However, this database does not contain any data for human-infecting *Plasmodium* species. While some human *Plasmodium* parasite genes lack rodent orthologs, nearly 75% have such orthologs, and integrating human and rodent *Plasmodium* phenotypes is likely to be highly informative. To allow such integration, any new database schema must be compatible with that of RMgmDB, which is broken into 6 different stages at which phenotypes can occur (asexual, gametocyte/gamete, fertilization & ookinete, oocyst, sporozoite and liver). RMgmDB also distinguishes cases in which a modification is not successful, which provide some implication of a possible role in asexual growth; we decided to call this quality *mutant viability*, though of course failure to obtain a mutant might also result from a technical failure. The database must also import phenotypes from the largest source of blood stage *Plasmodium* phenotyping data available to date, PlasmoGEM barcode-sequencing experiments (Bushell *et al.*, 2017).

### Orthology-based retrieval

The use of model systems has strongly facilitated attempts to understand human malaria parasites (Zuzarte-Luis *et al.*, 2014). These systems are valuable both because rodent malaria parasites are less technically challenging to genetically modify, and because their *in vivo* nature has allowed the study of some aspects of parasite biology in a more physiological setting than provided by *in vitro* culture. Critically, rodent models allow the recapitulation of the complete lifecycle with few technical hurdles, and therefore the vast majority of known non-blood-stage phenotypes come from rodent *Plasmodium* experiments. Nevertheless, rodent parasites do not contain orthologs of every *P. falciparum* gene, so these studies alone cannot provide a complete view of the parasite genetics causing human disease. Even where orthologs exist, phenotypes may not always be conserved, although available comparative data does suggest a high level of conservation (Bushell *et al.*, 2017).

We felt that an optimal approach would allow a researcher to search for gene IDs from any species but at a glance to see both results in this organism and for orthologous genes in other *Plasmodium* species. The database should also contain records for emerging new genetic models, such as *P. knowlesi* (Kocken *et al.*, 2002; Moon *et al.*, 2013), as well as *P. vivax* which, though currently genetically intractable, is of key medical importance, and where gene functions may be interpreted through orthology.

### Community contribution

The role for "crowd-sourcing" in biological databases is contentious (Good & Su, 2013; Karp, 2016). It is clear that community contributions cannot wholly replace curation for these types of datasets, but on the other hand manual curation is not easy to support through application to funding agencies, and suffers the problem of scale - a single person is unlikely to identify every single phenotype in the *Plasmodium* literature. An example of successful community contribution in parasite genetics comes from the EuPathDB databases (which include PlasmoDB, Aurrecoechea *et al.*, 2009). These have had thousands of user comments, many of which are now incorporated into annotations. We felt it important to provide a mechanism whereby a motivated researcher can add any missing phenotypes to the database. This requires the creation of an intuitive interface and easy interface for the entry of data, and means that the primary data source must be publications, with each phenotype backed up with a PubMed ID.

## Methods

### Database schema

The database comprises 4 main data tables. Phenotype data is stored in a "Phenotypes" table, containing the stage at which the phenotype is described, the phenotype itself (selected from a growing and defined taxonomy), a referenceable citation, details of the genetic system used to obtain the phenotype and any additional notes. Here, the gene itself is a reference to the "Genes" table, containing gene name and product data imported from PlasmoDB (Aurrecoechea *et al.*, 2009). Genes are linked to previous aliases by an "Aliases" table. They are also linked to one another by the "Orthology" table, which contains links between genes in which both OrthoMCL group and synteny is conserved.

### Display

The database can be queried either for a set of genes (Figure 1) or a single gene (Figure 2). The former provides a table with one line per gene, while the latter provides referencing for each claim and displays any additional notes. The search box on each page can be flexibly queried with a gene ID, symbol or description; but there is also an advanced search facility which allows the retrieval of, for example, only phenotypes backed up by evidence from conditional systems.

### Literature review

A scan was made of the *Plasmodium* literature using Google Scholar (which provides full-text search for a large proportion of publications) to identify reported attempts at *P. falciparum* gene disruption for curation. Terms for which complete literature scans were made included '"attempts to disrupt" falciparum', '"gene disruption of" falciparum' '"gene deletion construct" falciparum' and '"gene disruption construct" falciparum'. Numerous additional terms were used, and the first 10 pages of results for each search were manually curated and added to the database. In addition, genes with a suggested role in erythrocyte invasion were systematically curated by searching for all references to any version of their gene IDs, as discussed below.

One challenge when conducting literature searches into *Plasmodium* proteins is the fact that the numerous iterative improvements made to *Plasmodium* genome assemblies mean that a gene could be referred to by any of numerous current or historic gene IDs. To assist with this, the PhenoPlasm page for each gene contains a link to conduct a custom boolean search on Google Scholar, searching article full text for any of the historic gene identifiers which have been historically used to refer to the gene, as provided by PlasmoDB (Aurrecoechea *et al.*, 2009). Links are also provided to other databases, including searches for pathways and localization images on the Malaria Metabolic Pathways Site.

In addition to this curation, scripts were developed to regularly import data from RMgmDB and PlasmoGEM and transform it to the PhenoPlasm schema.

### Enrichment analysis

Genome-scale phenotyping data provides opportunities to integrate diverse genome-wide data sets and investigate how they relate to gene functionality. The PlasmoGEM dataset, which currently contains data for >50% of *P. berghei* genes, has been used to identify essential metabolic pathways and investigate the relationship between transcriptomics, evolution and phenotype (Bushell *et al.*, 2017).

To supplement these analyses, and further illustrate the utility of genome-scale phenotype data, we sought to identify protein domains whose presence in a gene was predictive of essentiality

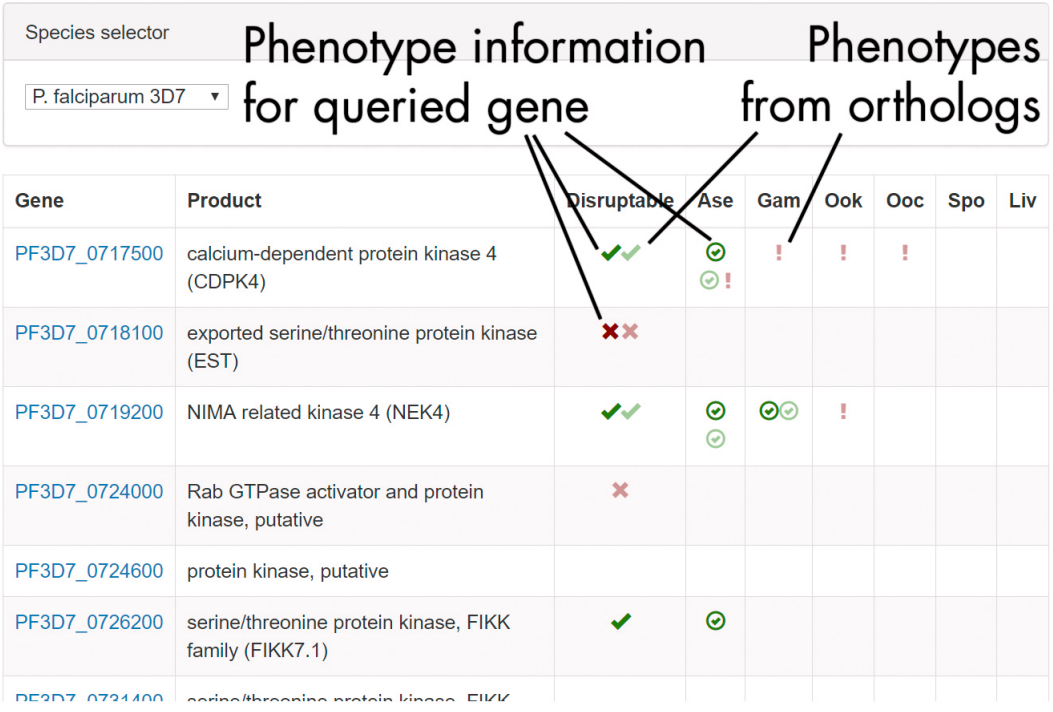


**Figure 1. The results of a search for 'kinase' genes, showing phenotype data, both from *P. falciparum* experiments and those in rodent models across multiple lifecycle stages.** Green ticks indicate mutant viability, and circled green ticks indicate wild-type phenotype. Red crosses indicate failure to disrupt the gene, and red exclamation marks indicate a phenotype different from wildtype. The icons are either shown in full opacity (indicating they apply to the gene in the species queried) or semi-transparent (indicating they refer to orthologous genes in other species).

or dispensability. We downloaded from PhenoPlasm the phenotypes relating to all *P. falciparum* genes, both directly assayed and inferred from orthologs, and added annotation information for InterPro signatures (Aurrecoechea *et al.*, 2009). We then used hypergeometric testing to identify signatures with members significantly enriched in essential or dispensable genes (Supplementary File S1).

## Results

### The extent of data now available on PhenoPlasm

At the time of manuscript preparation, some form of phenotyping information is available for 3,188 genes (Figure 3). Of these, 2,790 are from rodent malaria parasites, and so represent data imported from RMgmDB and PlasmoGEM. The remaining 398 are human parasite genes with phenotypes systematised by our curation,

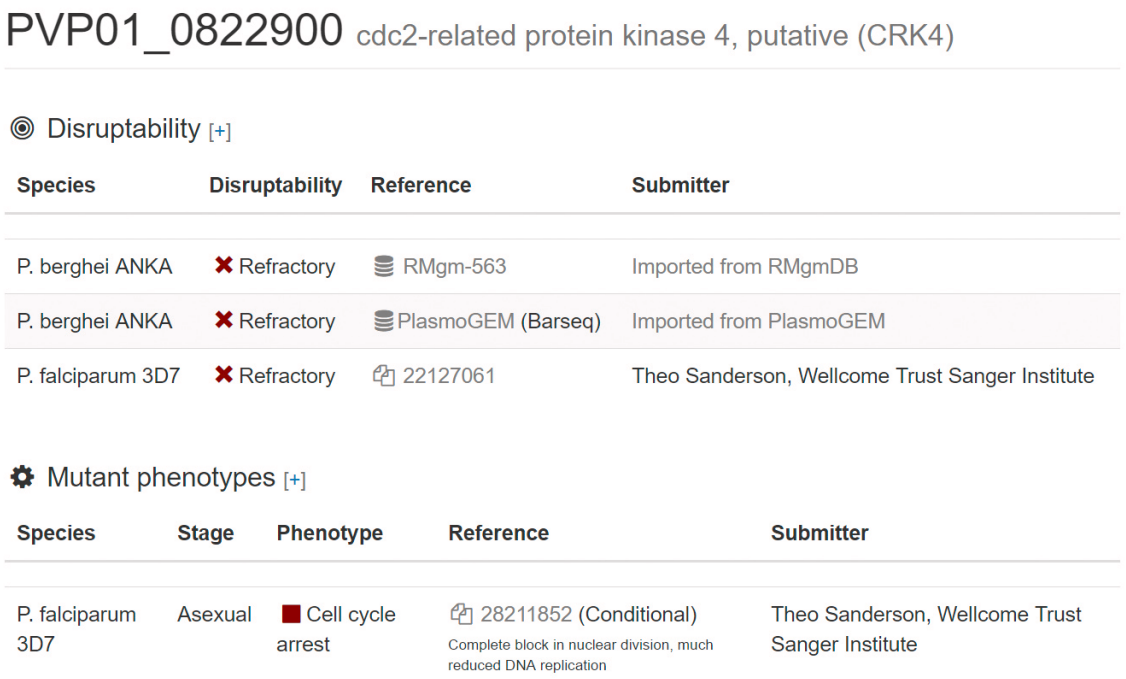


**Figure 2. The phenotype page for the *P. vivax* CRK4 gene.** Though no experimental data is available directly from *P. vivax*, published results are shown from *P. berghei* and *P. falciparum*, with references to the original datasets from which likely data in *P. vivax* could be inferred. This gene is essential and has therefore been refractory to all attempts to disrupt it by classical reverse genetics, but a conditional system has also been recently applied in *P. falciparum*, allowing a more detailed phenotype to be assigned to the gene from our taxonomy.

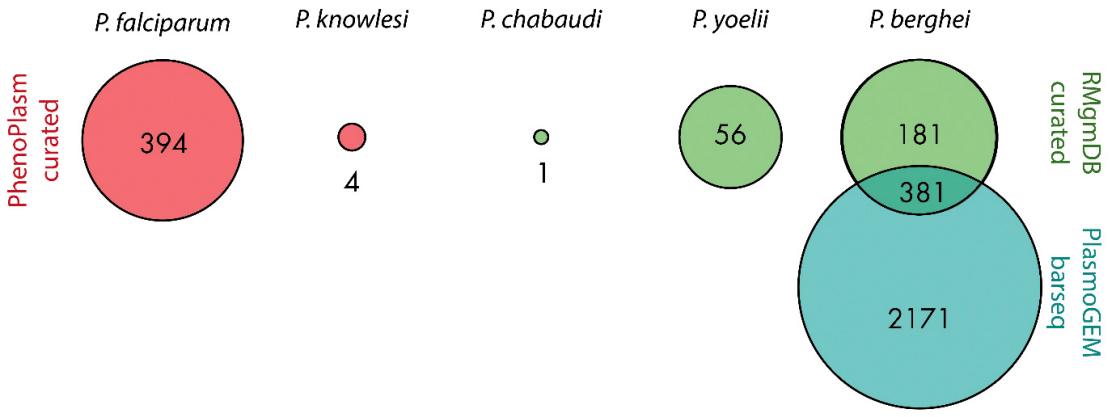


**Figure 3. The number of genes with phenotyping data available in PhenoPlasm for each *Plasmodium* species, and the source of these annotations.**

and brought together in a searchable format for the first time. For posterity, the complete data has been additionally deposited on Figshare (https://doi.org/10.6084/m9.figshare.5114017), and will be updated at least yearly.

There are 7,274 total phenotype datapoints (i.e. data for one life-stage, for one gene knock-out, in one study). The majority of the non-blood stage phenotype datapoints are from the rodent parasites, since relatively few *P. falciparum* genes have phenotyping data reported beyond the blood stage.

Given that some genes have been phenotyped in multiple *Plasmodium* species, the number of ortholog groups covered in at least one species is 2,778. This represents 60% of the core *Plasmodium* genome.

Since every phenotype in PhenoPlasm for the human malaria parasites is linked to a PubMed ID or other citation, we were able to informatically extract the dates associated with these publications and plot how the number of genes phenotyped in *P. falciparum* has increased over time (Figure 4). While this analysis has limitations (a portion of citations are review papers), it does reveal that progress remains slow in human-infecting parasites, and with no major acceleration apparent in the last decade. This illustrates the current importance of rodent models, and raises the question of the technologies needed to create a step-change in the rate of phenotype discovery for human malaria species.

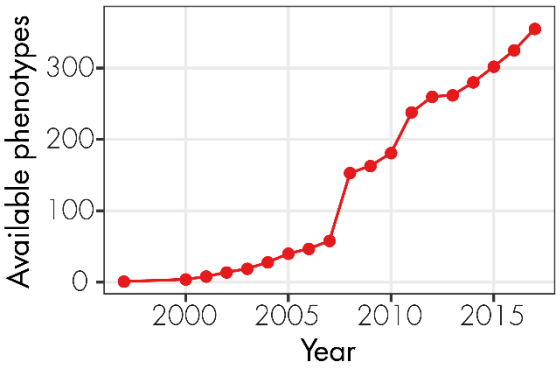


**Figure 4. A timeline of PubMed dates associated with publications reporting knock-out phenotypes for *P. falciparum* genes, from the first gene disruption in 1997 to mid-2017.** The values shown are cumulative from all previous years. Around 25 genes per year have had disruption attempts reported since the year 2000. The spikes that occur in 2008 and 2011 largely represent two individual publications systematically knocking out exported genes (Maier *et al.*, 2008) and kinases (Solyakov *et al.*, 2011) respectively.

### Applying the data to investigate genome-wide relationships between protein structure and knock-out phenotype reveals the importance of RAP-domain containing proteins

The InterPro signatures most enriched in genes producing viable mutants included the 6-cysteine domain (12/12 viable, p=7.46E-05), the MSP7 C-terminal domain (8/8 viable, p=0.0018) and the MFS transporter superfamily (21/24 viable, p=2.19E-05). These data confirm previous inferences, but does so in a systematic way for the first time (Supplementary Table S1).

The InterPro signatures enriched in apparently essential genes include expected results, such as the OB-fold nucleic acid domain and ribosomal protein S5 domain-2 type folds, but also identify the RAP (RNA-binding domain abundant in Apicomplexa) domain as functionally highly significant (Table 1, Supplementary Table S1). The RAP domain was named for its dramatic expansion in the Apicomplexa, as compared to other parts of the tree of life (Lee & Hong, 2004). Every one of the ten genes containing this domain for which knockouts have been attempted to date appears to be essential (there are 18–19 in most *Plasmodium* genomes). As a result of this analysis, we looked at the other apicomplexan taxon for which genome-scale data is available and found that all eleven proteins containing this domain in *Toxoplasma gondii* also have suggestions of essentiality in CRISPR-screening data (Sidik *et al.*, 2016). This functional data, combined with recent experimental observations of these proteins binding mRNA (Bunnik *et al.*, 2016), suggest these proteins as prominent candidates for future studies which may uncover a new realm of Apicomplexa-specific biology.

**Table 1. InterPro signatures most enriched in essential genes.**

| InterPro Family | Description | Viable mutant | Essential gene | Proportion of genes essential | p-value |
|---|---|---|---|---|---|
| **IPR012340** | Nucleic acid-binding, OB-fold | 3 | 24 | 89% | 0.0002 |
| **IPR020568** | Ribosomal protein S5 domain 2-type fold | 2 | 18 | 90% | 0.0008 |
| **IPR013584** | RAP domain | 0 | 10 | 100% | 0.0023 |
| **IPR006073** | GTP binding domain | 0 | 9 | 100% | 0.0043 |
| **IPR005225** | Small GTP-binding protein domain | 4 | 19 | 83% | 0.0049 |

## Discussion

A rapid growth in available genetic tools, coupled with decreasing costs of gene synthesis and sequencing, mean that *Plasmodium* experimental genetics is reaching the genome-level scale for the first time. Phenotypic data from these studies has the potential to shed light on the importance of the novel genes found in these early-branching eukaryotes. The development of large scale genetic modification programmes in *Plasmodium* species (Bronner *et al.*, 2016; Bushell *et al.*, 2017; Gomes *et al.*, 2015) is now shedding light on a large portion of the genome's functional importance in the asexual blood stage.

Nevertheless, no single approach is likely to reach saturation for some time, and exploring the complete parasite lifecycle in any system is likely to take even longer. In addition, the lack of a non-homologous end-joining pathway in *Plasmodium* parasites prevents the use of the conventional CRISPR-Cas9 screens, which have revolutionized genetics in other organisms (Sidik *et al.*, 2016). For these reasons, a complete view of the reverse-genetic landscape for a gene or pathway will require bringing together multiple datasets with the individual gene-by-gene studies that have characterized decades of research.

Until now, retrieving phenotyping data for a set of 80 genes in *Plasmodium* might have involved perhaps a day of work, requiring a separate literature search for each gene's *P. falciparum* ID, and all of its historic identifiers. To be comprehensive, the set of genes would additionally have to be transformed by orthology into each of the five available Plasmodium genetic systems, with further searches conducted. With the development of PhenoPlasm, all this data is available in a single batch search.

We hope that this database will assist in the prioritization of future large-scale studies, eliminating duplication of existing efforts, and allowing a focus on the portion of the genome which remains wholly unexplored by previous reverse genetic approaches. The availability of systematized data should allow the *Plasmodium* phenome to be bioinformatically queried in the same sort of routine way that transcriptomic data is used today to understand gene function. The application we present here, identifying RAP domain proteins as crucial for parasite survival, is just a hint of the wealth of information that well-organized phenotypic data can reveal at scale.

## Data availability

The database is accessible at http://phenoplasm.org/. Facilities are provided for download of batch searches in CSV form, and of the entire dataset. In addition, snapshots are available from Figshare (https://doi.org/10.6084/m9.figshare.5114017) which will be updated yearly.

### Competing interests

No competing interests were disclosed.

### Grant information

This work was supported by the Wellcome Trust [098051].

*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

## Supplementary material

**Supplementary Table S1: Full table of InterPro signatures, and the number of viable and non-viable mutants for proteins containing them, with p-values indicating statistical enrichment.**

Click here to access the data.

**Supplementary File S1: Reproducible analysis.** The analytical methods used to calculate the results presented here are preserved in an R-Markdown document to allow full reproduction of the results.

Click here to access the data.

## References

Aurrecoechea C, Brestelli J, Brunk BP, *et al.*: **PlasmoDB: a functional genomic database for malaria parasites.** *Nucleic Acids Res.* 2009; **37**(Database issue): D539–43.
PubMed Abstract | Publisher Full Text | Free Full Text

Bronner IF, Otto TD, Zhang M, *et al.*: **Quantitative insertion-site sequencing (QIseq) for high throughput phenotyping of transposon mutants.** *Genome Res.* 2016; **26**(7): 980–989.
PubMed Abstract | Publisher Full Text | Free Full Text

Bunnik EM, Batugedara G, Saraf A, *et al.*: **The mRNA-bound proteome of the human malaria parasite *Plasmodium falciparum*.** *Genome Biol.* 2016; **17**(1): 147.
PubMed Abstract | Publisher Full Text | Free Full Text

Bushell E, Gomes AR, Sanderson T, *et al.*: **Functional profiling of a Plasmodium genome shows a high incidence of essential genes in an intracellular parasite.** *Cell.* 2017.

de Koning-Ward TF, Gilson PR, Crabb BS: **Advances in molecular genetic**

**systems in malaria.** *Nat Rev Microbiol.* 2015; **13**(6): 373–387.
PubMed Abstract | Publisher Full Text

Gomes AR, Bushell E, Schwach F, *et al.*: **A genome-scale vector resource enables high-throughput reverse genetic screening in a malaria parasite.** *Cell Host Microbe.* 2015; **17**(3): 404–413.
PubMed Abstract | Publisher Full Text | Free Full Text

Good BM, Su AI: **Crowdsourcing for bioinformatics.** *Bioinformatics.* 2013; **29**(16): 1925–1933.
PubMed Abstract | Publisher Full Text | Free Full Text

Karp PD: **Crowd-sourcing and author submission as alternatives to professional curation.** *Database (Oxford).* 2016; **2016**: pii: baw149.
PubMed Abstract | Publisher Full Text | Free Full Text

Khan SM, Kroeze H, Franke-Fayard B, *et al.*: **Standardization in generating and reporting genetically modified rodent malaria parasites: the RMgmDB database.** *Methods Mol Biol.* 2013; **923**: 139–50.
PubMed Abstract | Publisher Full Text

Kocken CH, Ozwara H, van der Wel A, *et al.*: ***Plasmodium knowlesi* provides a rapid *in vitro* and *in vivo* transfection system that enables double-crossover gene knockout studies.** *Infect Immun.* 2002; **70**(2): 655–660.
PubMed Abstract | Publisher Full Text | Free Full Text

Lee I, Hong W: **RAP--a putative RNA-binding domain.** *Trends Biochem Sci.* 2004; **29**(11): 567–570.
PubMed Abstract | Publisher Full Text

Maier AG, Rug M, O'Neill MT, *et al.*: **Exported proteins required for virulence and rigidity of *Plasmodium falciparum*-infected human erythrocytes.** *Cell.* 2008; **134**(1): 48–61.
PubMed Abstract | Publisher Full Text | Free Full Text

Moon RW, Hall J, Rangkuti F, *et al.*: **Adaptation of the genetically tractable malaria pathogen *Plasmodium knowlesi* to continuous culture in human erythrocytes.** *Proc Natl Acad Sci U S A.* 2013; **110**(2): 531–6.
PubMed Abstract | Publisher Full Text | Free Full Text

Sidik SM, Huet D, Ganesan SM, *et al.*: **A Genome-wide CRISPR Screen in *Toxoplasma* Identifies Essential Apicomplexan Genes.** *Cell.* 2016; **166**(6): 1423–1435.e12.
PubMed Abstract | Publisher Full Text | Free Full Text

Solyakov L, Halbert J, Alam MM, *et al.*: **Global kinomic and phospho-proteomic analyses of the human malaria parasite *Plasmodium falciparum*.** *Nat Commun.* 2011; **2**: 565.
PubMed Abstract | Publisher Full Text

Zuzarte-Luis V, Mota MM, Vigário AM: **Malaria infections: What and how can mice teach us.** *J Immunol Methods.* 2014; **410**: 113–122.
PubMed Abstract | Publisher Full Text

# Open Peer Review

## Current Referee Status: ✔ ✔

**Version 1**

Referee Report 03 July 2017

**doi:**10.21956/wellcomeopenres.12855.r23827

✔ **Omar S. Harb**
Institute for Biomedical Informatics, Perelman School of Medicine, Department of Biology, University of Pennsylvania, Philadelphia, PA, USA

This is a nice paper describing the PhenoPlasm database. This database is unique in that in collects information about *Plasmodium falciparum* genetic modifications in one place and enables users to search the database using both systematic gene IDs and text search. Overall I believe this paper describes a very useful resource and will help scientists more easily access data about various types of mutants in *Plasmodium*. I have a number of minor suggestions and comments that should make this resource even more useful:

1. In this sentence “Nevertheless, rodent parasites do not contain orthologs of every *P. falciparum* gene, so these studies alone cannot provide a complete view of the parasite genetics **underlying causing** human disease.”, The words underlying causing right after each other don’t make sense. I think one of these words should be dropped.

2. Community contribution section – it would be useful to highlight the success of the community user comment system that has been in place at EuPathDB databases for over 10 years. Thousands of crowd-sourced user comments on genes have been entered and many of these have been used to guide curation teams to improve and edit annotation. This would serve as a nice example of how crowd sourcing works even with scientists!

3. “Genes are linked to previous aliases by an “Aliases” table, and to each other by orthology by the “Orthology” table, which contains links where both OrthoMCL group and synteny is conserved.” This sentence sounds a bit awkward – authors should think about rephrasing.

4. Once a search is run, a table of disrupted genes is reported. It would be useful if the number of rows (or genes) in the table is reported at the top of the page.

5. Results table: it would be nice if the header of the table is sticky (or locked) so when a user scrolls down they can still see the column names. I found myself having to keep scrolling up and down.

6. Once you run a search there the table contains various symbols eg. Check marks and exclamation marks. These also differ in color intensity/opacity. It wasn’t clear to me what these mean and I could not find a description of what they mean on the site. Maybe I missed it? Regardless it would be good if those can be provided on mouse over or on the site of the table as a legend. Perhaps taking the figure legend from figure 1 in the paper and making available to all table results on the

database.

7. How robust was the google scholar search to identify gene id aliases? Both GeneDB and PlasmoDB keep a list of aliases as well. Did you use these? What about genbank or UniProt accession numbers, have those been included as aliases?

8. In the results section the number or genes reported with information in PhenoPlasm differs slightly than what it on the website. I am assuming this is due to ongoing curation which is great! It would be good to state something like “At the time of manuscript preparation…”

9. “For posterity, the complete data has been additionally deposited on Figshare (https://doi.org/10.6084/m9. gshare.5114017), and will be regularly updated.” Can you define what regularly updated means? Once a month, once a year? How will this be done? Is there a mechanism in place to do this? How important is this anyway?

10. Data presented in Figure 4 is cute but not sure how much it adds to the paper. I am not against it, I just don’t think it adds anything useful.

11. Literature review using google scholar: I think it is important to provide a full list of terms used as a supplemental table. Also, it might even be better to provide a link to the advanced google scholar search so readers can try them out for themselves.

12. Manual curation: Needs more description as to how this was conducted. Was it only based on google scholar or did it also involved reading of papers. How many individuals were involved in reading papers and was there any attempt to quality control the results?

13. You should be able to extract additional mutant data from user comments in PlasmoDB. I just did a few quick searches showing that there are some in PlasmoDB not in PhenoPlasm but perhaps this can be a community curation/contribution: http://plasmodb.org/plasmo/im.do?s=94dda7211e47d33c

**Is the rationale for creating the dataset(s) clearly described?**
Yes

**Are the protocols appropriate and is the work technically sound?**
Yes

**Are sufficient details of methods and materials provided to allow replication by others?**
Partly

**Are the datasets clearly presented in a useable and accessible format?**
Yes

***Competing Interests:*** For full disclosure I work for the Eukaryotic Pathogen Databases which includes PlasmoDB. This database was used and is mentioned in the paper.

**I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Author Response 24 Jul 2017

**Theo Sanderson**, Wellcome Trust Sanger Institute, UK

We are very grateful for this rapid review and the helpful suggestions you have made. We have responded inline below for clarity.

*This is a nice paper describing the PhenoPlasm database. This database is unique in that in collects information about Plasmodium falciparum genetic modifications in one place and enables users to search the database using both systematic gene IDs and text search. Overall I believe this paper describes a very useful resource and will help scientists more easily access data about various types of mutants in Plasmodium. I have a number of minor suggestions and comments that should make this resource even more useful:*

*In this sentence "Nevertheless, rodent parasites do not contain orthologs of every P. falciparum gene, so these studies alone cannot provide a complete view of the parasite genetics underlying causing human disease.", The words underlying causing right after each other don't make sense. I think one of these words should be dropped.*

Thank you. This has been fixed.

*Community contribution section – it would be useful to highlight the success of the community user comment system that has been in place at EuPathDB databases for over 10 years. Thousands of crowd-sourced user comments on genes have been entered and many of these have been used to guide curation teams to improve and edit annotation. This would serve as a nice example of how crowd sourcing works even with scientists!*

This is a great point that we have put into the text. We are very keen on allowing community contribution, which is the reason we provide the facility, and it is great to have an example to draw on of how this can work in practice.

*"Genes are linked to previous aliases by an "Aliases" table, and to each other by orthology by the "Orthology" table, which contains links where both OrthoMCL group and synteny is conserved." This sentence sounds a bit awkward – authors should think about rephrasing.*

We have now done so.

*Once a search is run, a table of disrupted genes is reported. It would be useful if the number of rows (or genes) in the table is reported at the top of the page.*

Thank you for this suggestion. This has been added and indeed improves the interface.

*Results table: it would be nice if the header of the table is sticky (or locked) so when a user scrolls down they can still see the column names. I found myself having to keep scrolling up and down.*

This is a helpful point that we do accept. It currently poses technical challenges. As an intermediate solution we have added mouseover tooltip text to the icons which indicates which stage they correspond to as well as what the icon represents. We hope to make the column headers stick at a future date.

*Once you run a search there the table contains various symbols eg. Check marks and exclamation marks. These also differ in color intensity/opacity. It wasn't clear to me what these mean and I could not find a description of what they mean on the site. Maybe I missed it? Regardless it would be good if those can be provided on mouse over or on the site of the table as a legend. Perhaps taking the figure legend from figure 1 in the paper and making available to all table results on the database.*

This is a great point raised by both reviewers which we have now implemented. There is a page listing these on the site but this was somewhat buried.

*How robust was the google scholar search to identify gene id aliases? Both GeneDB and PlasmoDB keep a list of aliases as well. Did you use these? What about genbank or UniProt accession numbers, have those been included as aliases?*

We have clarified this section. The feature uses the PlasmoDB previous IDs (which are now acknowledged in the text) and provides a simple mechanism to conduct a search for "PF3D7_123456 OR MAL13P.123456 OR …." to identify references to the gene. GenBank and UniProt accession numbers have not been included at this time. We believe relatively few knock-out studies use such identifiers.

*In the results section the number or genes reported with information in PhenoPlasm differs slightly than what it on the website. I am assuming this is due to ongoing curation which is great! It would be good to state something like "At the time of manuscript preparation…"*

Your explanation is exactly correct, and we have implemented this suggestion. Thank you.

*"For posterity, the complete data has been additionally deposited on Figshare (https://doi.org/10.6084/m9. gshare.5114017), and will be regularly updated." Can you define what regularly updated means? Once a month, once a year? How will this be done? Is there a mechanism in place to do this? How important is this anyway?*

We were aware that a common concern upon the creation of databases is that should funding run out the database will become defunct and inaccessible, and this was our attempt to defuse such concerns by ensuring the data would be available in perpetuity regardless. We have clarified that this will take place at least yearly. We do not feel that this a major concern for *our* database, but then we suspect that the creators of now-defunct databases did not either, so this seems a sensible insurance policy. This process is currently manual.

*Data presented in Figure 4 is cute but not sure how much it adds to the paper. I am not against it, I just don't think it adds anything useful.*

We think this history is (somewhat) interesting, and are keen to make it publically available. We had no clear idea what this graph would like until we plotted it, because systematised data is needed to do so. We do accept it is a very minor application of the database which will only be useful if (for instance) cited in a review paper.

*Literature review using google scholar: I think it is important to provide a full list of terms used as a supplemental table. Also, it might even be better to provide a link to the advanced google scholar search so readers can try them out for themselves.*

As in response to the other reviewers, we do accept these critiques. Unfortunately, we do not have systematic records of all terms used. As we discuss below, and in response to the other reviewers, further searches yield very few additional genes, giving us confidence that our review has at least been thorough.

*Manual curation: Needs more description as to how this was conducted. Was it only based on google scholar or did it also involved reading of papers. How many individuals were involved in reading papers and was there any attempt to quality control the results?*

This curation generally involved reading the papers in question, with rare exceptions where results were very clear from the abstract. All were read by a single person with a handful of exceptions which are the result of community contributions. One way in which results have been quality-controlled (though this was not the primary motivation) has been looking for conflicts between our curated P. falciparum data and PlasmoGEM data (Bushell et al., 2017). These should be enriched for curatorial errors but did not uncover any. Nevertheless there are almost certainly curatorial errors in the database, which we hope will be resolved with the eyes of the community, given that publications are the ultimate source authority.

If you refer to attempts to control for the quality of the source data, in general we did not consider this to be necessary.

*You should be able to extract additional mutant data from user comments in PlasmoDB. I just did a few quick searches showing that there are some in PlasmoDB not in PhenoPlasm but perhaps this can be a community curation/contribution:*
*http://plasmodb.org/plasmo/im.do?s=94dda7211e47d33c*

This is a great suggestion, and thank you for providing the search strategy. We have worked through this list and added a number of datapoints. Reassuringly these are mostly cases of insertional mutagenesis or spontaneous deletions, which we did not conduct specifically conduct searches for in putting together the database.

***Competing Interests:*** No competing interests were disclosed.

Referee Report 29 June 2017

**doi:**10.21956/wellcomeopenres.12855.r23692

**Tania F. de Koning-Ward** , **Natalie Counihan**
School of Medicine, Deakin University, Waurn Ponds, VIC , Australia

The journal article by Sanderson and Rayner presents a new database termed ‘PhenoPlasm’ that provides a combined database of reverse genetics information extracted from the rodent *Plasmodium* databases RMgmDB and PlasmoGEM and manual curation of phenotype data from publications that have described reverse genetic experiments in *Plasmodium falciparum*.

This new database will provide an excellent resource for malaria researchers as it will enable them to very rapidly assess whether attempts to disrupt/knockdown a particular gene(s) in any of the various species

has already been attempted and if so what the resulting phenotypes are, without having to transform data by orthology. We actually recommend that the database link is included in the abstract.

After visiting the database, we noticed there was some additional and very useful features that are not described in the manuscript and we recommend that they are included. For example, the subcellular localisation data and the maps. This puts so much more at the fingertips of researchers. It is ambiguous as to what 'maps' refer to until you click on the link so we do recommend that this is called something else, such as 'pathways protein mapped to'.

We also recommend that when a search is undertaken that a key of the symbols is visible on the same page for greater clarity. This is clear in the manuscript but not at all evident when using the database.

Figure 1: For CDPK, there is a circled pale green tick indicating the knockout has a wildtype phenotype. However, next to this is an exclamation mark, meaning the phenotype is different from wildtype. Isn't the circled green tick and exclamation mark mutually exclusive?

Figure 2: The word 'conditional' is embedded in the text. We were wondering if it would be clearer to instead have an additional column where it could be stated what experimental approach has been used (ie. whether the mutant phenotype data is derived from a gene knockout or conditional knockout/knockdown, etc). For an example, if you search Pf3D7_0929400, the disruptability table indicates the gene is refractory to deletion in *P. falciparum* but in the mutant phenotypes table, the phenotype is ! Attenuated. This is somewhat confusing as parasites are not attenuated until after knockdown.

Some of the terms that were used in the Literature review section are indicated. However, it would be good to know what the numerous additional terms used were – for example, terms like 'knockout', 'knockdown', 'depletion', 'conditional' may have picked up additional publications. For example, HSP101 has been knocked out in *P. berghei,* hence why the cross is opaque in the search result (although this is not evident without a key on this page) but HSP101 has also been conditionally knocked down in *P. falciparum* (Beck *et al*, 2014 Nature) but this is missing from the database. Similarly, other conditional knockouts are missing from the database (Ito *et al*, eLIFE, 2017). (This is why we ticked 'partly' for sufficient details of methods as it wasn't clear what the terms used for inclusion criteria were.)

Interpro signature section. In the first paragraph (results, page 5), should this be specifically the MSP7 C-terminal domain rather than any of the other MSPs C-terminal domain (for example, MSP1)?

In Supp. Table 1, the signatures are from various platforms (eg. SMART, EMBL-EBI Pfam, TIGR, etc) and therefore the same domain may be represented more than once. It may be good to note this in the text. Thus, in some cases, the numbers do not match between search engines. (eg. AAA = 7/20 or 17/33).

This section also mentions the MFS transporter as 21/24 being viable but in the supplemental table MFS is listed twice with 8/9 (PF07690) and 3/4 (PS50850) being viable. Are we missing something in the way we are searching this Table? Moreover, we also couldn't see the 6 cysteine domain in Supp Table 1.

There were a couple of minor formatting issues:

1. Delete 'a' on line 6. Ie. across a the literature.
2. The quotation marks for terms used in the literature searches need correction.

3. In the results section, second paragraph, italicise *P. falciparum*. Also on Page 6, second paragraph.

**Is the rationale for creating the dataset(s) clearly described?**
Yes

**Are the protocols appropriate and is the work technically sound?**
Yes

**Are sufficient details of methods and materials provided to allow replication by others?**
Partly

**Are the datasets clearly presented in a useable and accessible format?**
Yes

***Competing Interests:*** No competing interests were disclosed.

**We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Author Response 21 Jul 2017

**Theo Sanderson**, Wellcome Trust Sanger Institute, UK

Thank you very much for this rapid and highly constructive review which has significantly improved the database. We have been working to address the points you have raised and have responded inline below.

*The journal article by Sanderson and Rayner presents a new database termed 'PhenoPlasm' that provides a combined database of reverse genetics information extracted from the rodent Plasmodium databases RMgmDB and PlasmoGEM and manual curation of phenotype data from publications that have described reverse genetic experiments in Plasmodium falciparum.*
*This new database will provide an excellent resource for malaria researchers as it will enable them to very rapidly assess whether attempts to disrupt/knockdown a particular gene(s) in any of the various species has already been attempted and if so what the resulting phenotypes are, without having to transform data by orthology. We actually recommend that the database link is included in the abstract.*

The database link is indeed included in the abstract.

*After visiting the database, we noticed there was some additional and very useful features that are not described in the manuscript and we recommend that they are included. For example, the subcellular localisation data and the maps. This puts so much more at the fingertips of researchers. It is ambiguous as to what 'maps' refer to until you click on the link so we do recommend that this is called something else, such as 'pathways protein mapped to'.*

We agree that "Maps" is confusing and have amended accordingly. We have also added a sentence in the text pointing out we include links out to these resources from the Malaria Metabolic Pathways database.

*We also recommend that when a search is undertaken that a key of the symbols is visible on the same page for greater clarity. This is clear in the manuscript but not at all evident when using the database.*

This is an excellent point raised by both reviewers which we have now addressed with an online key.

*Figure 1: For CDPK, there is a circled pale green tick indicating the knockout has a wildtype phenotype. However, next to this is an exclamation mark, meaning the phenotype is different from wildtype. Isn't the circled green tick and exclamation mark mutually exclusive?*

This is an important point illustrated by the phenotypes shown for CDPK4. These phenotypes are indeed in contradiction. This is a genuine case of mutually exclusive experimental reports. Our response to such situations is to display both contradictory phenotypes. This should alert the researcher to a controversy. Clicking on the gene in question allows the sources of the two phenotypes to be traced and a judgement made.

In this case there are two reports on RMgmDB of "No difference" for growth in the asexual bloodstage. However PlasmoGEM barseq data suggests a subtle 14% decrease in growth rate. It is possible that a relatively small attenuation like this might be missed with other methodologies, so this is a relatively minor contradiction. The database will also contain less reconcilable differences. At this stage our aim is to present the interested scientist with all reported phenotypes and allow them to draw their own conclusions.

*Figure 2: The word 'conditional' is embedded in the text. We were wondering if it would be clearer to instead have an additional column where it could be stated what experimental approach has been used (ie. whether the mutant phenotype data is derived from a gene knockout or conditional knockout/knockdown, etc). For an example, if you search Pf3D7_0929400, the disruptability table indicates the gene is refractory to deletion in P. falciparum but in the mutant phenotypes table, the phenotype is ! Attenuated. This is somewhat confusing as parasites are not attenuated until after knockdown.*

We do understand this concern. In light of this suggestion we tested implementing the suggestion of a column, but ultimately decided it was slightly more confusing simply because of the increased busyness of each phenotype. We will keep this suggestion in mind for the future as we receive more feedback.

*Some of the terms that were used in the Literature review section are indicated. However, it would be good to know what the numerous additional terms used were – for example, terms like 'knockout', 'knockdown', 'depletion', 'conditional' may have picked up additional publications. For example, HSP101 has been knocked out in P. berghei, hence why the cross is opaque in the search result (although this is not evident without a key on this page) but HSP101 has also been conditionally knocked down in P. falciparum (Beck et al, 2014 Nature) but this is missing from the database. Similarly, other conditional knockouts are missing from the database (Ito et al, eLIFE, 2017). (This is why we ticked 'partly' for sufficient details of methods as it wasn't clear what the terms used for inclusion criteria were.)*

Thank you for pointing out these omissions of specific studies, which we have now rectified. Many of these further search terms were in fact used, but we accept this critique as it would have been

useful to have kept an exhaustive list of queries used. We feel we are now beyond the point of diminishing returns where additional searches have yielded very few new results. For instance the ' depletion' search did not recover any new results in the first 10 pages. There are no doubt further missing phenotypes but we feel these will be best added by engaging with the community.

*Interpro signature section. In the first paragraph (results, page 5), should this be specifically the MSP7 C-terminal domain rather than any of the other MSPs C-terminal domain (for example, MSP1)?*

Thank you, we have changed this. We were accurately quoting the InterPro family name but we agree that more specificity is helpful in the text here.

*In Supp. Table 1, the signatures are from various platforms (eg. SMART, EMBL-EBI Pfam, TIGR, etc) and therefore the same domain may be represented more than once. It may be good to note this in the text. Thus, in some cases, the numbers do not match between search engines. (eg. AAA = 7/20 or 17/33).*

Yes - we had similar concern , we have now associated phenotypes not with the signature itself but with the parental InterPro family and re-run the analysis. This does not significantly impact the results scientifically but we agree that it makes it easier to interpret the data by avoiding duplicated rows and associating better descriptive names.

*This section also mentions the MFS transporter as 21/24 being viable but in the supplemental table MFS is listed twice with 8/9 (PF07690) and 3/4 (PS50850) being viable. Are we missing something in the way we are searching this Table? Moreover, we also couldn't see the 6 cysteine domain in Supp Table 1.*

The data here was correct, but limited annotation made it confusing to interpret. The signature SSF103473 was 21/24 viable, and does correspond to the MFS InterPro family but the SSF103473 family itself does not have a descriptive name. The same applies to 6-cysteine protein family. Now that we are using InterPro families instead all this confusion should be resolved, since the InterPro family name is listed in the table.

*There were a couple of minor formatting issues:*

1. *Delete 'a' on line 6. Ie. across a the literature.*
2. *The quotation marks for terms used in the literature searches need correction.*
3. *In the results section, second paragraph, italicise P. falciparum. Also on Page 6, second paragraph.*

Thank you, these have been fixed.

___
Theo Sanderson

***Competing Interests:*** No competing interests were disclosed.