

RESEARCH

Open Access



# Genome-wide interrogation of structural variation reveals novel African-specific prostate cancer oncogenic drivers

Tingting Gong<sup>1,2,3</sup>, Weerachai Jaratlerdsiri<sup>1,2</sup>, Jue Jiang<sup>1,2</sup>, Cali Willet<sup>4</sup>, Tracy Chew<sup>4</sup>, Sean M. Patrick<sup>5</sup>, Ruth J. Lyons<sup>2</sup>, Anne-Maree Haynes<sup>2</sup>, Gabriela Pasqualim<sup>6,7</sup>, Ilma Simoni Brum<sup>6</sup>, Phillip D. Stricker<sup>2,8</sup>, Shingai B. A. Mutambirwa<sup>9</sup>, Rosemarie Sadsad<sup>4</sup>, Anthony T. Papenfuss<sup>10,11</sup>, Riana M. S. Bornman<sup>5</sup>, Eva K. F. Chan<sup>2,12</sup> and Vanessa M. Hayes<sup>1,2,5,13\*</sup> 

## Abstract

**Background:** African ancestry is a significant risk factor for advanced prostate cancer (PCa). Mortality rates in sub-Saharan Africa are 2.5-fold greater than global averages. However, the region has largely been excluded from the benefits of whole genome interrogation studies. Additionally, while structural variation (SV) is highly prevalent, PCa genomic studies are still biased towards small variant interrogation.

**Methods:** Using whole genome sequencing and best practice workflows, we performed a comprehensive analysis of SVs for 180 (predominantly Gleason score  $\geq 8$ ) prostate tumours derived from 115 African, 61 European and four ancestrally admixed patients. We investigated the landscape and relationship of somatic SVs in driving ethnic disparity (African versus European), with a focus on African men from southern Africa.

**Results:** Duplication events showed the greatest ethnic disparity, with a 1.6- (relative frequency) to 2.5-fold (count) increase in African-derived tumours. Furthermore, we found duplication events to be associated with *CDK12* inactivation and *MYC* copy number gain, and deletion events associated with *SPOP* mutation. Overall, African-derived tumours were 2-fold more likely to present with a hyper-SV subtype. In addition to hyper-duplication and deletion subtypes, we describe a new hyper-translocation subtype. While we confirm a lower *TMPRSS2-ERG* fusion-positive rate in tumours from African cases (10% versus 33%), novel African-specific PCa ETS family member and *TMPRSS2* fusion partners were identified, including *LINC01525*, *FBXO7*, *GTF3C2*, *NTNG1* and *YPEL5*. Notably, we found 74 somatic SV hotspots impacting 18 new candidate driver genes, with *CADM2*, *LSAMP*, *PTPRD*, *PDE4D* and *PACRG* having therapeutic implications for African patients.

**Conclusions:** In this first African-inclusive SV study for high-risk PCa, we demonstrate the power of SV interrogation for the identification of novel subtypes, oncogenic drivers and therapeutic targets. Identifying a novel spectrum of SVs in tumours derived from African patients provides a mechanism that may contribute, at least in part, to the observed ethnic disparity in advanced PCa presentation in men of African ancestry.

\*Correspondence: [vanessa.hayes@sydney.edu.au](mailto:vanessa.hayes@sydney.edu.au)

<sup>1</sup> Ancestry and Health Genomics Laboratory, Charles Perkins Centre, School of Medical Sciences, Faculty of Medicine and Health, University of Sydney, Camperdown, NSW, Australia

Full list of author information is available at the end of the article



© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

**Keywords:** Chromosomal instability, Prostate cancer, African ancestry, Advanced disease, Ethnic disparity, Whole genome sequencing

## Background

Prostate cancer (PCa) is a significant health burden for men of African ancestry. In the USA, African American men are more likely to present with aggressive disease [1], with mortality rates 2.3- ( $\geq 65$  years) and 3.1-fold ( $< 65$  years) greater than men of European ancestry and as much as 5-fold greater than men of Asian ancestry [2]. Within sub-Saharan Africa, mortality rates are double global averages, reaching as much as 2.7-fold for southern Africa [3]. Previously, we have shown that southern Africans have a 2.1-fold greater risk for aggressive PCa at presentation than reported for African Americans (adjusting for age) [4]. Hypothesising that both genetic and non-genetic factors are driving ethnic disparity, we speculate that these differences are likely to be evident in the landscape of variants acquired during tumour growth. Still, little data is available for Africa. In an attempt to close this gap, we previously reported a 1.13 to 1.8-fold increase in tumour mutational burden (TMB), defined by the total number of somatic single nucleotide variants (SNVs) and small insertions and deletions (indels; length  $< 50$  bases) per megabase (Mb) of whole genome, in predominantly treatment-naïve high-risk (Gleason score  $\geq 8$ ) prostate tumours derived from men of southern African *versus* European ancestry [5, 6]. Observing a larger range of TMB in tumours derived from African (0.031 to 170.445 mutations/Mb) compared to European patients (0.015 to 2.145 mutations/Mb), we found mutational types to be strongly correlated and, as such, tumours harbouring the greatest number of structural variations (SVs; length  $\geq 50$  bases) were more likely to be derived from men of African ancestry [6]. To the best of our knowledge, no study has performed an in-depth interrogation of the range and type of SV that may be contributing to aggressive PCa presentation in patients from any region within sub-Saharan Africa.

Investigating SVs is critical for comprehensively describing and analysing the genomic burden of PCa [7, 8]. Notably, the most common somatic alteration in PCa involves an intrachromosomal translocation or 3Mb deletion on chromosome 21, resulting in fusion of the androgen-responsive gene *TMPRSS2* and members of the E26 transformation-specific (ETS) transcription factor family [9]. *TMPRSS2-ERG* gene fusions are common to roughly 50% of prostate tumours from men of European ancestry [10], dropping to 25% in African Americans [11] and 13% in Black men from South Africa [12]. We speculate that tumours from African patients may present with a

distinct SV landscape and associated fusion oncogenes. In addition to simple deletions, insertions, duplications, inversions and inter- or intra-chromosomal translocations, SVs appear to be uniquely complex in PCa, demonstrated by the phenomenon of chromoplexy, involving an abundance of interdependently occurring translocations and deletions [7]. While we reported chromoplexy to be more frequent in tumours from European (38%) *versus* African (33%) patients, conversely, African-derived tumours were more likely to present with a larger number of inter-chromosomal chained fusions (1-6 *versus* 1-2) [6].

Expanding on our earlier work to generate through deep whole genome sequencing (WGS) a tumour mutational profile for PCa in sub-Saharan Africa, describing a new molecular taxonomy [6], in this study we provide an in-depth interrogation for the type, frequency, distribution, ethnic disparities and associated clinical impact of the largely overlooked somatic SVs. Specifically, we interrogated the landscape of somatic SVs in treatment-naïve primary prostate tumours derived from 180 African *versus* European ancestral patients, with a bias towards high-risk disease. Including 114 African ancestral men from southern Africa (South Africa) makes this study, to the best of our knowledge, the largest of its kind for the region and greater sub-Saharan Africa. The inclusion of Europeans (predominantly Australians) allowed for direct comparison for prevalence of SVs in types and genomic regions using a single experimental and analytical pipeline. Ultimately, we elucidated the potential role of somatic SVs contributing to aggressive PCa in men of African ancestry, which may at least in part explain the significant ethnic-based disparity.

## Methods

### Patient clinicopathology and ancestry assignment

In this study, 180 clinicopathologically confirmed PCa patients were recruited from South Africa ( $n=120$ ), Australia ( $n=53$ ) and Brazil ( $n=7$ ). As previously described [6], South African men were recruited at the time of diagnosis from Southern African Prostate Cancer Study (SAPCS) participating urology clinics located within the greater Limpopo and Gauteng provinces [4]. All patients were treatment naïve at time of sampling. Australians attending the Prostate Cancer Clinic at St Vincent's Hospital in Sydney, and Brazilians attending a participating academic clinic in the greater State of Rio Grande do Sul, were recruited at the time of radical prostatectomy.

A single Australian patient (15178) had received one-month-long Ozurdex therapy prior to surgery, while only two Brazilian patients could be confirmed as treatment naïve prior to sampling. Three patients recruited from South Africa with no clinicopathological evidence for PCa and described in a previous study [6] were excluded. Overall, our study was biased towards advanced disease presentation, defined as a Gleason score  $\geq 8$  (138/180, 76.7%).

Germline WGS data provides clarification of genetic ancestral contributions [6], with ancestry informativeness assigned based on 7,472,833 biallelic SNVs across the genome using the population analysis tool fastSTRUCTURE v1.0 [13]. Consequently, 115 patients (63.9%) are African ancestral (114 South African, 1 Brazilian), with >78% African genetic contribution and 111 showing no non-African contributions; 61 patients (33.9%) are European ancestral (53 Australian, 4 South African, 4 Brazilian), of which five showed minimal Asian genetic contributions (3.3–26.3%) and a single patient minimal African ancestral contribution (15.7%) and four patients were classified ancestrally as admixed (2 South African, 2 Brazilian), demonstrating large African (31–63%) and European (37–59%) genetic fractions (Additional file 1: Table S1).

Clinicopathological features of the study participants defined by ancestry show a 5-year greater mean age and 6-fold greater prostate-specific antigen (PSA) level at presentation for African *versus* European patients (Additional file 1: Table S1). Our cohort concurs with our previous findings for significantly elevated PSA levels in Black men from South Africa, irrespective of PCa status, as well as an overall older age at presentation [4]. However, on pathological analysis, for the 138 (77%) cases presenting with high-risk Gleason score  $\geq 8$  PCa, 70.4% (81/115) are of African and 86.8% (53/61) of European ancestry.

### WGS data generation

As previously described [6], all samples underwent a single technical pipeline from DNA extraction to data generation. DNA was extracted from fresh-frozen prostate tumour tissues, derived either from biopsy core at diagnosis or from surgical tissue, as well as matched blood samples, using DNeasy blood and tissue kit protocol (Qiagen, Maryland). WGS was performed with 2 $\times$ 150 cycle paired-end mode on Illumina HiSeq X Ten (21 cases) or NovoSeq (159 cases) instruments at the Kinghorn Centre for Clinical Genomics (Garvan Institute of Medical Research, Australia). Following the BROAD's best practice recommendations for "data pre-processing for variant discovery", scalable FASTQ-to-BAM (v2.0) workflow with default settings was used to align sample

sequencing reads to the GRCh38 reference genome with alternative contigs [14]. The mean depth of coverage for the tumour and matched normal samples were 90 $\times$  (range 28–139 $\times$ ) and 46 $\times$  (range 30–97 $\times$ ), respectively. Tumour purities ranged from 13 to 88% (mean of 48%), as estimated by Sequenza (v2.1.2) [15].

### Somatic structural variant calling and gene annotation

Somatic SVs were called using Manta (v1.6.0) [16] and GRIDSS (v2.8.3) [17, 18] for each pair of tumour and normal samples. High-confidence SV calls from Manta were defined as those reported with 'PASS' in the VCF FILTER field in the output VCF file. SV types reported by Manta include deletions, tandem duplications, inversions, insertions and adjacent breakend (BND) for a fusion junction in an inter-chromosomal rearranged genome. Pairs of BND were annotated as inter-chromosomal translocations. High-confidence SV calls from GRIDSS were obtained using GRIDSS accompanied R script (gridss\_somatic\_filter.R). GRIDSS reports BND for all fusion junctions resulting from any SV event. Simple SV types, defined as deletions, duplications, insertions and inversions, were assigned using the accompanied R script: simple-event-annotation.R, while inter-chromosomal BND pairs were further annotated as translocation, in the same way as Manta. When integrating call sets from Manta and GRIDSS, two SV calls were considered as concordant if they were reported as high-confidence by one of the two callers (Additional file 1: Fig. S1) and have matching SV type and reported breakpoint positions within 5bp of each other. This new filtering method is able to overcome the limitation of the high-confidence definition of different SV callers and rescue more false negatives [19] and provides a more comprehensive SV call set compared to our earlier study [6]. Germline SVs were called by Manta (v1.6.0) with filtration of 'PASS' in FILTER field in the VCF file.

Gene annotation of all SV breakpoints was performed using the Ensembl human gene annotation (GRCh38 assembly, release 99). SV BND was annotated as 'interrupting' if it was located within a gene region. A SV event is classified as a gene fusion if both BNDs interrupt two different genes. Annotation of exons of SV breakpoints in *TMPRSS2-ERG* fusion-positive samples was based on the exon regions (exonStarts and exonEnds) for all transcripts in the UCSC Table *refFlat* [last updated 08/17/2020] from the NCBI RefSeq track for GRCh38. There are two transcripts of *TMPRSS2*, each with 14 exon regions and 10 transcripts of *ERG*, each with a different number of exon regions [5–12]. For each transcript of *ERG*, the upstream exon to the SV breakpoint interrupting *ERG* was identified. For each transcript of *TMPRSS2*, the downstream exon to the SV breakpoint interrupting

*TMPRSS2* was identified. This process was done for all combinations of transcripts of *TMPRSS2* and *ERG* genes and all SV breakpoints in *TMPRSS2-ERG* fusion-positive samples.

#### Germline and somatic mutation (SNVs and indels) calling and annotation

Following the BROAD's best practice recommendations for "germline short variant discovery (SNPs + Indels)" and "somatic short variant discovery (SNVs + Indels)", small germline and somatic mutations (SNVs and indels) were called using the scalable Germline-ShortV v.1.0 [20] and Somatic-ShortV v.1.0 workflows [21], respectively. Both germline and somatic variants were annotated using annovar (version 2019Oct24) with the RefSeq gene database (build version Hg38) [22].

#### Copy number variation calling

Somatic copy number variation (CNV) with discrete copy number segments were determined using the copy number calling pipeline of CNVKit [23]. These were further examined using GISTIC v2.0.23 [24] to identify CNV at the gene level. CN gains (amplifications) or losses (deletions) per gene were determined based on CN values estimated as 2 or -2 respectively from GISTIC output CN values (all\_threshold.by\_genes.txt). CN values were estimated as  $\pm 2$  if exceeding the high-level thresholds and  $\pm 1$  if exceeding the low-level thresholds, but not the high-level thresholds [24]. For CN gains and losses, the low-level threshold values are 0.1 and -0.1 respectively, while the high-level thresholds were calculated on a sample-by-sample basis by GISTIC.

#### Recurrent mutation in hyper-SV mutated tumours

Recurrent somatic mutations (SNVs, indels and CNVs) in hyper-SV mutated tumours were examined for 631 previously described PCa driver genes [6]. Logistic regression was used to test the null hypothesis of no correlation between the total count (and relative frequency) of each SV type and most recurrent mutated genes, based on variant types of SNVs, indels and CNVs. *P*-values were adjusted for multiple testing correction using the Benjamini-Hochberg method.

#### Gene biallelic inactivation classification

We examined three genetic inactivation types defined in the study of Campbell et al. [25], including 'Loss' (somatic or germline deletions), 'Break' (somatic or germline SVs) and 'Mutation' (somatic or germline SNVs) for *BRCA2* in hyper-deletion and *CDK12* in hyper-duplicated tumours. For a gene *G* with A and B alleles ( $G^{A/B}$ ), four classes of biallelic inactivation of both alleles ( $G^{-/-}$ ) were defined as (1) Loss/Mutation, loss of the A allele and

nonsynonymous driver mutation of the B allele; (2) Loss/Loss, two deletions overlapping an exon and CN derived allele count is 0 both for A and B alleles; (3) Loss/Break, loss of the A allele and SVs where one or both breakpoints interrupting an exon of B allele; and (4) Mutation/Mutation, a nonsynonymous germline SNV and a nonsynonymous somatic SNV of the same gene [25].

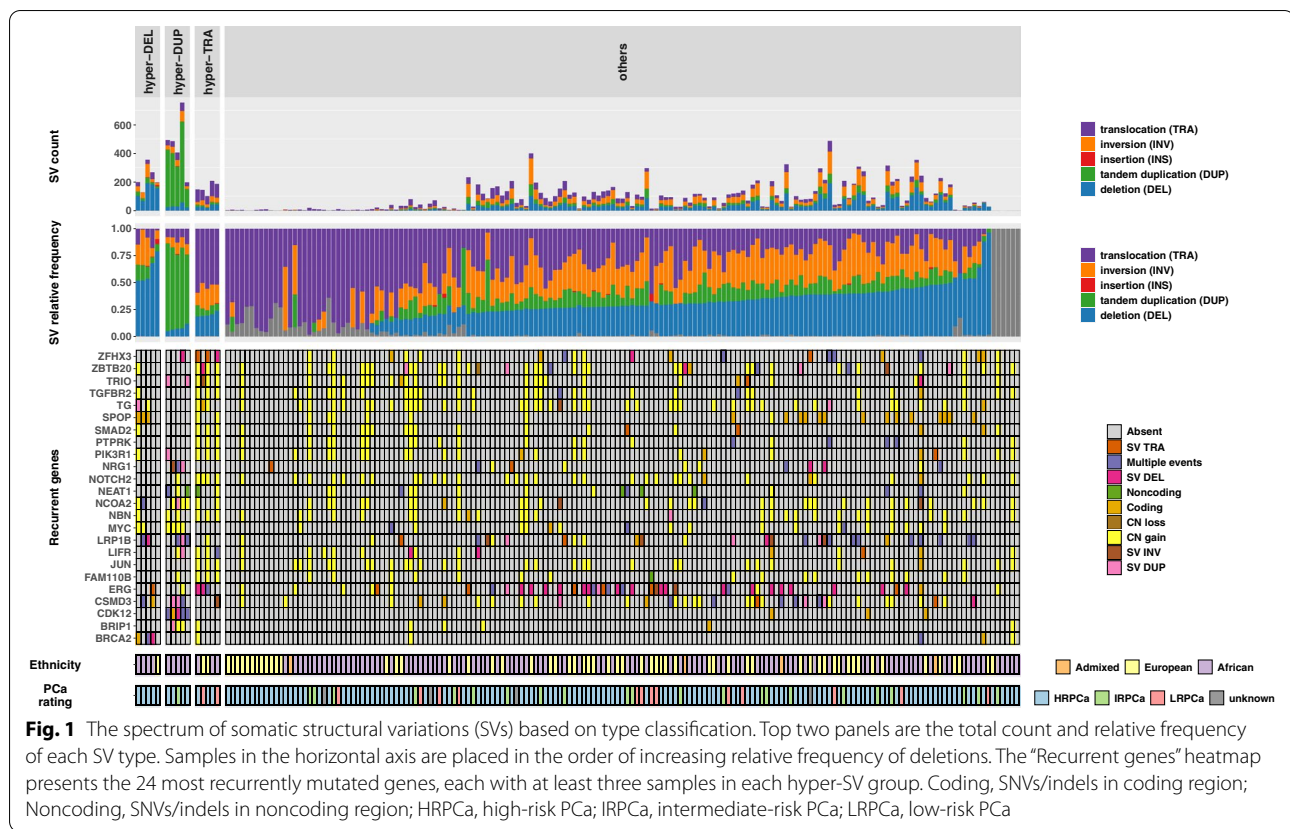
## Results

### Spectrum of somatic structural variant types

Defined by ethnicity and PCa risk at diagnosis or surgery, we observed large variability in the number of somatic SVs (range 0–754) per tumour (Fig. 1). In their simplest form, deletions and inversions were found to be the most common SV types, while consistent with other PCa studies, we observed a low frequency of insertion events (Additional file 1: Fig. S2). Only 15 tumours (8.3%) presented with at least one insertion, which may be due to the limitations of insertion detection using short-read next generation sequencing data [19]. We found duplication events to have the greatest variability by count and relative frequency among the ethnic groups, representing a 2.5- and 1.6-fold increase in tumours from African versus European patients, respectively. While not observed in our study, Quigley et al. reported a significant association between biallelic inactivating alterations in *TP53* and the frequency of inversions in mCRPC [26]. Here we found the relative frequency of inversions to be significantly associated with *SPOP* mutations (adjusted *p*-value = 0.04).

Furthermore, we define hyper-SV mutated tumours as having at least 100 (average SVs per tumour in this cohort) total count of SVs with at least 50% dominated by a single SV type. As such, we identified five hyper-duplicated (KAL070, N0067, SMU087, UP2050 and UP2133), five hyper-deleted (BRA08, KAL0013, UP2003, UP2103, UP2396) and five hyper-translocated (UP2187, 5656, 12596, UP2267 and SMU142) tumours, their CIRCOS plots are shown in Fig. S3 (Additional file 1). Hyper-SV tumours were notably biased towards tumours derived from men of African ancestry (Fig. 1), with the hyper-duplicated genomes African-specific (5/5), the hyper-deleted African-biased (4/5) and the hyper-translocated observed in tumours from both African (3) and European (2) patients. To the best of our knowledge, no study has reported the hyper-translocated PCa subtype to date [26–29].

In non-African studies, the hyper-duplicated mutation subtype has been reported in 3% and 7% of metastatic castration-resistant PCa (mCRPC) by Quigley et al. ( $n = 101$ ) [26] and Van Dessel et al. ( $n = 197$ ) respectively [29] and in 22% and 20% of smaller cohort studies by Viswanathan et al. ( $n = 23$ ) [30] and Wedge et al. ( $n =$



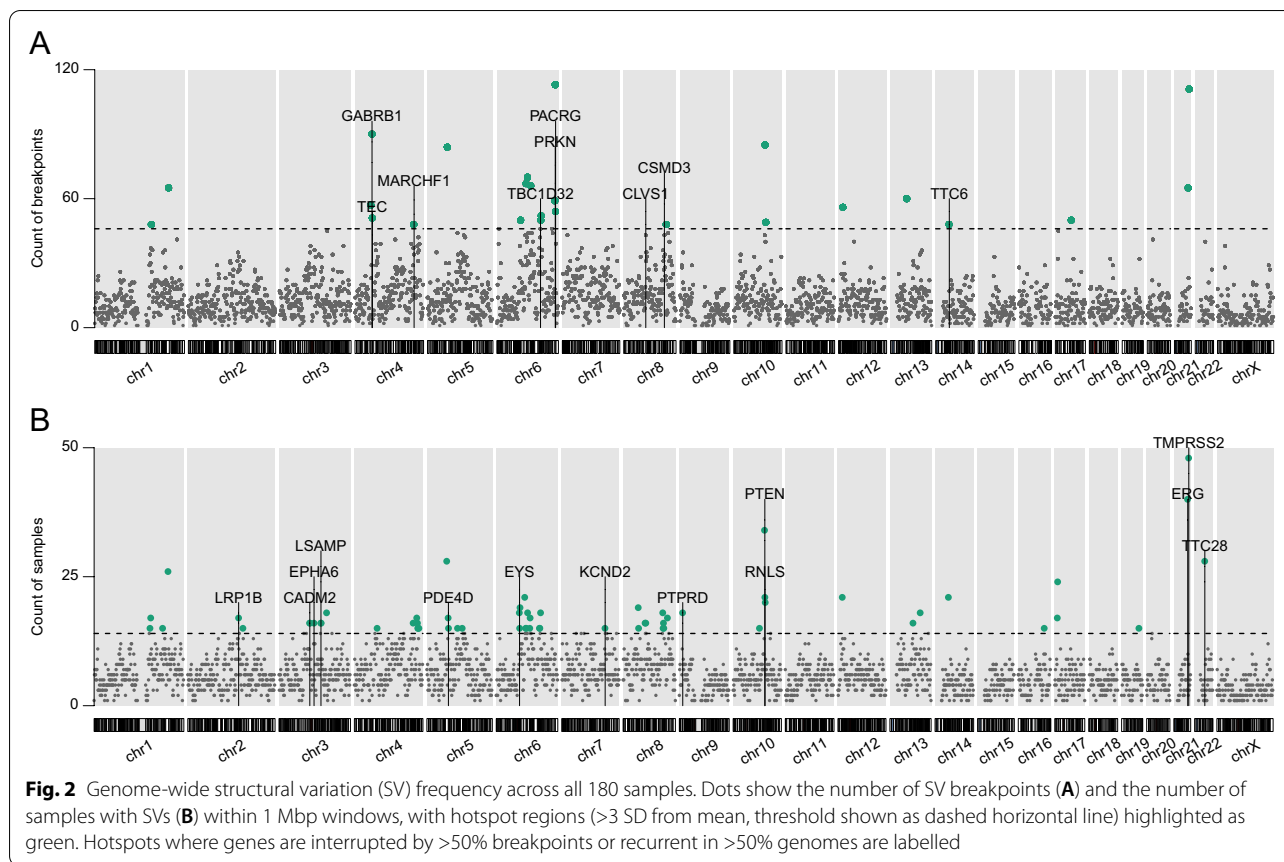
10) respectively [27]. Conversely, this subtype was notably absent in localised PCa from studies including The Cancer Genome Atlas (TCGA) WGS data ( $n = 20$ ) [31], Fraser et al. ( $n = 200$ ) [28] and Wedge et al. ( $n = 92$ ) [27]. Importantly, while sourced from primary tissue, the status of metastatic seeding is unknown for our African patients. Furthermore, enrichment of duplications has previously been associated with bi-allelic *CDK12* inactivation (*CDK12*<sup>-/-</sup>) in mCRPC [26, 29, 30]. Here we found the relative frequency of duplications per tumour to be significantly correlated with *CDK12* mutation (adjusted  $p$ -value = 0.001) with four hyper-duplicated tumours found to be *CDK12*<sup>-/-</sup> (Fig. 1 and Additional file 1: Table S2). Although tumour SMU087 has both somatic CN loss and deletion detected on *CDK12*, it did not satisfy the criteria for assessment of *CDK12* biallelic loss. In addition, we found *MYC* CN gains to be significantly associated with increased relative frequency of duplication per tumour (adjusted  $p$ -value = 0.03), with four of the hyper-duplicated tumours presenting with *MYC* CN gains (Fig. 1).

Enrichment for deletions (<100kbp) in non-African studies has been found to be associated with bi-allelic *BRCA2* mutation (*BRCA2*<sup>-/-</sup>) in mCRPC [26, 29]. In our study, we observed *BRCA2*<sup>-/-</sup> in three hyper-deleted

tumours from African patients. While the two remaining hyper-deleted tumours, BRA08 (European) and KAL0013 (African), showed no *BRCA2* loss, two or more nonsynonymous germline *BRCA2* mutations were observed for each patient (Additional file 1: Table S3), although defined by ClinVar [December 2020] [32] as 'benign' or of 'uncertain significance'. This suggests the hyper-deleted signature observed in these two patients is either unrelated to biallelic *BRCA2* loss or the clinical significance of these two germline SNVs is under-recognised. *BRCA2* mutation was not found statistically associated with count of deletions in this study. However, we found the count of deletions per tumour to be significantly associated with the presence of somatic *SPOP* mutations (adjusted  $p$ -value = 0.005, Fig. 1), which presented in three hyper-deleted tumours from African patients, including the single none-*BRCA2*<sup>-/-</sup> African-derived tumour (patient KAL0013).

#### Spectrum of somatic structural variant breakpoints

We identify SV hotspots based on (i) the total number of SV breakpoints and (ii) the number of samples with at least one SV breakpoint for each 1 Mb non-overlapping bin across the genome. Overall, each bin contained  $10 \pm 7.4$  (median  $\pm$  MAD) breakpoints from  $6.0 \pm 3.0$

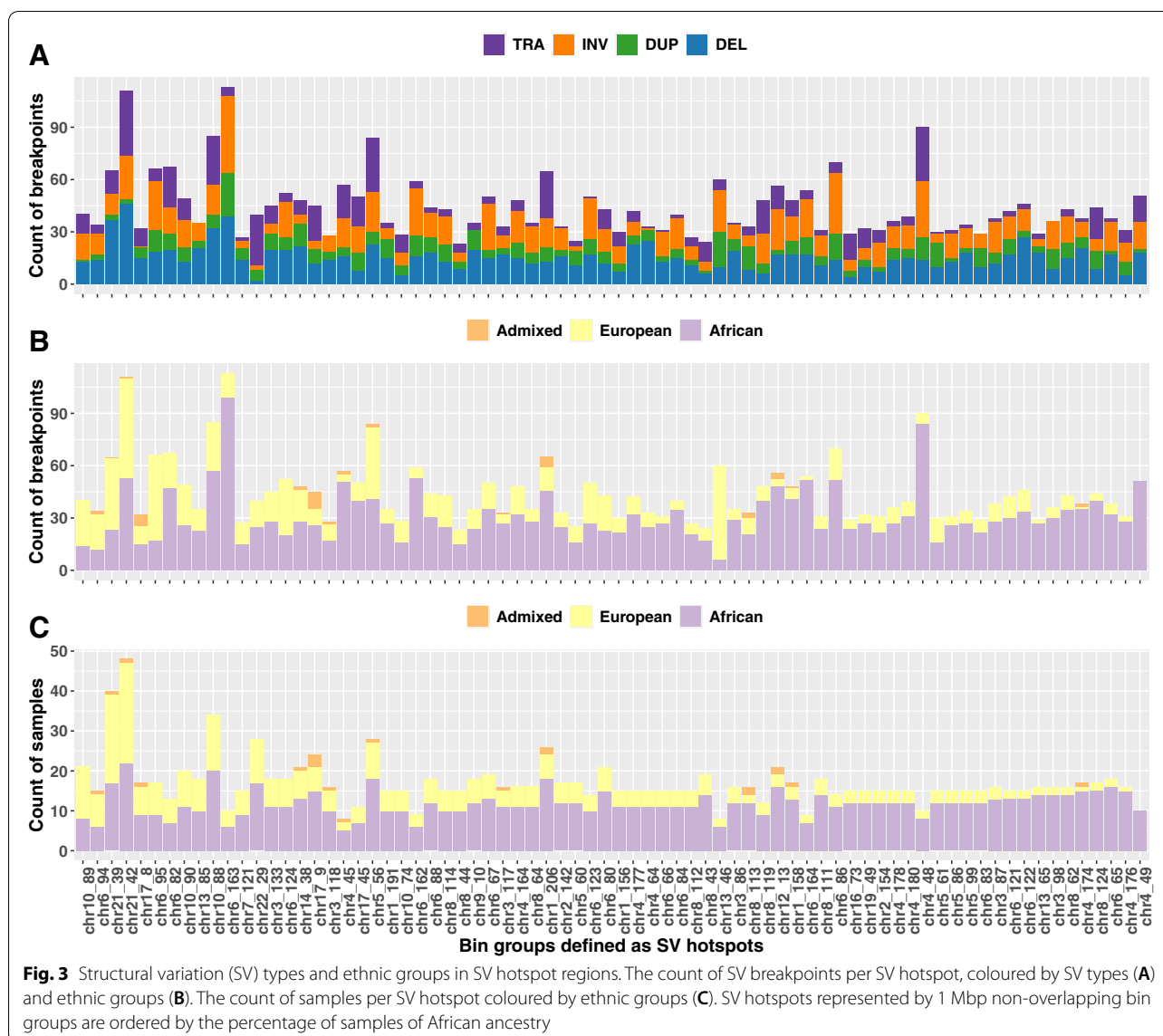


samples. SV hotspots were then defined as genomic regions most frequently ( $> Q_3 + k \times (Q_3 - Q_1)$ ) affected by SV breakpoints, either in the same genome or recurrent across genomes (Additional file 1: Fig. S4). Based on Tukey's fences approach,  $k = 1.5$  was used to find outliers in sample count (Additional file 1: Fig. S4B), while more stringent  $k = 3$  was used to define outliers for SV breakpoints count, considering clustered SV breakpoints such as chromothripsis can be attained in a single tumour (Additional file 1: Fig. S4A). In summary, 74 genomic bins (from a total of 2833) were identified as SV hotspots (Fig. 2), and 13 hotspots were found to be both frequently affected by SV breakpoints (>46 breakpoints) and recurrent among samples (>14 samples). Of all 74 hotspots, 26 presented with >50% SV breakpoint interrupting a single gene or with the single gene interruption recurrent in >50% genomes. *ERG*, *PTEN*, *CSMD3* and *LRP1B* were previously identified as driver genes associated with PCA using this sample data source and PCAWG cohorts [6], while our new method highlighted 18 additional potential driver genes, including *GABRB1*, *CLVS1*, *RNLS*, *TMPRSS2*, *TTC28*, *EYS*, *TTC6*, *PTPRD*, *PRKN*, *PACRG*, *TBC1D32*, *CADM2*, *LSAMP*, *MARCHF1*, *PDE4D*, *KCND2*, *EPHA6* and *TEC*. Among the gene candidates,

*EYS*, *PTPRD*, *PRKN*, *CSMD3*, *CADM2*, *LSAMP* and *PDE4D* are larger than the defined genomic bin (1 Mbp) and as such we cannot exclude for their co-location with the SVs being a chance event. Additionally, we recognise that *PACRG*, *LRP1B* and *PDE4D* are putative fragile sites [33]. Additionally, we observe three hotspots (chr6: 66–67 Mbp, chr6: 94–95 Mbp, chr5: 85–86 Mbp and chr13: 64–65 Mbp) with <5% breakpoints overlapping gene regions, which may indicate a different mechanism in promoting PCA.

All identified SV hotspots included multiple SV types (Fig. 3A). 12 hotspots contained >50% deletions of all SV events within the bin, including the *ERG* gene region (chr21: 38–39 Mbp). One hotspot (chr6: 66–67 Mbp) includes more inversion events (52%) and two hotspots (chr16: 72–73 Mbp and chr22: 28–29 Mbp) include more translocation events (52% and 73%, respectively). The hotspot chr8: 43–44 Mbp was found with an even distribution of deletions, duplications, inversions, and translocations of around 20% for each SV type.

Taking ethnicity into consideration, eight hotspots were found to have elevated number of SV breakpoints ( $\geq 90\%$ ) in patients representing a single ancestry, of which seven were specific to Africans (Fig. 3B). Notably,



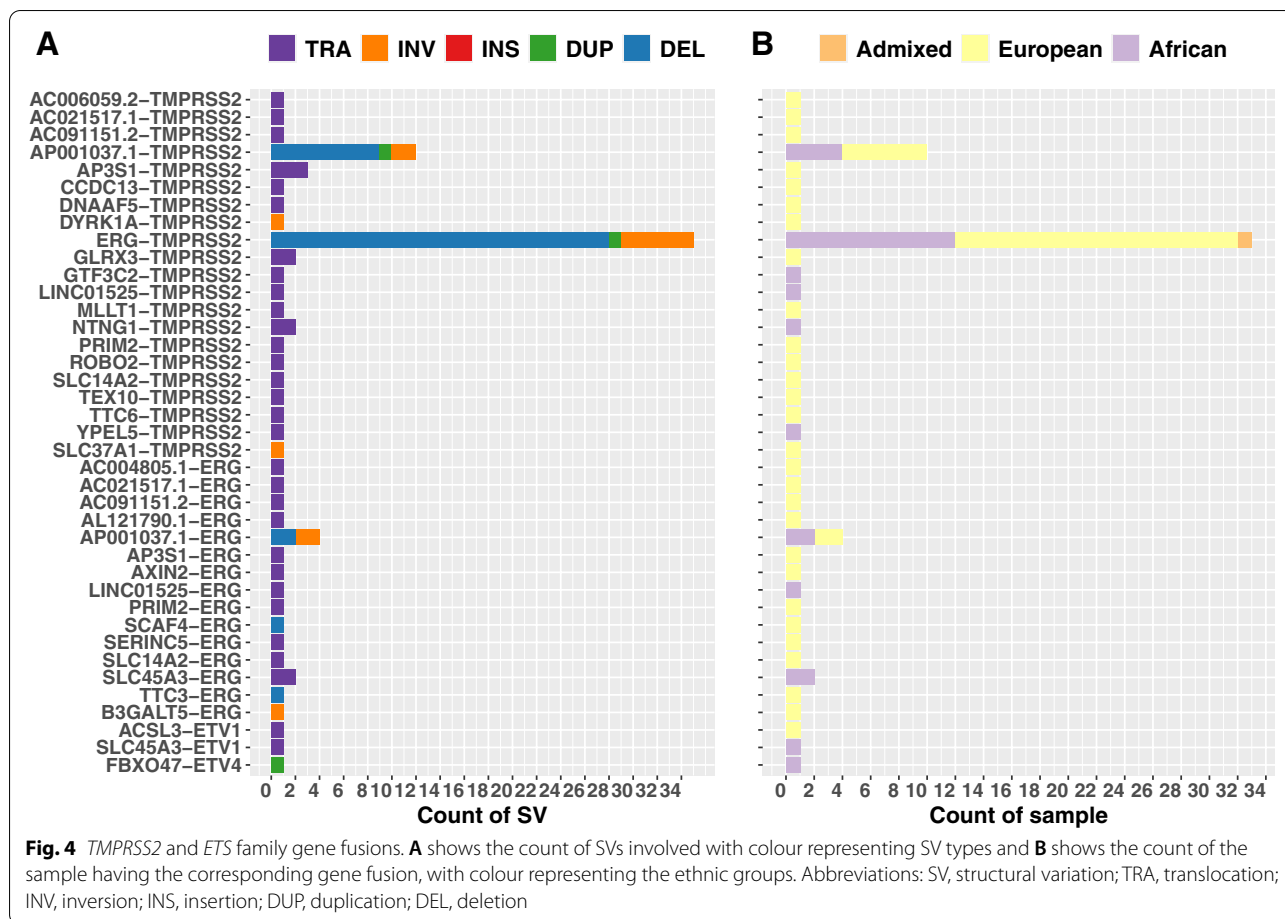
the single European-specific hotspot (chr13: 45–46 Mbp) was driven by a large number of SVs in a single tumour from a European patient. However, when considering patient count, we observed 6/8 tumours to be African-derived (Fig. 3C). Overall, 65/74 SV hotspots have > 50% breakpoints observed in >50% tumours of African ancestry. In hotspots chr4: 48–49 Mbp and chr4:175–176 Mbp, more than 90% of tumours are derived from African patients and significantly associated with African ancestry ( $p$ -value = 0.04 and 0.03 respectively by Chi-squared test).

Although *TEC* was found to be a candidate driver gene interrupted by 55% breakpoints in the African-dominant hotspot chr4: 48–49 Mbp, only 2/10 tumours have *TEC* disruption. Among the 22 candidate driver genes found

in SV hotspots, previously reported PCa-related gene *ERG* interruption at hotspot chr21: 38–39 Mbp (37/40 tumours) was biased towards European-derived tumours (21/37), representing 34.4% and 13.0% of tumours from European and African patients, respectively. Conversely, we found the previously identified PCa driver gene *LRP1B* (12/17) and new candidate genes *TTC28* (16/27), *CADM2* (12/16), *LSAMP* (11/16), *EYS* (16/18), *PTPRD* (12/18), *PACRG* (5/6) and *PDE4D* (12/17) to be predominantly interrupted in tumours from African patients.

#### Gene fusions: SV types and breakpoint clustering

Through investigation of gene regions impacted by SV BND pairs, we identified 6,617 gene fusions, in which 134 are recurrent in two genomes, 13 in three genomes and



6 in four or more genomes (Additional file 2: Table S4 and Table S5). 33 gene fusions were previously reported for PCa, including two of the top six recurrent gene fusions *ZBTB20-LSAMP* (4 tumours of which three are African-derived), and the well-established PCa fusion gene *TMPRSS2-ERG* (31 tumours) [34]. Among the novel gene fusions identified, 144 are recurrent (two or more) in tumours from African patients (Additional file 2: Table S4). The top novel African-associated recurrent gene fusions include *AC016822.1-PCDH15* (4/4), *AC098650.1-RBMS3* (3/3), *AC117473.1-EPHA6* (3/3), *AL513166.1-FPGT-TNNI3K* (3/3), *AL513166.1-TNNI3K* (3/3), *CASC19-PCAT1* (5/5), *DPYD-DPYD-AS1* (3/3), *PRKN-PACRG* (3/4) and *SATB1-TBC1D5* (3/3). Additionally, we observed 35 intra-chromosomal SVs within the African-derived *PRKN-PACRG* positive tumour N0081.

Taking a closer look at *TMPRSS2-ERG* and their alternative partners, as well as all previously reported PCa relevant *ETS* family members, namely *ETV1*, *ETV4* and *ETV5* [10], we identified besides *TMPRSS2-ERG*, 20 *TMPRSS2*, 15 *ERG*, 2 *ETV1* and 1 *ETV4* partners, largely

driven by translocations [35] and to a lesser extent by deletion [13], inversion [7] and duplication [2] events (Fig. 4A). While 23 were European-specific, 10 were African-derived (Fig. 4B). *SLC45A3-ERG*, identified in 2/115 (1.7%) tumours from African patients, after *TMPRSS2-ERG* is the second most common fusion event in *ERG* positive PCa tumours [35], while *SLC45A3-ETV1* fusion has also been reported [36]. In our tumours from African patients, we identified *FBXO47* as a novel PCa partner to *ETV4* and *LINC01525* to *ERG*. African-specific novel PCa partners to *TMPRSS2* include *GTF3C2*, *LINC01525*, *NTNG1* and *YPEL5*.

Further investigation of *TMPRSS2-ERG* fusion identified two additional tumours harbouring large (around 2.8 Mbp) deletion events, with BNDs 73,116 bp and 33,796 bp downstream of *ERG*, respectively. Of the 33 *TMPRSS2-ERG* gene fusion-positive patients, 20 are European (33% of 61), 12 African (10% of 115) and one of admixed African-European ancestry. The lower percentages observed across our study coincides with previously reported ethnic disparities [11, 12], as well as the increased presence of this fusion event reported for lower-grade tumours



[37]. Previously attributed to an interstitial deletion or an insertional chromosomal rearrangements [38], of the 33 tumours identified as *TMPRSS2-ERG* fusion-positive in this study, 16 are the result of a single deletion event, two present with a deletion and each additionally with two matching translocations (2 pairs of BNDs), indicating retention of the interstitial region (Additional file 1: Fig. S5), and 10 present with a deletion with additional overlapping SVs. Of the remaining tumours, two were the result of inversion events (European patients 15126 and 5902), one from a duplication (African patient TSH005), while a single tumour from a European (BRA10) and African (UP2103) patient involved four and seven overlapping SVs, respectively. For BRA10, the multi-event *TMPRSS2-ERG* fusions included: one deletion between *TTC3* and *ERG*, one inversion on *ERG*, one inversion between *ERG* and *TMPRSS2* and one inversion between *TMPRSS2* and *SLC37A1*. For UP2103: two translocations between *NTNG1* and *TMPRSS2*, one translocation between *LINC01525* and *ERG*, one translocation between *LINC01525* and *TMPRSS2*, one inversion between *ERG* and *TMPRSS2* and two inversions on *TMPRSS2*. It is unclear if the fusion is the result of the inversion or multiple SVs found in BRA10 and UP2103.

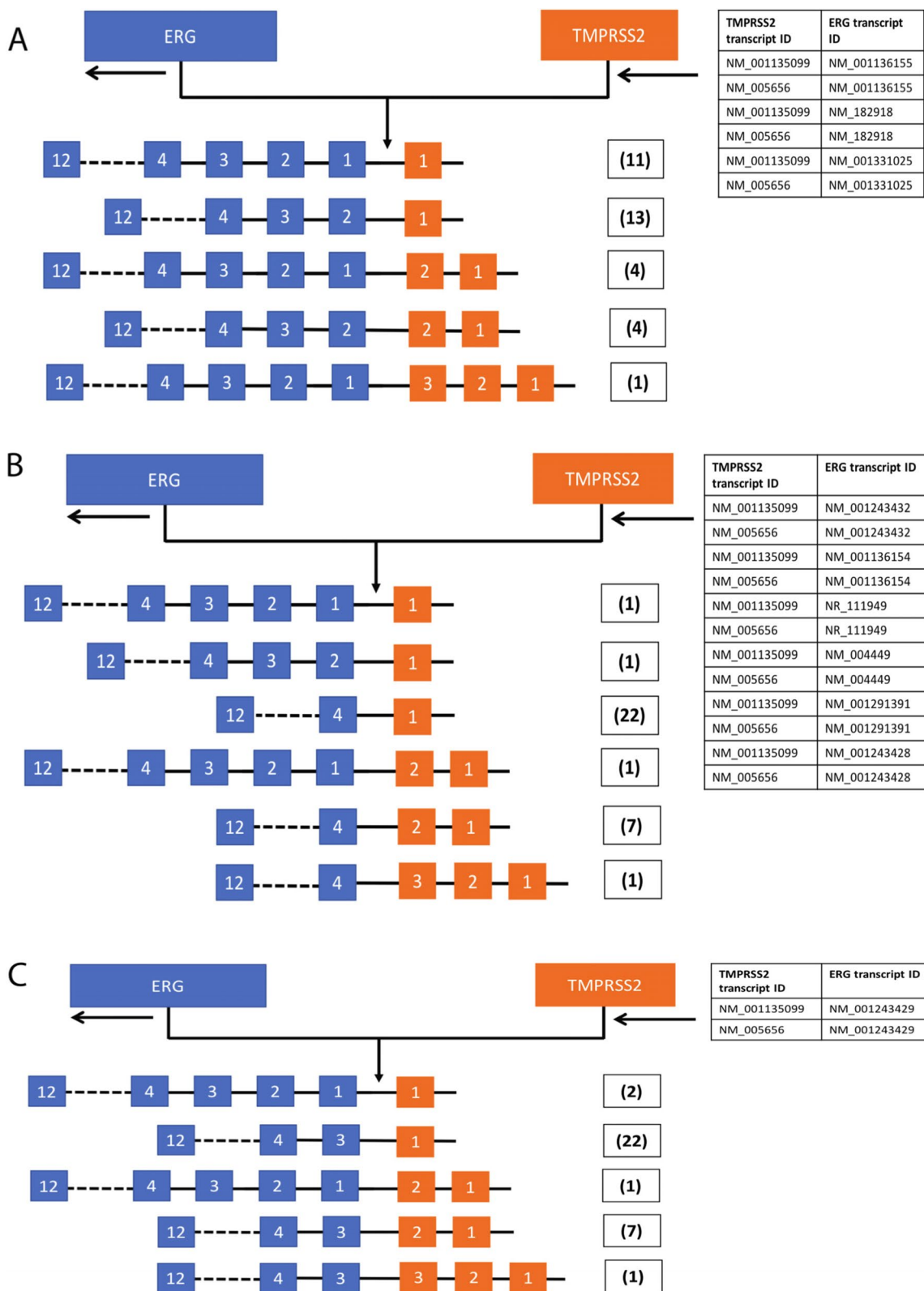
Investigating if SV breakpoints involved in *TMPRSS2-ERG* fusions cluster in any specific genomic position, we found breakpoints on *TMPRSS2* (2 transcripts) to be clustered predominantly 3' of exon 1 or 2, while breakpoints on *ERG* (10 transcripts) clustered 5' of exon 3 or 4 (Fig. 5). These observations are consistent with previous findings using RNA expression data [12, 38], while UP2103, UP2089 and UP2093 were previously included in a targeted RNA sequencing study aimed at defining the exact *TMPRSS2-ERG* fusion transcript junction coordinates [12]. The latter study detected three or four *TMPRSS2-ERG* fusion transcript junctions (isoforms) with different coordinates from all of the three samples, while we identified a single deletion in UP2089 and UP2093 and a single inversion in UP2103. Thus, these studies concur that a single genomic fusion can result in multiple fusion transcripts or isoforms.

## Discussion

To our knowledge, this is the first study to investigate the potential role of SVs in significant risk for aggressive PCa observed for men of African ancestry from sub-Saharan Africa, specifically southern Africa. Through direct ethnic-based comparative analysis using a single technical and informatic pipeline, we report a higher variability of somatic SVs for tumours from African (0–754) versus European (0–398) patients. Comparing the prevalence of SVs in types between tumours from African and European patients, duplication had the greatest difference

among the ethnic groups, showing a 2.5- and 1.6-fold increase in African versus European-derived tumours in its average count and relative frequency, respectively. Hyper-SV mutated tumours were largely restricted to African patients (12/15). While a study of 20 metastatic tumour types, including PCa, found stomach and oesophageal tumours to be highly enriched for translocations [39], this is the first report of a hyper-translocated PCa subtype. Additionally, this is the first study to identify hyper-duplicated and hyper-deleted non-treated primary tumours rather than mCRPC. While the metastatic status of the African patients is unknown, our study suggests that hyper-SV is likely an indicator of aggressive disease rather than a consequence of treatment response. Confirming a link between *CDK12*<sup>-/-</sup> and hyper-duplicated (4/5 tumours), we identify an additional association with *MYC* CN gain. While our study concurs with *BRCA2*<sup>-/-</sup> in hyper-deleted tumours (3/5), here we report further association with *SPOP* mutation.

SVs have previously been shown to be non-randomly distributed across cancer genomes [40], implying that SVs at certain loci may drive the expansion of some cancer clones. Applying an independent SV hotspot analysis approach, 74 hotspots were identified based on the number of breakpoints or recurrent tumours in each genomic window, revealing 18 new potential driver genes. Investigating the prevalence of SV hotspots in ethnic groups, hotspots chr4: 48–49 Mbp and chr4:175–176 Mbp were predominantly found in African-derived tumours. In addition, we found *ERG* as European-derived tumour-related driver and five new African-derived tumour-related driver genes, including *TTC28*, *CADM2*, *LSAMP*, *EYS*, *PACRG*, *PDE4D* and *PTPRD*. Large number of inter-chromosomal translocation inactivating *TTC28* has been reported in colorectal cancer [41]. *CADM2* acts as a tumour suppressor in renal cell carcinoma [42], prostate cancer [43] and hepatocellular carcinoma [44], and has been reported promoting brain metastasis in lung cancer and proposed as a potential molecular target [45]. Recurrent deletions of the *LSAMP* locus have been reported in tumours from African American men, identifying an African-specific aggressive PCa subset [46]. *PACRG* has previously been reported to be associated with poor prognosis of renal cell carcinoma [47]. The high expression of *PDE4D* has been reported to be associated with aggressive disease in multiple cancers, with therapeutic potential reported for pancreatic ductal adenocarcinoma [48], tamoxifen-resistant ER-positive breast cancer [49], lung cancer [50] and colon cancer [51]. In PCa, *PDE4D* has been implicated as proliferation-promoting factor and proposed as a biomarker and potential drug target [52, 53]. *PTPRD* was classified as a tumour suppressor gene, which has been reported to be highly mutated and



**Fig. 5** Spectrum of gene fusion junctions. **A–C** panels show the three forms of structural variation (SV) breakpoint clusters based on different transcripts of *TMPRSS2* and *ERG*, shown in the top right. The number of samples with breakpoint in different exon positions of *TMPRSS2-ERG* fusion junction is shown in brackets

correlated to the disease progression in colon [54] and gastric cancers [55] and found deleted in multiple types of cancers [56]. However, *PTPRD* has been reported as a significantly low-frequency mutated gene in PCa [57], indicating SV may be an alternate variant type activating *PTPRD* in African patients.

Investigation of gene fusions caused by SVs identified; the well-established PCa fusion gene *TMPRSS2-ERG* in 10% and 33% of tumours from African and European patients, respectively; the previously reported African-specific *ZBTB20-LSAMP*; and nine novel African-associated fusions. *LSAMP* was identified as a new potential driver gene in this study; other studies also reported a significantly higher number of inter-chromosomal rearrangements and exclusive association of *LSAMP* deletion/rearrangement for African American tumours, including *ZBTB20-LSAMP* gene fusion [46]. Investigating *TMPRSS2-ERG* and their alternative partners revealed six (out of 29) African-derived fusions with novel PCa partners to *ETV4*, *ERG* and *TMPRSS2*, including *FBXO7*, *LINC01525*, *GTF3C2* and *NTNG1*. Classified as a potential tumour suppressor gene, disruption of *FBXO47* has been reported in breast, ovarian and renal cancers [58, 59], while to the best of our knowledge this is the first report of *LINC01525* disruption in any cancer. *GTF3C2* is one member of the general transcription factor III family, which has been reported as a prognostic factor in liver cancer [60]. *NTNG1* belongs to the family of netrins, with an elevated level of *NTNG1* reported to result in cisplatin resistance in ovarian cancer [61]. *NTNG1* mutation has also been associated with poor prognosis in colorectal [62] and pancreatic cancer [63]. A study of more than 10,000 samples across more than 30 tumour types found *NTNG1* to have the highest mutation rate in the netrin family observing *NTNG1* fusion transcripts in multiple cancers, including breast, lung and skin [64]. Increased expression of *YPEL5*, coding for a member of the carboxy-terminal to LisH (CTLH) complex, has been reported in erlotinib-treated EGFR-mutant non-small cell lung cancer [65], while recurrent *YPEL5-PPP1CB* fusion has been reported for chronic lymphocytic leukaemia [66].

Deletion and translocation have been reported to cause 5' untranslated region of *TMPRSS2* to be fused with *ERG* on the same directions of the two genes, resulting in chimeric proteins [38]. With further investigation of *TMPRSS2-ERG* gene fusion-positive tumours, 17/33 of them harbour multiple SV events of different SV types in addition to deletion and translocation events. Furthermore, we found inversions with BND interrupting *TMPRSS2* and *ERG* genes in three tumours from European patients and one RNA-seq validated tumour from the African patient UP2103 [12]. The inversion resulted

in *TMPRSS2* and *ERG* genes fused in opposing coding strand directions, which may result in the formation of a chimeric transcript with a similar role to an antisense RNA [67]. Overall, our study demonstrates the complexity of SV events resulting in *TMPRSS2-ERG* fusion that cannot be attributed to simple DNA loss or translocation.

The use of short-read sequencing is a potential limitation of this study. We have used high-coverage WGS and employed the best-practise SV calling and filtering approach to achieve the balance of detection sensitivity and precision, but may have overlooked a fraction of real SVs, in particular the known to be 'difficult-to-detect' insertions and/or those present at the sub-clonal level [19]. Future studies using long-read sequencing may reveal a greater SV burden and additional hotspots. In addition, the discovered novel oncogenic drivers in this study have yet to be validated. The better prognosis and treatment for African PCa patients can benefit from further African-relevant validation and functional studies of the discovered hotspots and candidate drivers in this study.

## Conclusions

As a hallmark feature of its genome, SV is a major contributor to the development and progression of PCa by gene disruption and enabling genomic instability. Here we have described in a first-of-its-kind study the spectrum of simple SV types and SV-derived hotspots, including novel oncogenic drivers and gene fusions specific to African patients that may explain, at least in part, the observed disparity in PCa aggressiveness observed for men of African *versus* European ancestry. The identification of novel African-specific prognostic, including *PTPRD*, *LSAMP*, *PACRG*, *FBXO7*, *GTF3C2* and *NTNG1*, and therapeutic targets, including *CADM2*, *PDE4D* and *YPEL5*, emphasises the need for both African inclusion and SV interrogation to reduce advanced PCa ethnic disparity through tailored clinical management.

## Abbreviations

PCa: Prostate cancer; SV: Structural variation; TMB: Tumour mutational burden; SNV: Single nucleotide variants; indels: Small insertions and deletions; Mb: Megabase; ETS: E26 transformation-specific; DEL: Deletion; INS: Insertion; DUP: Duplication; INV: Inversion; TRA: Translocation; WGS: Whole genome sequencing; PSA: Prostate-specific antigen; BND: Breakend; CNV: Copy number variation; mCRPC: Metastatic castration-resistant prostate cancer.

## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13073-022-01096-w>.

**Additional file 1: Figure S1.** Concordant SV call generation from Manta and GRIDSS. **Figure S2.** Summary of SVs in each type, compared to other studies. **Figure S3.** CIRCOS plot of hyper-SV mutated tumours. **Figure S4.** The spread of SV breakpoints and samples in each 1 Mbp genomic

bin. **Figure S5.** *TMPRSS2-ERG* fusion with interstitial region retention.

**Table S1.** Clinical and pathological characteristics of 180 prostate cancer patients included in this study. **Table S2.** Biallelic assessment of *CDK12* in hyper-duplicated samples. **Table S3.** Biallelic assessment of *BRCA2* in hyper-deleted samples.

**Additional file 2: Table S4.** Summary of gene fusions identified from SVs. **Table S5.** SV calls resulting in gene fusions.

### Acknowledgements

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 Endocrine and Tumor Molecular Biology Laboratory (Brazil), who over many years have recruited patients and provided samples to these critical biosources.

### Authors' contributions

TG and VMH designed the study. TG performed formal analysis and WJ, JJ, CW, TC, RS and AP additional supportive data analysis. TG and VMH led the data interpretation. SMP, RJL, AH and DGP performed data curation. ISB, PDS, SBM and RMB recruited participants and collected the clinical data and samples. EKC and VMH supervised the study. TG and VMH wrote the manuscript. All authors read and approved the final manuscript.

### Funding

This research was supported by the Medical Health and Medical Research Council (NHMRC) of Australia through Project Grant APP1165762 (V.M. Hayes) and Ideas Grants APP2001098 (V.M. Hayes and M.S.R. Bornman) and APP2010551 (V.M. Hayes), University of Sydney Bridging Grant G199756 (V.M. Hayes), and partly through the USA. Department of Defense (DoD) Prostate Cancer Research Program (PCRP) Idea Development Award PC200390 (led by V.M. Hayes and including co-leads W. Jaratlerdsiri, S. Mutambirwa, M.S.R. Bornman). 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. Hayes, E.K.F. Chan and W. Jaratlerdsiri), the Intersect Computational Merit Allocation Scheme (V.M. Hayes), Intersect Australia Limited, and the Sydney Informatics Hub, Core Research Facility. 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, were supported by the Cancer Association of South Africa (CANSA) (M.S.R. Bornman and V.M. Hayes). V.M. Hayes was supported by the Petre Foundation via the University of Sydney Foundation, with additional support to W. Jaratlerdsiri and A.T. Paupenfuss provided by the Prostate Cancer Research Alliance Australian Government and Movember Foundation Collaboration PRECEPT (Prostate cancer prognosis and treatment study, led by N. Corcoran, University of Melbourne, Australia).

### Availability of data and materials

The datasets analysed in this study were obtained and accessible through Jaratlerdsiri et al [6], with sequence data deposited in the European Genome-Phenome Archive (EGA; <https://ega-archive.org>) under overarching accession EGAS00001006425 and including the Southern African Prostate Cancer Study (SAPCS) Dataset (EGAD00001009067) and Garvan/St Vincent's Prostate Cancer Database (EGAD00001009066). The computational code used to analyse SV subtypes, SV hotspots and gene fusions is available on GitHub [68].

### Declarations

#### Ethics approval and consent to participate

Irrespective of country of origin, all individuals provided informed consent to participate in the study. Conforming to the principles of the Helsinki Declaration, South African patients were recruited as part of the Southern African Prostate Cancer Study (SAPCS) with approval granted by the University of Pretoria Faculty of Health Sciences Research Ethics Committee (HREC, with US Federal wide assurance FWA00002567 and IRB00002235 IORG0001762; #43/2010). In Australia, participant recruitment was approved by the St

Vincent's HREC (#SVH/12/231) and in Brazil by the Grupo de Pesquisa e Pós-Graduação (GPPG) Scientific Committee and Research Ethical Commission (#20160539). Samples were shipped to the Garvan Institute of Medical Research in accordance with institutional Material Transfer Agreements (MTAs), as well as additional Republic of South Africa Department of Health Export Permit (National Health Act 2003; J1/2/4/2 #1/12). This study was approved by the St. Vincent's HREC (#SVH/15/227) for genomic interrogation.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

### Author details

<sup>1</sup>Ancestry and Health Genomics Laboratory, Charles Perkins Centre, School of Medical Sciences, Faculty of Medicine and Health, University of Sydney, Camperdown, NSW, Australia. <sup>2</sup>Genomics and Epigenetics Theme, Garvan Institute of Medical Research, Darlinghurst, NSW, Australia. <sup>3</sup>Human Phenome Institute, Fudan University, Shanghai, China. <sup>4</sup>Sydney Informatics Hub, University of Sydney, Sydney, NSW, Australia. <sup>5</sup>School of Health Systems and Public Health, University of Pretoria, Pretoria, South Africa. <sup>6</sup>Endocrine and Tumor Molecular Biology Laboratory, Instituto de Ciências Básicas da Saúde, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil. <sup>7</sup>Laboratory of Genetics, Instituto de Ciências Biológicas, Universidade Federal do Rio Grande, Rio Grande, Brazil. <sup>8</sup>Department of Urology, St. Vincent's Hospital, Darlinghurst, NSW, Australia. <sup>9</sup>Department of Urology, Sefako Makgatho Health Science University, Dr George Mukhari Academic Hospital, Medunsa, Ga-Rankuwa, South Africa. <sup>10</sup>Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia. <sup>11</sup>Department of Medical Biology, University of Melbourne, Melbourne, Victoria, Australia. <sup>12</sup>NSW Health Pathology, Sydney, Australia. <sup>13</sup>Faculty of Health Sciences, University of Limpopo, Turfloop Campus, Mankweng, South Africa.

Received: 12 April 2022 Accepted: 28 July 2022

Published online: 31 August 2022

### References

- Smith ZL, Eggener SE, Murphy AB. African-American Prostate Cancer Disparities. *Curr Urol Rep.* 2017;18(10):81.
- Taitt HE. Global Trends and Prostate Cancer: A Review of Incidence, Detection, and Mortality as Influenced by Race, Ethnicity, and Geographic Location. *Am J Mens Health.* 2018;12(6):1807–23.
- Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin.* 2021;71(3):209–49.
- Tindall EA, Monare LR, Petersen DC, van Zyl S, Hardie R-A, Segone AM, et al. Clinical presentation of prostate cancer in Black South Africans. *Prostate.* 2014;74(8):880–91.
- Jaratlerdsiri W, Chan EKF, Gong T, Petersen DC, Kalsbeek AMF, Venter PA, et al. Whole Genome Sequencing Reveals Elevated Tumor Mutational Burden and Initiating Driver Mutations in African Men with Treatment-Naïve, High-Risk Prostate Cancer. *Cancer Res.* 2018;78(24):6736–46. [canres.0254.2018](https://doi.org/10.1158/0008-5472.CCR-18-0254).
- Jaratlerdsiri W, Jiang J, Gong T, Patrick S, Willet C, Chew T, et al. African-specific prostate cancer molecular taxonomy. *Nature.* Accepted 28 July 2022.
- Baca SC, Prandi D, Lawrence MS, Mosquera JM, Romanel A, Drier Y, et al. Punctuated evolution of prostate cancer genomes. *Cell.* 2013;153(3):666–77.
- Ryan MJ, Bose R. Genomic alteration burden in advanced prostate cancer and therapeutic implications. *Front Oncol.* 2019;9:1287.
- Tomlins Scott A, Rhodes Daniel R, Perner S, Dhanasekaran Saravana M, Mehra R, Sun X-W, et al. Recurrent Fusion of *TMPRSS2* and *ETS* Transcription Factor Genes in Prostate Cancer. *Science.* 2005;310(5748):644–8.
- The Cancer Genome Atlas Research N. The molecular taxonomy of primary prostate cancer. *Cell.* 2015;163(4):1011–25.

11. Zhou CK, Young D, Yeboah ED, Coburn SB, Tettey Y, Biritwum RB, et al. TMPRSS2:ERG Gene Fusions in Prostate Cancer of West African Men and a Meta-Analysis of Racial Differences. *Am J Epidemiol*. 2017;186(12):1352–61.
12. Blackburn J, Vecchiarelli S, Heyer EE, Patrick SM, Lyons RJ, Jaratlersiri W, et al. TMPRSS2-ERG fusions linked to prostate cancer racial health disparities: A focus on Africa. *Prostate*. 2019;79(10):1191–6.
13. Raj A, Stephens M, Pritchard JK. fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics*. 2014;197:573–89.
14. Sadsad R, Samaha G, Chew T. Fastq-to-bam @ NCI-Gadi [Internet]. WorkflowHub. 2021 [Available from: <https://doi.org/10.48546/workflowhub.workflow.146.1>].
15. Favero F, Joshi T, Marquard AM, Birkbak NJ, Krzystanek M, Li Q, et al. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann Oncol*. 2015;26(1):64–70.
16. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Kallberg M, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*. 2016;32(8):1220–2.
17. Cameron DL, Schroder J, Penington JS, Do H, Molania R, Dobrovic A, et al. GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. *Genome Res*. 2017;27(12):2050–60.
18. Cameron DL, Baber J, Shale C, Valle-Inclan JE, Besselink N, van Hoeck A, et al. GRIDSS2: comprehensive characterisation of somatic structural variation using single breakend variants and structural variant phasing. *Genome Biol*. 2021;22(1):202.
19. Gong T, Hayes VM, Chan EKF. Detection of somatic structural variants from short-read next-generation sequencing data. *Brief Bioinform*. 2021;22(3):bbaa056.
20. Sadsad R, Samaha G, Chew T. Germline-ShortV @ NCI-Gadi [Internet] 2021 [Available from: <https://doi.org/10.48546/workflowhub.workflow.143.1>].
21. Sadsad R, Chew T. Somatic-ShortV @ NCI-Gadi [Internet] 2021 [Available from: <https://doi.org/10.48546/workflowhub.workflow.148.1>].
22. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38(16):e164–e.
23. Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput Biol*. 2016;12(4):e1004873.
24. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol*. 2011;12:R41.
25. Campbell PJ, Getz G, Korbel JO, Stuart JM, Jennings JL, Stein LD, et al. Pan-cancer analysis of whole genomes. *Nature*. 2020;578(7793):82–93.
26. Quigley DA, Dang HX, Zhao SG, Lloyd P, Aggarwal R, Alumkal JJ, et al. Genomic Hallmarks and Structural Variation in Metastatic Prostate Cancer. *Cell*. 2018;175(3):889.
27. Wedge DC, Gundem G, Mitchell T, Woodcock DJ, Martincorena I, Ghorri M, et al. Sequencing of prostate cancers identifies new cancer genes, routes of progression and drug targets. *Nat Genet*. 2018;50(5):682–92.
28. Fraser M, Sabelnykova VY, Yamaguchi TN, Heisler LE, Livingstone J, Huang V, et al. Genomic hallmarks of localized, non-indolent prostate cancer. *Nature*. 2017;541(7637):359–64.
29. van Dessel LF, van Riet J, Smits M, Zhu Y, Hamberg P, van der Heijden MS, et al. The genomic landscape of metastatic castration-resistant prostate cancer reveals multiple distinct genotypes with potential clinical impact. *Nat Commun*. 2019;10(1):5251.
30. Viswanathan SR, Ha G, Hoff AM, Wala JA, Carrot-Zhang J, Whelan CW, et al. Structural Alterations Driving Castration-Resistant Prostate Cancer Revealed by Linked-Read Genome Sequencing. *Cell*. 2018;174(2):433–47 e19.
31. Menghi F, Barthel FP, Yadav V, Tang M, Ji B, Tang Z, et al. The Tandem Duplicator Phenotype Is a Prevalent Genome-Wide Cancer Configuration Driven by Distinct Gene Mutations. *Cancer Cell*. 2018;34(2):197–210.e5.
32. Landrum MJ, Chitipiralla S, Brown GR, Chen C, Gu B, Hart J, et al. ClinVar: improvements to accessing data. *Nucleic Acids Res*. 2020;48(D1):D835–D44.
33. Kumar R, Nagpal G, Kumar V, Usmani SS, Agrawal P, Raghava GPS. HumCFS: a database of fragile sites in human chromosomes. *BMC Genomics*. 2019;19(9):985.
34. Wei T, Lu J, Ma T, Huang H, Kocher J-P, Wang L. Re-Evaluate Fusion Genes in Prostate Cancer. *Cancer Inform*. 2021;20:11769351211027592.
35. Esgueva R, Perner S, LaFargue CJ, Scheble V, Stephan C, Lein M, et al. Prevalence of TMPRSS2-ERG and SLC45A3-ERG gene fusions in a large prostatectomy cohort. *Mod Pathol*. 2010;23(4):539–46.
36. Rubin MA, Maher CA, Chinnaiyan AM. Common Gene Rearrangements in Prostate Cancer. *J Clin Oncol*. 2011;29(27):3659–68.
37. Farrell J, Young D, Chen Y, Cullen J, Rosner IL, Kagan J, et al. Predominance of ERG-negative high-grade prostate cancers in African American men. *Mol Clin Oncol*. 2014;2(6):982–6.
38. Murphy SJ, Kosari F, Karnes RJ, Nasir A, Johnson SH, Gaitatzes AG, et al. Retention of Interstitial Genes between *TMPPRSS2* and *ERG* Is Associated with Low-Risk Prostate Cancer. *Cancer Res*. 2017;77(22):6157.
39. Priestley P, Baber J, Lolkema MP, Steeghs N, de Bruijn E, Shale C, et al. Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature*. 2019;575(7781):210–6.
40. Lin Y-L, Gokcumen O. Fine-Scale Characterization of Genomic Structural Variation in the Human Genome Reveals Adaptive and Biomedically Relevant Hotspots. *Genome Biol Evol*. 2019;11(4):1136–51.
41. Yamagishi H, Kuroda H, Imai Y, Hiraishi H. Molecular pathogenesis of sporadic colorectal cancers. *Chinese J Cancer*. 2016;35(1):4.
42. He W, Li X, Xu S, Ai J, Gong Y, Gregg JL, et al. Aberrant methylation and loss of *CADM2* tumor suppressor expression is associated with human renal cell carcinoma tumor progression. *Biochem Biophys Res Commun*. 2013;435(4):526–32.
43. Chang G, Xu S, Dhir R, Chandran U, O'Keefe DS, Greenberg NM, et al. Hypoexpression and epigenetic regulation of candidate tumor suppressor gene *CADM-2* in human prostate cancer. *Clin Cancer Res*. 2010;16(22):5390–401.
44. Li D, Zhang Y, Zhang H, Zhan C, Li X, Ba T, et al. *CADM2*, as a new target of miR-10b, promotes tumor metastasis through FAK/AKT pathway in hepatocellular carcinoma. *J Exp Clin Cancer Res*. 2018;37(1):46.
45. Dai L, Zhao J, Yin J, Fu W, Chen G. Cell adhesion molecule 2 (*CADM2*) promotes brain metastasis by inducing epithelial-mesenchymal transition (EMT) in human non-small cell lung cancer. *Ann Transl Med*. 2020;8(7):465.
46. Petrovics G, Li H, Stümpel T, Tan S-H, Young D, Katta S, et al. A novel genomic alteration of *LSAMP* associates with aggressive prostate cancer in African American men. *eBioMedicine*. 2015;2(12):1957–64.
47. Yan L, Gong Y-Z, Shao M-N, Ruan G-T, Xie H-L, Liao X-W, et al. Distinct diagnostic and prognostic values of  $\gamma$ -aminobutyric acid type A receptor family genes in patients with colon adenocarcinoma. *Oncol Lett*. 2020;20(1):275–91.
48. Liu F, Ma J, Wang K, Li Z, Jiang Q, Chen H, et al. High expression of *PDE4D* correlates with poor prognosis and clinical progression in pancreatic ductal adenocarcinoma. *J Cancer*. 2019;10(25):6252–60.
49. Mishra RR, Belder N, Ansari SA, Kayhan M, Bal H, Raza U, et al. Reactivation of cAMP Pathway by *PDE4D* Inhibition Represents a Novel Druggable Axis for Overcoming Tamoxifen Resistance in ER-positive Breast Cancer. *Clin Cancer Res*. 2018;24(8):1987–2001.
50. Pullamsetti SS, Banat GA, Schmall A, Szibor M, Pomagruk D, Hänze J, et al. Phosphodiesterase-4 promotes proliferation and angiogenesis of lung cancer by crosstalk with HIF. *Oncogene*. 2013;32(9):1121–34.
51. Boyd A, Baskar G, Petty T, Keeton A, Piazza G, Richter W. cAMP-Phosphodiesterase *PDE4D* as a Target for Colon Cancer Therapy. *FASEB J*. 2017;31(S1):671.11.
52. Böttcher R, Dulla K, van Strijp D, Dits N, Verhoef EI, Baillie GS, et al. Human *PDE4D* isoform composition is deregulated in primary prostate cancer and indicative for disease progression and development of distant metastases. *Oncotarget*. 2016;7(43):70669–84.
53. Powers GL, Hammer KDP, Domenech M, Frantskevich K, Malinowski RL, Bushnan W, et al. Phosphodiesterase 4D inhibitors limit prostate cancer growth potential. *Mol Cancer Res*. 2015;13(1):149–60.
54. Funato K, Yamazumi Y, Oda T, Akiyama T. Tyrosine phosphatase *PTPRD* suppresses colon cancer cell migration in coordination with *CD44*. *Exp Ther Med*. 2011;2(3):457–63.
55. Bae WJ, Ahn JM, Byeon HE, Kim S, Lee D. *PTPRD*-inactivation-induced *CXCL8* promotes angiogenesis and metastasis in gastric cancer and is inhibited by metformin. *J Exp Clin Cancer Res*. 2019;38(1):484.

56. Zhao S, Sedwick D, Wang Z. Genetic alterations of protein tyrosine phosphatases in human cancers. *Oncogene*. 2015;34(30):3885–94.
57. Nunes-Xavier CE, Mingo J, López JJ, Pulido R. The role of protein tyrosine phosphatases in prostate cancer biology. *Biochim Biophys Acta Mol Cell Res*. 2019;1866(1):102–13.
58. Simon-Kayser B, Scoul C, Renaudin K, Jezequel P, Bouchot O, Rigaud J, et al. Molecular cloning and characterization of FBXO47, a novel gene containing an F-box domain, located in the 17q12 band deleted in papillary renal cell carcinoma. *Genes Chromosomes Cancer*. 2005;43(1):83–94.
59. Zheng S, Fu Y. Age-related copy number variations and expression levels of F-box protein FBXL20 predict ovarian cancer prognosis. *Transl Oncol*. 2020;13(12):100863.
60. Song Y-Z, Li X, Li W, Wang Z, Li K, Xie F-L, et al. Integrated genomic analysis for prediction of survival for patients with liver cancer using The Cancer Genome Atlas. *World J Gastroenterol*. 2018;24(28):3145–54.
61. Fang S, Luo Y, Zhang Y, Wang H, Liu Q, Li X, et al. NTNG1 Modulates Cisplatin Resistance in Epithelial Ovarian Cancer Cells via the GAS6/AXL/Akt Pathway. *Front Cell. Dev Biol*. 2021;9:652325.
62. Sho S, Court CM, Winograd P, Russell MM, Tomlinson JS. A prognostic mutation panel for predicting cancer recurrence in stages II and III colorectal cancer. *J Surg Oncol*. 2017;116(8):996–1004.
63. Francescone R, Barbosa Vendramini-Costa D, Franco-Barraza J, Wagner J, Muir A, Lau AN, et al. Netrin G1 Promotes Pancreatic Tumorigenesis through Cancer-Associated Fibroblast-Driven Nutritional Support and Immunosuppression. *Cancer Discovery*. 2021;11(2):446.
64. Hao W, Yu M, Lin J, Liu B, Xing H, Yang J, et al. The pan-cancer landscape of netrin family reveals potential oncogenic biomarkers. *Sci Rep*. 2020;10(1):5224.
65. Wu X. Up-regulation of YPEL1 and YPEL5 and down-regulation of ITGA2 in erlotinib-treated EGFR-mutant non-small cell lung cancer: A bioinformatic analysis. *Gene*. 2018;643:74–82.
66. Velusamy T, Palanisamy N, Kalyana-Sundaram S, Sahasrabudde AA, Maher CA, Robinson DR, et al. Recurrent reciprocal RNA chimera involving *YPEL5* and *PPP1CB* in chronic lymphocytic leukemia. *Proc Natl Acad Sci*. 2013;110(8):3035.
67. Sugimoto T, Tomita A, Abe A, Iriyama C, Kiyoi H, Naoe T. Chimeric anti-sense RNA derived from chromosomal translocation modulates target gene expression. *Haematologica*. 2012;97(8):1278–80.
68. Gong T. *StructuralVariantUtil*. GitHub. 2022. <https://github.com/tgong1/StructuralVariantUtil>.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

