Title: Genomic characterisation reveals a dominant lineage of SARS-CoV-2 in Papua New Guinea

Authors: Theresa Palou[*1], Mathilda Wilmot[*2], Sebastian Duchene[3], Ashleigh Porter[3], Ms Janlyn Kemoi[4], Dagwin Suarkia[5], Patiyan Andersson[2], Anne Watt[2], Norelle Sherry[2], Torsten Seemann[2], Michelle Sait[2], Charlie Turharus[6], Son Nguyen[8], Sanmarié Schlebusch[8], Craig Thompson[8], Jamie McMahon[8], Stefanie Vaccher[7], Chantel Lin[2], Esorom Daoni[1], Benjamin P Howden[#2], Melinda Susapu[#1]

Affiliations:

[1]National Control Centre, Ministry of Health, Papua New Guinea

[2]Microbiological Diagnostic Unit Public Health Laboratory, The University of Melbourne, at The Peter Doherty Institute for Infection and Immunity, Melbourne, Australia

[3]Department of Microbiology and Immunology, The University of Melbourne, at The Peter Doherty Institute for Infection and Immunity, Melbourne, Australia

[4]Central Public Health Laboratory, Port Moresby, Papua New Guinea

[5]Institute of Medical Research, Goroka, Papua New Guinea

[6]Ok Tedi Mining Limited, Papua New Guinea

[7]Independant consultant, Papua New Guinea

[8]Forensic and Scientific Services, Queensland Health, Brisbane, Australia

[*]Equal contributions

[#]Equal contributions

Contact: Prof Benjamin Howden; Microbiological Diagnostic Unit Public Health Laboratory, The University of Melbourne, at The Peter Doherty Institute for Infection and Immunity, 792 Elizabeth St, Melbourne, 3000, Australia

bhowden@unimelb.edu.au

Abstract

The COVID-19 pandemic has highlighted the utility of pathogen genomics as a key part of a comprehensive public health response to emerging infectious diseases threats, however the ability to generate, analyses and respond to pathogen genomic data varies around the world. Papua New Guinea (PNG), which has limited in-country capacity for genomics, has experienced significant outbreaks of SARS-CoV-2 with initial genomics data indicating a large proportion of cases were from lineages that are not well-defined within the current nomenclature. Through a partnership between in-country public health agencies and academic organisations, industry, and a public health genomics reference laboratory in Australia a system for routine SARS-CoV-2 genomics from PNG was established. Here we aim to characterise and describe the genomics of PNG's second wave and examine the sudden expansion of a lineage that is not well defined but very prevalent in the Western Pacific region. We generated 1797 sequences from cases in PNG and performed phylogenetic and phylodynamic analyses to examine the outbreak and characterise the circulating lineages and clusters present. Our results reveal the rapid expansion of the B.1.466.2 and related lineages within PNG, from multiple introductions into the country. We also highlight the difficulties that unstable lineage assignment causes when using genomics to assist with rapid cluster definitions.

Introduction

Coronavirus disease (COVID-19), caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has resulted in 2.7 million infections and over 40,000 deaths across the Western Pacific region (World Health Organisation 2021b). Papua New Guinea (PNG) was one of the first countries in the region to report a COVID-19 case in March 2020, with 21,896 reported cases and 243 deaths as of 6 October 2021 (World Health Organisation 2021b). Papua New Guinea experienced the first wave of infection and community transmission from April 2020, with the PNG Government moving rapidly to implement a range of public health measures, resulting in successful reduction and control the first wave of infection by August 2020 (The World Bank 2021). Despite this, a rapid increase of COVID-19 cases were detected in PNG in early 2021 resulting in a second wave of infection that saw cases rise from 1583 confirmed cases at the start of March 2021 to 17,774 by the end of July, even with renewed public health control measures.

Papua New Guinea has a population of approximately 8.8 million people living across 22 provinces on the mainland and islands, with 87% of Papua New Guineans living in rural areas. The geographical spread of the population creates significant logistical challenges for diagnostic testing and epidemiological investigation to monitor introduction and transmission of lineages, and surveillance of disease trends over time. Access to diagnostic testing has been variable across the country and hampered by staffing and logistical issues (Smaghi et al. 2021), impacting the ability to monitor and rapidly implement public health measures to reduce expansion of disease spread. The detection of cases and the collection of samples for sequencing by PNG's health system is therefore predominantly from the National Capital District, which encompasses the capital Port Moresby and from the most populous province, Morobe, in which the second largest city in the country, Lae, is located (Figure 1, Table 1). The majority of cases reported in the country, however, are identified through private testing carried out by Ok Tedi Mining Ltd, based in the Western province. As such, despite having only 2.8% of the population, the distribution of cases in PNG is heavily biased to the Western province.

The current typing nomenclature for SARS-COV-2, involves the assignment of lineages that reflect evolutionary relationships and are hierarchically organised following the phylogenetic tree structure. This nomenclature system describes major lineages with letters of the alphabet (e.g. A, B, etc.), with sub- and sub-sub-lineages being numbered and separated by dots (“.”). Thus, sub-lineage B.1.466.2 is contained within sub-lineage B.1.466, which is itself part of lineage B.1 and the direct parent lineage, B. For readability, only three sub-levels are recorded under this nomenclature system and sub-lineages beyond this level will be shorten by aliases using the next available alpha symbol. For instance, B.1.466.2.1 has been assigned the alias AU.1. A PANGO lineage of SARS-Cov-2 may be designated as a variant of concern (VOC) if there is evidence for epidemiological, pathological, or immunological features of concern (Public Health England 2021). These may be designated by international bodies, or potentially observed and designated as variants of concern locally. Currently the WHO classifies four lineages as VOCs: B.1.1.7, B.1.351 (and sublineages), P.1 and B.1.617.2 (World Health Organisation 2021a). All four variants display an unusually high number of mutations, including a number of variations in the genomic region encoding the spike protein thought to have the potential to increase transmissibility or confer immune evasion properties.

Emerging variants of concern and rapid virus evolution requires access to genomic surveillance to support the control and management of the pandemic. Genomic sequencing of SARS-CoV-2 allows for detection and identification of new and emerging lineages and variants of concern, assists with the identification of outbreaks and transmission events to contribute to public health interventions, and allows for an estimate of trends and expansion of disease spread. Here, we aim to characterise the circulating SARS-CoV-2 lineages in PNG and describe the dynamics of a genomic dataset that is unique in the region.

Methods

Genomic and epidemiological data

Positive SARS-CoV-2 samples from cases in PNG were submitted from the PNG Central Public Health Laboratory (CPHL) and Ok Tedi Mining Limited (OTML) to the Microbiological Diagnostic Unit Public Health Laboratory (MDU PHL), at the Doherty Institute, Melbourne, for genome sequencing, analysis, and integrated reporting. Ok Tedi Mining Limited operates predominantly in the Western Province of PNG, a remote, sparsely populated area bordering Indonesia. While only 2.8% of PNGs population reside in the Western Province, OTML routinely transports workers in and out of the mining sites, and sends samples collected as part of their workplace testing program, to Australia for diagnostic testing. All positive samples were referred to MDU PHL for sequencing. Samples referred from CPHL represent a subset of available samples, selected on the bases of temporal and geographic diversity and sample quality, and were sent directly to MDU PHL. Forensic Scientific Services (FSS) at Queensland Health also performed sequencing on additional PNG samples, submitted by CPHL. These sequences were shared with MDU PHL as part of a collaborative analysis agreement under the governance of the PNG NCC.

Limited epidemiological data were provided alongside the samples by the PNG NCC and by OTML. There are currently a number of challenges with COVID-19 data collection and recording in PNG and with the epidemiological data, resulting from incomplete or manually transcribed epidemiological records. Ok Tedi Mining Limited provided information on case nationality, whether a case was tested on arrival to the mining site (in-bound), or whether they were tested whilst working on site (outbound and monitoring). For non-OTML cases, the PNG NCC provided data on geographical location of a case, including province, region within province and town/village as well as information on symptoms, case contact (where known) and occupation. A case was assigned to a geographical province within PNG based on the data provided by the PNG NCC, or where that was unavailable, from the data provided by OTML.

Detailed genomics methods are described in Seemann et al. 2020 and Lane et all 2021 . Briefly, RNA extracted from SARS-CoV-2 RT-PCR positive samples underwent tiled amplicon

PCR using either ARTIC version 1 or 3 primers ("ARTIC-Ncov2019/Primer_schemes/NCoV-2019/V3 at Master · ARTIC-Network/ARTIC-Ncov2019 · GitHub" n.d.), following published protocols ("NCoV-2019 Sequencing Protocol" n.d.). Reads were aligned to the reference genome (Wuhan Hu-1; GenBank MN908947.3) and consensus sequences generated. Quality control (QC) metrics on consensus sequences included requiring ≥50% genome recovered (≥95% in FSS pipeline setting), ≤50 single nucleotide polymorphisms (SNPs) from the reference genome, and ≤50 ambiguous or missing bases. Genomic clusters were defined as two or more related sequences using a complete-linkage hierarchical clustering algorithm of pairwise genetic distances derived from a maximum likelihood phylogenetic tree. SARS-CoV-2 genomic lineages were defined using the PANGO lineage nomenclature (Rambaut et al. 2020; SARS-CoV-2 lineages).

Genomic epidemiology and phylodynamics

To quantify the dynamics of introductions, we used a set of 1587 genome samples from PNG (Supplementary Appendix B). This dataset included the genomes generated in this study with sufficient sequence quality and an associated date of collection, and a sample of global genomic diversity focussed on the region of Oceania by using the latest NextStrain Oceania build, that included of 489 genomes from other countries (as of 20 March 2021). We aligned the sequences using MAFFT v7 (Katoh and Standley 2013).

We use a previous approach (Duchene et al. 2020b) to obtain a time-scaled phylogenetic tree (Duchene et al. 2020a; To et al. 2016). We defined "genomic importation clusters" as monophyletic groups of at least two genomes sampled from PNG, whereas a "singleton" is a genome sampled from PNG that sits within a group of genomes sampled elsewhere. An importation cluster therefore corresponds to a putative introduction event that led to ongoing transmission, whereas a singleton represents a situation where there is no evidence of ongoing transmission (du Plessis et al. 2021). Importantly, whether an importation cluster corresponds to a single importation event is contingent on the data at hand. If the geographic area of interest is sampled at a much higher intensity that other areas, as is the case here, the number of importation clusters will tend to be an underestimate of the

number of importation events that gave rise to the data, such that they should be considered as a lower bound.

We calculated a range of genomic importation clusters statistics from the time-scaled tree. We focused on the number of importation lineages, their detection date, first introduction, putative importation date and the detection lag (the time from the origin of the importation cluster to the date of collection of the first genome). For the largest four genomic importation clusters we fit a coalescent exponential model in a Bayesian framework in BEAST2.5 (Bouckaert et al. 2019) to infer their exponential growth rate, sampling proportion and doubling time. The xml file, dated tree, and GISAID accession numbers are available at https://github.com/sebastianduchene/png_sars_cov_2_analyses.

Results

We sequenced 2981 positive samples at MDU PHL and FSS, collected up to 13 July 2021, yielding 1797 sequences that met internal quality control (QC) measures. Sequences used in this study are listed in Supplementary Appendix A. In total, 1184 samples failed internal QC and were not included in the phylogenetic analyses. From the 1797 samples that passed QC, 1672 were successfully linked to the epidemiological metadata provided by OTML and the PNG NCC. Of the samples with available epidemiological data, 59% (1053 of 1672 samples) were from the Western Province in PNG (the location of OTML operations), 14% (259 of 1672 samples) from the National Capital District, 6% (113 of 1672 samples) from Morobe and 4% (69 of 1672 samples) from East New Britain (Figure 1). The remaining samples (178 of 1672 samples) span 16 other provinces (Supplementary Appendix A).

Lineages

PANGO lineage assignment on the 1797 samples from PNG was found to be highly unstable, with constant shifts in the assignment of large numbers of samples across Pangolin versions, particularly across three highly related lineages, B.1.466/B.1.459/AU lineages (Table 2). Samples were frequently reassigned across and within the lineage groups, regardless of genome coverage or sequence quality.

Eighty-eight percent (1580/1797) of PNG sequences were identified as either AU.1, AU.3, B.1.466.2 or B.1.459 (Table 2). These five, highly related, lineage groups are associated with the Pacific and Southeast Asian region, particularly Indonesia, Malaysia, PNG and Australia, with the B.1.466.2 clade first proposed for definition by FSS and Queensland Health, after a rise in cases in returned travellers from PNG (16). AU.1/AU.2/AU.3 are all aliases of the B.1.466.2 sub-lineages, whilst B.1459 appears highly related to B.1.466.2/AU on phylogeny. Additionally, 2.4% (43/1797) of sequences typed as B, the first major haplotype to be discovered and B.1 (Table 2), a large European lineage linked to the Northern Italian outbreak in 2020 (17). The assignment of these recent samples to an early lineage is likely the result of limited analysis and sample representation in this area of the global tree and not the true persistence of such early versions of the virus. Five percent (94/1797) of sequences typed as B.6/B.6.8, early lineages predominantly seen in India (B.6) and PNG

(B.6.8). Despite the surge in cases seen in PNG during this period, and the large on-going outbreak, only one sample had been identified as a variant of concern (Delta- B.1.617.2) by 29 July 2021. Lineages for the remaining sequences are available in Appendix B.

Phylogenetic clusters

We performed a phylogenetic analysis and included publicly available sequences from the Solomon Islands, the Philippines, Guam, Timor-Leste, Australia, and Indonesia as well as publicly available PNG sequences, for context (Supplementary Appendix C, Figure 2). Five broad clusters were identified (Figure 3), containing a mix of lineages including intermingling of the AU, B.1.466.2, B.1.459 and B.1 samples within clusters, and closely related samples typing as different PANGO lineages (Figure 4).

Analysis of the temporal distribution of the phylogenetic clusters and PANGO lineages shows a shift from the B.6/B.6.8 lineages in mid-2020, to the described B.1 and AU/B.1.466.2/B.1.459 lineages in early 2021 (Figure 4). All B.6 and B.6.8 sequences identified in this data set cluster together ('cluster 1', Figure 2) and were collected between 2020-06-17 and 2021-03-24 (Figure 4). The majority (51%) of samples within this cluster with a recorded collection date were collected prior to 2020-12-21. No other lineages were found in 2020 samples, either in the data described in this paper or in the publicly available PNG sequences.

Despite the majority of samples in the data set originating in the Western Province or National Capital District, the phylogenetic clusters identified in this analysis were geographically diverse, with each of the clusters appears concentrated in different areas of PNG (Figure 3). The largest cluster, 'cluster 2' (Figure 2), appears to be connected to the OTML mine sites and the Western Province, whilst the smaller clusters appear linked to the National Capital District and larger surrounding provinces ('cluster 3'), the island of New Britain ('cluster 5') or spread from the highland provinces across to New Britain ('cluster 4').

Phylogenetic analysis of putative introductions

We estimate that there have been at least 55 introduction events into PNG based on the available genomic data (Supplementary table 1; Figure 5). Only three of these introductions

consisted of a single case, with no evidence of ongoing transmission. Importantly, the importation clusters were largely consistent with the broad genomic clusters identified above. We found that 24 genome importation clusters had at least five sequences, with the largest having 926 sequences included.

The first genomic importation cluster with at least five genomes was detected on 19 July 2020, while the last was detected on 9 March 2021. These estimated dates are likely to be later than the actual importation events, because the genomic signal lags behind actual introductions (du Plessis et al. 2021). Under this framework, we estimate that genome importation clusters with at least 5 genomes were introduced between February 2020 and March 2021. Their respective detection lags had a mean of 18 days (range from 1 day to three months). The largest cluster, with 926 sequences included, was probably introduced around mid-December 2020 and it was detected on 1st January 2021, with a detection lag of 20 days. The detection lag was shortest at the peak of the second wave in April and May 2021, with a mean of one day. We also estimate that most importations events occurred around March 2021.

The largest genomic importation cluster mostly consisted of PANGO lineages B.1.466.2.1 (AU.1), B.1.459, and B.1.466.2.3 (AU.3), with 387, 256, and 198 genomes respectively, such that these three lineages represented over 90% of all the genomes in the cluster.

Phylodynamic analyses of genomic importation clusters

We used a coalescent framework to infer population dynamic parameters for the four largest genome importation clusters. Our estimates of the coalescent growth rate were very similar among clusters at around 28 $year^{-1}$, which roughly corresponds to a reproductive number, R_e, of 2.5. The 95% credible interval of the four clusters excluded a 0, such that they all have evidence of epidemic growth. The corresponding doubling times overlapped for all genome importation clusters. The largest importation cluster, A, had the longest doubling time, at 9 days (95% credible interval: 8 to 11), while the smallest cluster, B, had the shortest time, at 8 days (95% credible interval: 7 to 10). We also estimated the sampling intensity, which is the number of genomes divided by the inferred infected population size when the last sample was collected. These estimates were very uncertain and below 0.02

(2%), with cluster A having the highest sampling intensity, at 0.011 (95% credible interval: 0.003 to 0.03). Although these estimates are very uncertain, probably due to the low genetic diversity, they suggest that the genome sampling represents a very small proportion of the outbreak associated with each importation cluster (Figure 6).

Discussion

In total, 1797 sequences generated by MDU PHL and FSS from PNG SARS-CoV-2 cases underwent PANGO lineage assignment and phylogenetic analysis to characterise the lineage distribution and genomic relatedness of SARS-CoV-2 in PNG. Analysis of the lineages within this data set found only one VOC sample present, however the lineages that have been identified are not well characterised by the Pangolin nomenclature, with intermingling of multiple lineages in the phylogenetic tree, and closely related samples and clusters containing numerous assignments.

Phylogenetic analysis of clusters and importations in the data generated at MDU PHL shows a marked shift in the lineage distribution and has identified 55 importation clusters, the majority of which resulted in multiple cases. Due to natural sampling biases in our data, the actual number of viral introductions is likely much higher. These importation clusters are consistent with the broad clusters we have described and the substructure within each of these. The results suggest that while the first introduction in July 2020 resulted in a large B.6/B.6.8 cluster ('cluster 1') this was rapidly replaced in 2021 with four distinct clusters made up of B.1 and AU sub-lineages, likely from multiple introductions. However, phylodynamic analysis of the data suggests that the sequences presented here represent a very small proportion of the likely cases associated with each cluster. This correlates with the known testing and sampling challenges within PNG and with the reported epidemiology of the COVID-19 outbreak, where a peak and then drop in case numbers in mid-late 2020 was followed by a sudden increase in early 2021, leading to the large-scale outbreak from which these sequences were predominantly sampled (World Health Organisation 2021b).

This data suggests that there has been rapid expansion and geographical spread of lineages in PNG (B.1.459, B.1.466.2 and AU) that are not recognised as a VOC or VOI and that there was effective replacement of B.6/B.6.8 with the currently circulating PANGO lineages. Publicly available sequences suggest that these lineages identified in PNG are also commonly observed in other countries in the region, particularly Indonesia (Cahyani et al. 2021; Zainulabid et al. 2021) which may explain why the B.1.466.2 and AU lineages are persisting and present in all unrelated clusters, despite multiple introductions into the country. The presence of only one VOC sample in this dataset suggests that at the end of

July 2021, the burden of disease in PNG was still predominantly caused by the B.1.466.2, B.1.459 and AU lineages. However, the sampling issues described above mean this is possibly an under-representation of the level of Delta present within the community at this time.

The characterisation of lineage distribution in PNG is made difficult by the described issues in lineage assignment and stability in this area of the tree. Large numbers of the PNG sequences type as early lineages ('B.1') and lineage assignment frequently, including a large proportion of samples that routinely switch between AU/B.1.466.2 and B.1.459. This impacts the utility of the genomics and prevents PNG from tracking the spread and transmission of SARS-CoV-2, without detailed genomic investigation, a process that is difficult given resource constraints within PNG. We would therefore argue for closer examination of this area of the global SARS-CoV-2 phylogeny, to resolve the classification issues for lineages routinely seen in the Western Pacific region (O'Toole et al. 2021).

This dataset provides a significant amount of new genomic data in an under sampled region (Chong et al. 2020) where attempts at representative sequencing have been hampered by resource and logistical issues (Kabuni 2020; Smaghi et al. 2021). The data presented here is relevant to the entire Western Pacific region as it shows how quickly lineages in the region can take hold, regardless of official VOC status and how issues related to under representation in databases like PANGO, can impact work being done in countries like PNG. However, we acknowledge the limitations of this data, including; the high sequencing failure rate, possibly due to age of samples on arrival in Australia, samples with low viral load, or issues with sample storage during transport; bias in sampling sites and regions within PNG; and the impact that limited testing has on the representativeness of this dataset. Our analysis was also impacted by the limited epidemiological data available to provide context for phylogenetic clusters, the time lag from collection to sequencing, and by the logistical constraints that mean only a small proportion of swabs from an already under sampled population can be sent for sequencing.

The genome sequencing and bioinformatic analyses for this program of work were undertaken offshore at MDU PHL in Australia, however significant consideration was given

to the training opportunities that this model of work afforded. While a longer-term goal will be in-country deployment of sequencing capacity, during this program of work, significant training in genomic sampling strategies, genomic and epidemiological data governance, combined genomic and epidemiological data analysis, and genomic reporting for public health were undertaken. The international referral of samples was identified as the only rapid, short-term solution for rapid generation of genome sequence data early in the pandemic from PNG, however the partnership between the laboratories and National Coordinating Centre in PNG and the offshore counterparts has significantly improved knowledge on the approach and use of genome sequence data which will inform future in-country strategies and improve the likelihood of success.

Analysis of a small set of sequences from SARS-CoV-2 cases in PNG has provided insight into how quickly lineages can take hold in a country or region, particularly where testing and response resources are limited. The on-going sequencing work with PNG also highlights the need for curation of PANGO lineages in all areas of the global SARS-CoV-2 tree to ensure stability in linage assignment, enabling countries with limited ability to undertake detailed genomic analysis to still utilise this important public health tool for outbreak and cluster characterisation. This has also demonstrated the value of equitable access to advanced technologies, including genomic sequencing, for informing public health decisions, particularly when necessary to rapidly identify or characterise certain pathogens.

Data Availability

SARS-CoV-2 genome sequences generated in this study have been deposited in the GISAID platform (https://www.gisaid.org/), accession number IDs are available in Supplementary Appendix C.

Acknowledgements:

We acknowledge and thank all the SARS-CoV-2 diagnostic and sequencing laboratories working in the region, for their contributions to this work. Sequence data and authors are available on GISAID and listed in Supplementary Appendix C.

Author contributions:

MS, TP, NS and BH implemented, established governance, and supervised the genomics program. TP, MS, JK, SV, DS, CT and EM provided the samples, epidemiological data and supported the capacity building for interpretation of genomic results. MW, PA, AW, NS, TS, MS, CL, BP, SS, CT, JM and SN generated, analysed and reported genomic sequence data. SD and AP conducted phylodynamic analyses. MW, SD, AP, TP, SV, CL and BH contributed to the preparation of the manuscript.

Funding:

The Australian Government Department of Foreign Affairs and Trade's Centre for Health Security and the PNG-Australia Transition to Health Initiative have provided funding to support WGS services provided by the MDU PHL. Ok Tedi Mining Limited also supported this work.

Tables

Table 1. PNG samples sent to Australia for sequencing by province of collection and proportion of the population the resides in each province for comparison.

REGION	NUMBER OF SAMPLES SENT FOR SEQUENCING	POPULATION BY % OF PNG TOTAL[1]
HIGHLANDS PROVINCES		
EASTERN HIGHLANDS	1 (0.03%)	8.00%
ENGA	6 (0.2%)	5.90%
HELA	27 (0.9%)	3.40%
JIWAKA	6 (0.3%)	4.70%
SIMBU (CHIMBU)	38 (1.3%)	5.20%
SOUTHERN HIGHLANDS	41 (1.4%)	7.00%
WESTERN HIGHLANDS	41 (1.4%)	5.00%
MOMASE REION		
EAST SEPIK	1 (0.03%)	6.20%
MADANG	1 (0.03%)	6.80%
MOROBE	137 (4.6%)	9.30%
SANDAUN (WEST SEPIK)	0	3.40%
SOUTHERN REGION		
CENTRAL	36 (1.2%)	3.70%
GULF	15 (0.5%)	2.20%
MILNE BAY	0	3.80%
NATIONAL CAPITAL DISTRICT	496 (16.6%)	5.00%
NORTHERN PROVINCE (ORO)	1 (0.03%)	2.60%
WESTERN PROVINCE	1812 (60.8%)	2.80%
ISLAND REGIONS		
BOUGAINVILLE (AUTONOMOUS REGION)	9 (0.3%)	3.40%
EAST NEW BRITAIN	95 (3.2%)	4.50%
MANUS	3 (0.1%)	0.80%
NEW IRELAND	16 (0.5%)	2.70%
WEST NEW BRITAIN	18 (0.6%)	3.60%

[1] Based on 2011 census data (National Statistical Office of Papua New Guinea 2011)

Table 2. Number of samples and mutational profile of lineages in PNG dataset

Lineage	Samples (n)	Characteristic Mutations[1]	
		Gene	Amino Acid
AU.1	**507**	N	T205I
		ORF1a	A776V
		ORF1a	P804L
		ORF3a	Q57H
		ORF8	S84L
		S	D614G
		ORF1b	P314L
		ORF1a	P1640L
		ORF1a	T1168I
		ORF10	P10S
		ORF1b	R2308C
		ORF1a	A690V
AU.3	**444**	N	T205I
		ORF1a	T2615I
		ORF1b	P314L
		ORF8	S84L
		S	P681R
		N	D348H
		ORF1a	S944L
		ORF1a	P1640L
		ORF1b	S1182L
		S	D614G
		ORF3a	Q57H
		ORF1a	L3644F
		ORF1a	T1168I
		ORF1b	T2040I
		S	N439K
B	**20**	ORF8	S84L
B.1	**23**	ORF8	S84L
		S	D614G
		ORF1b	P314L
B.1.459	**532**	ORF8	S84L
		S	D614G
		ORF1b	P314L
		ORF1a	P1640L
		ORF3a	Q57H
B.1.466.2	**148**	N	T205I
		S	D614G
		S	N439K

		ORF1b	P314L
		ORF8	S84L
		ORF3a	Q57H
		ORF1a	T1168I
		ORF1a	P1640L
		ORF1b	S1182L
		S	P681R
		ORF1a	S944L
		ORF1a	L3644F
B.6	**95**	ORF8	S84L
		ORF1b	A88V
		ORF1a	T2016K
		N	P13L
		ORF1a	L3606F
B.6.8	**2**	N	P13L
		ORF1a	T2016K
		ORF1b	A88V
		ORF8	L95F
		ORF8	S84L

[1]Data from GISAID and Outbreak.info (Julia L. Mullen 2020)

Figure Legends

Figure 1: Map of PNG showing administrative provinces and the proportion of samples originating from each, in this dataset.

Figure 2: Phylogenetic tree showing PNG samples in the context of publicly available international sequences from the Solomon Islands, the Philippines, Guam, Timor, Australia, and Indonesia. PNG sequences generated at MDU PHL and FSS are shown by the coloured tips.

Figure 3: PNG province of sequence origin, by phylogenetic cluster and date of collection. The described phylogenetic clusters are represented by different colours, with the size of the circle proportional to the number of samples collected in each province on that day. Note; WP = Western Province; WNB= West New Britain; WHP = Western Highlands Province; SHP = Southern Highlands Province; NOP = Northern (Oro) Province; NIP = New Ireland Province; NCD = National Capital District; MOR = Morobe; Man = Manus; MAD = Madang; JIW = Jiwaka; HLP = Hela Province; GF= Gulf; ESP = East Sepik; ENB = East New Britain; CHI = Chimbu (Simbu) ; CEP = Central Province; AROB = Autonomous Region of Bouganville

Figure 4: Timeline of each of the described phylogenetic clusters identified in the PNG sequence dataset as 2021-07-29. The different lineages identified in each cluster are represented by colour, while the size of the circle is proportional to the number of samples in each cluster collected on that day.

Figure 5: Phylogenetic analyses of importation clusters from maximum-likelihood dated trees. Top panel: bars corresponds to importation clusters, the y-axis denoting the number of genomes and their time span along the x-axis. Blue dots correspond to the first genome collected and green is the last genome from each cluster. Bottom panel: importation dynamics over time. The grey bars denote the number of importation events per month, while the orange bars show the detection lag; the number of days from the first inferred transmission event to the first collected genome.

Figure 6: Epidemiological estimates from top four importation clusters. Violin plots denote Bayesian posterior distributions of key parameters, the growth rate, epidemic doubling time, and the sampling intensity (number of genomes per infected case). In the first panel (growth rate) the dashed lines denote the corresponding values for reproductive numbers (R_e) of 1.5 and 2.5 assuming a duration of infection of 10 days.

References

Bouckaert, Remco, et al. (2019), 'BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis', *PLOS Computational Biology,* 15 (4), e1006650.

Cahyani, Inswasti, et al. (2021), 'Genome Profiling of SARS-CoV-2 in Indonesia, ASEAN, and the Neighbouring East Asian Countries: Features, Challenges, and Achievements', *bioRxiv*, 2021.07.06.451270.

Chong, Yoong Min, et al. (2020), 'SARS-CoV-2 lineage B.6 was the major contributor to early pandemic transmission in Malaysia', *PLoS neglected tropical diseases,* 14 (11), e0008744-e44.

du Plessis, Louis, et al. (2021), 'Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK', *Science,* 371 (6530), 708-12.

Duchene, Sebastian, et al. (2020a), 'Temporal signal and the phylodynamic threshold of SARS-CoV-2', *Virus Evolution,* 6 (2).

Duchene, Sebastian, et al. (2020b), 'The impact of early public health interventions on SARS-CoV-2 transmission and evolution', *medRxiv*, 2020.11.18.20233767.

Julia L. Mullen, Ginger Tsueng, Alaa Abdel Latif, Manar Alkuzweny, Marco Cano, Emily Haag, Jerry Zhou, Mark Zeller, Emory Hufbauer, Nate Matteson, Kristian G. Andersen, Chunlei Wu, Andrew I. Su, Karthik Gangavarapu, Laura D. Hughes, and the Center for Viral Systems Biology (2022), 'Outbreak.info', <https://outbreak.info/>, accessed 19 January.

Kabuni, Michael (2020), 'COVID-19: the situation so far and challenges for PNG', *DEVPOLICYBLOG* (2021; Canberra, Australia: Devlopment Policy Centre, Crawford School of Public Policy, Australian National University).

Katoh, Kazutaka and Standley, Daron M. (2013), 'MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability', *Molecular Biology and Evolution,* 30 (4), 772-80.

Lane, Courtney R., et al. (2021), 'Genomics-informed responses in the elimination of COVID-19 in Victoria, Australia: an observational, genomic epidemiological study', *The Lancet Public Health,* 6 (8), e547-e56.

National Statistical Office of Papua New Guinea 'Population', *Census 2011 Statistics* <https://www.nso.gov.pg/statistics/population/>, accessed 17 January 2021.

O'Toole, Áine, et al. (2021), 'Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool', *Virus Evolution*.

Public Health England (2021), 'SARS-CoV-2 variants of concern and variants under investigation in England. Technical briefing 20', in Public Health England (ed.).
Rambaut, Andrew, et al. (2020), 'A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology', *Nature Microbiology,* 5 (11), 1403-07.
SARS-CoV-2 lineages 'Pangolin lineage webpage', <https://cov-lineages.org/lineages.html. >, accessed 15 August 2021.
Smaghi, B. S., et al. (2021), 'Barriers and enablers experienced by health care workers in swabbing for COVID-19 in Papua New Guinea: A multi-methods cross-sectional study', *Int J Infect Dis*.
The World Bank 'Papua New Guinea COVID-19 Emergency Response Project ', <https://projects.worldbank.org/en/projects-operations/project-detail/P173834>, accessed 21 May 2021.
To, T. H., et al. (2016), 'Fast Dating Using Least-Squares Criteria and Algorithms', *Syst Biol,* 65 (1), 82-97.
World Health Organisation (2021a), 'COVID-19 Weekly Epidemiological Update 13 Oct 2021. ', *Situation Reports* (61 edn.).
World Health Organisation (2021b), 'COVID-19 in Papua New Guinea Situation Report 63', *Emergency Situational Updates*.
Zainulabid, U. A., et al. (2021), 'Near-Complete Genome Sequences of Nine SARS-CoV-2 Strains Harboring the D614G Mutation in Malaysia', *Microbiol Resour Announc,* 10 (31), e0065721.