Pf7: an open dataset of Plasmodium falciparum genome variation in 20,000 worldwide samples

Muzamil Mahdi Abdel Hamid, Mohamed Hassan Abdelraheem, Desmond Omane Acheampong, Ambroise Ahouidi, Mozam Ali, Jacob Almagro-Garcia, Alfred Amambua-Ngwa, Chanaki Amaratunga, Lucas Amenga-Etego, Ben Andagalu, et al.

RESEARCH ARTICLE

Pf7: an open dataset of *Plasmodium falciparum* genome variation in 20,000 worldwide samples [version 1; peer review: 3 approved]

MalariaGEN, Muzamil Mahdi Abdel Hamid[1], Mohamed Hassan Abdelraheem[1,2], Desmond Omane Acheampong[3], Ambroise Ahouidi[4], Mozam Ali[5], Jacob Almagro-Garcia[5], Alfred Amambua-Ngwa[5,6], Chanaki Amaratunga[7], Lucas Amenga-Etego [8,9], Ben Andagalu[10], Tim Anderson [11], Voahangy Andrianaranjaka[12], Ifeyinwa Aniebo[13], Enoch Aninagyei[14], Felix Ansah[8], Patrick O Ansah [9], Tobias Apinjoh[15], Paulo Arnaldo[16], Elizabeth Ashley [17,18], Sarah Auburn[19,20], Gordon A Awandare[8], Hampate Ba [21], Vito Baraka [22,23], Alyssa Barry[24-26], Philip Bejon [27], Gwladys I Bertin [28], Maciej F Boni [20,29], Steffen Borrmann[30], Teun Bousema [31,32], Marielle Bouyou-Akotet[33], Oralee Branch[34], Peter C Bull[27,35], Huch Cheah[36], Keobouphaphone Chindavongsa[37], Thanat Chookajorn [38], Kesinee Chotivanich[38], Antoine Claessens [6,39], David J Conway [31], Vladimir Corredor[40], Erin Courtier[5], Alister Craig [41,42], Umberto D'Alessandro [6], Souleymane Dama[43], Nicholas Day [17,18], Brigitte Denis[42], Mehul Dhorda[17,44], Mahamadou Diakite [43,45], Abdoulaye Djimde [43], Christiane Dolecek[20], Arjen Dondorp [17,18], Seydou Doumbia[43,45], Chris Drakeley [31], Eleanor Drury[5], Patrick Duffy[7], Diego F Echeverry [46,47], Thomas G Egwang[48], Sonia Maria Mauricio Enosse[16], Berhanu Erko [49], Rick M. Fairhurst[50], Abdul Faiz [51], Caterina A Fanello [17], Mark Fleharty[52], Matthew Forbes[5], Mark Fukuda[53], Dionicia Gamboa [54], Anita Ghansah[55], Lemu Golassa [49], Sonia Goncalves[5], G L Abby Harrison[24], Sara Anne Healy [7], Jason A Hendry[56], Anastasia Hernandez-Koutoucheva[5], Tran Tinh Hien[18,29], Catherine A Hill[57], Francis Hombhanje[58], Amanda Hott[59], Ye Htut[60], Mazza Hussein[1], Mallika Imwong[38], Deus Ishengoma [22,61], Scott A Jackson [62], Chris G Jacob[5], Julia Jeans[5], Kimberly J Johnson[5], Claire Kamaliddin [28,63], Edwin Kamau [64], Jon Keatley[5], Theerarat Kochakarn[38], Drissa S Konate [43], Abibatou Konaté[65], Aminatou Kone[43], Dominic P Kwiatkowski [5], Myat P Kyaw[66,67], Dennis Kyle [59,68], Mara Lawniczak [5], Samuel K Lee[52], Martha Lemnge[22], Pharath Lim[7,69],

Chanthap Lon[70], Kovana M Loua [ID][71,72], Celine I Mandara[22], Jutta Marfurt[19], Kevin Marsh [ID][20,27], Richard James Maude[17,18,73], Mayfong Mayxay [ID][18,74,75], Oumou Maïga-Ascofaré[43,76,77], Olivo Miotto [ID][5,17,78], Toshihiro Mita [ID][79], Victor Mobegi [ID][80], Abdelrahim Osman Mohamed[81], Olugbenga A Mokuolu[82], Jaqui Montgomery[42,83], Collins Misita Morang'a [ID][8], Ivo Mueller[24,84], Kathryn Murie[5], Paul N Newton[18,74], Thang Ngo Duc[85], Thuy Nguyen[5], Thuy-Nhien Nguyen [ID][18,29], Tuyen Nguyen Thi Kim[29], Hong Nguyen Van[85], Harald Noedl[86,87], François Nosten [ID][18,88], Rintis Noviyanti[89], Vincent Ntui-Njock Ntui [ID][15], Alexis Nzila[90], Lynette Isabella Ochola-Oyier[27], Harold Ocholla[91,92], Abraham Oduro [ID][9], Irene Omedo[5,27], Marie A Onyamboko [ID][93], Jean-Bosco Ouedraogo [ID][94], Kolapo Oyebola [ID][95,96], Wellington Aghoghovwia Oyibo[97], Richard Pearson [ID][5], Norbert Peshu[27], Aung P Phyo [ID][17,98], Christopher V Plowe[99], Ric N Price [ID][17-19], Sasithon Pukrittayakamee[38], Huynh Hong Quang[100], Milijaona Randrianarivelojosia[101,102], Julian C Rayner [ID][103], Pascal Ringwald[104], Anna Rosanas-Urgell [ID][105], Eduard Rovira-Vallbona[105], Valentin Ruano-Rubio[52], Lastenia Ruiz[106], David Saunders[107], Alex Shayo [ID][108], Peter Siba[109], Victoria J Simpson[5], Mahamadou S. Sissoko[43], Christen Smith[5], Xin-zhuan Su [ID][7], Colin Sutherland[31], Shannon Takala-Harrison[110], Arthur Talman[111], Livingstone Tavul[109], Ngo Viet Thanh[29], Vandana Thathy[27,112], Aung Myint Thu[88], Mahamoudou Toure[43], Antoinette Tshefu[113], Federica Verra[114], Joseph Vinetz [ID][54,115], Thomas E Wellems [ID][7], Jason Wendler[7,116], Nicholas J White[17,18], Georgia Whitton[5], William Yavo[65,117], Rob W van der Pluijm[17]

[1]Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK
[2]Department of Entomology, Purdue University, West Lafayette, USA
[3]Centre for Health Research & Diagnostics, Divine Word University, Madang, Papua New Guinea
[4]University of South Florida, Tampa, USA
[5]Department of Medical Research, Yangon, Myanmar
[6]East African Consortium for Clinical Research (EACCR), Dar es Salaam, Tanzania
[7]Center for Applied Genetic Technologies, University of Georgia, Athens, GA, USA
[8]The University of Calgary, Calgary, Canada
[9]U.S. Military HIV Research Program, Walter Reed Army Institute of Research, Silver Spring, MD, USA
[10]University Félix Houphouët-Boigny, Abidjan, Cote d'Ivoire
[11]Myanmar Oxford Clinical Research Unit, University of Oxford, Yangon, Myanmar
[12]MRC Centre for Genomics and Global Health, Big Data Institute, Oxford University, Oxford, UK
[13]Juntendo University, Tokyo, Japan
[14]Department of Biochemistry and Centre for Biotechnology and Bioinformatics, University of Nairobi, Nairobi, Kenya
[15]Faculty of Medicine, University of Khartoum, Khartoum, Sudan
[16]Department of Paediatrics and Child Health, University of Ilorin, Ilorin, Nigeria
[17]World Mosquito Program, Monash University, Melbourne, Australia
[18]University of Melbourne, Melbourne, Australia

[19]National Institute of Malariology, Parasitology and Entomology (NIMPE), Hanoi, Vietnam

[20]MARIB - Malaria Research Initiative Bandarban, Bandarban, Bangladesh

[21]Medical University of Vienna, Vienna, Austria

[22]Shoklo Malaria Research Unit, Mahidol-Oxford Tropical Medicine Research Unit, Faculty of Tropical Medicine, Mahidol University, Mae Sot, Thailand

[23]NYU School of Medicine Langone Medical Center, New York, USA

[24]Department of Pathology, University of Cambridge, Cambridge, UK

[25]National Center for Parasitology, Entomology and Malaria Control, Phnom Penh, Cambodia

[26]Center of Malariology, Parasitology and Entomology (CMPE), Vientiane, Lao People's Democratic Republic

[27]Mahidol University, Bangkok, Thailand

[28]LPHI, MIVEGEC, INSERM, CNRS, IRD, University of Montpellier, Montpellier, France

[29]National University of Colombia, Bogota, Colombia

[30]Liverpool School of Tropical Medicine, Liverpool, UK

[31]Malawi-Liverpool-Wellcome Trust Clinical Research Program, Blantyre, Malawi

[32]Malaria Research and Training Centre, University of Science, Techniques and Technologies of Bamako, Bamako, Mali

[33]WorldWide Antimalarial Resistance Network – Asia Regional Centre, Bangkok, Thailand

[34]University of Public Health, Yangon, Myanmar

[35]University of Georgia, Athens, USA

[36]Medical Care Development International, Maryland, USA

[37]National Institute of Allergy and Infectious Diseases, Phnom Penh, Cambodia

[38]University Gamal Abdel Nasser of Conakry, Conakry, Guinea

[39]Institut National de Santé Publique, Conakry, Guinea

[40]Harvard TH Chan School of Public Health, Harvard University, Boston, USA

[41]Lao-Oxford-Mahosot Hospital-Wellcome Trust Research Unit, Microbiology Laboratory, Mahosot Hospital, Vientiane, Lao People's Democratic Republic

[42]Institute of Research and Education Development (IRED), University of Health Sciences, Ministry of Health, Vientiane, Lao People's Democratic Republic

[43]Bernhard Nocht Institute for Tropical Medicine, Hamburg, Germany

[44]Research in Tropical Medicine, Kwame Nkrumah University of Sciences and Technology, Kumasi, Ghana

[45]National Institute for Medical Research (NIMR), Dar es Salaam, Tanzania

[46]Walter and Eliza Hall Institute, Melbourne, Australia

[47]Deakin University, Geelong, Australia

[48]Burnet Institute, Melbourne, Australia

[49]KEMRI Wellcome Trust Research Programme, Kilifi, Kenya

[50]Institute of Research for Development (IRD), Paris, France

[51]Oxford University Clinical Research Unit (OUCRU), Ho Chi Minh City, Vietnam

[52]Institute for Tropical Medicine, University of Tübingen, Tübingen, Germany

[53]London School of Hygiene and Tropical Medicine, London, UK

[54]Radboud University Medical Center, Nijmegen, The Netherlands

[55]Department of Parasitology-Mycology, Université des Sciences de la Santé, Libreville, Gabon

[56]Université d'Antananarivo, Antananarivo, Madagascar

[57]Health Strategy and Delivery Foundation, Lagos, Nigeria

[58]Department of Biomedical Sciences, School of Basic and Biomedical Sciences, University of Health & Allied Sciences, Ho, Ghana

[59]University of Buea, Buea, Cameroon

[60]Instituto Nacional de Saúde (INS), Maputo, Mozambique

[61]Mahidol-Oxford Tropical Medicine Research Unit (MORU), Bangkok, Thailand

[62]Centre for Tropical Medicine and Global Health, University of Oxford, Oxford, UK

[63]Menzies School of Health Research, Charles Darwin University, Darwin, Northern Territory, Australia

[64]Nuffield Department of Medicine, University of Oxford, UK

[65]Institut National de Recherche en Santé Publique, Nouakchott, Mauritania

[66]Department of Epidemiology, International Health Unit, Universiteit Antwerpen, Antwerp, Belgium

[67]Institute of Malariology, Parasitology, and Entomology (IMPE) Quy Nhon, Ministry of Health, Quy Nhon, Vietnam

[68]Institut Pasteur de Madagascar, Antananarivo, Madagascar

[69]Universités d'Antananarivo et de Mahajanga, Antananarivo, Madagascar

[70]Cambridge Institute for Medical Research, University of Cambridge, Cambridge, UK

[71]World Health Organization (WHO), Geneva, Switzerland

[72]

Institute of Tropical Medicine Antwerp, Antwerp, Belgium
[73]Universidad Nacional de la Amazonia Peruana, Iquitos, Peru
[74]Department of Medicine, Uniformed Services University, Bethesda, MD, USA
[75]Nelson Mandela Institute of Science and Technology, Arusha, Tanzania
[76]Papua New Guinea Institute of Medical Research, Goroka, Papua New Guinea
[77]Center for Vaccine Development and Global Health, University of Maryland, School of Medicine, Baltimore, MD, USA
[78]MIVEGEC, Université de Montpellier, IRD, CNRS, Montpellier, France
[79]Department of Microbiology and Immunology, Columbia University Irving Medical Center, New York, NY, USA
[80]University of Kinshasa, Kinsasha, Congo, Democratic Republic
[81]Sapienza University of Rome, Rome, Italy
[82]Yale School of Medicine, New Haven, CT, USA
[83]Seattle Children's Hospital, Seattle, USA
[84]Malaria Research and Control Center of the National Institute of Public Health, Abidjan, Cote d'Ivoire
[85]University Clinical Research Center (UCRC), Bamako, Mali
[86]Departamento de Microbiología, Universidad del Valle, Cali, Colombia
[87]Centro Internacional de Entrenamiento e Investigaciones Médicas - CIDEIM, Cali, Colombia
[88]Biotech Laboratories, Kampala, Uganda
[89]Aklilu Lemma Institute of Pathobiology, Addis Ababa University, Addis Ababa, Ethiopia
[90]National Institutes of Health (NIH), Maryland, USA
[91]Dev Care Foundation, Dhaka, Bangladesh
[92]Broad Institute of Harvard and MIT and Harvard, Cambridge, MA, USA
[93]Department of Immunology and Medicine, US Army Medical Component, Armed Forces Research Institute of Medical Sciences (USAMC-AFRIMS), Bangkok, Thailand
[94]Laboratorio ICEMR-Amazonia, Laboratorios de Investigacion y Desarrollo, Facultad de Ciencias y Filosofia, Universidad Peruana Cayetano Heredia, Lima, Peru
[95]Nogouchi Memorial Institute for Medical Research, Legon-Accra, Ghana
[96]Eijkman Institute for Molecular Biology, Jakarta, Indonesia
[97]King Fahid University of Petroleum and Minerals (KFUMP), Dhahran, Saudi Arabia
[98]KEMRI Centres for Disease Control and Prevention (CDC) Research Program, Kisumu, Kenya
[99]Centre for Bioinformatics and Biotechnology, University of Nairobi, Nairobi, Kenya
[100]Kinshasa School of Public Health, University of Kinshasa, Kinshasa, Congo, Democratic Republic
[101]Institut de Recherche en Sciences de la Santé, Ouagadougou, Burkina Faso
[102]Nigerian Institute of Medical Research, Lagos, Nigeria
[103]Parasitology and Bioinformatics Unit, Faculty of Science, University of Lagos, Lagos, Nigeria
[104]College of Medicine, University of Lagos, Lagos, Nigeria
[105]Shoklo Malaria Research Unit, Bangkok, Thailand
[106]University of Maryland School of Medicine, Maryland, USA
[107]Institute of Endemic Diseases, University of Khartoum, Khartoum, Sudan
[108]Nuclear Applications In Biological Sciences, Sudan Atomic Energy Commission, Khartoum, Sudan
[109]Department of Biomedical Sciences, School of Allied Health Sciences, University of Cape Coast, Cape Coast, Ghana
[110]Health Research Epidemiological Surveillance and Training Institute (IRESSEF), Université Cheikh Anta Diop, Dakar, Senegal
[111]Wellcome Sanger Institute, Hinxton, UK
[112]Medical Research Council Unit The Gambia at the London School of Hygiene and Tropical Medicine, Banjul, The Gambia
[113]National Institute of Allergy and Infectious Diseases (NIAID), NIH, Maryland, USA
[114]West African Centre for Cell Biology of Infectious Pathogens (WACCBIP), University of Ghana, Legon, Ghana
[115]Navrongo Health Research Centre, Ghana Health Service, Navrongo, Ghana
[116]United States Army Medical Research Directorate-Africa, Kenya Medical Research Institute/Walter Reed Project, Kisumu, Kenya
[117]Texas Biomedical Research Institute, San Antonio, USA

Open Peer Review

Approval Status ✔ ✔ ✔

Abstract

We describe the MalariaGEN Pf7 data resource, the seventh release of *Plasmodium falciparum* genome variation data from the MalariaGEN network. It comprises over 20,000 samples from 82 partner studies in 33 countries, including several malaria endemic regions that were previously underrepresented. For the first time we include dried blood spot samples that were sequenced after selective whole genome amplification, necessitating new methods to genotype copy number variations. We identify a large number of newly emerging *crt* mutations in parts of Southeast Asia, and show examples of heterogeneities in patterns of drug resistance within Africa and within the Indian subcontinent. We describe the profile of variations in the C-terminal of the *csp* gene and relate this to the sequence used in the RTS,S and R21 malaria vaccines. Pf7 provides high-quality data on genotype calls for 6 million SNPs and short indels, analysis of large deletions that cause failure of rapid diagnostic tests, and systematic characterisation of six major drug resistance loci, all of which can be freely downloaded from the MalariaGEN website.

	1	2	3
version 1 16 Jan 2023	✓ view	✓ view	✓ view

1. **Fabián Sáenz** (iD), Pontificia Universidad Católica del Ecuador, Quito, Ecuador

2. **David A. Fidock** (iD), Columbia University Irving Medical Center, New York, USA

3. **Cristian Koepfli** (iD), University of Notre Dame, Notre Dame, USA

Any reports and responses or comments on the article can be found at the end of the article.

Keywords

malaria, plasmodium falciparum, genomics, data resource, genomic epidemiology

---

Corresponding author: MalariaGEN (support@malariagen.net)

Author roles: Abdel Hamid MM: Investigation, Resources, Writing – Review & Editing; **Abdelraheem MH**: Investigation, Resources, Writing – Review & Editing; **Acheampong DO**: Investigation, Resources, Writing – Review & Editing; **Ahouidi A**: Investigation, Resources, Writing – Review & Editing; **Ali M**: Investigation, Writing – Review & Editing; **Almagro-Garcia J**: Data Curation, Formal Analysis, Investigation, Writing – Review & Editing; **Amambua-Ngwa A**: Investigation, Resources, Writing – Review & Editing; **Amaratunga C**: Investigation, Resources, Writing – Review & Editing; **Amenga-Etego L**: Investigation, Resources, Writing – Review & Editing; **Andagalu B**: Investigation, Resources, Writing – Review & Editing; **Anderson T**: Investigation, Resources, Writing – Review & Editing; **Andrianaranjaka V**: Investigation, Resources, Writing – Review & Editing; **Aniebo I**: Investigation, Resources, Writing – Review & Editing; **Aninagyei E**: Investigation, Resources, Writing – Review & Editing; **Ansah F**: Investigation, Resources, Writing – Review & Editing; **Ansah PO**: Investigation, Resources, Writing – Review & Editing; **Apinjoh T**: Investigation, Resources, Writing – Review & Editing; **Arnaldo P**: Investigation, Resources, Writing – Review & Editing; **Ashley E**: Investigation, Resources, Writing – Review & Editing; **Auburn S**: Investigation, Resources, Writing – Review & Editing; **Awandare GA**: Investigation, Resources, Writing – Review & Editing; **Ba H**: Investigation, Resources, Writing – Review & Editing; **Baraka V**: Investigation, Resources, Writing – Review & Editing; **Barry A**: Investigation, Resources, Writing – Review & Editing; **Bejon P**: Investigation, Resources, Writing – Review & Editing; **Bertin GI**: Investigation, Resources, Writing – Review & Editing; **Boni MF**: Investigation, Resources, Writing – Review & Editing; **Borrmann S**: Investigation, Resources, Writing – Review & Editing; **Bousema T**: Investigation, Resources, Writing – Review & Editing; **Bouyou-Akotet M**: Investigation, Resources, Writing – Review & Editing; **Branch O**: Investigation, Resources, Writing – Review & Editing; **Bull PC**: Investigation, Resources, Writing – Review & Editing; **Cheah H**: Investigation, Resources, Writing – Review & Editing; **Chindavongsa K**: Investigation, Resources, Writing – Review & Editing; **Chookajorn T**: Formal Analysis, Investigation, Writing – Review & Editing; **Chotivanich K**: Investigation, Resources, Writing – Review & Editing; **Claessens A**: Investigation, Resources, Writing – Review & Editing; **Conway DJ**: Investigation, Resources, Writing – Review & Editing; **Corredor V**: Investigation, Resources, Writing – Review & Editing; **Courtier E**: Data Curation, Investigation, Project Administration, Writing – Review & Editing; **Craig A**: Investigation, Resources, Writing – Review & Editing; **D'Alessandro U**: Investigation, Resources, Writing – Review & Editing; **Dama S**: Investigation, Resources, Writing – Review & Editing; **Day N**: Investigation, Resources, Writing – Review & Editing; **Denis B**: Investigation, Resources, Writing – Review & Editing; **Dhorda M**: Investigation, Resources, Writing – Review & Editing; **Diakite M**: Investigation, Resources, Writing – Review & Editing; **Djimde A**: Investigation, Resources, Writing – Review & Editing; **Dolecek C**: Investigation, Resources, Writing – Review & Editing; **Dondorp A**: Investigation, Resources, Writing – Review & Editing; **Doumbia S**: Investigation, Resources, Writing – Review & Editing; **Drakeley C**: Investigation, Resources, Writing – Review & Editing; **Drury E**: Investigation, Writing – Review & Editing; **Duffy P**: Investigation, Resources, Writing – Review & Editing; **Echeverry DF**: Investigation, Resources, Writing – Review & Editing; **Egwang TG**: Investigation, Resources, Writing – Review & Editing; **Enosse SMM**: Investigation, Resources, Writing – Review & Editing; **Erko B**: Investigation, Resources, Writing – Review & Editing; **Fairhurst RM**: Investigation, Resources, Writing – Review & Editing; **Faiz A**: Investigation, Resources, Writing – Review & Editing; **Fanello CA**: Investigation, Resources, Writing – Review & Editing; **Fleharty M**: Formal Analysis, Investigation, Software, Writing – Review & Editing; **Forbes M**: Formal Analysis, Investigation, Writing – Review & Editing; **Fukuda M**: Investigation, Resources, Writing – Review & Editing; **Gamboa D**: Investigation, Resources, Writing – Review &

Introduction

Despite global malaria eradication efforts in the mid-20th century and more recent advances in malaria control, *Plasmodium falciparum* remains endemic throughout Africa, Asia, South America and Oceania. According to the most recent World Malaria Report, each year over 200 million people suffer from malaria due to *P. falciparum* and over 600,000 die as a result[1]. Most of the disease burden falls on Africa, and particularly African children. There is international commitment to control malaria more effectively and many countries are working towards the long-term goal of malaria elimination. However the parasites are continually evolving to resist antimalarial drugs and to evade host immunity, and this is a major challenge to sustainable malaria control and elimination.

Our understanding of the evolutionary biology and population genomics of malaria parasites has advanced considerably over the past decade. There is now a substantial body of literature on the genomic diversity and global population structure of malaria parasites, on within-host genetic variation and what this tells us about superinfection and cotransmission, on the identification and monitoring of parasite drug resistance loci, on genetic variation in malaria vaccine antigens, and on methods of analysing genetic relatedness to understand patterns of malaria transmission. A useful summary of the current state of the field can be found in the recent review by Neafsey *et al.*[2].

This rapidly growing area of research is underpinned by open data on genome sequence variation in natural parasite populations. Here we report a new release of curated open data on *Plasmodium falciparum* genome variation from the MalariaGEN network[3]. It includes samples featured in previous MalariaGEN data releases including the Pf3k Project[4], the *Plasmodium falciparum* Community Project[5] and the GenRe Mekong Project[6]. To avoid confusion between different MalariaGEN datasets we now identify each by a version number. In this new nomenclature, the previous version[5] is called Pf6, and the version described here is called Pf7.

Whole genome sequencing of all the samples in the Pf7 dataset was performed at the Wellcome Sanger Institute and a standardised analysis pipeline was used for variant discovery and genotyping. The Pf7 analysis pipeline was broadly similar to that used for the Pf6 dataset with some improvements that are described in more detail below. Sequence data and genotype calls were returned to partners for use in their own analyses and publications in line with MalariaGEN's guiding principles on equitable data sharing[3].

The Pf7 dataset comprises 20,864 samples of *P. falciparum* collected by 82 partner studies from 33 countries in Africa, Asia, South America and Oceania between 1984 and 2018 (Table 1, Supplementary Tables 1 and 2, Supplementary Figure 1). Compared to the Pf6 dataset, this includes 13,752 new samples, 33 new partner studies and 5 additional countries. The majority of new samples (12,146) were collected since 2014, but there were also 379 samples collected prior to 2000 (Supplementary Figure 2).

The most significant technical advance in the Pf7 dataset is that most of the new samples (12,891/13,752, 94%) came from the MalariaGEN SpotMalaria Project[7]. SpotMalaria was designed to simplify and standardise the process of collecting dried blood spot (DBS) samples for parasite genetic analysis. The SpotMalaria protocol ensures that the vast majority of DBS samples are suitable for targeted genotyping of drug resistance loci, e.g. by amplicon sequencing, and that a significant proportion of samples are also suitable for whole genome sequencing. This requires an intermediate step known as selective whole genome amplification (sWGA), which makes it possible to obtain parasite genome sequence data from a very small sample, but at the cost of introducing considerable variability in sequencing coverage across the genome. This is the first study to analyse a large number of sWGA *P. falciparum* genomes and therefore it was important to establish that sWGA was not introducing significant biases, and to adapt our methods for calling structural variations which are particularly sensitive to artefactual variation in sequencing coverage.

We have performed a number of analyses to make the Pf7 dataset as useful as possible for a broad range of users. These include descriptions of global population structure, geographic

Table 1. Counts of samples in the dataset. Countries are grouped into ten major sub-populations based on their geographic and genetic characteristics. For each country, the table reports: the number of distinct sampling locations (first-level administrative divisions); the total number of samples sequenced; the number of high-quality samples included in the analysis; the percentage of samples collected between 2016–2018, the most recent sampling period in the dataset; and the percentage of samples which are new since the Pf6 release. There are 20,704 samples from natural infections with validated metadata and 160 classified as unverified identity, where this information is not available. The breakdown by admin division is reported in Supplementary Table 1 and the list of contributing studies in Supplementary Table 2.

Major sub-population	Country	Sampling locations	Sequenced samples	Analysis set samples	% analysis samples 2016-2018	% analysis samples new in Pf7
SA (South America)	Peru	1	21	21	0%	0%
	Colombia	4	159	135	59%	88%
	Venezuela	1	2	2	50%	100%

Major sub-population	Country	Sampling locations	Sequenced samples	Analysis set samples	% analysis samples 2016-2018	% analysis samples new in Pf7
AF-W (Africa - West)	Gambia	3	1,247	863	13%	74%
	Senegal	2	155	150	0%	44%
	Guinea	2	199	151	0%	1%
	Mauritania	3	104	92	0%	18%
	Côte d'Ivoire	1	71	71	0%	1%
	Mali	6	1,804	1,167	38%	63%
	Burkina Faso	1	58	57	0%	2%
	Ghana	7	4,145	3,131	54%	73%
	Benin	2	334	150	76%	76%
	Nigeria	2	140	110	74%	74%
	Gabon	1	59	55	0%	100%
	Cameroon	1	294	264	11%	11%
AF-C (Africa - Central)	Congo DR	1	573	520	18%	34%
AF-NE (Africa - Northeast)	Sudan	3	203	76	88%	100%
	Uganda	1	15	12	0%	8%
	Ethiopia	2	34	21	0%	0%
	Kenya, Kisumu	1	64	63	0%	5%
AF-E (Africa - East)	Kenya, Kilifi	1	662	627	0%	92%
	Malawi	2	371	265	0%	7%
	Tanzania	5	697	589	0%	46%
	Mozambique	1	91	34	0%	100%
	Madagascar	2	25	24	0%	0%
AS-S-E (Asia - South - East)	India, Odisha or West Bengal	2	244	233	100%	100%
AS-S-FE (Asia - South - Far East)	India, Tripura	1	72	67	100%	100%
AS-SE-W (Asia - Southeast - West)	Bangladesh	1	1,658	1,310	59%	94%
	Myanmar	8	1,260	985	69%	79%
	Thailand, Tak or Ranong	2	994	895	0%	3%
AS-SE-E (Asia - Southeast - East)	Thailand, Sisakhet	1	112	59	39%	66%
	Laos	5	1,052	991	87%	88%
	Cambodia	7	1,723	1,267	28%	30%
	Vietnam	10	1,733	1,404	62%	84%
OC-NG (Oceania - New Guinea)	Indonesia	1	133	121	25%	34%
	Papua New Guinea	3	251	221	46%	46%
Total natural infection with validated metadata	Various locations	97	20,704	16,203	42%	63%
Unverified identity	Various locations	0	160	0		
Total		97	20,864	16,203	42%	63%

patterns of drug resistance, haplotypic analysis of drug resistance loci, *hrp2* and *hrp3* deletions that can cause failure of rapid diagnostic tests, and variation in the C-terminal of the *csp* antigen used in the most advanced malaria vaccines. These analyses are not intended to be comprehensive and technical users of the dataset can download the analysis-ready dataset for more specialised or detailed investigations.

A high level view of the Pf7 dataset can be obtained from the data exploration tool at the MalariaGEN website. This shows the locations and years where the samples were collected, and the genotype-inferred drug resistance status of each sample. Most importantly it names the investigators who led the studies that contributed the samples at each location and thus made this global dataset possible.

Results
Variant discovery and genotyping
We used the Illumina platform to produce whole genome sequencing data on all samples and mapped sequence reads against the *P. falciparum* 3D7 v3 reference genome. The median depth of coverage was 107 sequence reads averaged across the whole genome and across all samples. We used an analysis pipeline for variant discovery and genotyping analogous to that used in Pf6, as outlined in the Methods section.

In the first stage of analysis we discovered genomic variations in nearly half of the 23Mb *P. falciparum* positions (10,145,661 in total, Supplementary Table 3), including 4,397,801 single nucleotide polymorphisms (SNPs).

For the analysis reported here, we excluded all variants in subtelomeric and internal hypervariable regions, mitochondrial and apicoplast genomes and applied stringent quality filters to the remaining variants as described in the Methods section. A total of 3,125,721 SNPs (of which 2,513,888 were biallelic) and 2,742,938 non-SNPs, i.e. short indels or SNP-indel combinations, passed all these filters. Some of the variant positions that were classified as SNPs in Pf6 are now classified as non-SNPs because they additionally include indel alleles.

We performed quality control checks to remove samples with: (i) unverified or incomplete sample collection information; (ii) evidence of co-infection with other *Plasmodium* species; (iii) more than one technical replicate or time course sampling; (iv) low coverage; (v) a higher than expected number of singleton SNPs. In total, we retained 16,203 high-quality samples (Table 1).

This analysis-ready dataset with details of all participating partner studies and a python package providing convenience methods for accessing is available here.

Effects of selective whole genome amplification (sWGA)
Unlike the previous version, nearly all samples that are new to this release (12,891/13,752, 94%) have been sequenced after undergoing selective whole genome amplification (sWGA). This process allows us to sequence samples collected as dried blood spots, which greatly simplifies many of the operational challenges in collecting venous blood[8].

An artefact introduced by sWGA is high variability in coverage across the genome[8]. This impacts on the use of local variation in genomic coverage as a way to identify large structural variations such as tandem duplications. We therefore developed a novel method based on GATK GermlineCNVCaller (gCNV) for typing duplications around *mdr1* and *plasmepsin 2–3* (associated with resistance to mefloquine and piperaquine, respectively) and deletions of *hrp2* and *hrp3* (associated with rapid diagnostic test failures). We started by compiling a list of observed breakpoints in and around the loci of interest. We then leveraged on the fact that the amplification bias introduced by sWGA, and the consequent variation in coverage, is relatively systematic and can be used for a cross-sample normalisation. Finally, we complemented the results with an analysis to detect presence of face-away reads around the known breakpoints and obtained a final set of calls. For *plasmepsin 2–3* duplications, concordance between gCNV and the face-away reads methods was high, with 99% of samples called as duplication by gCNV also being called as duplication by the face-away method, and the remaining 1% all called as missing. For *mdr1*, concordance was significantly lower, with 19% of samples called as duplication by gCNV being called as no duplication by the face-away method. This could be explained by the fact that the set of breakpoints used is likely not exhaustive, and also by some duplications not being tandem duplications[5]. For samples called as no duplication by gCNV, the vast majority were also called as no duplication by the face-away method (83%) or else missing (17%). For *hrp2* and *hrp3* deletions, we manually validated the results and identified evidence of breakpoints for all the deletion calls.

To ensure that sWGA is not introducing biases in population structure, we analysed four sets of samples from the same location and time periods for which we had a substantial number of both sWGA and non-sWGA samples, and could detect no apparent stratification (Supplementary Figure 3).

Global population structure
We grouped samples by location using the classification scheme known as first-level administrative division: we refer to these as *sampling locations*. Based on principal coordinate and neighbour-joining tree analyses of all samples, we identified ten major divisions of population structure: we refer to these as *major sub-populations*. We then determined the geographical range of each major sub-population by examining the sampling locations that it contained (Figure 1 and Supplementary Figure 4). We identified four major sub-populations in Africa: AF-W (8,610 samples from western Africa), AF-C (573 samples from Kinshasa, DRC), AF-NE (316 samples from Sudan, Ethiopia, Uganda, and Kisumu county in western Kenya), AF-E (1,846 samples from east Africa). In Asia we identified four major sub-populations: two in South Asia (AS-S-E, 244 samples from the Indian states of Odisha and West Bengal, and AS-S-FE, 1,730 samples from Bangladesh and the far-eastern Indian state of Tripura), and two in Southeast
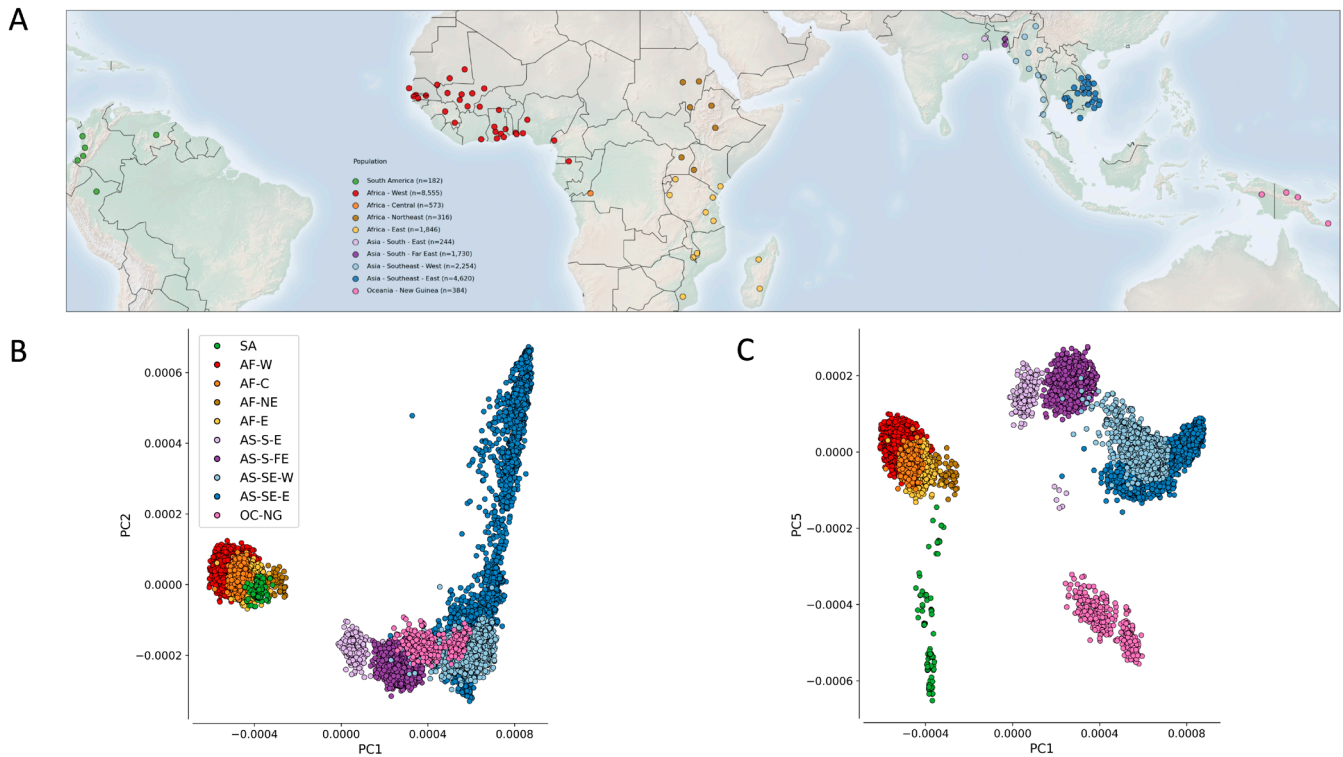
Figure 1. Geographic distribution of sampling locations and population structure. (**A**) Map shows the centres of the 97 first-level administrative divisions from where samples were collected. Points are coloured according to the major sub-population to which the location is assigned (Table 1). (**B**) First two components of a genome-wide principal coordinate analysis. The first axis (PC1, 17.6% of variance explained) captures the separation of African and South American from Asian and Oceanian samples. The second axis (PC2, 2.4% of variance explained) captures finer levels of population structure particularly in the eastern SE Asia population. Each point represents a QC pass sample and the colour legend is the same as in (**A**). (**C**) First and fifth (0.7% of variance explained) components of a genome-wide principal coordinate analysis. Here there is an approximate mapping between the principal components and the geographic location (latitude and longitude).

Asia (AS-SE-W, 2,254 samples from the part of Southeast Asia west of Bangkok, and AS-SE-E, 4,620 samples from the eastern part). The two remaining major sub-populations were OC-NG (384 samples from Oceanian island of New Guinea) and SA (182 samples from South America).

This geographical assignment of ten major sub-populations is a somewhat crude approximation of the underlying population structure, and it does not reflect international conventions for grouping countries or regions. However it provides a framework that allows a broad comparison of population genetic features between different parts of the world, such as the rate of decay of linkage disequilibrium (Supplementary Figure 5), nucleotide diversity (Supplementary Figure 6, lower panel) and complexity of infections (Supplementary Figure 6, upper panel). We also examined the fixation index between sub-populations (Supplementary Figure 7).

Geographic patterns of drug resistance
We classified samples as resistant or sensitive to major antimalarials and combinations based on genotyping of known

drug resistance alleles (Table 2 - see here for details of the heuristics used). At a regional level, the frequency of samples classified as resistant to each drug is broadly consistent with known and previously reported geographical patterns, with the highest prevalence of multidrug resistance observed in Southeast Asia. Interestingly, in South Asia, we find that the frequency of resistance to chloroquine, sulfadoxine and pyrimethamine appears to be much higher in the far-eastern sub-population (Bangladesh and Tripura) than in the eastern sub-population (Odisha and West Bengal).

Care is required when interpreting these findings, as most of the major sub-populations spanned a large geographic region, within which there could be considerable epidemiological diversity, and also because we aggregate samples that were collected over relatively long periods of time during which patterns of resistance may have changed (Supplementary Figure 8). To take West Africa as an example, if we consider samples collected between 2013 and 2016 (Figure 2), we observed levels of chloroquine resistance varying from 0% in Volta, Ghana to 100% in Atlantique, Benin.

Table 2. Frequency of different sets of polymorphisms associated with drug resistance in samples from different geographical regions. All samples were classified into different types of drug resistance based on published genetic markers, and represent best attempt based on the available data. Each type of resistance was considered to be either present, absent or unknown for a given sample. For each resistance type, the table reports: the genetic markers considered; the drug they are associated with; the proportion of samples in each major sub-population classified as resistant out of the samples where the type was not unknown. The number of samples classified as either resistant or not resistant varies for each type of resistance considered (e.g. due to different levels of genomic accessibility); numbers in brackets report the minimum and maximum number analysed while the exact numbers considered are reported in Supplementary Table 4. SP: sulfadoxine-pyrimethamine; treatment: SP used for the clinical treatment of uncomplicated malaria; IPTp: SP used for intermittent preventive treatment in pregnancy; AS-MQ: artesunate + mefloquine combination therapy; DHA-PPQ: dihydroartemisinin + piperaquine combination therapy. *dhfr* triple mutant refers to having all three of 51I, 59R and 108N in *dhfr*. *dhfr* and *dhps* sextuple mutant refers to having all five of 51I, 59R and 108N in *dhfr* and 437G and 540E in *dhps*, plus one of *dhfr*:164L, *dhps*:581G, *dhps*:613S or *dhps*:613T. Full details of the rules used to infer resistance status from genetic markers can be found on the resource page at https://www.malariagen.net/resource/34.

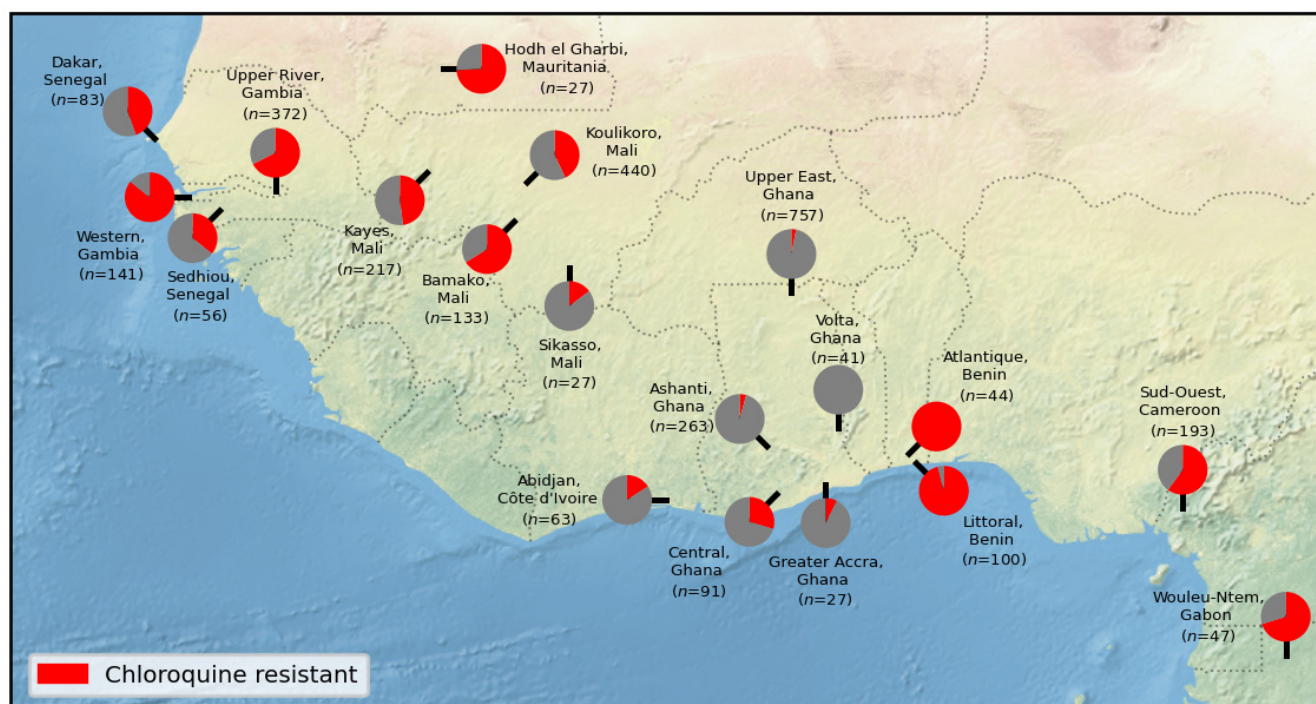| Marker	Associated with resistance to	South America (n=154-158)	Africa - West (n=5234-6233)	Africa - Central (n=397-520)	Africa - East (n=1373-1532)	Africa - Northeast (n=120-170)	Asia - South - East (n=164-233)	Asia - South - Far East (n=1212-1369)	Asia - Southeast - West (n=1657-1876)	Asia - Southeast - East (n=2059-3684)	Oceania - New Guinea (n=298-341)
crt 76T	Chloroquine	100%	29%	61%	24%	40%	31%	94%	99%	95%	96%
dhfr 108N	Pyrimethamine	64%	87%	100%	96%	98%	64%	100%	100%	99%	99%
dhps 437G	Sulfadoxine	60%	78%	97%	83%	82%	8%	89%	100%	83%	69%
mdr1 2+ copies	Mefloquine	0%	0%	0%	0%	0%	0%	0%	29%	5%	1%
kelch13 WHO list	Artemisinin	0%	0%	0%	0%	0%	0%	0%	36%	58%	1%
plasmepsin 2-3 2+ copies	Piperaquine	0%	0%	0%	0%	0%	0%	0%	0%	37%	0%
dhfr triple mutant	SP (treatment)	0%	77%	85%	80%	61%	1%	46%	86%	88%	0%
dhfr and *dhps* sextuple mutant	SP (IPTp)	0%	0%	2%	9%	2%	0%	13%	79%	14%	0%
kelch13 and *mdr1*	AS-MQ	0%	0%	0%	0%	0%	0%	0%	10%	4%	0%
kelch13 and *plasmepsin 2-3*	DHA-PPQ	0%	0%	0%	0%	0%	0%	0%	0%	35%	0%

Figure 2. Heterogenity of chloroquine resistance in west Africa. Inferred resistance levels to chloroquine between 2013 and 2016 in different administrative divisions within West Africa. We only include locations for which we have at least 25 samples with an unambiguous inferred chloroquine resistance phenotype. Note the very different chloroquine resistance profiles in nearby locations, e.g. Volta, Ghana vs Atlantique, Benin.

Amplifications of the genes *mdr1* and *plasmepsin 2-3* are markers of resistance to mefloquine and piperaquine, respectively. Interestingly, two samples collected in 1993 in Cambodia have tandem duplications of both genes, an event which is relatively rare in more recent samples (only 21 samples in total out of 1,959 that have evidence of amplification of either gene). In addition to presence/absence of the amplification, we also provide details of the associated breakpoints for all samples which shows these two samples having two distinct breakpoints in *plasmepsin*, one of which is identical to that most commonly found in contemporary samples.

Haplotype analysis of *kelch13* and *crt* drug resistance loci

Previous reports have shown that the current wave of multidrug-resistant *P. falciparum* in Southeast Asia is driven by the KEL1 lineage of the *kelch13* artemisinin resistance locus[9–12] and is associated with multiple new mutations in the *crt* resistance locus[9]. This dataset confirms the dramatic increase of KEL1 in Cambodia, Eastern Thailand, Laos and Vietnam that has occurred over the past ten years (Supplementary Figure 9). Analysis of the *crt* locus in samples with the K76T resistance variant reveals a major cluster of haplotypes on a common genetic background, the one observed in a widely used lab strain isolated in Asia in 1980 and commonly referred to as Dd2 (Figure 3A, Supplementary Table 5). In addition to the original Dd2 haplotype, we observe 31 additional mutations.

These are essentially restricted to eastern SE Asia with only four samples from outside this region, and they include mutations that have previously been associated with piperaquine resistance[9,13]. They are seen across all countries of eastern SE Asia and have risen rapidly in frequency leading us to consider them *newly emerging Dd2-background* haplotypes (Figure 3B). Most have a single mutation on a Dd2 background, but we observe 13 haplotypes with two mutations and one haplotype (in a single sample) with three mutations (Supplementary Table 5). These findings highlight the value in retrospective analysis of drug resistance mutations, as most of these samples were collected and sequenced before the relevance of *crt* mutations to piperaquine resistance was appreciated.

Variation in the c-terminal region of *csp*

In addition to the selective pressures due to antimalarial drugs highlighted in the previous section, another area of interest is selective pressure due to vaccines. Having a baseline understanding of genetic variation in vaccine genes is likely to be valuable.

The WHO has recently recommended the RTS,S vaccine for use in regions with moderate to high transmission which includes much of sub-Saharan Africa[14]. The vaccine targets the gene *csp* and has a construct based on the 3D7 reference sequence of part of the central NANP repeat region where
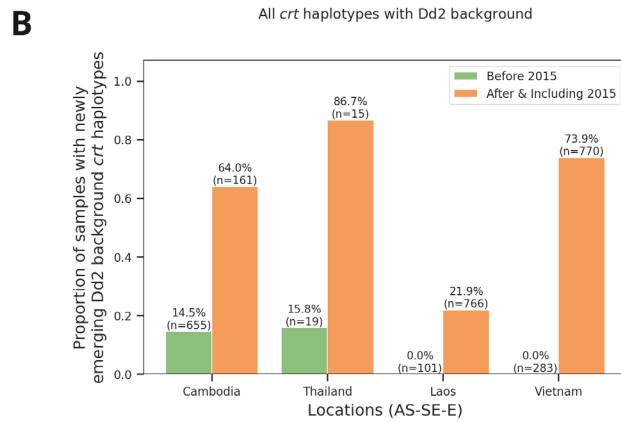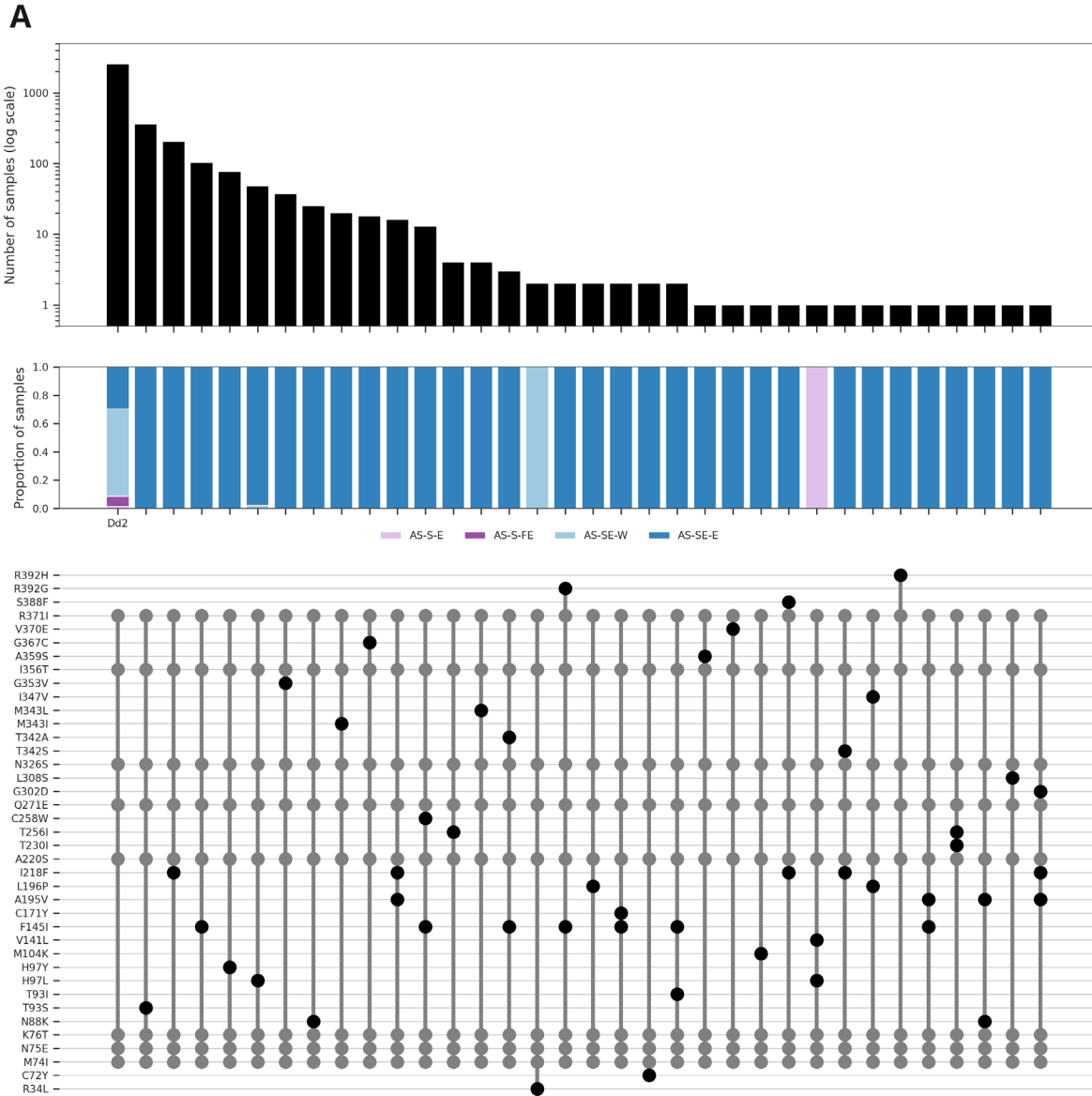
Figure 3. Newly emerging Dd2 background mutations in *crt*. (a) Top panel - frequency of different haplotypes with a genetic background identical to the lab strain Dd2. Dd2 is derived from an isolate taken from a patient in Indochina in 1980. Middle panel - breakdown of samples by major sub-population for each haplotype. Lower panel - amino acid mutations in the haplotypes (with respect to 3D7 reference). Mutations found in the Dd2 haplotype are shown in grey, all other mutations are shown in black. **(b)** Bar plots showing changing frequency of newly emerging Dd2 background *crt* haplotypes in different locations in the eastern SE region. Newly emerging Dd2 background haplotypes are defined as all haplotypes that have all mutations seen in Dd2 plus additional mutations.

antibodies bind and the c-terminal region which contains T cell epitopes[15]. Another vaccine based on the same region and sequence, R21, is also showing promise in early stage clinical trials[16]. Vaccine efficacy is likely to depend on a number of factors, both host and parasite, and clinical trials show some variability between different locations in Africa[17]. How similar the parasite is in the targeted region to the 3D7 sequence used in vaccine design could be a contributing factor to this variability. Genetic diversity in the construct region may or may not affect vaccine efficacy, and in order to understand this it will be important to monitor efficacy against diversity going forwards. Here we begin a systematic catalogue of population-level diversity in *csp*. While it is challenging if not impossible to genotype the central repeat region using short read data, we start here by looking at variation leading to amino acid changes in the c-terminal region of the protein.

We identified all non-synonymous mutations in the *csp* c-terminal region and analysed the frequency of these in different populations. Interestingly, the vast majority of the samples across the globe carry non-reference alleles, i.e. different from the 3D7 sequence used in the vaccine design, at amino acids 301, 317, 318, 321 and 361 (Supplementary Figure 10). We found a total of 248 unique amino acid haplotypes of $csp_{277-397}$ out of 11,254 samples with no ambiguous calls. Amino acid haplotype sequences for the c-terminal region of *csp* for all samples can be found at the MalariaGEN website.

Surprisingly, the most common haplotype in the dataset and the one with the second lowest number of differences from all other unique haplotypes is the one observed in lab strain Dd2, being found in 2,760/11,254 (25%) of the samples and having a mean of 4.7 differences from other haplotypes (Figure 4a). In contrast, the 3D7 haplotype used for both *csp*-based vaccines is only found in 3% of samples and has on average 6.9 differences from all other haplotypes.

Importantly, this striking difference also holds when examining each population separately, including in West Africa from where 3D7 is thought to have originated (Figure 4b).

Taken together, these results show perhaps surprising differences between the target haplotype used in the design of RTS,S and R21 and those circulating in natural parasite populations, and provide a systematic catalogue that can be used in future studies to elucidate any possible clinical significance of sequence diversity.

Genetic origins of *hrp2* and *hrp3* deletions

Most widely used rapid diagnostic tests (RDTs) rely on detection of the products of the *hrp2* and *hrp3* genes, and deletions in these genes is known to lead to RDT failure[18]. We used gCNV to call presence/absence of *hrp2* and *hrp3* deletions in 68% of QC passed samples. Frequencies of deletions vary greatly across countries, and deletions of *hrp3* (1.9%) are more common than those of hrp2 (0.14%) (Supplementary Table 6). The only countries where we see deletions in both *hrp2* and *hrp3* that would cause HRP RDTs to fail are Peru

(6/20 samples), Indonesia (2/115 samples) and Sudan (1/7 samples).

There have been numerous reports of such deletions in recent years, but to date there has been little detail on the mechanisms causing such deletions. We manually inspected reads around the apparent breakpoints in order to classify the types of events driving these deletions. For *hrp2*, all deletion events can be explained by a process of telomere healing whereby the end of a chromosome is deleted and a telomere repeat sequence attached to the breakpoint[19,20] (Figure 5). Telomere healing events can be determined with breakpoint precision and in almost all cases samples with the same breakpoints are from the same country (Supplementary Table 7). For *hrp3*, we also identified a number of telomere healing events, but also two other quite different types of event causing deletion of the gene (Figure 5, Supplementary Table 7). In many cases a new hybrid chromosome appears to have been created by a recombination between chromosome 13 and 11 at a cluster of rRNA genes that have orthologous copies on both chromosomes. In other cases a recombination between chromosome 13 and an inverted section from within chromosome 5 containing the gene *mdr1* can be identified. This remarkable event results in both deletion of *hrp3* and duplication of *mdr1*.

Other types of genetic variation: allelic forms of *eba175*

As with previous releases, in Pf7 we have created genotypes at SNPs genome-wide and CNVs in specific locations, but we intend to continue to expand the resource to consider other types of genetic variation. Various surface antigens, including vaccine candidate genes, have two distinctly different allelic forms[21]. Often the two forms are so divergent that reads from a non-reference form will not map to the 3D7 reference genome[22], hence necessitating an alternative to the mapping-based genotype calling approach described above. As an example, the gene *eba175* has two different allelic forms, known as the F- and C-types[23,24]. As a proof of concept for such a dimorphic gene, we used a novel kmer-based method to call these two types in each sample. We see that 7,380 (69%) samples have the F allele exclusively and 3,364 (31%) have the C allele. Although frequencies of each type vary between populations, we see >10% frequency of each type in all populations (Supplementary Figure 11). These results give weight to the argument that *eba175* is under balancing selection, most likely negative frequency-dependent selection driven by interactions with the human immune system. Analyses of other such dimorphic genes will be left to future work, as will more detailed analysis of variation within these different allelic types.

Discussion

The Pf7 dataset increases the amount of curated open data for population genomic analysis of *P. falciparum* by almost threefold, and greatly increases the number of samples collected within the last five years. With denser geographical coverage it is possible to undertake higher resolution analysis of epidemiological variation within a region, e.g. we observe
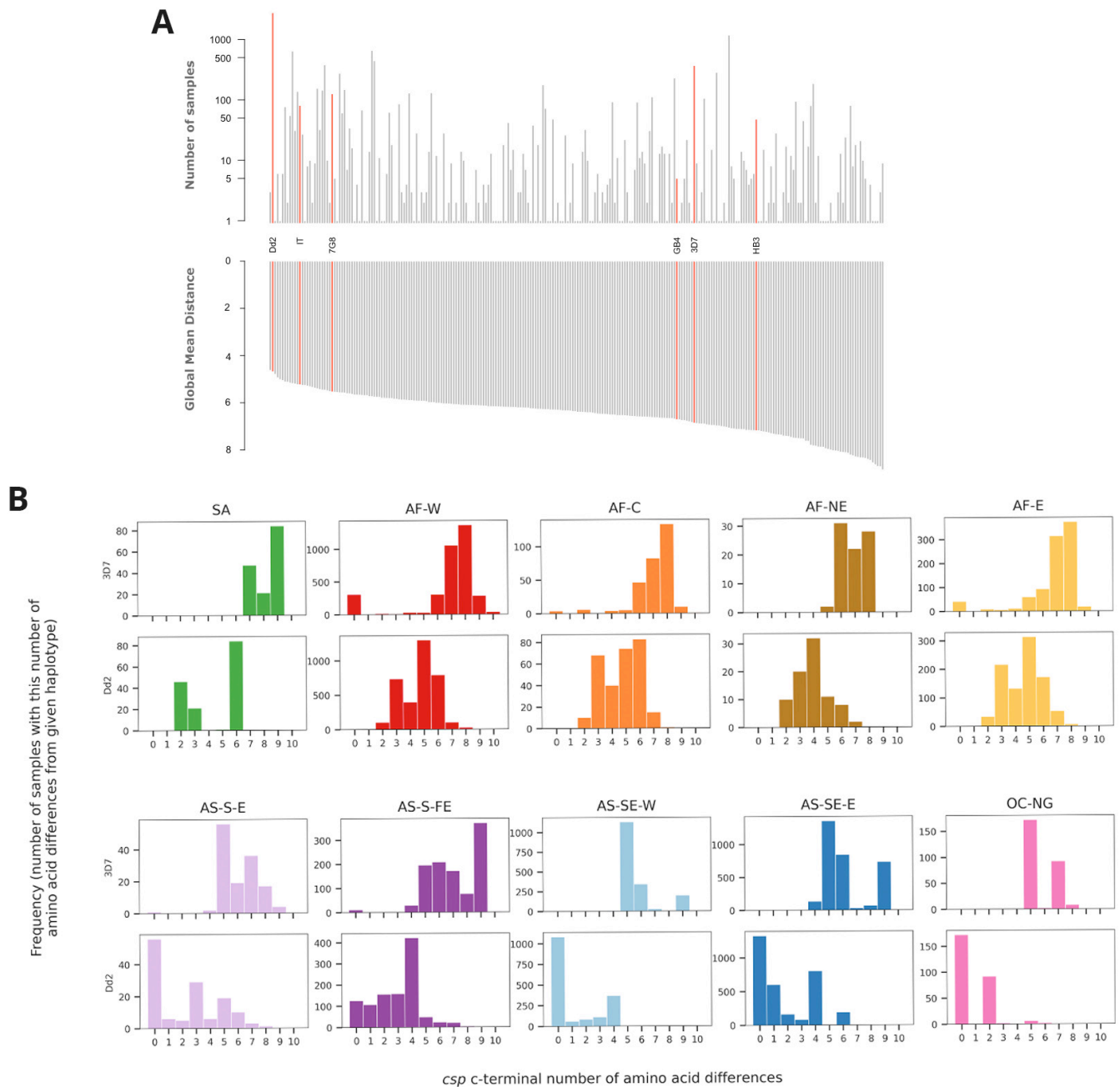
Figure 4. Analysis of c-terminal of *csp*. (a) Upper panel - frequency of different haplotypes of c-terminal of *csp*. Haplotypes found in some lab strains are named and highlighted in red. Haplotypes are ordered as per lower panel. Lower panel - global mean distance (number of amino acid differences) to all other haplotypes. **(b)** Histograms of number of amino acid differences between samples in each major sub-population and the 3D7 haplotype (upper plot) and Dd2 haplotype (lower plot).

considerable heterogeneity of inferred chloroquine resistance in West Africa (Figure 3), and it also allows us to identify new sub-populations with distinctive epidemiological features, e.g. we find two sub-populations in south Asia that have contrasting drug resistance profiles. There is useful historical information to be obtained from older samples that are included in this new data release, e.g. some samples collected

in Cambodia in the early 1990s appear to be resistant to both piperaquine and mefloquine, which is highly relevant to the ongoing evolution of multidrug resistance to artemisinin combination therapy in Southeast Asia.

An important technical advance is that Pf7 contains a large number of samples that were collected as dried blood spots in the
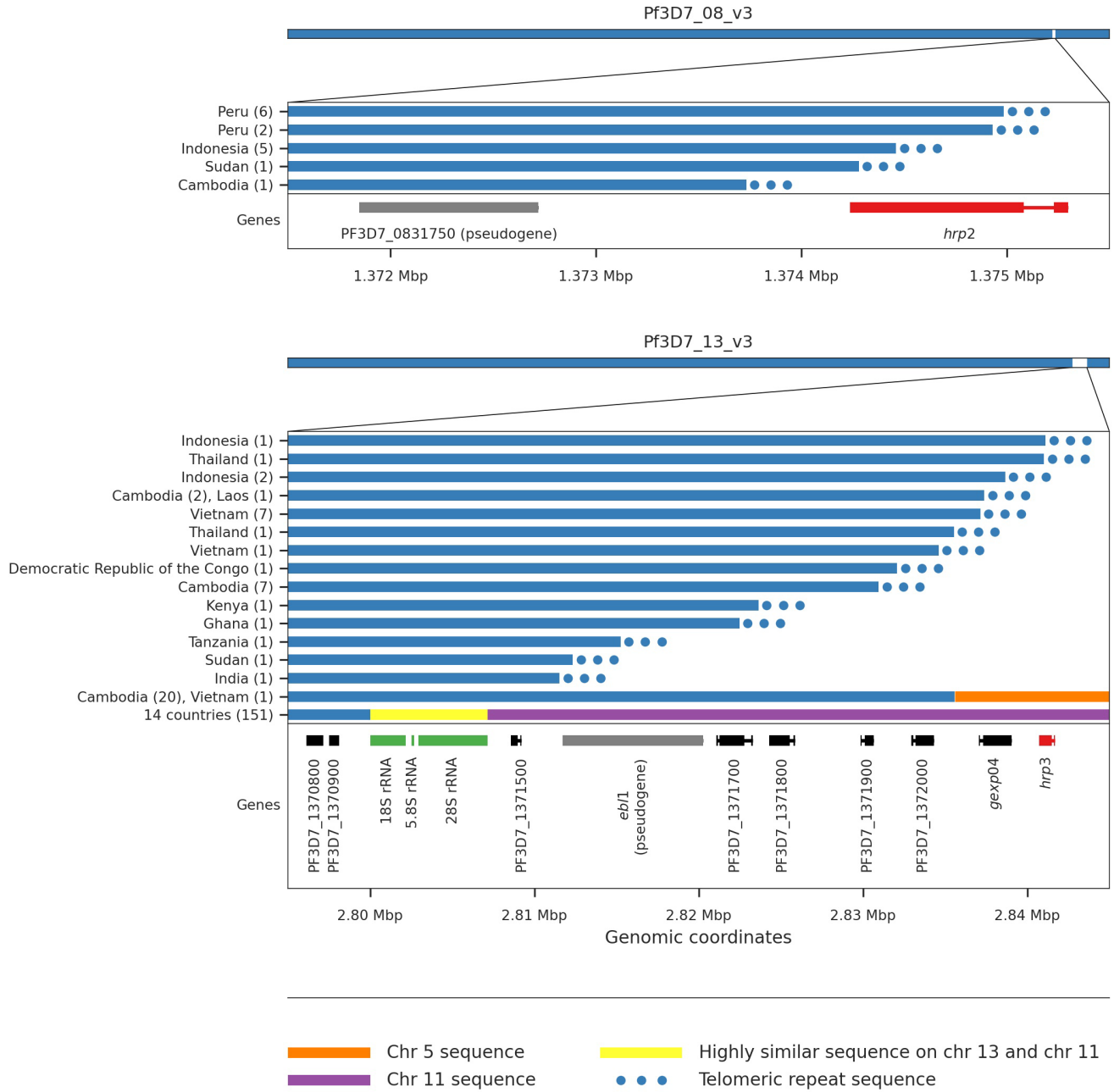
Figure 5. HRP deletion breakpoints. We see five different breakpoints resulting in the deletion of *hrp2*. Four of these are within exon 2 of the gene whereas the fifth is found between *hrp2* and the pseudogene PF3D7_0831750. For all five events we see evidence of telomeric healing from reads that contain part Pf3D7_08_v3 sequence and part telomeric repeat sequence (GGGTTCA/GGGTTTA). We see 16 different breakpoints resulting in the deletion of *hrp3*. For fourteen of these we see evidence of telomeric healing. Note that many of these events result in the deletion of other genes in addition to *hrp3*. For twenty samples from Cambodia and a single sample from Vietnam we see evidence of a recombination with chromosome 5 which results in a hybrid chromosome comprising mostly chromosome 13 sequence but a small inverted section of an internal portion of chromosome 5 containing the gene *mdr1*. We also see evidence of a recombination with chromosome 11 which results in a hybrid chromosome comprising mostly chromosome 13 sequence but also a section of the 3' end of chromosome 11. This is the most common deletion type, being seen in 151 samples from 14 different countries. Because the recombination occurs between highly similar sequences of a set of three orthologous ribosomal RNA genes found on both chromosomes, it is not possible to identify the exact breakpoints.

field. We and others have previously described successful whole genome sequencing of *P. falciparum* from DBS after selective whole-genome amplification but it was unclear how well this methodology would perform at scale. Here we show that the SpotMalaria protocol for sWGA of DBS samples allows us to generate whole genome sequence data of sufficient quality to genotype the vast majority of SNPs with sufficient accuracy and reliability for large-scale population genomic analysis. We have introduced improvements to our pipelines for calling copy number variants, necessitated by the greatly increased heterogeneity of sequencing coverage following sWGA. There remain hypervariable gene families and other regions of the parasite genome that cannot be accurately genotyped in field samples using current methods, and these difficulties are compounded by sWGA, but by working on sequencing and analysis methods we aim to continually improve genome coverage in future releases.

The knowledge that DBS samples can be used for whole genome analysis in large-scale studies is of practical importance, as it empowers field researchers and national malaria control programs to integrate population genomic information with other forms of epidemiological and public health data, and it paves the way to a global infrastructure for genomic surveillance of *P. falciparum*. Information about the processes and methods of the SpotMalaria Project can be obtained at the MalariaGEN website

The Pf7 dataset includes a range of analyses and sample annotations that are intended to increase the utility of the data for researchers working on practical problems in malaria control. Compared to the Pf6 dataset, we have made improvements to methods for calling CNVs at the *mdr1* and *pm2* drug resistance loci and for calling *hrp2* and *hrp3* deletions that can affect rapid diagnostic tests. Other new analyses included in Pf7 include more detailed descriptions of: (a) *hrp2* and *hrp3* deletion breakpoints; (b) drug resistance locus haplotypes and in particular newly emerging *crt* haplotypes; (c) profiles of variation in the *csp* vaccine antigen and the vaccine candidate *eba175*. In future releases we aim to improve and expand analyses that are relevant to malaria control programmes and policymakers.

The Pf7 dataset focuses entirely on genome sequencing data, but there is a growing body of data from amplicon sequencing and targeted genotyping approaches that is highly informative about multiple aspects of *P. falciparum* population genomics. For example, the GenRe-Mekong Project has used the SpotMalaria platform combined with amplicon sequencing to enable malaria control programmes in the Greater Mekong Region to conduct national genomic surveillance of multidrug resistance[6]. In future data releases we aim to integrate data from these different sources to greatly increase sample size and geographical coverage, and thus improve the resolution of population genomic analysis.

Methods
All samples in this study were derived from blood samples obtained from patients with *P. falciparum* malaria, collected with informed consent from the patient or a parent or guardian. At each location, sample collection was approved by the appropriate local and institutional ethics committees. The following local and institutional committees gave ethical approval for the partner studies: Walter and Eliza Hall Institute Human Research Ethics Committee, Australia; University of Antwerp, Belgium; Comite d'Ethique de la Recherche - Institut des Sciences Biomedicales Appliquees, Benin; Ministere de la Sante – Republique du Benin, Benin; Comité d'Éthique, Ministère de la Santé, Bobo-Dioulasso, Burkina Faso; Ministry of Health National Ethics Committee for Health Research, Cambodia; Institutional Review Board University of Buea, Cameroon; Comite Institucional de Etica de investigaciones en humanos de CIDEIM, Colombia; Research Ethics Committee of the Faculty of Medicine of the National University of Colombia; Comité National d'Ethique de la Recherche, Cote d'Ivoire; Comite d'Ethique Universite de Kinshasa, Democratic Republic of Congo; Armauer Hansen Research Institute Institutional Review Board, Ethiopia; Addis Ababa University, Aklilu Lemma Institute of Pathobiology Institutional Review Board, Ethiopia; Ghana Health Service Ethical Review Committee, Ghana; University of Ghana Noguchi Medical Research Institute, Ghana; Navrongo Health Research Centre Institutional Review Board, Ghana; Comite d'Ethique National Pour la Recherché en Santé, République de Guinée; KEMRI Scientific and Ethics Review Unit, Kenya; Ministry of Health National Ethics Committee For Health Research, Laos; College of Medicine, University of Lagos, Nigeria; Comité National d'Ethique auprès du Ministère de la Santé Publique, Madagascar; College of Medicine Regional Ethics Committee University of Malawi, Malawi; Faculté de Médecine, de Pharmacie et d'Odonto-Stomatologie, University of Bamako, Bamako, Mali; L'université des Sciences, des Techniques et des Technologies de Bamako, Mali; Ethics Committee of the Ministry of Health, Mali; Ethics committee of the Ministry of Health, Mauritania; National Bioethics Committees of Mozambique (CNBS); Institutional Review Board, Papua New Guinea Institute of Medical Research, Goroka, Papua New Guinea; PNG Medical Research Advisory Council (MRAC), Papua New Guinea; Institutional Review Board, Universidad Nacional de la Amazonia Peruana, Iquitos, Peru; Al Neelain University Institutional Review Board, Sudan; Federal Ministry of Health, Sudan; Ethics Committee of the Ministry of Health, Senegal; Medical Research Coordinating Committee of the National Institute for Medical Research, Tanzania; Ethics Committee, Faculty of Tropical Medicine, Mahidol University, Bangkok, Thailand; Gambia Government/MRC Joint Ethics Committee, Banjul, The Gambia; Liverpool School of Tropical Medicine, UK; London School of Hygiene and Tropical Medicine Ethics Committee, London, UK; Oxford Tropical Research Ethics Committee, Oxford, UK; University College London Hospitals Research Ethics Committee, UK; Walter Reed Army Institute of Research, USA; National Institute of Allergy and Infectious Diseases, Bethesda, MD, USA; University of Maryland School of Medicine IRB, USA; Ministry of Health Institute of Malariology-Parasitology-Entomology, Vietnam.

Standard laboratory protocols were used to determine DNA quantity and proportion of human DNA in each sample as previously described[22,25].

Here we summarise the bioinformatics methods used to produce and analyse the data; full details are available here.

Reads mapping to the human reference genome were discarded before all analyses, and the remaining reads were mapped to the *P. falciparum* 3D7 v3 reference genome using bwa mem[26]. "Improved" BAMs were created using the Samtools FixMateInformation, Picard MarkDuplicates, and GATK base quality score recalibration. All lanes for each sample were merged to create sample-level BAM files.

Putative variants were called in each sample independently using GATK (v4.1.4.0) HaplotypeCaller, then all samples were combined to jointly genotype the entire cohort using GATK GenotypeGVCFs[27].

SNPs and indels were filtered using GATK's Variant Quality Score Recalibration (VQSR). Variants with a VQSLOD score ≤ 2 were filtered out. Functional annotations were applied using snpEff[28] version 4.3. Genome regions were annotated using BCFtools v1.10.2 (http://www.htslib.org/doc/bcftools.html) and masked if they were outside the core genome. Unless otherwise specified, we used biallelic SNPs that pass all quality filters for all the analysis.

VCF files were converted to zarr v2.4.0 format and subsequent analyses were mainly performed using scikit-allel v1.2.1 and the zarr files.

We identified species using nucleotide sequence from reads mapping to six different loci in the mitochondrial genome, using custom java code. The loci were located within the *cox3* gene (PF3D7_MIT01400), as described in a previously published species detection method[29]. Alleles at various mitochondrial positions within the six loci were genotyped and used for classification.

We created a final analysis set of 16,203 samples after removing samples with unverified identity, mixed species, replicate and low coverage samples, and samples with excessive numbers of singleton SNPs.

We calculate genetic distance between samples using biallelic coding SNPs that pass filters using a method previously described[9].

The matrix of genetic distances was used to generate neighbour-joining trees and principal coordinates. Based on these observations we grouped the samples into ten major sub-populations: South America, West Africa, Central Africa, Northeast Africa, East Africa, an eastern part of South Asia, a far-eastern part of South Asia, the western part of Southeast Asia, the eastern part of Southeast Asia and Oceania, with samples assigned to region based on the geographic location of the sampling site.

F_{WS} was calculated using custom python scripts using the method previously described[7]. Nucleotide diversity (π) was calculated in non-overlapping 25kbp genomic windows, only considering coding SNPs to reduce the ascertainment bias

caused by poor accessibility of non-coding regions. LD decay (r^2) was calculated using the method of Rogers and Huff and biallelic SNPs with low missingness and regional allele frequency >10%. Mean F_{ST} between populations was calculated using Hudson's method.

To call duplication genotypes around *mdr1* and *plasmepsin 2-3* from binned coverage, we adapted the GATK GermlineCNVCaller (gCNV) pipeline, which features a probabilistic Bayesian model that jointly infers both copy-number activity and a model for denoising sequencing systematics. After breakpoint genotypes were called, we performed an initial, permissive round of annotation-based call filtering, using hard cuts to identify failing samples and demarcate duplication and reference genotypes. This was followed by a final round of curation, based on manual inspection of the denoised copy ratios, to discard spurious duplication calls. The resulting filtered gCNV call set was integrated with an analogous call set based on consideration of face-away read-pair evidence, in which we set the breakpoint to be that with the highest proportion of face-away reads.

Deletions in *hrp2* and *hrp3*, genes which are located in subtelomeric regions of the genome with very high levels of natural variation, were identified using the same breakpoint-genotyping framework introduced above. As before, an initial round of permissive, annotation-based filtering was performed, followed by a final round of curation to discard spurious deletion calls. We identified deletion breakpoints by manual inspection of custom plots.

The procedure used to map genetic markers to inferred resistance status classification is described in detail for each drug in the accompanying data release. In brief, we called amino acids at selected loci by first determining the reference amino acids and then, for each sample, applying all variations using the GT field of the VCF file. This same approach was used to identify haplotypes of $csp_{277-397}$. The amino acid and copy number calls generated in drug resistance genes were used to classify all samples into different types of drug resistance. Our methods of classification were heuristic and based on the available data and current knowledge of the molecular mechanisms. Each type of resistance was considered to be either present, absent or unknown for a given sample. *eba175* F- and C-type calls are made by identifying samples that have 19bp kmers present that are unique to the C and F haplotypes.

Data availability

Data are available under the MalariaGEN terms of use for the Pf Community Project: https://www.malariagen.net/data/terms-use/p-falciparum-community-project-terms-use. Depending on the nature, format and content of the data, appropriate mechanisms have been utilised for data access, as detailed below.

This project contains the following underlying data that are available as an online resource: https://www.malariagen.net/resource/34. Data are also available from Figshare.

Figshare: Supplementary data to: Pf7: an open dataset of Plasmodium falciparum genome variation in 20,000 worldwide samples. https://doi.org/10.6084/m9.figshare.21674321[30].

- Study information: Details of the 82 contributing partner studies, including description, contact information and key people.

- Sample provenance and sequencing metadata: sample information including partner study information, location and year of collection, ENA accession numbers, and QC information for 20,864 samples from 33 countries.

- Measure of complexity of infections: characterisation of within-host diversity (F_{ws}) for 16,203 QC pass samples.

- Drug resistance marker genotypes: genotypes at known markers of drug resistance for 16,203 samples, containing amino acid and copy number genotypes at six loci: crt, dhfr, dhps, mdr1, kelch13, plasmepsin 2-3.

- Inferred resistance status classification: classification of 16,203 QC pass samples into different types of resistance to 10 drugs or combinations of drugs and to RDT detection: chloroquine, pyrimethamine, sulfadoxine, mefloquine, artemisinin, piperaquine, sulfadoxine-pyrimethamine for treatment of uncomplicated malaria, sulfadoxine- pyrimethamine for intermittent preventive treatment in pregnancy, artesunate-mefloquine, dihydroartemisinin-piperaquine, hrp2 and hrp3 gene deletions.

- Drug resistance markers to inferred resistance status: details of the heuristics utilised to map genetic markers to resistance status classification.

- CRT haplotypes: Full *crt* gene haplotypes for 16,203 QC pass samples.

- CSP C-terminal haplotypes: Full *csp* C-terminal haplotypes for 16,203 QC pass samples plus 6 lab strains.

- EBA175 calls: *eba175* allelic type calls for 16,203 QC pass samples.

- Reference genome: the version of the 3D7 reference genome fasta file used for mapping.

- Annotation file: the version of the 3D7 reference annotation gff file used for genome annotations.

- Genetic distances: Genetic distance matrix comparing all 20,864 samples.

- Short variants genotypes: Genotype calls on 10,145,661 SNPs and short indels in all 20,864 samples from 33 countries, available both as VCF and zarr files.

Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

Author contributions
Data analysis group
Pearson, RD, Hernandez-Koutoucheva, A, Whitton, G, Lee, SK, Fleharty, M, Ruano-Rubio, V, Almagro-Garcia, J, Kochakarn, T, Chookajorn, T, Nguyen, T, Murie, K, Forbes, M, Hendry, JA, Miotto, O, Kwiatkowski, DP

Local study design, implementation and sample collection
Abdel Hamid, MM, Abdelraheem, MH, Acheampong, DO, Ahouidi, A, Amambua-Ngwa, A, Amaratunga, C, Amenga-Etego, L, Andagalu, B, Anderson, T, Andrianaranjaka, V, Aniebo, I, Aninagyei, E, Ansah, F, Ansah, PO, Apinjoh, T, Arnaldo, P, Ashley, E, Auburn, S, Awandare, GA, Ba, H, Baraka, V, Barry, A, Bejon, P, Bertin, GI, Boni, MF, Borrmann, S, Bousema, T, Bouyou-Akotet, M, Branch, O, Bull, PC, Cheah, H, Chindavongsa, K, Chotivanich, K, Claessens, A, Conway, DJ, Corredor, V, Craig, A, D'Alessandro, U, Dama, S, Day, N, Denis, B, Dhorda, M, Diakite, M, Djimde, A, Dolecek, C, Dondorp, A, Doumbia, S, Drakeley, C, Duffy, P, Echeverry, DF, Egwang, TG, Enosse, SMM, Erko, B, Fairhurst, RM, Faiz, A, Fanello, CA, Fukuda, M, Gamboa, D, Ghansah, A, Golassa, L, Harrison, GLA, Healy, SA, Hien, TT, Hill, CA, Hombhanje, F, Hott, A, Htut, Y, Hussein, M, Imwong, M, Ishengoma, D, Jackson, SA, Kamaliddin, C, Kamau, E, Konate, DS, Konaté, A, Kone, A, Kyaw, MP, Kyle, D, Lawniczak, M, Lemnge, M, Lim, P, Lon, C, Loua, KM, Maïga-Ascofaré, O, Mandara, CI, Marfurt, J, Marsh, K, Maude, RJ, Mayxay, M, Mita, T, Mobegi, V, Mohamed, AO, Mokuolu, OA, Montgomery, J, Morang'a, CM, Mueller, I, Newton, PN, Ngo Duc, T, Nguyen, T, Nguyen Thi Kim, T, Nguyen Van, H, Noedl, H, Nosten, F, Noviyanti, R, Ntui, VN, Nzila, A, Ochola-Oyier, LI, Ocholla, H, Oduro, A, Omedo, I, Onyamboko, MA, Ouedraogo, J, Oyebola, K, Oyibo, WA, Peshu, N, Phyo, AP, Plowe, CV, Price, RN, Pukrittayakamee, S, Quang, HH, Randrianarivelojosia, M, Rayner, JC, Ringwald, P, Rosanas-Urgell, A, Rovira-Vallbona, E, Ruiz, L, Saunders, D, Shayo, A, Siba, P, Sissoko, MS, Su, X, Sutherland, C, Takala-Harrison, S, Talman, A, Tavul, L, Thanh, NV, Thathy, V, Thu, AM, Toure, M, Tshefu, A, van der Pluijm, RW, Verra, F, Vinetz, J, Wellems, TE, Wendler, J, White, NJ, Yavo, W

Sequencing, data production and informatics
Pearson, RD, Nguyen, T, Keatley, J, Murie, K, Drury, E, Ali, M, Jacob, CG, Goncalves, S

Partner study support and coordination
Simpson, VJ, Goncalves, S, Johnson, KJ, Jeans, J, Smith, C, Miotto, O, Courtier, E, Pearson, RD, Kwiatkowski, DP

References

1. **World malaria report 2021**.
    Reference Source

2. Neafsey DE, Taylor AR, MacInnis BL: **Advances and opportunities in malaria population genomics.** *Nat Rev Genet.* 2021; **22**(8): 502–517.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

3. Malaria Genomic Epidemiology Network: **A global network for investigating the genomic epidemiology of malaria.** *Nature.* 2008; **456**(7223): 732–737.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

4. https://www.malariagen.net/parasite/pf3k

5. MalariaGEN: Ahouidi A, Ali M, *et al.*: **An open dataset of** *Plasmodium falciparum* **genome variation in 7,000 worldwide samples [version 2; peer review: 2 approved].** *Wellcome Open Res.* 2021; **6**: 42.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

6. Jacob CG, Thuy-Nhien N, Mayxay M, *et al.*: **Genetic surveillance in the Greater Mekong subregion and South Asia to support malaria control and elimination.** *eLife.* 2021; **10**: e62997.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

7. https://www.malariagen.net/parasite/spotmalaria

8. Oyola SO, Ariani CV, Hamilton WL, *et al.*: **Whole genome sequencing of** *Plasmodium falciparum* **from dried blood spots using selective whole genome amplification.** *Malar J.* 2016; **15**(1): 597.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

9. Hamilton WL, Amato R, van der Pluijm RW, *et al.*: **Evolution and expansion of multidrug-resistant malaria in southeast Asia: a genomic epidemiology study.** *Lancet Infect Dis.* 2019; **19**(9): 943–951.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

10. Amato R, Pearson RD, Almagro-Garcia J, *et al.*: **Origins of the current outbreak of multidrug-resistant malaria in southeast Asia: a retrospective genetic study.** *Lancet Infect Dis.* 2018; **18**(3): 337–345.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

11. Imwong M, Suwannasin K, Kunasol C, *et al.*: **The spread of artemisinin-resistant** *Plasmodium falciparum* **in the Greater Mekong subregion: a molecular epidemiology observational study.** *Lancet Infect Dis.* 2017; **17**(5): 491–497.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

12. Imwong M, Hien TT, Thuy-Nhien NT, *et al.*: **Spread of a single multidrug resistant malaria parasite lineage (***PfPailin***) to Vietnam.** *Lancet Infect Dis.* 2017; **17**(10): 1022–1023.
    PubMed Abstract | **Publisher Full Text**

13. Wicht KJ, Mok S, Fidock DA: **Molecular Mechanisms of Drug Resistance in** *Plasmodium falciparum* **Malaria.** *Annu Rev Microbiol.* 2020; **74**: 431–454.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

14. **WHO recommends groundbreaking malaria vaccine for children at risk**.
    Reference Source

15. Heppner DG, Kester KE, Ockenhouse CF, *et al.*: **Towards an RTS,S-based, multi-stage, multi-antigen vaccine against falciparum malaria: progress at the Walter Reed Army Institute of Research.** *Vaccine.* 2005; **23**(17–18): 2243–2250.
    PubMed Abstract | **Publisher Full Text**

16. Datoo MS, Natama HM, Somé A, *et al.*: **High Efficacy of a Low Dose Candidate Malaria Vaccine, R21 in 1 Adjuvant Matrix-M™, with Seasonal Administration to Children in Burkina Faso**. 2021.
    Publisher Full Text

17. RTS,S Clinical Trials Partnership: **Efficacy and safety of RTS,S/AS01 malaria vaccine with or without a booster dose in infants and children in Africa: final results of a phase 3, individually randomised, controlled trial.** *Lancet.* 2015; **386**(9988): 31–45.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

18. WHO: **False-negative RDT results and implications of new reports of** *P. falciparum histidine-rich protein 2/3* **gene deletions: information note.** *False-Negat RDT Results Implic New Rep P Falciparum Histidine-Rich Protein 23 Gene Deletions Inf Note.* 2016.
    Reference Source

19. Scherf A, Mattei D: **Cloning and characterization of chromosome breakpoints of** *Plasmodium falciparum*: **breakage and new telomere formation occurs frequently and randomly in subtelomeric genes.** *Nucleic Acids Res.* 1992; **20**(7): 1491–1496.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

20. Zhang X, Alexander N, Leonardi I, *et al.*: **Rapid antigen diversification through mitotic recombination in the human malaria parasite** *Plasmodium falciparum*. *PLoS Biol.* 2019; **17**(5): e3000271.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

21. Roy SW, Ferreira MU, Hartl DL: **Evolution of allelic dimorphism in malarial surface antigens.** *Heredity (Edinb).* 2008; **100**(2): 103–110.
    PubMed Abstract | **Publisher Full Text**

22. Miles A, Iqbal Z, Vauterin P, *et al.*: **Indels, structural variation, and recombination drive genomic diversity in** *Plasmodium falciparum*. *Genome Res.* 2016; **26**(9): 1288–1299.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

23. Ware LA, Kain KC, Lee Sim BK, *et al.*: **Two alleles of the 175-kilodalton** *Plasmodium falciparum* **erythrocyte binding antigen.** *Mol Biochem Parasitol.* 1993; **60**(1): 105–109.
    PubMed Abstract | **Publisher Full Text**

24. Wendler JP: **Accessing complex genomic variation in Plasmodium falciparum natural infections**. (Oxford University, UK, 2015).
    Reference Source

25. Manske M, Miotto O, Campino S, *et al.*: **Analysis of** *Plasmodium falciparum* **diversity in natural infections by deep sequencing.** *Nature.* 2012; **487**(7407): 375–379.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

26. Li H, Durbin R: **Fast and accurate short read alignment with Burrows–Wheeler transform.** *Bioinformatics.* 2009; **25**(14): 1754–1760.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

27. DePristo MA, Banks E, Poplin R, *et al.*: **A framework for variation discovery and genotyping using next-generation DNA sequencing data.** *Nat Genet.* 2011; **43**(5): 491–498.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

28. Cingolani P, Platts A, Wang LL, *et al.*: **A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w[1118]; iso-2; iso-3.** *Fly (Austin).* 2012; **6**(2): 80–92.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

29. Echeverry DF, Deason NA, Davidson J, *et al.*: **Human malaria diagnosis using a single-step direct-PCR based on the** *Plasmodium* **cytochrome oxidase III gene.** *Malar J.* 2016; **15**: 128.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

30. MalariaGEN: **Pf7: an open dataset of Plasmodium falciparum genome variation in 20,000 worldwide samples.** *figshare.* Dataset. 2022.
    https://www.doi.org/10.6084/m9.figshare.21674321.v2

Open Peer Review

Current Peer Review Status: ✓ ✓ ✓

Version 1

Reviewer Report 21 February 2023

https://doi.org/10.21956/wellcomeopenres.20716.r54149

✓   **Cristian Koepfli** [iD]

Department of Biological Sciences, Eck Institute for Global Health, University of Notre Dame, Notre Dame, IN, USA

This manuscript presents the latest version of the MalariaGEN P. falciparum dataset, with over 20,000 whole genomes. Given the very rich dataset, the manuscript necessarily needs to focus on brief analysis of a few key aspects, namely the diversity and population structure of markers of drug resistance, hrp2/3 deletion, and (potential) vaccine candidates csp and eba175.

While the results are overall clearly described despite the brevity, the global population structure shown in Figure 1 is difficult to follow. The authors define ten major sub-populations. Yet, in both PCoA plots shown, the separation of several populations is not apparent. For example, the four African populations seem to overlap, as do the two populations from SE-Asia. In contrast, the Oceania-New Guinea population seems to split into two (I am surprised there is no comment on that). Given the grouping of the sub-populations is maintained to describe the analysis throughout the manuscript, the authors might consider adding panels with further components to better show separation of sub-populations. Or, maybe, a PCoA with African isolates only might make the separation more evident.

As further point for clarification, in the section on sWGA the authors the analysis of face-away reads. This term seems not to be commonly known among people working with WGS data. Do the authors mean reads spanning the break-points? This section seems rather technical compared to the remainder of the manuscript.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Molecular epidemiology, diagnosis of P. falciparum and P. vivax, genotyping, amplicon sequencing, hrp2/2 deletion typing, gametocyte quantification, transmission

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

Reviewer Report 07 February 2023

✔ **David A. Fidock** (iD)

Department of Microbiology and Immunology, Columbia University Irving Medical Center, New York, NY, USA

This study by the MalariaGEN consortium and multiple partners across the malaria-endemic world is exceptional. The authors report the results of a remarkable study that genetically profiled over 20,000 Plasmodium falciparum samples in the current Pf7 database, including an additional 13,752 samples compared to their most recent data iteration (Pf6), adding an additional 33 partner sites and five countries. The study contains whole-genome sequencing data from over 16,000 samples, which required selective whole-genome amplification (sWGA) to generate sufficient data reads from the limited amounts of parasite DNA that could be recovered from dried blood spots. The authors describe a careful optimization and analysis of the methodology employed, as sWGA creates some uneven distribution in the sample reads across the genome. The authors conclude that this method is sufficiently reliable that it enables calling copy number variants across the genome, which is a new and important feature to this new analysis and data release.

The published article version highlights a few of the most important findings, including:
- Figure 1: population sub-structuring across their samples.

○ Figure 2: Heterogeneity of chloroquine-resistant P. falciparum across west Africa (as determined using the pfcrt K76T mutation that is a well validated genetic marker). This finding likely reflects different chloroquine and potentially amodiaquine usage across the region.

○ Figure 3: The finding of multiple new pfcrt mutations in Southeast Asia, arising on the chloroquine-resistant Dd2 background, which in a number of cases reflects a key role for these mutations in contributing to piperaquine resistance. A number of these mutations are also likely to compensate for fitness costs caused by pfcrt mutations.

○ Figure 4: Evidence of a high rate of non-synonymous mutations in CSP, which is important given that a segment of this protein forms the core parasite antigen used in the WHO-approved RTS,S and the clinically trialed R21 vaccines. That figure would benefit from additional annotation to show the location of the segments used in these vaccines.

○ Figure 5: Evidence of breakpoints that cause deletions of HRP2 and HRP3, whose detection is an essential component of the rapid diagnostic tests commonly used to diagnose malarial infection. Here, it would be good to point out the absence of samples from the Horn of Africa, where this is a major problem. That description should cite the relevant literature, e.g. PMID 34580442[1].

○ Table 1 lists the new samples by region.

○ Table 2 is a very useful list of the prevalence of genetic markers of resistance to multiple drugs. When citing amplification of plasmepsins 2 and 3 as a marker of piperaquine resistance, the authors should cite the two initial reports of their association: PMID 27818095[2] and 27818097[3]. This lists, by necessity, excludes pfcrt mutations as, unlike for chloroquine, there is no single mutation that is essential and instead there are several that can contribute to resistance. The authors cite a review but could include a recent study that provides more insight into mechanism: PMID 31776516[4].

The real power of this study can be found through their searchable website: https://www.malariagen.net/resource/34. This resource will provide a superb tool for the malaria research community. Now that this article is published, the following link can be updated to reflect that data release should no longer be forthcoming: https://www.malariagen.net/apps/pf7/. It will be important for readers to be able to interrogate the database in an easy way to collate information on variation in a particular gene. I was not able yet to figure out how to generate this search and more instruction on this point would be very helpful.

This is a remarkable study in the history of malaria research. The entire team should be commended for this extraordinary dataset.

References

1. Feleke SM, Reichert EN, Mohammed H, Brhane BG, et al.: Plasmodium falciparum is evolving to escape malaria rapid diagnostic tests in Ethiopia.*Nat Microbiol*. 2021; **6** (10): 1289-1299 PubMed Abstract | Publisher Full Text

2. Amato R, Lim P, Miotto O, Amaratunga C, et al.: Genetic markers associated with dihydroartemisinin-piperaquine failure in Plasmodium falciparum malaria in Cambodia: a

genotype-phenotype association study.*Lancet Infect Dis*. 2017; **17** (2): 164-173 PubMed Abstract | Publisher Full Text

3. Witkowski B, Duru V, Khim N, Ross LS, et al.: A surrogate marker of piperaquine-resistant Plasmodium falciparum malaria: a phenotype-genotype association study.*Lancet Infect Dis*. 2017; **17** (2): 174-183 PubMed Abstract | Publisher Full Text

4. Kim J, Tan YZ, Wicht KJ, Erramilli SK, et al.: Structure and drug resistance of the Plasmodium falciparum transporter PfCRT.*Nature*. 2019; **576** (7786): 315-320 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: Dr. Vandana Thathy, an author on this study who worked on this study in her previous position, is now a senior member of my group. I have nonetheless been able to remain impartial for this review.

Reviewer Expertise: Malaria, drug resistance, therapeutics mode of action, genetics, mechanism, gene editing, Plasmodium falciparum,

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

Reviewer Report 02 February 2023

https://doi.org/10.21956/wellcomeopenres.20716.r54147

✅

Fabián Sáenz [iD]

Centro de Investigación para la Salud en América Latina, Facultad de Ciencias Exactas y Naturales, Pontificia Universidad Católica del Ecuador, Quito, Ecuador

This manuscript by MalariaGEN and collaborators describes a new release set of genomic data called Pf7. The data includes curated genomic data of more than 20,000 *Plasmodium falciparum* samples, out of which 13750 were new for the Pf7 set. Samples were obtained from 33 countries in 4 continents collected between 1984 and 2018. Most of the new samples came from dried blood spots for which SWGA was used (and used a GATK based method to normalize data around certain genes). Variants were called and incomplete samples discarded. The authors divided the samples in 10 sub populations that matched quite well sample locations based in principal coordinate analysis and neighbor joining trees (four in Africa, four in Asia, one in Oceania and one in South America). In order to show relevance of the data, some interesting analysis specific to drug resistance, vaccine target diversity, rapid diagnosis test target deletion and invasion ligands was presented. Of particular interest, samples were classified according to the resistance level to some of the main drugs using well known markers and show how drug resistance has moved across SE Asia in unique genetic backgrounds. In addition, the variation in the c-terminal region and background mutations was analyzed and shown to be quite divergent from the reference 3D7 (used in the RTS,S vaccine). In addition to map the deletions of *hrp2* and *hrp3* in the samples around the world. the origin of these deletions is explored and the authors conclude that telomere healing plays an important role in the process. The available data also helps confirm that eba175 is an invasion ligand gene under balancing selection.

This work is a very important addition to the genomic *Plasmodium* datasets available for the scientific community. The data will help the development of new worldwide and regional malaria studies helping elimination in several areas of the world and ultimately helping in the ultimate goal that is eradication.

Here are some minor observations to this work:

Introduction:
- The analysis pipeline should be referenced through the text.

Results:
- Variant discovery and genotyping: When referring to 3D7 v3 reference genome a reference should be included. When referring to the variant discovery pipeline for Pf6, a reference should be included.

Effects of selective whole genome amplification (sWGA):
- The authors are well aware that variability in coverage across the genome can arise from sWGA and develop a method based on gCNV. It can be seen that for some genes such as mdr1, concordance was low. Would this method be usable for other markers in the genome? Limitations in this sense should be further discussed.

- To show that sWGA of DBS samples is not introducing a bias in population structure, the authors compare clustering of the same samples from whole blood and DBS. They do not see stratification by sample type (supl fig 3). From this figure, it is not clear to me that clustering is similar in both cases since samples of both collection types are analysed together. In addition, it is not clear how sample type clustering would be sufficient to

demonstrate no bias in population structure due to sWGA. It would be important to have further insight from the authors in this sense.

Global population structure:

- ○ In figure 1B and supplementary figure 4 it is of particular interest that some African samples and South American isolates appear to be more related than African samples between themselves, but this is not commented by the authors. Could the authors comment in this sense.

Genetic origins of *hrp2* and *hrp3* deletions:

- ○ The first sentence in this section reads: "Most widely used rapid diagnostic tests rely on detection of the products of the hrp2 and hrp3 genes". In my opinion this should be clarified or written differently since RDTs are designed to detect the product of *hrp2* gene but no *hrp3*. The deletion of hrp3 may affect the outcome of RDTs if *hrp2* is deleted because of the similarity between both of the products.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Infectious diseases: Malaria genetics and genomics

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