scientific reports

OPEN

Coding and regulatory somatic profiling of triple-negative breast cancer in Sub-Saharan African patients

Ricardo J. Pinto[1,2,3], Dylan Ferreira[4], Paulo Salamanca[5], Fernando Miguel[5], Pamela Borges[6], Carla Barbosa[6], Vitor Costa[6], Carlos Lopes[7], Lúcio Lara Santos[4,8,9,10] & Luisa Pereira[1,2]✉

The burden of triple-negative breast cancer (TNBC) may be shaped by genetic factors, particularly inherited and somatic mutation profiles. However, data on this topic remain limited, especially for the African continent, where a higher TNBC incidence is observed. In the age of precision medicine, cataloguing TNBC diversity in African patients becomes imperative. We performed whole exome sequencing, including untranslated regions, on 30 samples from Angola and Cape Verde, which allowed to ascertain on potential regulatory mutations in TNBC for the first time. A high somatic burden was observed for the African cohort, with 86% of variants being so far unreported. Recurring to predictive functional algorithms, 17% of the somatic single nucleotide variants were predicted to be deleterious at the protein level, and 20% overlapped with candidate cis-regulatory elements controlling gene expression. Several of these somatic functionally-impactful mutations and copy number variation (mainly in 1q, 8q, 6 and 10p) occur in known BC- and all cancer-driver genes, enriched for several cancer mechanisms, including response to radiation and related DNA repair mechanisms. *TP53* is the top of these known BC-driver genes, but our results identified possible novel TNBC driver genes that may play a main role in the African context, as *TTN*, *CEACAM7*, *DEFB132*, *COPZ2* and *GAS1*. These findings emphasize the need to expand cancer omics screenings across the African continent, the region of the globe with highest genomic diversity, accelerating the discovery of new somatic mutations and cancer-related pathways.

Keywords Triple-negative breast cancer, Sub-Saharan African ancestry, Exome and UTRs sequencing, Coding and regulatory somatic profile

Female breast cancer (BC) presently stands as one of the most pervasive malignancies globally, after lung cancer, with 2.31 million newly reported cases in 2022[1]. Moreover, it occupies the fourth position in cancer-related mortality, contributing to 6.9% (approximately 665,684 deaths) of the cancer fatalities recorded in the same year[1]. Despite this alarming scenario, data suggests that the global incidence and mortality trends of the disease have been stabilizing in high-income (HIC) countries over the past decades, where rates have historically been high[1]. Conversely, the opposite trend is currently observed in certain low- and middle-income (LIC and MIC) regions, such as Africa, where incidence rates are rapidly increasing, and estimates point to a doubling figure by 2050[2,3]. Increasing life-span, the adoption of more Westernized lifestyles, involving changes in reproductive factors (such as delayed childbearing, reduced overall parity, and consequently, decreased breastfeeding time), dietary habits, and body fat composition, are closing the gap in BC burden on an international scale[4]. In Africa, BC typically assumes the position of the primary or secondary malignancy in terms of incidence in most

[1]i3S, Instituto de Investigação e Inovação Em Saúde, Universidade do Porto, Porto, Portugal. [2]IPATIMUP, Instituto de Patologia e Imunologia Molecular da Universidade do Porto, Porto, Portugal. [3]ICBAS, Instituto de Ciências Biomédicas Abel Salazar, Universidade do Porto, Porto, Portugal. [4]Research Center of IPO-Porto (CI-IPOP) / RISE@CI-IPOP (Health Research Network), Portuguese Oncology Institute of Porto (IPO-Porto) / Porto Comprehensive Cancer Center (P.CCC) Raquel Seruca, Porto, Portugal. [5]Angolan Institute Against Cancer, Luanda, Angola. [6]Hospital Universitário Agostinho Neto, Praia, Cabo Verde. [7]Unilabs | Laboratório Anatomia Patológica, Porto, Portugal. [8]FP-I3ID, University Fernando Pessoa, Porto, Portugal. [9]Department of Surgical Oncology, Portuguese Oncology Institute of Porto, Porto, Portugal. [10]School of Medicine and Biomedical Sciences, University Fernando Pessoa, Gondomar, Portugal. ✉email: luisap@i3s.up.pt

countries, albeit there is notable diversity in age-adjusted incidence rates (per 100,000) across various African regions, ranging from 53.2 to 46.2 in Northern and Southern Africa, respectively, to 42.1 reported in Western Africa, and 31.9 or 26.7 in Eastern and Middle Africa, respectively[1,5]. While these figures remain 43% lower than those recorded in HIC, they should serve as a cause for concern among African health policymakers, particularly due to the elevated mortality rates associated with BC that women in these countries experience—a 35% increase compared to HICs—ranking them as the highest in the world[1]. The low survival rates of BC recorded in Africa may be attributed to multiple factors, with a significant contribution from late-stage disease at the time of diagnosis[6,7]. Additionally, the absence of organized population-based screening programs and diagnostic services, limited access to effective treatments, lack of population awareness, and unique inherent biological factors may collectively contribute to this issue[8,9].

Despite the huge gap in genomic information for BC (and cancer in general) between ethnicities[10,11], existing evidence supports the influence of ethnicity-specific determinants on the prevalence of distinct BC profiles at both intercontinental and African regional levels[12]. These studies are still largely based on comparisons involving African American (AA) and European American (EA) BC patients[10,13,14], but data begins to be collected for a few African countries, namely Nigeria[15,16]. Importantly, BC high burden in individuals of African ancestry is associated to a younger age at presentation, an increased mortality risk even after adjusting for socioeconomic status, and poor clinical outcomes[17,18]. This aggressive BC trend might potentially be correlated with the heightened incidence of triple-negative breast cancer (TNBC) in women of African descent[17]. In fact, BC is a highly heterogenous disease, clinically and genetically, and hence influencing treatment approaches. The World Health Organization (WHO) has outlined clinical criteria for disease diagnosing and prognosing, including histological type, tumour grade, stage, and the assessment of the immunohistochemistry (IHC) surrogate markers of oestrogen (ER) and progesterone (PR) receptors, and the amplification of the erb-b2 receptor tyrosine kinase 2 gene (*HER2*)[19]. The IHC stratification forms the basis for categorizing BC into five molecular subtypes: Luminal A-like, Luminal B-like (HER2-negative), Luminal B-like (HER2-positive), HER2-positive, and TNBC[20]. TNBC is characterized by the absence of hormone receptors expression and HER2 amplification, thereby constraining the applicability of endocrine therapy or targeted anti-HER2 drugs, such as tamoxifen, in treating these patients[21,22]. Furthermore, this BC subtype typically manifests an early onset pattern, marked by aggressive disease progression, frequent recurrence events, and a tendency to preferentially metastasize to the brain, liver, and lungs, features that jointly contribute to its adverse prognosis and low survival rates[23,24].

TNBC incidence is disproportionately higher in women of African ancestry, with accumulating evidence indicating a twofold increase in population-based incidence rates of TNBC in AA women compared to EA women[25–28]. Moreover, AAs diagnosed with TNBC face a higher likelihood of rapid relapse and exhibit shorter overall survival times than EA patients[29]. North America experienced significant influence from the trans-Atlantic slave trade originating in Western Sub-Saharan Africa, admixed with around 14%-21% European (EUR) and 1%-3% Native American ancestry[30]. So AA can be a proxy for Western Africans (WAfr), but not for the entire highly diverse African continent, so extrapolation of results from molecular or omics analyses conducted in AAs to the broader African context should be approached with caution[14]. Even the prevalence of TNBC is heterogeneous in the African continent, being higher among WAfr patients (53.2%) and AAs (29.8%), compared to EA (15.5%) and Eastern African (EAfr) patients (15.0%)[31]. Others have reported a similar trend, wherein a gradient of TNBC incidence is observed, ranging from the highest levels in Western Sub-Saharan Africa to the lowest ones in EAfr, Southern Africa, and Northern Africa[14,32,33].

Evidence suggests that the burden of TNBC may be influenced by ethnic factors, specifically ancestry-specific genomic and mutation profiles, although data in this field are limited in Africa[24]. In the era of precision medicine, propelled by advancements in sequencing technologies, particularly next-generation sequencing (NGS), harnessing the richness of African genetic diversity becomes pivotal. This endeavour is essential for unravelling the potential biological mechanisms driving TNBC pathogenesis and identifying innovative avenues for therapeutic intervention. Yet, the integration of such research initiatives in the African continent faces numerous challenges, ranging from a shortage of human, technological and technical resources to funding availability[34]. To disentangle the mutational landscape of TNBC patients across different regions of Africa, we performed a whole exome sequencing (WES) analysis on samples from Angola and Cape Verde. Angola is located in Southern Africa, whose population is largely of Bantu origin except for a few Khoisan groups in the frontier with Namibia. Individuals from the capital have an admixture of around 10% EA (own unpublished population data). Cape Verde is an archipelago in the Senegalese coast, with a highly admixed population between West Africans and EA of Mediterranean origin (in mean 50–50%[35,36]). The WES of these two African populations were compared with AA and EA TNBC samples from The Cancer Genome Atlas (TCGA).

When we designed our study, we decided to include the sequencing of the UTR regions in the WES library, as these regions are rich in regulatory sequences which have been overlooked in most studies. Algorithms to estimate the potential functional impact of the regulatory mutations are not so well developed as the ones for the protein-coding mutations, but begin to provide essential information, namely the ENCODE consortium[37,38].

Results

Cohorts' characterization

The TNBC cases included in the African cohort ($n = 30$) were diagnosed at a median age of 50 (mean age = 49.3), while for TCGA, European samples displayed a younger median age at diagnosis compared to African counterparts (51 vs. 57 years). Analysing the distribution per class of ages at diagnosis (Supplementary Fig. S1), it was possible to confirm that the African cohort had a higher proportion (23.3%) of samples younger than 40 years compared to the TCGA-AA (0%) and TCGA-EA (13.8%). The vast majority of African cases (83.3%) presented with invasive ductal carcinomas, as would be expected, and 82.1% of stage-classified cases were advanced (stages III + IV) (Supplementary Table S1). Regarding TCGA, the AA cohort presented more advanced disease stages than EA (20.8% vs. 14.5%; stages III + IV).

The somatic mutational TNBC profiles in African, TCGA-AA and TCGA-EA cohorts

In order to ascertain if the calling used in this work leads to identical results to the TCGA consortium, we analysed the African, TCGA-AA and TCGA-EA cohorts with the same workflow. Since TCGA exome enrichment library does not contain untranslated regions (UTRs), we had to restrict this initial analysis to the common exonic regions covered by both libraries, used here and by TCGA consortium. Basically, we ended up with the TCGA's library minus 12%. With our pipeline, among the 30 African TNBC samples analysed, a total of 7,372 somatic mutations were identified, compared with 1,152 and 7,402 found in TCGA-AA (*n* = 24) and TCGA-EA (*n* = 65), respectively. On the other hand, TCGA inferred 1,431 and 9,253 for AA and EA cohorts, respectively. The comparable contents in somatic mutations inferred by us and TCGA is reassuring. Even so, given that the African cohort is less than half the size of TCGA-EA cohort, it seems that the African cohort has a more diverse somatic profile than the TCGA-EA cohort. Of note, the content of somatic mutations in TCGA-AA cohort is extremely low compared with the other two, even correcting for the sample size.

The distributions of the somatic mutations (Fig. 1a) show a clear dominance of two classes, missense and intronic, followed by a third one of synonymous mutations. The ratio of missense mutations is 1.7 to 2.8 times higher than the synonymous ones. Other mutations, such as frameshift, nonsense and in-frame, are comparably rare. An amount of around 86% (n = 6,324) of the somatic mutations in the African cohort were not yet reported in the COSMIC database, demonstrating the potential to discover new somatic mutations when screening more African cancer patient samples.

The somatic mutations were located in 5,470, 1,062 and 5,465 genes in the African, TCGA-AA and TCGA-EA cohorts, respectively (Fig. 1b). A total of 216 somatically mutated genes were common to all groups, and are significantly enriched for essential molecular pathways: anatomical structure development; response to stimulus;

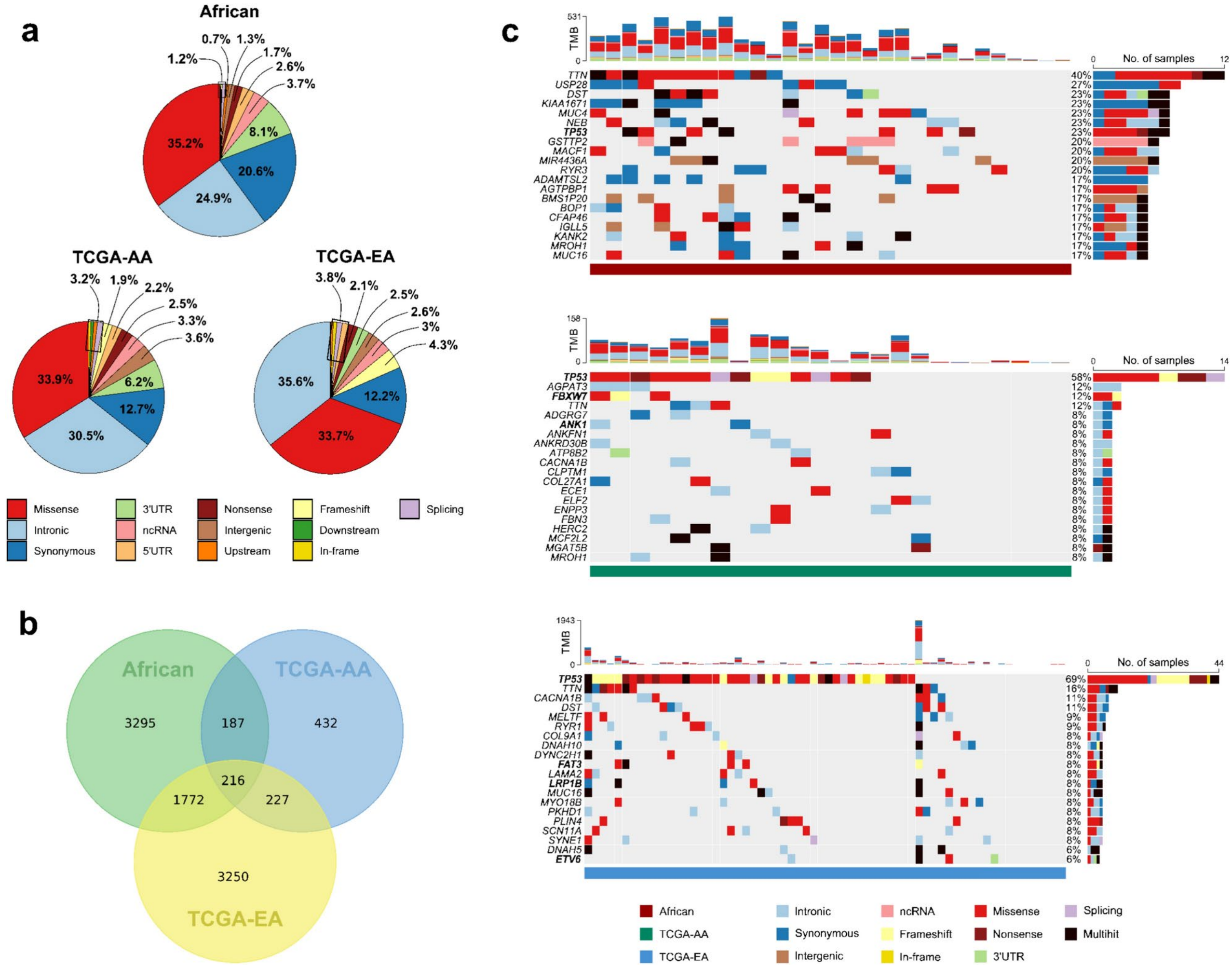

Fig. 1. The somatic landscape of the TNBC cohorts, when restricting to common exonic library coverage. (**a**) Frequency of the different categories of identified somatic variants. (**b**) Venn diagram for the somatically mutated genes in the three cohorts. (**c**) Oncoplots for top-20 most frequently mutated genes in each cohort (African, TCGA-AA and TCGA-EA); genes in bold are recognized cancer driver genes for all types of cancers.

cell motility; cell communication; cellular localization. Eight of these 216 shared genes are BC driver genes (*ARID1B*, *BIRC6*, *FAT1*, *FBXW7*, *MTOR*, *NF1*, *NOTCH2*, *TP53*) and 13 cancer driver genes associated with other tumour types (*ANK1*, *ANKRD11*, *ATRX*, *DDB2*, *ELL*, *HSP90AB1*, *NBEA*, *NSD1*, *PDGFRA*, *RET*, *ROS1*, *SMARCA4*, *TCF7L2*) (Supplementary Table S2). Besides these, in total, shared and non-shared, the African cohort displayed 99 somatic mutations in 57 BC driver genes (almost 50% of the 117 BC driver genes included in the IntOGen database). Among these, but always with very low frequencies, are *BRCA1*, *BRCA2*, *EGFR* and *PIK3CA*.

In the oncoplots (Fig. 1c), the TCGA cohorts, for both ethnicities, have a high frequency of somatic mutations in the *TP53* gene (58 and 69% for TCGA-AA and TCGA-EA, respectively), while this gene frequency was lower (23%) in the African cohort. Other published TNBC African datasets reported varying frequencies for *TP53* mutations: one Nigerian dataset[15] reported ~ 60%, while a combined Barbadian and Nigerian dataset[16] reported 32%. Inversely, *TTN* was the top gene in the African cohort, more than double the frequency observed in the TCGA cohorts. This gene encodes for titin, one of the largest known human proteins (~ 3 MDa) and a key structural component of striated muscle, particularly in the sarcomere[39].

The MutSig2CV analysis on the African cohort highlighted the genes *TP53*, *CNDP1*, *PRIM2, CEACAM7* and *COL6A2,* which had mutation frequencies higher than what could be attributed to random chance (adjusted *P*-value; $Q < 0.05$). When doing this analysis with both African and TCGA-AA cohorts together, to increase sample size, two other genes become significant, *DEFB132* and *COPZ2*. Only *TP53* is a recognized cancer driver gene, and this analysis reinforces the main role of this tumour suppressor in TNBC, but genes *CEACAM7, DEFB132* and *COPZ2* have already been shown to be associated with several cancers, playing a role in signalling, tumour immunity and tumour suppression[40–42].

We assessed the African and TCGA cohorts in terms of distribution of copy number variants (CNVs) along the exomes (Fig. 2). The overall CNV profiles (predominantly gains) of the three cohorts overlapped substantially in the long arms of chromosomes 1 and 8, and in the beginning of the short arm of chromosome 10. The African and TCGA-EA cohorts also shared CNV gains for parts of chromosome 6. The proportion of individuals per cohort presenting these CNV gains was around 20–25%. Some of these regions displaying CNV gains code for genes that have been identified as BC driver genes: *ABL2*, *MDM4*, *NTRK1*, *RGS7*, *CSMD3*, *FAM135B*, *PLAG1*, *UBR5* and *FGFR1*. Thus, the gain in copies of these genes can be important somatic hits in the tumorigenesis process.

Inferring the potential functional impact of protein-coding and regulatory somatic mutations in the African TNBC cohort

Considering the entire library used in the African cohort, enriched for UTRs, the total amount of somatic mutations in this cohort was 12,833. Some of these mutations will be passenger, neutral and possible artifacts. In order to focus attention on somatic mutations that may have a functional impact, we applied predictors of function for both protein-coding and regulatory functions.

The Combined Annotation Dependent Depletion (CADD) algorithm is a robust predictor of deleteriousness of protein-coding SNV mutations, with values equal of over 20 indicating potentially deleterious mutations. Clearly, the missense mutations dominate the potentially deleterious mutations (Fig. 3a), and they represent 59% of the total missense mutations of the African cohort. Almost all non-sense (96%) and splicing (87%) somatic mutations are potentially deleterious. On the other end, all the mutations in the other classes are almost non-deleterious. In summary, around 16% of the somatic SNV mutations in the African cohorts are potential causal variants, and when restricting these mutations to novel ones (not observed in COSMIC) our cohort has a significant higher proportion of potential causal variants (13%).

When focusing only in the genes that have somatic mutations with CADD indices for potential deleteriousness (Fig. 3c), the *TP53* gene is on the top of the African oncoplot, indicating that despite the lower frequency of mutations in this gene in this cohort, almost all these mutations are potentially deleterious (75%). The *TTN* gene shares a leading position with *TP53*, showing that also several of the mutations (59%) in this long gene have potential functional impact.

In order to estimate the functional impact of regulatory somatic mutations, we verified all the somatic mutations (WES + UTR) in the African cohort against the ENCODE data on candidate cis-regulatory elements (cCREs) (Fig. 3b). A total of 2,628 somatic mutations in the African cohort overlap with cCREs, representing 20% of the somatic mutation profile; focusing only in the UTRs. About 30% of the potentially regulatory somatic mutations are located in UTRs versus 70% in the other regions, supporting a more general inclusion of UTRs in cancer exonic screenings. As expected, proximal regions of the genes, like "upstream" and "5'UTR" are enriched for promoter like signatures (PLSs), while the "3'UTR" is enriched for distal enhancer like signatures (dELS). The "intronic" and the coding regions have a high number of cCREs, mostly enhancer like signatures, both proximal and distal.

The oncoplot of these somatic mutations overlapping cCREs (Fig. 3d) highlights genes involved in immune response (as *IGLL5*, *C7* and *DEFB132*), signalling (*CDS2* and *PITPNB*), cell cycle (*KIFC2*, *CDK11B* and *GAS1*; the later a putative tumour suppressor gene[43]) and protein deubiquitination (*USP31*). Considering all genes containing somatic mutations overlapping cCREs and shared by at least two samples for enrichment analysis, results indicate Wnt signalling pathway.

Inferring the putative driver mutations in the African cohort

We next checked if the putative impactful coding and regulatory somatic mutations, as well as the CNVs, in the African cohort occur in "BC driver genes" or in "all cancer driver genes" (from the IntOGen database). That information allows to infer the putative map of hits occurring in the cancer transformation of these patients.

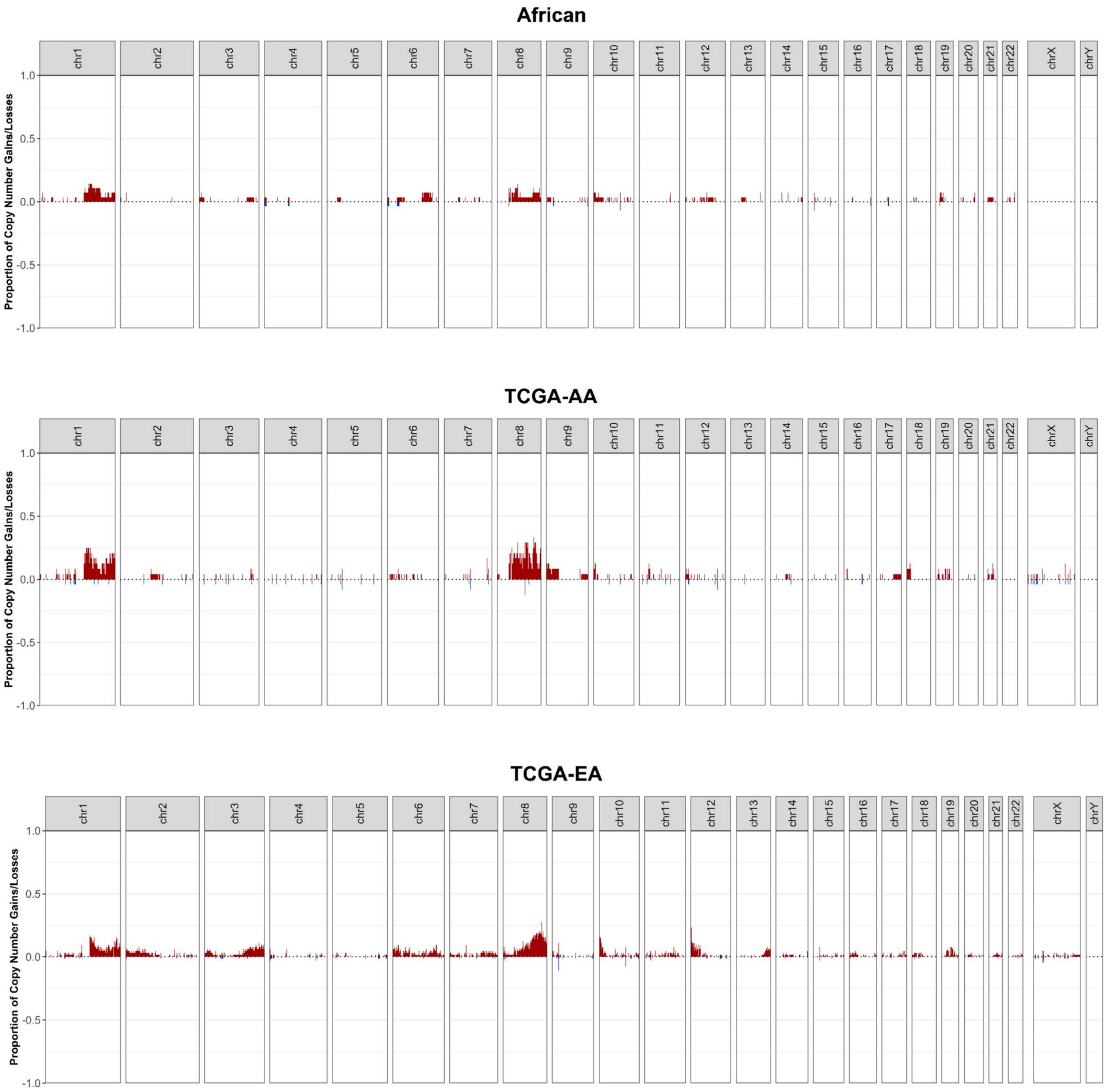

Fig. 2. An overview of CNV calls for each cohort of TNBC cases, showing the proportion or frequency of CNV gains or losses across the sequenced regions.

For the "BC driver genes" (Fig. 4), 25 out of the 30 samples have at least a hit in 1 to 14 "BC driver genes". The impactful coding and regulatory somatic mutations are located in 42 "BC driver genes", that are involved in key pathways such as kinase activity, transcription regulation, mammary gland development, response to radiation, and DNA repair complex (Supplementary Table S3). CNVs occur in 42 "BC driver genes", involved in similar pathways to the previous ones, except for the signature for response to radiation and DNA repair complex (Supplementary Table S4).

For the "all cancer driver genes" (Supplementary Fig. S2), 27 out of 30 samples have at least one hit in 1 to 49 "all cancer driver genes". The impactful coding and regulatory somatic mutations are located in 170 "all cancer driver genes", while CNVs occur in 185 "all cancer driver genes". These genes reinforce the signature on key pathways as when only considering the BC driver genes, but now also highlighting immune-related pathways, by mutations in driver genes usually associated with leukaemia and lymphomas, (Supplementary Tables S5 and S6).

Conclusions

The omics characterization conducted in this work, on Angola and Cape Verde patients, enriched significantly the catalogue for TNBC characterization in the Sub-Saharan African region, which was previously limited to Nigeria, Kenya, Ghana and Ethiopia[15,44,45]. The overall profile of the somatic mutations in the African cohort

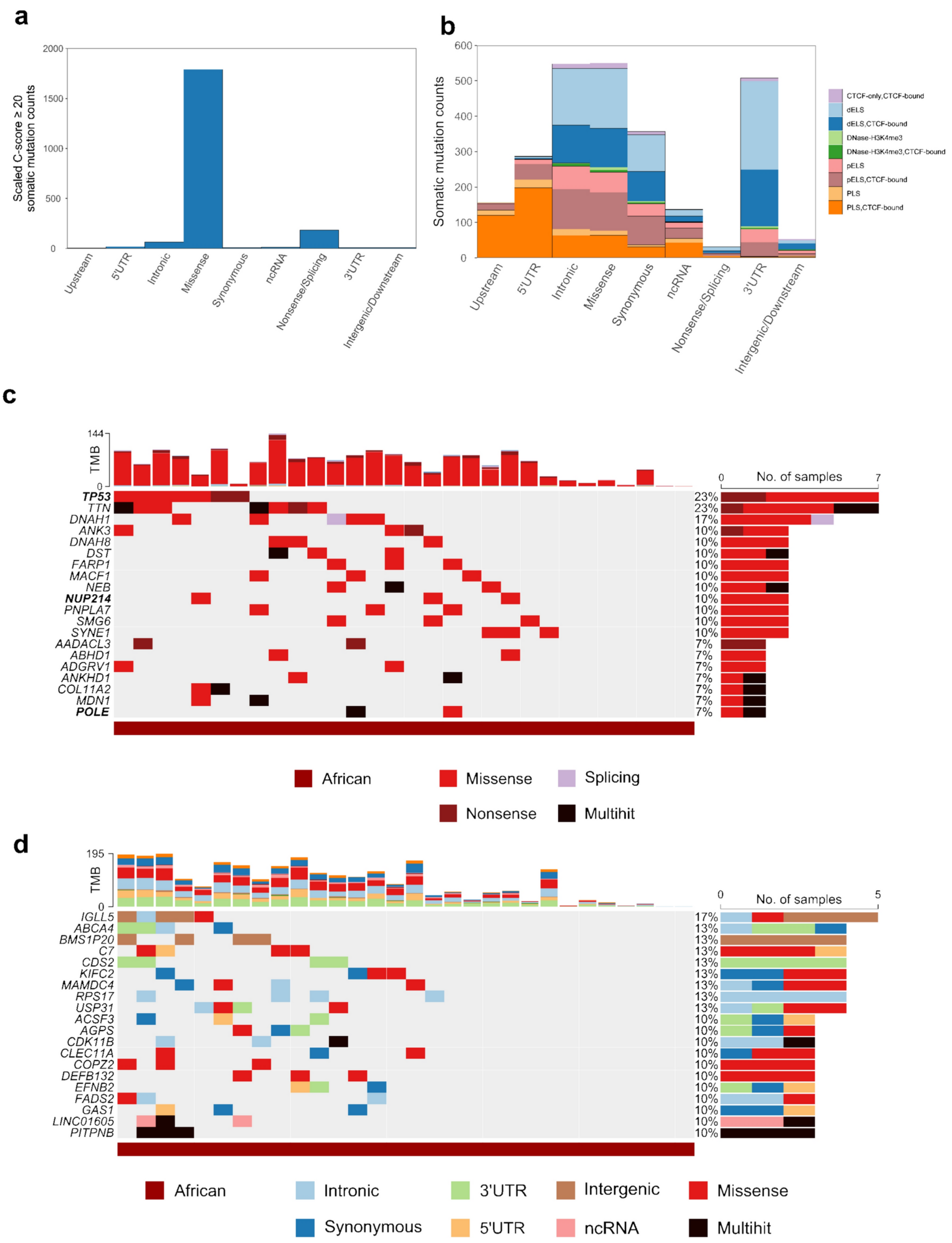

Fig. 3. Potential functional impact of protein-coding (recurring to CADD algorithm) and regulatory (recurring to ENCODE information) somatic mutations in the African TNBC cohort. (**a**) Number of somatic mutations with CADD ≥ 20 per genomic region (ANNOVAR annotation). (**b**) Number of somatic mutations overlapping ENCODE cCREs per genomic region (ANNOVAR annotation). (**c**) Oncoplot for genes where somatic mutations with CADD ≥ 20 are located; genes in bold are recognized cancer driver genes for all types of cancers. (**d**) Oncoplot for genes where somatic mutations overlapping ENCODE cCREs are located; genes in bold are recognized cancer driver genes for all types of cancers.

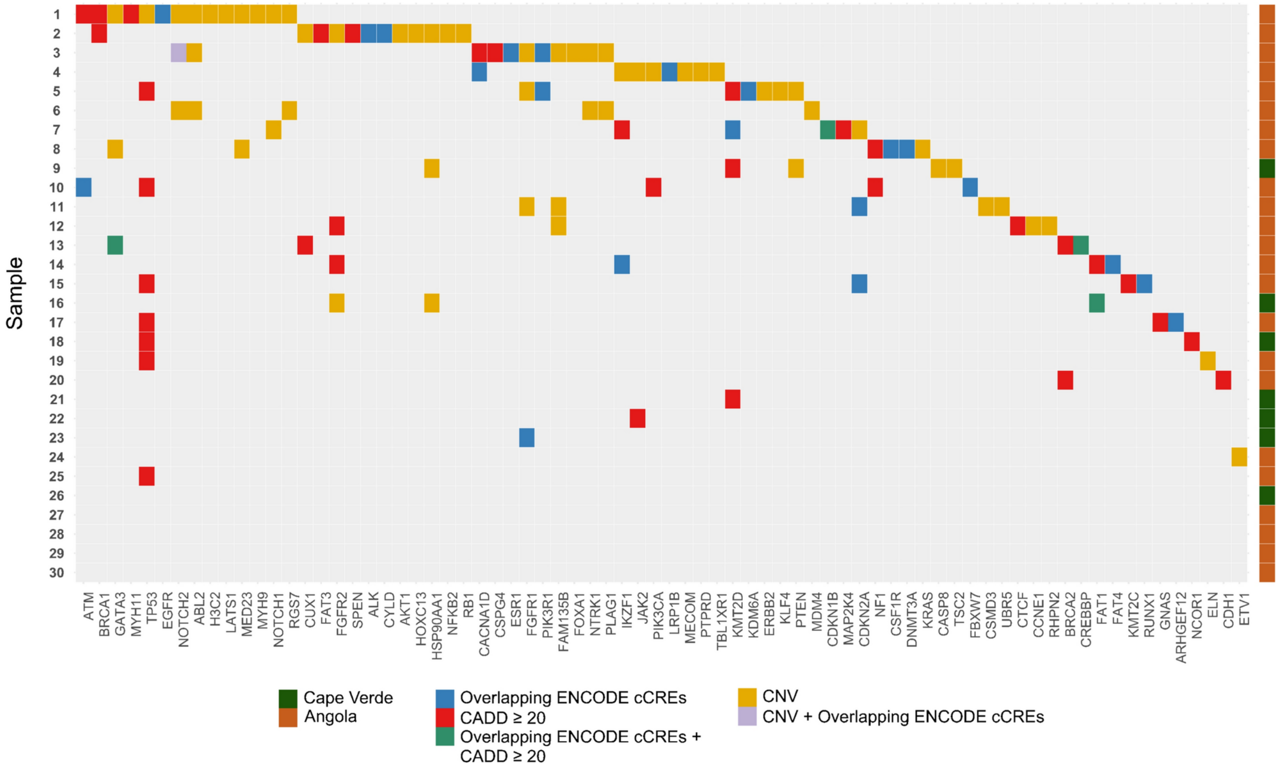

Fig. 4. Heatmap illustrating the somatic hits per TNBC African sample for potentially functional impacting coding and regulatory, as well as CNVs in breast cancer driver genes.

indicated a higher burden of somatic mutations than in the TCGA TNBC EA and AA cohorts. Of note, applying a robust deleteriousness predictive tool for the coding mutations, it was possible to infer that 17% of the somatic mutations have probable coding functional impact, and most of these (79%) were not yet described in the COSMIC database. The design of our work, by including UTRs sequencing together with the WES, allowed us for the first time to have a careful investigation of potential regulatory mutations, as UTRs are rich in this kind of functional impactful mutations. We observed that 20% of the somatic mutations overlap cCREs and may play a role in regulating the expression of genes. Some of these coding and regulatory impactful somatic mutations are located in known BC- or all cancer-driver genes. This observation empowers our work when trying to disentangle between potential driver/probable passenger mutations. With the somatic functional (coding and regulatory) prediction and CNV analyses, we could infer the putative driver mutations in almost all individuals of the African cohort. These impactful mutations in BC-driver genes play a role in cancer mechanisms, as control of cell division and differentiation, kinase activity, transcription regulation, mammary gland development, and also response to radiation and related DNA repair mechanisms. These last signatures in response to radiation and related DNA repair mechanisms reinforces previous observations that these mechanisms are important in AA and Hispanic BC cohorts[46], and can be modulated to promote radiosensitisation in TNBC tumours[47]. Interestingly, radiation-triggered molecular mechanisms may contribute to TNBC progression, which is significant in a context where radiotherapy is recommended as a standard of care[48]. In particular, these BC-driver genes involved in response to radiation in the African cohort are *ATM, BRCA1, BRCA2, CREBBP, GATA3, NF1, TP53, DNMT3A, EGFR, FBXW7, PIK3R1*. Some of those genes play a role in the DNA damage response, in particular *ATM, BRCA1, BRCA2, TP53,* joined by *CDKN1B* and *CDKN2A*.

It is clear the *TP53* dominance in terms of TNBC driver genes when attending to the putative functional impact of the somatic mutations. The heterogeneity in the frequency that this gene is mutated between the published African-ancestry cohorts[15,16] is probably due to the still low sample sizes of these cohort. Also, as most of these works do not take into account the functional prediction, they are unable to underrate the passenger mutations also occurring in this gene.

Our results on known driver genes relied on the IntOGen database[49], which is biased towards non-African populations. But we gathered evidence on putative TNBC driver genes, that may play an important role in African TNBC cohorts. One is *TTN*, for which there is growing evidence of its high mutation burden in a pan-cancer context, a trend that is also observed in TNBC[50–53]. MutSig2CV[54] indicated *CEACAM7* as having a significantly enriched somatic mutation burden in the African TNBC cohort. This gene encodes a cell surface glycoprotein, is a member of the carcinoembryonic antigen family of proteins and is downregulated in colon and rectal cancer[40]. Rare inherited protein-truncating variants in this gene have been associated with increased breast cancer susceptibility in AA women[55]. MutSig and cCRE analyses pointed to *DEFB132*, a member of the defensin family said to play an important role in tumour immunity, by chemoattracting multiple types of immune cells, leading to chemokine and cytokine release in the tumour microenvironment that have a direct cytotoxic effect on tumour cells[41]. MutSig and cCRE analyses also revealed *COPZ2*, which per se was shown to not display tumour-suppressive activities, but it harbours microRNA (miR)-152, which is silenced in tumour cells concurrently with *COPZ2*, thus acting as a tumour suppressor in vitro and in vivo[42]. The regulatory mutations detected in the African cohort are in a promoter, very close to the miR-152. The downregulation of miR-152/*COPZ2* in tumour cells may serve as a potential marker for therapeutic targeting *COPZ* isoforms[42]. *GAS1*, which in the African cohort harbours mutations within cCREs, is a direct inhibitor of the cell cycle during the G0/S transition[56,57], and regulates apoptosis in a context-dependent manner[56,57], supporting its role as a tumour suppressor in cancer. Conflicting evidence, however, has shown that *GAS1* is involved in the stemness of mesenchymal-like BC stem cells (BCSC) through its association with *NOTCH4*, which is implicated in TNBC[58]. In a very recent study, *GAS1* was found to be specifically associated with a higher risk of TNBC, probably through the maintenance of a chemotherapy-resistant BCSC phenotypes via paracrine signalling mediated by hedgehog (Hh)-activated cancer associated fibroblasts[59]. *GAS1* is an important co-receptor of Hh signalling[59].

Our overall results underscore the importance of enlarging cancer omics screenings for the African ancestry, as it potentiates the discovery of novel somatic mutations and associated cancer-related pathways. Obviously, further research is needed to ascertain the role of these new candidate functional mutations and driver genes in TNBC pathogenesis, and their suitability as drug targets specifically for this cancer type. But given its bad prognosis, this research is essential to make a positive impact in clinical outcomes.

Material and methods

TNBC cohorts

Immunohistochemical records of Angolan and Cape Verdean BC patients admitted to the Angolan Institute of Cancer Control and Clínica Sagrada Esperança (Luanda, Angola), and Hospital Agostinho Neto (Praia, Cape Verde), respectively, between 2018 and 2023 were reviewed for selection of TNBC cases. Subsequently, cases with missing formalin-fixed paraffin-embedded (FFPE) specimens or those with only biopsied tumour material were also excluded. For the remaining cases, a histological section was taken for haematoxylin and eosin (H&E) staining, which was then carefully inspected by an experienced pathologist for confirmation and selection of tumour and adjacent histologically normal tissue areas. The Angola cases were from Luanda region (high rate of migration from other Angolan regions), while the Cape Verdean cases were from Santiago Island (270,000 inhabitants in 2021 census) where the capital Praia is located. As Cape Verde is a more admixed population than Angola, we tried at least avoiding including recently admixed individuals by confirming the origin of the four grandparents: five patients confirmed that the four grandparents being of Cape Verdean origin; two did not provide information. In the Angolan cohort, we acquired tumour and matched adjacent normal DNA for 27 samples, with only tumour DNA obtained for two other samples. In the Cape Verdean sample set, six samples yielded both tumour and normal DNA, while for one sample, only tumour genomic material was collected. This study was approved by the Angolan Ministry of Health and by the Cape Verdean ethics national committee that also waived the need of obtaining informed consent due to the retrospective nature of the study. All experiments were performed according to the Helsinki Declaration principles.

TCGA TNBC cases were selected after careful inspection of clinical data of patients included in the BC TCGA cohort[60]. Tumours with a negative status of ER, PR and HER2 receptors met our inclusion criteria, for a total of 116 cases. African- and European-American TCGA TNBC samples, henceforth designated by TCGA-AA and TCGA-EA, were identified by mining the ancestral components estimated in[61], and having > 70% of the respective ancestry, totalling 92 cases. Out of these, only 89 had available WES data, which were consequently integrated into this study. Within this final cohort, 24 cases were found to bear an African ancestral background (≥ 70% of African component in[61]), while 65 were classified as European (≥ 70% of African component).

DNA extraction and WES

Sequential (to the H&E slide, maintaining orientation) and 10 µm slide-mounted tissue sections were macrodissected, and nucleic acids were extracted using the Maxwell® RSC DNA FFPE Kit and Maxwell® RSC Instrument (Promega, Madison, WI, USA). The extracted DNA underwent quality control (QC) assessment resorting to Qubit™ 1X dsDNA HS Assay Kit and Qubit® 3.0 Fluorometer (Thermo Fisher Scientific, Waltham, MA, USA).

Target exome enrichment was carried out using Agilent SureSelect Human All-Exon V5 + UTRs and sequenced as 2 × 151 base pair paired-end reads on an Illumina HiSeq platform (Illumina, San Diego, CA, USA) according to manufacturer's protocol. The intended coverage was of 100X. WES + UTRs covers 74.66 Mb, while the library used in TCGA (Nimblegen SeqCap EZ Human Exome Library v3.0) only covers 64.23 Mb.

Raw data alignment

For African TNBC samples, raw paired-end reads underwent trimming using trimmomatic (v0.39)[62] and were subsequently aligned to the human reference genome GRCh38 from NCBI using the Burrows-Wheeler Aligner (BWA)-MEM algorithm (v 0.7.15)[63]. Aligned reads were converted and stored in the BAM format and sorted using samtools (v1.6)[64]. TCGA TNBC BAM files were retrieved from the Genomic Data Commons (GDC) Data Portal. Alignment metrics were evaluated using samtools, Alfred[65] and Bamdst (https://github.com/shiquan/bamdst) and all files successfully passed this QC step. Using the Genome Analysis Toolkit (GATK; v4.2.0.0)[66], we implemented several procedures following GATK best practices. These included identifying and marking duplicate reads with MarkDuplicates, rectifying base quality scores with the Base Quality Score Recalibration method, and verifying and correcting read mate-pair information with FixMateInformation. We also corrected invalid BAM files' MD tags using samtools (v1.6). After applying QC to our African cohorts, six Angolan samples were excluded from subsequent analysis due to low coverage.

Variant calling, filtering, and annotation

For somatic short variant calling, we used Mutect2 in both tumour with matched normal and tumour-only modes, restricted to each WES capture genomic region. This included all sites marked as germline and observed in a panel of normal tissue (PoN), which was constructed based on available normal samples for both sample groups following GATK best practices. Variants were submitted to FilterMutectCalls (part of GATK v4.2.0.0) using default parameters. After excluding sites flagged as "multiallelic", we applied stringent filtering criteria for variant exclusion in tumours: variants with a depth (DP) of less than 7 were excluded; for DP between 8 and 10, only sites with a variant allele fraction (VAF) of at least 30% were included; and for DP of 10 or greater, a VAF threshold of at least 20% was applied. The final somatic dataset was obtained by including variants passing the FilterMutectCalls QC step and by excluding those with a frequency above 0.1% in 1000 Genomes database[67] as these will be germline variants. The resulting variant list was annotated with ANNOVAR[68].

Calling of CNV

Segment-level copy number calls were generated utilizing the R package *PureCN*, with internal copy number normalization and segmentation[69]. Specifically designed for targeted short-read sequencing data, such as WES, this algorithm estimates copy number while adjusting for sample tumour purity and ploidy, irrespective of the availability of matched normal samples. Amplifications were identified if the segment's integer copy number was ≥ 6 for focal CNVs or ≥ 7 for non-focal CNVs; deletions were considered when the integer was inferior to a cutoff of 0.5.

Significantly mutated genes

MutSig2CV[54] was used to pinpoint significantly mutated genes in the African and African + TCGA-AA cohorts. This version calculates a *P*-value based on three metrics: 1) MutSigCV determines the *P*-value for observing the given quantity of non-silent mutations in the gene, given the background model determined by silent (and noncoding) mutations in the same gene and the neighbouring genes of covariate space that form its 'bagel'; 2) MutSigCL measures the significance of the positional clustering (mutation hotspots in the genes) of the mutations observed; 3) and MutSigFN measures the significance of the tendency for mutations to occur at positions that are highly evolutionarily conserved (using conservation as a proxy for probably functional impact).

Inferring functional impact of somatic mutations

The CADD v.1.7 tool was used to evaluate the potential deleteriousness of somatic variants, aiding in the prioritization of causal variants[70]. This tool incorporates diverse annotations such as conservation and functional data to compute a scaled C score, which rank variants according to their probability of being harmful, thereby offering a comprehensive assessment of variant impact[70]. We have implemented an integrative scaled C-score cutoff of 20, above which variants were categorised as deleterious, as suggested by the authors. In order to reduce redundancy in CADD results, we filtered the output by the Ensembl canonical transcripts.

To infer possible regulatory functions, we verified if the somatic mutations were located on the human 926,535 candidate cis-regulatory elements (cCREs) identified by ENCODE and reported as Supplementary Table 10 in their seminal work[71]. This database of cCREs includes: 1) active and poised cCRE-ELS, with high DNase and H3K27ac signals, and falling within 2,000 bp of an annotated transcription start site (with low relative H3K4me3 signal), partitioned in proximal ELS (pELS) and dELS; 2) active and poised cCRE-PLS, with high DNase signals and high H3K4me3 signals; 3) CCCTC-binding factor (CTCF)-only elements (CTCF or 11-zinc finger protein), with high DNase and CTCF signals but low signals for H3K4me3 and H3K27ac, which are candidates for insulators and looping functions in which CTCF participates. Other regulatory elements (ELS and PLS) can also be bound by CTCF, where this protein may also participate in those roles.

Enrichment analysis

The online tool g:Profiler[72] was used for functional enrichment analysis, to aid in the identification of relevant molecular mechanisms, biological processes and pathways where the mutated genes play a role. g:Profiler evaluates the functional enrichment of the input gene list by using the cumulative hypergeometric test, a well-proven method in the field. To mitigate false-positive findings, we used the g:Profiler default method of g:SCS for multiple testing correction. We only considered molecular pathways containing between 15 and 500 genes.

Information on driver genes

Information on the driver cancer genes was extracted from the IntOGen database[49]. This database provides a compendium of driver genes which appropriately reflects the consensus from seven driver identification methods applied to the somatic point mutations. Most of these identification methods focus on protein sequence, functional domains and 3D structure.

Data visualization

Data plotting was performed in RStudio version 2023.12.1 + 402 (R version 4.2.3). Package *ggplot2* was used for simple graphs. The oncoplots were obtained with *maftools*. The frequency proportions of CNV amplifications and deletions were visualized using the *GenVisR* package[73]. The Python module *venn* was also used.

Data availability

The new whole exome + UTR data obtained in this study can be accessed from the EGA repository (European Genome-Phenome Archive) with the identifier EGAC50000000486, under the title "WES in Angolan and Cape Verdean triple-negative breast cancer samples".

Received: 20 September 2024; Accepted: 17 March 2025
Published online: 25 March 2025

References

1. Bray, F. et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **74**, 229–263. https://doi.org/10.3322/caac.21834 (2024).
2. Azubuike, S. O., Muirhead, C., Hayes, L. & McNally, R. Rising global burden of breast cancer: the case of sub-Saharan Africa (with emphasis on Nigeria) and implications for regional development: a review. *World J. Surg. Oncol.* **16**, 63. https://doi.org/10.1186/s12957-018-1345-2 (2018).
3. Anyigba, C. A., Awandare, G. A. & Paemka, L. Breast cancer in sub-Saharan Africa: The current state and uncertain future. *Exp. Biol. Med. (Maywood)* **246**, 1377–1387. https://doi.org/10.1177/15353702211006047 (2021).
4. Joko-Fru, W. Y. et al. The evolving epidemic of breast cancer in sub-Saharan Africa: Results from the African Cancer Registry Network. *Int. J. Cancer* **147**, 2131–2141. https://doi.org/10.1002/ijc.33014 (2020).
5. Pace, L. E. & Shulman, L. N. Breast cancer in Sub-Saharan Africa: challenges and opportunities to reduce mortality. *Oncologist* **21**, 739–744. https://doi.org/10.1634/theoncologist.2015-0429 (2016).
6. Espina, C., McKenzie, F. & Dos-Santos-Silva, I. Delayed presentation and diagnosis of breast cancer in African women: a systematic review. *Ann. Epidemiol.* https://doi.org/10.1016/j.annepidem.2017.09.007 (2017).
7. Jedy-Agba, E., McCormack, V., Adebamowo, C. & Dos-Santos-Silva, I. Stage at diagnosis of breast cancer in sub-Saharan Africa: a systematic review and meta-analysis. *Lancet Glob. Health* **4**, e923–e935. https://doi.org/10.1016/S2214-109X(16)30259-5 (2016).
8. Birnbaum, J. K., Duggan, C., Anderson, B. O. & Etzioni, R. Early detection and treatment strategies for breast cancer in low-income and upper middle-income countries: a modelling study. *Lancet Glob. Health* **6**, e885–e893. https://doi.org/10.1016/S2214-109X(18)30257-2 (2018).
9. McCormack, V. et al. Breast cancer survival and survival gap apportionment in sub-Saharan Africa (ABC-DO): a prospective cohort study. *Lancet Glob. Health* **8**, e1203–e1212. https://doi.org/10.1016/S2214-109X(20)30261-8 (2020).
10. El Jaddaoui, I. et al. Cancer omics in Africa: Present and prospects. *Front. Oncol.* **10**, 606428. https://doi.org/10.3389/fonc.2020.606428 (2020).
11. Spratt, D. E. et al. Racial/ethnic disparities in genomic sequencing. *JAMA Oncol.* **2**, 1070–1074. https://doi.org/10.1001/jamaoncol.2016.1854 (2016).
12. Wright, N., Rida, P., Rakha, E., Agboola, A. & Aneja, R. Panoptic overview of triple-negative breast cancer in Nigeria: Current challenges and promising global initiatives. *J. Glob. Oncol.* **4**, 1–20. https://doi.org/10.1200/JGO.17.00116 (2018).
13. Yuan, J. et al. Integrated analysis of genetic ancestry and genomic alterations across cancers. *Cancer Cell* https://doi.org/10.1016/j.ccell.2018.08.019 (2018).
14. Pereira, L., Mutesa, L., Tindana, P. & Ramsay, M. African genetic diversity and adaptation inform a precision medicine agenda. *Nat. Rev. Genet.* https://doi.org/10.1038/s41576-020-00306-8 (2021).
15. Pitt, J. J. et al. Characterization of Nigerian breast cancer reveals prevalent homologous recombination deficiency and aggressive molecular features. *Nat. Commun.* **9**, 4181. https://doi.org/10.1038/s41467-018-06616-0 (2018).
16. Hercules, S. M. et al. Analysis of the genomic landscapes of Barbadian and Nigerian women with triple negative breast cancer. *Cancer Causes Control* **33**, 831–841. https://doi.org/10.1007/s10552-022-01574-x (2022).
17. Newman, L. A. Breast cancer disparities: high-risk breast cancer and African ancestry. *Surg. Oncol. Clin. N. Am.* **23**, 579–592. https://doi.org/10.1016/j.soc.2014.03.014 (2014).
18. Jiagge, E. et al. Breast cancer and African ancestry: lessons learned at the 10-year anniversary of the Ghana-Michigan research partnership and international breast registry. *J. Glob. Oncol.* **2**, 302–310. https://doi.org/10.1200/JGO.2015.002881 (2016).
19. Tan, P. H. et al. The 2019 World Health Organization classification of tumours of the breast. *Histopathology* **77**, 181–185. https://doi.org/10.1111/his.14091 (2020).
20. Senkus, E. et al. Primary breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* **26**(Suppl 5), v8-30. https://doi.org/10.1093/annonc/mdv298 (2015).
21. Foulkes, W. D., Smith, I. E. & Reis-Filho, J. S. Triple-negative breast cancer. *N. Engl. J. Med.* **363**, 1938–1948. https://doi.org/10.1056/NEJMra1001389 (2010).
22. Prat, A. et al. Clinical implications of the intrinsic molecular subtypes of breast cancer. *Breast* **24**(Suppl 2), S26-35. https://doi.org/10.1016/j.breast.2015.07.008 (2015).
23. Garrido-Castro, A. C., Lin, N. U. & Polyak, K. Insights into molecular classifications of triple-negative breast cancer: Improving patient selection for treatment. *Cancer Discov.* **9**, 176–198. https://doi.org/10.1158/2159-8290.CD-18-1177 (2019).
24. Nwagu, G. C., Bhattarai, S., Swahn, M., Ahmed, S. & Aneja, R. Prevalence and mortality of triple-negative breast cancer in West Africa: biologic and sociocultural factors. *JCO Glob. Oncol.* **7**, 1129–1140. https://doi.org/10.1200/GO.21.00082 (2021).
25. Bauer, K. R., Brown, M., Cress, R. D., Parise, C. A. & Caggiano, V. Descriptive analysis of estrogen receptor (ER)-negative, progesterone receptor (PR)-negative, and HER2-negative invasive breast cancer, the so-called triple-negative phenotype: a population-based study from the California cancer Registry. *Cancer* **109**, 1721–1728. https://doi.org/10.1002/cncr.22618 (2007).
26. Lund, M. J. et al. Race and triple negative threats to breast cancer survival: a population-based study in Atlanta, GA. *Breast Cancer Res. Treat.* **113**, 357–370. https://doi.org/10.1007/s10549-008-9926-3 (2009).
27. Amirikia, K. C., Mills, P., Bush, J. & Newman, L. A. Higher population-based incidence rates of triple-negative breast cancer among young African-American women : Implications for breast cancer screening recommendations. *Cancer* **117**, 2747–2753. https://doi.org/10.1002/cncr.25862 (2011).
28. Kohler, B. A. et al. Annual report to the nation on the status of cancer, 1975–2011, featuring incidence of breast cancer subtypes by race/ethnicity, poverty, and state. *J. Natl. Cancer Inst.* https://doi.org/10.1093/jnci/djv048 (2015).
29. Huo, D. et al. Population differences in breast cancer: survey in indigenous African women reveals over-representation of triple-negative breast cancer. *J. Clin. Oncol.* **27**, 4515–4521. https://doi.org/10.1200/JCO.2008.19.6873 (2009).
30. Patin, E. et al. Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. *Science* **356**, 543–546. https://doi.org/10.1126/science.aal1988 (2017).
31. Jiagge, E. et al. comparative analysis of breast cancer phenotypes in African American, White American, and West Versus East African patients: Correlation between African ancestry and triple-negative breast cancer. *Ann. Surg. Oncol.* **23**, 3843–3849. https://doi.org/10.1245/s10434-016-5420-z (2016).
32. Newman, L. A. & Kaljee, L. M. Health disparities and triple-negative breast cancer in African American women: A review. *JAMA Surg.* **152**, 485–493. https://doi.org/10.1001/jamasurg.2017.0005 (2017).
33. Hercules, S. M. et al. Triple-negative breast cancer prevalence in Africa: a systematic review and meta-analysis. *BMJ Open* **12**, e055735. https://doi.org/10.1136/bmjopen-2021-055735 (2022).
34. Adebamowo, S. N. et al. Implementation of genomics research in Africa: challenges and recommendations. *Glob. Health Action* **11**, 1419033. https://doi.org/10.1080/16549716.2017.1419033 (2018).
35. Beleza, S. et al. The admixture structure and genetic variation of the archipelago of Cape Verde and its implications for admixture mapping studies. *PLoS ONE* **7**, e51103. https://doi.org/10.1371/journal.pone.0051103 (2012).
36. Laurent, R. et al. A genetic and linguistic analysis of the admixture histories of the islands of Cabo Verde. *Elife* https://doi.org/10.7554/eLife.79827 (2023).

37. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315. https://doi.org/10.1038/ng.2892 (2014).
38. Consortium & E. P.,. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74. https://doi.org/10.1038/nature11247 (2012).
39. Peled, Y. et al. Titin mutation in familial restrictive cardiomyopathy. *Int. J. Cardiol.* **171**, 24–30. https://doi.org/10.1016/j.ijcard.2013.11.037 (2014).
40. Messick, C. A. et al. CEACAM-7: a predictive marker for rectal cancer recurrence. *Surgery* **147**, 713–719. https://doi.org/10.1016/j.surg.2009.10.056 (2010).
41. Adyns, L., Proost, P. & Struyf, S. Role of defensins in tumor biology. *Int. J. Mol. Sci.* https://doi.org/10.3390/ijms24065268 (2023).
42. Shtutman, M. et al. Tumor-specific silencing of COPZ2 gene encoding coatomer protein complex subunit zeta 2 renders tumor cells dependent on its paralogous gene COPZ1. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 12449–12454. https://doi.org/10.1073/pnas.1103842108 (2011).
43. Evdokiou, A. & Cowled, P. A. Tumor-suppressive activity of the growth arrest-specific gene GAS1 in human tumor cell lines. *Int. J. Cancer* **75**, 568–577 (1998).
44. Saleh, M. et al. Comparative analysis of triple-negative breast cancer transcriptomics of Kenyan, African American and Caucasian Women. *Transl. Oncol.* **14**, 101086. https://doi.org/10.1016/j.tranon.2021.101086 (2021).
45. Martini, R. et al. African ancestry-associated gene expression profiles in triple-negative breast cancer underlie altered tumor biology and clinical outcome in women of African descent. *Cancer Discov.* **12**, 2530–2551. https://doi.org/10.1158/2159-8290.CD-22-0138 (2022).
46. Dutta, P., Keung, M. Y., Wu, Y. & Vadgama, J. V. Genetic variants in African-American and Hispanic patients with breast cancer. *Oncol. Lett.* **25**, 51. https://doi.org/10.3892/ol.2022.13637 (2023).
47. Pesch, A. M., Pierce, L. J. & Speers, C. W. Modulating the radiation response for improved outcomes in breast cancer. *JCO Precis. Oncol.* https://doi.org/10.1200/po.20.00297 (2021).
48. Lin, Y. et al. Radiation exposure triggers the progression of triple negative breast cancer via stabilizing ZEB1. *Biomed. Pharmacother.* **107**, 1624–1630. https://doi.org/10.1016/j.biopha.2018.08.026 (2018).
49. Martinez-Jimenez, F. et al. A compendium of mutational cancer driver genes. *Nat. Rev. Cancer* **20**, 555–572. https://doi.org/10.1038/s41568-020-0290-x (2020).
50. Oh, J. H. et al. Spontaneous mutations in the single TTN gene represent high tumor mutation burden. *NPJ Genom. Med.* **5**, 33. https://doi.org/10.1038/s41525-019-0107-6 (2020).
51. Saravia, C. H. et al. Patterns of mutation enrichment in metastatic triple-negative breast cancer. *Clin. Med. Insights Oncol.* **13**, 1179554919868482. https://doi.org/10.1177/1179554919868482 (2019).
52. Ademuyiwa, F. O., Tao, Y., Luo, J., Weilbaecher, K. & Ma, C. X. Differences in the mutational landscape of triple-negative breast cancer in African Americans and Caucasians. *Breast Cancer Res. Treat.* **161**, 491–499. https://doi.org/10.1007/s10549-016-4062-y (2017).
53. Lips, E. H. et al. Next generation sequencing of triple negative breast cancer to find predictors for chemotherapy response. *Breast Cancer Res.* **17**, 134. https://doi.org/10.1186/s13058-015-0642-8 (2015).
54. Lawrence, M. S. et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501. https://doi.org/10.1038/nature12912 (2014).
55. Huskey, A. L. W., McNeely, I. & Merner, N. D. CEACAM gene family mutations associated with inherited breast cancer risk - a comparative oncology approach to discovery. *Front. Genet.* **12**, 702889. https://doi.org/10.3389/fgene.2021.702889 (2021).
56. Wang, H. et al. Growth arrest-specific gene 1 is downregulated and inhibits tumor growth in gastric cancer. *FEBS J.* **279**, 3652–3664. https://doi.org/10.1111/j.1742-4658.2012.08726.x (2012).
57. Martinelli, D. C. & Fan, C. M. The role of Gas1 in embryonic development and its implications for human disease. *Cell Cycle* **6**, 2650–2655. https://doi.org/10.4161/cc.6.21.4877 (2007).
58. Zhou, L. et al. NOTCH4 maintains quiescent mesenchymal-like breast cancer stem cells via transcriptionally activating SLUG and GAS1 in triple-negative breast cancer. *Theranostics* **10**, 2405–2421. https://doi.org/10.7150/thno.38875 (2020).
59. Smith-Byrne, K. et al. Identifying therapeutic targets for cancer among 2074 circulating proteins and risk of nine cancers. *Nat. Commun.* **15**, 3621. https://doi.org/10.1038/s41467-024-46834-3 (2024).
60. Cancer Genome Atlas. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70. https://doi.org/10.1038/nature11412 (2012).
61. Carrot-Zhang, J. et al. Comprehensive analysis of genetic ancestry and its molecular correlates in cancer. *Cancer Cell* **37**(639–654), e636. https://doi.org/10.1016/j.ccell.2020.04.012 (2020).
62. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170 (2014).
63. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324 (2009).
64. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* https://doi.org/10.1093/gigascience/giab008 (2021).
65. Rausch, T., Hsi-Yang Fritz, M., Korbel, J. O. & Benes, V. Alfred: interactive multi-sample BAM alignment statistics, feature counting and feature annotation for long- and short-read sequencing. *Bioinformatics* **35**, 2489–2491. https://doi.org/10.1093/bioinformatics/bty1007 (2019).
66. van der Auwera, G. & O'Connor, B. D. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra* (O'Reilly Media Incorporated, 2020).
67. Genomes Project et al. A global reference for human genetic variation. *Nature* **526**, 68–74. https://doi.org/10.1038/nature15393 (2015).
68. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164. https://doi.org/10.1093/nar/gkq603 (2010).
69. Riester, M. et al. PureCN: copy number calling and SNV classification using targeted short read sequencing. *Source Code Biol. Med.* **11**, 13. https://doi.org/10.1186/s13029-016-0060-z (2016).
70. Schubach, M., Maass, T., Nazaretyan, L., Roner, S. & Kircher, M. CADD v1.7: using protein language models, regulatory CNNs and other nucleotide-level scores to improve genome-wide variant predictions. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkad989 (2024).
71. Consortium et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710. https://doi.org/10.1038/s41586-020-2493-4 (2020).
72. Kolberg, L. et al. g:Profiler-interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). *Nucleic Acids Res.* **51**, W207–W212. https://doi.org/10.1093/nar/gkad347 (2023).
73. Skidmore, Z. L. et al. GenVisR: Genomic visualizations in R. *Bioinformatics* **32**, 3012–3014. https://doi.org/10.1093/bioinformatics/btw325 (2016).

Acknowledgements

We thank all the patients that kindly provided the biological material for this work. RJP was funded by a PhD grant (SFRH/BD/145217/2019) from FCT—Fundação para a Ciência e a Tecnologia, I.P., and a research grant

(Bolsas LPCC-NRN 2024) from Liga Portuguesa Contra o Cancro—Núcleo Regional do Norte. DF was funded by a PhD grant (2020.09384.BD) from FCT. The work was funded regional funds through FCT, under the research project PTDC/BIA-MOL/3986/2021.

Author contributions

L.L.S. and L.P. conceived the study. R.J.P. did the lab work and performed the bioinformatics analyses, under L.P. supervision. D.L. processed the tumor blocks for DNA extraction and HE staining for guidance. P.S., F.M., P.B., C.B. and V.C. selected and provided the tumor samples and compiled the associated metadata. C.L. provided pathological second opinion and selected between normal and tumour tissues. R.J.P. and L.P. wrote the manuscript with inputs from all co-authors. All authors reviewed and approved the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-94707-6.

Correspondence and requests for materials should be addressed to L.P.

Reprints and permissions information is available at www.nature.com/reprints.

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