African-specific prostate cancer molecular taxonomy

Vanessa Hayes (vanessa.hayes@sydney.edu.au)
University of Sydney https://orcid.org/0000-0002-4524-7280

Weerachai Jaratlerdsiri
University of Sydney

Jue Jiang
Garvan Institute of Medical Research https://orcid.org/0000-0003-0920-8310

Tingting Gong
University of Sydney

Sean Patrick
University of Pretoria

Cali Willet
University of Sydney

Tracy Chew
University of Sydney

Ruth Lyons
Garvan Institute of Medical Research

Anne-Maree Haynes
Garvan Institute of Medical Research

Gabriela Pasqualim
Universidade Federal do Rio Grande do Sul

Melanie Louw
National Health Laboratory Services

James Kench
University of Sydney

Raymond Campbell
Kalafong Academic Hospital

Eva Chan
New South Wales Health Pathology https://orcid.org/0000-0002-6104-3763

David Wedge
University of Manchester https://orcid.org/0000-0002-7572-3196

Rosemarie Sadsad
University of Sydney

Ilma Brum

Universidade Federal do Rio Grande do Sul
Shingai Mutambirwa
Sefako Makgatho Health Science University
Phillip Stricker
St Vincnet's Hospital
Riana Bornman
University of Pretoria https://orcid.org/0000-0003-3975-2333
Lisa Horvath
Chris O'Brien Lifehouse

Biological Sciences - Article

Keywords: Prostate cancer, tumour genome profiling, Global Mutational Subtypes

Posted Date: December 1st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1122619/v1

Version of Record: A version of this preprint was published at Nature on August 31st, 2022. See the published version at https://doi.org/10.1038/s41586-022-05154-6.

African-specific prostate cancer molecular taxonomy

Weerachai Jaratlerdsiri[1,2], Jue Jiang[1,2], Tingting Gong[1,2], Sean M. Patrick[3], Cali Willet[4], Tracy Chew[4], Ruth J. Lyons[2], Anne-Maree Haynes[5], Gabriela Pasqualim[6,7], Melanie Louw[8], James G. Kench[9], Raymond Campbell[10], Lisa G. Horvath[5,11], Eva K.F. Chan[2], David C. Wedge[12], Rosemarie Sadsad[4], Ilma Simoni Brum[6], Shingai B.A. Mutambirwa[13], Phillip D. Stricker[5,14], M.S. Riana Bornman[3], Vanessa M. Hayes[1,2,3,15*]

[1]Ancestry and Health Genomics Laboratory, Charles Perkins Centre, School of Medical Sciences, Faculty of Medicine and Health, University of Sydney, Camperdown, NSW, Australia; [2]Human Comparative and Prostate Cancer Genomics Laboratory, Garvan Institute of Medical Research, Darlinghurst, NSW, Australia; [3]School of Health Systems & Public Health, University of Pretoria, South Africa; [4]Sydney Informatics Hub, University of Sydney, Darlington, NSW, Australia; [5]Genomics and Epigenetics Theme, Garvan Institute of Medical Research, Darlinghurst, NSW, Australia; [6]Endocrine and Tumor Molecular Biology Laboratory (LABIMET), Instituto de Ciências Básicas da Saúde, Universidade Federal do Rio Grande do Sul, Brazil; [7]Laboratory of Genetics, Instituto de Ciências Biológicas, Universidade Federal do Rio Grande, Brazil; [8]National Health Laboratory Services, Johannesburg, South Africa; [9]Department of Tissue Pathology and Diagnostic Oncology, Royal Prince Alfred Hospital and Central Clinical School, University of Sydney, Sydney, NSW, Australia; [10]Kalafong Academic Hospital, Pretoria, South Africa; [11]Medical Oncology, Chris O'Brien Lifehouse, Royal Prince Alfred Hospital and Faculty of Medicine and Health, University of Sydney Camperdown, NSW, Australia; [12]Division of Cancer Sciences, University of Manchester, United Kingdom; [13]Department of Urology, Sefako Makgatho Health Science University, Dr George Mukhari Academic Hospital, Medunsa, South Africa; [14]Department of Urology, St. Vincent's Hospital,

Darlinghurst, NSW, Australia; [15]Faculty of Health Sciences, University of Limpopo, Turfloop Campus, South Africa.

*e-mail: vanessa.hayes@sydney.edu.au

Abstract

Prostate cancer is characterised by significant global disparity; mortality rates in Sub-Saharan Africa are double to quadruple those in Eurasia[1]. Hypothesising unknown interplay between genetic and non-genetic factors, tumour genome profiling envisages contributing mutational processes[2,3]. Through whole-genome sequencing of treatment-naïve prostate cancer from 183 ethnically/globally distinct patients (African *versus* European), we generate the largest cancer genomics resource for Sub-Saharan Africa. Identifying ~2 million somatic variants, Africans carried the greatest burden. We describe a new molecular taxonomy using all mutational types and ethno-geographic identifiers, including Asian. Defined as Global Mutational Subtypes (GMS) A–D, although Africans presented within all subtypes, we found GMS-B to be 'African-specific' and GMS-D 'African-predominant', including Admixed and European Africans. Conversely, Europeans from Australia, Africa and Brazil predominated within 'mutationally-quiet' and ethnically/globally 'universal' GMS-A, while European Australians shared a higher mutational burden with Africans in GMS-C. GMS predicts clinical outcomes; reconstructing cancer timelines suggests four evolutionary trajectories with different mutation rates (GMS-A, low 0.968/year *versus* D, highest 1.315/year). Our data suggest both common genetic factors across extant populations and regional environmental factors contributing to carcinogenesis, analogous to gene-environment interaction defined here as a different effect of an environmental surrounding in persons with different ancestries or vice versa. We

anticipate GMS acting as a proxy to intrinsic and extrinsic mutational processes in cancers, promoting global inclusion in landmark studies.

Main

Prostate cancer is a common heterogeneous disease, responsible annually for more than 1,400,000 new diagnoses and 375,000 male-associated deaths worldwide[1]. Characterised by a highly variable natural history and diverse clinical behaviours[4], it is not surprising that genome profiling has revealed extensive intra- and inter-tumour heterogeneity and complexity[5,6]. The identification of oncogenic subtypes[7] and actionable drug targets[8] are moving prostate cancer management a step closer to the promise of precision medicine[7,9-13]. While high-income European ancestral countries are well along the road to incorporating cancer genomics in all aspects of cancer care[14], the rest of the world lags behind, with a notable absence in Sub-Saharan Africa[15]. Prostate cancer is no different, with a single large-scale study out of China[12]; in 2018, we provided the first snapshot for Sub-Saharan Africa, reporting an elevated mutational density in a mere six cases[16]. With mortality rates over double high-income countries and quadrupled for greater Asia, Sub-Saharan Africa prostate cancer is the top-ranked male-associated cancer both by diagnosis and deaths, including southern Africa with age-standardised rates of 65.9 and 22 per 100,000, respectively[1]. Through the Southern African Prostate Cancer Study (SAPCS), we report a 2.1-fold increase in aggressive disease compared to African Americans[17].

Here we describe, to our knowledge, the largest cancer and prostate cancer genomics data for Sub-Saharan Africa, including 123 South African men. Controlling for study artefacts, an additional 60 non-Africans were passed simultaneously through the same high-depth whole-genome sequencing (WGS), mutation-calling and analytical framework. Focusing on treatment-naïve aggressive tumours (mostly Grades 4-5, Extended Data Fig. 1a) and patient-matched blood achieving coverages of

88.69±14.78 and 44.34±8.11, respectively (median±s.d., Supplementary Table 1), we uniformly generated, called and assessed about 2 million somatic variants. We show a greater number of acquired genetic alterations within Africans, while identifying both globally relevant and African-specific genomic subtypes. Through combining our somatic variant dataset with that published for European-ancestral[7,8,18,19] and Chinese[12] prostate cancer genomes, we reveal a novel prostate cancer taxonomy with different clinical outcomes. The inclusion of 2,658 cancer genomes from the ICGC/TCGA Pan-Cancer Analysis of Whole Genomes (PCAWG)[14] led to expanding our global mutational subtyping between cancer types. Using known clock-like mutational processes in each subtype, we infer mutation timing of oncogenic drivers in broad periods of tumour evolution and calculate mutation rates for each subtype that had a distinctive tumour evolution pattern. Combined, these analyses allow us to demonstrate how global inclusion in cancer genomics can unravel unseen heterogeneity in prostate cancer in terms of its genomic and clinical behaviours.

Genetic ancestry

Genetic ancestries were estimated for the 183 patient donors using a joint dataset in a unified analysis aggregated from a collection of geographically matched African (n=64) and European (n=4) deep-coverage reference genomes[20,21]. Ancestries were assigned using 7,472,833 markers as: African (n=113), with greater than 98% contribution; European (n=61), allowing for up to 10% Asian contribution (with a single outlier of 26%); and African-European Admixed (n=9), with as little as 4% African or European contribution (Extended Data Fig. 1b).

Total somatic mutations

In 183 prostate tumours, we identified 1,067,885 single nucleotide variants (SNVs), 11,259 dinucleotides, 307,263 small insertions/deletions (indels <50 bp), 419,920 copy number alterations (CNAs) and 22,919 structural variants (SVs), with each mutational type elevated in African derived tumours (Fig. 1a). A median of 37.54%±5.51 of SNVs were C-to-T mutations, and the transition and transversion ratio was 1.282 cohort-wise. African derived tumours harboured a higher rate of small mutations (SNVs and indels), with a median of 1.197 mutations/Mb (0.031-170.445), compared to those of Europeans (1.061 mutations/Mb, *P*-value = 0.013, two-sample t-test). Percent genome alteration (PGA) was similarly greater in Africans (0.073 *versus* 0.028, *P*-value = 0.021). Correlation tests of ethnicity and total somatic mutations also supported the findings (FDR=0.009 and 0.032 for SNVs and PGA, respectively, Extended Data Fig. 1d). The top six highest estimates of SV breakpoints per sample were observed among African patients (928-2,284 breakpoints). Intrachromosomal SV breakpoints were 52-55% positive for chromothripsis among Africans and Europeans (median, 3 and 2 high-confidence events, respectively). Chromoplexy was more frequent in Europeans than in Africans (38% *versus* 33%, *P*-value=0.536), with the number of interchromosomal chains more likely to be elevated in Africans than Europeans (1-6 *versus* 1-2, *P*-value=0.748). Moreover, the magnitude of all types of mutations was strongly correlated to one another (Fig. 1b). Thus, the more mutations a prostate tumour has of any given type, the more mutations it is likely to have of all types.

Candidate oncogenic drivers

Prostate cancer is known to have a long tail of oncogenic drivers[19] across the spectrum of different mutational types[8] (Extended Data Fig. 2). Protein-coding

mutations, including probably and possibly damaging, were significantly greater in Africans (PolyPhen-2, 14 *versus* 11 mutations in Europeans, *P*-value=0.022, two-sample t-test). We identified 482 coding and 167 noncoding drivers defined by the PCAWG Consortium[22] (Extended Data Fig. 3a). A median of 2±22.5 coding drivers was observed in this study (Supplementary Table 2), with 1±5.4 appearing to be prostate cancer-specific[7,8,18,19]. The coding driver genes significantly mutated among 183 patients were *FOXA1*, *PTEN*, *SPOP* and *TP53* (10-25 patients, FDR=1.34e-21–9.44e-05), while noncoding driver elements were the *FOXA1* 3′-UTR, *SNORD3B-2* small RNA and a regulatory miRNA promoter at chromosome 22:38,381,983 (FDR=9.12e-13, 6.16e-09 and 0.070, respectively). Recurrent CNAs of all the patients included 137 gains and 129 losses (GISTIC2, FDR <0.10, Supplementary Table 3) with some spanning driver genes (Extended Data Fig. 3b), such as *DNAH2* (FDR=2.18e-07), *FAM66C* (1.30e-09), *FOXP1* (0.005), *FXR2* (2.18e-07), *PTEN* (9.61e-13), *SHBG* (2.18e-07), and *TP53* (2.18e-07).

In addition, a fraction of somatic SVs (2 breakpoints each; 1,328 breakpoints in total) overlapped with 156 driver genes reported as altered by significantly recurrent breakpoints in the PCAWG study[22], while using a generalised linear model with adjustable background covariates we identified an additional 100 genes to be significantly impacted by SV breakpoints (FDR=1.3e-43–0.097, Extended Data Fig. 3c, Supplementary Table 4). For over 20% of tumours, SV breakpoints coexisted with other mutational types within *DNAH2*, *ERG*, *FAM66C*, *FXR2*, *PTEN*, *SHBG*, and *TP53*. Using optical genome mapping (OGM), an alternative non-sequencing method to interrogate for chromosomal abnormalities[23], we validated recurrent breakpoints in novel HLA regions (*DQA1* and *DQB1* genes), identifying translocations between the

3-Mb HLA complex at chromosome 6 and its corresponding HLA alternate contigs (Extended Data Fig. 3d).

Integrative clustering analysis of molecular subtypes

Molecular subtyping of tumours is a standard approach in cancer genomics to stratify patients into different degrees of somatic alterations in a homogeneous population, with an implication for clinical use[9-12]. Identifying five of the seven TCGA oncogenic driver-defined subtypes in our study[7], European patients were 25% more likely than African patients to be classified (Supplementary Table 5, Extended Data Fig. 4a-d). While *TMPRSS2-ERG* fusions (predominantly 3-Mb deletions) and *FOXA1* coding mutations (forkhead domain) occurred at higher frequencies in our European over African patients, 37.7% and 8.2% *versus* 13.3% and 5.3%, respectively (OR=0.255, *P*-value=0.0004 and OR=0.854, *P*-value=0.771), *SPOP* coding mutations (MATH and BTB domains) were more common in the African (8.8%) *versus* European patients (6.6%, OR=1.688, *P*-value=0.426).

For further molecular classification, we performed iCluster analysis on all mutational types (small mutations, copy number and SVs) identifying four subtypes, A to D (Supplementary Table 6, Fig. 2a, b). We found Subtype A to be mutationally quiet (1.01 mutations/Mb, 0.50 breakpoints/10Mb, 2% PGA); conversely Subtype D showed the greatest mutational density (1.91 mutations/Mb, 1.08 breakpoints/10Mb, 31% PGA) with a mixture of copy number (CN) gains and losses, while Subtypes B and C were marked by substantial CN gains or losses, respectively (Fig. 3b). The quiet subtype seems to be common in prostate cancer studies[7,9,24], while the number of pan-cancer consensus drivers[22] increased from Subtype A (median, 2 drivers) to B (3), C (3) and D (4).

Using all mutational types in the analysis, 124 genes were significantly mutated across the four subtypes (FDR=3.742e-13–0.067; Fig. 3a), occurring in 31 to 183 patients (frequency, 0.17-1). Among them, 100 genes were reported as oncogenic drivers in the PCAWG[22] and *FOXA1* and *SPOP* genes acting as the TCGA subtypes were also replicated in this analysis, while the 24 novel recurrently mutated genes were predominantly impacted by SV breakpoints and CNAs. The median number of mutated genes ranged from 28 (range 3-105) for Subtype A to 82, 98 and 93 for Subtypes B, C and D, respectively (42-109, 72-112, 49-107). While different mutational types tended to co-occur within genes and/or patients (Supplementary Table 7), small mutations (coding and noncoding) were noticeably observed in the quiet subtype, supporting acquisition early in tumourigenesis[25]. Our preferentially mutated genes within tumour subtypes resemble the long tail of prostate cancer drivers[19], with some highly impacting many tumours, but most only impacting a few tumours.

The 124 preferentially mutated genes within our tumour subtypes corresponded to eight TCGA/ICGC cancer pathways (see Supplementary Methods, Extended Data Fig. 5). While six showed slightly elevated mutational frequencies in African derived tumours, genes impacting epigenetic mechanisms were significantly biased towards Europeans (OR=0.179, *P*-value=2.9e-07, Extended Data Fig. 6b). Pathway enrichment analysis supported five functional networks of the cancer pathways, with two of them involved in signal transduction and DNA checkpoint processes to which five of the eight pathways were interacted (Extended Data Fig. 6a; Supplementary Table 8).

Global molecular subtypes

196 Through combining molecular profiling and patient demographics, ethnicity and
197 geography, we identify a new prostate cancer taxonomy we define as 'Global
198 Mutational Subtypes (GMS)' (Fig. 2b). While all European patients from Australia
199 (n=53) and Brazil (n=3) were limited to GMS-A and C, African derived tumours were
200 dispersed across all four subtypes. We found GMS-B and D to predominate in
201 Africans, with GMS-B including a single patient of admixed ancestry (92% African)
202 and GMS-D including a single admixed (63% African) and a single European
203 ancestral patient. The latter was one of only five Europeans in our study born and
204 raised in Africa. Compared to other patients of European ancestry, this patient showed
205 the highest mutational density across all types. Alternative consensus clustering of
206 individual mutational types mostly recapitulated the subtypes by integrative analysis
207 (Supplementary Table 6). Through further inclusion of Chinese Asian high-risk
208 prostate cancer data[12] (n=93, Extended Data Fig. 7a), we found GMS-A to be
209 ethnically and geographically 'universal', while GMS-D remained 'African-specific'
210 with a new 'African-Asian' GMS-E emerging. GMS-B remained 'African-specific'
211 and GMS-C 'European-African'. While all patients were treatment naïve at the time
212 of sampling, our European cohort was recruited with extensive follow-up data
213 (median±s.d., 122.5±44.4 months). Interestingly, biochemical relapse (Fig. 3c) and
214 death-free survival probability (Fig. 3d) explains better clinical outcomes for patients
215 presenting with the 'universal' over the 'European-African' GMS (A *versus* C, log-
216 rank *P*-value=0.008 and 0.041, respectively).

217 Our novel GMS taxonomy could leverage pan-cancer studies in the following ways.
218 First, a sampling strategy of patients from the PCAWG project was rather
219 homogeneous in each cancer, therefore inhibiting the discovery of globally restricted
220 subtypes[3,14] (Extended Data Fig. 7b). Second, ancestral[26] and geographic data of

patients should be included in molecular profiling of cancers. Lastly, the inclusion of ethnic disparity in cancer studies would need to properly address admixture in a sampling cohort, with too low ancestral cut-off appearing to create highly admixed, but similar ancestry among individuals, therefore discouraging ethnically diverse samples.

Novel and known mutational signatures

Approximating the contribution of mutational signatures to individual cancer genomes facilitates an association of the signatures to exogenous or endogenous mutagen exposures that contribute to the development of human cancer[3]. Here, we generated a novel list of copy number (CN) and SV signatures and their contributions to prostate cancer using nonnegative matrix factorisation[27] (Extended Data Fig. 8a, b). Combined with a known catalogue of small mutational signatures, including single base substitutions (SBS), doublet base substitutions (DBS) and small insertions and deletions (ID), we observed not only a substantial variation in the number of mutational features, but also over-representation in African derived tumours (Extended Data Fig. 8c). Overall, 96 SBS, 78 DBS and 83 ID features examined had significantly higher totals in Africans (SBS, 3,399 *versus* 2,840 in Europeans, *P*-value=0.014; DBS, 42 *versus* 32, *P*-value=0.006; ID, 374 versus 360, *P*-value=0.016, two-sample t-test). We generated six *de novo* signatures for each small signature type (median cosine similarity 0.986, 0.856, and 0.976, respectively), corresponding to 12, seven and eight global signatures, respectively (0.966, 0.850, and 0.946, respectively; Extended Data Fig. 9), with 26 likely to be of biological origin (SBS47, possible sequencing artefacts). DBS substitutions accounted for about 1% of the prevalence of SBS. The CN features were also greater in Africans (CN, 3,971 *versus* 2,721, *P*-

value=1.92e-08; SV, 94 *versus* 88, *P*-value=0.100). The SV features defined in a recent pan-cancer study[27] were each mutually exclusive and included simple SVs (split according to size, replication timing and occurrence at fragile sites), templated insertions (split by size), local n-jumps and local–distant clusters. The factorisation of a sample-by-mutation spectrum matrix identified six CN signatures (CN1-6) and eight SV signatures (SV1-8), as well as their contributions to each tumour.

We found the full spectrum of mutational signatures (SBS, DBS, ID, CN and SV) to support our newly described GMS. Enrichment records of the top signatures in each tumour were significantly associated type by type with the taxonomic subtypes, except for DBS (*P*-values=5.1e-07–0.017, one-way ANOVA or Fisher's exact test, Extended Data Fig. 8d). Regardless of signature type, 13/40 mutational signatures showed either inverse or proportionate correlations with our GMS (FDR=4.97e-13–0.095, Spearman's correlation, Fig. 4). Duplication signatures, including CN1 (tandem duplication), CN4 (whole genome duplication), SV2 (insertion) and SV5 (large duplication), were biased to the most mutationally noisy subtype (Extended Data Fig. 8a, b), with CN4 and SV5 frequent in Africans (rho=-0.24, FDR=0.005-0.006). The mutational density of 30 out of 32 genes highly mutated in our GMS and reported in prostate cancer was also significantly correlated with different somatic signatures, with most observed in CN2, CN6 and SV6 signatures that were mainly caused by deleted genomes. Small-size signatures were inversely significant among 20 mutated genes, indicating a higher number of mutations towards lesser mutated tumours (FDR=1.05e-08–0.099).

Life history of globally mutated subtypes

Timeline estimates of individual somatic events reflect evolutionary periods that differ from one patient to another; for example, a cluster of identical alterations derived from clones in one patient presented as subclonal events in another patient (Extended Data Fig. 10a, b). However, they provide in part the order of driver mutations and CNAs present in each sample[25]. The reconstruction of aggregating single-sample ordering of all drivers and CNAs reveals different evolutionary patterns unique to each GMS (Extended Data Fig. 10c, Fig. 5a, b). We draw approximate cancer timelines for each GMS portraying the ordering of driver genes, recurrent CNAs and signature activities chronologically interleaved with whole-genome duplication (WGD) and the emergence of the most recent common ancestor (MRCA) leading up to diagnosis. Basically, significantly co-occurring interactions of the drivers and CNAs are shown (OR=2.6–97.8, *P*-values = 2.04e-30–0.01), supporting their clonal and subclonal ordering states within the reconstructed timelines. SBS and ID signatures that are abundant in each GMS display changes of their mutational spectrum between the clonal and subclonal state, suggesting a difference in mutation rates. The plot of clock-like CpG-to-TpG mutations and patient age adjustment shows the median mutation rate as little as 0.968 per year for the 'universal' GMS, but the highest rate at 1.315 per year observed in the 'African-specific' GMS-D. GMS-B and C have rates of 1.144 and 1.092 per year, respectively. Assessing the relative timing of somatic driver events, *TP53* mutations and accompanying 17p loss are of particular interest, occurring early in GMS-C progression and at a later stage in GMS-A. League model relative timing of driver events (see Supplementary Methods) is in line with a fraction of probability distribution of the *TP53* alterations at the early stage, but most are at an intermediate state of evolution (Extended Data Fig. 10d). This basic knowledge of *in vivo* tumour development suggests that some tumours could have a

shorter latency period before reaching their malignant potential, so known genomic heterogeneity of their primary clones is paramount to pave a way for early detection.

Discussion

To our knowledge, our study represents the first, if not, the largest whole-genome prostate cancer, and likely any cancer, genome resource for Sub-Saharan Africa. Here we describe a novel prostate cancer molecular taxonomy, identifying ethnically and geographically distinctive Global Mutational Subtypes (GMS). Compared to previous taxonomy using significantly mutated genes in prostate cancer[7,19], we found GMS to compliment known subtypes such as *SPOP* and *FOXA1* mutations, in contrast to underrepresented subtypes in this study, including gene fusions (Extended Data Fig. 4a). We also found GMS to correlate with mutational signatures reported in the known catalogue of somatic mutations in cancer, where each tumour is represented by different degrees of exogenous and endogenous mutagen exposures[3]. Our study has leveraged the analysis of evolution across 38 cancer types by the PCAWG Consortium[25], recognising that each GMS represents a unique evolutionary history with drivers and mutational signatures varied between cancer stages and linking somatic evolution to a patient's demographics. Therefore, some represent 'rare or geographically restricted signatures' that are still a myth in pan-cancer studies[3,14].

We consider two extreme cases, 'universal' GMS-A *versus* 'African-specific' GMS-B and D, that would have been influenced by two different mutational processes for conceptual simplicity (Fig. 5c). One is predisposing genetic factors that are known for prostate tumourigenesis across ethnolinguistic groups[28-30]. This factor contributes to endogenous mutational processes, especially those with significant germline-somatic interactions, such as the *TMPRSS2-ERG* fusion less frequently observed in men of

317 African and Asian ancestry[12,31], germline *BRCA2* mutations and the somatic *SPOP*
318 driver co-occurred with their respective counterparts[32,33]. Another factor is modifiable
319 environmental attributes specific to certain circumstances or geographic regions that,
320 until now, have been elusive to prostate cancer. They act as mutagenic forces leading
321 to the positive selection of point mutations throughout life in healthy tissues[34,35] and
322 cancers[36], forming fluid boundaries between normal ageing and cancer tissues.
323 According to Ottman[37], the above-mentioned model of gene-environment interaction
324 is observed when there is a different effect of a genotype on disease in individuals
325 with different environmental exposures or, alternatively, a different effect of an
326 environmental exposure on disease in individuals with different genotypes. Other
327 GMS subtypes would be a combination of the two processes, warranting a need for
328 larger populations of different ethnicities from different geographical localities to be
329 studied for a breakthrough in nature *versus* nurture. As such, the study directly
330 accounts for the large spatio-genomic heterogeneity of prostate cancer and its
331 associated evolutionary history in understanding the disease aetiology.

332 Our study suggests that larger genomic datasets of ethnically and geographically
333 diverse populations in a unified analysis will continue to identify rare and
334 geographically restricted subtypes in prostate cancer and potentially other cancers.
335 We are the first to demonstrate that ancestral and geographic attributes of patients
336 could facilitate those studies on cancer population genomics, an alternative to cancer
337 personalised genomics, for a better scientific understanding of nature *versus* nurture.

Figure legends

Fig. 1 | Mutational density in prostate tumours of different ancestries. a, Distribution of somatic aberration (event number or number of base pairs) for seven mutational types across 183 tumour-blood WGS pairs. **b,** Different types of mutational burden observed in this cohort. Samples are percentile ranked and then ordered based on the sum of percentiles across the mutational types observed in each ethnic group (left panel). Spearman's correlation is shown between mutation types, with dot size representing the magnitude of correlation and background colour giving statistical significance of FDR values (right panel).

Fig. 2 | Prostate cancer taxonomy of ethnically diverse populations. a, Integrative clustering analysis reveals four distinct molecular subtypes of prostate cancer. The molecular subtypes are illustrated by small somatic mutations (coding regions and noncoding elements), somatic copy number alterations and somatic SVs. The percentage and association between the iCluster membership and patient ancestry are illustrated in square brackets. A, African ancestry; Ad, Admixed; and E, European ancestry. **b,** Total somatic mutations across four molecular subtypes in this study. Dashed lines indicate the median values of mutational densities across the four subtypes. For each subtype, patients are ordered based on their ethnicity.

Fig. 3 | Aberration of driver genes in four diverse subtypes. a, Analysis of the long tail of driver genes using different mutation data combined. A total of 124 genes are associated with four prostate cancer subtypes, and all have previously been reported as significantly recurrent mutations/SV breakpoints in the PCAWG Consortium[22], except for ones marked by asterisks, where they are assigned to be significantly mutated using whole-genome data in this study. The Y-axis shows corrected *P*-values in $-\log_{10} P$. CDS, coding driver data; NC, noncoding driver data; SV, significantly recurrent breakpoint data; and CN, gene-level copy number data. **b,** Unsupervised hierarchical clustering of known and putative driver genes identified within four prostate cancer subtypes (A-D, a bottom-up direction). Rows are patients, and columns represent 124 driver genes (alphabetical order) identified using different mutational types. **c,** Kaplan-Meier plot of biochemical relapse (BCR)-free survival proportion of European patients in subtype A (n=161) *versus* C (n=19). **d,** Kaplan-Meier plot of cancer survival probability of European patients in subtype A (n=82) *versus* C (n=17).

Fig. 4 | Estimates of genomic aberrations contributed by each mutational signature. The size of each dot represents FDR values of Spearman correlation *P*-values using BH correction. The colours of each dot represent correlation coefficient (rho). GMS is assigned as 1-4 for Subtypes A-D, respectively; African, Admixed and European are recorded as 1-3, respectively. The correlation of 32 significantly mutated genes in prostate cancer is shown in the X-axis.

Fig. 5 | Evolutionary history of globally mutated subtypes. a, The cancer timeline of the universal subtype begins from the fertilised egg to the age of the patients at a cohort. **b,** that of GMS-C. Estimates for major events, such as WGD (whole-genome duplication) and the emergence of the MRCA (the most recent common ancestor), are used to define early, variable, late and subclonal stages of tumour evolution approximately in chronological time. When early and late clonal stages are uncertain, the variable stage is assigned. The variable/constant time period includes events that are ranked before the WGD event and also begins shortly after another break in the timeline. The late period does have a definite start, as this includes events that are ranked after WGD, when it occurs. Driver genes and CNAs are shown in each stage if present in previous studies[8,22] and defined by MutationTime.R program. Mutational signatures (Sigs) that, on average, change over the course of tumour evolution, or are substantially active but not changing, are shown in the epoch in which their activity is rather greatest. Dagger symbols denote alterations that are found to have different timing. Significant pairwise interaction events between the mutations and copy number alterations were computed using Odds Ratio (OR). Either co-occurrence or mutually exclusive event is considered if OR >2 or <0.5, respectively. Median mutation rates of CpG-to-TpG burden per Gb are calculated using age-adjusted branch length of cancer clones and maximally branching subclones. **c,** Schematic representation of a world map with the distribution of GMS (A–D) among ethnically/globally diverse populations. The gene-environment interaction model of globally mutated subtypes is shown in the right panel. The contingency table of number of patients with different ancestries (germline variants) stratified by subtypes and associated with certain geography or environmental exposure (two-sided *P*-value= 0.0005, Fisher's exact test with 2,000 bootstraps).

Methods

Patient cohorts and whole-genome sequencing

Our study included ~180 treatment naïve prostate cancer patients recruited under informed consent and appropriate ethics approval (Supplementary Methods, Section 2) from Australia (n=53), Brazil (n=7) and South Africa (n=123). DNA extracted from fresh tissue and matched blood underwent 2x150 bp sequencing on the Illumina NovaSeq instrument (Kinghorn Centre for Clinical Genomics, Garvan Institute of Medical Research).

WGS processing and variant calling

Each lane of raw sequencing reads was aligned against human reference hg38 + alternate contigs using bwa v0.7.15[38]. Lane-level BAMs from the same library were merged, and duplicate reads were marked. The Genome Analysis Toolkit (GATK v4.1.2.0) was used for base quality recalibration[39]. Contaminated and duplicate samples (n=8) were removed. We implemented three main pipelines for the discovery of germline and somatic variants, with the latter including small (SNV and indel) to large genomic variation (CN and SV). Complete pipelines and tools used are available from the Sydney Informatics Hub (SIH), Core Research Facilities, University of Sydney (see Code availability). Scalable bioinformatic workflows are described in Supplementary Methods, Section 4.

Genetic ancestry was estimated using fastSTRUCTURE v1.0[40], Bayesian inference for the best approximation of marginal likelihood of a very large variant dataset. Reference panels for African and European ancestry compared in this study were retrieved from previous whole-genome databases[20,21].

Analysis of chromothripsis and chromoplexy

Clustered genomic rearrangements of prostate tumours were identified using ShatterSeek v0.4[41] and ChainFinder v1.0.1[42]. Our somatic SV and somatic CNA callsets were prepared and co-analysed using custom scripts (see Code availability, Supplementary Methods, Section 6).

Analysis of mutational recurrence

We used three approaches to detect recurrently mutated genes or regions based on three mutational types, including small mutations, SVs and CNAs (see Supplementary Methods, Section 7). In brief, small mutations were tested within a given genomic element as being significantly more mutated than adjacent background sequences. The genomic elements retrieved from syn5259886, the PCAWG Consortium[22] were a group of coding sequences and 10 groups of noncoding regions. SV breakpoints were tested in a given gene for their statistical enrichment using Gamma-Poisson regression and corrected by genomic covariates[13]. Focal and arm-level recurrent CNAs were examined using GISTIC v2.0.23[43]. Known driver mutations in coding and noncoding regions published in PCAWG[22,44,45] were additionally recorded in our 183 tumours, and those specific to prostate cancer genes were also included[7,8,13,18,19].

Integrative analysis of prostate cancer subtypes

Integrative clustering of three genomic data types for 183 patients was performed using iClusterplus[12,46] in R, with the following inputs: *i*) driver genes and elements; *ii*) somatic CN segments; and *iii*) significantly recurrent SV breakpoints. We ran iClusterPlus.tune with clusters ranging from 1-9. We also performed unsupervised consensus clustering on each of the three data types individually. Association analysis

of genomic alteration with different iCluster subtypes was performed in detail in Supplementary Methods, Section 8. Differences in drivers, recurrent breakpoints and somatic CNAs across different iCluster subtypes were reported.

Comparison of iCluster with Asian and pan-cancer data

To compare molecular subtypes between extant human populations, the Chinese Prostate Cancer Genome and Epigenome Atlas (CPGEA, PRJCA001124)[12] was merged and processed with our integrative clustering analysis across the three data types described above, with some modifications. Moreover, we leveraged the PCAWG Consortium[14] to define molecular subtypes across different ethnic groups in other cancer types using published data of somatic mutations, SV and GISTIC results by gene. Four cancer types that consisted of breast, liver, ovarian, and pancreatic cancers were considered due to existing primary ancestries of African, Asian and European with at least 70% contribution. Full details are given in Supplementary Methods, section 8.4.

Prostate cancer subjects of PCAWG[14] were retrieved to compare with Australian data with clinical follow-up. Only those of European ancestry greater than 90% (n=139) were analysed for the three genomic data types of iCluster subtyping, as well as individual consensus clustering. Clustering results identical to the larger cohort size mentioned above were chosen for association analyses. Differences in the biochemical relapse and lethal prostate cancer of the subjects across the subtypes were assessed using the Kaplan–Meier plot followed by a log-rank test for significance.

Analysis of mutational signatures

Mutational signatures (SBS, DBS and ID), as defined by the PCAWG Mutational Signatures Working Group[3], were fit to individual tumours with observed signature activity using SigProfiler[47]. Nonnegative matrix factorisation (NMF) was implemented to detect *de novo* and global signature profiles among 183 patients and their contributions. Novel mutational genome rearrangement signatures (CN and SV) were also performed using the NMF, with 45 CN and 44 SV features examined across 183 tumours. We followed the PCAWG working classification and annotation scheme for genomic rearrangement[27]. Two SV callers were used to obtain exact breakpoint coordinates. Replication timing scores influencing on SV detection were set at >75, 20-75, and <20 for early, mid, and late timing, respectively[48]. Full details of analysis steps, parameters and relevant statistical tests are given in Supplementary Methods, Section 9.

Reconstruction of cancer timelines

Timing of copy number gains and driver mutations (SNVs and indels) into four epochs of cancer evolution (early clonal, unspecified clonal, late clonal, and subclonal) was conducted using MutationTimeR[25]. CN gains including 2+0, 2+1, and 2+2 (1+1 for a diploid genome) were considered for a clearer boundary between epochs instead of solely information of variant allele frequency. Confidence intervals ($t_{lo} - t_{up}$) for timing estimates were calculated with 200 bootstraps. Mutation rates for each subtype were calculated following Gerstung, et al[25] that CpG-to-TpG mutations were counted for the analysis because they were attributed to spontaneous deamination of 5-methyl-cytosine to thymine at CpG dinucleotides, therefore acting as a molecular clock.

League model relative ordering was performed to aggregate across all study samples to calculate the overall ranking of driver mutations and recurrent CNAs. The information for the ranking was derived from the timing of each driver mutation and that of clonal and subclonal CN segments, as described above. Full description is provided in Supplementary Methods, Section 10.

Data availability

Alignments, somatic and germline variant calls, annotations and derived datasets are available for general research use for browsing and download through the European Genome-Phenome Archive (accession number EGA0000000000). Other supporting data are available upon request from the corresponding author.

Code availability

The core computational pipelines used in this study for read alignment, quality control and variant calling are available to the public at https://github.com/Sydney-Informatics-Hub/Bioinformatics. Analysis code for chromothripsis and chromoplexy is available through another GitHub page, https://github.com/tgong1/Code_HRPCa.

Acknowledgements

The work presented was supported by the National Health and Medical Research Council (NHMRC) of Australia through a Project Grant (APP1165762, V.M.H.), NHMRC Ideas Grant (APP2001098, V.M.H. and M.S.R.B.), University of Sydney Bridging Grant (G199756, V.M.H.), and partly through the U.S. Department of Defense (DoD) Prostate Cancer Research Program (PCRP) Idea Development Award (PC200390, including W.J., S.M.P., D.C.W., S.M., M.S.R.B. and V.M.H.). The

authors acknowledge the use of the National Computational Infrastructure (NCI) which is supported by the Australian Government, and accessed through the National Computational Merit Allocation Scheme (V.M.H., E.K.F.C and W.J.), the Intersect Computational Merit Allocation Scheme (V.M.H.), Intersect Australia Limited, and the Sydney Informatics Hub, Core Research Facility, while we acknowledge the Garvan Institute of Medical Research's Kinghorn Centre for Clinical Genomics (KCCG) core facility for data generation. Recruitment, sampling and processing for the Southern African Prostate Cancer Study (SAPCS), as required for the purpose of this study, was supported by the Cancer Association of South Africa (CANSA, M.S.R.B. and V.M.H.). V.M.H. was supported by Petre Foundation via the University of Sydney Foundation, A-M.H. and W.J. by a Cancer Institute of New South Wales (CINSW) Program Grant (TPG172146 to L.G.H., J.G.K., P.D.S. and V.M.H.), with additional support to W.J. provided by the Prostate Cancer Research Alliance Australian Government and Movember Foundation Collaboration PRECEPT (Prostate cancer prognosis and treatment study, led by A/Prof. N. Corcoran, University of Melbourne, Australia). T.G. is now located at the Human Phenome Institute, Fudan University, Shanghai, China and E.K.F.C. at NSW Health Pathology, Sydney, Australia. We are forever grateful to the patients and their families who have contributed to this study; without their contribution, this research would not be possible. We acknowledge the contributions of the many clinical staff across the SAPCS (South Africa), the St Vincent's Hospital Sydney (Australia) and 6LEndocrine and Tumor Molecular Biology Laboratory (Brazil), who over many years have recruited patients and provided samples to these critical bioresources, with special recognition of Professor Philip Venter (retired), Dr's Richard L. Monare

(retired) and Dr Smit van Zyl, previously from the University of Limpopo, South Africa, for their critical contributions as inaugural members of the SAPCS.

Authors' contributions

V.M.H. designed the experiments and supervised the project; W.J. led the bioinformatic and statistical analyses, while both W.J. and V.M.H. performed data interpretation. S.M.P., R.J.L., A-M.H., and D.G.P. prepared the samples and managed phenotypic data. M.L. and J.G.K. performed pathological grading, while R.C., L.G.H., I.S.B., S.B.A.M., P.D.S. and M.S.R.B. managed patient recruitments and consents, as well as clinical interpretation. V.M.H., S.B.A.M. and M.S.R.B. codirect the Southern African Prostate Cancer Study (SAPCS). W.J., J.J., T.G., C.W., T.C. and R.S. developed the pipelines and performed the efficient and scalable high-performance computational variant calling, with critical advice provided by E.K.F.C and V.M.H. W.J., J.J. and T.G. performed complex variant annotation, while R.J.L. generated the optical genome mapping (OGM) data. W.J. performed mutational signature and tumour evolution analysis, with critical advice provided by D.C.W. W.J. and V.M.H. wrote the manuscript. W.J. generated the figures, while all authors contributed to the final editing and approval.

Competing interest declaration

The authors declare no competing interests.

Supplementary Tables

Supplementary Table 1 | Clinical cohort characteristics and sequencing quality

Supplementary Table 2 | Driver information by patient

Supplementary Table 3 | GISTIC2 results of all genomic lesions under 99% confidence level

Supplementary Table 4 | List of significantly recurrent SV breakpoints at FDR lower than 0.10

Supplementary Table 5 | TCGA prostate cancer taxonomy identified in this study

Patient by driver mutation and patient by driver structural variation summary matrices are provided.

Supplementary Table 6 | Integrative iCluster analysis of 183 prostate tumours

Supplementary Table 7 | List of 124 preferentially mutated genes within four tumour subtypes

Supplementary Table 8 | Pathway enrichment analysis of 124 preferentially mutated genes

Supplementary Table 9 | Total mutational signature profiles across 183 tumours

The table shows data matrices of SBS feature by patient, DBS feature by patient, ID feature by patient, CN feature by patient, and SV feature by patient.

Supplementary Table 10 | Cross-individual contamination level

Supplementary Table 11 | Cancer evolution analysis of prostate cancer

Clonal architecture by PhyloWGS and timing of gains and drivers by MutationTimeR is provided per tumour

Extended data legends

Extended Data Fig. 1 | Clinical cohorts and statistical metrics. a, Clinical and pathological patient characterisation. **b,** STRUCTURE analysis of bi-allelic germline variants with the logistic prior model. Model components used to explain structure in the plot are K=5. All spectrum of African contributions are summed and assigned as African ancestry. **c,** Saturation curve for all driver types across 183 patients. Recurrent copy number gains and losses were measured using GISTIC v2 (Supplementary Methods). CDS, coding sequence; SV, structural variation. **d,** Spearman's correlation between different variables measured in this cohort. Dot sizes represent the magnitude of correlation, with significant *P*-values <0.01.

Extended Data Fig. 2 | Somatic driver mutations in 183 prostate cancer patients. The covariates on the left show mutational types and statistical significance (FDR) from ActiveDriverWGS and GISTIC2. **a,** The top 300 driver genes in PCAWG discovered in primary prostate tumours among 183 specimens. The top barplot shows the distribution of the number of prostate cancer drivers and/or that of PCAWG. The heatmap shows drivers found in this study (rows) for each patient (columns). Heatmaps are coloured by mutational type. Bottom covariates show the clinical features of patients. The percentage of transition/transversion mutations across 183 patients shows 1,364,210 small somatic mutations across chromosomes 1-Y. **b,** The bottom heatmap shows the top 75 of previously reported coding driver genes in prostate cancer observed in this study[7,8,18,19]. The right barplot shows the number of patients for each driver.

Extended Data Fig. 3 | Discovery of prostate cancer drivers. a, The number and types of PCAWG driver genes and elements studied in our cohort. **b,** Recurrent copy number alterations among 183 prostate tumours identified with a 99% confidence level using GISTIC v2 (Supplementary Methods). The figure shows GISTIC peaks of significant regions of recurrent amplification (red) or deletion (blue) supported by FDR <0.01. **c,** Genome-wide scan for significantly recurrent breakpoints in our study. The quantile-quantile plot shows *P*-values for mutational densities across 183 prostate cancer patients. Generalised linear modelling (GLM) of somatic mutation densities along the genome with significant background mutational processes adjusted in the model is also shown. **d,** Bionano Genomics optical genome mapping at the HLA complex. Examples of HLA translocations from a

European patient (ID 12543) and an African patient (ID UP2360) studied in this cohort are characterised by pairs of optical maps, each carrying a fusion junction with flanking fragments aligning to one side of the two reference breakpoints. Using the recurrent HLA breakpoints identified in this study, the genome map of the African specimen is found to have a low-end fusion function matched with chromosome 6 through a manual inspection of unfiltered consensus maps using Bionano Access v15.2. Note that the HLA alternate contig fused in the European tumour is different from one suggested by short-read sequencing (chr6_GL000252v2_alt). The reference genome map is an *in silico* digest of the human reference hg38 with the DLE-1 enzyme. Genome map sizes are indicated on the horizontal axis in megabase (Mb) units. Matching fluorescent labels between sample and reference genome map are connected by gray lines.

Extended Data Fig. 4 | TCGA molecular taxonomy. a, Seven important oncogenic drivers identified by TCGA within our African and European patients. **b,** Coding mutations observed within *SPOP* and *FOXA1* genes. Rarely, a mutation at the BTB domain of *SPOP* gene is shown (R221C in an African patient, KAL0072). FH, forkhead. **c,** *ETV1* fusions within positive patients caused by copy number (CN) losses and/or structural variants (DEL, deletion; ICX, interchromosomal translocation; and INV, unbalanced or balanced inversion). CN changes in chromosome 7 show the *ETV1* loss with $\log_2$ CN ratio less than -0.2. **d,** *ERG* fusions caused by CN losses and/or structural variants.

Extended Data Fig. 5 | Prostate cancer genes and pathways. The search is carried out using the TCGA and ICGC cancer databases. The top affected genes for each pathway are present with lollipop plots to show their hotspots of simple coding mutations if they existed.

Extended Data Fig. 6 | Major biological pathways and networks of prostate cancer. a, Networks of functional interactions between driver genes are shown for each cancer pathway. Nodes represent Gene Ontology biological processes and Reactome pathways and edges show functional interactions. **b,** Pathway alteration frequencies between African and European. A sample was considered altered in a given pathway if at least a single gene in the pathway had a genomic alteration. *P*-values indicate the level of significance (two-sided Fisher's exact test).

Extended Data Fig. 7 | Molecular subtypes in prostate cancer and pan-cancers. a, Unsupervised hierarchical clustering of primary prostate tumours across three major ethnic groups was performed using total somatic mutations present within WGS normalised data. Admixed individuals were also tested in prostate cancer subtypes to which they belonged. **b,** Molecular subtyping of total somatic mutations within pan-cancer studies, namely pancreatic, ovarian, breast and liver cancers. Raw data of small somatic mutations, structural variants and copy number alterations acquired per cancer were retrieved from the PCAWG[14]. For each subtype, patients are ordered based on their ethnicity. Ethnic groups are assigned using a cut-off of ancestral contribution greater than 70%; otherwise, considered as Admixed.

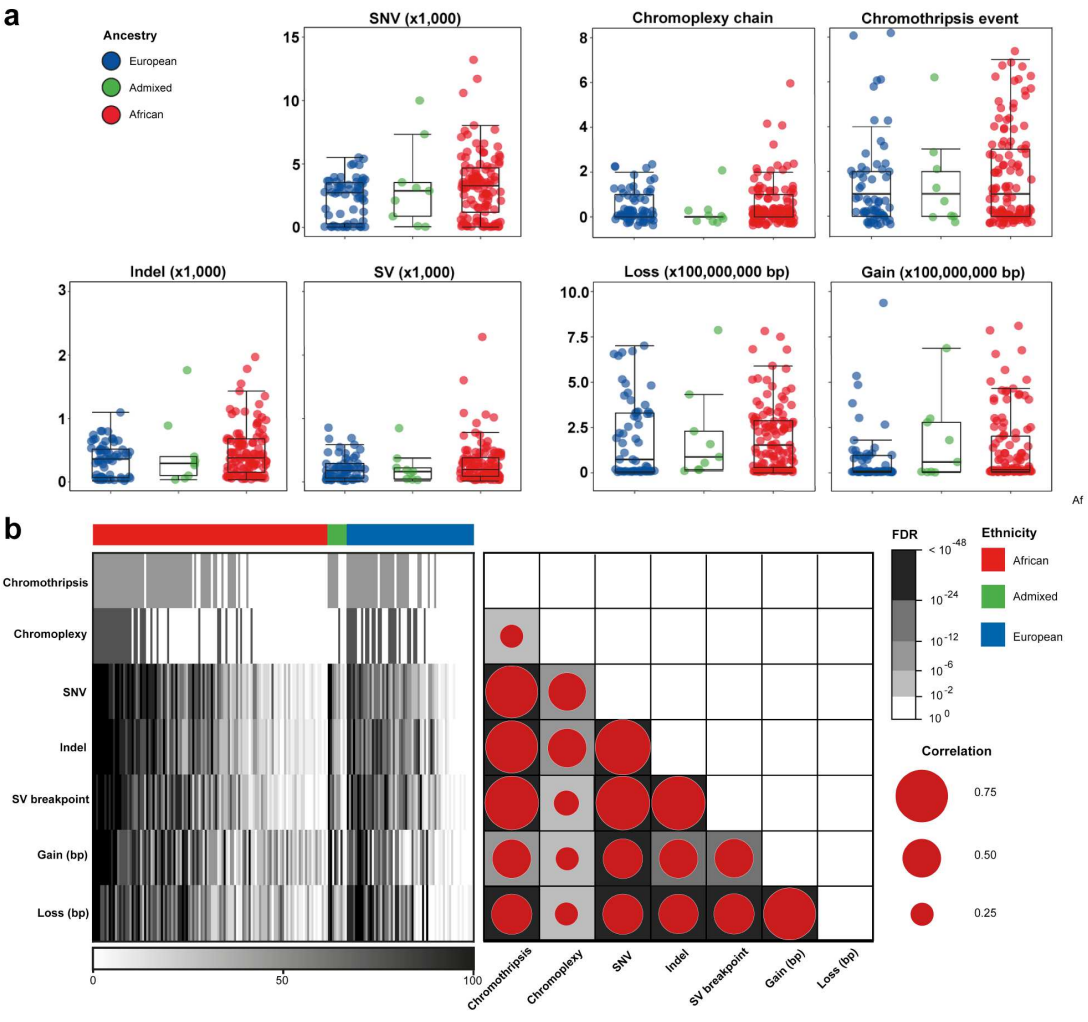
Extended Data Fig. 8 | Known and novel mutational signatures in prostate cancer. a, Copy number signatures in prostate cancer across 45 CN features ranked by mutational processes observed. The six most distinctive signatures and their important components extracted by the NMF algorithm were run on the sample size of 183 genomes. Bar charts represent the estimated proportion of each event feature assigned to each signature (rows sum to one). **b,** Structural variation signatures in prostate cancer ranked by mutational processes observed from small deletion to reciprocal rearrangement. The eight most distinctive signatures and their important components extracted from 44 features using the NMF algorithm were run on the sample size of 183 genomes. Bar charts represent the estimated proportion of each event feature assigned to each signature (rows sum to one). **c,** Frequency of SBS, DBS, ID, CN and SV features across 183 tumours. Colours at the bottom panel show the following ethnic groups: *i*) African, red; *ii*) Admixed, green; and *iii*) European, blue. **d,** Stacked barplots of multiple signature exposures for each mutational type enriched per patient and ranked by ethnic group. Copy number and structural variation signatures (CN1-6 and SV1-8, respectively) are the first identified in this study for prostate cancer, and their enrichment in a patient appears to be significantly associated (P-values <0.05) with our GMS, considering either *de novo* or global mutational signatures discovered in the Catalogue of Somatic Mutations in Cancer (COSMIC).

Extended Data Fig. 9 | Total profiles of SBS, DBS, ID, CN and SV signatures. The classification of each signature type (SBS, 96 classes; DBS, 78 classes; ID, 83 classes; CN, 45 classes; and SV, 44

652 classes) is described in Supplementary Methods. The plotted data are available in digital form
653 (Supplementary Table 9).

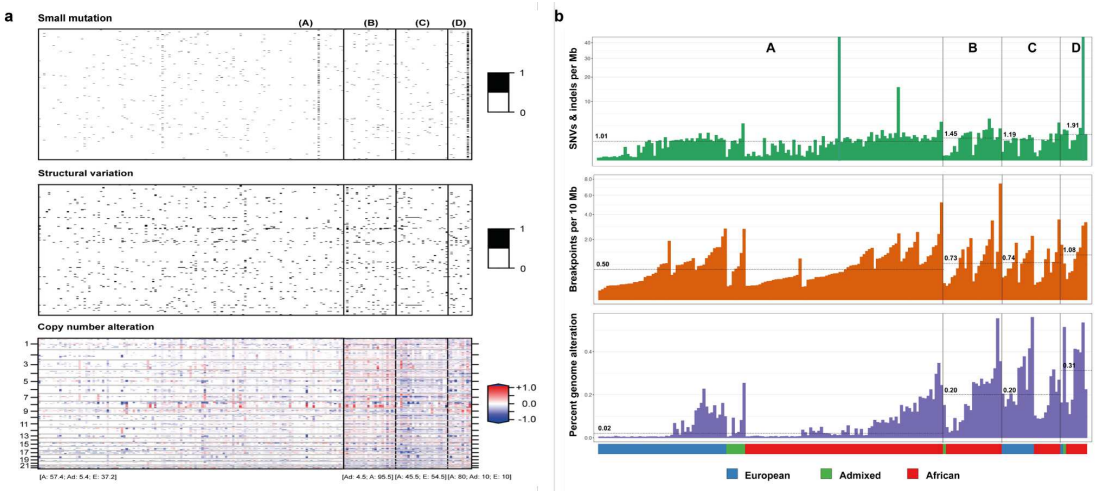
654 **Extended Data Fig. 10 | Stages of prostate tumour development. a,** Clonal architecture and its
655 frequency in prostate cancer between Africans and Europeans. Tumours are divided into three groups:
656 monoclonal, linear and branching polyclonal. The number of small somatic mutations (SSM) and CNA
657 as percentage of genome alteration (PGA) is provided as median and range in bracket. Cancer cell
658 fraction (CCF) in each clone and/or subclone is shown in a circular node. Tumours that show
659 characteristics consistent with being polytumours or with multiple independent primary tumors are
660 excluded to remain conservative. **b,** Unbiased hierarchical clustering of CNA between clonal (trunk)
661 and subclonal (branch) mutations. Trunk mutations encompass those that occur between the root node
662 (normal) and its only child node, while all others are classified to have occurred in branch. Red
663 indicates gain; blue indicates loss; and rows indicate patients. Unidentified regions in trunk and branch
664 are assumed to have neutral copy number. ConsensusClusterPlus showed seven CNA clusters among
665 our patients to be optimal. The figure shows that a trunk alteration from one patient is mutationally
666 similar to a branch alteration from another, rather than to other trunk ones from different patients in a
667 cohort. **c,** Cancer timelines of GMS-B and D identified in this study. Detailed explanation is provided
668 in Fig. 5. **d,** Relative ordering model (PhylogicNDT LeagueModel) results for a cohort of samples
669 (n=66). The samples can be analysed if they have somatic events of interest prevalent greater than 5%
670 of the sample size and have informative clonal status available for each event (16 events). Probability
671 distributions show the uncertainty of timing for specific events in the cohort.

672 **Figures**



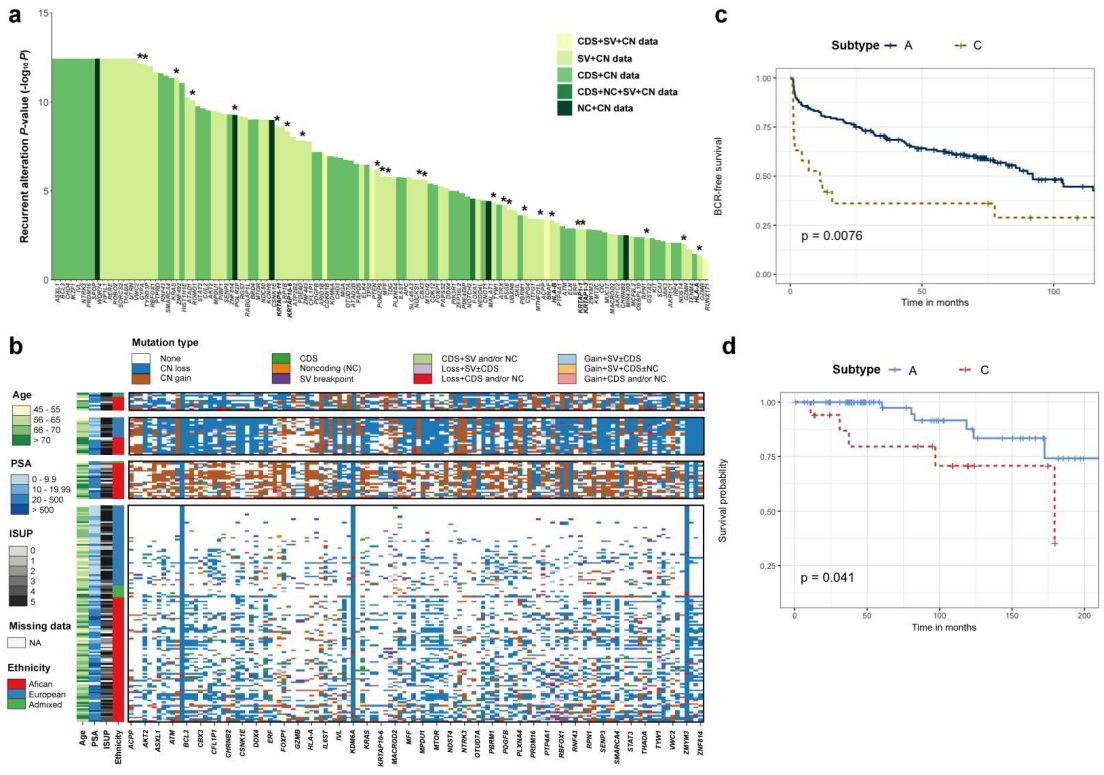
673

674 **Fig. 1 | Mutational density in prostate tumours of different ancestries. a,** Distribution of somatic
675 aberration (event number or number of base pairs) for seven mutational types across 183 tumour-blood
676 WGS pairs. **b,** Different types of mutational burden observed in this cohort. Samples are percentile
677 ranked and then ordered based on the sum of percentiles across the mutational types observed in each
678 ethnic group (left panel). Spearman's correlation is shown between mutation types, with dot size
679 representing the magnitude of correlation and background colour giving statistical significance of FDR
680 values (right panel).



681

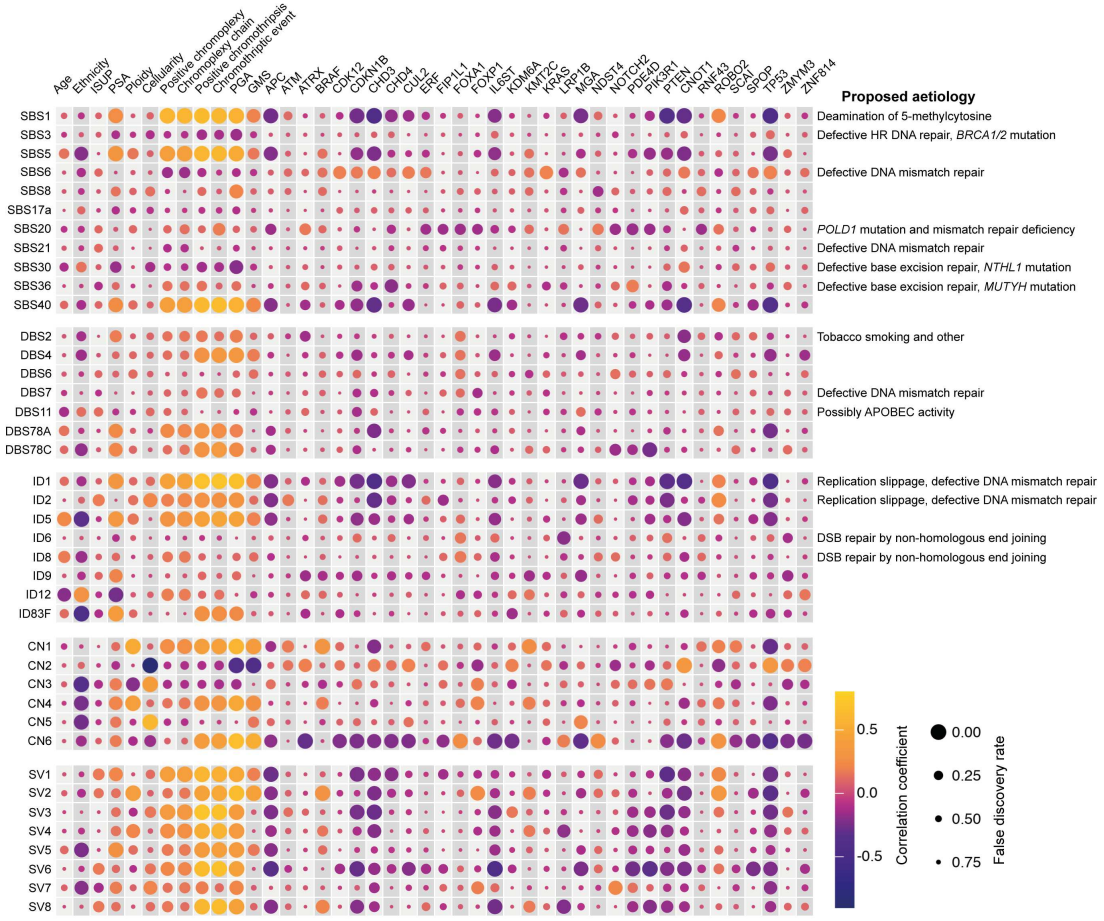
682 **Fig. 2 | Prostate cancer taxonomy of ethnically diverse populations. a,** Integrative clustering
683 analysis reveals four distinct molecular subtypes of prostate cancer. The molecular subtypes are
684 illustrated by small somatic mutations (coding regions and noncoding elements), somatic copy number
685 alterations and somatic SVs. The percentage and association between the iCluster membership and
686 patient ancestry are illustrated in square brackets. A, African ancestry; Ad, Admixed; and E, European
687 ancestry. **b,** Total somatic mutations across four molecular subtypes in this study. Dashed lines indicate
688 the median values of mutational densities across the four subtypes. For each subtype, patients are
689 ordered based on their ethnicity.



690

691 **Fig. 3 | Aberration of driver genes in four diverse subtypes. a,** Analysis of the long tail of driver
692 genes using different mutation data combined. A total of 124 genes are associated with four prostate
693 cancer subtypes, and all have previously been reported as significantly recurrent mutations/SV
694 breakpoints in the PCAWG Consortium[22], except for ones marked by asterisks, where they are
695 assigned to be significantly mutated using whole-genome data in this study. The Y-axis shows
696 corrected *P*-values in $-\log_{10} P$. CDS, coding driver data; NC, noncoding driver data; SV, significantly
697 recurrent breakpoint data; and CN, gene-level copy number data. **b,** Unsupervised hierarchical
698 clustering of known and putative driver genes identified within four prostate cancer subtypes (A-D, a
699 bottom-up direction). Rows are patients, and columns represent 124 driver genes (alphabetical order)
700 identified using different mutational types. **c,** Kaplan-Meier plot of biochemical relapse (BCR)-free
701 survival proportion of European patients in subtype A (n=161) *versus* C (n=19). **d,** Kaplan-Meier plot
702 of cancer survival probability of European patients in subtype A (n=82) *versus* C (n=17).

703

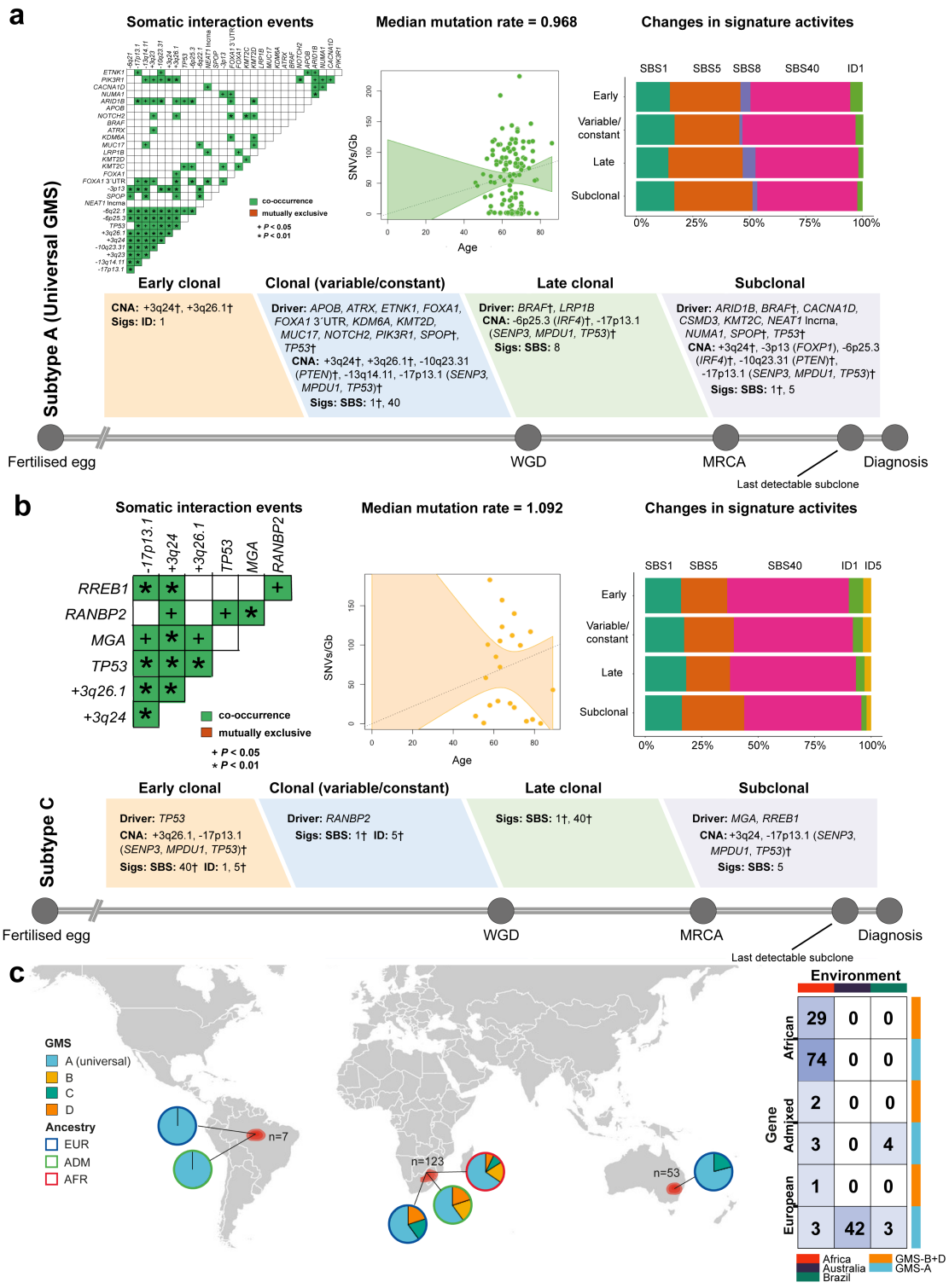


704

705 **Fig. 4 | Estimates of genomic aberrations contributed by each mutational signature.** The size of
706 each dot represents FDR values of Spearman correlation *P*-values using BH correction. The colours of
707 each dot represent correlation coefficient (rho). GMS is assigned as 1-4 for Subtypes A-D,
708 respectively; African, Admixed and European are recorded as 1-3, respectively. The correlation of 32
709 significantly mutated genes in prostate cancer is shown in the X-axis.

710

711



712

713 **Fig. 5 | Evolutionary history of globally mutated subtypes. a,** The cancer timeline of the universal

714 subtype begins from the fertilised egg to the age of the patients at a cohort. **b,** that of GMS-C.

715 Estimates for major events, such as WGD (whole-genome duplication) and the emergence of the

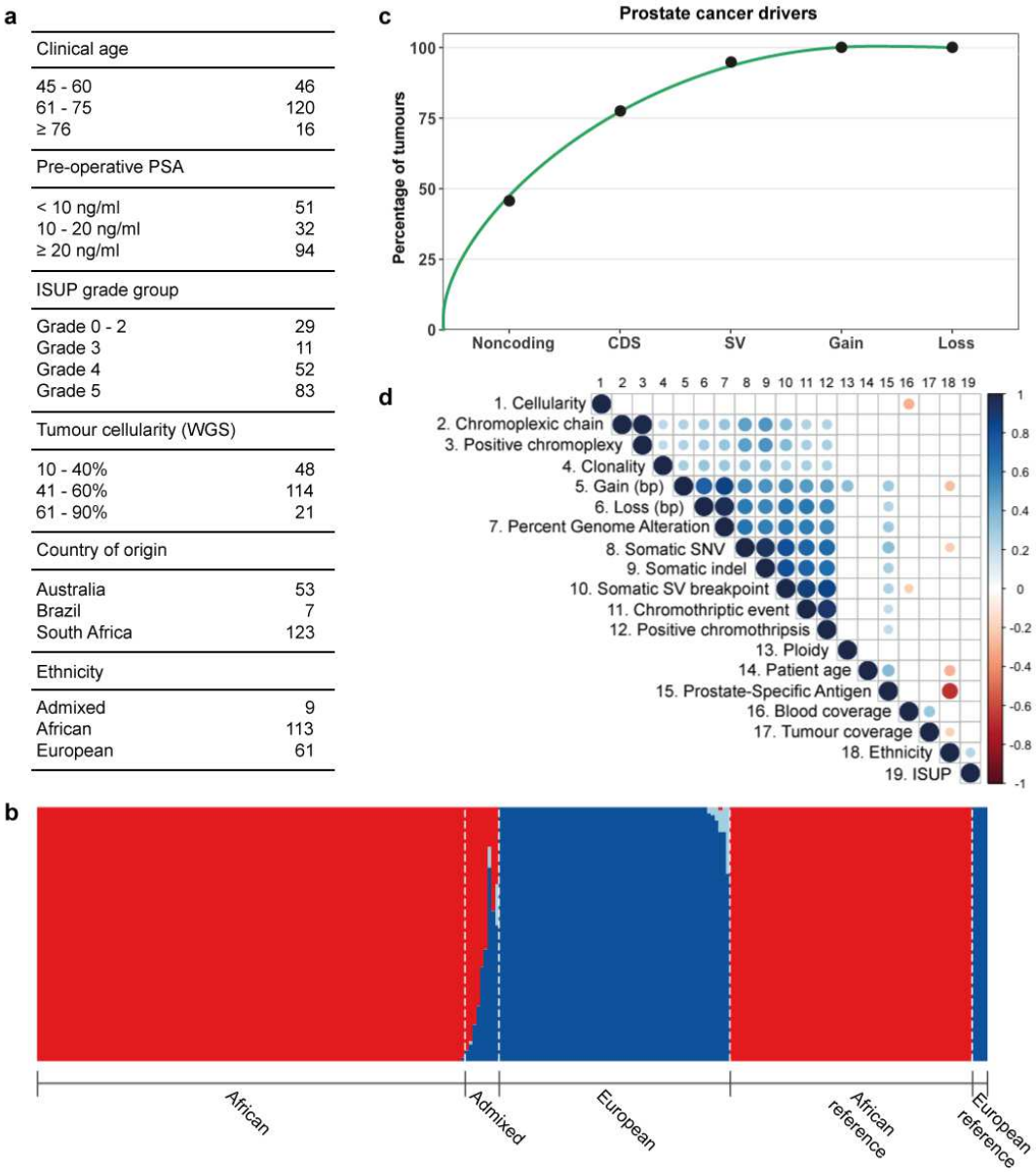
716 MRCA (the most recent common ancestor), are used to define early, variable, late and subclonal stages

of tumour evolution approximately in chronological time. When early and late clonal stages are uncertain, the variable stage is assigned. The variable/constant time period includes events that are ranked before the WGD event and also begins shortly after another break in the timeline. The late period does have a definite start, as this includes events that are ranked after WGD, when it occurs. Driver genes and CNAs are shown in each stage if present in previous studies[8,22] and defined by MutationTime.R program. Mutational signatures (Sigs) that, on average, change over the course of tumour evolution, or are substantially active but not changing, are shown in the epoch in which their activity is rather greatest. Dagger symbols denote alterations that are found to have different timing. Significant pairwise interaction events between the mutations and copy number alterations were computed using Odds Ratio (OR). Either co-occurrence or mutually exclusive event is considered if OR >2 or <0.5, respectively. Median mutation rates of CpG-to-TpG burden per Gb are calculated using age-adjusted branch length of cancer clones and maximally branching subclones. **c,** Schematic representation of a world map with the distribution of GMS (A–D) among ethnically/globally diverse populations. The gene-environment interaction model of globally mutated subtypes is shown in the right panel. The contingency table of number of patients with different ancestries (germline variants) stratified by subtypes and associated with certain geography or environmental exposure (two-sided *P*-value= 0.0005, Fisher's exact test with 2,000 bootstraps).

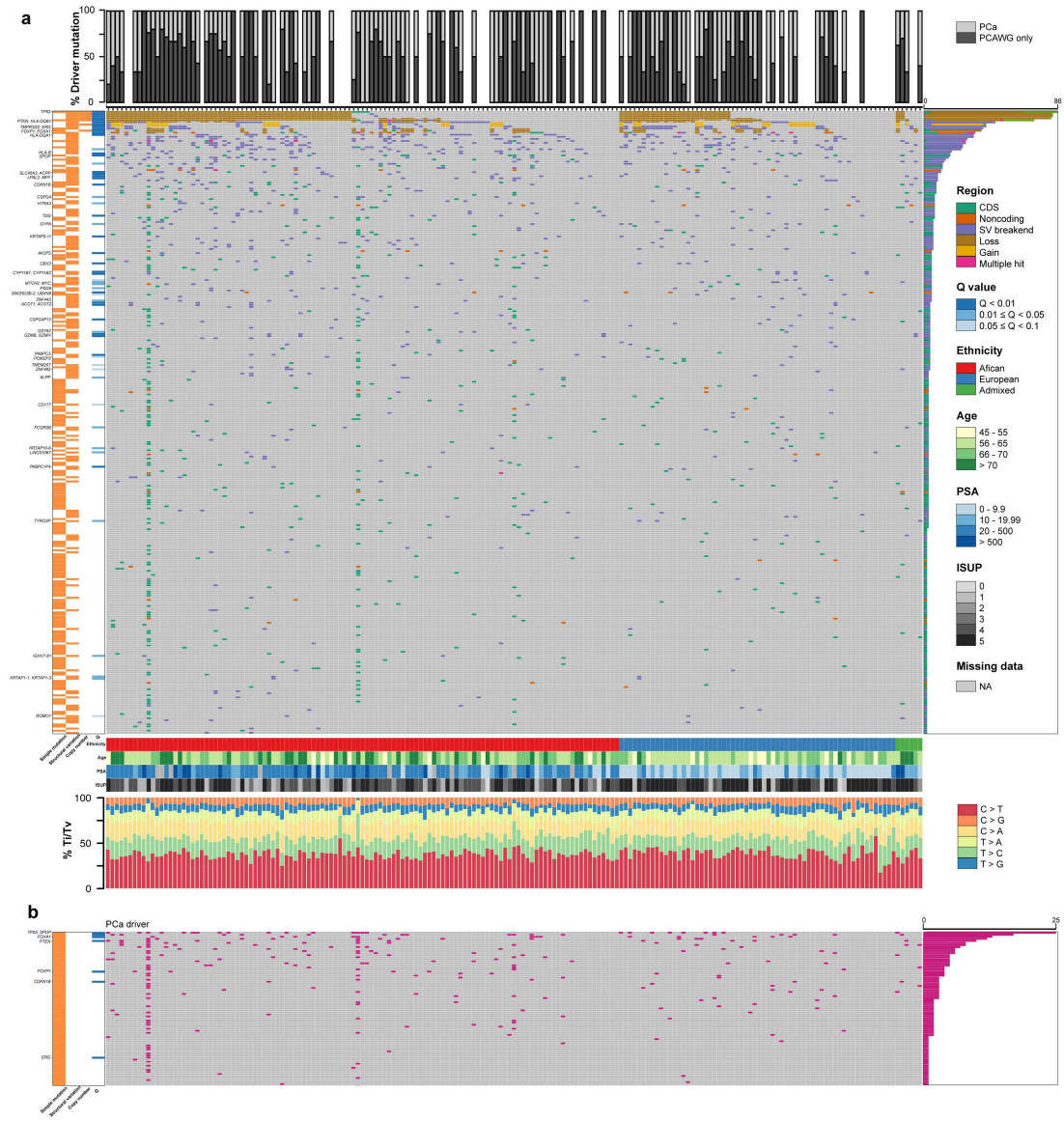
Extended data

a

Clinical age	
45 - 60	46
61 - 75	120
≥ 76	16
Pre-operative PSA	
< 10 ng/ml	51
10 - 20 ng/ml	32
≥ 20 ng/ml	94
ISUP grade group	
Grade 0 - 2	29
Grade 3	11
Grade 4	52
Grade 5	83
Tumour cellularity (WGS)	
10 - 40%	48
41 - 60%	114
61 - 90%	21
Country of origin	
Australia	53
Brazil	7
South Africa	123
Ethnicity	
Admixed	9
African	113
European	61



Extended Data Fig. 1 | Clinical cohorts and statistical metrics. a, Clinical and pathological patient characterisation. **b,** STRUCTURE analysis of bi-allelic germline variants with the logistic prior model. Model components used to explain structure in the plot are K=5. All spectrum of African contributions are summed and assigned as African ancestry. **c,** Saturation curve for all driver types across 183 patients. Recurrent copy number gains and losses were measured using GISTIC v2 (Supplementary Methods). CDS, coding sequence; SV, structural variation. **d,** Spearman's correlation between different variables measured in this cohort. Dot sizes represent the magnitude of correlation, with significant *P*-values <0.01.



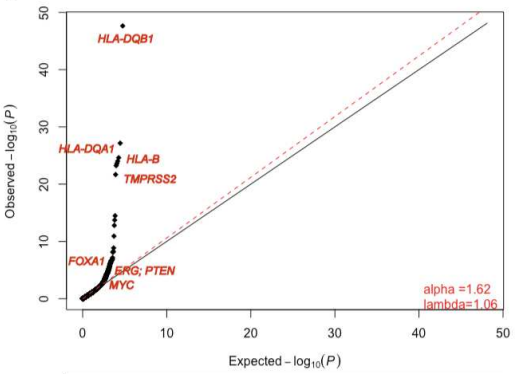
744

745 **Extended Data Fig. 2 | Somatic driver mutations in 183 prostate cancer patients.** The covariates
746 on the left show mutational types and statistical significance (FDR) from ActiveDriverWGS and
747 GISTIC2. **a,** The top 300 driver genes in PCAWG discovered in primary prostate tumours among 183
748 specimens. The top barplot shows the distribution of the number of prostate cancer drivers and/or that
749 of PCAWG. The heatmap shows drivers found in this study (rows) for each patient (columns).
750 Heatmaps are coloured by mutational type. Bottom covariates show the clinical features of patients.
751 The percentage of transition/transversion mutations across 183 patients shows 1,364,210 small somatic
752 mutations across chromosomes 1-Y. **b,** The bottom heatmap shows the top 75 of previously reported
753 coding driver genes in prostate cancer observed in this study[7,8,18,19]. The right barplot shows the number
754 of patients for each driver.

a

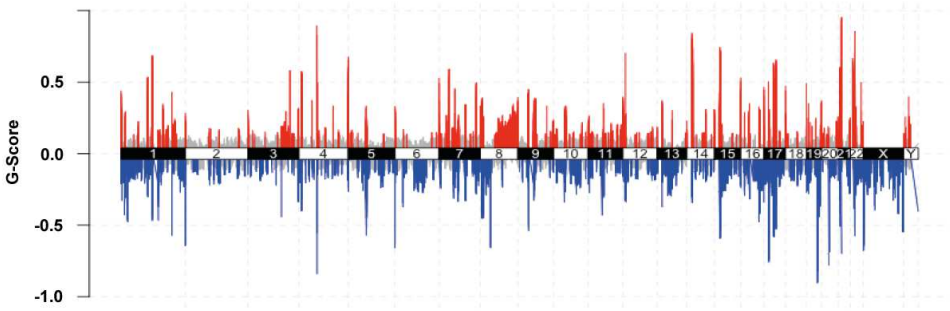
Type	PCAWG (n = 2,658)	Our study (n = 183)	Top driver	FDR
CDS	955	482	*TP53, SPOP, PTEN, FOXA1*	< 9.4e-05
Gene promoter	311	85		0.37
Small RNA	129	18	*SNORD3B-2*	< 0.04
5' UTR	118	23		0.79
Premature miRNA	95	5		0.82
Enhancer	45	4		1.00
lincRNA promoter	25	8		1.00
lincRNA exon	22	12		1.00
3' UTR	16	7	*FOXA1*	9.1e-13
miRNA promoter	12	5	chr22:38,381,983-99	0.07
Mature miRNA	2	0		1.00

c

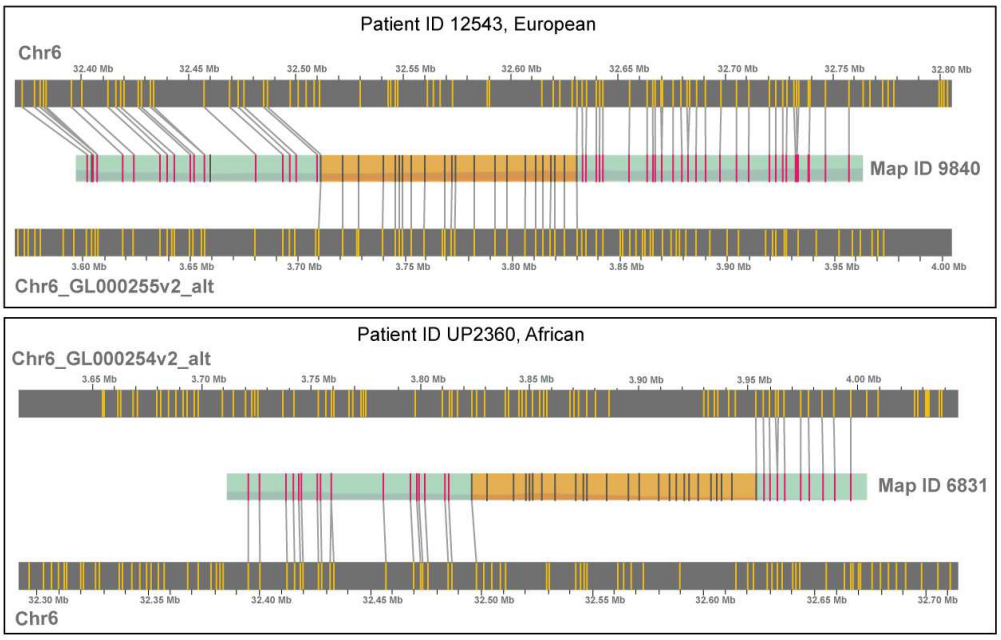


	Estimate	Std Error	Z val	Pr(>|Z|)
Intercept	-2.99	0.18	-16.27	< 2e-16**
Replication timing	-0.04	0.02	-2.51	0.01*
Gene density	0	0	12.72	< 2e-16**
Mappability (100 mer)	-8.04	0.16	-51.37	< 2e-16**
C content	-2.93	0.48	-6.09	1.1e-09**

b

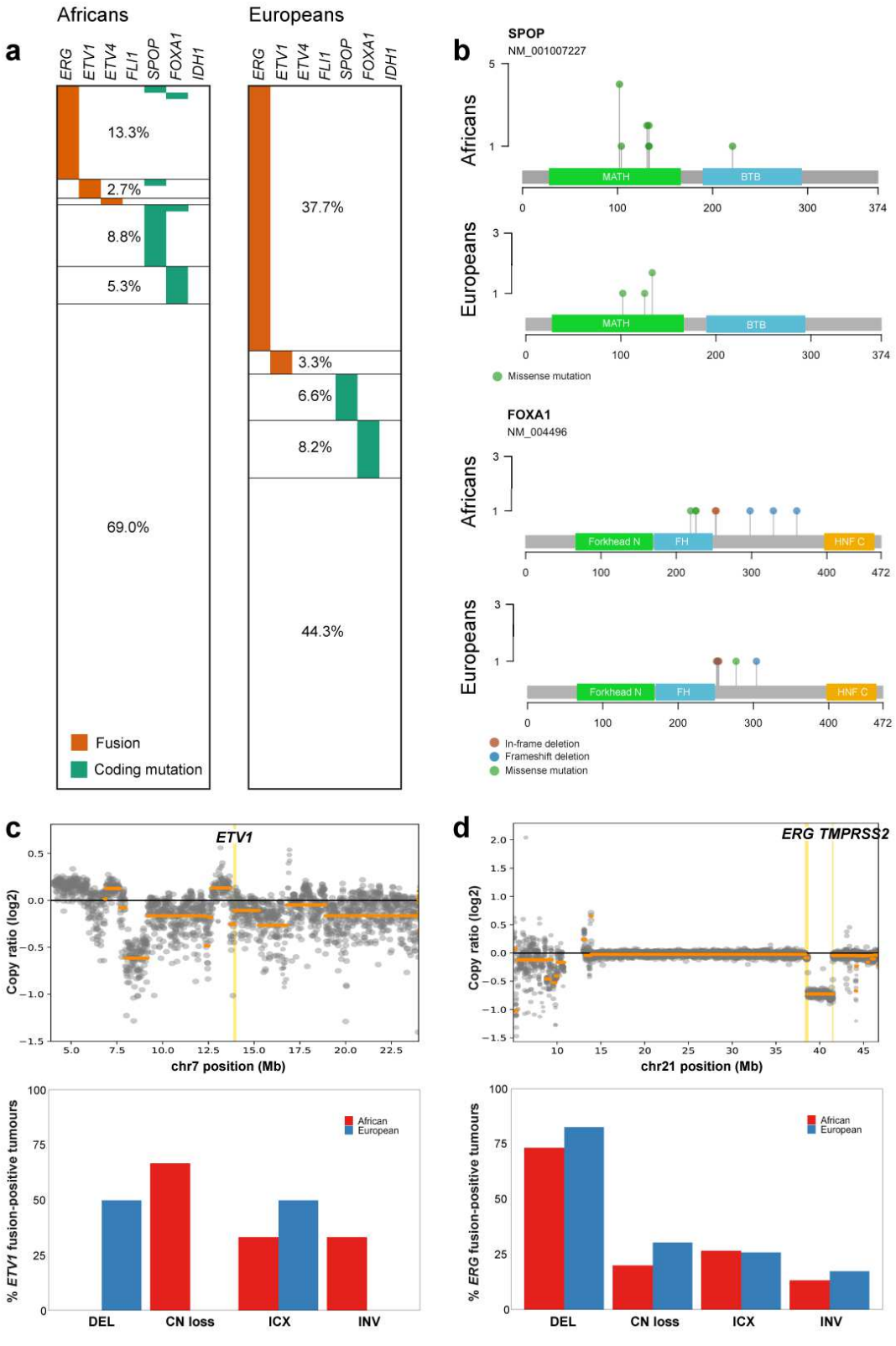


d



Extended Data Fig. 3 | Discovery of prostate cancer drivers. a, The number and types of PCAWG driver genes and elements studied in our cohort. **b,** Recurrent copy number alterations among 183 prostate tumours identified with a 99% confidence level using GISTIC v2 (Supplementary Methods).

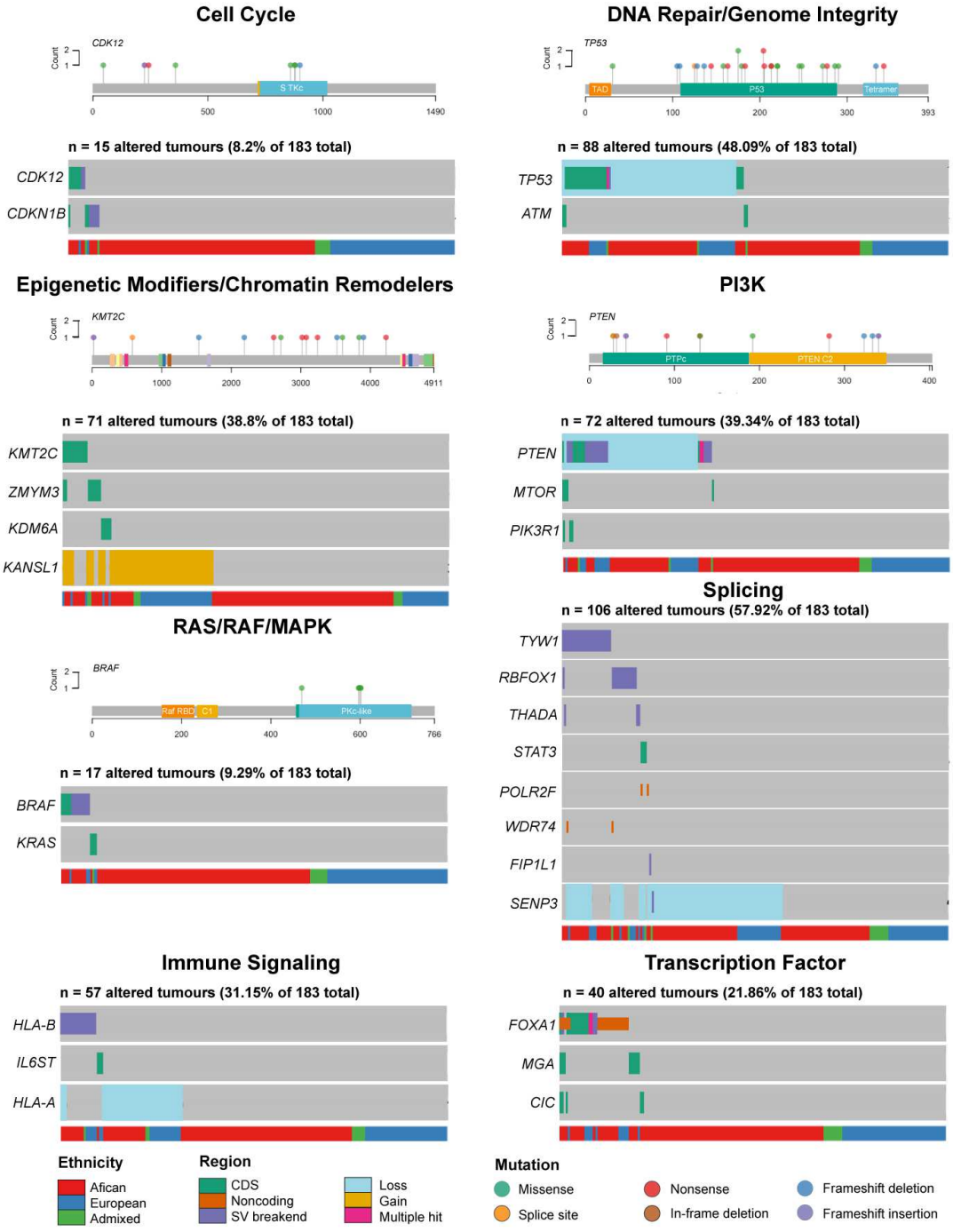
The figure shows GISTIC peaks of significant regions of recurrent amplification (red) or deletion (blue) supported by FDR <0.01. **c,** Genome-wide scan for significantly recurrent breakpoints in our study. The quantile-quantile plot shows *P*-values for mutational densities across 183 prostate cancer patients. Generalised linear modelling (GLM) of somatic mutation densities along the genome with significant background mutational processes adjusted in the model is also shown. **d,** Bionano Genomics optical genome mapping at the HLA complex. Examples of HLA translocations from a European patient (ID 12543) and an African patient (ID UP2360) studied in this cohort are characterised by pairs of optical maps, each carrying a fusion junction with flanking fragments aligning to one side of the two reference breakpoints. Using the recurrent HLA breakpoints identified in this study, the genome map of the African specimen is found to have a low-end fusion function matched with chromosome 6 through a manual inspection of unfiltered consensus maps using Bionano Access v15.2. Note that the HLA alternate contig fused in the European tumour is different from one suggested by short-read sequencing (chr6_GL000252v2_alt). The reference genome map is an *in silico* digest of the human reference hg38 with the DLE-1 enzyme. Genome map sizes are indicated on the horizontal axis, in megabase (Mb) units. Matching fluorescent labels between sample and reference genome map are connected by gray lines.



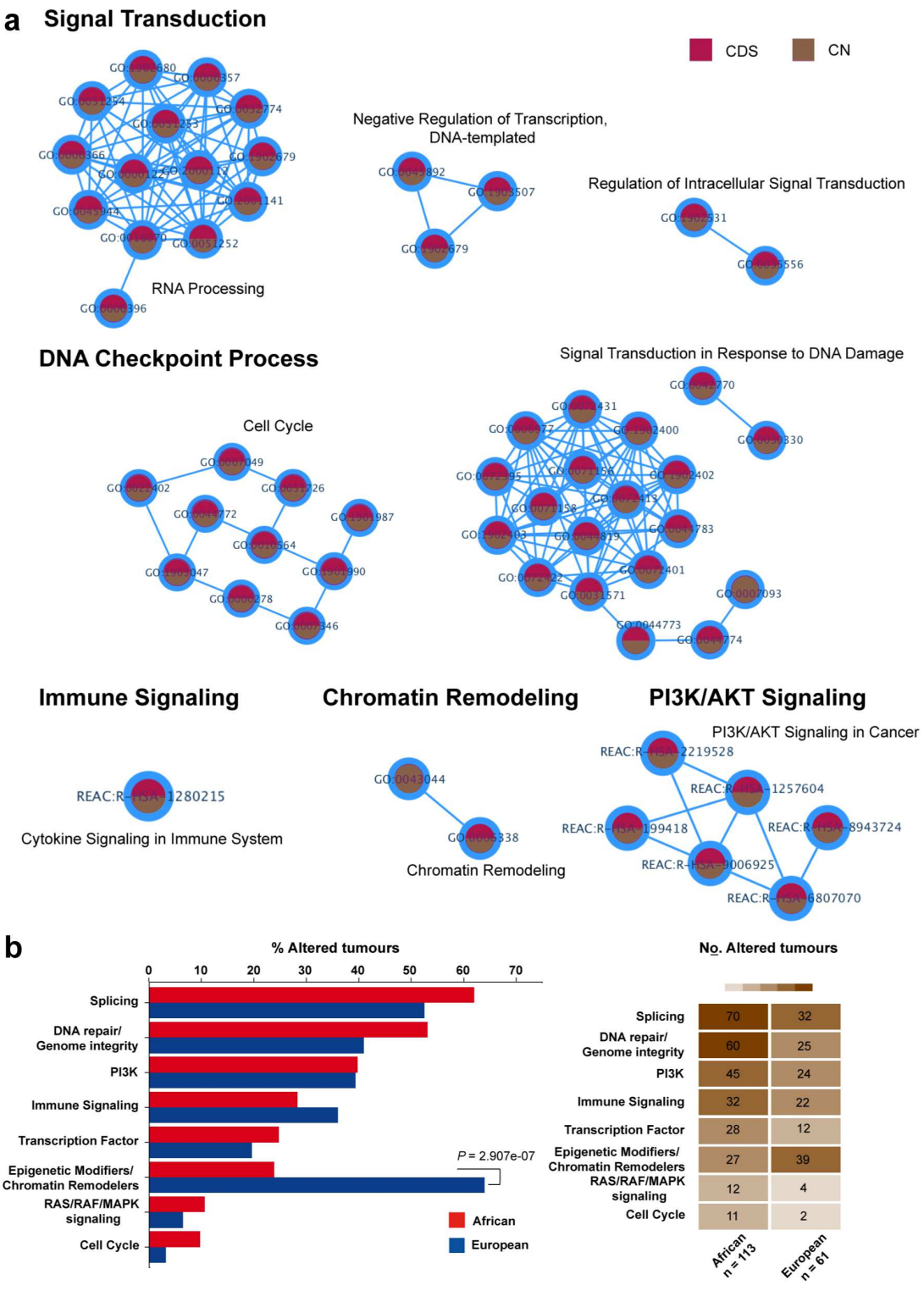
775

776 **Extended Data Fig. 4 | TCGA molecular taxonomy. a,** Seven important oncogenic drivers identified
777 by TCGA within our African and European patients. **b,** Coding mutations observed within *SPOP* and
778 *FOXA1* genes. Rarely, a mutation at the BTB domain of *SPOP* gene is shown (R221C in an African

patient, KAL0072). FH, forkhead. **c,** *ETV1* fusions within positive patients caused by copy number (CN) losses and/or structural variants (DEL, deletion; ICX, interchromosomal translocation; and INV, unbalanced or balanced inversion). CN changes in chromosome 7 show the *ETV1* loss with $\log_2$ CN ratio less than -0.2. **d,** *ERG* fusions caused by CN losses and/or structural variants.



Extended Data Fig. 5 | Prostate cancer genes and pathways. The search is carried out using the TCGA and ICGC cancer databases. The top affected genes for each pathway are present with lollipop plots to show their hotspots of simple coding mutations if they existed.



788

789 **Extended Data Fig. 6 | Major biological pathways and networks of prostate cancer. a,** Networks

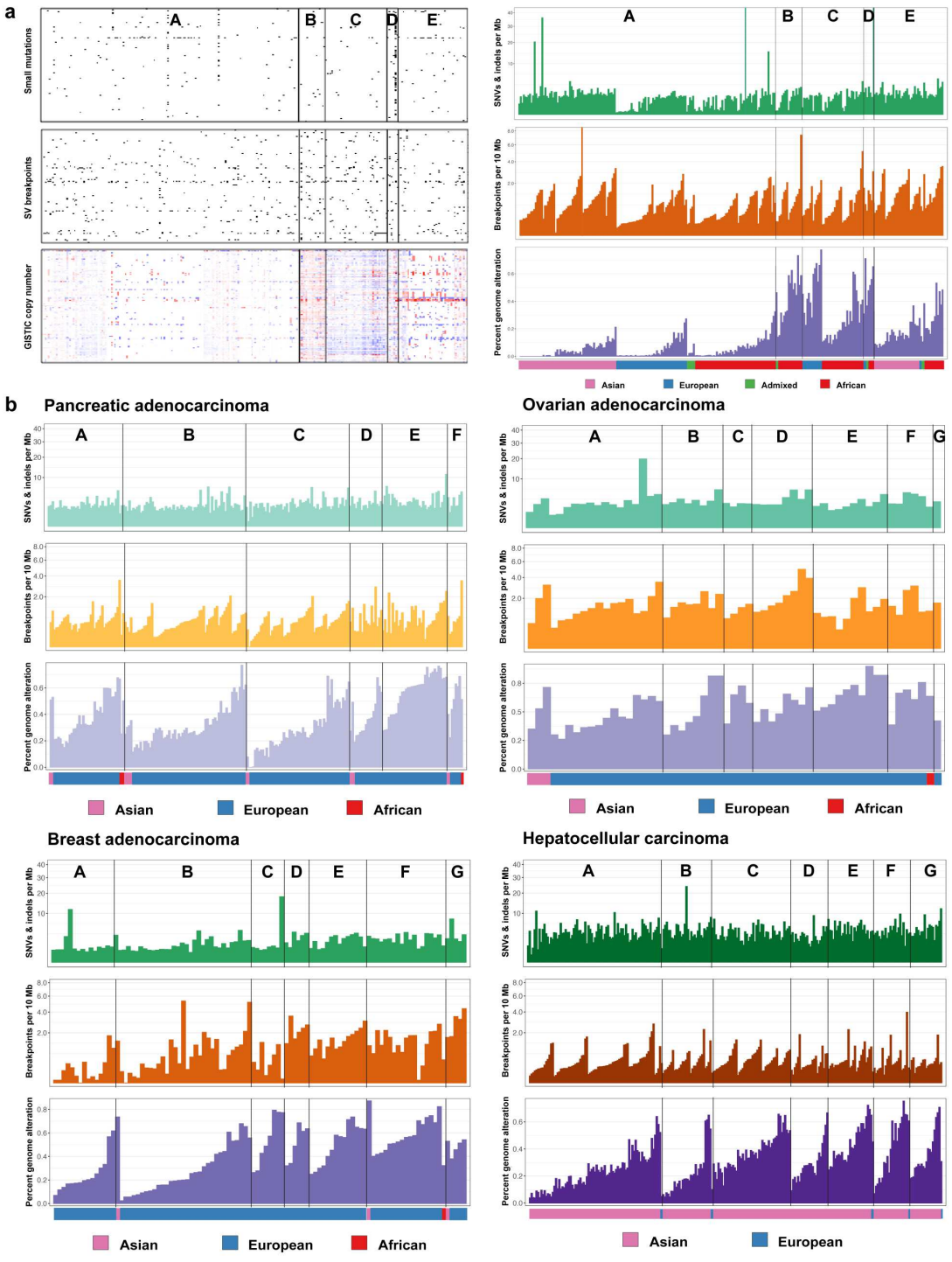
790 of functional interactions between driver genes are shown for each cancer pathway. Nodes represent

791 Gene Ontology biological processes and Reactome pathways and edges show functional interactions.

792 **b,** Pathway alteration frequencies between African and European. A sample was considered altered in

793 a given pathway if at least a single gene in the pathway had a genomic alteration. *P*-values indicate the

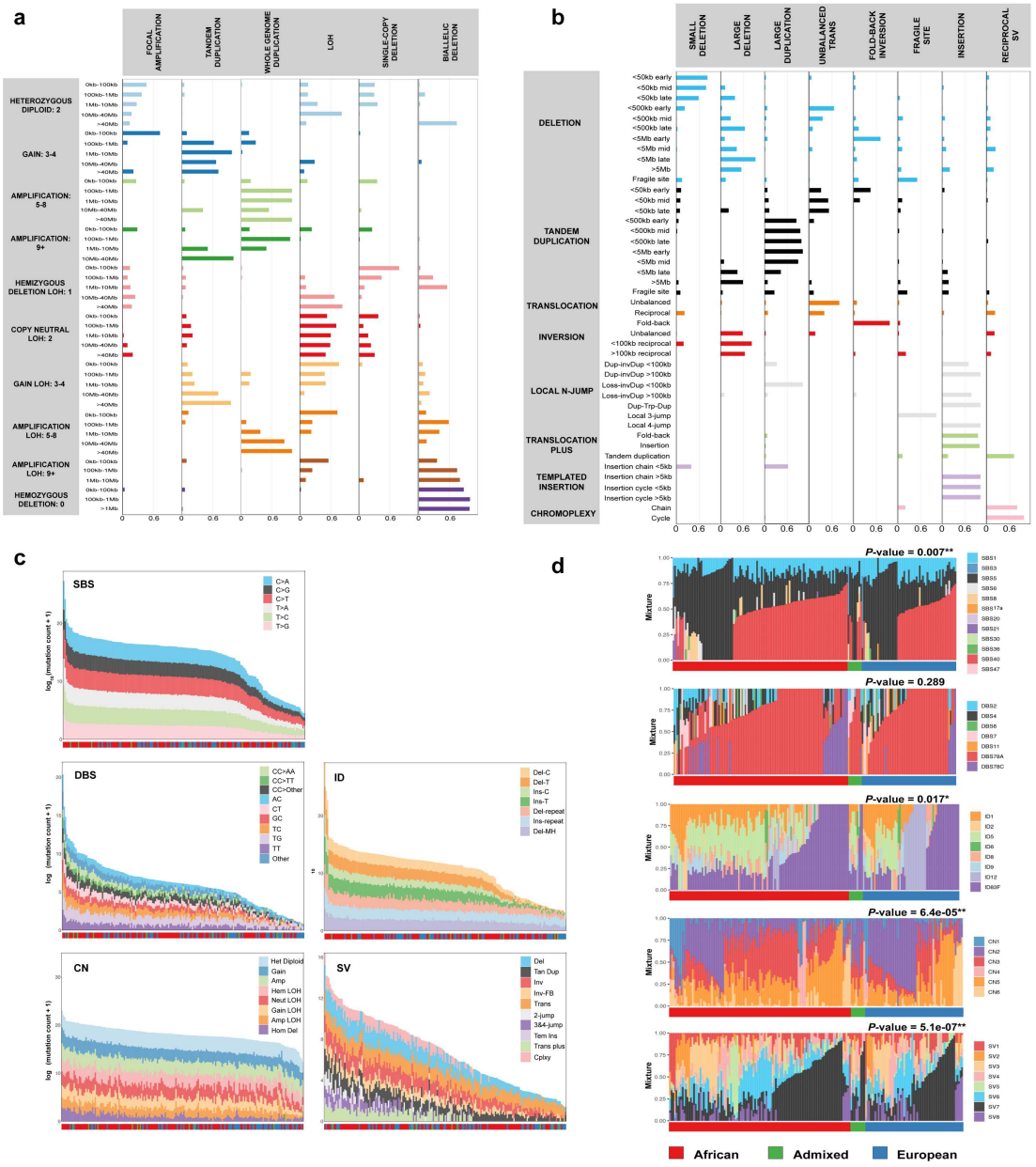
794 level of significance (two-sided Fisher's exact test).



795

796 **Extended Data Fig. 7 | Molecular subtypes in prostate cancer and pan-cancers. a,** Unsupervised
797 hierarchical clustering of primary prostate tumours across three major ethnic groups was performed
798 using total somatic mutations present within WGS normalised data. Admixed individuals were also
799 tested in prostate cancer subtypes to which they belonged. **b,** Molecular subtyping of total somatic
800 mutations within pan-cancer studies, namely pancreatic, ovarian, breast and liver cancers. Raw data of

small somatic mutations, structural variants and copy number alterations acquired per cancer were retrieved from the PCAWG[14]. For each subtype, patients are ordered based on their ethnicity. Ethnic groups are assigned using a cut-off of ancestral contribution greater than 70%; otherwise, considered as Admixed.



805

806 **Extended Data Fig. 8 | Known and novel mutational signatures in prostate cancer. a,** Copy
807 number signatures in prostate cancer across 45 CN features ranked by mutational processes observed.
808 The six most distinctive signatures and their important components extracted by the NMF algorithm
809 were run on the sample size of 183 genomes. Bar charts represent the estimated proportion of each
810 event feature assigned to each signature (rows sum to one). **b,** Structural variation signatures in prostate
811 cancer ranked by mutational processes observed from small deletion to reciprocal rearrangement. The
812 eight most distinctive signatures and their important components extracted from 44 features using the
813 NMF algorithm were run on the sample size of 183 genomes. Bar charts represent the estimated
814 proportion of each event feature assigned to each signature (rows sum to one). **c,** Frequency of SBS,

DBS, ID, CN and SV features across 183 tumours. Colours at the bottom panel show the following ethnic groups: *i*) African, red; *ii*) Admixed, green; and *iii*) European, blue. **d,** Stacked barplots of multiple signature exposures for each mutational type enriched per patient and ranked by ethnic group. Copy number and structural variation signatures (CN1-6 and SV1-8, respectively) are the first identified in this study for prostate cancer, and their enrichment in a patient appears to be significantly associated (*P*-values <0.05) with our GMS, considering either *de novo* or global mutational signatures discovered in the Catalogue of Somatic Mutations in Cancer (COSMIC).

823

824 **Extended Data Fig. 9 | Total profiles of SBS, DBS, ID, CN and SV signatures.** The classification of

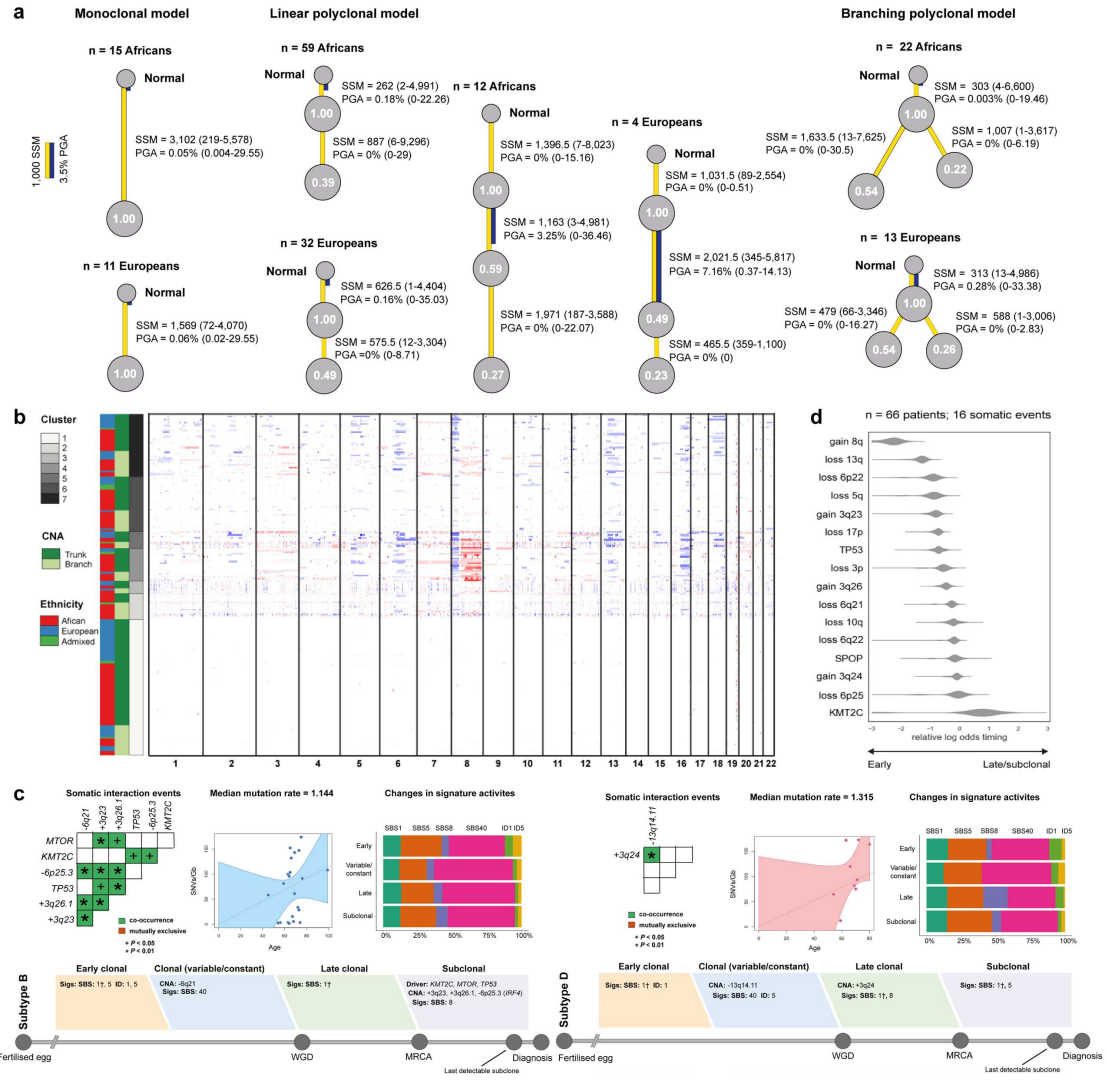
825 each signature type (SBS, 96 classes; DBS, 78 classes; ID, 83 classes; CN, 45 classes; and SV, 44

826 classes) is described in Supplementary Methods. The plotted data are available in digital form

827 (Supplementary Table 9).

828

829



Extended Data Fig. 10 | Stages of prostate tumour development. a, Clonal architecture and its frequency in prostate cancer between Africans and Europeans. Tumours are divided into three groups: monoclonal, linear and branching polyclonal. The number of small somatic mutations (SSM) and CNA as percentage of genome alteration (PGA) is provided as median and range in bracket. Cancer cell fraction (CCF) in each clone and/or subclone is shown in a circular node. Tumours that show characteristics consistent with being polytumours or with multiple independent primary tumors are excluded to remain conservative. **b,** Unbiased hierarchical clustering of CNA between clonal (trunk) and subclonal (branch) mutations. Trunk mutations encompass those that occur between the root node (normal) and its only child node, while all others are classified to have occurred in branch. Red indicates gain; blue indicates loss; and rows indicate patients. Unidentified regions in trunk and branch are assumed to have neutral copy number. ConsensusClusterPlus showed seven CNA clusters among

our patients to be optimal. The figure shows that a trunk alteration from one patient is mutationally similar to a branch alteration from another, rather than to other trunk ones from different patients in a cohort. **c,** Cancer timelines of GMS-B and D identified in this study. Detailed explanation is provided in Fig. 5. **d,** Relative ordering model (PhylogicNDT LeagueModel) results for a cohort of samples (n=66). The samples can be analysed if they have somatic events of interest prevalent greater than 5% of the sample size and have informative clonal status available for each event (16 events). Probability distributions show the uncertainty of timing for specific events in the cohort.

References

1 Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin* **71**, 209-249 (2021).

2 Alexandrov, L. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-421 (2013).

3 Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94-101 (2020).

4 Sandhu, S. *et al.* Prostate cancer. *Lancet* **398**, 1075-1090 (2021).

5 Boutros, P. C. *et al.* Spatial genomic heterogeneity within localized, multifocal prostate cancer. *Nat Genet* **47**, 736-745 (2015).

6 Berger, M. F. *et al.* The genomic complexity of primary human prostate cancer. *Nature* **470**, 214-220 (2011).

7 The-Cancer-Genome-Atlas-Network. The molecular taxonomy of primary prostate cancer. *Cell* **163**, 1011-1025 (2015).

8 Wedge, D. C. *et al.* Sequencing of prostate cancers identifies new cancer genes, routes of progression and drug targets. *Nat Genet* **50**, 682-692 (2018).

9 Lalonde, E. *et al.* Tumour genomic and microenvironmental heterogeneity for integrated prediction of 5-year biochemical recurrence of prostate cancer: a retrospective cohort study. *Lancet Oncol* **15**, 1521-1532 (2014).

10 Kamoun, A. *et al.* Comprehensive molecular classification of localized prostate adenocarcinoma reveals a tumour subtype predictive of non-aggressive disease. *Ann Oncol* **29**, 1814-1821 (2018).

11 Yamaguchi, T. N. *et al.* Molecular and evolutionary origins of prostate cancer grade. .

12 Li, J. *et al.* A genomic and epigenomic atlas of prostate cancer in Asian populations. *Nature* **580**, 93-99 (2020).

13 Crumbaker, M. *et al.* The Impact of Whole Genome Data on Therapeutic Decision-Making in Metastatic Prostate Cancer: A Retrospective Analysis. *Cancers (Basel)* **12**, E1178 (2020).

14 ICGC/TCGA-Pan-Cancer-Analysis-of-Whole-Genomes-Consortium. Pan-cancer analysis of whole genomes. *Nature* **578**, 82-93 (2020).

15 Rotimi, S. O., Rotimi, O. A. & Salhia, B. A Review of Cancer Genetics and Genomics Studies in Africa. *Front Oncol* **10**, 606400 (2021).

16 Jaratlerdsiri, W. *et al.* Whole Genome Sequencing Reveals Elevated Tumor Mutational Burden and Initiating Driver Mutations in African Men with Treatment-Naïve, High-Risk Prostate Cancer. *Can Res* **78**, 6736-6746 (2018).

17 Tindall, E. A. *et al.* Clinical presentation of prostate cancer in black South Africans. *Prostate* **74**, 880-891 (2014).

18 Robinson, D. *et al.* Integrative clinical genomics of advanced prostate cancer. *Cell* **161**, 1215-1228 (2015).

19 Armenia, J. *et al.* The long tail of oncogenic drivers in prostate cancer. *Nat Genet* **50**, 645-651 (2018).

20 Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201-206 (2016).

21 Jaratlerdsiri, W. *et al.* KhoeSan Genome Project, a catalogue of ancient human genome variation.

22 Rheinbay, E. *et al.* Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature* **578**, 102-111 (2020).

899 23 Xia, L. *et al.* Multiplatform discovery and regulatory function analysis of
900 structural variations in non-small cell lung carcinoma. *Cell Rep* **36**, 109660
901 (2021).
902 24 Taylor, B. S. *et al.* Integrative genomic profiling of human prostate cancer.
903 *Cancer Cell* **18**, 11-22 (2010).
904 25 Gerstung, M. *et al.* The evolutionary history of 2,658 cancers. *Nature* **578**,
905 122-128 (2020).
906 26 Li, C. H., Haider, S. & Boutros, P. C. Ancestry Influences on the Molecular
907 Presentation of Tumours. *bioRxiv*.
908 27 Li, Y. *et al.* Patterns of somatic structural variation in human cancer genomes.
909 *Nature* **578**, 112-121 (2020).
910 28 Houlahan, K. E. *et al.* Germline determinants of the prostate tumor genome.
911 29 Schumacher, F. R. *et al.* Association analyses of more than 140,000 men
912 identify 63 new prostate cancer susceptibility loci. *Nat Genet* **50**, 928-936
913 (2018).
914 30 Al-Olama, A. A. *et al.* A meta-analysis of 87,040 individuals identifies 23 new
915 susceptibility loci for prostate cancer. *Nat Genet* **46**, 1103-1109 (2014).
916 31 Huang, F. W. *et al.* Exome Sequencing of African-American Prostate Cancer
917 Reveals Loss-of-Function ERF Mutations. *Cancer Discov*, doi:10.1158/2159-
918 8290 (2017).
919 32 Romanel, A. *et al.* Inherited determinants of early recurrent somatic mutations
920 in prostate cancer. *Nat Commun* **8**, 48 (2017).
921 33 Taylor, R. A. *et al.* Germline BRCA2 mutations drive prostate cancers with
922 distinct evolutionary trajectories. *Nat Commun* **8**, 13671 (2017).

34 Cairns, J. Mutation selection and the natural history of cancer. *Nature* **255**, 197-200 (1975).

35 Martincorena, I. & Campbell, P. J. Somatic mutation in cancer and normal cells. *Science* **349**, 1483-1489 (2015).

36 Alexandrov, L. B. *et al.* Clock-like mutational processes in human somatic cells. *Nat Genet* **47**, 1402-1407 (2015).

37 Ottman, R. Gene–Environment Interaction: Definitions and Study Designs. *Prev Med* **25**, 764–770 (1996).

38 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics* **25**, 1754-1760 (2009).

39 Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* **11**, 11.10.11-33 (2013).

40 Raj, A., Stephens, M. & Pritchard, J. K. fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics* **197**, 573-589 (2014).

41 Cortés-Ciriano, I. & Lee JJ, X. R., Jain D, Jung YL, Yang L, Gordenin D, Klimczak LJ, Zhang CZ, Pellman DS; PCAWG Structural Variation Working Group, Park PJ; PCAWG Consortium. Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nat Genet* **52**, 331–341 (2020).

42 Baca, S. C. *et al.* Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666-677 (2013).

43 Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* **12**, R41 (2011).

44 Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029-1041.e1021 (2017).

45 Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495-501 (2014).

46 Mo, Q. *et al.* Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proc Natl Acad Sci U S A* **110**, 4245-4250 (2013).

47 Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47-54 (2016).

48 Du, Q. *et al.* Replication timing and epigenome remodelling are associated with the nature of chromosomal rearrangements in cancer. *Nat Commun* **10**, 416 (2019).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- HRPCaSupplementaryMETHODS.pdf
- S1Clinicalcohortcharacteristicsandsequencingquality.xlsx
- S2Driverinformationbypatient.xlsx
- S3GISTIC2resultsofallgenomiclesionsunder99Xconfidencelevel.xlsx
- S4ListofsignificantlyrecurrentSVbreakpointsatFDRlowerthan0.10.xlsx
- S5TCGAprostatecancertaxonomyidentifiedinthisstudy.xlsx
- S6IntegrativeiClusteranalysisof183prostatetumours.xlsx
- S7Listof124preferentiallymutatedgeneswithinfourtumoursubtypes.xlsx
- S8Pathwayenrichmentanalysisof124preferentiallymutatedgenes.xlsx
- S9Totalmutationalsignatureprofilesacross183tumours.xlsx
- S10Crossindividualcontaminationlevel.xlsx
- S11Cancerevolutionanalysisofprostatecancer.xlsx